Mémoire de Master

Master Informatique – Parcours Intelligence artificielle

Prédiction du stock de carbone organique du sol

Application aux villages de Bary, Diohine/Sassem et Sop

Présenté par :
Magaye Ndiaye

Encadrants :
Dr. Mandicou Ba (UMMISCO, IRD)
Pr. Marie Beurton-Aimar (Université de Bordeaux)

Avril – Septembre 2025

Université de Bordeaux – Institut de Recherche pour le Développement (IRD) – UMMISCO Sénégal

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à Dieu Tout-Puissant pour m'avoir donné la force, la santé et la persévérance nécessaires à la réalisation de ce mémoire.
Je remercie sincèrement mon encadrant Dr. Mandicou Ba pour son accompagnement bienveillant, ses conseils éclairés et son exigence scientifique, qui ont été essentiels à la bonne conduite de ce travail. J'adresse également mes remerciements à Professeure Marie Beurton-Aimar et à M. Zemmari pour leur disponibilité, leurs encouragements et leurs orientations précieuses.
Mes remerciements vont également à l'ensemble de l'équipe de l'UMMISCO Sénégal et du projet SLAM-B, pour l'accueil chaleureux et l'environnement scientifique stimulant qu'ils m'ont offert durant ce stage.
Je souhaite aussi exprimer ma reconnaissance au programme Numerics de l'Université de Bordeaux, qui m'a permis de bénéficier d'un cadre de formation académique de haut niveau et de ressources précieuses pour la réalisation de ce travail.
Je n'oublie pas mes camarades et collègues de recherche, dont les échanges m'ont permis d'enrichir mes réflexions et d'élargir mes horizons.
Enfin, j'adresse une pensée particulière à ma famille et à mes proches, pour leur soutien moral et affectif constant. Leur confiance et leurs encouragements m'ont donné la motivation de toujours aller de l'avant.

Résumé

Résumé —Le carbone organique du sol (SOC) est un levier central pour la fertilité, la rétention en eau et l'atténuation du changement climatique. Dans le contexte sahélien, la variabilité pluviométrique, l'érosion et la pression anthropique exigent des outils de cartographie fiables et reproductibles. Ce travail propose un pipeline de prédiction du SOC dans trois villages du centre du Sénégal (Bary, Diohine/Sassem, Sop) combinant données pédologiques (AfSP et IRD), télédétection optique (Sentinel-2, Landsat-8) et apprentissage automatique. Les images ont été filtrées et agrégées en composites trimestriels (2013–2025 pour Landsat-8 ; 2015–2025 pour Sentinel-2), puis enrichies en indices spectraux (NDVI, NDWI, BSI, CI-Green) et en covariables topographiques et pédologiques globales. Plusieurs modèles ont été entraînés (RF, Gradient Boosting, XGBoost, HistGradientBoosting, MLP PyTorch) sous validation stricte par groupes (GroupKFold) afin de limiter les fuites spatiales. Les meilleurs résultats sont obtenus avec un MLP (stacking en appui), avec des performances atteignant $R^2 = 0.80$ et $RMSE = 0.52$ *g/kg* pour 0–10 cm, et $R^2 = 0.65$ et $RMSE = 0.66$ *g/kg* pour 10–30 cm. Des cartes d'incertitude (écart-type et intervalles à 90 %) accompagnent les cartes prédictives pour appuyer la décision et orienter de futures campagnes terrain. Enfin, nous discutons les limites (synchronisation temporelle données sol / satellites, hétérogénéités locales) et proposons des pistes d'amélioration (hybrides géostatistiques, capteurs complémentaires, enrichissement des données locales).

Mots-clés : carbone organique du sol ; télédétection ; Sentinel-2 ; Landsat-8 ; apprentissage automatique.

Abstract

Abstract — Soil Organic Carbon (SOC) underpins soil fertility, water retention and climate mitigation. In the Sahel, strong rainfall variability, erosion and human pressure call for reliable and reproducible SOC mapping tools. We design a prediction pipeline for three villages in central Senegal (Bary, Diohine/Sassem, Sop) that blends soil observations (IRD), optical remote sensing (Sentinel-2, Landsat-8) and machine learning. Satellite archives are quality-filtered and aggregated into quarterly composites (2013–2025 for Landsat-8; 2015–2025 for Sentinel-2) and complemented with spectral indices (NDVI, NDWI, BSI, CI-Green) plus topographic and global soil covariates. We train multiple models (RF, Gradient Boosting, XGBoost, HistGradientBoosting, PyTorch MLP) under strict group-based validation (GroupKFold) to avoid spatial leakage. The best performance is achieved by an MLP (with stacking support), reaching $R^2 = 0.80$ and $RMSE = 0.52$ *g/kg* at 0–10 cm, and $R^2 = 0.65$ and $RMSE = 0.66$ *g/kg* at 10–30 cm. Predictive maps are paired with uncertainty layers (standard deviation and 90% confidence intervals) to guide decisions and plan additional field sampling. We discuss limitations (temporal alignment between soil and satellite data, local heterogeneity) and outline improvements (hybrid geostatistical methods, complementary sensors, and richer local datasets).

Keywords: soil organic carbon; remote sensing; Sentinel-2; Landsat-8; machine learning.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 problématique et contexte

Le carbone organique du sol (SOC) joue un rôle fondamental dans la fertilité[1], la régulation hydrique et l'atténuation du changement climatique. Dans les zones sahéliennes, et en particulier au Sénégal, la pression agricole sur les terres du bassin arachidier, combinée à une forte variabilité pluviométrique et à l'érosion, menace la durabilité des systèmes de production. Or, le suivi du SOC constitue un levier majeur pour améliorer la productivité agricole, préserver les ressources en eau et contribuer aux engagements climatiques internationaux. Malgré son importance, la cartographie du SOC reste complexe. Les mesures directes par prélèvements et analyses en laboratoire sont précises, mais elles demeurent coûteuses, chronophages et spatialement limitées. En parallèle, la télédétection et l'apprentissage automatique offrent aujourd'hui des alternatives puissantes pour estimer et cartographier le SOC à grande échelle, mais leur mise en œuvre en contexte sahélien reste encore peu explorée.

Dans ce cadre, il est crucial de concevoir des outils de prédiction fiables, reproductibles et adaptés aux réalités locales, afin de fournir des cartes utiles aux chercheurs, décideurs et agriculteurs.

1.2 Carbone organique du sol

1.2.1 Cycle du carbone et puits de carbone

Le SOC représente un compartiment essentiel du cycle global du carbone[1]. Il agit comme un puits de carbone, en stockant une partie du CO_2 atmosphérique sous forme de matière organique. La dynamique de ce stockage dépend des pratiques agricoles, des conditions climatiques et des propriétés physico-chimiques des sols.

1.2.2 Rôle du SOC dans les écosystèmes

Au-delà de son rôle climatique, le SOC conditionne la structure du sol, sa porosité et sa capacité de rétention en eau. Il favorise la productivité agricole en soutenant la fertilité et en renforçant la résilience des sols face aux stress climatiques (sécheresse, érosion). La dégradation du SOC entraîne au contraire un appauvrissement des sols et une vulnérabilité accrue des systèmes agricoles.

1.2.3 Séquestration du carbone

La séquestration [2] du carbone organique dans les sols est un levier d'atténuation du changement climatique. Des pratiques telles que les rotations culturales, l'agroforesterie, l'apport de résidus organiques ou le maintien de couvertures végétales permettent d'accroître les stocks de SOC. Toutefois, ces gains restent hétérogènes dans l'espace et se matérialisent lentement, ce qui rend nécessaire le suivi spatial et temporel du SOC à haute résolution. Les pratiques de gestion durable des terres, telles que l'agroforesterie, la couverture végétale et l'application de matières organiques, peuvent favoriser la séquestration du carbone dans le sol. Cependant, les gains en matière de SOC sont souvent lents à se matérialiser et varient considérablement en fonction des contextes locaux.

1.3 Objectifs scientifiques et originalité du travail

L'objectif général de ce stage est de concevoir un pipeline opérationnel et reproductible pour estimer et cartographier le SOC (0–10 cm et 10–30 cm) dans trois villages pilotes du centre du Sénégal (Bary, Diohine/Sassem et Sop).

Plus spécifiquement, ce travail vise à :

- Intégrer des données multi-sources : profils pédologiques (AfSP, IRD, WoSIS), imagerie satellitaire optique (Sentinel-2, Landsat-8), indices spectraux et covariables environnementales (topographie, sols globaux).
- Évaluer plusieurs approches de modélisation : géostatistique, méthodes d'apprentissage automatique (forêts aléatoires, boosting) et réseaux de neurones (MLP PyTorch), avec une validation rigoureuse par groupes pour limiter les fuites spatiales.
- Cartographier l'incertitude associée aux prédictions, afin de fournir aux utilisateurs non seulement une estimation du SOC mais aussi un indicateur de fiabilité.
- Déployer une application web interactive (Django/PostGIS) intégrant une API de prédiction et une interface cartographique pour rendre les résultats accessibles aux chercheurs et décideurs.

1.4 Conclusion

Cette introduction met en évidence l'importance du SOC dans la fertilité des sols et la lutte contre le changement climatique, tout en soulignant les limites des approches classiques de mesure. Elle justifie le recours à la télédétection et à l'apprentissage automatique pour développer des outils adaptés au contexte sahélien.

Le chapitre suivant présentera un état de l'art sur les approches existantes de cartographie du SOC, les données mobilisées (profils pédologiques, télédétection), ainsi que les principaux projets et travaux antérieurs qui ont inspiré et guidé ce mémoire.

Chapitre 2

État de l'art

2.1 Concepts et enjeux

La mesure directe du SOC repose sur des prélèvements de sol suivis d'analyses en laboratoire (ex. combustion sèche, oxydation humide). Bien que précises, ces approches sont coûteuses, localisées et difficilement généralisables.

Des méthodes de proxidétection, notamment la spectroscopie de terrain, offrent des alternatives rapides mais restent limitées par des contraintes logistiques et une calibration complexe. Ces limites justifient l'essor des approches indirectes, fondées sur la télédétection et l'apprentissage statistique ou automatique, qui permettent de relier des observations ponctuelles à des covariables spatiales plus facilement disponibles.

2.2 Données d'observation optique

2.2.1 Sentinel-2 (MSI)

Lancé en 2015, le programme Sentinel-2 [3] fournit des images multispectrales à 10–20 m de résolution et à une fréquence de 5 jours. Ces images corrigées atmosphériquement (produits L2A) permettent de calculer des indices biophysiques pertinents pour la dynamique du SOC : NDVI (végétation), NDWI (humidité), BSI (sol nu), CI-Green (chlorophylle). Afin de réduire le bruit lié aux nuages et aux variations locales, des composites temporels trimestriels sont souvent utilisés.

2.2.2 Landsat-8 (OLI/TIRS)

Avec une résolution plus modeste (30 m) mais une archive longue (depuis 2013), Landsat-8 [4] complète Sentinel-2 en apportant une profondeur temporelle utile pour détecter des

tendances interannuelles. L'harmonisation des indices spectraux avec Sentinel-2 permet de constituer des séries temporelles cohérentes

2.2.3 Covariables exogènes

Au-delà de l'imagerie optique, des covariables environnementales enrichissent la prédiction du SOC :

Topographie (MNT, pente, exposition, via SRTM/Copernicus DEM) [**copernicusdem**, 5],

Climat (pluviométrie, température),

Bases pédologiques globales (SoilGrids, iSDAsoil).

Ces proxys fournissent des informations complémentaires sur la productivité végétale, la redistribution hydrique ou les propriétés texturales du sol.

2.3 Projets internationaux de cartographie du SOC

2.3.1 iSDAsoil

iSDAsoil [6] propose des cartes pédologiques à haute résolution ($\approx 30\,\text{m}$) pour l'Afrique, obtenues par l'intégration de profils de sol, spectroscopie et covariables de télédétection, avec des modèles d'apprentissage automatique.

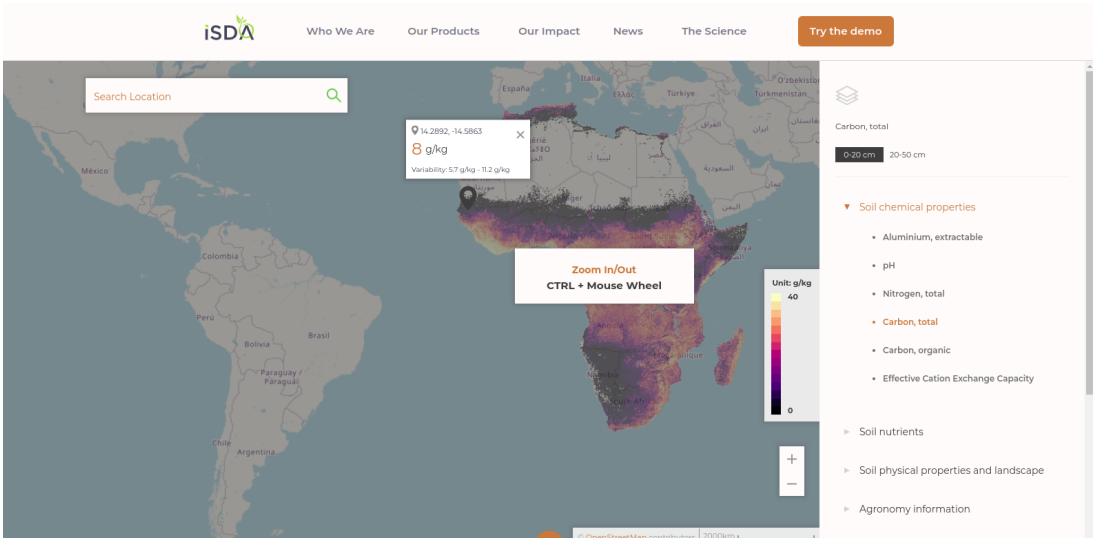


Figure 2.1 : iSDAsoil

2.3.2 SoilGrids

SoilGrids [7] (ISRIC) produit des cartes mondiales du SOC à des résolutions typiques de 250 m à 1 km, basées sur des profils harmonisés et des covariables multiples. Si ces

produits constituent une référence internationale, leur manque de finesse locale limite leur applicabilité dans les contextes sahéliens, où l'hétérogénéité spatiale est forte.

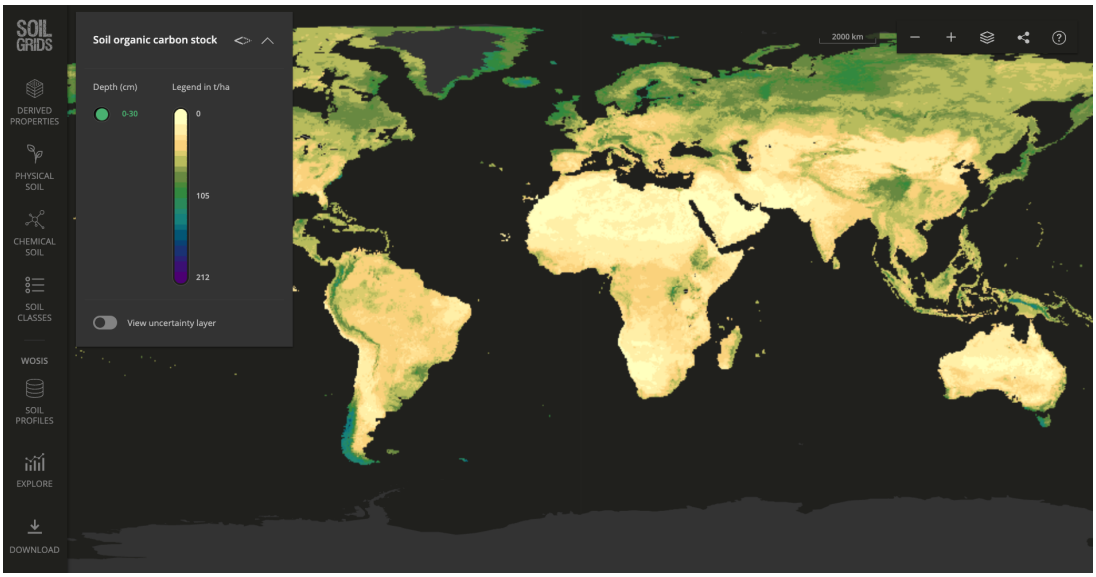


Figure 2.2 : SoilGrids

2.4 Approches de cartographie pédologique

2.4.1 Géostatistique

La géostatistique [8] (variogramme, krigeage) reste une méthode classique pour interpoler le SOC à partir de données ponctuelles. Elle est efficace dans des contextes où la densité d'échantillonnage est élevée, mais sa performance chute lorsque les données sont peu nombreuses ou hétérogènes.

2.4.2 Apprentissage automatique

Les modèles d'apprentissage supervisé (forêts aléatoires, boosting, réseaux de neurones) exploitent [9, 10] la relation entre SOC et covariables optiques, topographiques ou climatiques. Contrairement à la géostatistique, ils capturent mieux les relations non linéaires et s'adaptent à la complexité des interactions sol-végétation-climat. Des approches d'assemblage (stacking) renforcent encore la généralisabilité.

2.5 Protocoles d'évaluation et métriques

La validation est un enjeu crucial[11]. Les validations aléatoires surestiment souvent les performances en raison de fuites spatiales. Pour y remédier, des schémas plus rigoureux (ex.

GroupKFold par site/profil) sont utilisés, garantissant une séparation stricte des données. Les métriques standards incluent le R^2et le RMSE , exprimés en g/kg.

2.6 Incertitude et interprétabilité

La cartographie du SOC ne peut se limiter à une prédiction ponctuelle : il est essentiel de fournir une mesure d'incertitude[12]. Celle-ci peut être estimée par :

- des ensembles et bootstrap (dispersion inter-modèles),
- le Monte Carlo Dropout pour les réseaux de neurones,
- l'analyse des résidus et de l'importance des variables, afin d'identifier les zones sous-contraintes en données.

2.7 Travaux de Thiam (2022–2023)

En 2022, Thiam a mené une étude sur le même sujet [13].

2.7.1 géostatistique vs apprentissage automatique

Les travaux de Thiam comparent des interpolateurs géostatistiques (krigeage) et des modèles de machine learning (forêts et boosting) sur $\approx 3\,600$ entrées ($\approx 1\,800$ profils) pour deux horizons (0–10 cm et 10–30 cm). Les modèles de machine learning surpassent globalement le krigeage, surtout en 0–10 cm. La précision diminue en 10–30 cm, cohérente avec la moindre sensibilité des capteurs optiques aux couches profondes. Les métriques doivent être rapportées avec unités (g/kg) pour assurer la comparabilité.

Modèles	0-10 cm		10-30 cm	
	RMSE	R^2	RMSE	R^2
XGBRegressor	1.456	0.561	2.213	0.373
LGBMRegressor	1.377	0.607	2.263	0.344
GradientBoostingRegressor	1.354	0.620	2.244	0.355
HistGradientBoostingRegressor	1.375	0.608	2.233	0.361
RandomForestRegressor	1.478	0.548	2.355	0.290

Tableau 2.1 – Performance des modèles de régression pour les profondeurs de 10 cm et 30 cm

Modèles	0-10 cm	10-30 cm
	RMSE	RMSE
Sphérique	1.5812	5.2579
Gaussien	1.5811	5.3068
Exponentiel	1.5762	5.2647

Tableau 1.1 – Performance des modèles de variogramme pour les profondeurs de 10 cm et 30 cm

Figure 2.3 : Tiam 2022 : Comparaison entre apprentissage automatique et géostatistique

2.7.2 Modèles utilisés

L'étude met en regard des interpolateurs géostatistiques (krigeage ordinaire et variantes) et des modèles d'apprentissage supervisé . Le protocole retient deux horizons (0–10 et 10–30 cm) et un jeu de $\sim 3\,600$ entrées correspondant à $\sim 1\,800$ profils.

2.7.3 Résultats

L'étude de Thiam constitue une référence en contexte sénégalais.

- Comparaison géostatistique vs apprentissage automatique : sur $\approx 3\,600$ entrées ($\approx 1\,800$ profils) pour deux horizons (0–10 cm et 10–30 cm), les modèles d'apprentissage automatique surpassent globalement le krigeage, notamment en surface (0–10 cm).
- Limites identifiées : volume de données encore limité au regard de l'hétérogénéité pédologique, perte de précision marquée dans la couche 10–30 cm, absence de cartographie explicite de l'incertitude.

Ces résultats confirment la pertinence du machine learning mais soulignent la nécessité de protocoles plus stricts et de l'intégration de nouvelles sources de données.

Métrique	0-10 cm	10-30 cm
RMSE	1.337	2.208
R^2	0.630	0.375

Tableau 2.2 – Performance du modèle de stacking pour les profondeurs de 10 cm et 30 cm

Figure 2.4 : Résultats de Thiam 2022

2.8 Synthèse et lacunes

La littérature récente met en évidence :

- la puissance des approches d'apprentissage automatique pour améliorer la précision des cartes de SOC,
- la nécessité de combiner données locales et globales pour réduire les biais,
- l'importance d'intégrer validation spatiale et incertitude pour garantir la robustesse,
- la difficulté persistante de prédire les couches profondes (10–30 cm) avec l'optique seule.

Ces constats ouvrent la voie à des pipelines reproductibles et enrichis, combinant bases locales (IRD), données satellitaires récentes (Sentinel-2, Landsat-8), et outils de déploiement applicatif.

C'est dans ce cadre que s'inscrit le présent travail, qui vise à dépasser ces limites en

développant un pipeline robuste, intégrant incertitude et reproductibilité, et en proposant une application opérationnelle adaptée au contexte sénégalais.

Chapitre 3

Cadre du stage

3.1 Laboratoire d'accueil et encadrement

Le stage s'est déroulé au sein de l'UMMISCO Sénégal [14] (Unité Mixte Internationale de Modélisation Mathématique et Informatique des Systèmes Complexes), basé à l'Université Cheikh Anta Diop (UCAD II) de Dakar.

UMMISCO est une structure pluridisciplinaire qui associe mathématiques appliquées, informatique et sciences environnementales pour répondre à des enjeux de développement durable. Ses axes de recherche incluent la télédétection, l'intelligence artificielle et la modélisation des écosystèmes complexes.

L'unité fonctionne dans un cadre de coopération internationale, en lien étroit avec l'IRD [15] (Institut de Recherche pour le Développement).

L'encadrement scientifique a été assuré par des chercheurs spécialisés en pédologie, télédétection et apprentissage automatique. Des réunions hebdomadaires de suivi ont permis d'orienter les choix méthodologiques, de garantir la cohérence avec les objectifs du projet SLAM-B [16], et d'assurer un équilibre entre production scientifique et mise en œuvre applicative.

3.2 Ferlo-Sine

Ce stage s'inscrit dans le cadre du projet SLAM-B (Soil Land and Agroecosystem Management in the Ferlo Basin), qui vise à développer des outils innovants pour la gestion durable des terres dans la région du Ferlo.
Au sein de SLAM-B, le Scenario Lab Ferlo-Sine joue un rôle central. Il constitue un espace collaboratif où chercheurs, institutions locales et décideurs co-construisent des approches de suivi environnemental et socio-économique. L'accent est mis sur :

- l'intégration de données de télédétection et de terrain,
- la simulation des dynamiques agro-écologiques,
- et le développement d'outils d'aide à la décision pour la gestion des ressources naturelles.

Le présent travail s'inscrit directement dans ces objectifs, en proposant un pipeline de prédiction et de cartographie du SOC, assorti d'une application interactive.

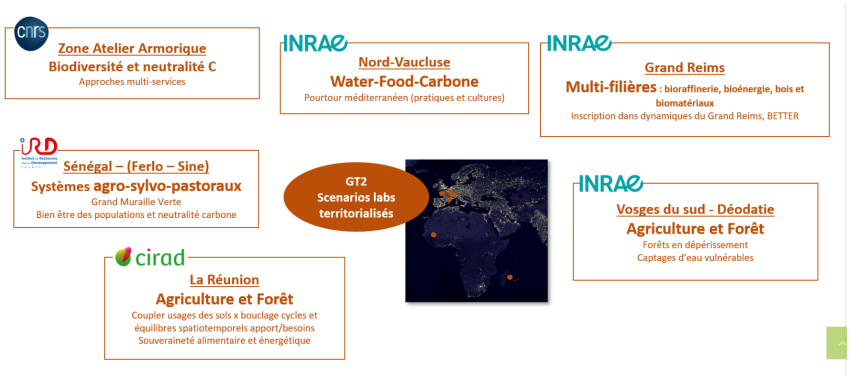


Figure 3.1 : Ferlo-Sine

3.3 Zone d'étude : Bary, Diohine/Sassem et Sop

Trois villages du centre du Sénégal ont été retenus comme sites pilotes :

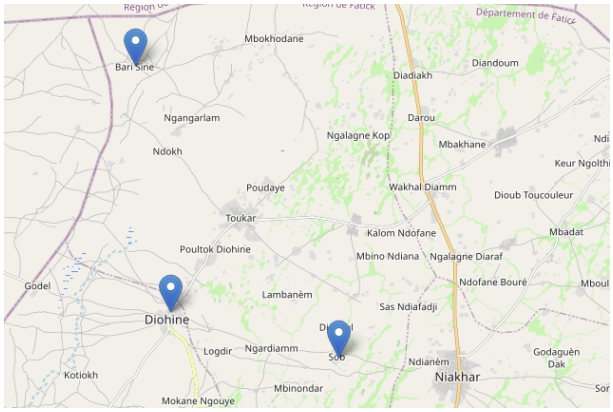


Figure 3.2 : Zone d'étude

Ces trois villages offrent une représentativité intéressante des contextes pédologiques et agroécologiques du bassin arachidier. Leur sélection permet d'évaluer la généralisabilité des modèles de prédiction du SOC dans des environnements variés.

3.4 Objectifs opérationnels et jalons

Le stage avait pour objectif opérationnel la mise en place d'un pipeline complet de prédiction du SOC, depuis la collecte des données jusqu'au déploiement applicatif. Les jalons principaux étaient :

1. **Revue bibliographique et cadrage scientifique** (avril–mai) :analyse des enjeux, méthodologies existantes et identification des lacunes.

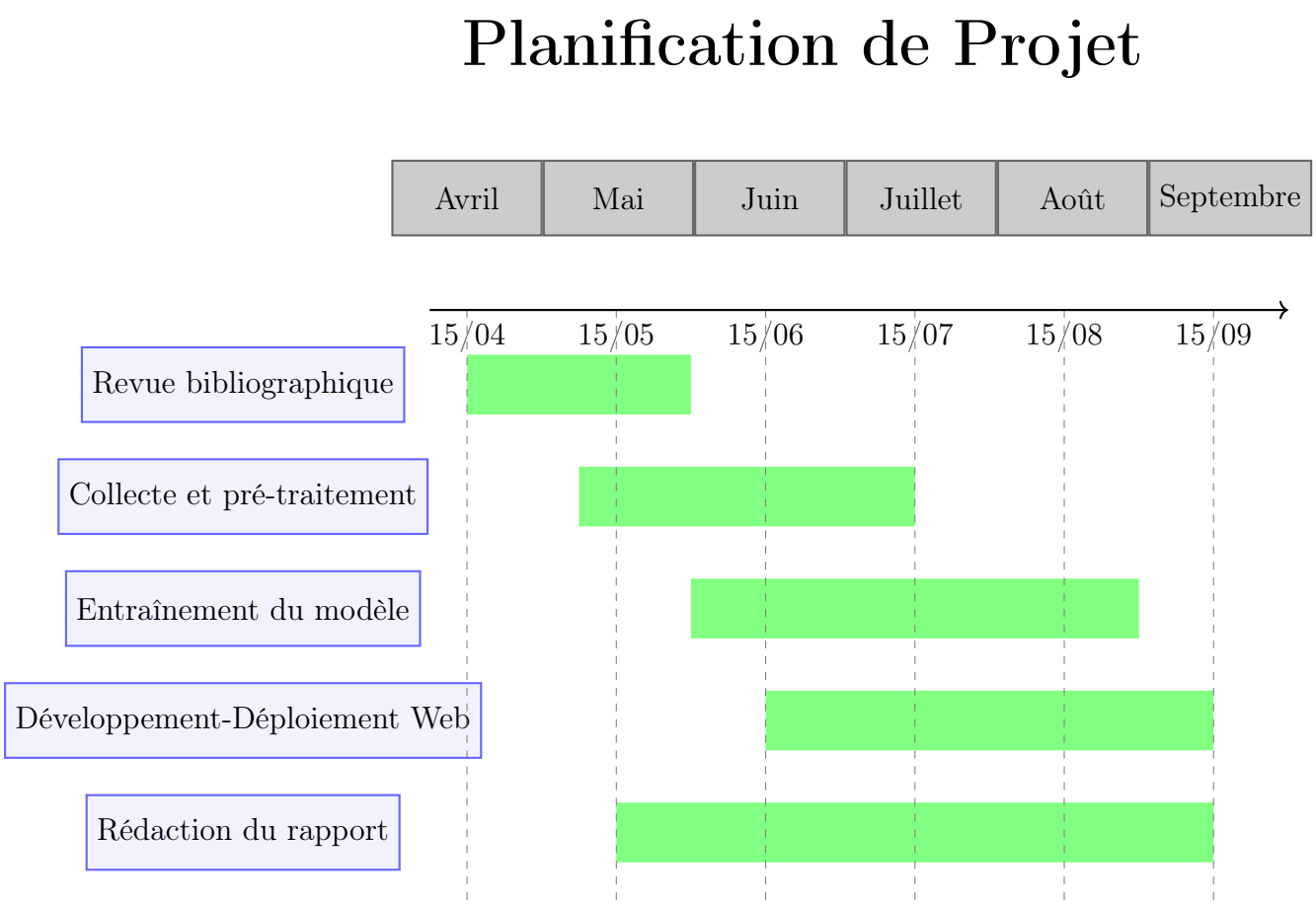
2. **Collecte et harmonisation des données** (mai–juin) : intégration des profils pédologiques (AfSP, IRD, WoSIS), préparation des images Sentinel-2 et Landsat-8, construction des covariables dérivées.

3. **Modélisation et évaluation** (juin–juillet) : entraînement des modèles (forêts aléatoires, boosting, MLP), validation spatiale stricte, analyse des performances et incertitudes.

4. **Déploiement applicatif** (juillet–août) : développement d'une API /predict et d'une interface cartographique interactive (Django + PostGIS).

5. **Rédaction du mémoire et restitution** (août) : synthèse scientifique, analyse critique, discussion des perspectives.

Cette planification a permis de garantir un équilibre entre recherche et application, avec un produit final à la fois scientifique (pipeline reproductible, résultats chiffrés) et pratique (application opérationnelle).



Chapitre 4

Données et outils

4.1 Sources et caractéristiques des données

La prédiction du stock de carbone organique du sol (SOC) repose sur la combinaison de données pédologiques de référence et de covariables satellitaires et environnementales.

4.1.1 Profils AfSP

La base Africa Soil Profiles [17] (AfSP) constitue une ressource continentale de référence. Elle regroupe 18 533 profils pédologiques collectés à travers l'Afrique entre 1938 et 2011, chacun comprenant jusqu'à quatre couches de profondeur.

- Avantage : taille du jeu de données, diversité spatiale et richesse des attributs physico-chimiques.
- Limite : hétérogénéité temporelle (décalage par rapport aux satellites récents) et variabilité dans les méthodes de mesure.

Seuls les horizons 0–10 cm et 10–30 cm ont été retenus pour rester cohérents avec les profils IRD et WoSIS.

Figure 4.1 : Profils AfSP

4.1.2 Données locales de l'IRD

L'**Institut de Recherche pour le Développement (IRD)** a fourni un jeu de données local spécifiquement centré sur le Sénégal (Ferlo-Sine).

- Taille : environ 1 800 profils (deux horizons : 0–10 cm et 10–30 cm).
- Période d'échantillonnage : 2016–2017, contemporaine des séries Sentinel-2 et Landsat-8.
- Avantage : synchronisation temporelle avec les satellites, permettant une calibration robuste.
- Limite : couverture restreinte à quelques villages pilotes (Bary, Diohine/Sassem, Sop), ce qui limite la généralisabilité directe.

Ces données constituent la base de validation principale du pipeline développé.

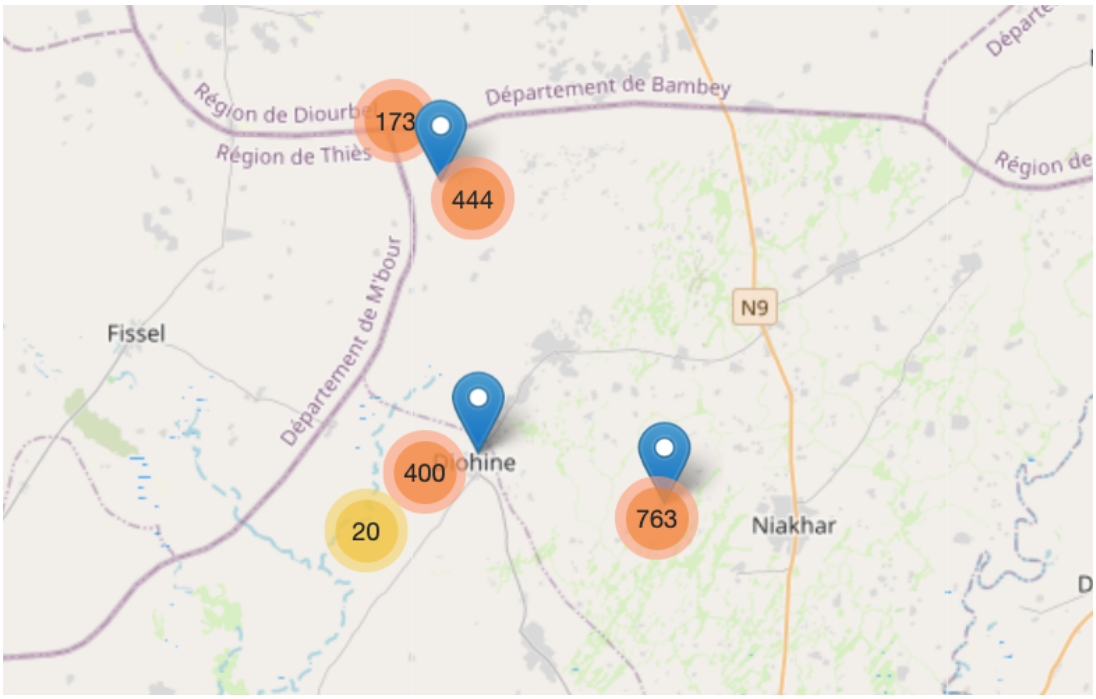


Figure 4.2 : Zone d'étude

4.1.3 Profils WoSIS

La base **WoSIS (World Soil Information Service)** [18] compilée par l'ISRIC, rassemble plus de 196 297 profils pédologiques mondiaux.

- Caractéristiques : nombre de couches variable selon les profils ; période de collecte souvent inconnue ou hétérogène.

- Avantage : dimension internationale et homogénéisation des profils, utile pour situer les valeurs locales dans un cadre global.
- Limite : incohérences dans les profondeurs et manque d'informations temporelles.

Dans ce travail, WoSIS est mobilisé comme cadre de comparaison global et pour enrichir l'entraînement des modèles.

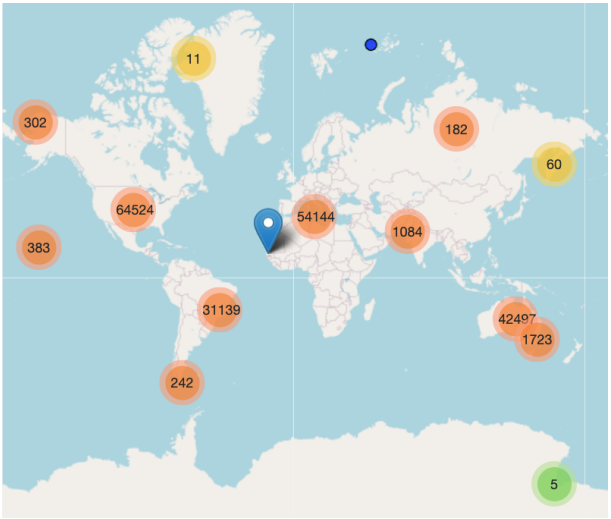


Figure 4.3 : Profils WoSIS

4.1.4 Imagerie Sentinel-2 et Landsat-8

L'imagerie satellitaire constitue la principale source de covariables spatiales :

- **Sentinel-2 [3] (MSI)** : données multispectrales à 10–20 m de résolution spatiale, couvrant la période 2015–2025.
- **Landsat-8 [4] (OLI/TIRS)** : données multispectrales à 30 m, disponibles depuis 2013, permettant d'allonger la série temporelle.

Ces deux familles de capteurs offrent une complémentarité : haute résolution spatiale pour Sentinel-2 et profondeur temporelle pour Landsat-8.

4.1.5 Variables dérivées (indices spectraux, topographie, sols globaux)

À partir des images satellites et de bases complémentaires, plusieurs covariables ont été construites :

- **Indices spectraux** : NDVI (végétation), NDWI (humidité), BSI (sols nus), CI-Green (chlorophylle).

- **Variables topographiques** : altitude, pente, exposition issues de modèles numériques de terrain (SRTM, Copernicus DEM).

4.2 Préparation et qualité des données

4.2.1 Contrôles et nettoyage

Un ensemble de contrôles a été mis en place pour garantir la fiabilité des données :

- Détection et suppression des doublons ou valeurs aberrantes dans les profils.
- Masquage des nuages dans Sentinel-2 et Landsat-8.
- Vérification de la cohérence .

4.2.2 Harmonisation spatiale et temporelle

Les données ont été harmonisées afin de permettre leur intégration dans un même pipeline :

- Projection dans un système de coordonnées commun (WGS84 / UTM).
- Normalisation des profondeurs en 0–10 cm et 0–30 cm pour assurer une cohérence avec les cibles retenues.

Chapitre 5

Méthodologie

5.1 Pipeline de collecte et préparation des données

Le pipeline de données a été construit afin de relier les observations pédologiques ponctuelles avec les variables de télédétection et environnementales. L'ensemble du processus a été automatisé via **Google Earth Engine (GEE)** et complété par des scripts Python. Les principales étapes sont :

1. **Filtrage spatial (zones d'intérêt)** : Bary, Diohine/Sassem, Sop),
2. **Filtrage temporel** (2013–2025 pour Landsat-8, 2015–2025 pour Sentinel-2),
3. **Sélection des capteurs et période** : Sentinel-2 (2015–2025) et Landsat-8 (2013–2025).
4. **Nettoyage : Filtres de qualité** : masques de nuages (Sentinel-2 QA60).
5. **Agrégation trimestrielle (médiane)** : médianes trimestrielles pour réduire le bruit et capturer la saisonnalité.
6. **Extraction ponctuelle pour chaque profil pédologique**,
7. **Extraction des variables dérivées** : indices spectraux (NDVI, NDWI, BSI, CI-Green), polarisation radar (VV, VH, ratio), variables topographiques et pédologiques.
8. **Harmonisation et intégration** : reprojection sur une grille commune et intégration dans PostGIS.

Le pipeline s'appuie uniquement sur l'optique (Sentinel-2 L2A, Landsat-8 OLI/TIRS). Après filtrage qualité (masques nuages) et agrégation en composites trimestriels, nous extrayons NDVI, NDWI, BSI, CI-Green ainsi que des covariables topographiques (MNT, pente, exposition) et pédologiques globales. Ces variables alimentent des modèles tabulaires (RF, boosting) et un MLP PyTorch, avec normalisation adéquate et GroupKFold par site.

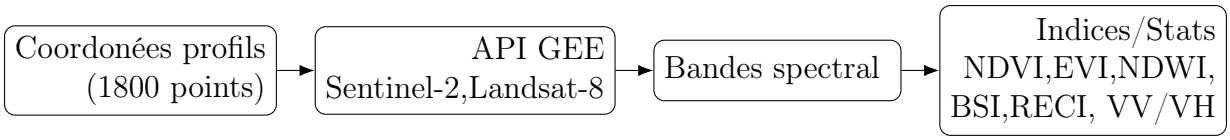


Figure 5.1 : pipeline de collecte des données satellites.

5.1.1 Optimisation et réglage des modèles

Une optimisation systématique des hyperparamètres a été réalisée :

- Recherche en grille et optimisation bayésienne (Optuna).
- Paramètres explorés : profondeur maximale, nombre d'arbres, taux d'apprentissage, régularisation.
- Validation croisée imbriquée pour sélectionner les configurations robustes.

5.2 Multi-Layer Perceptron

Un modèle avancé a été implémenté sous PyTorch afin d'exploiter la capacité des réseaux de neurones à modéliser des relations complexes.

- **Architecture** : couches denses (256–128–64 neurones), activation ReLU, dropout (0.2–0.3).
- **Normalisation** : standardisation des variables continues et encodage des variables catégorielles.
- **Fonction de perte** : erreur quadratique moyenne (MSE).
- **Stratégie OOF (Out-Of-Fold)** : entraînement croisé avec sauvegarde des prédictions par pli.

5.3 Quantification et cartographie de l'incertitude

L'incertitude des prédictions a été intégrée pour renforcer la valeur décisionnelle des cartes produites :

- **Ensembles et bootstrap** : variance des prédictions entre modèles.
- **Dropout bayésien** : approximation par Monte Carlo Dropout.
- **Cartes d'écart-type** : représentation spatiale de la dispersion des prédictions.
- **Intervalles de confiance (90%)** : bornes inférieure et supérieure des prédictions de SOC.

5.4 Limites méthodologiques et ajustements

Une approche initiale consistait à entraîner un modèle générique sur la base AfSP (18 533 profils, 4 couches, 1938–2011), puis à effectuer un fine-tuning local sur les données IRD.

Cependant, le décalage temporel entre AfSP et les satellites récents (Sentinel-2 à partir de 2015) a limité la transférabilité directe. Les ajustements ont donc consisté à :

privilégier les données synchronisées temporellement (IRD 2016–2017),

restreindre les horizons aux couches superficielles (0–10 cm, 10–30 cm),

intégrer les bases WoSIS et SoilGrids comme complément de calibration.

5.5 Innovations méthodologiques

Par rapport aux travaux antérieurs (ex. Thiam, 2022–2023), ce mémoire introduit plusieurs innovations :

- Validation Leave-Area-Out (LAO)
- Chaque village (Bary, Diohine/Sassem, Sop) est exclu à tour de rôle du jeu d'entraînement et utilisé comme zone de test indépendante.
- Cela permet d'évaluer la généralisabilité spatiale des modèles et de détecter les biais liés à la structure des données.
- Prédiction hiérarchique SOC10 $\rightarrow$ SOC30
- La prédiction du SOC à 10–30 cm intègre comme covariable le SOC prédit en surface (0–10 cm).
- Ce transfert améliore la stabilité des prédictions dans les couches profondes, moins bien captées par la télédétection optique.
- Approche par site
- Des modèles spécifiques ont été calibrés pour chaque village pilote afin de capturer les spécificités locales, avant d'être comparés à un modèle global.

Ces innovations renforcent à la fois la robustesse méthodologique et la pertinence opérationnelle du pipeline.

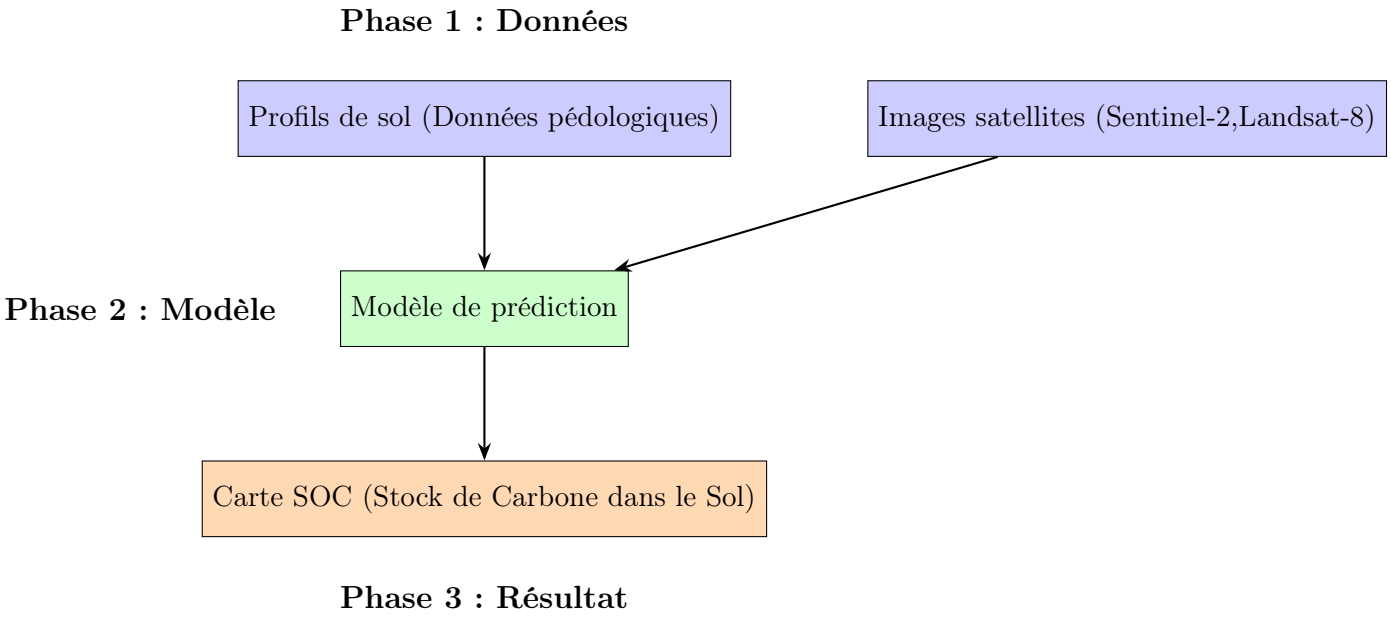


Figure 5.2 : Écosystème du modèle de prédiction du stock de carbone dans le sol

5.6 Modèles de référence

Afin d'établir une base comparative solide, plusieurs modèles supervisés ont été entraînés :

- Régression linéaire régularisée (Ridge, Lasso),
- Forêts aléatoires (RF),
- Gradient Boosting (XGB, LightGBM, HistGradientBoosting).

Les validations ont suivi un schéma rigoureux :

- **GroupKFold par site** : les profils d'un même village sont exclus du jeu d'apprentissage lorsqu'ils sont utilisés pour le test, afin d'éviter toute fuite spatiale.
- **Métriques** : coefficient de détermination (R2), erreur quadratique moyenne (RMSE) .

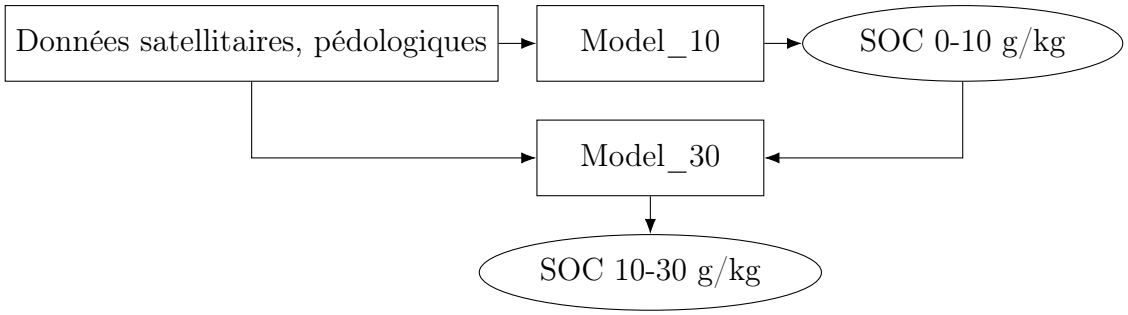


Figure 5.3 : Apprentissage des modèles pour les couches de SOC 0-10 cm et 10-30 cm

Chapitre 6

Résultats

6.1 Modèles de référence

Les modèles de référence incluent plusieurs méthodes d'apprentissage supervisé basées sur des arbres de décision (Random Forest, Gradient Boosting, XGBoost, LightGBM, HistGradientBoosting), ainsi qu'un assemblage par stacking. Les résultats (Tableau 6.1) montrent des performances variables selon la profondeur du sol.

Modèle	**0-10**		**0-30**	
	R^2	**RMSE**	**R^2**	**RMSE**
LGBMR	0.63	0.97	0.42	0.72
GBR	0.54	1.20	0.40	0.75
XGBR	0.57	1.12	0.39	0.76
HistGradientBoosting	0.65	0.90	0.43	0.72
RandomForest	0.61	1.02	0.45	0.70
stacking	0.57	1.05	0.41	0.74

Table 6.1 : Résultats des modèles de référence.

Pour la couche 0–10 cm, les coefficients de détermination (R^2) se situent entre 0.54 (Gradient Boosting) et 0.65 (HistGradientBoosting), avec des RMSE compris entre 0.90 et 1.20 g/kg. Ces valeurs indiquent une capacité prédictive correcte, mais encore limitée pour capturer les variations fines du SOC en surface. Parmi ces modèles, HistGradientBoosting apparaît comme le plus performant ($R^2 = 0.65$; RMSE = 0.90), probablement en raison de sa capacité à gérer les distributions hétérogènes et les relations non linéaires complexes.

Pour la couche 10–30 cm, les performances sont globalement plus faibles, avec des R^2 variant de 0.39 (XGBR) à 0.45 (Random Forest). Les RMSE se situent autour de 0.70–0.76 g/kg. Ces résultats confirment que les indices spectraux issus de l'imagerie optique sont moins sensibles aux horizons profonds, ce qui complique la prédiction du SOC à 30 cm.

Néanmoins, le Random Forest obtient les meilleurs résultats relatifs en profondeur (R^2 = 0.45 ; RMSE = 0.70), traduisant une meilleure robustesse aux données bruitées.

L'assemblage par stacking, censé combiner les forces des différents modèles, n'apporte pas d'amélioration significative dans ce cas (R^2 = 0.57 et 0.41 pour 0–10 cm et 10–30 cm respectivement). Cela peut s'expliquer par la forte corrélation entre les modèles de base (tous issus d'arbres de décision), qui limite la complémentarité des prédictions.

En résumé, les modèles de référence parviennent à capturer une partie de la variabilité du SOC, mais leurs performances demeurent modestes, en particulier pour les couches profondes. Ces résultats justifient l'exploration de modèles plus puissants et flexibles, tels que les réseaux de neurones artificiels, afin de mieux exploiter la richesse des covariables et de dépasser les limites observées.

6.2 MLP

Le réseau de neurones Multi-Layer Perceptron (MLP), implémenté sous PyTorch, a montré des performances nettement supérieures à celles des modèles de référence. Les résultats par profondeur et par site (Tableau 6.2) révèlent plusieurs tendances Pour la couche 0–10

Table 6.2 : Performances par profondeur et site .

Profondeur	Site	R^2 (test)	RMSE (test)
10 cm	0	0.78	0.55
10 cm	1	0.87	0.43
10 cm	2	0.76	0.54
10 cm	all	0.80	0.52
30 cm	0	0.59	0.76
30 cm	1	0.68	0.62
30 cm	2	0.62	0.60
30 cm	all	0.65	0.66

cm, les prédictions atteignent un coefficient de détermination global de R^2 = 0.80 avec un RMSE de 0.52 g/kg, soit une amélioration notable par rapport aux modèles d'arbres (meilleurs $R^2 \approx 0.65$). Localement, les performances varient légèrement entre villages : Diohine/Sassem obtient les scores les plus élevés (R^2 = 0.87 ; RMSE = 0.43), tandis que Bary et Sop restent solides (R^2 entre 0.76 et 0.78). Ces résultats confirment la sensibilité de l'imagerie optique aux couches superficielles et la capacité du MLP à exploiter pleinement les interactions complexes entre indices spectraux et covariables pédologiques.

Pour la couche 10–30 cm, les performances restent plus modestes mais néanmoins meilleures

que celles des modèles de référence : R^2 global = 0.65 et RMSE = 0.66 g/kg. Les scores varient de 0.59 (Bary) à 0.68 (Diohine/Sassem). Comme attendu, les horizons profonds sont plus difficiles à prédire, la télédétection optique captant surtout les caractéristiques de surface. Toutefois, l'approche hiérarchique (intégration du SOC prédit à 0–10 cm comme covariable pour 10–30 cm) contribue à stabiliser les résultats et limite la dégradation de la performance.

6.2.1 Cartographie des prédictions du SOC et analyse des incertitudes

Les figures ci-dessous présentent les cartes de prédictions du stock de carbone organique du sol (SOC) pour les deux profondeurs étudiées, obtenues après agrégation sur une maille régulière de 250 m.

SOC 0–10 cm

La carte SOC_{10} (Fig. 6.1, gauche) met en évidence une forte hétérogénéité spatiale, avec des valeurs comprises entre 2 et 8 g·kg^{-1}. Plusieurs zones localisées, notamment au centre et au nord de la zone d'étude, présentent des teneurs plus élevées en carbone. Cette variabilité est cohérente avec les conditions pédologiques et l'occupation des sols, la couche superficielle étant directement influencée par la végétation, les pratiques agricoles et les apports organiques. Le modèle semble donc capturer efficacement les contrastes de surface détectables par les données de télédétection.

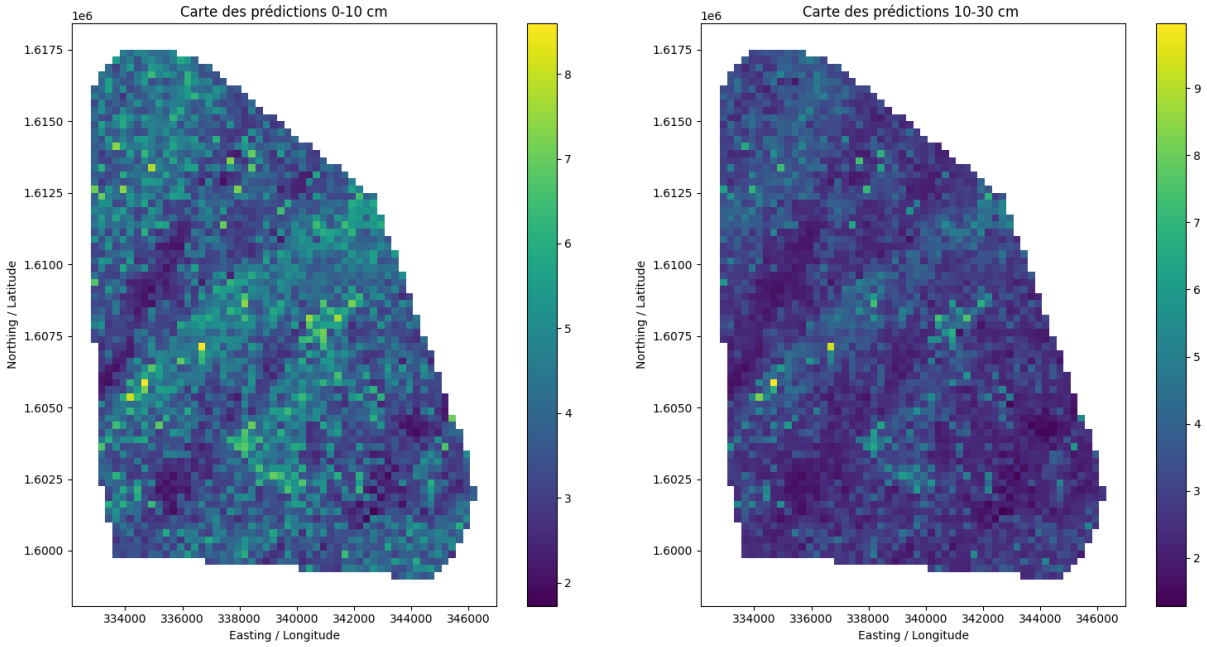
SOC 0–30 cm

La carte SOC_{30} (Fig. 6.1, droite) présente une structure spatiale globalement similaire mais avec des valeurs moyennes légèrement plus faibles et une variabilité plus atténuée. L'échelle des prédictions s'étend de 2 à 9 g·kg^{-1}, mais la majorité des mailles se situe entre 2 et 6 g·kg^{-1}. Cette homogénéisation relative reflète la difficulté à estimer le carbone en profondeur, les signaux satellitaires étant principalement sensibles aux couches superficielles.

Analyse des incertitudes

Afin d'évaluer la robustesse des prédictions, la distribution des moyennes (μ), des écarts-types (σ) et de la largeur des intervalles de confiance à 90% (IC90) a été examinée pour chaque profondeur (Fig. 6.2).

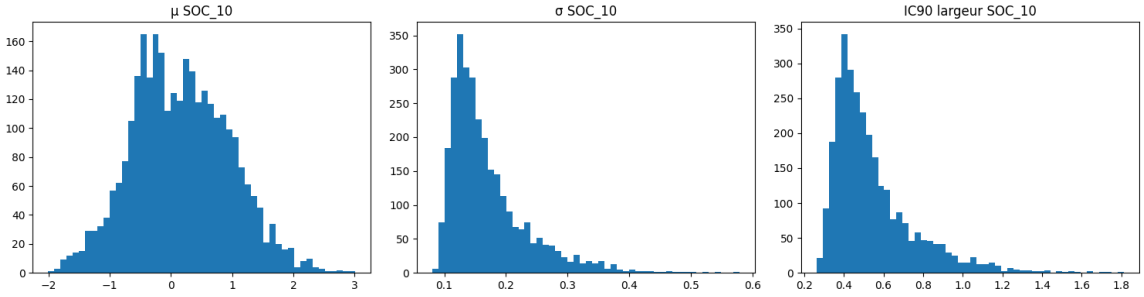
Pour le SOC_{10}, les prédictions présentent des incertitudes faibles, avec σ centré autour de 0.15–0.20 et des IC90 étroits ($\approx$ 0.4–0.6). En revanche, pour le SOC_{30}, les incertitudes sont plus marquées : σ atteint souvent 0.20–0.25 et les IC90 sont en moyenne deux fois



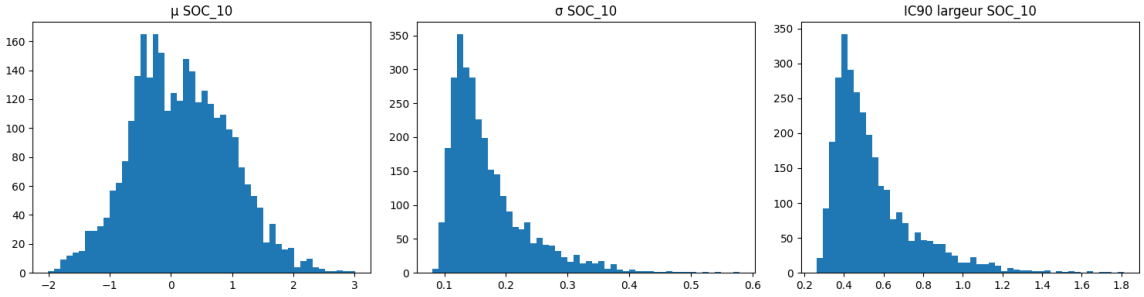
(a) Prédictions SOC_{10} (0–10 cm) **(b)** Prédictions SOC_{30} (0–30 cm)

FIGURE 6.1 : Cartes de prédictions du SOC obtenues à partir du modèle sur une maille de 250 m.

plus larges ($\approx$ 0.7–1.0). Cela confirme que les estimations en surface sont plus fiables, tandis que celles en profondeur doivent être interprétées avec davantage de prudence.



(a) Distribution des incertitudes pour SOC_{10} (0–10 cm)



(b) Distribution des incertitudes pour SOC_{30} (0–30 cm)

FIGURE 6.2 : Histogrammes des moyennes (μ), des écarts-types (σ) et des largeurs d'IC90 des prédictions SOC.

6.3 Analyse comparative des modèles

Modèle	0-10		0-30	
	R^2	RMSE	R^2	RMSE
LGBMR	0.63	0.97	0.42	0.72
GBR	0.54	1.20	0.40	0.75
XGBR	0.57	1.12	0.39	0.76
HistGradientBoosting	0.65	0.90	0.43	0.72
RandomForest	0.61	1.02	0.45	0.70
stacking	0.57	1.05	0.41	0.74
PyTorch MLP	0.80	0.52	0.65	0.66

Table 6.3 : Comparaison des performances entre les modèles.

6.4 Conclusion

La comparaison des modèles montre que le MLP sous PyTorch surpasse nettement les modèles de référence, avec des R^2 de 0.80 (0–10 cm) et 0.65 (10–30 cm), et des RMSE respectifs de 0.52 g/kg et 0.66 g/kg. Les erreurs sont plus faibles en surface (RMSE $\approx$ 0.52 g/kg) qu'en profondeur (RMSE $\approx$ 0.66 g/kg).
Les zones à faible densité d'échantillonnage présentent des erreurs plus élevées, confirmant l'importance de la répartition spatiale des profils.
Les valeurs extrêmes de SOC (très faibles ou très élevées) sont parfois mal prédites, reflétant une limite d'extrapolation des modèles.
Les modèles à base d'arbres ont tendance à sous-estimer les fortes valeurs de SOC, du fait de leur effet régularisateur.

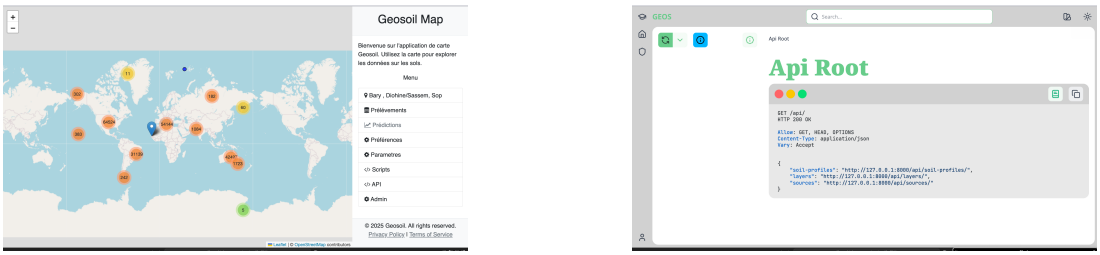
L'analyse des erreurs résiduelles met en évidence plusieurs éléments clés :

- Les erreurs sont globalement plus faibles pour la couche 0–10 cm ($RMSE = 0{,}52$) que pour la couche 10–30 cm ($RMSE = 0{,}66$), ce qui reflète la plus grande sensibilité des capteurs optiques et radar à la surface du sol.
- Les zones à faible densité d'échantillonnage présentent des erreurs plus élevées, ce qui confirme un biais spatial lié à l'inégale répartition des profils pédologiques.
- Certains profils avec des valeurs extrêmes de SOC (très faibles ou très élevées) sont mal prédits par les modèles, suggérant une capacité d'extrapolation limitée au-delà du domaine d'apprentissage.
- Les modèles basés sur les arbres (RF, XGB, LGBM) ont tendance à sous-estimer les valeurs très élevées de SOC, en raison de leur effet régularisateur.

Ces résultats soulignent la pertinence des estimations en surface pour la cartographie opérationnelle du SOC et la nécessité de compléter les prédictions en profondeur par des données pédologiques locales pour fiabiliser les estimations. web développée, permettant une consultation interactive et une interrogation par coordonnées GPS.

Chapitre 7

Déploiement applicatif



(a) Page d'accueil. **(b)** API.

Figure 7.1 : Interface de l'application et API associée.

7.1 Objectifs fonctionnels

Le déploiement applicatif vise à rendre les résultats du modèle accessibles et exploitables par différents utilisateurs (chercheurs, décideurs, acteurs du développement rural). Les objectifs principaux sont :

- Fournir une interface conviviale permettant de visualiser les cartes de stock de carbone organique du sol (SOC).
- Mettre à disposition une API permettant d'interroger le modèle sur de nouveaux points géolocalisés.
- Garantir la reproductibilité et la portabilité grâce à l'utilisation de conteneurs Docker.
- Assurer une intégration avec la base de données spatiale PostGIS pour stocker et gérer efficacement les données.

7.2 Architecture (Django, PostGIS, API `/predict`)

L'architecture logicielle repose sur une application web développée avec le framework **Django** :

- **Backend Django** : gestion des requêtes, communication avec la base de données et orchestration des prédictions.
- **Base de données PostGIS** : stockage des profils de sol, des covariables spatiales et des résultats de prédiction.
- **API REST /predict** : service qui prend en entrée des coordonnées géographiques et renvoie la prédiction du SOC avec un intervalle de confiance.
- **Frontend** : interface cartographique interactive (Leaflet/Mapbox) permettant de naviguer, zoomer et interroger les cartes.
- **Conteneurisation Docker** : chaque composant (Django, PostgreSQL/PostGIS, worker ML) est isolé et orchestré pour simplifier le déploiement.

7.3 Sécurité, performances et limites

Sécurité

- Authentification par jetons (JWT) pour l'accès à l'API.
- Configuration SSL/TLS pour sécuriser les échanges.
- Gestion des droits utilisateurs (lecture seule pour les décideurs, écriture pour les chercheurs).

Performances

- Mise en cache des prédictions fréquentes pour réduire la latence.
- Optimisation des requêtes spatiales PostGIS (indexation GiST et BRIN).
- Possibilité de paralléliser l'inférence via des workers ML.

Limites

- Dépendance à la connectivité Internet pour l'accès à l'application.
- Coût computationnel élevé pour l'inférence sur de grandes surfaces en haute résolution.

- Nécessité de mettre à jour régulièrement les modèles avec de nouvelles données de sols et images satellitaires.

Chapitre 8

Conclusion et perspectives

8.1 Bilan des contributions

Ce stage a permis de concevoir et d'expérimenter un pipeline complet, reproductible et opérationnel pour la prédiction du stock de carbone organique du sol (SOC) dans trois villages pilotes du Sénégal (Bary, Diohine/Sassem et Sop).

Les principales contributions sont :

- Intégration multi-sources : combinaison de données pédologiques locales (IRD), continentales (AfSP) et globales (WoSIS), enrichies par l'imagerie Sentinel-2 et Landsat-8 ainsi que des covariables topographiques et pédologiques globales.
- Évaluation comparative : mise en œuvre et comparaison de plusieurs modèles de prédiction (Random Forest, XGBoost, Gradient Boosting, HistGradientBoosting, MLP PyTorch), avec des protocoles stricts de validation spatiale (GroupKFold et Leave-Area-Out).
- Innovation méthodologique : introduction d'une prédiction hiérarchique (SOC10 $\rightarrow$ SOC30) et de modèles par site, afin d'améliorer la robustesse sur les couches profondes et capturer les spécificités locales.
- Cartographie d'incertitude : production de cartes probabilistes (écart-type, intervalles de confiance à 90
- Déploiement applicatif : développement d'une application web (Django + PostGIS) intégrant une API /predict et une interface cartographique interactive, permettant un accès pratique et reproductible aux résultats.

Ces contributions confirment la pertinence des approches d'apprentissage automatique, en particulier des réseaux de neurones entraînés localement, pour la cartographie fine et fiable du SOC en contexte sahélien.

8.2 Limitations identifiées

Plusieurs contraintes méthodologiques et structurelles ont été mises en évidence :

- Décalage temporel entre les profils pédologiques historiques (AfSP, WoSIS) et les séries satellites récentes (Sentinel-2, Landsat-8), réduisant la pertinence du transfert continental → local.
- Difficulté accrue pour la couche 10–30 cm, où la sensibilité des capteurs optiques diminue, entraînant une baisse de performance ($R^2 \approx 0.65$ contre 0.80 en surface).
- Distribution hétérogène des données : la densité limitée des profils IRD (1 800 profils concentrés sur trois villages) crée des biais spatiaux et accroît l'incertitude dans certaines zones.

Contraintes computationnelles : les modèles avancés (MLP) nécessitent des ressources GPU importantes, ce qui limite leur application à de très grandes surfaces.

8.3 Perspectives scientifiques et opérationnelles

À la lumière de ces résultats et limites, plusieurs axes d'amélioration se dégagent :

- Alignement temporel des données : planifier de nouvelles campagnes de prélèvements pédologiques synchronisées avec les séries satellitaires récentes, afin de renforcer la cohérence spatio-temporelle.
- Enrichissement des covariables : intégrer des données hyperspectrales (PRISMA, EnMAP) et radar (Sentinel-1) pour améliorer la prédiction, notamment en profondeur.
- Approches hybrides : combiner apprentissage automatique et géostatistique (krigeage résiduel, co-krigeage) pour mieux capturer la structure spatiale.
- Méthodes avancées d'incertitude : recourir à des approches bayésiennes plus robustes ou à des modèles probabilistes (Gaussian Processes) pour fournir des cartes de confiance plus fines.
- Extension territoriale : appliquer le pipeline à d'autres territoires agricoles du Sénégal, voire de l'Afrique de l'Ouest, afin de produire des cartes régionales utiles aux décideurs publics et aux agriculteurs.
- Renforcement de l'interopérabilité : maintenir et enrichir l'application web pour en faire un outil partagé entre chercheurs, institutions agricoles et acteurs du développement rural.

8.4 Conclusion générale

En définitive, ce travail démontre qu'il est possible de construire un pipeline robuste, reproductible et opérationnel pour la prédiction du carbone organique du sol en contexte sahélien. Il apporte à la fois :

- une contribution scientifique, en confirmant la supériorité des approches d'apprentissage automatique (MLP, validations spatiales strictes),
- une contribution méthodologique, avec l'introduction de nouvelles approches (prédiction hiérarchique, validation LAO, cartographie d'incertitude),
- et une contribution opérationnelle, grâce au développement d'une application web interactive.

Ce mémoire constitue ainsi une base solide pour le développement d'outils d'aide à la décision en matière de gestion durable des sols et de planification agricole, contribuant à la sécurité alimentaire et à l'adaptation au changement climatique en Afrique de l'Ouest.

Annexes

Code source

Le code source complet du pipeline, incluant l'extraction des données, la préparation, l'entraînement des modèles, la quantification de l'incertitude et le déploiement applicatif, est disponible sur GitHub : https://github.com/venire-ute/stage-rendu

Bibliographie

[1] B. Minasny et al. “Soil carbon 4 per mille”. In : *Geoderma* 292 (15 avr. 2017), p. 59-86. ISSN : 0016-7061. DOI : 10.1016/j.geoderma.2017.01.002. URL : https://www.sciencedirect.com/science/article/pii/S0016706117300095 (visité le 10/09/2025).

[2] R. Lal. “Soil carbon sequestration impacts on global climate change and food security”. In : *Science (New York, N.Y.)* 304.5677 (11 juin 2004), p. 1623-1627. ISSN : 1095-9203. DOI : 10.1126/science.1097396.

[3] European Space Agency (ESA). *Sentinel-2 User Guide*. https://sentinel.esa.int/web/sentinel/missions/sentinel-2. Mission Sentinel-2 MSI, produits L2A. 2025.

[4] United States Geological Survey (USGS). *Landsat-8 (OLI/TIRS) Data Users Handbook*. https://www.usgs.gov/landsat-missions/landsat-8. Satellite Landsat-8, OLI/TIRS. 2025.

[5] NASA JPL. *Shuttle Radar Topography Mission (SRTM) DEM*. https://www2.jpl.nasa.gov/srtm/. Modèle numérique de terrain global, résolution 30 m. 2000.

[6] Innovative Solutions for Decision Agriculture (iSDA). *iSDAsoil - African Soil Property Maps*. https://isda-africa.com/isdasoil/. Cartes pédologiques de l’Afrique à haute résolution (30 m). 2020.

[7] ISRIC - World Soil Information. *SoilGrids - global predictions of soil properties*. https://soilgrids.org/. Cartes mondiales de propriétés du sol (250 m – 1 km). 2020.

[8] N. Cressie. “Geostatistics”. In : *The American Statistician* 43.4 (1er nov. 1989), p. 197-202. ISSN : 0003-1305. DOI : 10.1080/00031305.1989.10475658. URL : https://doi.org/10.1080/00031305.1989.10475658 (visité le 11/09/2025).

[9] L. Liu, M. Ji et M. Buchroithner. “Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery”. In : *Sensors* 18.9 (sept. 2018), p. 3169. ISSN : 1424-8220. DOI : 10.3390/s18093169. URL : https://www.mdpi.com/1424-8220/18/9/3169 (visité le 10/09/2025).

[10] Z. Dong, L. Yao, Y. Bao, J. Zhang, F. Yao, L. Bai et P. Zheng. "Prediction of Soil Organic Carbon Content in Complex Vegetation Areas Based on CNN-LSTM Model". In : *Land* 13.7 (juill. 2024), p. 915. ISSN : 2073-445X. DOI : 10.3390/land13070915. URL : https://www.mdpi.com/2073-445X/13/7/915 (visité le 10/09/2025).

[11] D. R. Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In : *Ecography* 40.8 (2017), p. 913-929. ISSN : 1600-0587. DOI : 10.1111/ecog.02881. URL : https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881 (visité le 10/09/2025).

[12] K. Vaysse et P. Lagacherie. "Using quantile regression forest to estimate uncertainty of digital soil mapping products". In : *Geoderma* 291 (1er avr. 2017), p. 55-64. ISSN : 0016-7061. DOI : 10.1016/j.geoderma.2016.12.017. URL : https://www.sciencedirect.com/science/article/pii/S001670611631059X (visité le 10/09/2025).

[13] Thiam. "Cartographie du carbone organique du sol à partir des images Sentinel-2 dans le bassin du Ferlo". In : (2023).

[14] *UMMISCO : Unité Mixte Internationale de Modélisation Mathématique et Informatique des Systèmes Complexes*. Unité Mixte Internationale IRD/UCAD. IRD & Université Cheikh Anta Diop (UCAD). 2025. URL : https://ummisco.ird.fr/ (visité le 11/09/2025).

[15] *Institut de Recherche pour le Développement (IRD)*. Organisme public de recherche. IRD. 2025. URL : https://www.ird.fr/ (visité le 11/09/2025).

[16] *SLAM-B : Soil Land and Agroecosystem Management in the Ferlo Basin*. Programme de recherche sur la gestion durable des terres. Institut de Recherche pour le Développement (IRD). 2025. URL : https://www.ird.fr/ (visité le 11/09/2025).

[17] International Soil Reference and Information Centre (ISRIC) and partners. *Africa Soil Profiles Database (AfSP v1.2)*. https://www.isric.org/projects/africa-soil-profiles-database. Base de données de 18 533 profils pédologiques à l'échelle continentale. 2013.

[18] ISRIC - World Soil Information. *World Soil Information Service (WoSIS)*. https://www.isric.org/explore/wosis. Base mondiale de profils pédologiques harmonisés. 2020.

[19] W. H. Cloete, G. du Preez et G. M. Van Zijl. "How to map soil organic carbon stocks at field scale in South Africa ?" In : *Soil Advances* 3 (1er juin 2025), p. 100047. ISSN : 2950-2896. DOI : 10.1016/j.soilad.2025.100047. URL : https://www.sciencedirect.com/science/article/pii/S2950289625000156 (visité le 10/09/2025).

[20] T. Behrens, H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies et M. Goldschmitt. "Digital soil mapping using artificial neural networks". In : *Journal of Plant Nutrition and Soil Science* 168.1 (2005), p. 21-33. ISSN : 1522-2624. DOI : 10.1002/jpln.200421414. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/jpln.200421414 (visité le 10/09/2025).

[21] H. Zayani, Y. Fouad, D. Michot, Z. Kassouk, N. Baghdadi, E. Vaudour, Z. Lili-Chabaane et C. Walter. "Using Machine-Learning Algorithms to Predict Soil Organic Carbon Content from Combined Remote Sensing Imagery and Laboratory Vis-NIR Spectral Datasets". In : *Remote Sensing* 15.17 (jan. 2023), p. 4264. ISSN : 2072-4292. DOI : 10.3390/rs15174264. URL : https://www.mdpi.com/2072-4292/15/17/4264 (visité le 10/09/2025).

[22] A. B. McBratney, M. L. Mendonça Santos et B. Minasny. "On digital soil mapping". In : *Geoderma* 117.1 (1er nov. 2003), p. 3-52. ISSN : 0016-7061. DOI : 10.1016/S0016-7061(03)00223-4. URL : https://www.sciencedirect.com/science/article/pii/S0016706103002234 (visité le 10/09/2025).

[23] European Space Agency (ESA) and EU Copernicus Programme. *Copernicus DEM*. https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model. Modèle numérique de terrain global, résolution 30 m et 90 m. 2020.

[24] Institut de Recherche pour le Développement (IRD). *SLAM-B Project : Soil Land and Agroecosystem Management in the Ferlo Basin*. https://www.ird.fr/. Programme de recherche sur la gestion durable des terres. 2025.

[25] IRD and Université Cheikh Anta Diop (UCAD). *UMMISCO : Unité Mixte Internationale de Modélisation Mathématique et Informatique des Systèmes Complexes*. https://ummisco.ird.fr/. Unité Mixte Internationale IRD/UCAD. 2025.

[26] IRD. *Institut de Recherche pour le Développement (IRD)*. https://www.ird.fr/. Organisme public de recherche français. 2025.

[27] Agafonkin, Vladimir and contributors. *Leaflet : An open-source JavaScript library for mobile-friendly interactive maps*. https://leafletjs.com/. Version 1.9.x. 2025.

[28] Paszke, Adam and contributors. *PyTorch : An Open Source Machine Learning Framework*. Version 2.x. PyTorch Foundation. 2025. URL : https://pytorch.org/ (visité le 11/09/2025).

[29] Django Software Foundation. *Django Web Framework*. Version 5.x. 2025. URL : https://www.djangoproject.com/ (visité le 11/09/2025).

[30] PostGIS Project. *PostGIS : Spatial and Geographic Objects for PostgreSQL*. Version 3.x. OSGeo. 2025. URL : https://postgis.net/ (visité le 11/09/2025).

[31] Docker, Inc. *Docker : Empowering App Development for Developers*. Version 25.x. 2025. URL : https://www.docker.com/ (visité le 11/09/2025).

[32] Agafonkin, Vladimir and contributors. *Leaflet : An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps*. Version 1.9.x. 2025. URL : https://leafletjs.com/ (visité le 11/09/2025).