ROYAUME DU MAROC
INSTITUT AGRONOMIQUE ET VETERINAIRE HASSAN II

المملكة المغربية
معهد الحسن الثاني للزراعة والبيطرة

Projet de fin d'étude

Pour l'obtention de diplôme d'Ingénieur d'État en Agronomie

Option : Ingénierie Data Science en Agriculture

Utilisation du *Machine Learning* pour la cartographie de la répartition annuelle des espèces végétales sur le Sahel selon les facteurs environnementaux.

Présenté et soutenu publiquement par

GRIMAJ Meryem

Pr. DEHHAOUI Mohammed	IAV Hassan II	Président
Pr. BENSIALI Saloua	IAV Hassan II	Rapporteuse
Dr. TAUGOURDEAU Simon	CIRAD	Rapporteur
Pr. BEN GHABRIT Salmane	IAV Hassan II	Examinateur

Juillet 2024

Adresse : Madinat Al Irfane, B.P. 6202. Rabat – Maroc
Tél : (00 212) 0537 77 17 58/59
Fax : (00 212) 0537 77 58 45
Site web: http://www.iav.ac.ma

العنوان: ص. ب 6202 الرباط المعاهد الرباط - المغرب
الهاتف: 59 / 58 17 77 0537 (00 212)
الفاكس: (00 212) 0537 77 58 45
موقع الأنترنت: http://www.iav.ac.ma

DEDICACES

Je dédie ce travail à ma chère famille, dont le soutien indéfectible et l'amour constant ont été ma source d'inspiration et de motivation. À mes parents, pour leurs encouragements et leur confiance inébranlable en moi, ainsi qu'à mon frère et ma sœur, pour leur soutien et leurs conseils avisés.

Je tiens également à dédier ce mémoire à mes amis proches, dont l'amitié et l'encouragement m'ont accompagné tout au long de ce parcours.

Enfin, je dédie ce travail à tous mes professeurs, qui m'ont guidé, enseigné et inspiré tout au long de mes études.

REMERCIMENTS

Je tiens à exprimer ma profonde gratitude à tous ceux qui ont contribué à la réalisation de ce travail.

Tout d'abord, je remercie vivement **Professeur DEHHAOUI Mohammed** pour avoir accepté de présider cette soutenance. Votre présence et votre expertise sont d'une grande valeur pour ce projet.

Je souhaite également exprimer ma reconnaissance à **M. TAUGOURDEAU Simon**, mon encadrant au sein du CIRAD. Je vous remercie pour vos conseils avisés, votre soutien constant ainsi que votre disponibilité pour la réussite de ce projet. Merci infiniment pour votre encadrement exceptionnel.

Mes sincères remerciements vont également au **Professeur BENSIALI Saloua**, mon encadrante à l'IAV Hassan II. Je vous remercie professeure pour votre soutien indéfectible, vos encouragements et vos précieux conseils tout au long de cette étude aboutissant à ce travail. Je vous en suis profondément reconnaissante.

Je remercie également **Professeur BEN GHABRIT Salmane** pour l'honneur qu'il m'a fait en acceptant d'évaluer mon travail. Votre participation en tant que membre du jury est un honneur, et je vous remercie sincèrement pour vos précieux retours.

Je tiens également à exprimer ma gratitude à l'équipe du CIRAD, en particulier l'équipe **SELMET**, pour m'avoir accueilli et aidé durant mon séjour en France. Votre soutien a été inestimable et a grandement contribué à la réussite de ce projet. Un grand merci à **Madame BAZAN Samantha** pour m'avoir accueillie sur le site de SELMET et à toute l'unité SELMET pour avoir financé mon stage.

Je remercie aussi l'équipe HPC Marwan de CNRST, pour son suivi technique lors de l'utilisation du servie de calcul haute performance, sa disponibilité et tous ses efforts fournis pour me faciliter l’accès.

Je remercie aussi les professeurs du département de Statistiques et Informatique Appliquées pour la qualité de la formation proposée et pour leur encadrement. Vos enseignements et votre soutien ont été essentiels dans ma préparation et ma réussite.

Merci à vous tous pour votre soutien et votre confiance.

RESUME

La région du Sahel, une bande semi-aride située entre le désert du Sahara et la savane soudanienne, est soumise à des conditions environnementales extrêmes. Elle se caractérise par des précipitations irrégulières et souvent insuffisantes, entrainant des périodes de sècheresse récurrentes et une dégradation significative des sols. Ces conditions exacerbent les problèmes existants tels que l'insécurité alimentaire, la perte de biodiversité et la dégradation des terres, rendant la région particulièrement vulnérable aux changements climatiques. Or, la végétation sahélienne joue un rôle crucial dans la stabilisation des sols, la régulation du climat local et le soutien des moyens de subsistance des populations locales. Ainsi, prédire la répartition des espèces végétales dans une région aussi vaste et variée est essentiel. Pour répondre à cette problématique, notre étude vise à atteindre trois objectifs principaux. Notamment, prédire la répartition spatiale et temporelle des espèces végétales dominantes dans le Sahel, tester différents modèles de *Machine Learning* pour la spatialisation des données de présence des espèces végétales, et trouver le modèle de distribution adéquat pour chaque type de végétation (herbacées et arborées). Des cartes de distribution annuelles exploitables seront générées. Les données utilisées proviennent principalement de la base de données FLOTROP (Taugourdeau et al, 2019), contenant huit herbacées et deux arbres. Nous l'avons enrichie de variables environnementales telles que les précipitations, la température et le type de sol ainsi que l'indice de pluviométrie. Les techniques de modélisation incluent des méthodes de *Machine Learning* tels que les forêts aléatoires, MaxEnt, SVM, GLM, GAM et CNN, avec une validation des résultats à partir de données de l'herbier disponible au CIRAD. La méthodologie repose sur la collecte, le prétraitement et l'analyse de données environnementales et d'occurrence des espèces végétales pour modéliser leur distribution. Les résultats montrent que les forêts aléatoires (RF) sont particulièrement adaptées pour modéliser la distribution des herbacées, tandis que les réseaux de neurones convolutifs (CNN) sont plus performants pour les espèces arborées. Cependant, les deux modèles peuvent être utilisés pour les deux types d'espèces. Les modèles prédictifs ont permis de générer des cartes de distribution annuelles détaillées et exploitables, indiquant les zones où les conditions environnementales sont favorables ou défavorables à la présence des espèces. Ces cartes sont essentielles pour la gestion et la conservation des ressources végétales dans le Sahel.

Mots clés : SDM, *Machine Learning*, Classification, Présence-absence, SAHEL, Cartographie, CNN, RF.

ABSTRACT

The Sahel region, a semi-arid strip located between the Sahara Desert and the Sudanian savanna, is subject to extreme environmental conditions. It is characterized by irregular and often insufficient precipitation, leading to recurrent drought periods and significant soil degradation. These conditions exacerbate existing problems such as food insecurity, biodiversity loss, and land degradation, making the region particularly vulnerable to climate change. Sahelian vegetation plays a crucial role in soil stabilization, local climate regulation, and supporting the livelihoods of local populations Therefore, predicting the distribution of plant species in such a vast and varied region is essential. To address this issue, our study aims to achieve three main objectives. Specifically, to predict the spatial and temporal distribution of dominant plant species in the Sahel, to test different Machine Learning models for spatializing the presence data of plant species, and to find the appropriate distribution model for each type of vegetation (herbaceous and tree species). Additionally, usable annual distribution maps will be generated. The data used come mainly from the FLOTROP database (Taugourdeau et al., 2019). It contains eight herbaceous species and two trees. We enriched it with environmental variables such as precipitation, temperature, soil type, and rainfall index. The modeling techniques include machine learning methods such as random forests, MaxEnt, SVM, GLM, GAM, and CNN, with validation of results using herbarium data available from CIRAD. The methodology involves the collection, preprocessing, and analysis of environmental and occurrence data to model the distribution of plant species. The results show that Random Forests (RF) are particularly well-suited for modeling the distribution of herbaceous species, while Convolutional Neural Networks (CNN) perform better for tree species. However, either model can be used for both types of species. The predictive models have enabled the generation of detailed and usable annual distribution maps, indicating areas where environmental conditions are favorable or unfavorable for species presence. These maps are essential for the management and conservation of plant resources in the Sahel.

Keywords: SDM, *Machine Learning*, Classification, Presence-absence, SAHEL, Mapping, CNN, RF.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

- ACP : *Principal Component Analysis (PCA)*
- AFCM : *Multiple Correspondence Analysis (MCA)*
- AUC : *Area Under the Curve*
- Bio1 : *Annual Mean Temperature*
- Bio2 : *Mean Diurnal Range (Mean of monthly (max temp - min temp))*
- Bio3 : *Isothermality (Bio2/Bio7) (*100)*
- Bio4 : *Temperature Seasonality (standard deviation *100)*
- Bio5 : *Max Temperature of Warmest Month*
- Bio6 : *Min Temperature of Coldest Month*
- Bio7 : *Temperature Annual Range (Bio5-Bio6)*
- Bio8 : *Mean Temperature of Wettest Quarter*
- Bio9 : *Mean Temperature of Driest Quarter*
- Bio10 : *Mean Temperature of Warmest Quarter*
- Bio11 : *Mean Temperature of Coldest Quarter*
- Bio12 : *Annual Precipitation*
- Bio13 : *Precipitation of Wettest Month*
- Bio14 : *Precipitation of Driest Month*
- Bio15 : *Precipitation Seasonality (Coefficient of Variation)*
- Bio16 : *Precipitation of Wettest Quarter*
- Bio17 : *Precipitation of Driest Quarter*
- Bio18 : *Precipitation of Warmest Quarter*
- Bio19 : *Precipitation of Coldest Quarter*

- CIRAD : Centre de coopération internationale en recherche agronomique pour le développement
- CNN : *Convolutional Neural Network*
- FN : *False Negatives*
- FP : *False Positives*
- GAM : *Generalized Additive Model*
- GLM : *Generalized Linear Model*
- Lat : Latitude
- Long : Longitude
- Maxent : *Maximum Entropy Modeling*
- RF : *Random Forest*
- SDM : *Species Distribution Models*
- SVM : *Support Vector Machine*
- TN : *True Negatives*
- TP : *True Positives*

INTRODUCTION GENERALE

1. Mise en contexte

Les écosystèmes sahéliens, situés dans une zone de transition climatique entre le désert du Sahara et la savane soudanienne, sont soumis à des conditions environnementales extrêmes influençant fortement la distribution et la dynamique de végétation (Nicholson, 2013). En effet, le Sahel est connu par ses précipitations irrégulières et souvent insuffisantes, entrainant des périodes de sècheresse récurrentes et des dégradations importantes des sols. Toutes ces conditions font que cette région demeure très vulnérable aux changements climatiques, exacerbant les défis existants tels que l'insécurité alimentaire, la perte de biodiversité et la dégradation des terres.

La stabilisation des sols, la régulation du climat local et le soutien des moyens de subsistance des populations locales sont très liés à la végétation sahélienne. Cependant, les connaissances sur la distribution actuelle des espèces végétales dans cette région restent limitées en raison du manque de données de terrain systématiques et continues (Hijmans et al.,2005). En effet, la cartographie précise de la distribution des espèces végétales est essentielle pour la conservation de la biodiversité, la gestion durable des ressources naturelles et l'élaboration de stratégies d'adaptation au changement climatique.

Mais en raison de l'indisponibilité de certaines données sur le Sahel, les scientifiques ont tendance à utiliser les modèles de distribution des espèces (SDM) pour prédire leur répartition spatiale et temporelle en réponse aux conditions environnementales (Hutchinson, 1957). Cette approche repose sur des modèles statistiques utilisant des données d'occurrence d'espèces et des variables environnementales disponibles pour extrapoler leurs distributions. Il s'agit d'une technique particulièrement précieuse dans le contexte des changements climatiques globaux et de la perte d'habitats naturels, qui posent des menaces significatives à la biodiversité. En effet, cette technique ne se limite pas seulement à décrire la répartition des espèces, mais elle peut également interpoler des distributions d'espèces là où les données directes sont limitées ou absentes, fournissant ainsi des cartes prédictives essentielles pour la gestion des ressources naturelle et la planification de la conservation. En revanche, elle permet de prédire les changements futurs en fonction des scénarios climatiques (Thuiller et al.,2005).

Cela est particulièrement pertinent dans le Sahel, ou les effets du changement climatique ce manifestent par une variation des régimes de précipitations et des températures. Par exemple, l'augmentation des températures et la variabilité des précipitations peuvent affecter la phénologie des plantes, leur capacité de survie et, par conséquent, leur distribution géographique (Foley et al.,2005).

Comprendre les divers niveaux et dimensions de la biodiversité est essentiel pour élaborer des modèles prédictifs précis de la distribution des espèces végétales, particulièrement dans des zones écologiquement sensibles comme le Sahel (Tilman et al., 2006; Bai et al., 2004; Hooper et al., 2005).

2. Problématique

La biodiversité, concept central en écologie, joue un rôle crucial dans la stabilité et la résilience des écosystèmes. Prédire avec précision la distribution des espèces végétales dans le Sahel est un enjeu majeur de cette étude.

Pour atteindre cet objectif plusieurs défis se posent liés essentiellement au manque de données. En effet, les données de localisation directe pour de nombreuses espèces sont incomplètes ou absentes, ce qui est le cas de la base de données FLOTROP (Taugourdeau et al,2019) qui ne contient que les données de présences des espèces végétales pour quelques positions, rendant difficile l'évaluation précise de leurs distributions. En plus, le Sahel est soumis à une grande variabilité climatique et environnementale, compliquant la modélisation des niches écologiques des espèces.

Ces variabilités climatiques et les modifications des régimes de précipitations et des températures dues au changement climatique peuvent affecter les habitats et la distribution des espèces nécessitant des outils robustes pour anticiper ces impacts.

3. Objectifs

Les principaux objectifs de cette étude sont :

- Prédiction de la répartition spatiale et temporelle des espèces végétales dominantes dans le Sahel.

- Test des différents modèles de *Machine Learning* pour la spatialisation des données de présences des espèces végétales.

- Trouver le modèle de distribution adéquat pour chaque type de végétation (herbacées et arborées).

4. Hypothèses du projet

4.1 Hypothèses admises du projet

Pour mener à bien cette étude en tenant compte de la durée prévue et des données disponibles, nous avons établi plusieurs hypothèses fondamentales sur lesquelles nous baserons notre analyse :

- Absence d'interaction entre les espèces végétales : Cela signifie que chaque espèce est modélisée indépendamment, sans tenir compte des effets de la compétition ou de toute autre forme d'interaction écologique. Cette simplification permet de concentrer l'analyser sur les facteurs environnementaux influençant la présence des espèces, bien que cela puisse omettre certains aspects de la dynamique écologique réelle.
- Traitement des absences : Pour chaque relevé (ensemble d'enregistrements collectés pour des espèces données), nous supposons que toute espèce non mentionnée comme présente est absente. Cela signifie que si une espèce n'est pas observée dans un relevé spécifique, elle est enregistrée comme absente. Cette hypothèse repose sur l'idée que les relevées sont exhaustifs et que toute absence d'enregistrement indique réellement l'absence de l'espèce.

4.2 Hypothèses à tester du projet

Dans le cadre de notre projet visant à modéliser et prédire la distribution des espèces végétales dans la région du Sahel, nous émettons l'hypothèse que les méthodes de *Machine Learning* peuvent être efficacement utilisées pour spatialiser les présences des espèces végétales que ça soit pour les herbacées ou les arbres.

5. Organisation du document

Le document est structuré en quatre chapitres :

- Introduction générale : On commence par introduire le contexte de notre étude, les problématiques essentielles auxquelles nous allons répondre dans les étapes à suivre et les objectifs de notre étude.

- Chapitre1 : Étude bibliographique où nous allons initier toutes les notions importantes de notre étude, à savoir la biodiversité et ses différents axes, la région du Sahel sur laquelle nous allons travailler, sa géographie et ses conditions climatiques et environnementales en plus des descriptions de ses espèces végétales les plus dominantes et leurs adaptations aux conditions climatiques. Et nous allons finir par expliquer les concepts et les outils importants pour les SDM, leurs principes et les différents types de modèles ainsi que leurs applications.

- Chapitre2 : Matériels et méthodes. Ce chapitre décrit les données utilisées, les méthodes de collecte et de prétraitement ainsi que la méthodologie détaillée de la modélisation.

- Chapitre3 : Résultats. Dans cette partie nous allons présenter les résultats des SDM, ainsi que ceux de leur validation par les données de l'herbier.

- Chapitre 4 : Discussion. Cette partie inclura des discussions d'une part, sur la qualité et la précision des données et leurs effets sur les prédictions, d'autre part, sur les autres méthodes utilisées surtout pour les partenaires du CIRAD. On abordera aussi le problème du surapprentissage auquel nous avons été confrontés et nous finirons par une analyse des résultats obtenues.

- Le document se termine par une conclusion résumant les principaux résultats de l'étude et des perspectives pour les recherches futures

CHAPITRE I : Étude Bibliographique

Introduction

Ce chapitre explore les concepts fondamentaux nécessaires à la compréhension et à la modélisation de la distribution des espèces végétales dans le Sahel. Nous débuterons par une définition détaillée de la biodiversité, ses niveaux, ses dimensions spatiales et temporelles, ainsi que les diverses mesures utilisées pour évaluer cette biodiversité. Ensuite, nous présenterons une vue d'ensemble de la région du Sahel, en abordant ses caractéristiques géographiques, écologiques, et les défis environnementaux auxquels elle est confrontée. Enfin, nous introduirons les modèles de distribution des espèces (SDM), en détaillant leur définition, les principes sous-jacents, et les types de modèles utilisés dans ce domaine.

Partie 1 : Concepts liés à la biodiversité

Pour mieux comprendre et prédire la distribution des espèces végétales dans le Sahel, il est essentiel de maitriser les concepts fondamentaux de la biodiversité. Cette section explore les notions clés de cette dernière, telles que son historique, ses définitions, ses niveaux, ses dimensions spatiales et temporelles. Ces concepts sont intrinsèquement liés aux modèles de distribution des espèces (SDM), qui utilisent ces informations pour modéliser et prévoir la répartition des espèces en fonction des variables environnementales. Une bonne compréhension de la biodiversité permet en effet d’améliorer la précision et la robustesse des SDM, essentiels pour la conservation des espèces et la gestion des écosystèmes dans des zones écologiquement sensibles comme le Sahel.

1.1 Définition de la biodiversité

Dans sa forme la plus simple, la biodiversité représente la vie sur terre. On peut définir donc la biodiversité comme étant la variété des espèces vivantes que renferme l'ensemble des écosystèmes terrestres et aquatiques, se rencontrant actuellement sur la planète (Blondel, 2006).

La biodiversité, comme la décrit la convention des Nations Unies sur la diversité biologique (CDB), est "La diversité au sein des espèces, entre les espèces et dans les écosystèmes, y compris les plantes, les plantes, les animaux, les bactéries et les champignons"

Le concept de biodiversité est récent. En 1988, Edward O. Wilson publie « *Biological diversity* » qui met en avant pour la première fois l'idée de diversité biologique.

Mais ce nouveau concept n'a vraiment pris son essor qu'avec la signature de la Convention sur la diversité biologique lors du Sommet de la Terre de Rio en 1992. Dans son Article 2, cette convention définit la biodiversité comme étant la « variabilité des organismes vivants de toute origine, y compris, entre autres, les écosystèmes terrestres, marins et autres écosystèmes aquatiques et les complexes écologiques dont ils font partie.

Cela comprend la diversité au sein des espèces, et entre les espèces et ainsi que celle des écosystèmes ». L'écologue Robert Barbault résume ainsi cette définition : c'est « la vie, dans ce qu'elle a de divers ».

Le terme biodiversité est défini par la variabilité des organismes vivants de toutes origines y compris, entre autres, les écosystèmes terrestres, marins et autres écosystèmes aquatiques et les complexes écologiques dont ils font partie (Gaston & Spicer, 2006).

La biodiversité est définie comme étant la nature utile, c'est-à-dire l'ensemble des espèces ou des gènes que l'homme utilise à son profit, qu'ils proviennent du milieu naturel ou de la domestication. Ce concept désigne la variété des formes de vie comprenant les plantes, les animaux et les micro-organismes, les gènes qu'ils contiennent et les écosystèmes qu'ils forment (Ramade, 2009).

Dans ce contexte, on se focalise sur la biodiversité végétale et plus exactement sur celle en relation avec les arbres et les herbacées en concordance avec la base de données FLOTROP.

1.2 Les niveaux de la biodiversité

Le rôle de la diversité biologique dans un écosystème s'apprécie à trois niveaux d'intégration (Bouterfas, 2020-2021):

La diversité intraspécifique concerne la variabilité génétique des populations. Héritage de l'histoire de l'espèce, elle constitue une richesse distribuée entre individus pour répondre aux changements de l'environnement

La diversité des espèces, vue sous l'angle de leurs fonctions écologiques au sein de l'écosystème. Il existe une grande variété de formes, de tailles, et de caractéristiques biologiques parmi les espèces. Mises en jeu individuellement ou par groupes au sein des réseaux trophiques, ces propriétés ont une influence sur la nature et l'importance des flux de matière et d'énergie au sein de l'écosystème. Les interactions entre espèces, considérées non seulement sous l'angle de la compétition mais également sous celui du mutualisme et des symbioses, apportent une contribution intégrée de la diversité biologique à la dynamique des écosystèmes.

La diversité des écosystèmes correspond à la variété et à la variabilité temporelle des habitats.

1.3 Les différentes dimensions de la biodiversité

Comprendre la biodiversité exige de la saisir à différentes échelles spatiales et temporelles (Bouterfas, 2020-2021).

Une dimension spatiale

La biodiversité, cette richesse du monde vivant, ne peut être comprise qu'en l'observant à différentes échelles. Trois échelles principales d'interprétation nous permettent de saisir la complexité de la biodiversité :

Diversité alpha : la diversité locale, à l'échelle d'une communauté.

Diversité bêta : la diversité entre plusieurs communautés.

Diversité gamma : la diversité à l'échelle d'un ensemble de communautés.

Une dimension temporelle

La biodiversité n'est pas statique, elle évolue constamment. Le patrimoine génétique des espèces se modifie, les écosystèmes se transforment sous l'influence du temps, des changements climatiques, des perturbations naturelles et des actions humaines. Cette évolution peut être prise en compte pour mesurer la biodiversité et définir les actions de conservation efficaces.

1.4 Les différentes mesures de la biodiversité

Les opinions divergent quant à la manière de mesurer la biodiversité. Il n'existe pas de mesure universelle, et celles qui sont utilisées dépendent des objectifs visés. Également, tous les aspects de la biodiversité devraient être évalués dans un système donné sur un plan théorique. Cependant, cette tâche est pratiquement irréalisable, et est nécessaire de se contenter d'une estimation approximative en se basant sur des indicateurs. Ces indicateurs peuvent porter sur la génétique, les espèces ou les populations, la structure de l'habitat, ou toute combinaison qui fournit une évaluation relative mais pertinente de la diversité biologique

(Bouterfas, 2020-2021).

La richesse spécifique (nombre d'espèces) :

C'est l'unité de mesure la plus courante, mais la simple addition d'espèces ne suffit pas. Il faut aussi tenir compte de leur abondance relative.

La diversité spécifique

C'est la richesse spécifique à qui on rajoute l'abondance relative des espèces dans un assemblage donné. Parmi ces indices on a l'indice de shannon-Weaver , indice de Simpson , indice de Mill … etc.

La diversité dans l'espace

Diversité alpha, beta, gamma.

La diversité taxonomique

La phylogénétique (espèces, genres, familles) : La diversité taxonomique prend en compte les liens de parenté entre les espèces, une communauté avec des espèces de genres différents est plus diversifiée qu'une communauté avec des espèces du même genre.

La diversité fonctionnelle : le rôle des espèces

La diversité fonctionnelle s'intéresse au rôle des espèces dans l'écosystème, on regroupe les espèces en groupes fonctionnels selon leurs traits communs. La redondance fonctionnelle est importante pour la stabilité de l'écosystème.

Partie 2 : La région du Sahel : Géographie, Écologie et Adaptation au changement climatique.

2.1 Géographie et Climat

Le Sahel est une bande semi-aride de terre située entre la forêt tropicale au sud et le désert aride au nord de l'Afrique, couvrant environ 3 millions km2 traversant plusieurs pays d'Afrique de l'Ouest et de l'Est. Cependant, la délimitation exacte des pays inclus dans le Sahel varie selon les auteurs et les études.

Selon plusieurs sources, le Sahel traverse principalement les pays suivants : Sénégal, Mauritanie, Mali, Burkina Faso, Niger, Nigeria, Tchad, Soudan et l'Érythrée. (Nicholson, 2013; Le Houérou, 1989; OCHA, 2014). Cette classification est largement acceptée, mais il existe des variations dans la définition exacte du Sahel en fonction des critères climatiques, géographiques et politiques.

En effet, pour certains auteurs, le Sahel est défini comme une région recevant entre 200 et 500 mm de précipitations annuelles, avec une nette diminution des précipitations du Sud vers le nord. Cette région se caractérise par une saison des pluies courte et variable et une saison sèche prolongée (Epule et al., 2017 ; Nicholso,2001).

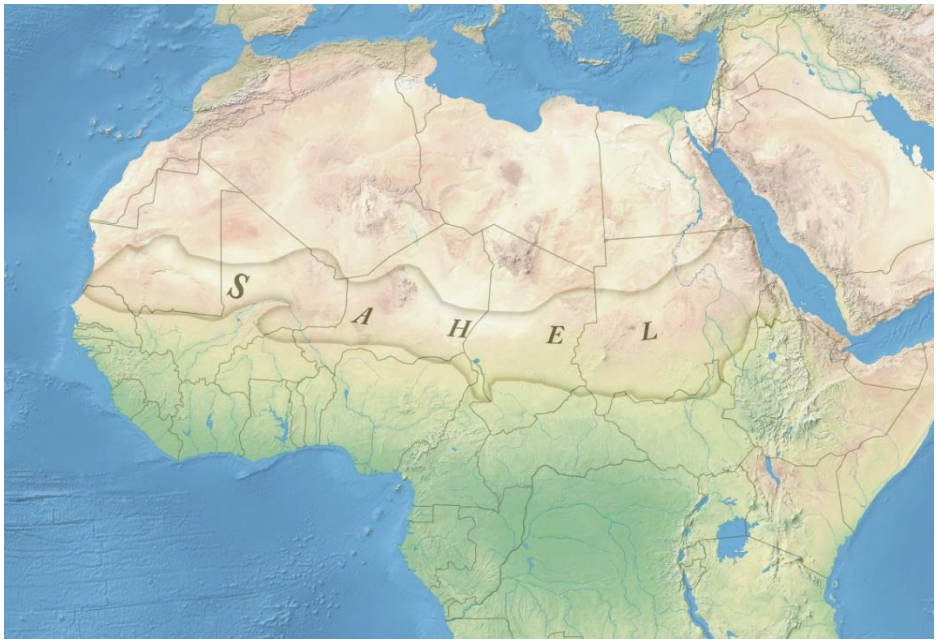


Figure 1: Carte de la Région du Sahel[1].

1 : Map_of_the_Sahel - Institute of Current World Affairs (icwa.org)

Pour d'autres, cette région est souvent décrite comme une zone de transition entre les savanes humides au sud et le désert aride au nord. La couverture végétale typique inclut des herbes basses et des arbustes épineux, avec des arbres tels que l'acacia et le baobab (Britannica,2024).

Certaines études incluent des régions périphériques en raison de similitudes socio-économiques et politiques Par exemple, le Sahel englobe des zones ou les conflits et les défis de développement sont communs, tels que les problèmes de sécurité alimentaire et de migration forcée. (Akinola et Ramontja, 2021).

Les frontières historiques et culturelles peuvent également influencer la délimitation du Sahel. Les régions avec des pratiques agricoles ou pastorales similaires sont souvent incluses pour des raisons de cohérence dans les études socio-économiques (Blench, 2001).

2.2 Écologie et biodiversité

La biodiversité du Sahel comprend une variété d'espèces adaptés aux conditions arides. La flore est dominée par des graminées et des arbustes épineux tels que *Cenchrus biflorus* et *Balanites aegyptiaca.*

La faune comprend des espèces migratrices, notamment des oiseaux qui utilisent les zones humides saisonnières du Sahel comme étape importante de leur migration entre l'Afrique et l'Eurasie. Les écosystèmes du Sahel sont essentiels pour le pâturage et l'agriculture nomade. (Anyamba & Tucker, 2005; Giannini et al., 2008)

2.3 Enjeux environnementaux et développement durable

Le Sahel est particulièrement vulnérable aux effets du changement climatique, tels que l'augmentation des températures et la variabilité des précipitations, ce qui exacerbe les problèmes de désertification et de dégradation des terres. Les initiatives de développement durable se concentrent sur l'amélioration de la résilience climatique, la gestion durable des terres et la restauration des écosystèmes.

Des projets tels que la Grande Muraille Verte vise à lutter contre la désertification en reboisant la région et en renforçant les pratiques agricoles durables. (Epule et al., 2017; World Bank, 2013).

2.4 Évolution de la distribution des espèces et du Climat dans le Sahel (1920 -2012)

2.4.1 Évolution du climat dans le Sahel entre 1920 et 2012

On peut diviser l'évolution climatique su Sahel sur cette période en plusieurs phases clés :

- **1920-1950** : Période relativement stable avec des précipitations modérées. Cette période a été marquée par une agriculture traditionnelle prospère, avec des communautés utilisant des techniques agricoles adaptées aux conditions locales.
- **1950- 1980** : Début d'une phase de sécheresse sévère, particulièrement marquée dans les années 1970, avec des précipitations réduites et des températures en augmentation. Les sécheresses des années 1970 ont provoquées une dégradation massive des terres, une famine généralisée et des migrations de populations vers des régions plus humides (Nicholson, 2001).
- **1980-2000** : Certaines années humides, mais la variabilité interannuelle des précipitations est restée élevée. Cette période a vu des efforts accrus pour comprendre et gérer les impacts climatiques, avec une attention accrue sur la reforestation et les techniques agricoles résilientes.
- **2000-2012** : Poursuite des fluctuations climatiques avec une tendance générale à l'augmentation des températures et à des épisodes de sécheresse récurrentes.

Des initiatives comme la Grande Muraille Verte ont été mises en œuvre pour combattre la désertification et restaurer les écosystèmes dégradés (Anyamba & Tucker, 2005).

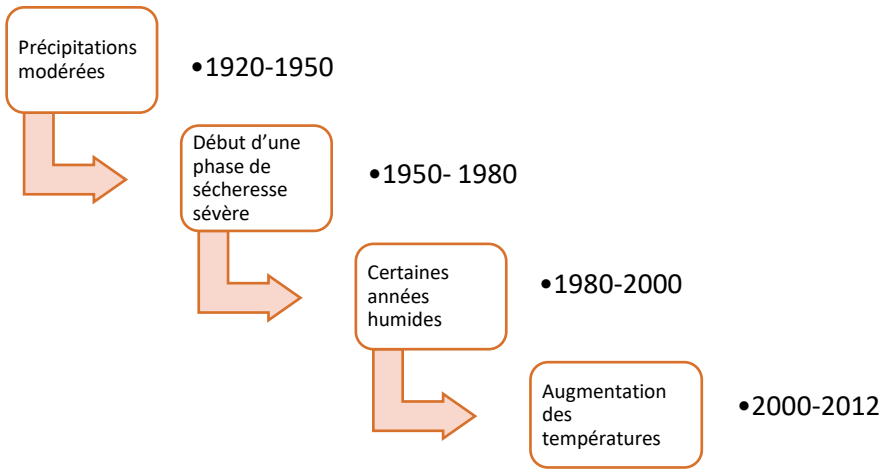


Figure 2: Évolution des conditions climatiques dans le Sahel entre 1920 et 2012

2.4.2 Évolution de la distribution des espèces végétales dans le Sahel entre 1920 et 2012

L'évolution de la distribution des espèces végétales dans le Sahel entre 1920 et 2012 a été largement influencée par les variations climatiques, les pratiques agricoles et les pressions anthropiques. Les espèces végétales les plus dominantes dans la région du Sahel comprennent principalement des graminées et des arbustes qui sont bien adaptés aux conditions climatiques arides et aux pratiques de pâturage intensif. Parmi les graminées dominantes, on retrouve *Cenchrus biflorus*, *Chloris prieurii*, *Diodella sarmentosa*, et *Zornia glochidiata*. Ces espèces herbacées annuelles sont essentielles pour le pâturage et jouent un rôle crucial dans l'écosystème de la savane sahélienne. (Gonzalez, 2001; Lykke, 1998).

En ce qui concerne les espèces ligneuses, les familles *Balanitaceae*, *Combretaceae*, et *Mimosaceae* sont particulièrement dominantes, fournissant des ressources vitales en bois et en fourrage pendant la saison sèche. (Ganaba et al., 1998; Gonzalez, 1997; Lykke, 2000) Intéressant nous aux dix espèces sur lesquels on voudrait travailler et qui représentent les 10 espèces les plus abondantes de notre base de données FLOTROP (Taugourdeau et al., 2019) :

- ***Cenchrus biflorus***: cette espèce a montré une grande résilience aux conditions climatiques arides du Sahel. Durant les périodes de sècheresse intense des années 1970 et 1980, cette espèce a maintenu sa présence grâce à sa capacité d'adaptation a des sols pauvres et a faibles humidités. Les années plus humides des décennies suivantes ont permis une légère expansion de sa distribution. Les variations de précipitations ont modifié la densité des populations de cette espèce, mais elle a globalement réussi à persister même dans les conditions les plus extrêmes. (Giannini, Biasutti, & Verstraete, 2008).

Figure 3 : Cenchrus biflorus[2]

2: Cenchrus biflorus Roxb. | Plants of the World Online | Kew Science 1

- ***Schoenefeldia gracilis :*** la distribution de cette espèce a fluctué en fonction des précipitations. Pendant les sècheresses sévères des années 1970, la plante a colonisée des zones désertifiées, remplaçant des espèces moins résistantes. Les années 1980 ont vu une réduction de sa couverture due à une légère augmentation des précipitations. En effet cette espèce prospère particulièrement bien en conditions de sècheresse prolongée à cause de son adaptation aux environnements arides. (Anyamba & Tucker, 2005).

Figure 4: Schoenefeldia gracilis[3]

- ***Aristida mutabilis*** : Cette graminée a profité des conditions arides, augmentant sa distribution pendant les décennies de sécheresse. Sa capacité à coloniser rapidement le sol dégradé a été un atout majeur pour sa survie et expansion. En plus que la capacité de cette espèce à s'établir dans des sols perturbés lui a permis de se répandre largement durant les périodes de faible précipitation (Nicholson, 2001).

3: West African Plants - A Photo Guide - Schoenefeldia gracilis Kunth (senckenberg.de)

Figure 5 : Aristida mutabilis[4]

- ***Dactyloctenium aegyptium*** : cette espèce est connue pour sa tolérance aux sols perturbés, a vu sa distribution augmenter avec l'intensification des activités agricoles et la dégradation des terres. Les sècheresses récurrentes ont favorisé son expansion dans des zones ou d'autres espèces ne pouvaient survivre (Mortimore & Adams, 2001).

Figure 6 : Dactyloctenium aegyptium[5]

- ***Eragrostis tremula :*** La distribution de cette espèce a été relativement stable, avec une légère augmentation observée durant les périodes de sécheresse prolongée. Cette graminée est bien adaptée aux conditions arides et a pu survivre malgré les fluctuations climatiques (Giannini et al., 2008).

4 : Aristida mutabilis (jircas.go.jp)

5: full.JPG (2048×1536) (mpg.de)

Figure 7 : Eragrostis tremula[6]

- ***Balanites aegyptiaca* :** Un arbre épineux, qui a réussi à maintenir sa distribution grâce à ses racines profondes capables d'accéder à des réserves d'eau souterraines. Sa distribution a été stable mais a légèrement augmentée dans les zones reboisées pour la lutte contre la désertification. En effet, cet arbre a bénéficié des projets de reforestation visant à lutter contre la désertification et à restaurer les écosystèmes dégradés (Nicholson, 2001).

Figure 8 : Balanites aegyptiaca[7]

6: Eragrostis tremula Hochst. ex Steud. [Espèce] - Images (plantnet-project.org)

7 : Balanites aegyptiaca - Plant Biodiversity of South-Western Morocco (teline.fr)

- ***Alysicarpus ovalifolius*** : Espèce annuelle fortement influencée par les précipitations saisonnières. Sa distribution a varié avec les fluctuations climatiques, augmentant durant les années plus humides et se contractant durant les périodes de sécheresse. Ainsi, les précipitations saisonnières influencent fortement sa présence, avec des populations fluctuantes selon les années (Anyamba & Tucker, 2005).

Figure 9 : Alysicarpus ovalifolius[8]

- ***Zornia glochidiata*** : Cette espèce a montré une grande résilience, maintenant une distribution stable malgré les changements climatiques. Ses graines peuvent rester dormantes pendant de longues périodes, lui permettant de survivre aux conditions difficiles (Giannini et al., 2008).

8: Alysicarpus ovalifolius - Photos - ISB: Atlas of Florida Plants - ISB: Atlas of Florida Plants (usf.edu)

Figure 10 : Zornia glochidiata[9]

- ***Combretum glutinosum*** : Cet arbuste a vu sa distribution réduire dans les zones de surexploitation mais a bénéficiée de la déforestation dans certains projets de conservation Les efforts de restauration écologiques ont aidées à stabiliser et parfois augmenter sa distribution (Nicholson, 2001).

Figure 11 : Combretum glutinosum[10]

9 : ZornigloM1_l.jpg (900×662) (jircas.go.jp)

10: West African Plants - A Photo Guide - Combretum glutinosum Perr. ex DC. (senckenberg.de)

- ***Andropogon gayanus*** : La distribution de cette graminée a fluctué avec les précipitations, augmentant durant les périodes humides. Elle est souvent utilisée dans les projets de restauration des terres grâce à sa capacité à améliorer la structure du sol (Mortimore & Adams, 2001).

Figure 12 : Andropogon gayanus[11]

11 : Hospedagem de Sites | Página não encontrada (bioseeds.com.br)

2.5 Type de sol

Le Sahel présente une grande diversité de types de sols influencées par les conditions climatiques, géologiques et les pratiques de gestion des terres. Parmi les principaux types de sol, on trouve les sols ferrugineux tropicaux qui sont typiques des régions semi-arides et arides du Sahel. Ces sols, souvent rouges à jaunes et riches en fer, présentent une structure granulaire et se développent principalement sous une végétation herbacée. Ils ont une faible teneur en matière organique et en nutriments, limitant leur fertilité, bien que leur texture sableuse facilite le drainage de l'eau. Ces sols sont généralement utilisés pour l'agriculture pluviale, notamment la culture de mil, de sorgho et d'arachides (FAO, 2004).

Les sols aérosols, très répandus dans le Sahel, particulièrement dans les régions de dunes et de déserts, sont des sols sableux. Ils possèdent une très faible capacité de rétention d'eau et sont pauvres en nutriments, ce qui les rend peu fertiles et sujettes à l'érosion éolienne. En raison de leur faible fertilité, ces sols sont souvent laissés en jachère ou utilisés pour des pâturages extensifs. Dans certaines zones, des techniques de gestion des sols sont employées pour améliorer leur capacité agricole (Tchakerian et Payne, 1997).

Un troisième type de sol, les sols vertisols qui sont des sols argileux, caractérisés par leur couleur noire ou gris foncé, qui se forment dans des conditions de drainage imparfait. Ces sols ont une haute capacité de rétention d'eau et sont fertiles, mais leur teneur élevée en argile peut causer des problèmes de drainage et de fissuration en saison sèche.

Et finalement, les sols hydromorphes, fréquents dans les zones de dépression et les plaines inondables du Sahel, qui sont saturés d'eau pendant une partie de l'année. Riche en matière organique et en nutriments, ces sols sont très fertiles. Cependant, leur saturation en eau peut limiter l'aération des racines des plantes (FAO, 2001).

Ces différents types de sols comme illustré dans la figure 13, montrent la diversité et la complexité des conditions pédologiques du Sahel, influençant la répartition des espèces végétale

Partie 3 : Modélisation de la distribution des espèces (SDM)

3.1 Définition des SDM

Les modèles de distribution des espèces sont des outils essentiels en écologie, biogéographie, et conservation de la biodiversité. Voici plusieurs définitions issues de la littérature scientifique:

- « La modélisation prédictive de la végétation peut être définie comme la prédiction de la distribution géographique de la composition de la végétation dans un paysage à partir de variables environnementales spatialisées » (Franklin, 1995)
- « Les SDM utilisent des techniques statistiques et des algorithmes d'apprentissage automatique pour modéliser la distribution spatiale des espèces à partir de données de présence et absence ainsi que de variables environnementales »

  (Guisan et Thuiller, 2005)
- « En général, la modélisation de niche écologique consiste à convertir les données brutes d'occurrence de l'espèce considérée en cartes de distribution géographique potentielle en indiquant sa présence probable ou son absence »

  (Guisan et Zimmermann, 2006)
- « Un SDM est un cadre analytique qui permet de relier les données d'occurrence des espèces à des variables explicatives environnementales, permettant ainsi de prédire les aires de répartition potentielles des espèces » (Phillips et Dudik, 2008)
- « Les SDM sont des outils statistiques et informatiques utilisés pour prédire la distribution géographique des espèces en fonction de variables environnementales et de données d'occurrence des espèces. » (Elith et Leathwick, 2009)
- « Les SDM, ou modèles de niche écologique, sont des méthodes qui relient les occurrences d'une espèce avec les conditions environnementales actuelles pour prédire leur distribution potentielle » (Franklin,2010)
- « Les SDM sont utilisés pour projeter les distributions futures des espèces sous différents scenarios de changement climatique, en se basant sur des modèles corrélatifs entre les observations des espèces et les variables environnementales » (Komori et al., 2020).

3.2 Notions et outils essentiels pour les SDM

3.2.1. Aire de répartition d'une espèce

L'aire de répartition d'une espèce peut être définie comme les limites géographiques ou cette espèce est trouvée, incluant les habitats ou elle peut vivre, se reproduire, et interagir avec d'autres organismes. (Elith et al.,2009). Cette aire est influencée par des facteurs abiotiques comme le climat et les sols, ainsi que par des facteurs biotiques comme la prédation et la compétition. (Franklin, 2010).

3.2.2. Niche écologique d'une espèce :

On peut définir la niche écologique d'une espèce comme le rôle que joue cette espèce dans son environnement, incluant ses interactions avec d'autres espèces et ses besoins en ressources. Elle englobe les conditions abiotiques et biotiques nécessaires à la survie et à la reproduction de l'espèce.

Elle est définie par les conditions environnementales et les ressources qui permettent à une espèce de maintenir des populations viables (Hutchinsin, 1957) et inclue à la fois les niches fondamentales, ou une espèce peut potentiellement survivre, et les niches réalisées, ou elle survit effectivement en présence de compétiteurs et de prédateurs. (Guisan et Zimmermann, 2006). La compréhension de la niche écologique est cruciale pour prédire comment les espèces répondent aux changements environnementaux. (Environmental Evidence, 2021).

3.3. Principe des modèles de distribution des espèces

Les modèles de distribution des espèces sont des outils essentiels en écologie et en biogéographie pour prédire la répartition géographique des espèces en fonction de variables environnementales et de données d'occurrence. En effet, ces modèles utilisent des données biologiques (présence, absence, abondance) et des données environnementales (climatiques, topographiques, édaphiques) pour établir des relations permettant de prédire les habitats favorables aux espèces (Elith et Leathwick,2009 ; Guisan et Zimmermann, 2006).

Les données nécessaires pour les SDM proviennent de diverses sources : relevées de terrain, observations opportunistes, bases de données de biodiversité, et technologies de télédétections. La quantité des données d'occurrence et environnementales influencent directement la précision des modèles. Les données doivent être géoréférencées. (Franklin, 2010 ; Komori et al.,2020).

La sélection des variables environnementales pertinentes est cruciale pour le succès des SDM. Les variables climatiques (température, précipitations), les caractéristiques du sol, l'altitude, et la végétation sont couramment utilisées. La sélection des variables doit être guidée par la connaissance écologique des espèces et les objectifs du modèle (Phillips et Dudik, 2008).

3.4. Les types de modèles

Jusqu'à présent, diffèrent types de modèles de distribution des espèces pour prédire et comprendre la distribution des espèces en fonction des variables environnementales existent. Ces modèles se divisent principalement en deux catégories : les modèles écologiques et les modèles d'apprentissage automatique.

3.4.1. Modèles écologiques

Ces modèles sont des méthodes statistiques et théoriques qui décrivent les relations entre les espèces et leur environnement. Ces modèles se basent sur des principes écologiques pour prédire où les espèces sont susceptibles de se trouver en fonction de diverses variables environnementales.

Parmi les principaux modèles écologiques on trouve :

- **Modèles linéaires généralisés (GLM)** : ces modèles sont une extension des modèles linéaires qui permettent de modéliser des réponses non normales (par exemple : présence/absence) en utilisant des distributions de la famille exponentielle. Ils sont utilisés pour comprendre la relation entre une espèce et les variables environnementales qui influencent sa distribution. (Elith et Leathwick, 2009)

- **Modèles additifs généralisés (GAM)** : Les GAM sont une extension des GLM qui permettent de modéliser des relations non linéaires entre les variables explicatives et la variable de réponse en utilisant des fonctions lisses, comme les splines. Cela permet une flexibilité accrue dans la modélisation des interactions complexes entre les espèces et leur environnement. (Guisan et Zimmermann, 2006).

- **Modèles de Régression Multiples (MARS)** : Les MARS utilisent des splines pour modéliser les relations non linéaires et les interactions entre les variables. Ils sont particulièrement efficaces pour capturer des relations complexes dans les données écologiques. (Friedman, 1991).

- **Modèles de Niche Écologique (ENM)** : Les ENM modélisent la niche écologique d'une espèce en se basant sur ses occurrences et les variables environnementales associées. Cela permet de prédire la distribution potentielle des espèces. (Peterson et al., 2011).

- **Modèles de Régression Logistique :** Ces modèles modélisent la probabilité de présence d'une espèce en fonction des variables explicatives. Ils sont utilisés pour les réponses binaires et offrent une compréhension des facteurs environnementaux influençant la présence des espèces. (Hosmer et Lemeshow, 2000).

- **Arbres de Classification et de Régression (CART)** : Ces modèles construisent des arbres de décision pour modéliser les relations entre les variables et la présence/absence des espèces. Ils sont simples à comprendre et interpréter. (Breiman et al., 1984).

- **Boosted Regression Trees (BRT)** : Les BRT combinent plusieurs arbres de décision en utilisant des techniques de boosting pour améliorer la précision des prédictions. Ils sont efficaces pour capturer les interactions complexes entre les variables. (Friedman, 2001).

3.4.2. Modèles de *Machine learning*

Les modèles d'apprentissage automatique utilisent des algorithmes avancés pour identifier des motifs complexes dans les données et faire des prédictions précises sur la distribution des espèces. Parmi les principaux modèles on a :

- MaxEnt (Maximum Entropy Modeling) : Ce modèle utilise principalement des données de présence pour estimer la distribution probable des espèces en maximisant l'entropie sous les contraintes fournies par les données environnementales. Maxent peut également être améliorée en utilisant des pseudo-absences pour affiner les prédictions. (Phillips et Dudik, 2008).

- **Forêts aléatoires (Random Forest)** : Cet algorithme d'apprentissage automatique construit une multitude d'arbres de décision sur des sous-échantillons des données et utilise la moyenne des prédictions pour améliorer la précision et éviter le surapprentissage. Les forêts aléatoires sont robustes et performantes pour les données écologiques complexes. (Cutler et al., 2007).

- **Support Vector Machines (SVM)** : Les SVM trouvent l'hyperplan optimal qui sépare les différentes classes de données (présence/absence) dans un espace de caractéristiques multidimensionnel. Efficaces pour des problèmes de classification binaire avec des données de haute dimension et des échantillons limités (Chang et Lin, 2011).

- **Réseaux de neurones profond (Deep Learning)** : Les réseaux de neurones convolutifs et autres architectures de deep learning peuvent modéliser des relations complexes et non linéaires entre les variables environnementales et la distribution des espèces. Ils sont utilisés pour analyser de grandes bases de données et capturer des relations subtiles entre les variables (Beery et al., 2020).

- Algorithme Génétique pour la Production de Règles (GARP) : GARP utilise des algorithmes génétiques pour créer des règles de distribution des espèces à partir des données environnementales. Il est particulièrement utile pour les données de présence uniquement. (Stockwell et Peters, 1999).

3.5 Applications des SDM

Les modèles de distribution des espèces ont des applications variées et cruciales dans plusieurs domaines de l'écologie, de la biogéographie et de la conservation. Voici quelques exemples des principales applications des SDM :

- **Conservation et gestion des espèces** : Les SDM sont utilisés pour identifier les habitats critiques et planifier des stratégies de conservation. En prédisant la distribution potentielle des espèces, les gestionnaires de la faune peuvent prioriser les zones pour la protection et la restauration des habitats naturels (Elith et Leathwick, 2009).

- **Études de biogéographie** : Les SDM permettent de comprendre les schémas de distribution des espèces à travers l'espace et le temps. Ils aident à reconstruire les distributions historiques des espèces et à prévoir les changements futurs sous différents scénarios climatiques. (Guisan et Thuiller, 2005)
- **Prévisions des impacts du changement climatique** : Les SDM sont utilisés pour prédire comment les distributions des espèces pourraient changer en réponse aux scenarios futurs de changement climatique. Ces prédictions aident à anticiper les déplacements d'espèces et les changements dans la composition des communautés écologiques. (Leta et al., 2013).
- **Gestion des espèces invasives** : Les SDM aident à prédire les zones susceptibles d'être envahies par des espèces exotiques nuisibles, permettant aux gestionnaires de prendre des mesures préventives pour contrôler ou éradiquer ces espèces avant qu'elles ne causent des dommages écologiques ou économiques importants. (Peterson et al., 2003).

Conclusion

L'étude bibliographique a fourni une base théorique solide pour comprendre la biodiversité et ses différentes dimensions, ainsi que le contexte environnemental spécifique de la région du Sahel. Nous avons également examiné les principes et les applications des modèles de distribution des espèces (SDM), ce qui nous permet de mieux appréhender les techniques et les outils nécessaires pour notre étude. Cette revue de la littérature nous a préparés à aborder les aspects méthodologiques de notre recherche, en nous assurant que nous comprenons les concepts clés et les défis associés à la modélisation de la distribution des espèces dans une région aussi complexe que le Sahel.

CHAPITRE II : MATÉRIEL ET MÉTHODES

Introduction

Dans ce chapitre, nous détaillons les méthodes et le matériel utilisés pour modéliser la distribution des espèces végétales dans le Sahel. Nous commençons par présenter la zone d'étude et les jeux de données utilisés, incluant les données d'occurrence des espèces et les variables environnementales. Ensuite, nous décrivons les logiciels et outils utilisés pour l'analyse, avant de détailler la méthodologie générale et spécifique adoptée pour collecter, prétraiter, et analyser les données, ainsi que pour modéliser la distribution des espèces.

2.1. Présentation de la zone d'étude

La zone d'étude pour notre projet est la région du Sahel, une bande de terre semi-aride. Elle s'étend approximativement de 12°N à 18°N de latitude et de 20 °W à 40 °E de longitude.

En raison de l'indisponibilité des données pour l'Érythrée dans notre base de données FLOTROP, nous nous limitons à la partie Est du Soudan. Ainsi, notre zone d'étude se situera approximativement entre 12°N et 18°N de latitude, et entre -17°W et 36°E de longitude, comme indiqué par la boîte rouge dans la figure 14.

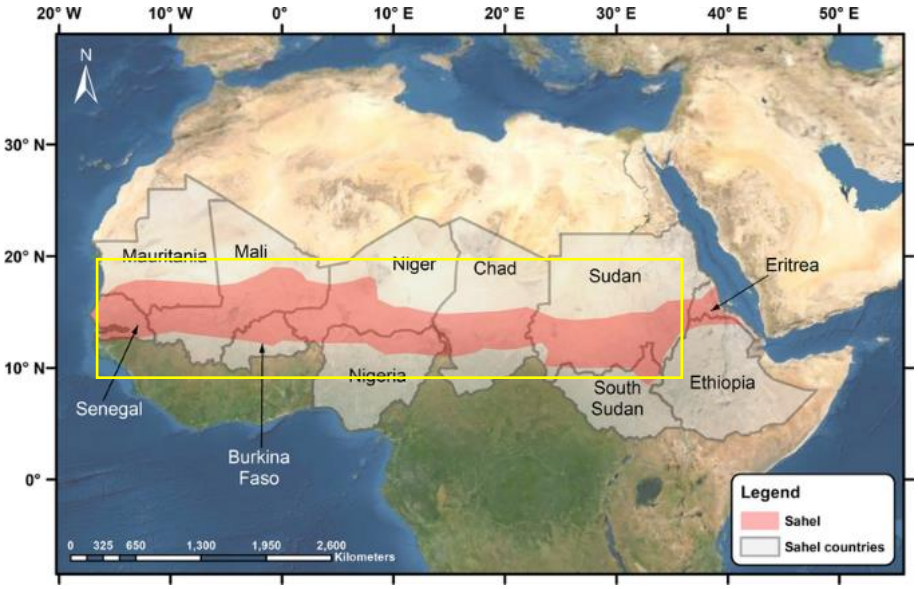


Figure 13: Carte de la région du Sahel et de ses pays12

12:Map of the Sahel region and countries | Download Scientific Diagram (researchgate.net)

2.2 Jeux de données utilisées

Comme mentionné et illustré à la figure 15, les modèles de distribution des espèces végétales nécessitent deux types de données d'entrée : les données biologiques décrivant la distribution déjà connue des espèces, et les données environnementales décrivant l'espace dans lequel les espèces se trouvent. Dans cette partie, nous décrirons les types de données utilisées ainsi que leurs sources.

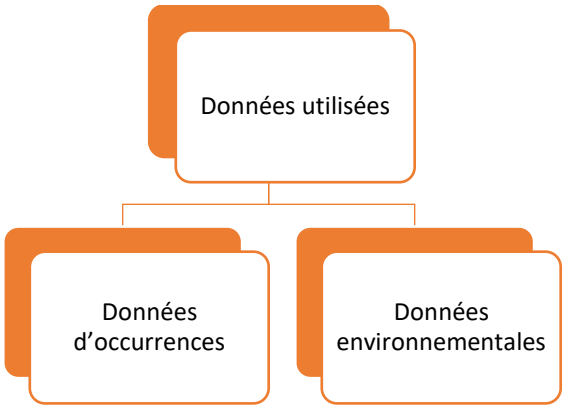


Figure 14:Types de données utilisées en entrée des modèles.

2.2.1 Données d'occurrences

Pour les données biologiques on se base sur la base de données FLOTROP, disponibles dans la base de données du *Global Biodiversity Information Facility* (GBIF) (Taugourdeau et al., 2019) et sur le site de Tela Botanica, considérée comme ressource majeure pour les données sur la diversité des plantes dans les écosystèmes ouverts du nord de l'Afrique tropicale. Cette base de données a été créer pour rassembler des observations botaniques collectées entre 1920 et 2012 dans les écosystèmes ouverts, tels que les savanes et les steppes. La collecte de données a été initiée par l'IEMVT (Institut d'Elevage et de Médecine Vétérinaire des Pays Tropicaux) et plutard par le CIRAD et le CNRS.

La base de données couvre une vaste région entre le 5 -ème et le 25 -ème parallèle nord, augmentant de 40% les données disponibles dans le GBIF pour cette région, et multipliant par 10 les occurrences disponibles pour certains pays.

En effet la base de données contient environ 340 000 occurrences de plantes provenant de diverses sources, y compris des relevées phytosociologies, des thèses de master et de doctorat, des rapports techniques, et des livres.

Les relevées de la base de données FLOTROP ont été classées en huit catégorie, allant des relevées phytosociologies non pondérés aux relevés écologiques avec estimation de la couverture ou du nombre d'individus. Les observations sont géoréférencées, bien que pour de nombreux relevés, les coordonnées aient été extrapolées à partir de description textuelle.

La base de données FLOTROP englobe plusieurs variables pour chaque occurrence :

- **Base d'enregistrement** : spécifie que toutes les enquêtes ont été réalisées sur la base d'observations humines.

- **Date de l'observation :** Date précise des relevés, parfois limitée à l'année ou au mois. Les données couvrent principalement les années 1960 – 2000.

-**Pays** : codé selon les normes ISO 3166, incluant plusieurs pays de l'Afrique notemment les pays du Sahel en plus du Cameroun, Bénin, Cote d'ivoire, Togo, Djibouti, Guinée, Cap-Vert et la République centre africaine avec une contribution majeure aux données pour le Burkina Faso, le Sénégal, le Niger, le Tchad, le Mali, la Mauritanie, et autres pays du Sahel.

-**Les coordonnées de l'observation (Latitude décimale et Longitude décimale)** : Géolocalisation précise pour tous les levés avec une précision à la minute près (environ 1.83 km dans ces régions). Les coordonnées sont en décimal pour la latitude et longitude et toutes les coordonnées sont en WGS 84.

- **Le nom de l'espèce (Nom scientifique et Rang taxon)** : Dans la base de données 4 372 espèces sont répertoriées, avec 64 espèces ayant plus de 1 000 occurrences et 19 avec plus de 2000 occurrences.

- **Le nom de l'observateur** : Noms de famille des observateurs inclus 222 auteurs différents.

- **Numéro d'enquête Flotrop** : Les observations de la base de données ont été regroupées par enquête c'est à dire en observant les espèces ensemble ainsi regroupées dans 26 932 enquêtes différentes.

- **Remarque d'occurrence** : Méthodes d'enquête originales notées, classées en 8 catégories de protocoles principalement des enquêtes phytosociologiques pondérées.

- **Remarque d'identification** : Noms de référence utilisés pour les noms d'espèces, avec validation par botanistes et correction des erreurs d'encodage.

Pour notre projet, nous utilisant une version plus compète de cette base disponible sur le CIRAD (Taugourdeau et al.,2019), qui a été organisée en mentionnant pour chaque enregistrement le relevé correspondant avec les informations complémentaires. Cette base de données inclue les colonnes représentées sur le tableau 1.

Tableau 1 Description des colonnes de la base de données FLOTROP utilisée dans l'étude

COLONNE	SIGNIFICATION
gbifID	Identifiant unique attribué par le GBIF (Global Biodiversity Information Facility) à chaque occurrence.
license	Licence sous laquelle les données sont publiées (par exemple, CC BY 4.0 pour une licence Creative Commons).
publisher	L'entité ou l'organisation qui a publié ces données.
basisOfRecord	Le type d'enregistrement (par exemple, spécimen, observation, etc.).
occurrenceID	Identifiant unique pour chaque occurrence spécifique.
recordNumber	Numéro d'enregistrement associé à l'occurrence.
recordedBy	Nom de la personne ou de l'équipe qui a collecté l'occurrence.
occurrenceStatus	Statut de l'occurrence (par exemple, présente, absente, douteuse, etc.).
occurrenceRemarks	Remarques supplémentaires sur l'occurrence.
eventDate	Date de l'événement (collecte, observation, etc.).
startDayOfYear et endDayOfYear	Jours de l'année où l'événement a eu lieu.
year, month, day	Année, mois et jour de l'événement.
continent	Continent où l'occurrence a été enregistrée.
countryCode	Code du pays où l'occurrence a été enregistrée.
decimalLatitude et decimalLongitude	Coordonnées géographiques de l'occurrence.
identificationRemarks	Remarques sur l'identification de l'espèce.
acceptedNameUsageID	Identifiant de l'espèce acceptée.
scientificName, kingdom, phylum, class, order, family, genus, species	Informations taxonomiques sur l'espèce.
datasetKey	Clé d'identification du jeu de données.
publishingCountry	Pays où les données ont été publiées.

lastInterpreted	Date de la dernière interprétation des données.
distanceFromCentroidInMeters	Distance de l'occurrence par rapport au centre géographique.
issue	Problèmes potentiels avec l'occurrence.
hasCoordinate, hasGeospatialIssues	Indicateurs de la qualité des coordonnées géographiques.
taxonKey, acceptedTaxonKey, kingdomKey, phylumKey, classKey, orderKey, familyKey, genusKey, speciesKey	Clés d'identification taxonomique.
verbatimScientificName	Nom scientifique tel qu'il apparaît dans l'occurrence.
protocol	Protocole utilisé pour collecter les données.
lastParsed, lastCrawled	Dates de la dernière analyse et du dernier parcours des données.
repatriated, isSequenced	Indicateurs de rapatriement et de séquençage des données.
gbifRegion, publishedByGbifRegion	Région GBIF et région de publication.
level0Gid, level0Name, level1Gid, level1Name, level2Gid, level2Name, level3Gid, level3Name	Identifiants et noms de niveaux géographiques.
iucnRedListCategory	Catégorie de la Liste rouge de l'UICN (Union internationale pour la conservation de la nature).

2.2.2 Données environnementales

a. Variables climatiques

Nous collectons les données météorologiques à partir des fichiers raster au format GeoTiff téléchargés à partir du site WorldClim. Ces fichiers contiennent sept variables climatiques : la température minimale, maximale, la radiation solaire, la vitesse de vent, la pression de la vapeur d'eau, les précipitations et l'élévation avec une résolution spatiale de 2.5 minutes. De plus, il existe également 19 variables bioclimatiques reprise dans le tableau 2.

Tableau 2 Description des variables bioclimatiques utilisées dans l'étude

VARIABLE	SIGNIFICATION
Bio1	Température annuelle moyenne
Bio2	Écart diurne moyen (moyenne mensuelle de la température maximale – température minimale)
Bio3	Isothermie (Bio2/Bio7) (×100)
Bio4	Saisonnière des températures (écart-type ×100)
Bio5	Température maximale du mois le plus chaud
Bio6	Température minimale du mois le plus froid
Bio7	Amplitude annuelle de la température (Bio5-Bio6)
Bio8	Température moyenne du trimestre le plus humide
Bio9	Température moyenne du trimestre le plus sec
Bio10	Température moyenne du trimestre le plus chaud
Bio11	Température moyenne du trimestre le plus froid
Bio12	Précipitations annuelles
Bio13	Précipitations du mois le plus humide
Bio14	Précipitations du mois le plus sec
Bio15	Saisonnière des précipitations (coefficient de variation)
Bio16	Précipitations du trimestre le plus humide
Bio17	Précipitations du trimestre le plus sec
Bio18	Précipitations du trimestre le plus chaud
Bio19	Précipitations du trimestre le plus froid
Bio18	Précipitations du trimestre le plus chaud
Bio19	Précipitations du trimestre le plus froid

Ainsi ces 26 variables ont été rajoutées à notre base de données en se basant sur les données de latitude et longitude.

b. Indices de pluviométrie

Afin d'intégrer la dimension temporelle dans notre analyse, nous avons ajouté des données relatives à l'année de collecte des relevés : l'indice de pluviométrie, un indicateur essentiel pour estimer si une année est sèche ou humide. Nous importons ces indices à partir d'une base de données disponible sur le site de l'université de Washington, couvrant la période de 1901 à 2019.

Ces indices sont importants pour comprendre les conditions climatiques sur une période étendue et enrichir notre analyse spatiale. Les données relatives à l'indice de pluviométrie relevés sont ajoutées en quatre colonnes distinctes, chaque colonne reflétant les conditions d'une année spécifique, de l'année antécédente, des cinq dernières années ou des dix dernières années.

Pour interpréter ces indices, une année est considérée comme humide si l'indice est positif, indiquant des conditions de précipitations favorables. En revanche, si l'indice est négatif, l'année est généralement considérée comme sèche, avec des précipitations insuffisantes.

Afin de prendre en compte l'effet historique des données, nous avons ajouté à notre base de données ces informations sur quatre colonnes distinctes. Chaque colonne indique si l'année est sèche ou humide, mais en se basant sur différentes références temporelles :

- La première colonne s'est basée uniquement sur l'année elle-même.
- La deuxième s'est basée sur l'année antécédente.
- La troisième colonne s'est basée sur la moyenne des cinq années antécédentes.
- Et la quatrième s'est basée sur la moyenne des dix années antécédentes.

La figure 16 montre la distribution des années sèches et humides selon quatre critères différents: Humidité_a (année courante), Humidité_ap (année précédente), Humidité_5ap (moyenne des 5 années précédentes), et Humidité_10ap (moyenne des 10 années précédentes).

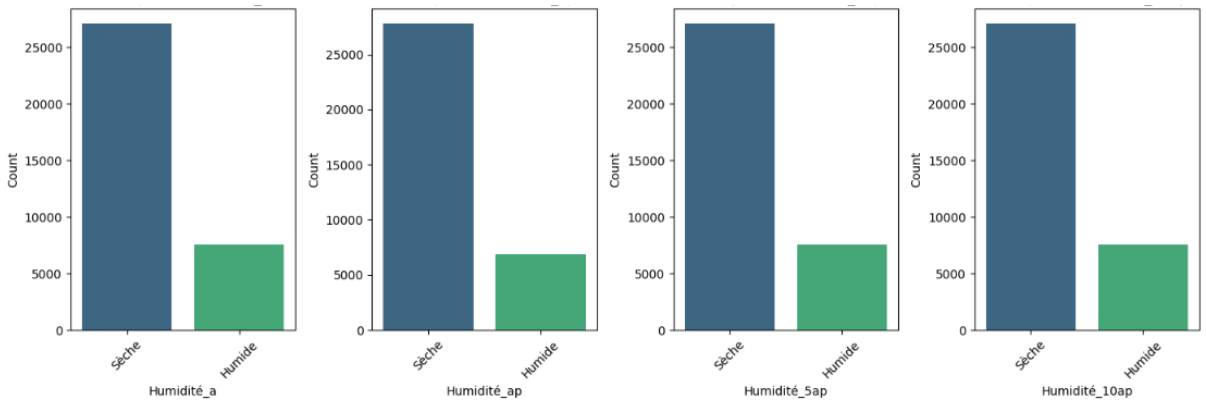


Figure 15: Distribution de l'Humidité pour Différentes Périodes dans la Base de Données

Ces indices relatifs permettent d'apporter une compréhension plus approfondie des conditions climatiques sur une période étendue, ce qui enrichira notre analyse spatiale en tenant compte de l'évolution temporelle des données.

c. Types de sol

Nous avons également intégré des données relatives au type du sol à partir de la base de données de La FAO (UNEP, n.d.). Le shapefile de la base de données a été transformé en raster avec la même résolution et la même étendue spatiale que les variables climatiques pour assurer la compatibilité et la cohérence des analyses.

La figure 17 montre le nombre d'enregistrements pour les dix types de sols les plus fréquents dans la base de données FAOSOIL, illustrant la distribution des types de sols sur la zone d'étude et le tableau 3 présente la signification des codes de types de sol extraite de la base de données FAO pour l'Afrique.

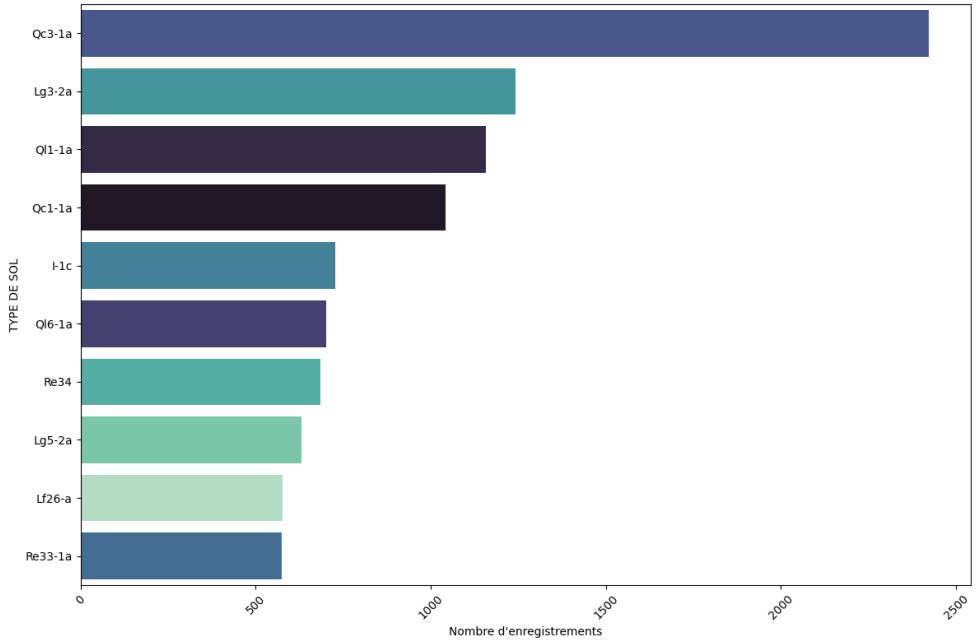


Figure 16: Les 10 types de sols les plus fréquents dans la base de données FAOSOIL

Tableau 3 Tableau des abréviations et significations des types de sol

Abréviation type du sol	**Signification**
Qc3-1a	Combisols calcaires
Lg3-2a	Gleysols ferrugineux
QI1-1a	Ferralsols hapliques
Qc1-1a	Cambisols calcaires
I-1c	Ferralsols rhodiques
QI6-1a	Ferralsols gérics
Re34	Acrisols plinthiques
Lg5-2a	Gleysols hapliques
Lf26-a	Leptosols calciques
Re33-1a	Acrisols ferriques

Ainsi on obtient notre base de données contenant les colonnes relatives aux 10 espèces végétales dominantes en plus des données de latitude, longitude et année correspondantes liées à nos 32 variables environnementales correspondantes à celles citées dans la figure 18.

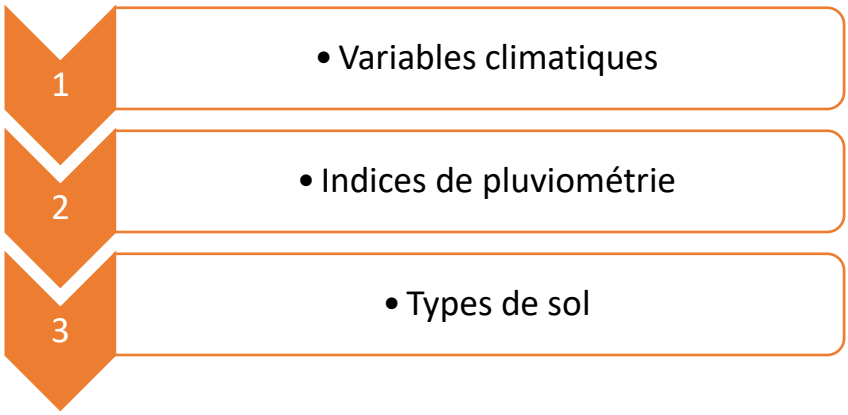


Figure 17 Données environnementales utilisées dans les modèles.

2.2.3 Données de l'herbier

Un herbier est une collection de spécimens de plantes séchées et pressées, souvent montées sur des feuilles de papier, avec des informations associées telles que le lieu et la date de collecte, le nom du collecteur, et des descriptions de l'habitat (Bridson & Forman, 1999).

Les herbiers sont utilisés comme référence pour l'identification des plantes et pour des études sur la diversité végétale, la distribution géographique et l'écologie des espèces (Thiers, 2021).

Dans cette étude, les données de l'herbier disponibles sur le site du CIRAD ont été utilisées pour valider les résultats de modélisation des distributions des espèces.

2.3 Logiciels et outils utilisés

Pour mener à bien cette étude de modélisation de la distribution des espèces dans la région du Sahel, plusieurs logiciels et outils ont été utilisés, chacun offrant des fonctionnalités spécifiques adaptées aux besoins du projet.

2.3.1. Langage de Programmation

a. Python

Python est un langage de programmation polyvalent, utilisé pour une variété de tâches allant de l'analyse de données au développement web. Dans cette étude, Python a été choisi pour ses bibliothèques spécialisées et sa large communauté de support. Les bibliothèques les plus utilisées sont :

- Pandas : Utilisé pour la manipulation et l'analyse des données, offrant des structures de données flexibles et performantes.
- NumPy : Utilisé pour le calcul scientifique et la gestion des tableaux multidimensionnels.
- Scikit-learn : Utilisé pour le *Machine Learning* offrant des outils simples et efficaces pour l'analyse prédictive des données.
- PyGAM : Utilisé pour l'ajustement des modèles additifs généralisés.
- Matplotlib et Seaborn : Utilisés pour la visualisation des données.

b. Langage R

R (R Core Team, 2024) est un langage de programmation et un environnement logiciel pour les statistiques et le graphisme. Il est particulièrement apprécié pour les analyses statistiques complexes et la visualisation des données. Les packages les plus utilisés dans cette étude incluent :

- raster : Utilisé pour le traitement et l'analyse des données raster.
- sp : Utilisé pour la manipulation et l'analyse des données spatiales.
- sdm : Un package dédié à la modélisation de la distribution des espèces, permettant d'intégrer divers modèles statistiques et d'apprentissage automatique.

Une comparaison entre R et Python a été effectuée en annexe (Tableau II). R est souvent préféré pour les analyses statistiques et les visualisations avancées, tandis que Python est apprécié pour son intégration facile avec les systèmes de production et ses vastes bibliothèques de *Machine Learning*

2.3.2 Environnement d'Exécution

a. Google Colab

Google Colab est un environnement de développement intégré qui permet d'exécuter du code Python directement dans le navigateur. Il est particulièrement utile pour le traitement des grandes quantités de données et offre un accès gratuit à des GPU pour l'entraînement de modèles de *Machine Learning*.

b. RStudio

RStudio est un environnement de développement intégré pour R.

2.3.3. Analyse et Visualisation Géospatiale

a. QGIS

QGIS est un système d'information géographique (SIG) utilisé pour la visualisation, la gestion et l'analyse des données géographiques. Cet outil a été utilisé principalement pour l'exploration et l'analyse de nos différents types de données raster en plus qu'il était utile pour la génération de cartes.

En utilisant ces outils et logiciels, nous avons pu exploiter leurs différentes fonctionnalités pour répondre aux besoins spécifiques de notre projet, de la manipulation des données climatiques à l'entraînement de modèles prédictifs complexes.

2.4 Méthodologie générale

La méthodologie de cette étude repose sur la collecte, le prétraitement et l'analyse de données environnementales et d'occurrence des espèces pour modéliser la distribution des dix espèces végétales dominantes de notre base donnée sur la région du Sahel. Nous utilisons après des techniques de modélisation de distribution des espèces (SDM) intégrant des approches de *Machine Learning* et de *Deep Learning* pour prédire la présence ou l'absence des espèces cibles.

Nous nous basons aussi sur les données de l'herbier pour valider la qualité de nos prédictions.

L'organigramme de la figure 19 ci-dessous présente les étapes générales que nous allons suivre dans notre démarche pour atteindre nos objectifs.

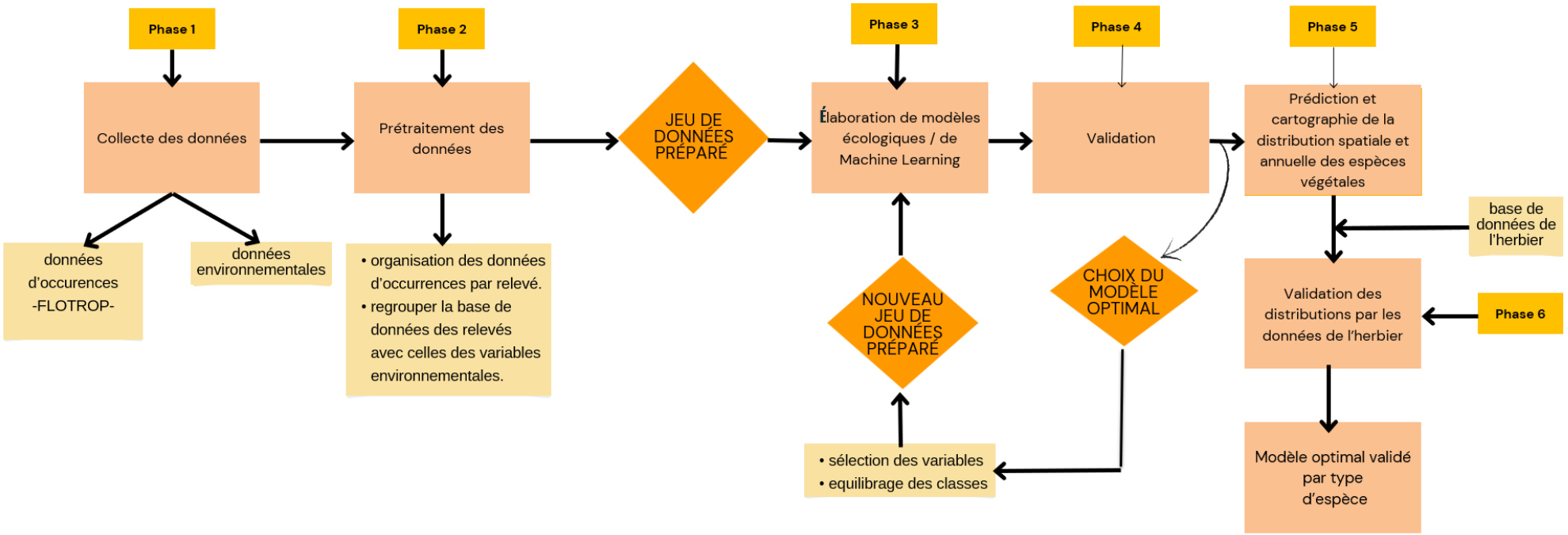


Figure 18:Processus de création et de validation des cartes de distribution des espèces végétale

2.5 Méthodologie détaillée

2.5.1 Collecte des données

2.5.1.1 Données d'occurrences

Après avoir défini les données que nous allons utiliser : données d'occurrence des espèces végétales et données environnementales, nous commençons par importer les données de la base de données FLOTROP. Pour répondre aux besoins spécifiques de notre modèle, on a sélectionné les colonnes suivantes :

- **L'Id Relevé** : correspond à l'identifiant du relevé de l'enregistrement en question.
- **lat,long** : coordonnées de l'espèce en question dans l'enregistrement correspondant.
- **Année** : correspond à l'année de collecte de cet enregistrement.
- **GENRE et ESPECE** : représentent le genre et le nom de l'espèce en question dont la combinaison correspond à l'identifiant de l'espèce étudiée.

La figure 20 présente ainsi la distribution spatiale des données disponible sur FLOTROP sur notre zone d'étude.

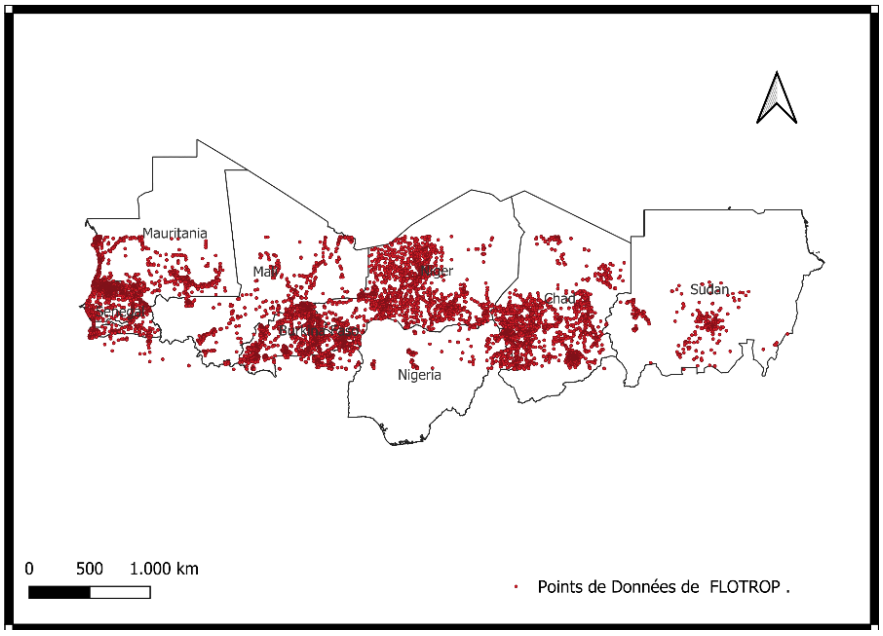


Figure 19:Distribution spatiale des points de données de FLOTROP sur la zone d'étude entre 1920 et 2012..

Nous avons ensuite restructuré notre base de données pour optimiser l'analyse des occurrences des espèces. Initialement, chaque enregistrement de notre base représentait une observation unique, avec plusieurs enregistrements possibles pour un même relevé correspondant à différentes espèces ou localisations. Pour cette étude, nous avons sélectionné uniquement les données relatives aux dix espèces les plus dominantes : *'Cenchrus biflorus', 'Schoenefeldia gracilis', 'Aristida mutabilis', 'Dactyloctenium aegyptium', 'Eragrostis tremula', 'Balanites aegyptiaca', 'Alysicarpus ovalifolius', 'Zornia glochidiata', 'Combretum glutinosum', 'Andropogon gayanus'*, ayant respectivement les codes suivant :153, 672, 162, 262, 324, 93, 42, 807, 186, 58, comme présenté dans le tableau 4, pour avoir suffisamment de données pour pouvoir entrainer nos modèles. La figure 21 ainsi présente le nombre d'enregistrement pour les 10 espèces dominantes.

Tableau 4 Codes et Noms des 10 Espèces Dominantes

Code	*93*	*42*	*807*	*186*	*58*
Espèce	*Balanites aegyptiaca*	*Alysicarpus ovalifolius*	*Zornia glochidiata*	*Combretum glutinosu*	*Andropogon gayanus*

Code	*93*	*42*	*807*	*186*	*58*
Espèce	*Balanites aegyptiaca*	*Alysicarpus ovalifolius*	*Zornia glochidiata*	*Combretum glutinosu*	*Andropogon gayanus*

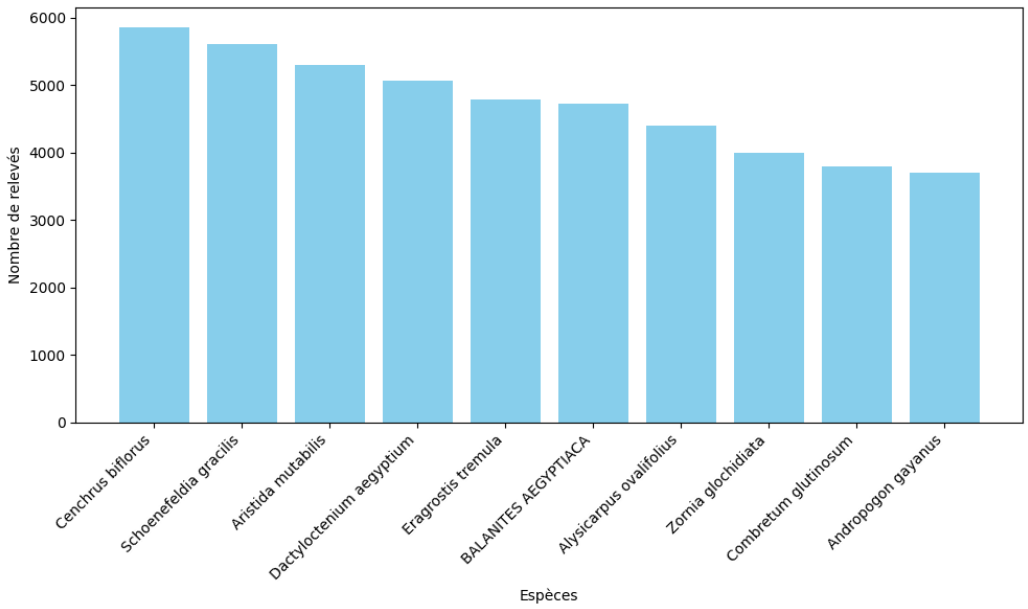


Figure 20:Nombre de relevés pour les 10 espèces les plus abondantes

Nous avons d'abord extrait les enregistrements correspondant aux dix espèces les plus abondantes parmi l'ensemble des données disponibles. Puis nous organisons notre base de données de façon à ce que chaque ligne de la nouvelle base de données représente un relevé unique, assurant que chaque relevé est identifié par une seule ligne dans la matrice de données.

Ainsi notre nouvelle matrice de données comporte 14 colonnes : une pour l'identifiant du relevé et dix pour les espèces dominantes en plus des 3 colonnes de latitude, longitude et année correspondants. Pour chaque espèce, un '1' indique sa présence dans le relevé correspondant, tandis qu'un '0' indique son absence supposée (pseudo-absence). En effet l'absence d'une espèce dans les relevés sélectionnés est interprétée comme une pseudo-absence en supposant que les relevés sont exhaustifs, c'est-à-dire que pour le même relevé on a collecté les informations de présence ou d'absence de toutes nos espèces végétales. Ainsi, notre base de données finale contient 34 744 occurrences avec 14 colonnes. Cette organisation, comme illustré dans la figure 22 facilite non seulement l'analyse des modèles de présence-absences mais fournit également une base solide pour la modélisation prédictive de la distribution des espèces en réponse à divers facteurs environnementaux.

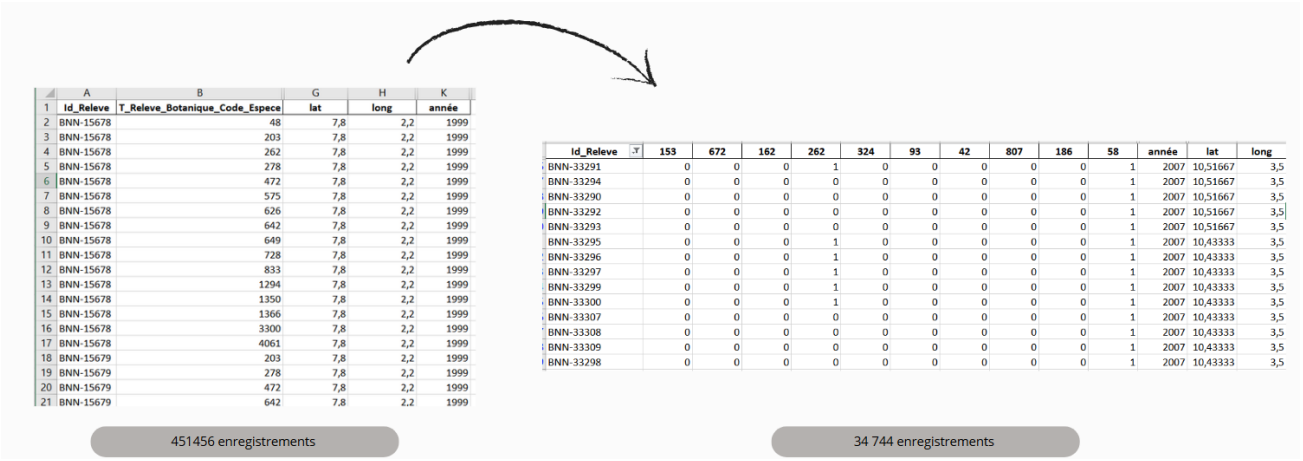


Figure 21. Transformation des données d'observation : de l'enregistrement initial à la matrice de présence/absence

2.5.1.2 Données environnementales

Nous avons commencé par télécharger les fichiers Geotiff depuis WorldClim correspondant à nos 7 variables climatiques et 19 variables bioclimatiques, puis nous les avons stockés dans un format structuré pour faciliter leur manipulation et leur intégration dans notre base de données.

Ensuite, les données de l'indice de pluviométrie sont lues à partir d'un fichier ASCII, où chaque ligne contient une année et un indice de pluviométrie. Un dictionnaire est utilisé pour stocker ces données, permettant un accès rapide par année.

Enfin, les données de sol sont intégrées à partir d'un fichier shapefile contenant les informations des types de sol. Les coordonnées des relevés des espèces sont converties en géométrie de type Point, et une intersection spatiale est effectuée pour associer les points aux polygones de sol les plus proches. Les valeurs de type de sol sont ajoutées au DataFrame contenant les relevés.

2.5.2 Prétraitement des Données

2.5.2.1 Préparation et nettoyage des Données

Avant d'entraîner les modèles, les données ont été prétraitées pour garantir leur qualité et leur pertinence. Voici les étapes détaillées du prétraitement :

- **Suppression des Colonnes Non Pertinentes** : Les colonnes non nécessaires à l'analyse ont été supprimées pour simplifier la base de données et se concentrer sur les variables explicatives pertinentes.
- **Gestion des Valeurs Manquantes** : Les enregistrements contenant des valeurs manquantes ont été retirés pour éviter les biais dans les modèles prédictifs.
- **Normalisation des Variables Numériques** : Les variables numériques ont été normalisées pour garantir une échelle uniforme. Cela permet de s'assurer que toutes les variables contribuent de manière équitable au modèle.
- **Encodage des Variables Catégorielles** : L'encodage des variables catégorielles par fréquence a été appliqué pour la variable du type de sol afin de convertir les données catégorielles en un format numérique approprié pour l'analyse.

  Extraction des Données Climatiques et Bioclimatiques : La fonction 'extract' de la bibliothèque 'raster' a été utilisée pour obtenir les valeurs des variables bioclimatiques aux coordonnées spécifiques des relevés. Lorsque plusieurs valeurs étaient extraites pour un seul point, la fonction 'mean' a été appliquée pour calculer la moyenne, lissant ainsi les variations.

- **Intégration des Données de Pluviométrie** : Les données de l'indice de pluviométrie ont été stockées dans un dictionnaire pour un accès rapide par année. Une fonction a été définie pour déterminer si une année est humide ou sèche en comparant les indices de pluviométrie de l'année en question, de l'année précédente, et des moyennes sur 5 et 10 ans. Les résultats ont été ajoutés au DataFrame existant contenant les données des espèces.
- **Association des Relevés aux types de sol** : Les coordonnées des relevés des espèces ont été converties en géométrie de type Point. Une intersection spatiale a été effectuée pour associer les points aux polygones de sol les plus proches. Les valeurs de type de sol ont été ajoutées au DataFrame contenant les relevés.

- **Enregistrement des DataFrames Modifiés** : Les DataFrames modifiés, contenant désormais les informations d'humidité/sécheresse et de types de sol, ont été enregistrés dans de nouveaux fichiers Excel. Cela permet une utilisation ultérieure dans l'analyse des données et la modélisation, garantissant que toutes les variables environnementales pertinentes sont intégrées de manière cohérente et structurée.

2.5.2.2 Équilibrage des classes

Dans le cas de la modélisation de la distribution des espèces, un défi majeur est le déséquilibre des classes dans les données. Ce déséquilibre survient lorsque l'une des classes présence ou absence d'une espèce est beaucoup plus représentée que l'autre. Dans notre cas, les données d'occurrence montrent souvent un déséquilibre significatif entre les deux classes, avec beaucoup plus de données d'absence que de présence pour certaines espèces comme présenté dans la figure 23. Ce déséquilibre peut biaiser les modèles d'apprentissage automatique, conduisant à des performances médiocres sur la classe minoritaire, qui est la classe des présences, la plus intéressante à prédire.

Pour atténuer ces problèmes, nous utilisons des techniques de rééquilibrage des classes, dont l'une des méthodes les plus efficaces est le sous-échantillonnage de la classe majoritaire '**RandomUnderSampling**'.

Autre méthode d'équilibrage étaient testées tels que le sur-échantillonnage de façon à augmenter les données pour la classe minoritaire, celle des présences mais cette technique ne donnait pas des résultats crédibles vu qu'il s'agit de faux présences sur lesquelles notre modèle va se baser.

Par contre le sous-échantillonnage réduit le nombre d'occurrence de la classe majoritaire pour équilibrer le jeu de données, ce qui force le modèle à accorder plus d'attention à la classe minoritaire.

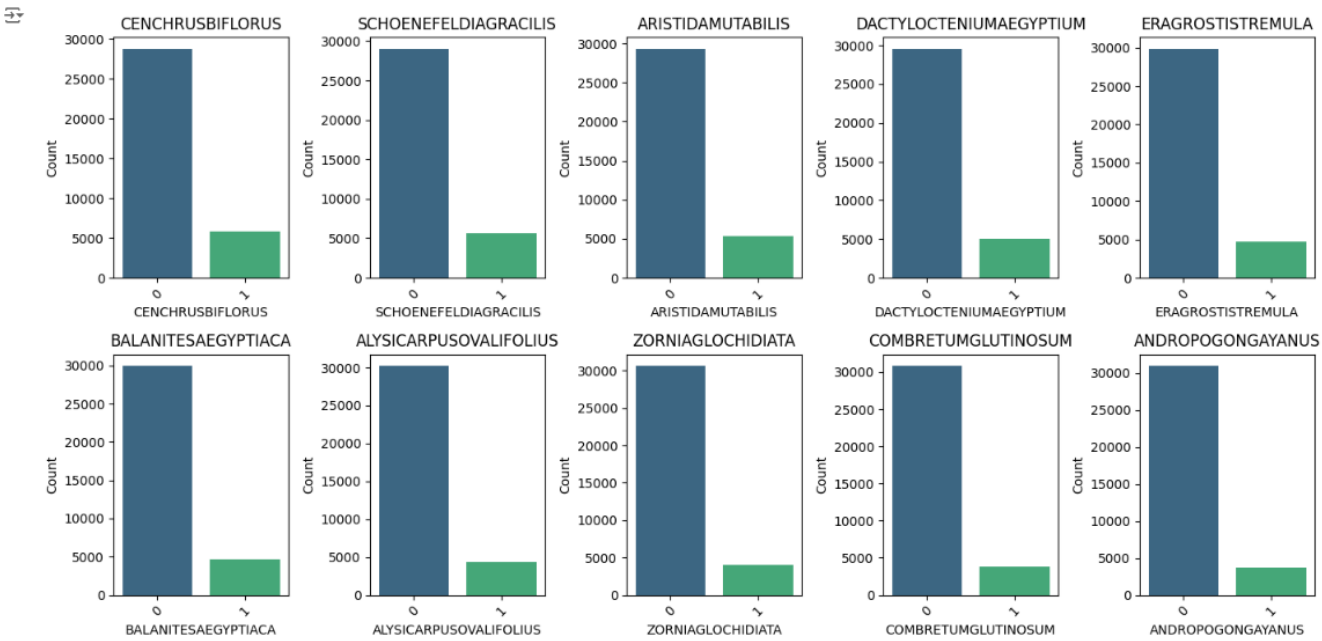


Figure 22:Nombre d'observation pour chaque classe des espèces les plus abondantes

2.5.2.3 Sélection des variables

Avant de procéder à l'élaboration de nos modèles, nous avons appliqué plusieurs méthodes de sélection de variables pour déterminer les variables environnementales les plus pertinentes afin de prédire la présence des espèces étudiées. Dans cette étude, plusieurs méthodes de sélection de variables ont été utilisées pour identifier les variables les plus significatives pour chaque espèce végétale étudiée :

- **Analyse en Composantes Principales (ACP)** : cette méthode statistique permet de réduire la dimensionnalité des données en transformant les variables d'origine en un nouvel ensemble de variables non corrélées appelées composantes principales. Les premières composantes capturent la majeure partie de la variance des données, facilitant ainsi l'interprétation et la visualisation des données multivariées.
- **Importance des caractéristiques** : Utilisée avec la méthode 'RandomForestClassifier', cette méthode évalue l'importance relative de chaque variable en fonction de sa contribution à la précision du modèle. Les variables ayant une importance supérieure à la moyenne sont sélectionnées pour la modélisation.
- **Information mutuelle** : Cette méthode mesure la dépendance entre les variables d'entrée et la variable cible. Les variables présentant la plus forte information mutuelle avec la variable cible sont sélectionnées pour la modélisation.

Ces différentes techniques de sélection des variables permettent d'optimiser les modèles en réduisant le nombre de variables inutiles et en améliorant la performance prédictive.

Grâce à cette approche méthodique de collecte et de prétraitement des données, nous avons pu garantir une base solide pour nos analyses et modélisations, assurant ainsi des résultats fiables et précis pour la prédiction de la distribution des espèces dans la région du Sahel.

2.5.3 Modélisation

Pour notre étude sur la prédiction de la distribution spatiale des espèces végétales en fonction des variables environnementales, nous avons choisi d'utiliser plusieurs modèles en raison de leurs capacités spécifiques à gérer les types de données disponibles : présence et pseudo-absence. La figure 24 présente les différents modèles testés lors de notre étude.

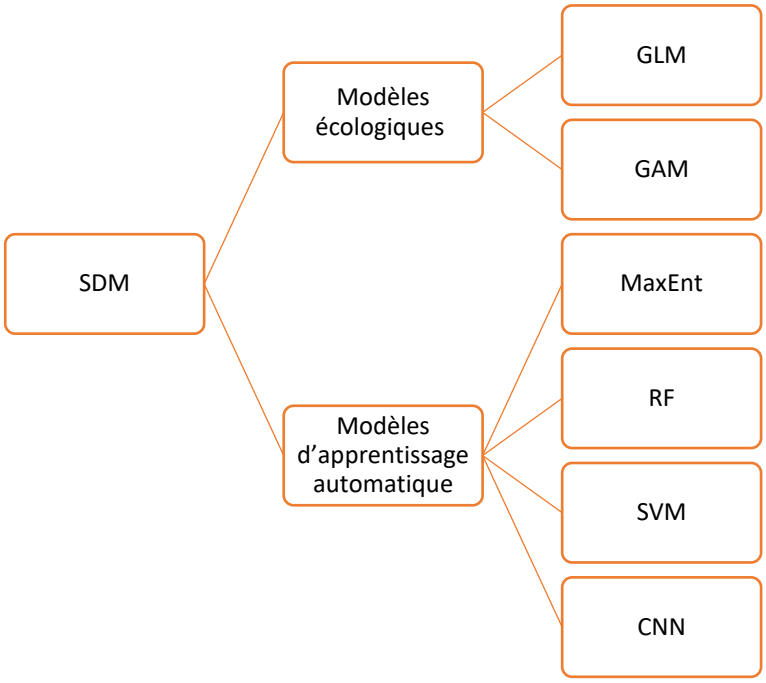


Figure 23. Modèles testés pour la cartographie de la répartition spatio-temporelle des espèces végétales.

2.5.3.1 Modèles Écologiques

a. GLM (Modèles Linéaires Généralisés)

Les GLM ont été choisis pour leur capacité à modéliser des relations linéaires entre les variables environnementales et la présence des espèces. Ils sont particulièrement utiles pour identifier et quantifier l'impact de chaque variable environnementale sur la distribution des espèces. Cette capacité est essentielle pour notre étude afin de comprendre les facteurs environnementaux clés influençant la présence des espèces végétales.

Nous utilisons 'LogisticRegression' de 'scikit-learn' pour appliquer ces modèles, modélisant la relation entre une variable de réponse et plusieurs variables explicatives à l'aide d'une fonction de lien, permettant de comprendre les relations linéaires entre les variables.

b. GAM (Modèles Additifs Généralisés)

Les GAM ont été sélectionnés en raison de leur flexibilité à modéliser des relations non linéaires complexes. Ils permettent de capturer les interactions subtiles et non linéaires entre les variables environnementales et la présence des espèces, offrant ainsi une meilleure précision dans les prédictions de distribution. Cette caractéristique est cruciale pour notre étude, qui nécessite une modélisation détaillée des interactions complexes.

Les GAM, qui sont des extensions des GLM, modélisent les relations non linéaires en associant des fonctions lisses à chaque prédicteur. Cette méthode permet de capturer efficacement les interactions complexes entre les variables environnementales et la présence des espèces. Pour notre étude, nous utilisons 'LogisticGAM' de 'pygam' afin de modéliser ces relations non linéaires de manière précise et détaillée.

2.5.3.2 Modèles de *Machine Learning*

a. MaxEnt (*Maximum Entropy Modeling*)

MaxEnt a été choisi pour sa robustesse et son efficacité avec des données de présence et des pseudo-absences. Il maximise l'entropie sous les contraintes fournies par les données environnementales, ce qui en fait un outil puissant pour estimer la distribution probable des espèces en se basant principalement sur les données de présence. Cette méthode est particulièrement adaptée à notre étude en raison de la disponibilité de données de présence.

Les variables catégorielles sont transformées en variables binaires, les coordonnées sont extraites et ajoutées sous forme de nouvelles colonnes Latitude et Longitude. Les doublons sont vérifiés et, si nécessaire, une agrégation est effectuée. Les variables catégorielles sont ensuite encodées à l'aide de 'dummyVars' de 'caret', et les données sont divisées en ensembles d'entraînement et de test. La formule du modèle est créée en incluant toutes les variables explicatives et les coordonnées. Les données sont préparées pour le package 'sdm' et le modèle Maxent est appliqué aux données d'entraînement. Enfin, le modèle est utilisé pour prédire la distribution des espèces sur l'ensemble de test, et les résultats des prédictions sont sauvegardés et visualisés.

b. Random Forest (Forêts Aléatoires)

Random Forest a été sélectionné pour sa capacité à gérer des interactions complexes et des données variées. Il offre une robustesse et une précision améliorées en évitant le surapprentissage, ce qui est essentiel pour les données écologiques complexes de notre étude. La méthode construit une multitude d'arbres de décision et utilise la moyenne des prédictions pour obtenir des résultats fiables.

Nous utilisons ‘RandomForestClassifier’ de ‘scikit-learn’ avec 100 arbres, chaque arbre étant formé sur un échantillon bootstrap des données. Les prédictions sont obtenues en moyennant les résultats des arbres individuels, ce qui permet d'assurer des estimations robustes et précises de la distribution des espèces.

c. SVM (*Support Vector Machines*)

Les SVM ont été choisis pour leur efficacité dans les problèmes de classification binaire et leur utilité avec des données de haute dimension. Ils trouvent l’hyperplan optimal pour séparer les classes de présence et d'absence dans un espace de caractéristiques multidimensionnel, offrant ainsi des prédictions précises et robustes. Cette méthode est adaptée à notre étude en raison de la complexité des données environnementales disponibles.

Nous utilisons ‘SVC’ de ‘scikit-learn’ avec le noyau RBF pour capturer les relations non linéaires et appliquer le modèle Support Vector Machine (SVM), un classificateur qui trouve l’hyperplan optimal séparant les classes dans un espace de haute dimension.

d. Deep Learning (Réseaux de Neurones Profonds)

Les réseaux de neurones profonds ont été choisis pour leur capacité à traiter de grandes quantités de données et à capturer des relations complexes et subtiles entre les variables environnementales et la distribution des espèces.

Nous utilisons ‘PyTorch’ pour définir et entraîner un modèle CNN personnalisé, comprenant des couches convolutionnelles, de pooling et des couches entièrement connectées. Le modèle est défini dans la classe ‘CNNModel’ avec des couches spécifiques et des fonctions d'activation ‘ReLU’, suivies d'une couche de pooling et de convolution. Les données sont standardisées, converties en objets ‘DataLoader’, et le modèle est entraîné et évalué sur des ensembles de validation et de test, avec une fonction de perte ‘CrossEntropyLoss’ et un optimiseur ‘Adam’.

Le modèle commence par une première couche convolutionnelle avec 1 canal d'entrée et 6 canaux de sortie, suivie d'une couche de pooling pour réduire la dimensionnalité. Une deuxième couche convolutionnelle avec 6 canaux d'entrée et 16 canaux de sortie est ensuite appliquée, suivie d'une autre couche de pooling. Enfin, les données sont aplaties et passées à travers trois couches entièrement connectées pour la classification finale. Le schéma général du modèle est ainsi représenté dans la figure 25.

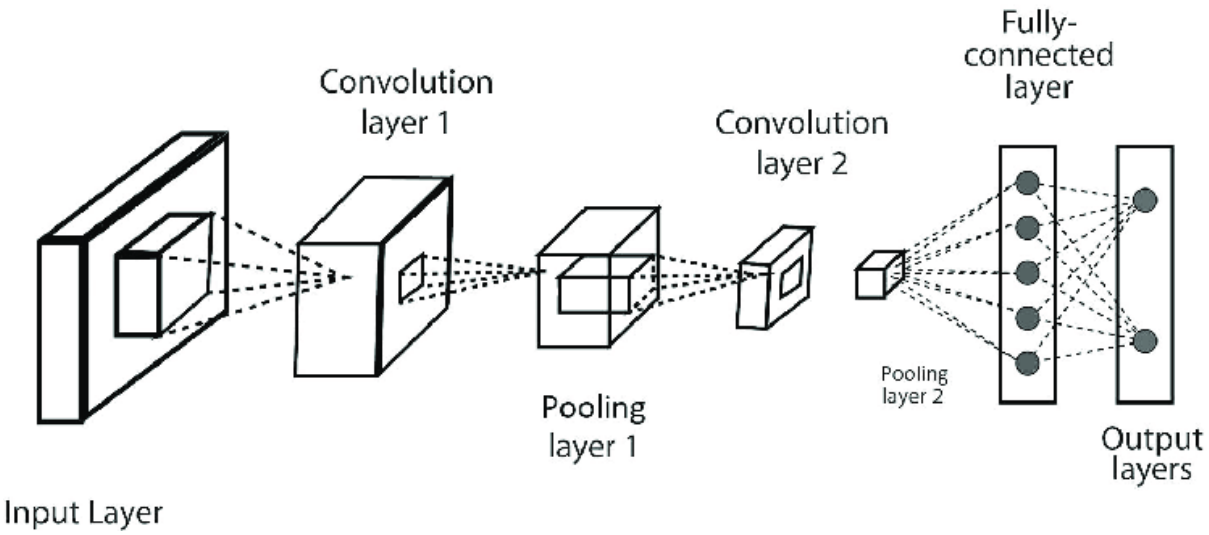


Figure 24. Architecture d'un réseau de neurones convolutionnels (Khan et al,2021).

La sélection de ces modèles pour notre étude est basée sur leurs capacités spécifiques à traiter les données disponibles et à fournir des prédictions précises sur la distribution spatiale des espèces végétales. L'intégration de ces modèles permet de capturer une variété de relations écologiques et environnementales, améliorant ainsi la robustesse et la précision de nos prévisions.

2.5.4 Évaluation des modèles :

Pour s’assurer que les modèles se généralisent bien aux nouvelles données et ne surapprennent pas les données d’entrainement, la validation des modèles a été effectuée en utilisant une validation croisée stratifiée à 10 plis. Pour chaque pli, les données sont divisées en ensemble d’apprentissage et de test, les variables sont normalisées et standardiser puis les modèles sont entrainés sur les ensembles d’apprentissage et les performances sont évaluées sur l’ensemble de test, différentes métriques ont été utilisées :

- **Exactitude** : proportion de prédictions correctes (positives et négatives) sur l'ensemble des prédictions. Bien que couramment utilisée, l'exactitude peut être trompeuse dans les jeux de données déséquilibrés, car elle ne prend pas en compte la répartition des classes. (Jiménez-Valverde et al., 2009).

$$Exactitude = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Précision** : pourcentage de prédictions positives correctes parmi toutes les prédictions positives effectuées. Elle se calcule comme le ratio entre le nombre de prédictions correctes et le nombre total de prédictions. Elle est utile pour avoir une vue d'ensemble de la performance du modèle. (Van Rijsbergen, 1979).

$$\text{Précision } = \frac{TP}{TP + TN}$$

- **Rappel macro (sensibilité)** : Rappel moyen pondéré par le nombre de vraies instances pour chaque étiquette. Le rappel mesure la capacité du modèle à identifier correctement toutes les instances positives de chaque classe. Le rappel macro assure que les performances de toutes les classes sont considérées de manière équitable (Lobo et al., 2008) .

$$Rappel = \frac{TP}{TP + FN}$$

- **F1 score macro** : Moyenne harmonique de la précision et du rappel macro. Il combine à la fois la précision et le rappel, fournissant une mesure unique qui équilibre les deux. Il est particulièrement utile lorsqu'on a besoin d'un compromis entre les faux positifs et les faux négatifs (Powers, 2011).

$$F1 - Score = 2 * \frac{\text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

- **Matrice de confusion** : Un tableau permettant de visualiser les performances du modèle en termes de véritables positifs, faux positifs, véritables négatifs et faux négatifs pour chaque classe comme illustré dans le tableau 5. En fournissant une vue détaillée des erreurs de classification, la matrice de confusion aide à comprendre les faiblesses spécifiques du modèle et à ajuster les stratégies de modélisation en conséquence. (Provost & Kohavi, 1998).

Tableau 5:Matrice de confusion

	Prédictions	
Réalité	TN(Vrais négatifs)	FP(Faux positifs)
	FN(Faux négatifs)	TP(Vrais positifs)

2.5.5 Cartographie de la distribution spatiale et annuelle des espèces végétales.

Après avoir sélectionné le modèle optimal pour chaque type d'espèce végétale, nous avons procédé à la prédiction de la distribution spatiale et annuelle des espèces sur notre zone d'étude entre 1920 et 2012. Nous avons utilisé les modèles *Random Forest* pour les espèces herbacées et CNN pour les espèces arborées.

Les prédictions ont été visualisées sur des cartes annuelles de distribution, permettant d'identifier les zones potentiellement favorables pour chaque espèce.

2.5.6 Vérification des prédictions par les données de l'herbier

Pour vérifier les résultats de nos modèles, nous avons utilisé les données de présence extraites des données d'herbier disponibles sur le Sahel. Bien que la quantité de données soit limitée, elles offrent une validation précieuse des prédictions réalisées par nos modèles.

2.5.7 Conclusion

Après avoir discuté les résultats des modèles, nous concluons sur le choix des modèles adéquats pour chaque type d'espèce. Nous comparons les résultats de la distribution des espèces aux données de la bibliographie.

Conclusion

Le chapitre sur le matériel et les méthodes a permis de clarifier les différentes étapes suivies pour collecter, prétraiter et analyser les données nécessaires à la modélisation de la distribution des espèces. Les outils et logiciels spécifiques utilisés, ainsi que les méthodologies adoptées, ont été détaillés, garantissant ainsi une reproductibilité et une transparence dans notre approche. Ces éléments méthodologiques nous préparent à présenter et interpréter les résultats de notre étude dans le prochain chapitre.

Chapitre III : Résultats et interprétations

Introduction

Dans cette section, nous allons présenter les résultats numériques obtenus pour les différents modèles appliqués pour la prédiction de la distribution des espèces végétales dans le Sahel. Les modèles évalués comprennent MaxENT, *Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM), et Convolutional Neural Network (CNN*).

Nous débutons par évaluer les performances de ces modèles sans équilibrer les classes ni appliquer de méthodes de sélection de variables. Cette première évaluation nous permet de mesurer les performances des modèles en utilisant l'ensemble des informations fournies par toutes les variables environnementales et en respectant les proportions de classes de nos données de présence. Les résultats obtenus sont discutés afin d'identifier le modèle optimal pour chaque type d'espèce. Pour simplifier nos interprétations, nous nous concentrons sur deux types de végétation : une herbacée (*Cenchrus biflorus*) et un arbre (*Balanites aegyptiaca*).

Après avoir identifié les modèles optimaux pour ces deux types d'espèces, nous procédons à l'équilibrage des classes. Pour chaque modèle, nous effectuons des expérimentations en incluant toutes les variables, puis en appliquant des méthodes de sélection de variables : la méthode d'importance des caractéristiques, l'information mutuelle, et l'analyse en composantes principales (ACP). Les deux premières méthodes nous permettent d'identifier les variables les plus influentes, tandis que la troisième est utilisée pour réduire la dimensionnalité des variables tout en conservant la majorité de l'information.

Enfin, nous présentons les résultats des modèles après équilibrage des classes et sélection des variables par ACP, afin d'optimiser les performances prédictives des modèles de distribution des espèces.

3.1 Résultats des différents modèles sans sélection de variables et sans équilibrage de classes

Pour commencer nous présentons les résultats des modèles sans sélection de variables et sans équilibrage de classe pour pouvoir évaluer la qualité des prédictions avec toutes les variables disponibles et avec les proportions de classe de notre base de données.

Pour notre étude, nous avons choisi d'utiliser la validation croisée stratifiée à 10 strates en raison du déséquilibre entre les classes de présence et d'absence. Cette méthode garantit que chaque bloc de validation contient une proportion égale de classes, ce qui permet d'obtenir des estimations plus robustes et fiables des performances du modèle. La validation croisée stratifiée aide à atténuer les effets des déséquilibres de classes, offrant une meilleure représentation des performances du modèle sur l'ensemble des données.

3.1.1 Pour la 1 ère espèce : *Cenchrus biflorus*

Le tableau 6 ci-dessous présente les métriques de performance pour différents modèles de prédiction appliqués à l'espèce *Cenchrus biflorus.* Les modèles évalués comprennent MaxEnt, *Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM), Generalized Additive Model (GAM), et Convolutional Neural Network (CNN)*. Les métriques considérées sont l'exactitude, la précision, le rappel, le F1-Score, ainsi que le nombre de vrais négatifs (TN), vrais positifs (TP), faux positifs (FP), et faux négatifs (FN).

Modèle	Exactitude	Précision	Rappel	F1-Score	TN	TP	FN	FP
MaxEnt	0.89	0.66	0.72	0.69	2661	422	163	217
RF	0.9	0.83	0.814	0.82	2719	399	185	157
SVM	0.88	0.8	0.778	0.79	2712	347	237	165
GLM	0.87	0.79	0.77	0.77	2683	358	226	194
GAM	0.89	0.8	0.79	0.8	2689	387	197	188
CNN	0.89	0.81	0.81	0.81	2680	403	181	197

Tableau 6:Tableau Comparatif des performances des modèles de distribution de Cenchrus biflorus.

Les résultats indiquent que le modèle Random Forest (RF) a la performance la plus élevée parmi les modèles évalués, avec une exactitude de 0.90, une précision de 0.83, un rappel de 0.814 et un F1-Score de 0.82. Ce modèle montre également un nombre élevé de vrais négatifs (2719) et un nombre relativement faible de faux positifs (157). Le modèle CNN, bien qu'ayant une précision et un rappel élevés, montre un nombre de faux négatifs (181) légèrement supérieur à celui du modèle RF, ce qui peut indiquer une légère tendance à manquer des présences de l'espèc

Les graphiques suivants de la figure 26 permettent de visualiser et de comparer les performances des différents modèles de prédiction pour l'espèce *Cenchrus biflorus*. Chaque graphique représente une métrique de performance spécifique : le rappel, le F1-Score, l'exactitude et la précision. Ces visualisations offrent une compréhension plus intuitive des différences de performances entre les modèles.

L'exactitude : Les modèles Random Forest (RF) et Convolutional Neural Network (CNN) montrent des exactitudes similaires et élevées. Cela signifie qu'ils font globalement moins d'erreurs de prédiction comparés aux autres modèles.

Précision : La précision la plus élevée est atteinte par le modèle RF (0.83), suivi de CNN (0.81). Cela indique que, parmi les présences prédites, la majorité sont correctes.

Rappel : Le modèle RF se distingue par son rappel (0.814), ce qui signifie qu'il est particulièrement efficace pour identifier les présences réelles de l'espèce. Le modèle MaxEnt, bien que performant en termes de rappel, présente une précision plus faible, ce qui signifie qu'il a tendance à prédire des présences incorrectes.

F1-Score : Le F1-Score, qui combine la précision et le rappel, est le plus élevé pour le modèle RF (0.82), soulignant son bon équilibre entre les deux métriques. Les autres modèles montrent des F1-Scores inférieurs, indiquant soit une précision soit un rappel moins efficace

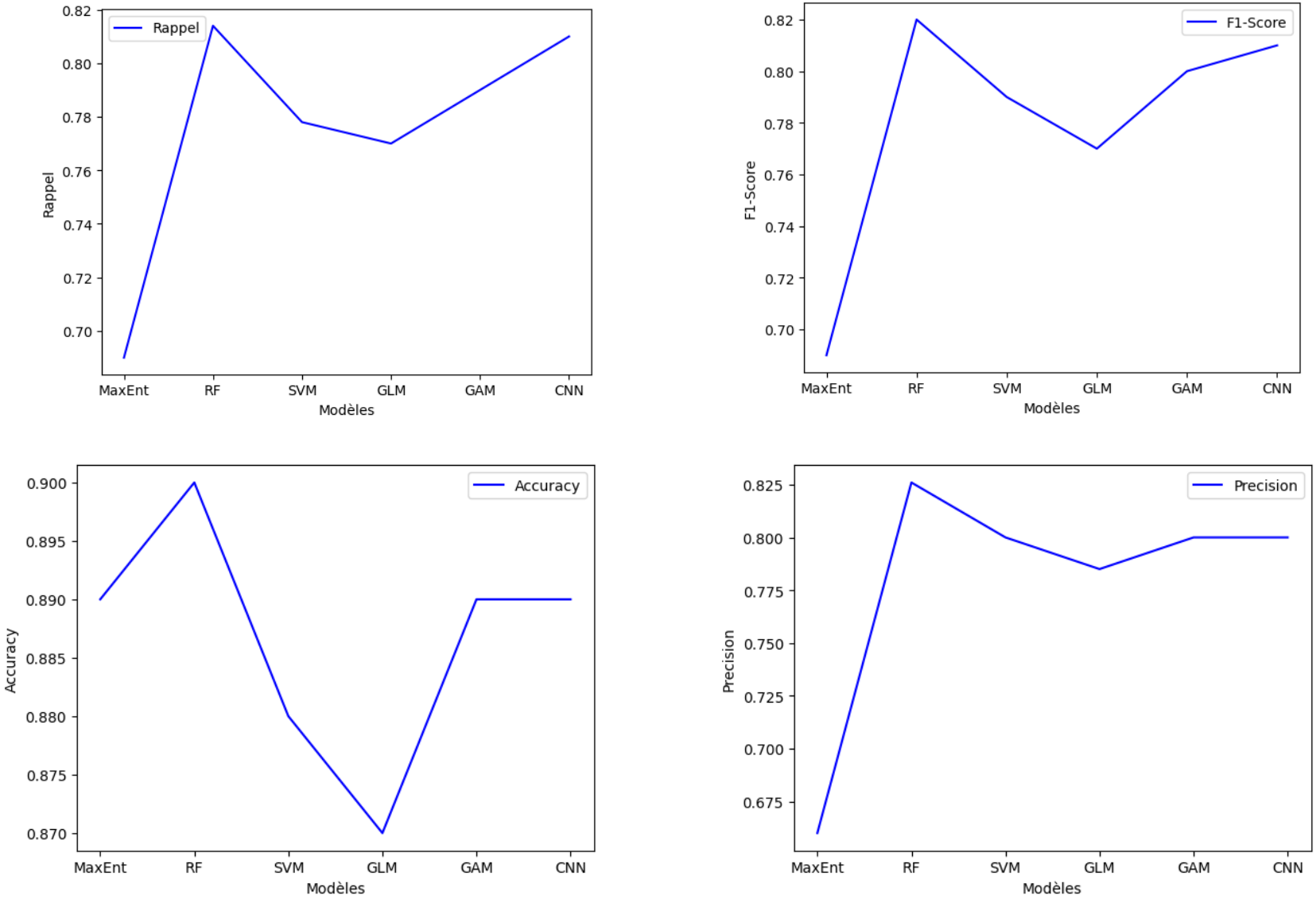


Figure 25. Comparaison de différentes métriques des différents modèles appliqués pour l'espèce Cenchrus biflorus.

En ce qui concerne les métriques spécifiques de TN, TP, FP, et FN qui sont présentés dans la figure 27 :

Vrais Négatifs (TN) : Le nombre de vrais négatifs est important pour comprendre combien d'absences réelles ont été correctement identifiées. Le modèle RF a le plus grand nombre de TN (2719), ce qui signifie qu'il est très efficace pour identifier correctement les absences de *Cenchrus biflorus*. Les autres modèles, bien que proches, montrent une performance légèrement inférieure.

Vrais Positifs (TP) : Le modèle CNN a le plus grand nombre de TP (403), ce qui signifie qu'il identifie correctement plus de présences de l'espèce. Cependant, ce modèle a une précision légèrement inférieure à celle de RF, ce qui signifie qu'il a également plus de faux positifs.

Faux Positifs (FP) : Les faux positifs représentent les présences incorrectement prédites. Le modèle RF a le plus petit nombre de FP (157), indiquant qu'il fait moins d'erreurs en prédisant la présence de l'espèce lorsqu'elle n'est pas réellement présente.

Faux Négatifs (FN) : Les faux négatifs représentent les présences réelles non détectées. Le modèle MaxEnt a le nombre de FN le plus bas (163), ce qui signifie qu'il manque moins de présences réelles. En revanche, le modèle RF a un nombre relativement élevé de FN (185), indiquant qu'il pourrait ne pas détecter certaines présences de l'espèce.

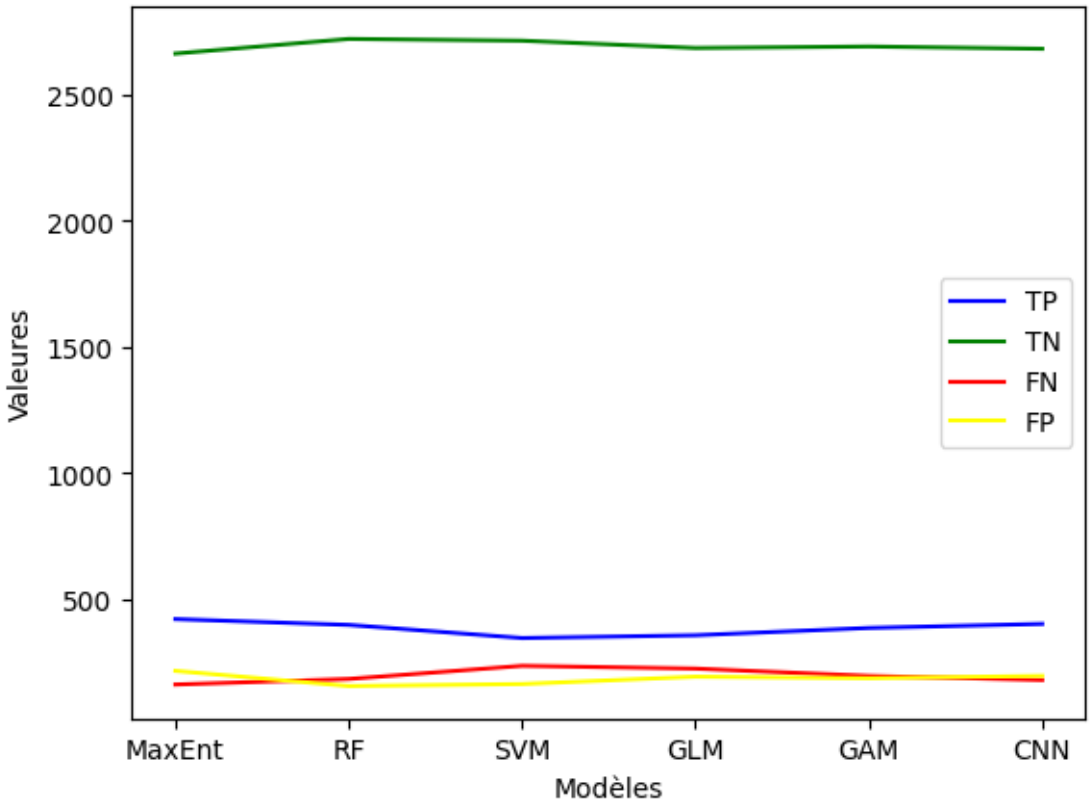


Figure 26. Comparaison du nombre de TP, TN, FN et FP des différents modèles pour Cenchrus biflorus.

En conclusion, le **modèle *Random Forest* (RF)** s'avère être le plus performant pour prédire la présence de *Cenchrus biflorus*, offrant un bon équilibre entre précision et rappel, et minimisant les erreurs de prédiction.

3.1.2 6ième espèce : *Balanites aegyptiaca*

Le tableau 7 présente les métriques de performance pour différents modèles de prédiction appliqués à l'espèce *Balanites aegyptiaca*. Les modèles évalués comprennent MaxEnt, *Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM), Generalized Additive Model (GAM), et Convolutional Neural Network (CNN).* Les métriques considérées sont l'exactitude, la précision, le rappel, le F1-Score, ainsi que le nombre de vrais négatifs (TN), vrais positifs (TP), faux positifs (FP), et faux négatifs (FN).

Tableau 7. Tableau Comparatif des Performances des Modèles de Distribution de Balanites aegyptiaca.

Modèle	Exactitude	Précision	Rappel	F1-Score	TN	TP	FP	FN
MaxEnt	0.84	0.44	0.59	0.51	2646	278	348	191
RF	0.89	0.76	0.73	0.74	2833	239	160	229
SVM	0.87	0.79	0.53	0.52	2981	30	12	439
GLM	0.86	0.68	0.54	0.54	2950	40	43	428
GAM	0.87	0.74	0.61	0.64	2918	111	75	375
CNN	0.88	0.77	0.66	0.69	2869	167	97	301

Les résultats indiquent que le modèle Random Forest (RF) a la performance la plus élevée parmi les modèles évalués, avec une exactitude de 0.89, une précision de 0.76, un rappel de 0.73 et un F1-Score de 0.74. Ce modèle montre également un nombre élevé de vrais négatifs (2833) et un nombre relativement faible de faux positifs (160). Le modèle CNN, bien qu'ayant une précision et un rappel élevés, montre un nombre de faux négatifs (301) légèrement supérieur à celui du modèle RF, ce qui peut indiquer une légère tendance à manquer des présences de l'espèce.

Les graphiques suivants de la figure 28 permettent de visualiser et de comparer les performances des différents modèles de prédiction pour l'espèce *Balanites aegyptiaca*. Chaque graphique représente une métrique de performance spécifique : l'exactitude, la précision, le

rappel, le F1-Score, ainsi que les valeurs de TN, TP, FP, et FN. Ces visualisations offrent une compréhension plus intuitive des différences de performances entre les modèles.

L'exactitude : Les modèles CNN et RF montrent des exactitudes similaires et élevées. Cela signifie qu'ils font globalement moins d'erreurs de prédiction comparés aux autres modèles.

Précision : La précision la plus élevée est atteinte par le modèle CNN (0.77), suivi de près par les modèles SVM, GLM et GAM (0.769), ce qui signifie que parmi les présences prédites, la majorité sont correctes.

Rappel : Le modèle RF se distingue par son rappel (0.73), ce qui signifie qu'il est particulièrement efficace pour identifier les présences réelles de l'espèce. Le modèle MaxEnt, bien que performant en termes de rappel, présente une précision plus faible, ce qui signifie qu'il a tendance à prédire des présences incorrectes.

F1-Score : Le F1-Score, qui combine la précision et le rappel, est le plus élevé pour le modèle CNN (0.74), soulignant son bon équilibre entre les deux métriques. Les autres modèles montrent des F1-Scores inférieurs, indiquant soit une précision soit un rappel moins efficace.

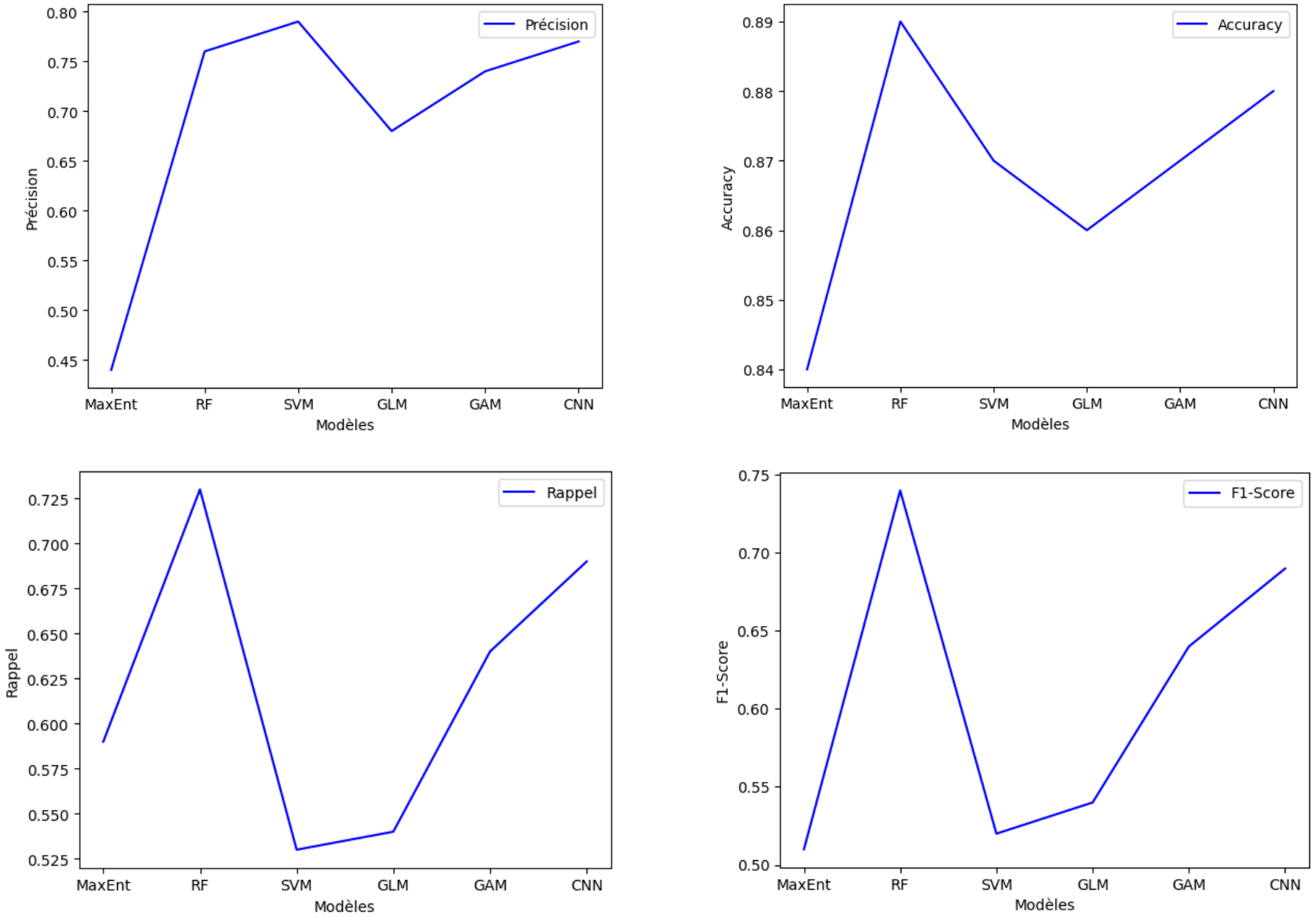


Figure 27. Comparaison de différentes métriques des différents modèles appliqués pour l'espèce *Balanites aegyptiaca*

En ce qui concerne les métriques spécifiques de TN, TP, FP, et FN sont présentées dans figure 29 comme suit :

Vrais Négatifs (TN) : Les valeurs des vrais négatifs (TN) sont les plus élevées pour le modèle SVM (2981), suivi de GLM (2950) et CNN (2869).

Vrais Positifs (TP) : En ce qui concerne les vrais positifs (TP), le modèle CNN a la valeur la plus élevée (167), suivi de MaxEnt (278).

Faux Positifs (FP) : Les faux positifs représentent les présences incorrectement prédites. Le modèle CNN a un nombre relativement élevé de FP (97), ce qui peut indiquer une tendance à prédire incorrectement la présence de l'espèce.

Faux Négatifs (FN) : Les faux négatifs représentent les présences réelles non détectées. Le modèle SVM a le nombre de FN le plus bas (12), ce qui signifie qu'il manque moins de présences réelles. En revanche, le modèle RF a un nombre relativement élevé de FN (229), indiquant qu'il pourrait ne pas détecter certaines présences de l'espèce.

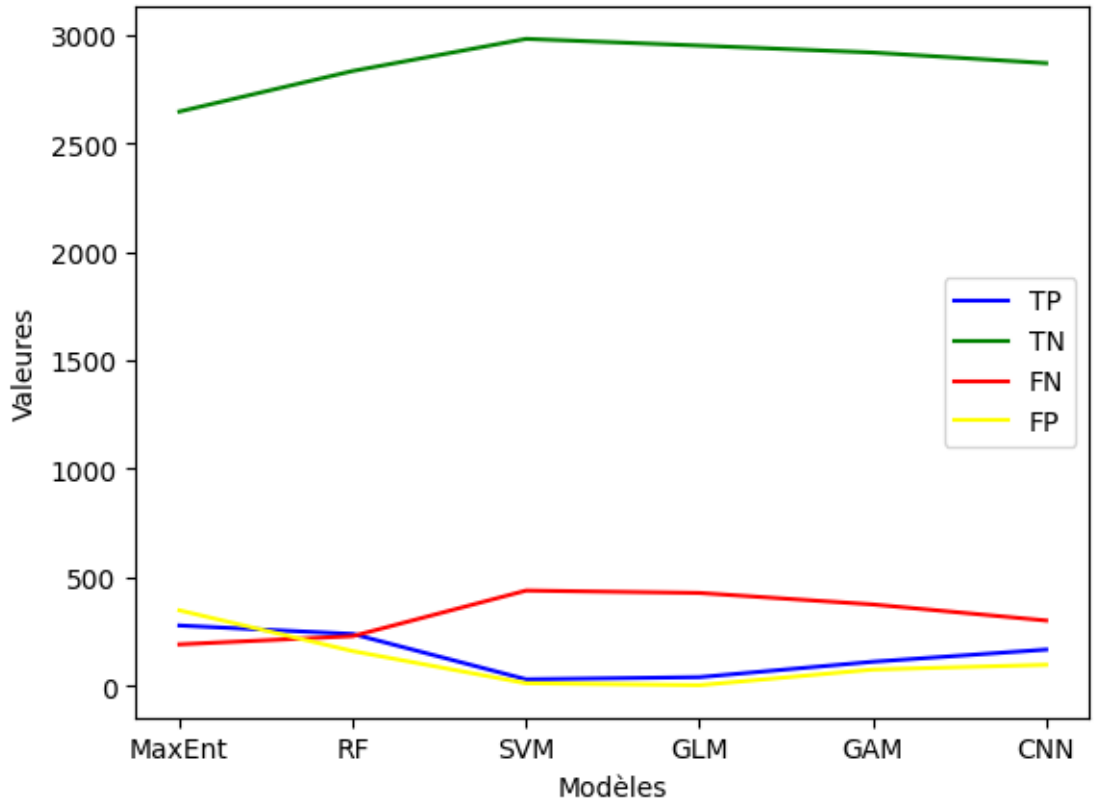


Figure 28. Comparaison des différentes métriques des modèles appliqués pour l'espèce *Balanites aegyptiaca*

Le modèle *Convolutional Neural Network* (CNN) s'avère être le plus performant pour prédire la présence de Balanites aegyptiaca, offrant un bon équilibre entre précision et rappel, et minimisant les erreurs de prédiction. Le modèle *Random Forest* (RF), bien qu'ayant une précision et un rappel élevés, présente une légère tendance à manquer certaines présences et à prédire incorrectement la présence de l'espèce. Les modèles SVM et GAM montrent également de bonnes performances, mais avec des limitations spécifiques.

Il est important de noter les différences entre les métriques calculées directement à partir des valeurs de la matrice de confusion (TN, TP, FP, FN) et celles obtenues par la validation croisée stratifiée. La validation croisée stratifiée calcule les métriques pour chaque bloc et les moyenne, ce qui peut lisser les valeurs extrêmes et offrir une estimation plus équilibrée des performances du modèle. En revanche, les métriques calculées à partir de la matrice de confusion unique agrègent toutes les prédictions des blocs, ce qui peut introduire des variations dues à des contributions inégales des blocs, particulièrement pour des métriques comme la précision et le F1-score, qui sont sensibles aux déséquilibres et variations entre les blocs de validation.

3.2 Choix du modèle optimal par espèce

L'analyse des performances des différents modèles de prédiction pour les espèces *Cenchrus biflorus* et *Balanites aegyptiaca* montre que les modèles ***Random Forest (RF)*** *et* ***Convolutional Neural Network (CNN)*** sont globalement les plus performants, bien que chaque modèle présente ses propres forces et faiblesses.

Pour *Cenchrus biflorus*, le modèle RF se distingue par son excellent équilibre entre précision et rappel, ainsi qu'un F1-Score élevé. Cela indique que le modèle est capable de prédire efficacement la présence et l'absence de l'espèce avec un nombre minimal d'erreurs. Le modèle CNN présente également de bonnes performances, notamment en termes de rappel et de précision, mais montre une légère tendance à manquer certaines présences de l'espèce.

Pour *Balanites aegyptiaca*, le modèle CNN a les meilleures performances globales, avec des valeurs élevées en précision, rappel et F1-Score, indiquant une bonne détection des présences et absences de l'espèce. Le modèle RF montre également de bons résultats, mais avec une tendance à manquer certaines présences de l'espèce. Les modèles SVM et GAM offrent un bon équilibre entre précision et rappel, mais présentent certaines limitations spécifiques.

3.3 Résultats des différents modèles avec équilibrage de classes et avec sélection de variables

Après avoir équilibré nos classes, pour améliorer la performance des modèles testés, nous avons utilisé plusieurs méthodes de sélection de variables, à savoir l'importance des caractéristiques, l'information mutuelle et l'analyse en composantes principales. Ceci afin d'identifier les variables les plus pertinentes et réduire la dimensionnalité de nos données d'entrée.

3.3.1 Résultats avec la méthode se basant sur l'importance des caractéristiques

Pour identifier les variables les plus importantes affectant la distribution de *Cenchrus biflorus* et *Balanites aegyptiaca*, nous avons utilisé la méthode d'importance des caractéristiques avec le modèle 'Random Forest'. Cette méthode permet de quantifier l'importance de chaque variable en termes de contribution à la précision du modèle.

Pour chaque espèce, nous avons entraîné un modèle Random Forest pour estimer l'importance des caractéristiques.

Un modèle de sélection basé sur les caractéristiques importantes (SelectFromModel) a été appliqué en utilisant un seuil fixé à la moyenne.

a- Pour *Cenchrus biflorus*

Les caractéristiques sélectionnées, présentées par ordre d'importance sur la figure 30, sont présentés sur le tableau 8.

Tableau 8. Importance des Caractéristiques Sélectionnées pour Cenchrus biflorus

Année
Latitude (lat)
Longitude (long)
Température moyenne annuelle (Annual_Mean_Temperature)
Température moyenne du trimestre le plus humide (Mean_Temperature_Wettest_Quarter)
Précipitations annuelles (Annual_Precipitation)
Précipitations du mois le plus humide (Precipitation_Wettest_Month)
Saison des précipitations (Precipitation_Seasonality)
Précipitations du trimestre le plus humide (Precipitation_Wettest_Quarter)
Température maximale annuelle (Tmax_annuelle)
Vitesse du vent annuelle (Vitesse_vent_annuelle)
Précipitations (Precipitation)
Type de sol (FAOSOIL_freq_encoded)

Le nombre de caractéristiques sélectionnées est égal à treize.

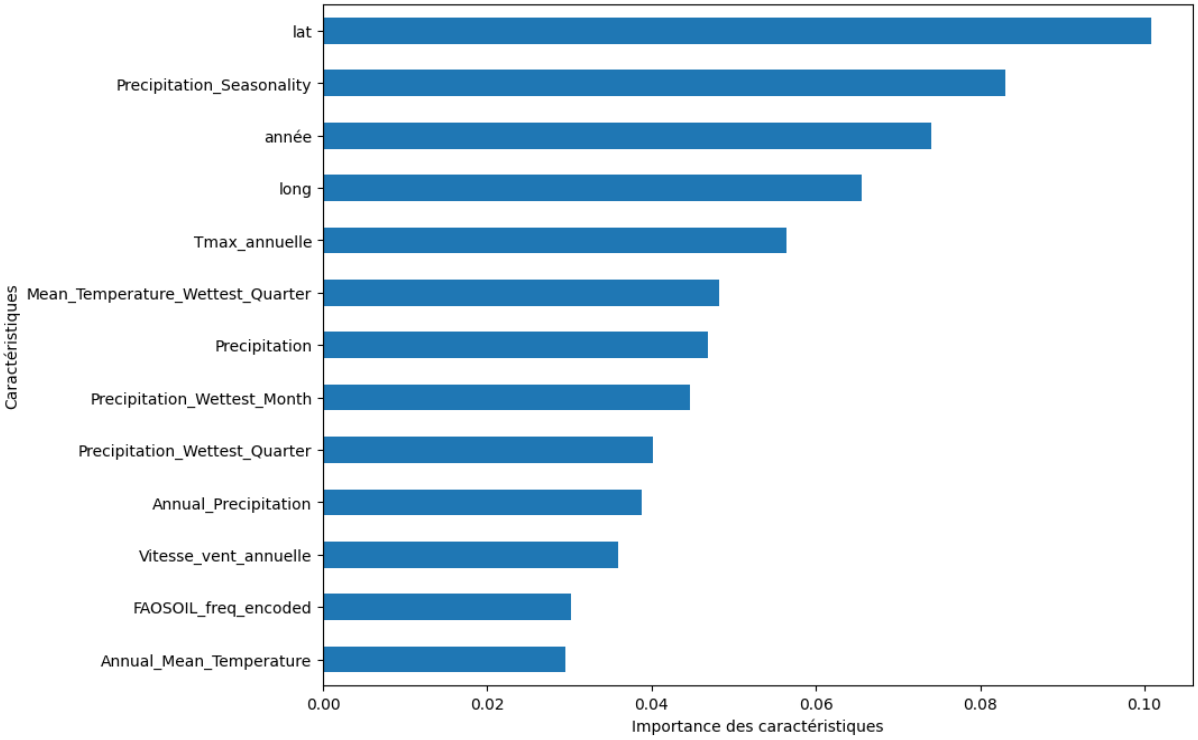


Figure 29. Importance des caractéristiques dans le modèle Random Forest pour l'espèce Cenchrus biflorus.

b- Pour *Balanites aegyptiaca* :

Les caractéristiques sélectionnées, présentées par ordre d'importance sur la figure 31, sont présentés sur le tableau 9.

Tableau 9. Importance des Caractéristiques Sélectionnées pour *Balanites aegyptiaca*

Année
Latitude (lat)
Longitude (long)
Température moyenne annuelle (Annual_Mean_Temperature)
Écart diurne moyen (Mean_Diurnal_Range)
Température maximale du mois le plus chaud (Max_Temperature_Warmest_Month)
Saison des précipitations (Precipitation_Seasonality)
Température maximale annuelle (Tmax_annuelle)
Radiation solaire annuelle (Solar_radiation_annuelle)

Le nombre de caractéristiques sélectionnées est égal à neuf.

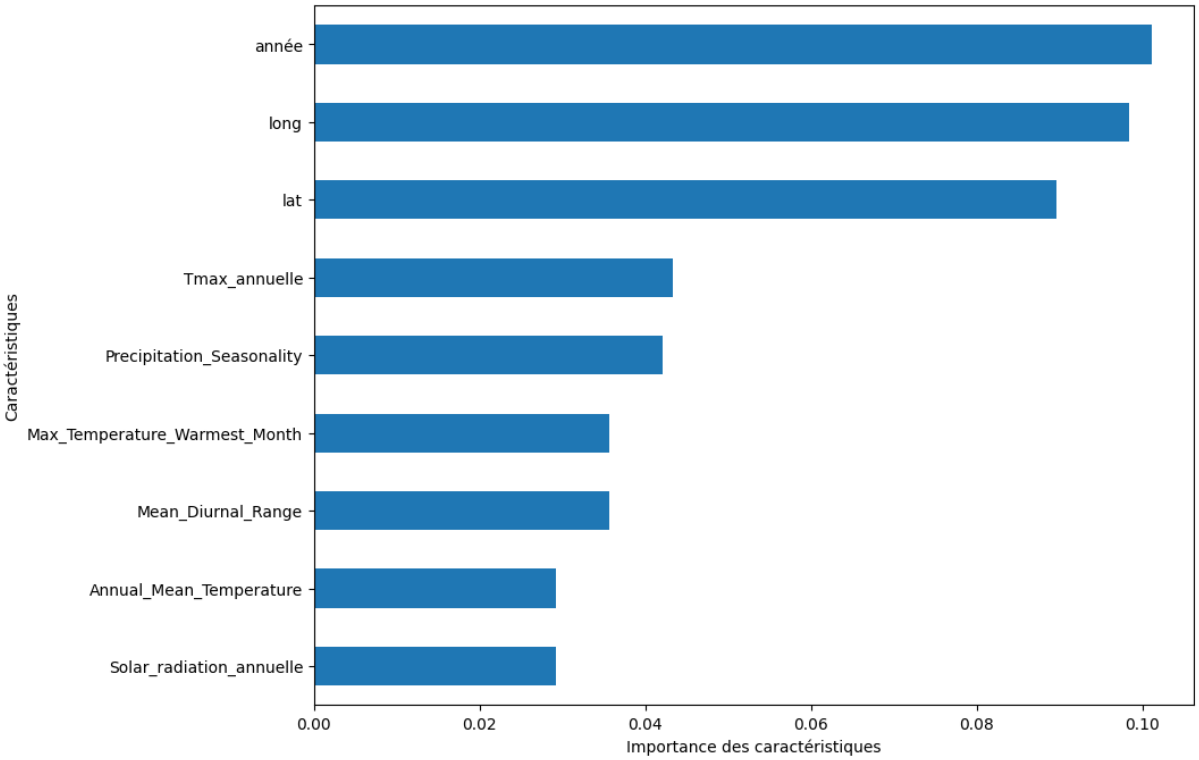


Figure 30. Importance des caractéristiques dans le modèle Random Forest pour l'espèce *Balanites aegyptiaca*

3.3.2 Résultats avec la méthode se basant sur l'information mutuelle

L'information mutuelle a été utilisée pour évaluer la dépendance entre les variables environnementales et la présence des espèces. Cette méthode mesure la quantité d'information partagée entre deux variables, permettant de détecter les relations non linéaires entre les variables et la présence des espèces.

L'information mutuelle entre chaque variable environnementale et la présence de l'espèce a été calculée

Les variables avec des valeurs d'information mutuelle élevée ont été sélectionnées comme étant les plus pertinentes pour la prédiction de la distribution des espèces.

Les graphiques ci-dessous montrent l'importance des variables sélectionnées pour les espèces *Cenchrus biflorus*, figure 32 et *Balanites aegyptiaca*, figure 33, en utilisant la méthode

de l'information mutuelle. Ces graphiques permettent d'identifier les variables environnementales les plus influentes pour la prédiction de la présence de chaque espèce.

a. Pour *Cenchrus biflorus*

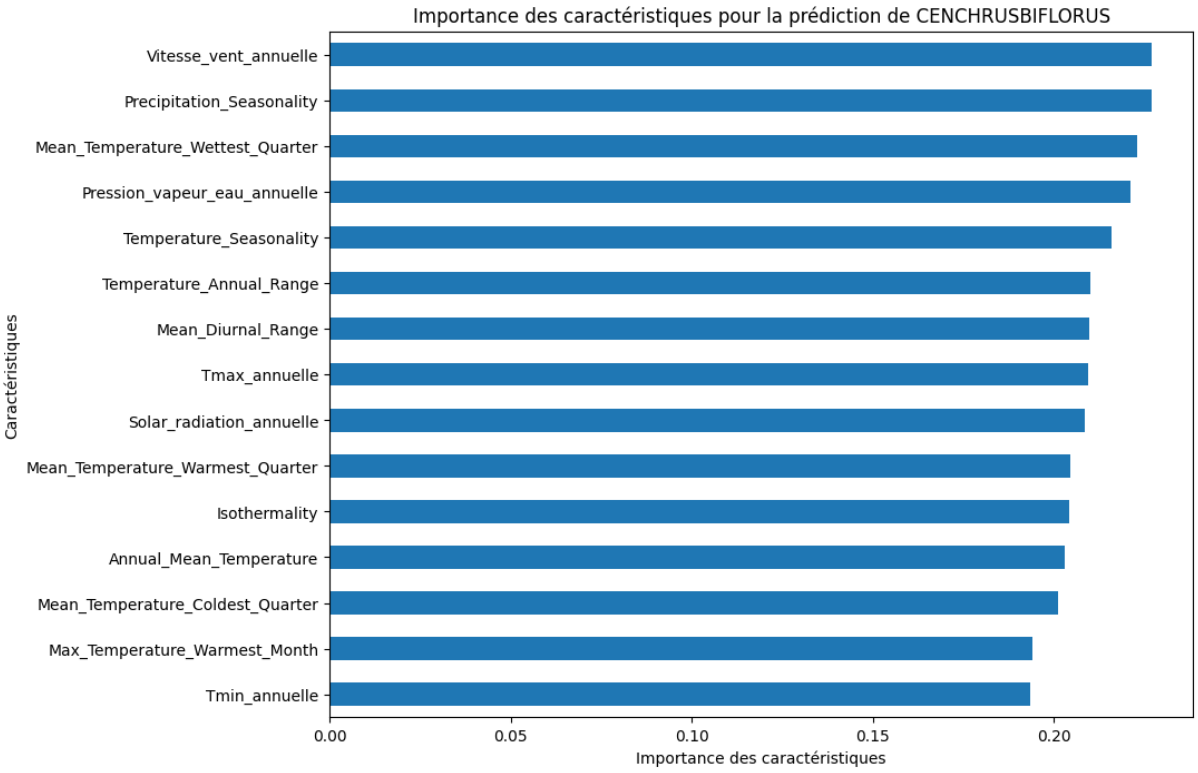


Figure 31.Importance des Caractéristiques Sélectionnées pour Cenchrus biflorus

Les caractéristiques les plus importantes pour la prédiction de *Cenchrus biflorus* incluent :

-**Vitesse du vent annuelle (Vitesse_vent_annuelle)** : La vitesse du vent annuelle est une variable clé, indiquant peut-être que cette espèce est influencée par les conditions venteuses.

-**Saison des précipitations (Precipitation_Seasonality)** : La saisonnalité des précipitations est également cruciale, suggérant que la distribution de l'espèce est fortement liée aux variations saisonnières des précipitations.

-**Température moyenne du trimestre le plus humide (Mean_Temperature_Wettest_Quarter)** : Cette variable indique que les températures durant les périodes les plus humides sont significatives pour l'espèce.

-Pression de la vapeur d'eau annuelle (Pression_vapeur_eau_annuelle) : La pression de la vapeur d'eau annuelle est une autre variable importante, reflétant peut-être l'humidité de l'air qui peut affecter la présence de l'espèce.

-Saison des températures (Temperature_Seasonality) : La saisonnalité des températures montre également une importance notable, suggérant une adaptation de l'espèce aux variations de température au cours de l'année.

Ces variables montrent que *Cenchrus biflorus* est sensible à une variété de facteurs climatiques, y compris la température, les précipitations et le vent.

b. Pour *Balanites aegyptiaca*

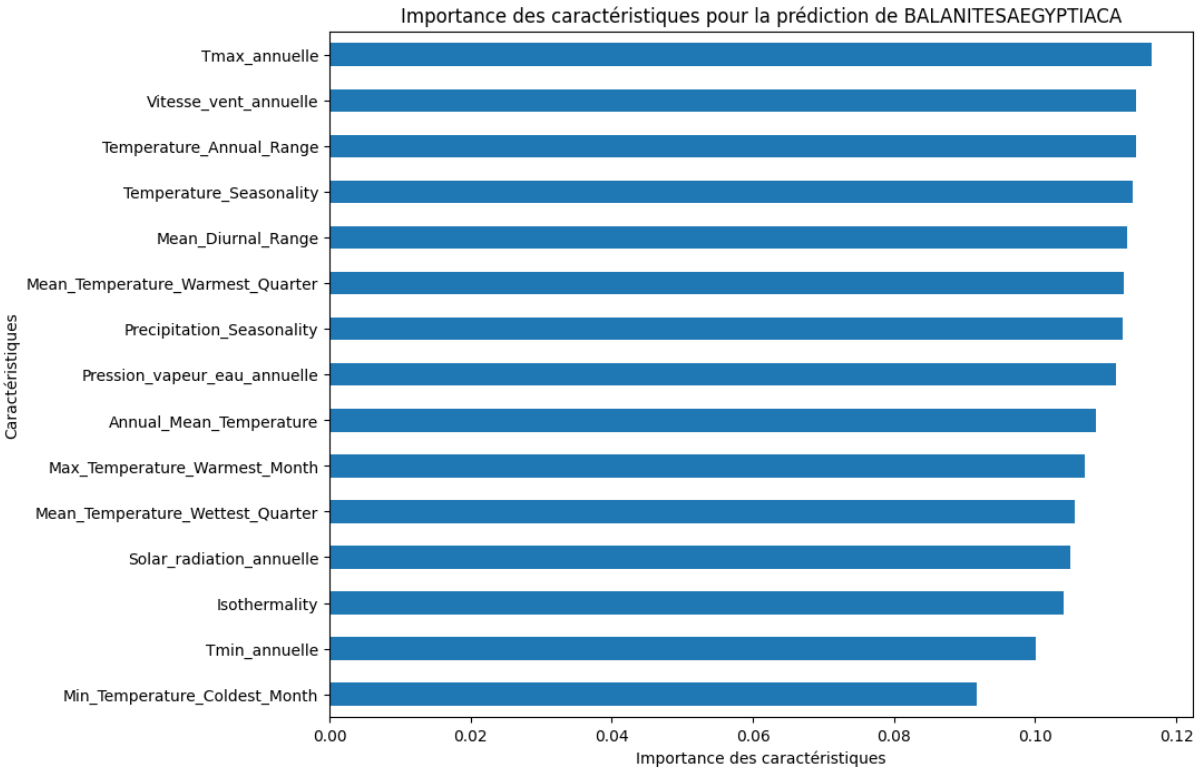


Figure 32. Importance des Caractéristiques Sélectionnées pour Balanites aegyptiaca.

Les caractéristiques les plus importantes pour la prédiction de *Balanites aegyptiaca* incluent :

-**Température maximale annuelle (Tmax_annuelle)** : La température maximale annuelle est la variable la plus influente, indiquant que cette espèce est particulièrement affectée par les températures élevées.

-Vitesse du vent annuelle (Vitesse_vent_annuelle) : Comme pour *Cenchrus biflorus*, la vitesse du vent annuelle est une variable importante pour Balanites aegyptiaca, suggérant une influence du vent sur cette espèce également.

-Écart de température annuel (Temperature_Annual_Range) : L'écart de température annuel montre l'importance des variations de température pour cette espèce.

-**Saison des températures (Temperature_Seasonality)** : La saisonnalité des températures est également une variable cruciale, indiquant que cette espèce est affectée par les fluctuations saisonnières des températures.

-Écart diurne moyen (Mean_Diurnal_Range) : L'écart diurne moyen, ou la différence entre les températures diurnes et nocturnes, est significatif pour cette espèce.

Les résultats montrent que *Balanites aegyptiaca* est influencée par des variables climatiques similaires à celles de *Cenchrus biflorus*, avec une importance particulière accordée aux températures maximales et aux variations de température.

3.3.3 Résultats de l'Analyse en Composantes Principales (ACP)

Pour réduire la dimensionnalité des données tout en conservant l'essentiel de l'information, nous avons appliqué l'Analyse en Composantes Principales (ACP). Cette méthode nous a permis de transformer les variables d'origine en un ensemble de nouvelles variables non corrélées (composantes principales), tout en conservant le maximum de variance possible des données d'origine.

Les données ont été normalisées pour s'assurer que toutes les variables sont sur une échelle comparable. L'ACP a été appliquée sur les données normalisées pour extraire les composantes principales. Les composantes principales avec les plus grandes variances ont été retenues pour l'analyse.

Pour les deux espèces les 20 premières composantes principales ont été retenu vu qu'il expliquaient plus de 95% de la variabilité du modèle.

En conclusion, l'Analyse en Composantes Principales (ACP) a été choisie pour son efficacité à réduire la dimensionnalité tout en capturant l'essentiel de l'information, son impact positif sur les performances des modèles, et sa capacité à offrir une méthodologie consistante et reproductible. L'ACP a permis de simplifier nos modèles, de minimiser le risque de surapprentissage et d'améliorer la précision des prédictions, rendant nos analyses plus robustes et fiables. Ces avantages font de l'ACP un choix optimal pour la sélection des variables environnementales dans nos modèles de prédiction de la distribution des espèces.

3.4 Evaluation des modèles après équilibrage des classes et sélection des variables par ACP

Après avoir équilibré les classes en appliquant le sous-échantillonnage et après avoir appliqué l'Analyse en Composantes Principales (ACP) pour sélectionner les 20 composantes principales représentant 95 % de la variabilité, nous obtenons les résultats présentés dans le tableau 10 pour le modèle *Random Forest* appliqué à *Cenchrus biflorus* et pour le modèle CNN appliqué à *Balanites aegyptiaca.*

Tableau 10. Performances des deux Modèles RF et CNN Après Équilibrage des Classes et Sélection des Variables

Modèle/Métrique	RF appliquée à *Cenchrus biflorus*	CNN appliquée à *Balanites aegyptica*
Exactitude	0.86	0.76
Précision	0.76	0.66
Rappel	0.87	0.81
F1-Score	0.79	0.67
TP	524	411
TN	2458	2212
FP	419	781
FN	60	57

3.4.1 Modèle *Random Forest* appliqué à *Cenchrus biflorus* :

- **Exactitude** : Avec une exactitude de 0.86, le modèle *Random Forest* montre une bonne performance globale, indiquant qu'il fait relativement peu d'erreurs de prédiction.
- **Précision** : La précision de 0.76 signifie que, parmi les présences prédites, 76 % sont correctes. Cela montre que le modèle est raisonnablement bon pour éviter les fausses alertes.
- **Rappel** : Un rappel élevé de 0.87 indique que le modèle est très efficace pour identifier les présences réelles de l'espèce, ce qui est crucial pour les études de biodiversité.

- **F1-Score** : Le F1-Score de 0.79 reflète un bon équilibre entre la précision et le rappel, montrant que le modèle gère bien les deux aspects.

- **Vrais Positifs (TP)** : 524, ce qui montre que le modèle identifie correctement un grand nombre de présences de *Cenchrus biflorus*

- **Vrais Négatifs (TN)** : 2458, indiquant une bonne capacité à identifier correctement les absences.

- **Faux Positifs (FP)** : 419, ce qui est un nombre modéré et indique que le modèle fait quelques erreurs en prédisant la présence quand il n'y en a pas.

- **Faux Négatifs (FN)** : 60, un nombre assez faible, montrant que le modèle manque rarement une présence réelle.

3.4.2 Modèle CNN appliqué à *Balanites aegyptiaca* :

- **Exactitude** : Avec une exactitude de 0.76, le modèle CNN montre une performance acceptable, bien que plus faible que le modèle RF pour *Cenchrus biflorus*.

- **Précision** : La précision de 0.66 indique que, parmi les présences prédites, 66 % sont correctes. Cela montre une plus grande tendance à produire des fausses alertes par rapport au modèle RF.

- **Rappel** : Un rappel élevé de 0.81 indique que le modèle est efficace pour identifier les présences réelles de l'espèce.

- **F1-Score :** Le F1-Score de 0.67 montre que le modèle maintient un bon équilibre entre la précision et le rappel, bien que légèrement inférieur au modèle RF.

- **Vrais Positifs (TP)** : 411, indiquant que le modèle identifie correctement un grand nombre de présences de *Balanites aegyptiaca.*

- **Vrais Négatifs (TN)** : 2212, indiquant une bonne capacité à identifier correctement les absences, bien que légèrement inférieure au modèle RF.

- **Faux Positifs (FP)** : 781, ce qui est assez élevé, montrant que le modèle fait beaucoup d'erreurs en prédisant la présence quand il n'y en a pas.

- **Faux Négatifs (FN)** : 57, un nombre assez faible, montrant que le modèle manque rarement une présence réelle.

Conclusion

Les résultats après l'application de l'équilibrage des classes et la sélection des variables montrent que le modèle *Random Forest* pour *Cenchrus biflorus* et le modèle CNN pour *Balanites aegyptiaca* ont des performances solides, bien que chaque modèle présente des forces et des faiblesses spécifiques, principalement dues à la qualité des données d'occurrences utilisées. Le modèle *Random Forest* pour *Cenchrus biflorus* offre une meilleure précision globale et un bon équilibre entre les métriques, tandis que le modèle CNN pour *Balanites aegyptiaca* montre une meilleure capacité de rappel mais avec une précision légèrement inférieure. Ces résultats confirment l'importance de l'équilibrage des classes et de la sélection des variables pour améliorer la performance des modèles de prédiction de la distribution des espèces.

3.5 Prédiction et élaboration de cartes de distribution

3.5.1 Préparation des données

Pour établir des cartes de distribution des espèces sur l'ensemble de la région du Sahel, nous utilisons les mêmes variables que notre modèle de base ; les données bioclimatiques, climatiques, les données de type de sol ainsi que l'indice humidité. Afin de maintenir la cohérence avec notre modèle, nous avons compilé une base de données intégrant toutes les données environnementales pertinentes pour l'ensemble de notre zone d'étude avec une résolution de 0.1 degré en latitude et en longitude.

Cette approche permet de capturer les variations environnementales fines nécessaires pour des prédictions précises de la distribution des espèces. Ainsi on obtient une première base de données avec 12 758 enregistrements et 31 variables similaires à celles de notre modèle de base et qui contient toutes les variables sauf les 4 colonnes relatif à l'indice d'humidité qui ont été importé dans une autre base de données. Pour avoir des cartes annuelles notre première base de données regroupée est fusionnée avec la base de données des indices d'humidités, en effet pour chaque année on rajoute l'indice d'humidité relative à cette année pour avoir une base de données contenant toutes nos variables explicatives.

3.5.2 Entraînement et évaluation des modèles

Pour la totalité des espèces, le modèle *Random Forest* avec sélection de variables et équilibrage des classes semble offrir les meilleures performances globales pour les espèces herbacées, tandis que le modèle *Convolutional Neural Network* (CNN) montre de bonnes performances pour les espèces arborées.

Nous avons choisi de travailler avec le modèle *Random Forest* avec sélection de variables et équilibrage des classes pour les espèces herbacées, et le modèle CNN pour les espèces arborées. Cette décision est motivée par plusieurs raisons liées à la qualité et à la précision des données, ainsi qu'à notre hypothèse sur les pseudo-absences. Nous préférons un modèle qui puisse offrir une meilleure couverture des présences potentielles, même au risque d'augmenter le nombre de faux positifs.

L'équilibrage des classes permet au modèle de mieux gérer les déséquilibres dans les données de présences et d'absences, en évitant que le modèle ne soit biaisé par les zones où les données de présence sont plus abondantes. Cela nous permet de générer des prédictions plus robustes et d'éviter de limiter notre prédiction aux zones où nous avons le plus de certitudes quant à la présence des espèces.

Les données ont été chargées et prétraitées, incluant le retrait des colonnes non nécessaires et le traitement des valeurs manquantes. Les données environnementales ont été standardisées et les variables catégorielles ont été encodées en utilisant une méthode d'encodage par fréquence. Ensuite, nous avons appliqué une Analyse en Composantes Principales pour réduire la dimensionnalité des données tout en conservant les informations les plus pertinentes.

Pour chaque espèce, nous avons utilisé une validation croisée stratifiée pour entraîner et évaluer le modèle. Les résultats de la validation croisée pour chaque espèce ont été agrégés pour calculer les métriques de performance moyenne. Puis, les modèles entraînés ont été sauvegardés pour chaque espèce afin de pouvoir être utilisés pour des prédictions ultérieures. Cela inclut la sauvegarde du modèle, du scaler utilisé pour standardiser les données, et de l'ACP pour la réduction de la dimensionnalité.

3.5.3 Génération des prédictions et cartes de distribution des espèces

Les modèles sauvegardés ont été chargés pour prédire la distribution des espèces pour chaque année. Les résultats de prédiction ont été visualisés sur des cartes montrant la présence prédite des espèces.

Les cartes de distribution générées montrent les zones prévues de présence des espèces pour chaque année, offrant ainsi une visualisation claire et intuitive des prédictions de distribution. Ces cartes peuvent être utilisées pour des analyses plus approfondies de la répartition des espèces et pour élaborer des stratégies de conservation adaptées.

Les figures 34 et 35 illustrent les prédictions de distribution des espèces obtenues à partir des modèles *Random Forest* appliquée à *Cenchrus biflorus*et et CNN appliquée à *Balanites aegyptica*, respectivement. Les cartes montrent les variations spatiales des présences prédites, permettant d'identifier les zones potentiellement favorables pour chaque espèce.

Dans la figure 34, l'évolution de la répartition de *Cenchrus biflorus* au cours des années révèle une expansion significative et une densification croissante de sa présence dans la région du Sahel. En 1920, la présence de l'espèce est prédite de manière limitée, principalement dans les parties centrale et orientale du Sahel. En 1950, une augmentation notable de la répartition est observée, avec une extension vers l'ouest et l'est, marquant une expansion géographique significative par rapport à 1920.

En 1970, la distribution de *Cenchrus biflorus* s'étend davantage, couvrant presque toute la région du Sahel avec une concentration dense dans les zones centrale et occidentale. Cette tendance se poursuit en 1980, où la densité de présence augmente encore, indiquant que l'espèce colonise de plus en plus de territoires dans la région.

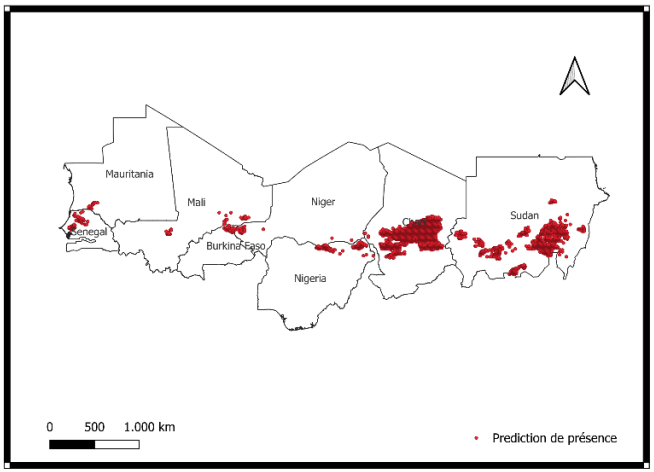
L'an 2000 marque une période où la distribution maintient une haute densité de présence à travers la majeure partie du Sahel, avec une présence particulièrement forte dans les zones centrale et occidentale. Enfin, en 2012, la présence de *Cenchrus biflorus* est dense et largement répartie à travers tout le Sahel, indiquant une stabilité et une prolifération continue de l'espèce au fil des décennies.

L'évolution de la répartition de *Balanites aegyptiaca,* dans la figure 35, au cours des années montre une tendance similaire à celle de *Cenchrus biflorus,* avec une expansion progressive et une densification de la présence de l'espèce dans la région du Sahel. En 1920, la présence de *Balanites aegyptiaca* est limitée, avec des observations prédominantes dans les parties centrale et orientale du Sahel, particulièrement au Niger et au Soudan.

En 1950, la répartition s'étend significativement vers l'ouest et le centre, couvrant désormais de grandes parties du Sahel, y compris la Mauritanie, le Mali et le Burkina Faso. En 1970, la distribution devient encore plus étendue, avec une concentration dense de la présence à travers toute la région du Sahel, démontrant une augmentation continue de la densité de présence.

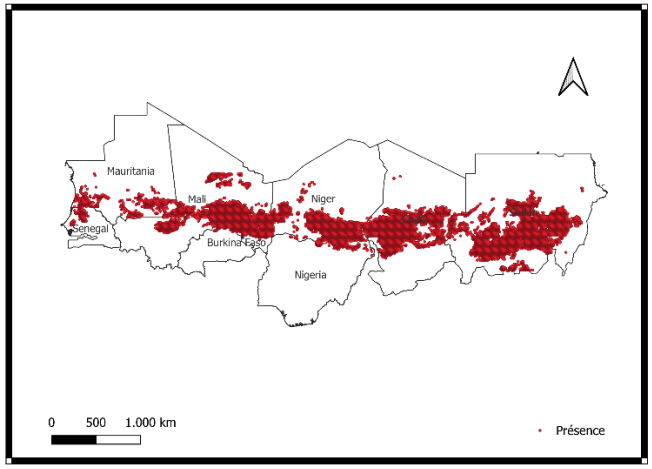
En 1980, la présence prédite de *Balanites aegyptiaca* montre une densité encore plus grande, s'étendant largement à travers le Sahel, y compris les régions occidentales et centrales. En 2000, la répartition de l'espèce reste étendue et dense, couvrant la plupart des zones du Sahel, avec une présence particulièrement forte dans les régions centrale et occidentale.

Enfin, en 2012, la présence de *Balanites aegyptiaca* est dense et largement répartie à travers toute la région du Sahel, montrant une stabilité et une prolifération continue de l'espèce au fil des décennies



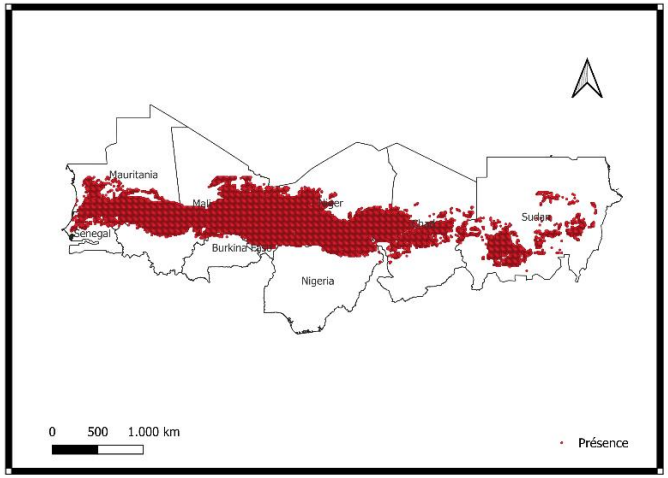
Cartes de la distribution de Cenchrus biflorus à 1920.

(a)



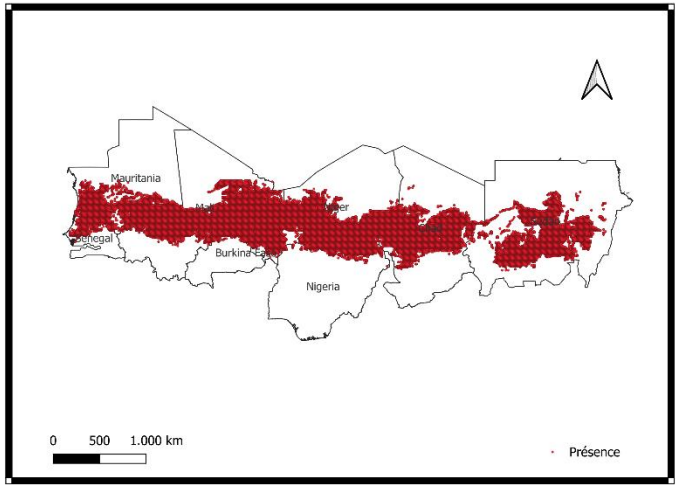
Cartes de la distribution de Cenchrus biflorus à 1950.

(b)



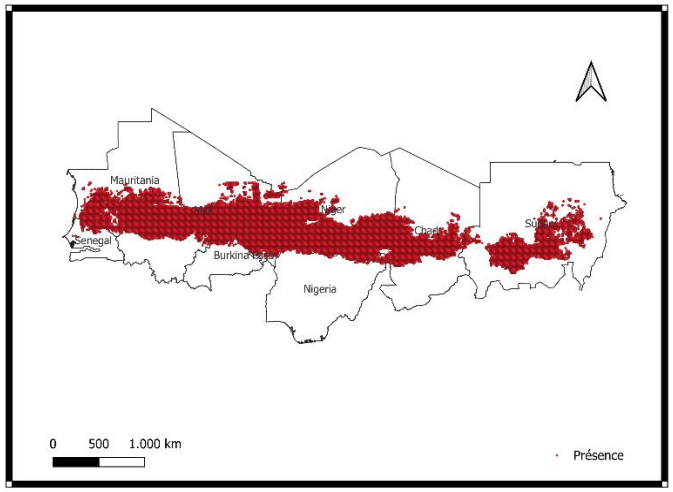
Cartes de la distribution de Cenchrus biflorus à 1970.

(c)



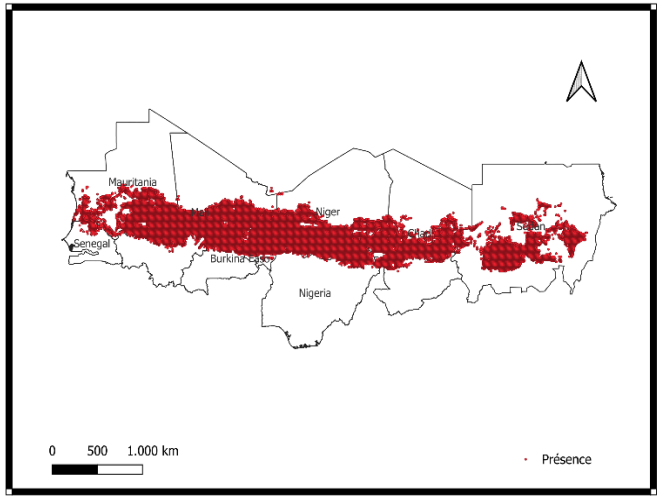
Cartes de la distribution de Cenchrus biflorus à 1980.

(d)



Cartes de la distribution de Cenchrus biflorus à 2000.

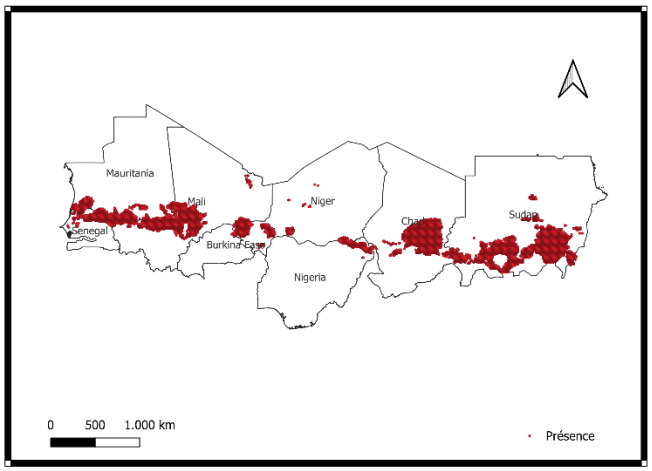
(e)



Cartes de la distribution de Cenchrus biflorus à 2012.

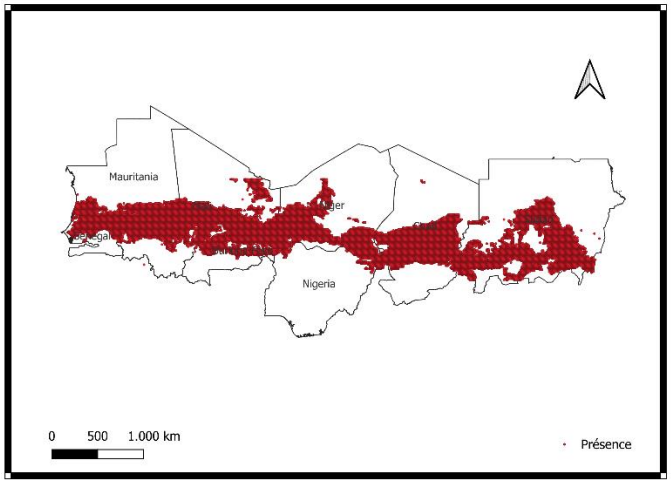
(f)

Figure 33. Cartes de la distribution spatio-temporelle de Cenchrus biflorus de 1950 à 2012 *au niveau du Sahel* .



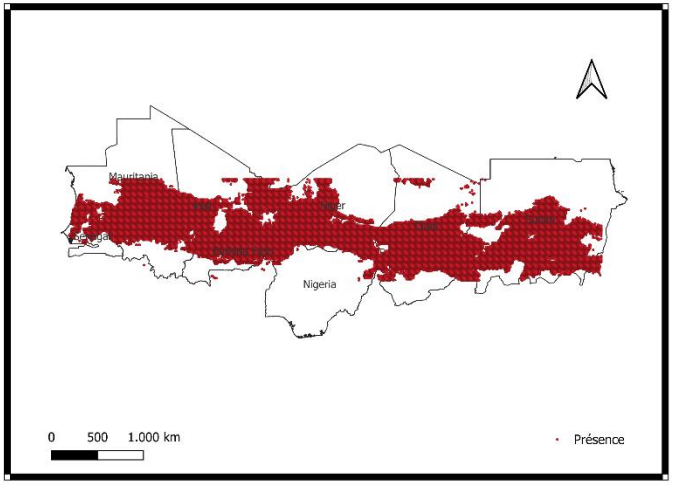
Cartes de la distribution de Balanites aegyptiaca à 1920.

(a)



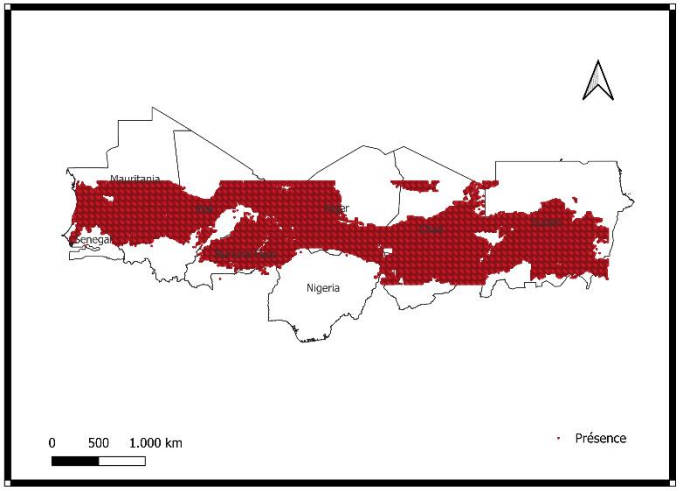
Cartes de la distribution de Balanites aegyptiaca à 1950.

(b)

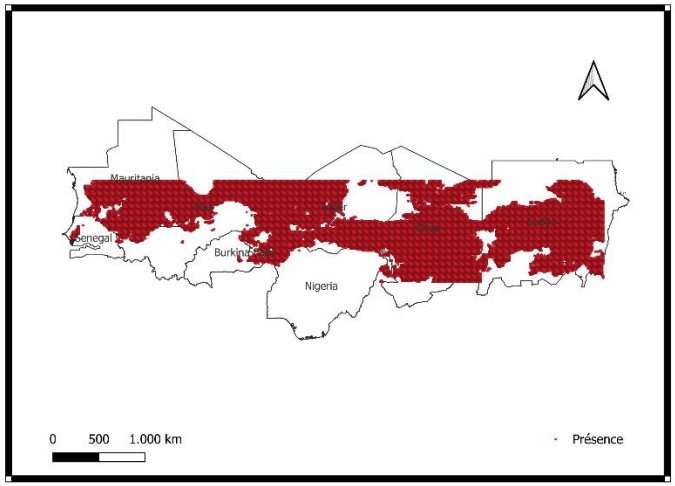


Cartes de la distribution de Balanites aegyptiaca à 1970.

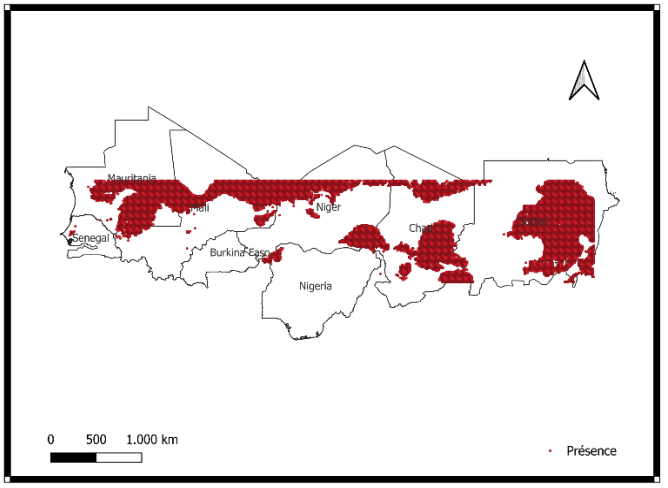
(c)



Cartes de la distribution de Balanites aegyptiaca à 1980.



Cartes de la distribution de Balanites aegyptiaca à 2000.



Cartes de la distribution de Balanites aegyptiaca à 2012.

Figure 34.. Cartes de la distribution spatio-temporelle de *Balanites aegyptiaca au niveau du Sahel.*

3.6 Validation par les données de l'herbier :

Pour valider les cartes élaborées, nous utilisons les données de présence d'espèce présente sur l'herbier disponible sur le site du CIRAD. Cette base de données contient les enregistrements qui marquent la présence des espèces végétales sur lesquelles nous travaillons sur la zone du Sahel et dont le nombre d'observation par espèce est présenté sur la figure 26. Elle contient 23 101 enregistrements avec 23 colonnes englobant les différentes présences des espèces végétales sur le Sahel entre 1899 et 2012.

Après avoir organisé et prétraité la base de données, nous avons éliminé les doublons, géré les données manquantes et codé les données. Parmi les 23 colonnes, nous nous intéressons uniquement aux quatre suivantes :

- **Taxon** : qui représente le nom de l'espèce en question.
- **Longitude et Latitude** : qui représentent les coordonnées de l'enregistrement qui marque la présence d'une espèce donnée.
- **Date Collecte** : cette colonne représente la date exacte de collecte de la donnée parfois en année ou en mois. Ainsi, on remarque que les données sont plus disponibles entre 1960 et 1980, comme justifié par la Figure 37.

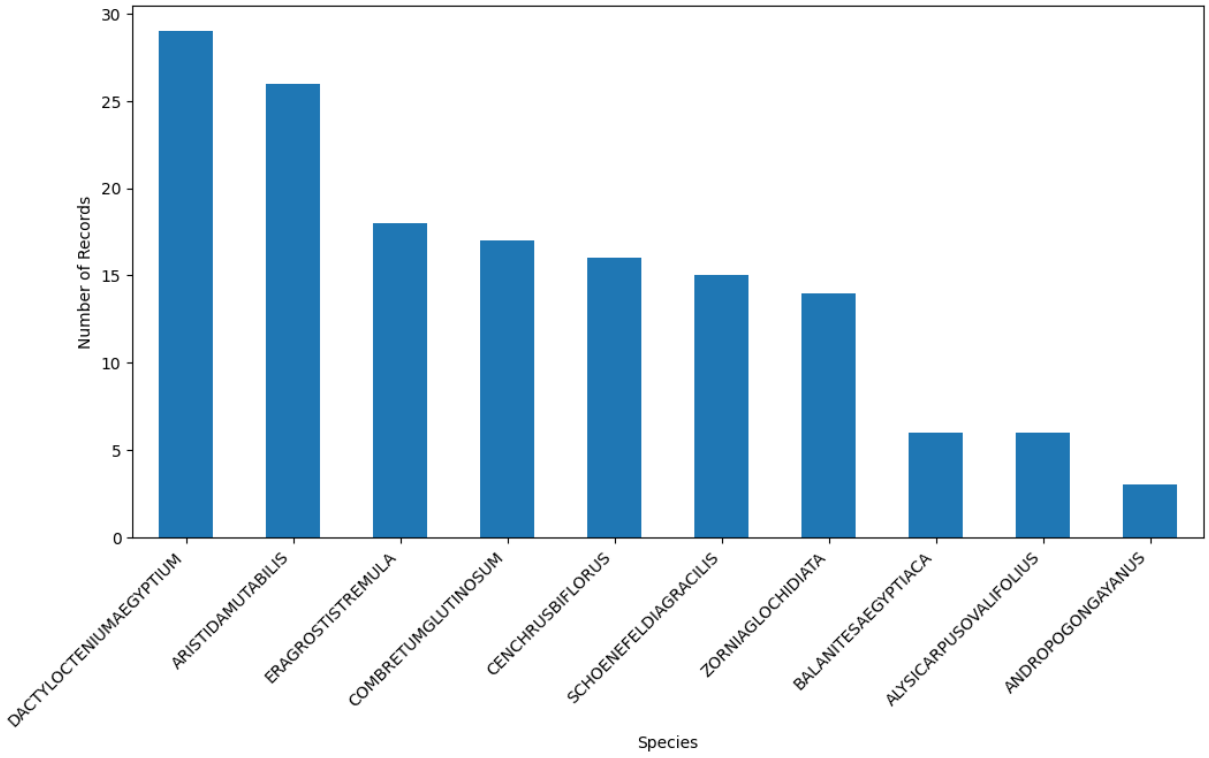


Figure 35: Nombre de Relevés pour les 10 Espèces Sélectionnées de l'herbier.

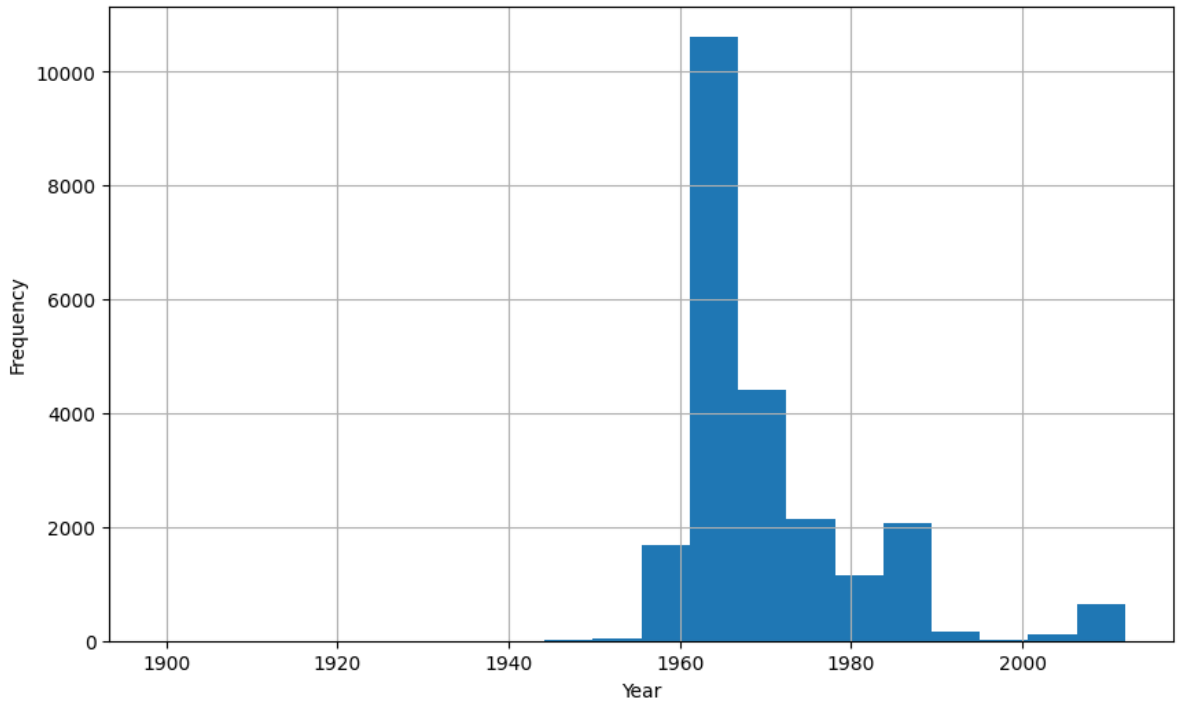


Figure 37:Distribution des Années de Données de l'herbier.

Ainsi nous présentons quelques cartes de distribution des espèces végétales qui comparent les valeurs de présences prédites par nos modèles avec celles de l'herbier.

a- Pour *Cenchrus biflorus* en 1964 :

La carte ci-dessous montre la superposition des prédictions de présence de *Cenchrus biflorus* par RF avec les présences de l'herbier en 1964. Les points verts indiquent les présences correctement prédites par le modèle, tandis que les points rouges indiquent celles incorrectement prédites.

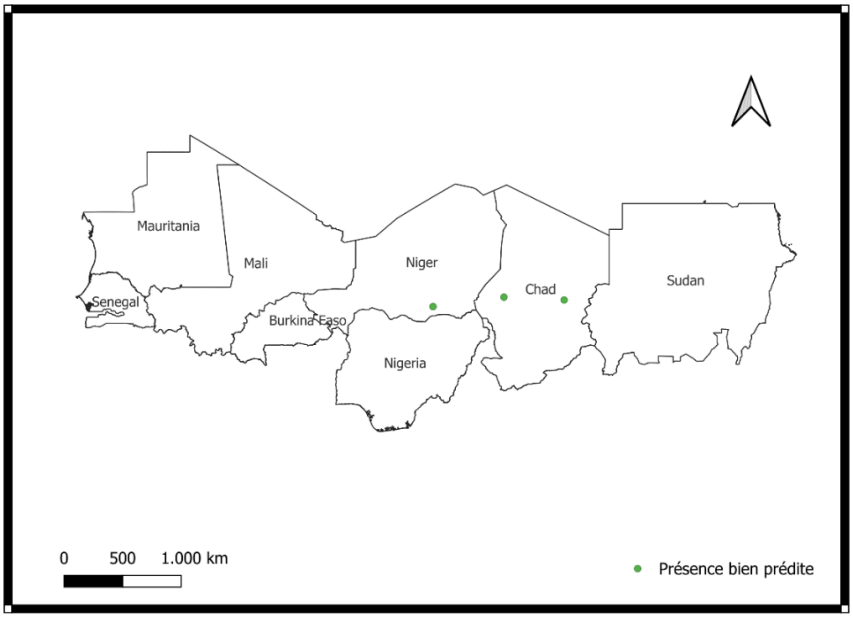


Figure 36: Prédiction de Présence avec données de l'herbier pour Cenchrus biflorus en 1964.

On note que les prédictions de présence correspondent bien aux données de l'herbier, confirmant ainsi l'efficacité du modèle pour cette espèce

b- Pour *Balanites aegyptiaca* en 1964:

La carte ci-dessous montre la superposition des prédictions de présence de ***Balanites aegyptiac*** par CNN avec les présences de l'herbier en 1964. Les points verts indiquent les présences correctement prédites par le modèle.

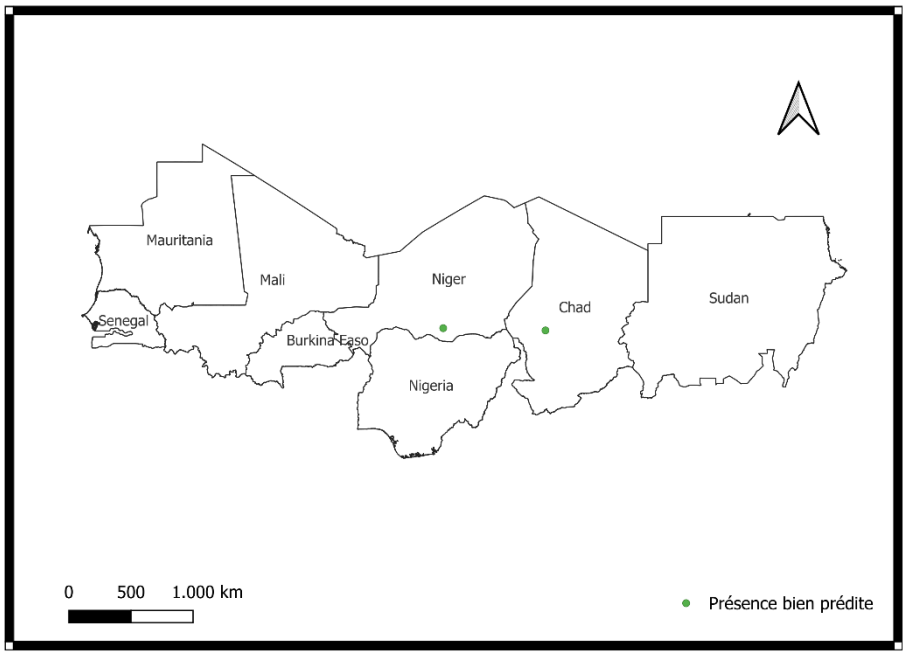


Figure 37: Prédiction de Présence avec données de l'herbier pour BALANITES AEGYPTIACA en 1964.

On observe que les prédictions de présence sont bien alignées avec les données de l'herbier, ce qui valide l'efficacité du modèle pour cette espèce particulière.

Conclusion

Cette étude visait à prédire la distribution des espèces végétales dans la région du Sahel en utilisant plusieurs modèles de *Machine Learning* et d'écologie, notamment MaxENT, *Random Forest* (RF), *Support Vector Machine* (SVM), *Generalized Linear Model* (GLM), *Generalized Additive Model* (GAM) et *Convolutional Neural Network* (CNN). À travers une approche rigoureuse de collecte, de prétraitement des données et de sélection de variables, nous avons identifié les modèles RF et CNN comme étant les plus performants pour prédire la distribution des espèces herbacées, *Cenchrus biflorus* et arborées, *Balanites aegyptia*.

Les résultats ont démontré que l'équilibrage des classes et la sélection des variables sont des étapes cruciales pour améliorer la précision et la robustesse des modèles. L'analyse en composantes principales (ACP) s'est révélée particulièrement efficace pour réduire la dimensionnalité des données tout en conservant l'essentiel de l'information. Les modèles optimaux, une fois entraînés et validés, ont permis de générer des cartes de distribution annuelles précises, offrant des insights précieux pour la conservation et la gestion des écosystèmes.

La validation des prédictions à l'aide des données de l'herbier a confirmé la fiabilité de nos modèles. Les résultats obtenus montrent que les modèles RF et CNN sont bien adaptés à la complexité des données environnementales et des distributions des espèces dans la région du Sahel.

En conclusion, cette étude fournit une base solide pour la modélisation prédictive de la distribution des espèces végétales dans des environnements complexes.

CHAPITRE IV : DISCUSSION

Introduction

Dans ce chapitre, nous discutons de la qualité et de la précision des données utilisées, ainsi que des défis rencontrés, tels que le surapprentissage des modèles. Nous analyserons les résultats obtenus par rapport aux niveaux, dimensions et mesures de la biodiversité, en mettant en perspective les conclusions tirées de nos modèles avec celles de la littérature scientifique. Enfin, nous explorerons d'autres méthodes potentielles pour les SDM et proposerons des pistes pour de futures recherches.

4.1 Qualité et précision des données :

La qualité et la précision des données sont des aspects cruciaux pour le succès de toute modélisation prédictive. Dans cette étude, les données utilisées proviennent de diverses sources incluant des variables bioclimatiques, climatiques, des données de sol et des indices d'humidité ainsi que les données d'occurrence de GBIF. Bien que ces données offrent une base riche pour la modélisation, elles présentent également les défis tels que :

Dans un premier temps, l'incomplétude et l'imprécision des données d'occurrence : ce problème est particulièrement prononcé pour les données d'occurrence des espèces. Ces données n'ont pas toujours été collectées selon des protocoles rigoureux et standardisés au fil des années. Par exemple, les données de présence extraites pour une zone donnée n'étaient souvent pas exhaustives. Cela signifie que les collecteurs pouvaient se concentrer sur l'étude de la présence d'une espèce spécifique sans prêter attention aux autres espèces présentes dans la même zone. Cette sélectivité introduit un biais dans les données, réduisant ainsi la fiabilité et la représentativité des observations enregistrées. Pour contourner ce problème, nous avons suggéré que pour le même relevé si une espèce n'est pas présente elle est considérée comme absence, c'est ce qu'on appelle les pseudo-absences.

Dans un second temps la fiabilité des collectes : La fiabilité des données collectées varie également, certaines collectes étant moins précises ou sujettes à des erreurs. Parfois, les informations sur les coordonnées géographiques ou les dates de collecte œuvrent être incorrecte

ou approximatives, introduisant du bruit dans les données. De telles erreurs affectent la qualité des prédictions, car elles influencent directement l'entrainement et la validation des modèles.

Par la suite, l'hétérogénéité des sources de données : les données utilisées proviennent de multiple source, chacune avec ses propres méthodes de collecte et niveau de précision Cette hétérogénéité peut compliquer l'intégration et l'analyse des données, entrainant des incohérences et des erreurs potentielles dans les modèles prédictifs

Enfin l'échantillonnage et le biais géographique : Les données d'occurrence peuvent également souffrir d'in biais d'échantillonnage, ou certaines régions sont suréchantillonnées par rapport à d'autres. Cela peut être du a la facilité d'accès, a l'intérêt de recherche, ou à d'autres facteurs logistiques. Ce biais géographique peut conduire à une mauvaise représentation de la distribution réelle des espèces, affectant la capacité du modèle à généraliser correctement les prédictions à l'ensemble de la région du Sahel.

4.2 Surapprentissage des modèles :

L'analyse des résultats pour différents modèles révèle des schémas de prédiction particuliers qui peuvent être attribués à des phénomènes de surapprentissage qui se produit lorsque le modèle apprend des détails et du bruit spécifique aux données d'entrainement au lieu de généraliser à de nouvelles données.

Ainsi on remarque par exemple que les prédictions de présence sont souvent plus abondantes dans certaines zones géographiques précises. Cela est principalement dû au fait que pour certaines zones, nous disposons de plus de données de présence que pour d'autres. Une surévaluation des présences dans les zones mieux documentées et une sous-évaluation dans les zones moins documentées sont entrainées par ce déséquilibrage. Pour atténuer ce problème, et comme déjà expliqué, nous avons utilisés l'équilibrage des classes, fournissant un nombre équilibré de données de présence et d'absence au modèle. Cette approche réduit le biais et améliore la généralisation du modèle.

Le problème *d'overfitting* est également amplifié par l'utilisation des colonnes année, latitude et longitude comme variables d'entrée. Ces colonnes peuvent introduire un biais

temporel et spatial, car le modèle peut apprendre des tendances spécifiques à certaines années et positions géographiques. Par exemple, si nous avons plus de données de présence pour une année et positions précises, le modèle peut prédire des présences en fonction de ces tendances historiques plutôt qu'en fonction de facteurs environnementaux pertinents. Pour résoudre ce problème, nous avons examiné des méthodes de régularisation et de validation croisée. La régularisation permet de pénaliser les modèles trop complexes, tandis que la validation croisée évalue les performances du modèle sur des sous-ensembles de données indépendants, réduisant ainsi le risque de surapprentissage.

La complexité des modèles, notamment en utilisant un grand nombre de variables explicatives, peut également contribuer au surapprentissage. La sélection de variables pertinentes et la réduction de la dimensionnalité par l'analyse des composantes principales sont des stratégies efficaces pour simplifier les modèles et améliorer leur capacité de généralisation. En effet, en réduisant le nombre de variables, nous minimisons le risque que le modèle apprenne des détails spécifiques aux données d'entrainement qui ne se généralisent pas bien à de nouvelles données.

4.3 Discussion des résultats

Analysons les résultats du modèle Random Forest pour une espèce herbacée : *Cenchrus biflorus* et celles du modèle CNN pour un arbre : *Balanites aegyptiaca* aux cours des différentes périodes durant lesquelles le climat du Sahel a changé de manière significative notamment pendant les périodes de changements climatiques extrêmes au Sahel, par exemple les sècheresses sévères des années 1970 et 1980, liées à des fluctuations dans les précipitations saisonnières, et les augmentations de précipitations depuis les années 1980 montrant une tendance vers un climat plus humide.

a- Pour *Cenchrus biflorus* :

En 1970, les prédictions illustrées dans la figure 34(c) indiquent une présence significative de cette espèce, principalement concentrée dans la région centrale du Sahel. Cette présence relativement homogène couvre une large bande horizontale à travers plusieurs pays de la région comme le Sénégal, le Mali, le Niger et le Tchad.

Cette distribution extensive pourrait être attribuée à la résilience de *Cenchrus biflorus* face aux conditions de sécheresse sévère qui ont marqué cette décennie, suggérant que cette espèce herbacée a une forte capacité d'adaptation aux environnements arides.

En 1980, comme illustré dans la figure 34(d), la présence prédite de *Cenchrus biflorus* reste étendue et dense, malgré les conditions climatiques extrêmement sèches qui ont persisté. Cette stabilité dans la distribution pourrait indiquer une certaine tolérance de l'espèce aux fluctuations climatiques, en particulier aux périodes de sécheresse prolongée. Les données montrent que l'espèce continue de se maintenir dans les régions centrales et occidentales du Sahel, avec une présence notable au Mali, au Burkina Faso et au Niger.

b- Pour *Balanites aegyptiaca* :

Les résultats du modèle CNN pour *Balanites aegyptiaca* montrent une tendance différente. En 1970, comme illustré dans la figure 35(c), la présence de cette espèce est principalement concentrée dans les zones centrale et orientale du Sahel. Contrairement *à Cenchrus biflorus*, *Balanites aegyptiaca* semble avoir une distribution plus fragmentée, avec des poches de densité élevée. Cela pourrait être dû à sa nature arborée et à ses exigences environnementales spécifiques qui la rendent plus vulnérable aux changements drastiques des précipitations.

En 1980, la figure 35(d) montre une expansion de la distribution de *Balanites aegyptiaca* malgré les sécheresses persistantes. Cette augmentation de la présence pourrait être liée à une certaine capacité d'adaptation à des conditions de faible humidité, bien que de manière plus localisée que *Cenchrus biflorus*. Les zones de présence se concentrent encore principalement dans les régions centrale et orientale, indiquant que *Balanites aegyptiaca* a pu trouver des niches favorables malgré le climat défavorable.

4.4 Résultats par rapport aux niveaux, dimensions et mesures de la biodiversité

Les résultats obtenus à partir de nos modèles de distribution des espèces (SDM) permettent de mieux comprendre les différents niveaux de biodiversité et leur impact sur la dynamique des espèces végétales dans le Sahel. Par exemple, la diversité interspécifique se révèle cruciale pour la résilience et l'adaptation des espèces face aux variations environnementales.

Pour *Cenchrus biflorus*, nos modèles ont démontré une grande résilience aux conditions arides du Sahel, probablement grâce à sa variabilité génétique. Cette espèce a pu s'adapter à différents types de sols et niveaux d'humidité, ce qui lui a permis de maintenir sa présence même pendant les périodes de sécheresse intense. Les résultats de modélisation montrent que *Cenchrus biflorus* est capable de survivre dans des conditions extrêmes grâce à ses adaptations génétiques, illustrant l'importance de la diversité intraspécifique pour la survie des espèces dans des environnements difficiles (Giannini, Biasutti et Verstraete, 2008).

Les types de sol et les conditions climatiques influencent également la composition des communautés végétales. Par exemple, des espèces comme *Schoenefeldia gracilis* prospèrent dans des zones désertifiées et remplacent les espèces moins résistantes durant les sécheresses. Les modèles de distribution révèlent que *Schoenefeldia gracilis* tend à coloniser les zones où les précipitations sont irrégulières et les températures élevées, suggérant une capacité d'adaptation aux conditions extrêmes.

De plus, les SDM permettent d'analyser la distribution géographique des espèces dans différentes régions, révélant des variations significatives en fonction des conditions environnementales locales. Cela souligne l'importance de la dimension spatiale dans la compréhension des dynamiques de biodiversité. Par exemple, nos modèles montrent que certaines espèces se déplacent vers des zones plus humides en réponse aux précipitations irrégulières et à l'augmentation des températures.

Enfin, ces modèles montrent comment les distributions des espèces ont évolué de 1920 à 2012 en réponse aux changements climatiques. Par exemple, l'augmentation des précipitations depuis les années 1980 a favorisé l'expansion de certaines espèces végétales vers des zones plus humides, modifiant ainsi la composition des communautés végétales.

Ces résultats mettent en lumière la relation étroite entre la biodiversité et les modèles de distribution des espèces. Ils soulignent l'importance de la diversité génétique, de la diversité des habitats et de la dimension temporelle dans la compréhension et la gestion de la biodiversité dans le Sahel. Il est possible que ces observations soient dues à la capacité des espèces à s'adapter aux variations environnementales grâce à leur variabilité génétique et à leur plasticité phénotypique. En somme, nos résultats confirment que la diversité des espèces et leur capacité d'adaptation sont essentielles pour la résilience des écosystèmes face aux changements climatiques.

4.5 Autres méthodes utilisées pour les SDM :

Les chercheurs du CIRAD travaillent sur les modèles de distribution des espèces. Parmi les méthodes qu'ils explorent, on trouve :

- **Malpolon** : Ce framework utilise uniquement des données de présence. Cela est principalement dû à la disponibilité des données sur lesquelles ils travaillent, constitués à partir de différents types d'images mises à leur disposition. Bien que ce package ne soit pas spécifiquement adapté aux SDM, Il peut être utilisé dans divers domaines. En effet, comme décrit sur leur site web, Malpolon est un Framework flexible conçu pour manipuler et analyser les données de biodiversité provenant de différentes sources, y compris les observations d'espèces et les données environnementales. En comparaison avec les méthodes utilisées dans cette étude, Malpolon se distingue par sa capacité à intégrer de vastes ensembles de données et à appliquer des techniques de modélisation avancées pour explorer les relations entre les espèces et leur environnement.

- **Modèles mathématiques de processus ponctuels** : Toujours en cours de construction, ce modèle se base sur des concepts purement mathématiques et utilise uniquement des données de présences. Il s'appuie sur les 19 variables bioclimatiques de WorldClim divisées en quartes périodes, afin d'entrainer le modèle sur une période et de prédire avec les données bioclimatiques des périodes suivantes. Cette approche permet d'évaluer la capacité du modèle à généraliser les prédictions sur des périodes temporelles différentes, fournissant ainsi une perspective dynamique de la distribution des espèces.

En résumé, ces approches présentent des avantages et des inconvénients distincts par rapport aux méthodes utilisées dans cette étude. Cependant, ces méthodes peuvent être limités par l'indisponibilité des données de présences. Les modèles mathématiques de processus ponctuels, quant à eux, fournissent une perspective plus détaillée des interactions spatiales, mais nécessitent une expertise mathématique approfondie pour leur mise en œuvre et leur interprétation.

Conclusion

La discussion a mis en lumière les forces et les faiblesses de notre approche méthodologique, tout en soulignant la pertinence de nos résultats dans le contexte plus large de la biodiversité et des changements climatiques dans le Sahel. Nos modèles ont démontré une capacité à prédire avec précision la distribution des espèces, bien que certaines limitations, comme le surapprentissage, aient été identifiées. En confrontant nos résultats à la littérature existante, nous avons pu confirmer l'importance de la diversité génétique et des conditions environnementales dans la résilience des espèces. Ces conclusions ouvrent la voie à des recherches futures visant à affiner les modèles et à mieux comprendre les dynamiques écologiques dans des environnements en mutation rapide.

CONCLUSION

L’étude sur la modélisation des espèces végétales dans la région du Sahel a permis d’identifier et de comparer plusieurs modèles de prédiction, en mettant en avant leurs performances respectives selon diverses métriques. Les résultats montrent que le modèle Random Forest avec sélection de variables s’avère être le plus performant pour prédire la distribution des espèces, en équilibrant efficacement la précision, le rappel et le score F1. Le modèle CNN, bien que légèrement moins performant, se présente comme une alternative solide, surtout pour les espèces herbacées.

Cependant, la comparaison entre la performance des modèles pour les espèces herbacées et les espèces arborées a révélée des différences notables. Les modèles CNN ont montré une meilleure adaptabilité pour les espèces herbacées, en raison de la complexité et de la variabilité spatiale moins prononcée de ces espèces par rapport aux arbres En revanche, pour les espèces arborées, le modèle Random Forest a démontré une robustesse accrue

Les cartes de distribution générées grâce à ces modèles offrent une vision précieuse de la répartition spatiale et temporelle des espèces végétales dans le Sahel, et peuvent servir d’outils essentiels pour la conservation de la biodiversité, la gestion durable des ressources naturelles et l’adaptation aux changements climatiques.

La validation par les données de l’herbier a confirmé la fiabilité des prédictions, renforçant ainsi la pertinence des modèles utilisés. Les résultats de cette étude soulignent l’importance de combiner des techniques avancées de modélisation avec des données environnementales et biologiques de haute qualité pour obtenir des prédictions précises et utiles pour la gestion environnementale.

En somme, cette étude contribue à combler les lacunes dans les connaissances sur la distribution des espèces végétales dans le Sahel et fournit une base pour des recherches futures et initiatives de conservation dans cette région vulnérable aux changements climatiques.

RECOMMANDATIONS

Suite aux conclusions de cette étude, plusieurs recommandations peuvent être formulées pour améliorer les futures recherches en modélisation des espèces végétales dans le Sahel et ailleurs :

- Amélioration de la qualité des données : Investir dans la collecte de données de haute précision, tant pour les variables biologiques que pour les variables environnementales, est essentiel. L'utilisation de capteurs avancés et de télédétection peut améliorer la qualité des données climatiques et de sol.
- Gestion des pseudo-absences : Développer des méthodes plus robustes pour la sélection de pseudo-absence afin de minimiser les biais introduits pour cette approche. L'utilisation d'algorithmes sophistiqués pour la génération de pseudo-absences peut améliorer la fiabilité des modèles.
- Comparaison et intégration des modèles : explorer des approches hybrides qui combinent les forces des CNN et des modèles *Random Forest* pour tirer parti des avantages de chaque méthode selon les caractéristiques des espèces étudiées.
- Études spécifiques pour chaque espèce végétale : mener des études spécifiques pour chaque espèce végétale en question. Une analyse approfondie et ciblée permettrait de mieux comprendre les facteurs écologiques et environnementaux influençant la distribution de chaque espèce, ce qui pourrait améliorer les modèles prédictifs.
- Prendre en considération l'interaction entre les espèces : Intégrer les interactions entre les espèces dans les modèles. Les interactions écologiques, telles que la compétition et la facilitation, jouent un rôle crucial dans la distribution des espèces. La prise en compte de ces interactions pourrait améliorer la précision des prédictions et offrir une vision plus holistique des dynamiques écologiques.
- Intégration d'autres variables environnementales : Enrichir les modèles, en intégrant d'autres variables environnementales, permettrait de capturer une gamme plus large de conditions environnementales affectant la distribution des espèces et ainsi améliorer la précision des prédictions.

- Utilisation des indices de biodiversité : Intégrer des indices de biodiversité, tels que la richesse spécifique et les indices de diversité (par exemple, l'indice de Shannon-Weaver), dans les analyses pour obtenir une évaluation plus complète de la biodiversité et de ses variations spatiales et temporelles.

En suivant ces recommandations, les futures études sur la distribution des espèces végétales pourront bénéficier de données plus précises contribuant ainsi à une meilleure compréhension et gestion de la biodiversité dans le Sahel et au-delà. De plus, ces modèles et conclusions peuvent être appliqués à d'autres régions, comme le Maroc en adaptant les données spécifiques à chaque zone.

REFERENCES BIBLIOGRAPHIQUES

- Adams, W. M. (2018). Estimation of Biodiversity Values and their Implications. In J. E. Mooney & P. V. Wilson (Eds.), Biodiversity Economics: Principles, Methods and Applications (pp. 150-170). Springer.
- Akinola, A. O., & Ramontja, N. (2021). "Violent Conflict in the Sahel: Causes, Dynamics, and Actors." SpringerLink.
- Anderson, R. P., & González, I. (2020). Advances in Species Distribution Modeling. Methods in Ecology and Evolution, 11(8), 1012-1023.
- Bai, Y. et al. (2004). Ecosystem stability and compensatory effects in the Inner Mongolia grassland.
- Bationo, A., et al. (2006). "Soil Fertility Improvement and Integrated Nutrient Management." Springer.
- Bouterfas, K. (2020-2021). Biodiversité Végétale. Université Djillali Liabes de Sidi Bel Abbès, Faculté des Sciences de la Nature et de la Vie, Département des Sciences de l'Environnement.
- Britannica. (2024). "Sahel." Encyclopaedia Britannica.
- Bridson, D., & Forman, L. (1999). The Herbarium Handbook. Royal Botanic Gardens, Kew.
- Charpentier, A. (1995). Introduction à la modélisation statistique. Paris: Springer.
- Cutler, D. R., et al. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.
- Dupont, Y., & Durand, P. (2010). Évaluation Économique de la Biodiversité. Cairn.
- Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677-697.

- Elith, J., & Leathwick, J. R. (2014). Performance of Species Distribution Models. Methods in Ecology and Evolution, 5(3), 215-227.
- Epule, T. E., et al. (2017). "Climate change adaptation in the Sahel." Springer.
- FAO. (2001). "Hydromorphic Soils of the Sahel." Food and Agriculture Organization of the United Nations.
- FAO. (2004). "Soil Classification: A global desk reference." Food and Agriculture Organization of the United Nations.
- Foley, J. A., et al. (2005). Global Consequences of Land Use. Science, 309(5734), 570-574.
- Franklin, J. (1995). Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography, 19, 474-499.
- Franklin, J. (2010). Mapping species distributions: spatial inference and prediction. Cambridge University Press.
- Franklin, J., & Miller, J. (2010). Species Distribution Modeling: The Next Generation. Journal of Biogeography, 37(11), 2145-2160.
- Ganaba, S., Ouadba, J.-M., & Bognounou, O. (1998). Les ligneux à usage de bois d'énergie en région sahélienne du Burkina Faso: préférence des groupes ethniques. Sécheresse, 9: 261–268.
- Giannini, A., Biasutti, M., & Verstraete, M. M. (2008). A climate model-based review of drought in the Sahel: Desertification, the re-greening and climate change. Global and Planetary Change, 64(3-4), 119-128.
- Gonzalez, P. J. (1997). Dynamics of biodiversity and human carrying capacity in the Senegal Sahel. Ph.D. Thesis, University of California, Berkeley, California.
- Gonzalez, P. (2001). Desertification and a shift of forest species in the west African Sahel. Climate Research, 17: 217–228.
- Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8(9), 993-1009.

- Guisan, A., & Thuiller, W. (2011). Using Remote Sensing for Species Distribution Models.Methods in Ecology and Evolution, 2(3), 312-320.
- Guisan, A., & Zimmermann, N. E. (2006). Predictive habitat distribution models in ecology. Ecological Modelling, 135, 147-186.
- Hillebrand, H. (2009). On the Generality of the Latitudinal Diversity Gradient. Annual Review of Ecology, Evolution, and Systematics, 40(1), 213-239.
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25(15), 1965-1978.
- Hooper, D. U. et al. (2005). Effects of biodiversity on ecosystem functioning: A consensus of current knowledge and needs for future researchHutchinson, G. E. (1957). Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology, 22, 415-427.
- Jiménez-Valverde, A., & Lobo, J. M. (2009). "The pitfalls of using the area under the ROC curve to evaluate the performance of species distribution models". Journal of Biogeography.
- Jones, M. H. (1998). Plant Community Responses to Experimental Warming. Oikos, 81(2), 309-322.
- Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., & Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear Poisson point process. Ecological Informatics, 55, 1-11.
- Le Houérou, H.N. (1989). The Grazing Land Ecosystems of the African Sahel. Ecological Studies, vol 75. Springer, Berlin, Heidelberg.
- Leta, S., DeClerq, E. M., & Madder, M. (2013). High-resolution predictive mapping for Rhipicephalus appendiculatus (Acari: Ixodidae) in the Horn of Africa. Experimental and Applied Acarology, 60, 531-542.
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). "AUC: a misleading measure of the performance of predictive distribution models". Global Ecology and Biogeography.

- Lykke, A. M. (1998). Assessment of species composition change in savanna vegetation by means of woody plants' size class distributions and local information. Biodiversity and Conservation, 7: 1261–1275.
- Lykke, A. M. (2000). Local perceptions of vegetation change and priorities for conservation of woody savanna vegetation in Senegal. Journal of Environmental Management, 59: 107–120.
- Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography, 36(10), 1058-1069.
- Miller,C. (2023). sdmProfiling. GitHub.
- Mortimore, M. J., & Adams, W. M. (2001). Farmer adaptation, change and crisis in the Sahel. Global Environmental Change, 11(1), 49-57.
- Nicholson, S. E. (2001). Climatic and environmental change in Africa during the last two centuries. Climate Research, 17, 123-144.
- Nicholson, S. E. (2013). The West African Sahel: A Review of Recent Studies on the Rainfall Regime and Its Interannual Variability. ISRN Meteorology, 2013, Article ID 453521.
- Peterson, A. T., Papes, M., & Kluza, D. A. (2003). Predicting the potential invasive distributions of four alien plant species in North America. Weed Science, 51(6), 863-868.
- Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31, 161-175.
- Phillips, S. J., & Dudík, M. (2017). Species Distribution Models: Theory and Applications. PeerJ, 5, e4095.
- Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3-4), 231-259.
- Powers, D. M. (2011). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation". Journal of Machine Learning Technologies.

- Provost, F., & Kohavi, R. (1998). "The confusion matrix: The importance of classifying correctly".
- R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Robertson, M. P., & Hirzel, A. H. (2004). Species Distribution Modeling: State of the Art.CiteSeerX.
- Smith, J., & Brown, L. (2023). Climate Change and Agriculture: Impacts and Adaptations. Sustainability, 15(9), 7128.
- Smith, A. B., & Santos, M. J. (2019). A New Method for Species Distribution Modeling. Ecological Informatics, 52, 68-74.
- Thiers, B. (2021). The world's herbaria 2021: A summary report based on data from Index Herbariorum. **New York Botanical Garden**.
- Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences, 102(23), 8245-8250.
- Tilman, D. et al. (2006). Biodiversity and Ecosystem Stability.
- Tchakerian, V. P., & Payne, M. (1997). "Wind Erosion in the Sahelian Zone of Africa: Implications for desertification and resource management." Journal of Arid Environments.
- UNEP. (n.d.). Africa Soil Units (FAO).
- United Nations Office for the Coordination of Humanitarian Affairs (OCHA). (2014). Sahel: Overview of Humanitarian Needs and Requirements. OCHA, New York.
- Van Rijsbergen, C. J. (1979). "Information retrieval and filtering: Precision, recall, and F-measure".
- Walther, O. (2017). "Wars and conflicts in the Sahara-Sahel." OECD.
- Wickens, G. E. (1998). The uses of the desert date tree (Balanites aegyptiaca) in Africa. Economic Botany, 52(2), 186-194.

- White, F. (1983). The vegetation of Africa: a descriptive memoir to accompany the UNESCO/AETFAT/UNSO vegetation map of Africa. Paris: UNESCO

Annexes

Tableau 11: Comparaison entre R et Python pour la modélisation de la distribution des espèces

Caractéristiques	R	Python
Logiciel pour spatialisation	Package : ‘dismo’ pour la modélisation de la distribution des espèces, ’raster’ pour la manipulation des donnees raster et’sp’ pour la manipulation de donnees spatiales.	La structure d’ArcGIS facilite la compilation de nombreux scripts Python en une seule boîte à outils qui peut être facilement distribuée et incorporée dans le logiciel
Outils Statistiques	- Statistiques récapitulatives univariées: package stats - ANOVA: package stats - Régression linéaire: package stats - Apprentissage automatique: packages caret, mlr	- Statistiques récapitulatives univariées: bibliothèques Pandas, NumPy – ANOVA: bibliothèque scikit-learn - Régression linéaire: bibliothèque statsmodels - Apprentissage automatique: bibliothèques scikit-learn, tensorflow, keras

Exemple de package / outils	ENMeval : réglage de modèle largement accessibles pour l'algorithme Maxent + fourni plusieurs méthodes pour partitionner les données d'occurrence + a signalé diverses mesures de performance + dispose d'une nouvelle structure orientée objet pour l'ajout d'autres algorithmes https://github.com/jamiemkass/ENMeval	ELAPID : Suite d'outils de modélisation statistique pour les biogéographes -prend en charge l'utilisation de formats de données géospatiales modernes et utilise des approches contemporaines pour l'entraînement des modèles statistiques GitHub - earth-chris/elapid: Species distribution modeling tools, including a python implementation of Maxent
	megaSDM (Shipley et al., 2022): Ce package est efficace pour traiter simultanément de nombreuses espèces, périodes et cas d'utilisation.	SDMtoolbox : une boîte à outils SIG basée sur python pour l'analyse de modèles de génétique, biogéographique et de distribution des espèces dans le paysage
Popularité pour ML et DL	Moins populaire pour ML et DL, mais dispose de package pour l'analyse statistique (ex :Caret,randomForest)	Très populaire pour ML et DL avec de nombreuses bibliothèques et outils dédiés (ex :TensorFlow,Keras)
Flexibilité	Spécialisé dans l'analyse statistique et peut être moins flexible pour d'autres applications.	Très flexible pour diverses taches

Performance	Peut-être moins performant pour les gros volumes de donnees en raison de sa nature plus axée sur la statistique.	Performant pour le traitement de gros volumes de donnees et les calculs intensifs
Apprentissage Automatique	Possède de package pour l'apprentissage automatique mais moins d'outils pour le deep learning .	Dispose de bibliothèques et outils puissants pour l'apprentissage automatique (ex :Scikit-learn)
Deep Learning	Moins adapté au deep learning, mais peut être utilisé avec des packages externes pour le deep learning.	Très adapté au deep learning avec des frameworks populaires (ex :TensorFlow,PyTorch)
Interprétabilité	Peut-être plus axé sur l'interprétabilité des modèles ML, favorisant la clarté et la compréhension des résultats.	Peut-être moins axé sur l'interprétabilité des modèles ML.
Communauté de Développeurs	Communauté active dans l'analyse statistique, mais moins étendue dans le domaine de ML et DL.	Grande communauté active dans le domaine du ML et DL, offrant un support et des ressources variées.

s

الملخص

تُعد منطقة الساحل، التي تقع بين الصحراء الكبرى والسافانا السودانية، منطقة شبه قاحلة تتعرض لظروف بيئية قاسية. تتميز بتساقط أمطار غير منتظم وغالبًا غير كافٍ، مما يؤدي إلى فترات جفاف متكررة وتدهور كبير في التربة. هذه الظروف تزيد من تفاقم المشاكل القائمة مثل انعدام الأمن الغذائي وفقدان التنوع البيولوجي وتدهور الأراضي، مما يجعل المنطقة عرضة بشكل خاص لتغير المناخ. يلعب الغطاء النباتي في منطقة الساحل دورًا حيويًا في تثبيت التربة، وتنظيم المناخ المحلي، ودعم سبل عيش السكان المحليين. لذلك، فإن التنبؤ بتوزيع الأنواع النباتية في منطقة شاسعة ومتنوعة كهذه أمر ضروري. لمعالجة هذه القضية، تهدف دراستنا إلى تحقيق ثلاثة أهداف رئيسية. وهي:التنبؤ بالتوزيع المكاني والزماني للأنواع النباتية السائدة في منطقة الساحل، اختبار نماذج تعلم الآلة المختلفة لتحديد التوزيع المكاني لبيانات تواجد الأنواع النباتية، وإيجاد النموذج المناسب لتوزيع كل نوع من النباتات (العشبية والشجرية). بالإضافة إلى ذلك، سيتم إنشاء خرائط توزيع سنوية قابلة للاستخدام. تعتمد البيانات المستخدمة بشكل رئيسي على قاعدة بيانات فلوتروب، المتكونة أساسا من ثمانية أنواع-عشبية ونوعين من الأشجار، معززة بمتغيرات بيئية مثل الأمطار، ودرجة الحرارة، ونوع التربة، وكذلك مؤشر هطول الأمطار. تشمل تقنيات النمذجة أساليب التعلم الآلي والتعلم العميق مثل الغابات العشوائية ونماذج أخرى، مع التحقق من النتائج باستخدام بيانات من المعشبة المتاحة بالمركزالدولي للبحث الزراعي والتنمية. تستند المنهجية المعتمدة إلى جمع ومعالجة وتحليل البيانات البيئية وبيانات تواجد الأنواع النباتية من اجل نمذجة توزيعها.تظهر النتائج أن الغابات العشوائية مناسبة بشكل خاص لنمذجة توزيع الأنواع العشبية، بينما تحقق الشبكات العصبية الالتفافية أداءً أفضل لأنواع الأشجار.

ومع ذلك، يمكن استخدام كلا النموذجين لكلا النوعين من الأنواع. أتاحت النماذج التنبؤية توليد خرائط توزيع سنوية مفصلة وقابلة للاستخدام، مما يشير إلى المناطق التي تكون فيها الظروف البيئية ملائمة أو غير ملائمة لوجود الأنواع. هذه الخرائط ضرورية لإدارة وحفظ الموارد النباتية في منطقة الساحل.

الكلمات المفتاحية: نماذج توزيع الأنواع، التعلم الآلي، التصنيف، وجود-غياب، الساحل، خريطة، الشبكات العصبية التلافيفية، الغابات العشوائ

ROYAUME DU MAROC
INSTITUT AGRONOMIQUE ET VETERINAIRE HASSAN II

المملكة المغربية
معهد الحسن الثاني
للزراعة والبيطرة

مشروع نهاية الدراسة لنيل دبلوم مهندس دولة في الزراعة

تخصص: هندسة علم البيانات في الزراعة

استخدام التعلم الآلي لرسم خرائط التوزيع السنوي للأنواع النباتية في منطقة الساحل وفقًا للعوامل البيئية

قدم وأدعم علنًا من قبل :

كريمج مريم

اللجنة :

الأستاذ محمد الدحاوي	**معهد الحسن الثاني للزراعة والبيطرة -**	**رئيس**
الأستاذة سلوى بنسعلي	**معهد الحسن الثاني للزراعة والبيطرة -**	**مقررة**
د. سيمون تاغوردو	**مركز التعاون الدولي للبحوث الزراعية من أجل التنمية**	**مقرر**
الأستاذ سلمان بن غبريت	**معهد الحسن الثاني للزراعة والبيطرة -**	**مُقَيِّم**

يوليوز 2024

Adresse : Madinat Al Irfane, B.P. 6202. Rabat – Maroc
Tél : (00 212) 0537 77 17 58/59
Fax : (00 212) 0537 77 58 45
Site web: http://www.iav.ac.ma

العنوان: ص. ب 6202 الرباط المعاهد الرباط - المغرب
الهاتف: 59 / 58 17 77 0537 (00 212)
الفاكس: (00 212) 0537 77 58 45
موقع الأنترنت: http://www.iav.ac.ma