

Cartographie numérique de la fertilité des sols du bassin arachidier sénégalais par modélisation spatiale de l'azote, du phosphore et du potassium : une étude exploratoire combinant krigeage de régression et apprentissage automatique

François Ngor Sene

Yancouba Sane

Université Assane SECK de Ziguinchor, Département de Géographie,
Laboratoire de Géomatique et d'Environnement (LGE), Ziguinchor, Sénégal

Alexandre Badiane

Université Assane SECK de Ziguinchor, Département de Géographie,
Laboratoire de Géomatique et d'Environnement (LGE), Ziguinchor, Sénégal
Centre de coopération internationale en recherche agronomique pour le
développement (CIRAD)

Aïdara Chérif Amadou Lamine Fall

Université Assane SECK de Ziguinchor, Département de Géographie,
Laboratoire de Géomatique et d'Environnement (LGE), Ziguinchor, Sénégal

Modou Sene

Bureau d'études et de recherches sur la gestion des ressources naturelles
pour le développement rural (BERGRN-DR), Thiès, Sénégal

[Doi:10.19044/esj.2026.v22n21p182](https://doi.org/10.19044/esj.2026.v22n21p182)

Submitted: 22 May 2026

Accepted: 09 July 2026

Published: 31 July 2026

Copyright 2026 Author(s)

Under Creative Commons CC-BY 4.0

OPEN ACCESS

Cite As:

Sene, F.N., Sane, Y., Badiane, A., Fall, A.C.A.L., & Sene, M. (2026). *Cartographie numérique de la fertilité des sols du bassin arachidier sénégalais par modélisation spatiale de l'azote, du phosphore et du potassium : une étude exploratoire combinant krigeage de régression et apprentissage automatique*. European Scientific Journal, ESJ, 22 (21), 182.

<https://doi.org/10.19044/esj.2026.v22n21p182>

Résumé

La fertilité des sols semi-arides d'Afrique de l'Ouest, en particulier du bassin arachidier sénégalais, est soumise à une dégradation progressive liée aux pressions anthropiques et climatiques. En l'absence des cartes pédologiques actualisées à haute résolution, la gestion raisonnée des intrants

agricoles demeure difficile à l'échelle parcellaire. Cette étude vise à prédire et spatialiser les teneurs en azote total (N), phosphore disponible (P) et potassium échangeable (K) sur 148 433 polygones du bassin arachidier du Sénégal (Observatoire de Niakhar). Elle propose une approche de cartographie numérique des sols (CNS) combinant neuf algorithmes d'apprentissage automatique. Les données d'entraînement proviennent de 10 classes de sols prédictives en fonction des covariables SCORPAN (Végétation, topographie, climat et géologie), où 100 échantillons ont été prélevés, soit 10 échantillons par classe de sol. Ces échantillons ont été regroupés en 10 échantillons composites soit un échantillon composite par classe de sol, ce qui a donné un effectif réel pour la modélisation : $n = 10$). La validation est réalisée par validation croisée leave-one-out (LOO-CV) inter-clusters. Les résultats montrent que MARS obtient les meilleures performances pour N ($R^2 = 0,457$; $RMSE = 0,0187\%$; $RPD = 1,43$) et que la MLR est la plus robuste pour K ($R^2 = 0,386$; $RPD = 1,35$). La prédiction du phosphore s'avère particulièrement délicate avec $R^2 < 0$ pour tous les modèles en LOO-CV, $CV = 150\%$. Le krigeage de régression produit des cartes choroplèthes couvrant l'intégralité du domaine. Les valeurs médianes prédites sont $N = 0,054\%$; $P = 5,17\text{ ppm}$; $K = 0,085\text{ meq/100 g}$. Le cluster 9 se distingue comme un hotspot de phosphore ($P\text{ moyen} = 31\text{ ppm}$). Ces résultats constituent une première base cartographique exploratoire. En raison du faible effectif d'entraînement ($n = 10$) et de l'absence de jeu de validation indépendant, les cartes ne peuvent pas encore être utilisées pour des recommandations de fertilisation précises à l'échelle parcellaire.

Mots-clés: Sols, Cartographie numérique des sols, Modélisation, Fertilité des sols, Sénégal

Digital Soil Mapping of Soil Fertility in the Senegalese Groundnut Basin Through Spatial Modeling of Nitrogen, Phosphorus, and Potassium: An Exploratory Study Combining Regression Kriging and Machine Learning

François Ngor Sene

Yancouba Sane

Assane Seck University of Ziguinchor, Department of Geography,
Laboratory of Geomatics and the Environment (LGE), Ziguinchor, Senegal

Alexandre Badiane

Assane Seck University of Ziguinchor, Department of Geography,
Laboratory of Geomatics and the Environment (LGE), Ziguinchor, Senegal
Centre for International Cooperation in Agricultural Research for
Development (CIRAD)

Aïdara Chérif Amadou Lamine Fall

Assane Seck University of Ziguinchor, Department of Geography,
Laboratory of Geomatics and the Environment (LGE), Ziguinchor, Senegal

Modou Sene

Research and Planning Office for Natural Resource Management for Rural
Development (BERGRN-DR), Thiès, Senegal

Abstract

The fertility of semi-arid soils in West Africa, particularly in the Senegalese groundnut basin, is subject to progressive degradation linked to anthropogenic and climatic pressures. In the absence of updated high-resolution pedological maps, sound management of agricultural inputs remains difficult at the field scale. This study aims to predict and spatialize total nitrogen (N), available phosphorus (P), and exchangeable potassium (K) content across 148,433 polygons of Senegal's groundnut basin (Niakhar Observatory). It proposes a digital soil mapping (DSM) approach combining nine machine learning algorithms. Training data come from 10 predictive soil classes defined according to SCORPAN covariates (Vegetation, topography, climate, and geology), for which 100 samples were collected, i.e., 10 samples per soil class. These samples were pooled into 10 composite samples, one composite sample per soil class, resulting in an actual modeling sample size of $n = 10$. Validation was carried out through inter-cluster leave-one-out cross-validation (LOO-CV). Results show that MARS achieved the best performance for N ($R^2 = 0.457$; RMSE = 0.0187%; RPD = 1.43), while MLR was the most robust for K ($R^2 = 0.386$; RPD = 1.35). Phosphorus prediction proved particularly challenging, with $R^2 < 0$ for all models under LOO-CV

(CV = 150%). Regression kriging produced choropleth maps covering the entire study domain. Predicted median values were N = 0.054%; P = 5.17 ppm; K = 0.085 meq/100 g. Cluster 9 stood out as a phosphorus hotspot (mean P = 31 ppm). These results constitute a first exploratory cartographic baseline. Given the small training sample size (n = 10) and the absence of an independent validation dataset, the maps cannot yet be used for precise fertilization recommendations at the field scale.

Keywords: Soils, Digital Soil Mapping, Modeling, Soil Fertility, Senegal

Introduction

Les sols des zones ouest-africaines constituent un capital agronomique de première importance pour les populations rurales dont l'alimentation et les revenus reposent en grande partie sur une agriculture pluviale de subsistance (arachide, mil, sorgho) (Banque mondiale, 2025). En zone sahélienne sénégalaise, la combinaison d'une faible pluviométrie (450-800 mm/an), de sols ferrugineux tropicaux sableux à faible capacité de rétention hydrique et d'une pression démographique croissante conduit à une minéralisation accélérée de la matière organique et à un appauvrissement progressif des éléments nutritifs majeurs (azote (N), phosphore (P) et potassium (K)) qui conditionnent directement les rendements agricoles (Bationo et al., 2004 ; Tiftonell & Giller, 2013).

La cartographie de la fertilité des sols à haute résolution spatiale représente un prérequis indispensable à toute démarche d'agriculture de précision. Cependant, les approches conventionnelles fondées sur des sondages systématiques et des analyses de laboratoire sont coûteuses, chronophages et difficiles à généraliser à de larges étendues spatiales, en particulier dans les contextes à ressources limitées d'Afrique subsaharienne (Mulder et al., 2011 ; Hengl et al., 2014). Des initiatives continentales telles que l'Africa Soil Information Service (AfSIS) ont démontré le potentiel des approches par télédétection pour cartographier les propriétés et la dégradation des sols à large échelle en Afrique (Vågen et al., 2013). La cartographie numérique des sols (CNS) offre une alternative prometteuse en combinant un nombre réduit de données de terrain avec des covariables environnementales

Depuis les travaux fondateurs de Jenny (1941) et le cadre conceptuel SCORPAN (McBratney et al., 2003), la CNS a connu un essor considérable grâce aux progrès de l'apprentissage automatique, à la disponibilité de données satellitaires à haute résolution et à la puissance de calcul croissante. De nombreuses études ont démontré la pertinence des algorithmes tels que la forêt aléatoire (RF), le gradient boosting, le krigeage et leurs combinaisons pour la prédiction des propriétés du sol à différentes échelles (Grimm et al., 2008 ; Keskin & Grunwald, 2018 ; Wadoux et al., 2020). En particulier, le

krigeage de régression (KR) a démontré sa capacité à améliorer la précision cartographique en capturant à la fois la relation sol-environnement et la dépendance spatiale résiduelle (Hengl et al., 2007 ; Behrens et al., 2010).

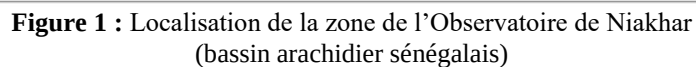
La zone du bassin arachidier sénégalais, caractérisée par une mosaïque complexe de sols ferrugineux tropicaux lessivés (Lixisols), de sols peu évolués sableux (Arénosols) et de sols hydromorphes dans les bas-fonds, présente une variabilité spatiale des propriétés chimiques particulièrement élevée et peu documentée à l'échelle fine. La dynamique des nitrates, la mobilité du phosphore dans les milieux acides et la lixiviation du potassium sous les effets conjugués de l'acidification et des pratiques culturales intensives confèrent à ces environnements un caractère pédologique singulier qui mérite une analyse dédiée.

La présente étude poursuit trois objectifs. Premièrement, comparer les performances de neuf modèles d'apprentissage automatique pour prédire N, P et K. Deuxièmement, appliquer le krigeage de régression pour cartographier ces nutriments sur les 148 433 polygones du bassin arachidier (Observatoire de Niakhar). Troisièmement, analyser la distribution spatiale des nutriments par cluster pédologique afin d'identifier les zones à statut nutritif critique

Matériel et Méthodes

Localisation et contexte pédoclimatique

La zone d'étude (Figure 1) est localisée dans le bassin arachidier du Sénégal, regroupant les régions administratives de Diourbel, Fatick, Kaolack et Kaffrine, entre 13,5° N et 15,0° N en latitude, et 14,5° W et 17,5° W en longitude. Elle couvre une superficie d'environ 51 315 km² et englobe un gradient pédologique allant des sols ferrugineux tropicaux peu ou non lessivés aux sols ferrugineux tropicaux lessivés, ainsi que des sols hydromorphes des dépressions (bas-fonds) et de certains profils halomorphes. Le climat est de type sahélien au Nord et soudano-sahélien vers le Sud, marqué par une saison sèche longue (octobre-juin) et une saison des pluies concentrée sur quatre mois (juillet-octobre). La pluviométrie annuelle est comprise entre 500 et 800 mm sur les 30 dernières années (gradient nord-sud marqué). Le substrat géologique, principalement constitué de grès continentaux du Continental Terminal (Mio-Pliocène), de sables quaternaires et d'affleurements de latérites, conditionne fortement la distribution spatiale des propriétés des sols.



Les données pédologiques proviennent de 10 classes de sol prédites, définies par une classification non supervisée k-means appliquée aux covariables SCORPAN standardisées ; le nombre de classes ($k = 10$) a été fixé a priori en fonction du budget d'échantillonnage disponible (100 sites, soit 10 sites par classe), et non selon un critère statistique d'optimalité tel que la méthode du coude ou l'indice de silhouette, où 100 sites d'échantillons de sols ont été choisis à travers un échantillonnage hypercube Latin basé sur les covariables SCORPAN et stratifié en fonction des différentes classes de sols prédites. Après prélèvement, les échantillons ont été regroupés en échantillon composite (10 échantillons composites) soit un échantillon composite par classe de sols. Le choix de faire des échantillons composites est lié au coût exorbitant de l'analyse au laboratoire des échantillons de sols, et pour pallier cela, nous avons procédé par agrégation en échantillon composite même si elle masque la variabilité intra-classe des propriétés pédologiques. Ce choix méthodologique, bien que limitant la puissance statistique des modèles (effectif réel $n = 10$), demeure courant dans les études de cartographie numérique des sols menées en contexte de ressources limitées, où le coût d'analyse en laboratoire impose un compromis entre couverture spatiale et résolution intra-classe (Mulder et al., 2011). Il est important de souligner que

l'effectif réel utilisé comme unités d'entraînement pour les neuf algorithmes est donc $n=10$, fondamentale pour interprétation des performances des modèles et la portée statistique des conclusions. Chaque classe représente une unité de sol homogène, caractérisée par ses valeurs moyennes de nutriments et de propriétés physico-chimiques. Les analyses de laboratoire ont porté sur : le N total par la méthode Kjeldahl, le P disponible par la méthode Bray-1, le K échangeable par extraction à l'acétate d'ammonium (pH 7), le pH à l'eau (rapport sol/eau = 1/2,5), la matière organique (MO) par la méthode de Walkley-Black, la capacité d'échange cationique (CEC) par la méthode à l'acétate d'ammonium. Les analyses ont été réalisées à l'Institut National de la Pédologie (INP) du Sénégal selon les normes ISO en vigueur.

Covariables spectrales et topographiques

Quinze covariables environnementales ont été dérivées d'images satellitaires Sentinel-2 (résolution 10 m, période 2024) composite médian annuel construit à partir de l'ensemble des scènes disponibles sur l'année 2024, ce qui lisse le bruit atmosphérique mais ne capture pas la variabilité phénologique intra-annuelle et d'un MNT SRTM à 30 m : NDVI, SAVI, CCI, RCI, NDWI, indice de quartz (QuartzIndex), indice de fer (FeIndex), indice de silice (SilicaIndex), absorption des argiles (ClayAbsorption), indice de couleur du sol (SoilColorIndex), NDOMI, NDSOCI, humidité du sol (SoilMoisture), pente (Slope) et altitude (Altitude). Ces covariables ont été extraites pour les 148 433 polygones sur la zone d'étude, permettant la prédiction exhaustive sur tout le domaine d'étude. Le calcul des centroïdes de chaque polygone a défini la grille de krigeage, projetée en UTM zone 28N (EPSG : 32628).

Sélection des covariables par le facteur d'inflation de la variance (VIF)

La multicollinéarité entre les covariables spectrales constitue un problème récurrent en CNS. L'analyse du facteur d'inflation de la variance (VIF) a été utilisée pour identifier et éliminer les covariables redondantes. Le VIF d'une covariable v est défini par $VIF_v = \frac{1}{1 - R^2_v}$, où R^2_v est le coefficient de détermination de la régression de v sur toutes les autres covariables. Un seuil de $VIF > 10$ a été retenu comme critère d'exclusion (Kutner et al., 2004 ; Dormann et al., 2013). Cette procédure itérative garantit une matrice de covariables orthogonale pour la modélisation.

Modèles d'apprentissage automatique

La prédiction des teneurs en azote total (N), phosphore disponible (P) et potassium échangeable (K) a été réalisée à l'aide de neuf algorithmes de régression couramment utilisés en cartographie numérique des sols. Ces

modèles ont été choisis afin de comparer différentes approches de modélisation, allant des méthodes linéaires classiques aux algorithmes capables de représenter des relations non linéaires complexes entre les propriétés du sol et les covariables environnementales. Avant la modélisation, les covariables retenues après l'analyse de multicollinéarité ($VIF < 10$) ont été standardisées afin de rendre comparables les variables exprimées dans des unités différentes. Les variables cibles (N, P et K) ont ensuite été associées aux valeurs des covariables extraites pour chacun des dix clusters pédologiques utilisés comme unités d'apprentissage.

Les modèles évalués comprennent : (1) la régression linéaire multiple (MLR), utilisée comme modèle de référence ; (2 et 3) les régressions Ridge et Lasso, permettant de limiter les effets de surajustement par pénalisation des coefficients ; (4) la régression par moindres carrés partiels (PLSR), adaptée aux situations de corrélation entre variables explicatives ; (5) le modèle MARS (Multivariate Adaptive Regression Splines), capable de représenter des relations non linéaires par fonctions spline ; (6) la forêt aléatoire (Random Forest, RF), fondée sur l'agrégation d'arbres de décision ; (7) Cubist, qui combine arbres de règles et régressions linéaires locales ; (8) XGBoost, basé sur l'apprentissage séquentiel d'arbres optimisés par gradient boosting ; et (9) les machines à vecteurs de support (SVM) utilisant un noyau radial pour modéliser les relations non linéaires. Compte tenu du très faible effectif ($n = 10$), toute procédure de calibration interne des hyperparamètres (recherche par grille, validation croisée imbriquée) aurait été statistiquement fragile ; les hyperparamètres de RF, XGBoost, SVM, Cubist et MARS ont donc été fixés aux valeurs par défaut de leurs packages R respectifs plutôt qu'optimisés, ce qui constitue une limite de cette étude exploratoire.

Compte tenu du faible effectif disponible ($n = 10$ clusters), une stratégie de validation croisée Leave-One-Out (LOO-CV) a été mise en œuvre. À chaque itération, un cluster était retiré de l'ensemble d'apprentissage et utilisé pour la validation, tandis que les neuf clusters restants servaient à l'entraînement du modèle. Cette procédure a été répétée jusqu'à ce que chaque cluster ait été utilisé une fois comme observation de validation. Cette approche permet de maximiser l'utilisation de l'information disponible tout en fournissant une estimation indépendante de la capacité prédictive des modèles. Les performances des différents algorithmes ont été évaluées à partir de plusieurs indicateurs statistiques complémentaires : le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE), le biais moyen (Bias), le ratio de performance à la déviation (RPD), le ratio de performance à l'intervalle interquartile (RPIQ), le coefficient de concordance de Lin (CCC) et l'indice d'accord de Willmott (d). Ces indicateurs permettent d'évaluer simultanément la précision, la robustesse et la capacité de généralisation des modèles.

Krigeage de régression

Le krigeage de régression (KR) combine une tendance déterministe $m(x)$ issue du modèle RF et une composante stochastique $\varepsilon^*(x)$ estimée par krigeage ordinaire des résidus (Hengl et al., 2007) : $\hat{Z}(x) = \hat{m}(x) + \hat{\varepsilon}^*(x)$. La tendance RF est entraînée sur l'ensemble des 10 classes pédologiques. Les résidus aux sites d'entraînement sont calculés par différence entre valeurs observées et prédictions RF. Le variogramme expérimental des résidus est ajusté automatiquement par `automap::autofitVariogram` (sélection du modèle théorique optimal, ajustement par moindres carrés pondérés). Le krigeage est réalisé vers les 148 433 centroïdes par `gstat::krige`. La surface prédite finale et l'écart-type de krigeage associé $\sigma_k(x)$ sont calculés pour chaque polygone. La qualité du variogramme est diagnostiquée par le rapport pépite/palier (nugget effect).

Outils logiciels

L'ensemble des analyses a été conduit sous R version 4.3 (R Core Team, 2024). Les packages principaux sont : `sf`, `sp` (manipulation spatiale), `ranger`, `xgboost`, `e1071`, `Cubist`, `pls`, `glmnet`, `earth` (apprentissage automatique), `gstat`, `automap` (géostatistique), `ggplot2`, `ggspatial`, `patchwork` (visualisation), `openxlsx` (export Excel). Le pipeline complet est implémenté dans un script unique garantissant la reproductibilité intégrale des résultats.

Résultats

Statistiques descriptives des propriétés du sol

Le Tableau 1 présente les statistiques descriptives des principales propriétés chimiques du sol et des covariables spectrales. Les teneurs en azote total sont faibles (N moyen = $0,062 \pm 0,027$ %), avec un CV de 43,2 % et une asymétrie positive ($\gamma = 1,58$) indiquant une distribution dominée par des valeurs basses avec quelques classes pédologiques à teneur élevée. Ces niveaux sont caractéristiques des sols sahéliens appauvris en matière organique, cohérents avec les valeurs rapportées par Feller et Beare (1997) pour les sols sableux d'Afrique de l'Ouest (N < 0,10 %).

Le phosphore présente la plus forte variabilité de l'ensemble des variables analysées (P moyen = $8,34 \pm 12,52$ ppm, CV = 150,1 %, $\gamma = 2,60$). La médiane (4,74 ppm) est très inférieure à la moyenne, témoignant d'une distribution très asymétrique dominée par quelques valeurs extrêmes. Le seuil de carence en P selon les normes FAO étant de 10 ppm (Bray-1), la majorité des classes pédologiques est en situation de déficit phosphorique. Le potassium échangeable montre une variabilité intermédiaire (K moyen = $0,086 \pm 0,054$ meq/100 g, CV = 62,2 %). Ces niveaux sont généralement considérés comme faibles à très faibles (seuil de carence < 0,12 meq/100 g), suggérant un

risque de carence potassique généralisé sur les sols sableux à faible CEC (CEC moyenne = 6,18 meq/100 g).

Tableau 1 : Statistiques descriptives des propriétés chimiques du sol et des covariables spectrales (n = 10 clusters pédologiques)

Variable	n	Moy	ET	Min	Méd	Max	CV (%)	γ
N total (%)	10	0,062	0,027	0,035	0,054	0,129	43,2	1,58
P disponible (ppm)	10	8,34	12,52	1,81	4,74	43,73	150,1	2,60
K échangeable (meq/100 g)	10	0,086	0,054	0,016	0,088	0,208	62,2	0,96
pH (eau)	10	6,02	0,75	5,20	5,75	7,60	12,4	0,90
MO (%)	10	1,013	0,537	0,469	0,860	2,345	52,9	1,58
CEC (meq/100 g)	10	6,18	0,68	5,10	5,95	7,30	11,0	0,17
NDVI	10	0,143	0,031	0,071	0,142	0,183	22,0	-0,96

ET : écart-type ; CV : coefficient de variation ; γ : coefficient d'asymétrie

Sélection des covariables par le VIF

L'analyse VIF a identifié plusieurs groupes de covariables fortement colinéaires : NDVI/SAVI ($r > 0,97$), CCI/RCI ($r > 0,88$) et QuartzIndex/SilicaIndex ($r > 0,85$). Après procédure itérative d'élimination, un sous-ensemble de covariables non corrélées ($VIF < 10$) a été retenu. L'ACP sur les covariables retenues (Figure 2) révèle deux axes principaux : un gradient végétation-humidité (axe 1 : NDVI, SAVI, NDMI, corrélé positivement avec N et MO) et un axe lithologie-texture (axe 2 : QuartzIndex, FeIndex, ClayAbsorption, lié à P et K). Les clusters pédologiques se distribuent de manière cohérente dans l'espace factoriel, validant la pertinence écologique des covariables sélectionnées. La matrice de corrélation de Pearson (Figure 3) révèle une association significative entre N et NDVI ($r = 0,56$; $p < 0,05$), cohérente avec le couplage N-MO-végétation bien documenté dans les sols tropicaux. Le phosphore montre une corrélation positive avec QuartzIndex, indiquant un contrôle lithologique. Le potassium est corrélé positivement avec ClayAbsorption ($r = 0,61$) et négativement avec Slope ($r = -0,53$).

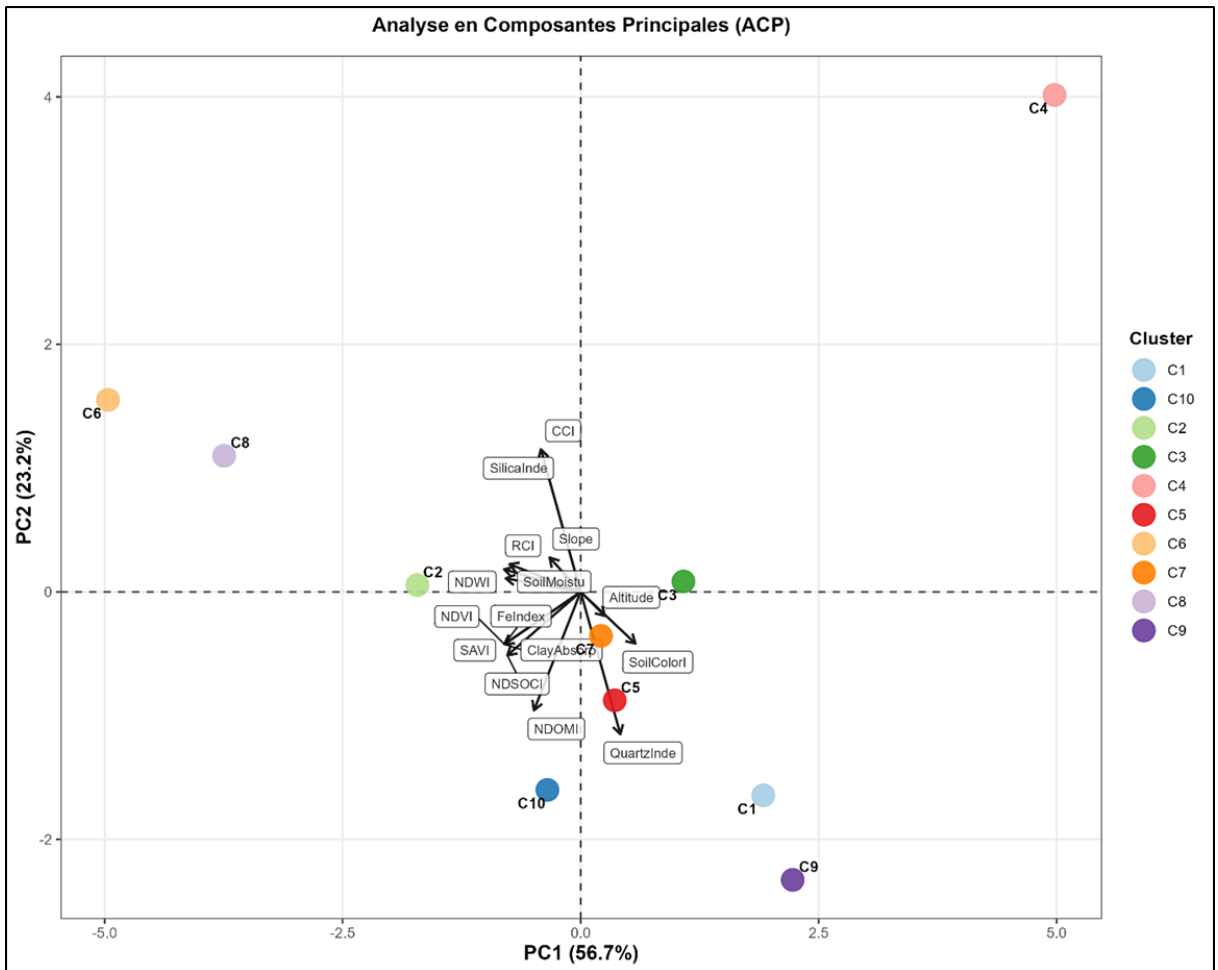


Figure 2 : Analyse en Composantes Principales (ACP) des covariables environnementales et des propriétés chimiques du sol

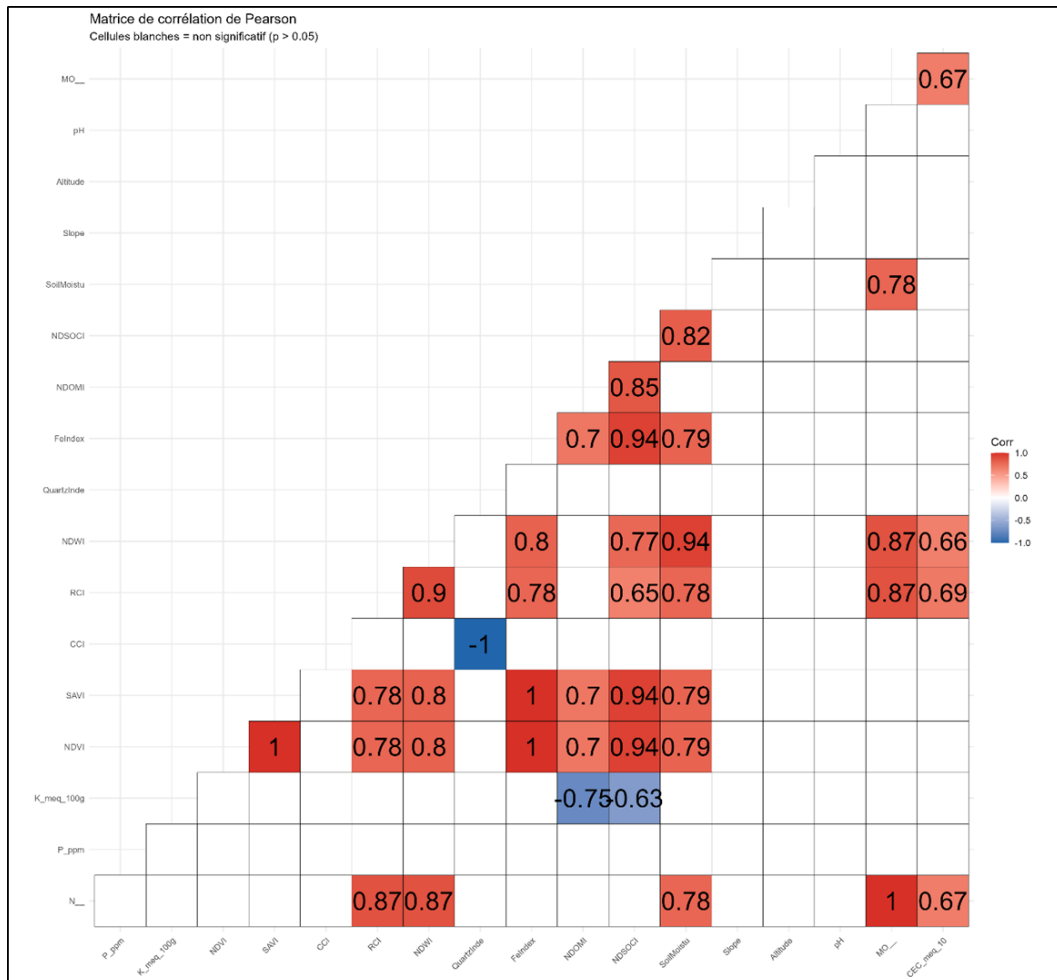


Figure 3 : Matrice de corrélation de Pearson entre les propriétés chimiques du sol et les covariables retenues

Performances des modèles d'apprentissage automatique

Le Tableau 2 synthétise les performances LOO-CV des neuf modèles pour N et K. Pour l'azote, MARS obtient les meilleures performances ($R^2 = 0,457$; $RMSE = 0,0187\%$; $RPD = 1,43$; $CCC = 0,667$), suivi de Ridge ($R^2 = 0,434$; $RPD = 1,40$) et Lasso ($R^2 = 0,416$). La MLR montre des performances très dégradées ($R^2 = -1,289$), confirmant le surajustement attendu avec $n = 10$. Les algorithmes ensemblistes RF ($R^2 = 0,166$), Cubist ($R^2 = 0,130$) et SVM ($R^2 = 0,133$) présentent des performances inférieures aux modèles linéaires régularisés, ce qui s'explique par le faible effectif ($n = 10$) défavorable aux méthodes nécessitant un grand nombre de données pour explorer les interactions non linéaires.

Pour le potassium, la MLR obtient le meilleur R^2 (0,386) malgré son instabilité apparente sur N, suivie de Ridge ($R^2 = 0,357$) et XGBoost ($R^2 =$

0,283). MARS produit un R^2 négatif pour K (-0,107), illustrant sa sensibilité aux valeurs atypiques avec peu d'observations. Ces résultats soulignent l'absence d'un modèle universellement supérieur et la nécessité d'une sélection par nutriment. La prédiction du phosphore constitue le défi majeur : tous les modèles produisent des R^2 négatifs en LOO-CV, le meilleur étant Cubist ($R^2 = -0,121$; RMSE = 12,57 ppm). Cette situation reflète la non-linéarité des processus de rétention du phosphore et l'insuffisance de 10 observations.

Tableau 2 : Performances LOO-CV des neuf modèles pour N total et K échangeable. Pour P, tous les modèles présentent un $R^2 < 0$ (voir texte). RPD > 1,4: performance satisfaisante pour la cartographie

Modèle	R^2 (N)	RMSE (N)	RPD (N)	CCC (N)	R^2 (K)	RMSE (K)	RPD (K)
MLR	-1,289	0,0385	0,697	-0,06	0,386	0,040	1,345
Ridge	0,434	0,0191	1,401	0,640	0,357	0,041	1,315
Lasso	0,416	0,0195	1,379	0,635	0,235	0,045	1,205
PLSR	0,309	0,0211	1,268	0,665	0,298	0,043	1,258
MARS	0,457	0,0187	1,431	0,667	-0,107	0,054	1,002
RF	0,166	0,0232	1,155	0,349	0,148	0,047	1,142
Cubist	0,130	0,0237	1,130	0,469	-0,350	0,059	0,907
XGBoost	0,346	0,0206	1,304	0,498	0,283	0,043	1,245
SVM	0,133	0,0237	1,132	0,262	0,121	0,048	1,124
P (tous modèles)	< 0	-	< 1,0	< 0	-	-	-

L'analyse de l'importance des variables par permutation dans RF (Figure 4) confirme des hiérarchies cohérentes avec les corrélations bivariées : pour N, RCI et NDWI dominant, en cohérence avec l'influence de l'humidité et de la végétation sur le cycle de l'azote ; pour K, RCI, NDVI et SAVI sont prédominants. Les valeurs SHAP (Figure 5) confirment ces tendances et révèlent des effets non linéaires de l'indice de fer et des indices de végétation (RCI, SAVI, NDVI) sur la distribution de K, ainsi que l'influence de la géologie (indice de silice, indice de carbonate de calcium, indice de quartz) sur la distribution de P. Les graphes de dépendance partielle (PDP) montrent une relation positive monotone entre le NDVI et N (Figure 6) et une relation en cloche entre les indices de végétation (NDVI et SAVI) et K (Figure 8). Pour le phosphore (Figure 7), la relation avec les indices géologiques est monotone lorsque la teneur en argile est élevée et devient non linéaire dans les sols sableux à forte proportion de sable. Ces analyses (importance RF, SHAP, PDP) étant calculées sur un modèle entraîné sur $n = 10$ observations, leurs interprétations sont considérées comme exploratoires et indicatives des tendances écologiques plausibles ; elles ne peuvent pas être considérées comme statistiquement robustes.

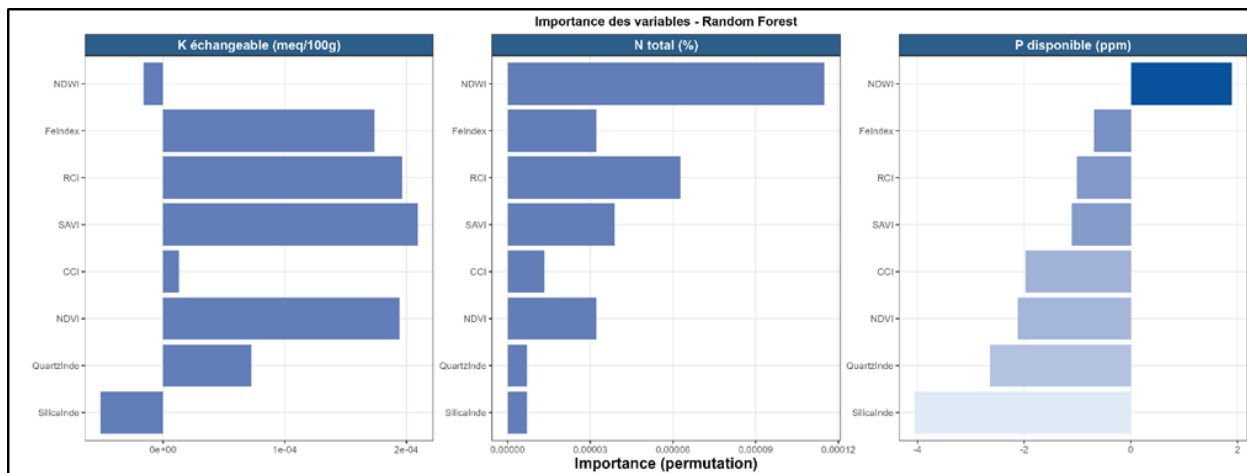


Figure 4 : Importance des variables dans la prédiction de N, P et K (Random Forest, importance par permutation)

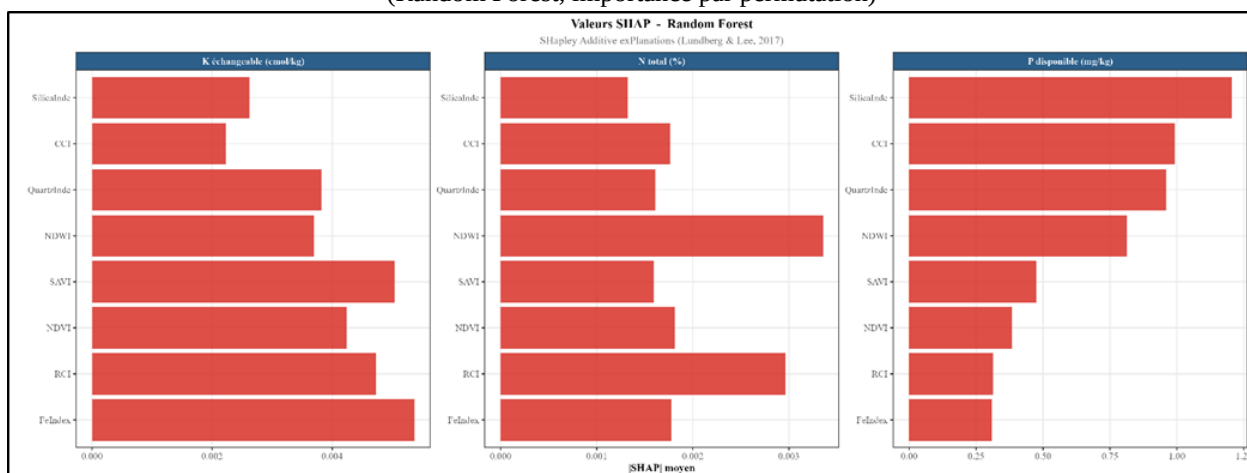


Figure 5 : Valeurs SHAP des variables dans la prédiction de N, P et K (Random Forest)

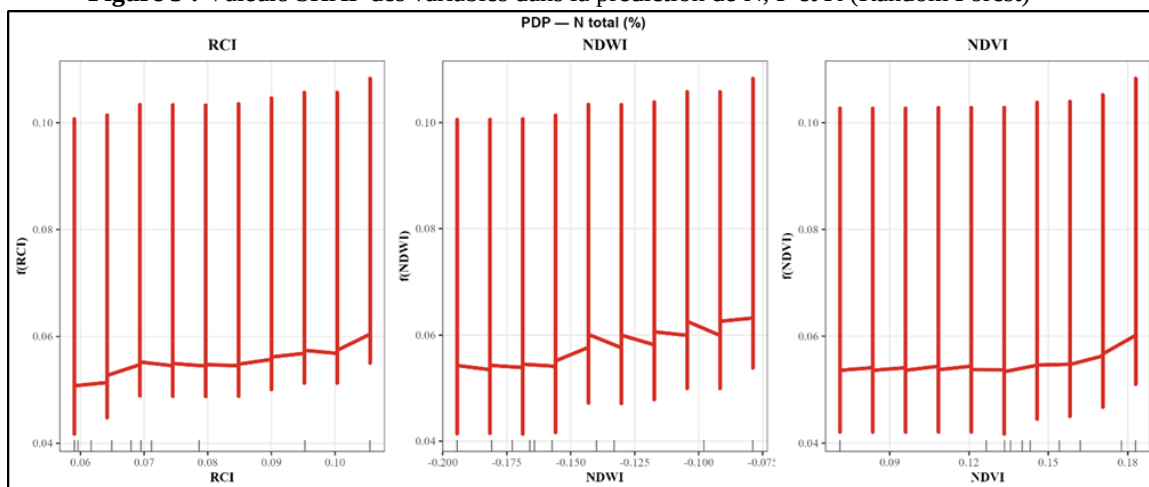


Figure 6 : Graphes de dépendance partielle (PDP) de l'azote total N (%)

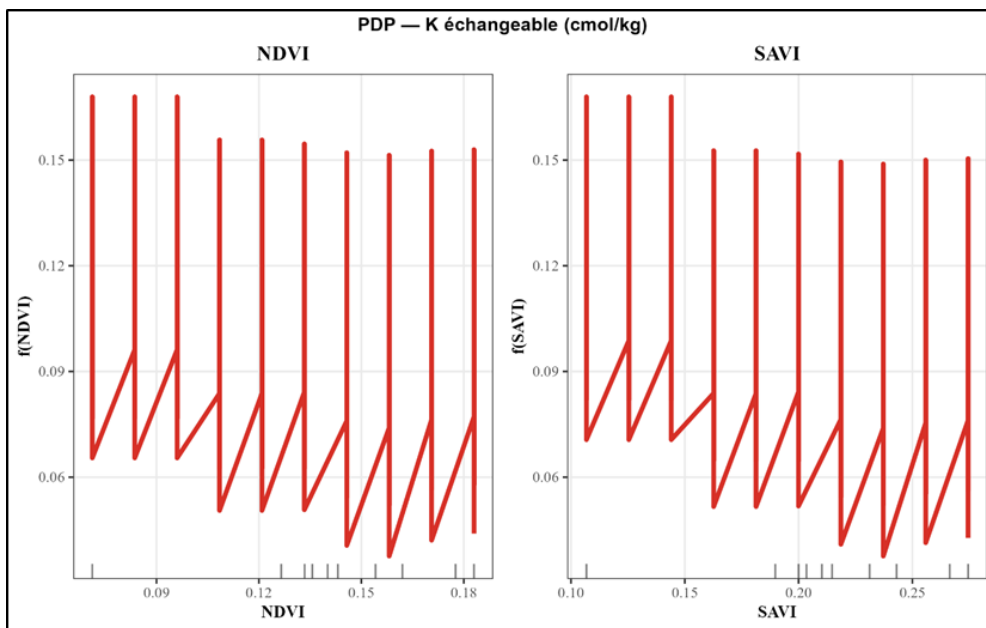


Figure 7 : Graphes de dépendance partielle (PDP) du phosphore disponible P (ppm)

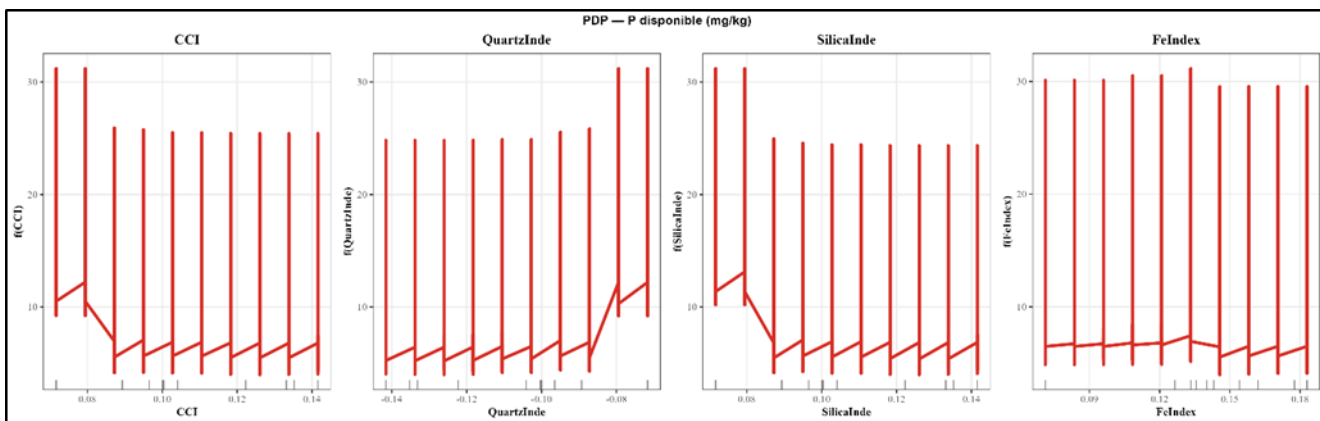


Figure 8 : Graphes de dépendance partielle (PDP) du potassium échangeable K (meq/100 g)

Cartographie par krigeage de régression

Tendance RF et variogrammes des résidus

La composante de tendance RF explique 94,1 % de la variance observée pour N ($R^2_{\text{tendance}} = 0,941$). Ces R^2 de tendance (0,941 pour N ; 0,968 pour P ; 0,996 pour K, Tableau 3) sont des métriques in-sample: ils sont calculés en réappliquant le modèle RF sur ses propres données d'entraînement ($n = 10$ clusters). Ils témoignent du bon ajustement du modèle RF pour la composante d'interpolation du krigeage de régression, mais ne constituent pas une mesure de la capacité prédictive, laquelle est uniquement évaluée par le LOO-CV (Tableau 2). Ces valeurs élevées de R^2 in-sample reflètent le

surajustement attendu d'un RF sur $n = 10$ points. Les résidus de la tendance RF présentent des écarts-types faibles ($\sigma_{\text{résidus}} \approx 0,009\%$ pour N ; $\approx 0,021$ meq/100 g pour K) et sont spatialement structurés selon des variogrammes ajustés par des modèles sphériques avec des portées de 1 000 à 2 000 m, indiquant une autocorrélation résiduelle à l'échelle hectométrique (Tableau 3). Le rapport nugget/palier ($C_0/(C_0+C)$) varie entre 0,15 et 0,35 selon le nutriment, confirmant une structuration spatiale significative des résidus qui justifie l'utilisation du krigeage plutôt qu'une simple prédiction RF.

Tableau 3 : Performances des modèles de prédiction spatiale des propriétés du sol (tendance RF, variogramme des résidus et krigeage de régression)

Propriété	Tendance RF			Variogramme			Krigage RK		
	Modèle	R ²	$\sigma_{\text{résidus}}$	Modèle	Psill	Range (m)	Intervalle	σ_k moy	n polygones
N total (%)	RF	0,941	0,0087	Sphérique	0	1858	[0,041; 0,111]	0,0098	148 433
P disponible (ppm)	RF	0,968	5,0473	Exponentiel	9,5895	1858	[-0,27; 41,62]	3,1008	148 433
K échangeable (meq/100 g)	RF	0,996	0,0173	Exponentiel	0	1858	[0,036; 0,171]	0,0209	148 433

Note : les R² de la colonne « Tendance RF » sont des métriques in-sample sur $n = 10$ et ne reflètent pas la capacité prédictive (LOO-CV, Tableau 2)

Distribution spatiale de l'azote total

Les valeurs prédites de N total (Tableau 4) couvrent une plage de 0,041 à 0,111 % sur les 148 433 polygones, avec une valeur médiane de 0,054 % et un écart-type de 0,014 % (CV = 16,8 %). Ces niveaux confirment la pauvreté générale en azote des sols de la zone, caractéristique des systèmes sableux sahéliens. La Figure 9 révèle un gradient spatial marqué : les teneurs les plus élevées ($N > 0,10\%$) sont concentrées dans le cluster 6 (N moyen RK = 0,109 %), correspondant aux zones de bas-fonds et de dépressions à sol plus riche en matière organique. Les teneurs les plus faibles ($N < 0,045\%$) s'observent dans le cluster 7 (N moyen = 0,042 %), associé aux sols Dior sableux dégradés sous culture continue.

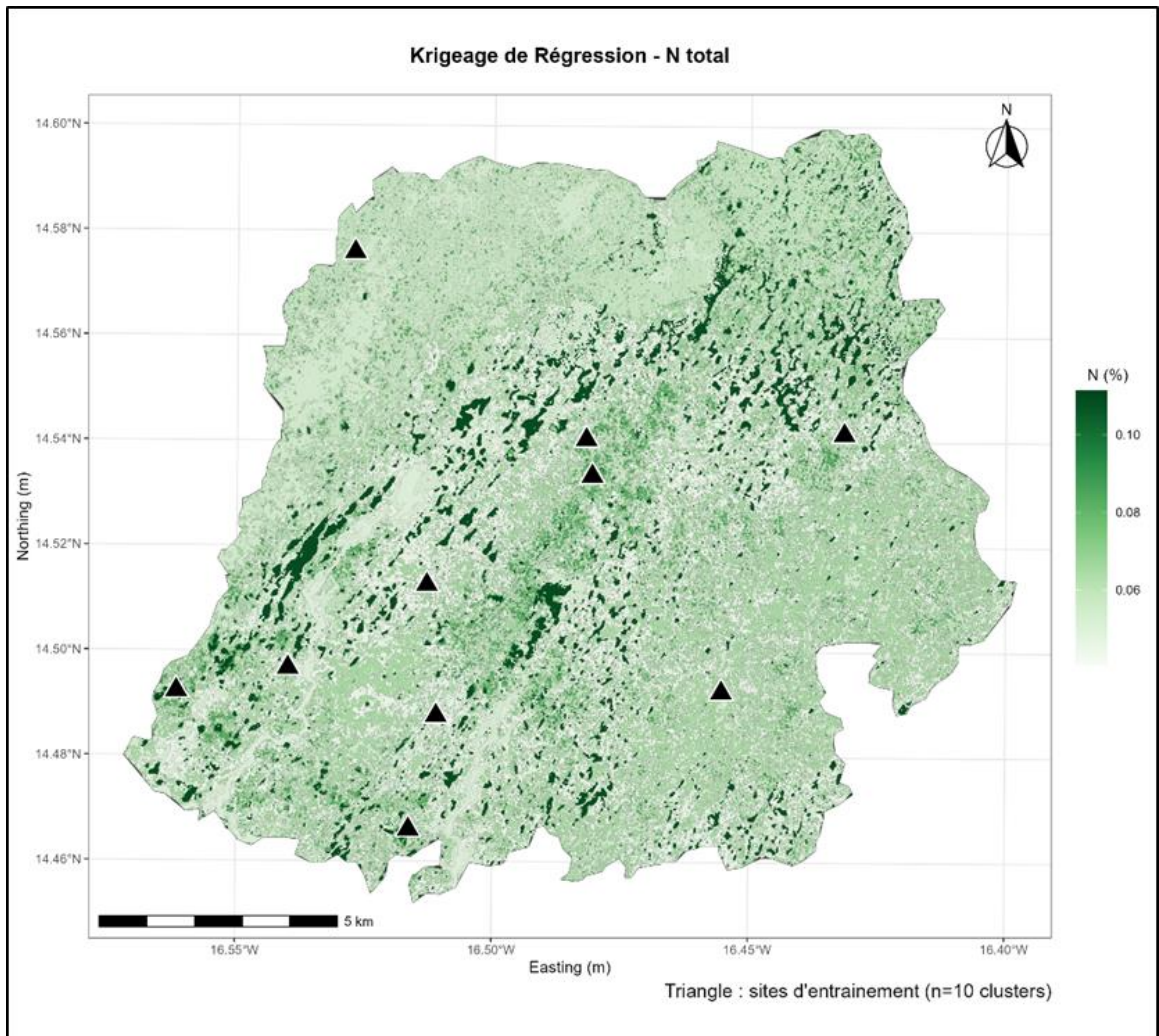


Figure 9 : Distribution spatiale de l'azote total prédit (N, %) par krigeage de régression

Distribution spatiale du phosphore disponible

Les prédictions RK du phosphore (Tableau 4) présentent la plus forte hétérogénéité spatiale (plage : -0,27 à 41,62 ppm ; médiane = 5,17 ppm ; CV = 62,8 %). Le cluster 9 se démarque nettement avec une valeur moyenne de 31,0 ppm, soit cinq fois supérieure à la moyenne zonale, constituant un hotspot de phosphore probablement associé à des affleurements de phosphates de chaux ou à des accumulations de matière organique phosphatée (Figure 10). Les valeurs légèrement négatives prédites par le krigeage non contraint pour quelques polygones (min = -0,27 ppm) seront ramenées à zéro pour les applications agronomiques.

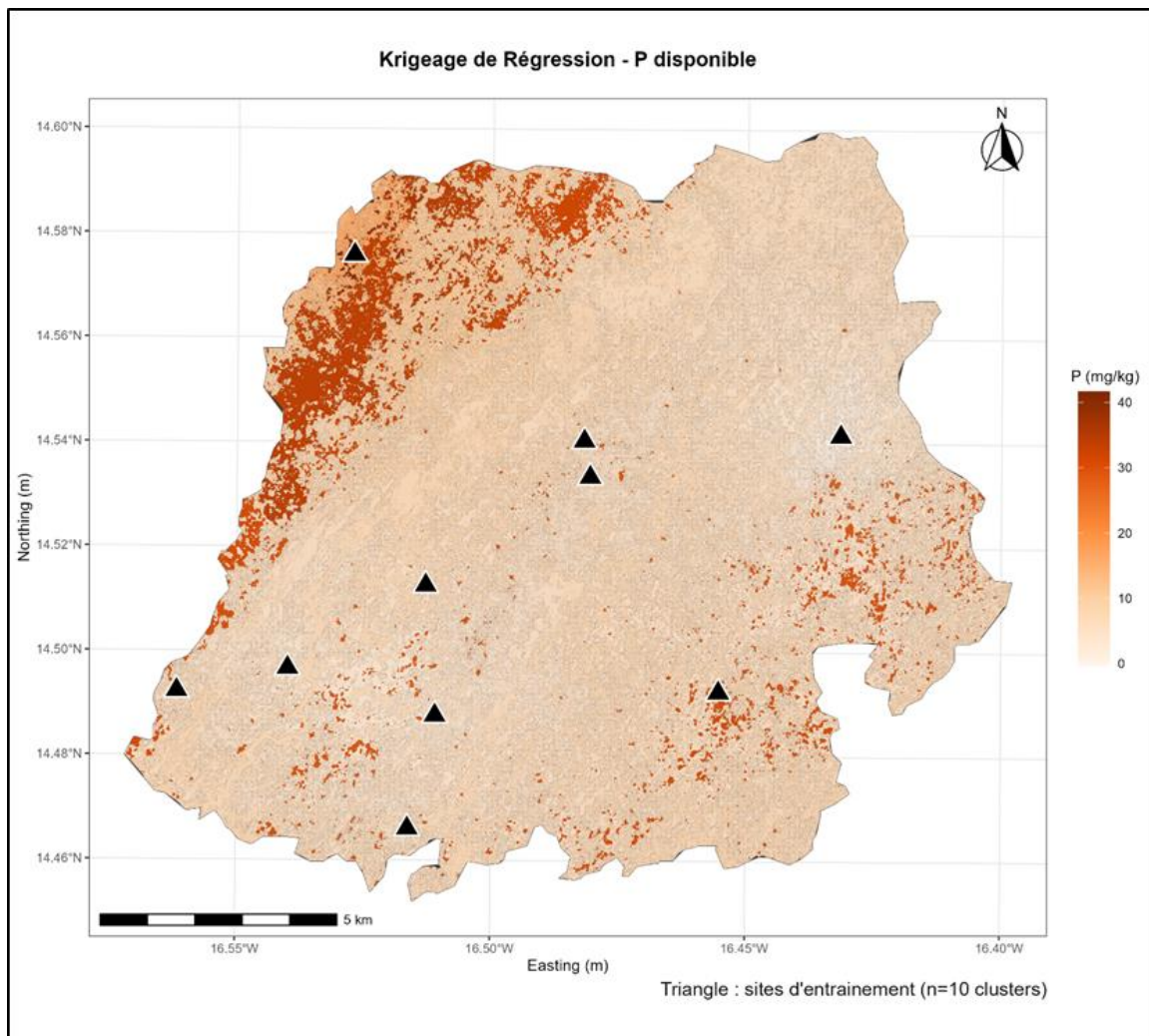


Figure 10 : Distribution spatiale du phosphore disponible prédit (P, ppm) par krigeage de régression. Carte exploratoire, sans valeur prédictive démontrée (R^2 LOO-CV < 0 pour tous les modèles) à ne pas utiliser pour des décisions de fertilisation. Elle sert de support pour un échantillonnage ciblé pour l'étude du phosphore des sols de la zone

Distribution spatiale du potassium échangeable

Le potassium prédit (Tableau 4) varie de 0,036 à 0,171 meq/100 g (médiane = 0,085 meq/100 g ; CV = 30,5 %). La Figure 11 montre une structuration spatiale nette : les positions topographiques basses (clusters 4 et 6, K moyen > 0,10 meq/100 g) sont enrichies en K par les processus de lessivage et d'accumulation dans les dépressions. Le cluster 4 (11 744 polygones) présente les teneurs les plus élevées (K moyen = 0,168 meq/100 g), proches du seuil de suffisance agronomique (0,2 meq/100 g). Le cluster 10 (K = 0,037 meq/100 g) représente la situation la plus déficitaire.

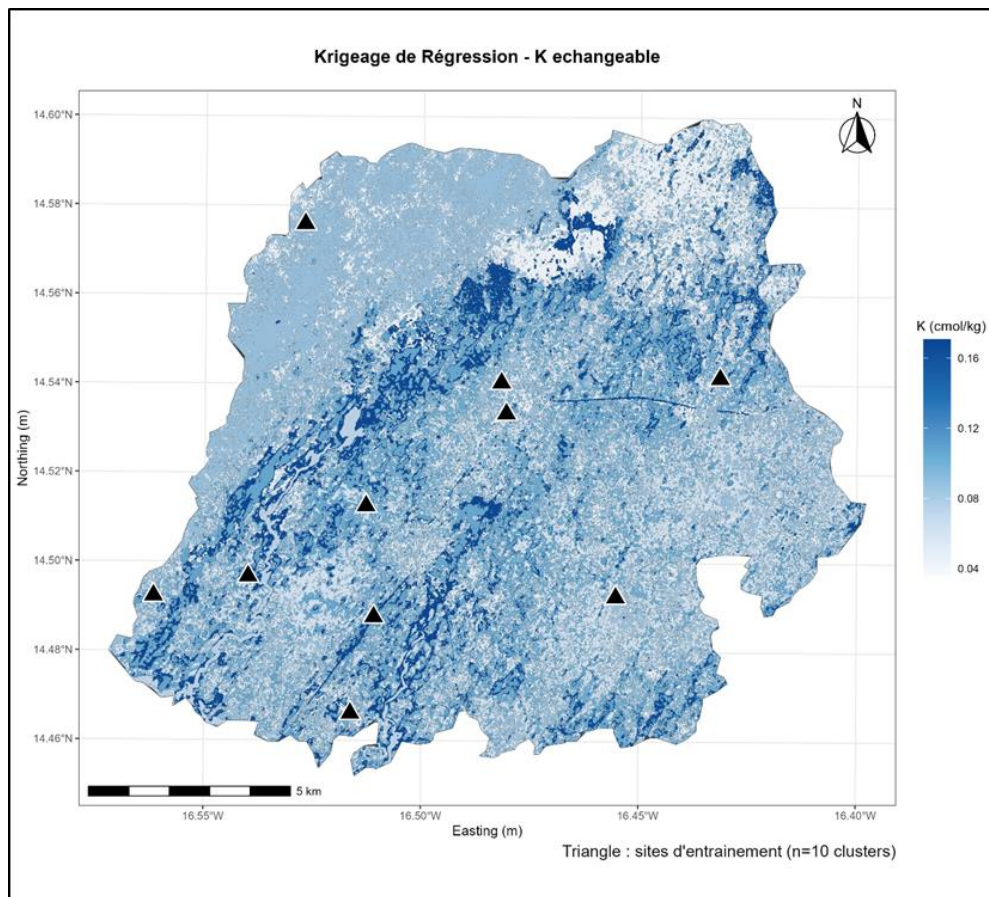


Figure 11 : Distribution spatiale du potassium échangeable prédit (K, meq/100 g) par krigage de régression

Tableau 4 : Prédictions moyennes de N, P et K par krigage de régression pour chaque cluster pédologique et la zone totale (148 433 polygones). En gras: valeurs de la zone totale

Cluster	n polygones	N RK moyen (%)	P RK moyen (ppm)	K RK moyen (meq/100 g)
1	17 193	0,0550	7,27	0,085
2	39 631	0,0651	4,06	0,055
3	282	0,0467	7,26	0,072
4	11 744	0,0502	5,97	0,168
5	10 162	0,0638	4,93	0,055
6	3 479	0,1091	4,83	0,105
7	23 085	0,0416	7,67	0,103
8	22 547	0,0827	4,81	0,099
9	3 590	0,0510	31,01	0,086
10	16 720	0,0556	6,07	0,037
ZONE TOTALE	148 433	0,0613	6,22	0,081

Incertitudes spatiales

L'écart-type de krigage (σ_k) fournit une mesure directe de l'incertitude locale. Pour N, σ_k est quasi-constant (environ 0,010 %), reflétant une portée variographique supérieure aux distances inter-centroïdes.

Pour P, σ_k varie entre 1,5 et 3,5 ppm avec des maxima dans les zones les plus distantes des sites d'entraînement. Les cartes d'incertitude (Figure 12) identifient les zones prioritaires pour des efforts d'échantillonnage supplémentaires : principalement les clusters 3 et 9, de faible superficie mais à forte variabilité interne.

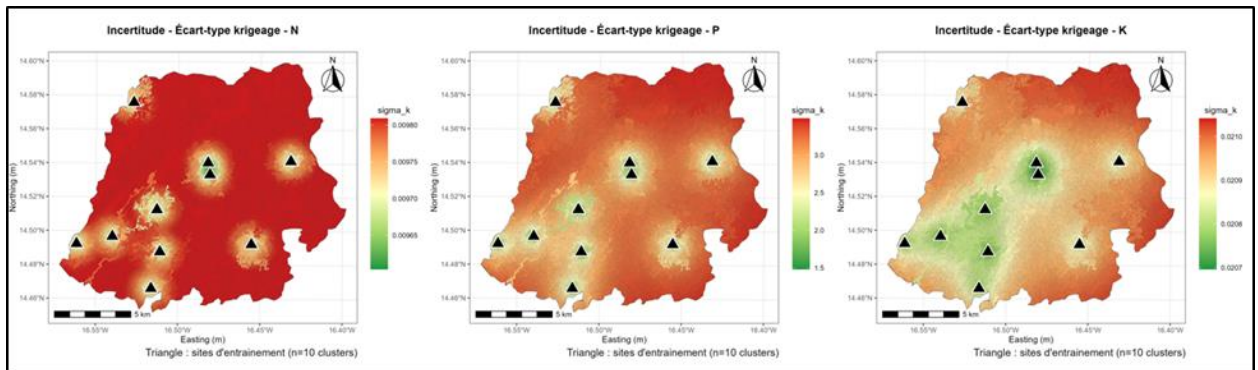


Figure 12 : Cartes des incertitudes spatiales (σ_k de krigeage) pour N, P et K

Discussion

Variabilité spatiale des nutriments et contrôles pédogénétiques

La forte variabilité des propriétés chimiques observée, particulièrement pour P (CV = 150 %) et K (CV = 62 %), est caractéristique du bassin arachidier soumis à une mosaïque de processus pédogénétiques contrastés. Ces niveaux sont cohérents avec ceux rapportés par Vohland et al. (2011) et Cambule et al. (2012) pour des écosystèmes analogues d'Afrique subsaharienne. La structure spatiale identifiée par l'ACP confirme l'opérationnalité du cadre SCORPAN (McBratney et al., 2003) : N et K sont principalement contrôlés par la biomasse et l'activité biologique (facteurs O et B du SCORPAN), tandis que P répond davantage au substrat lithologique (facteur P).

L'identification du cluster 9 comme hotspot de phosphore (P moyen = 31 ppm) constitue un résultat scientifiquement significatif. Sa localisation dans une zone associée à des formations phosphatées vers la région de Thiès (Diallo et al., 2015) ouvre des perspectives pour une valorisation agricole différenciée. Cependant, la haute concentration en P ne garantit pas sa phytodisponibilité effective, qui dépend du pH, de l'Al et du Fe échangeables et de l'activité microbienne (Frossard et al., 2000). Des analyses de phosphore fractionné (méthodes de Chang & Jackson) seraient nécessaires pour préciser la nature des composés phosphorés dominants.

Performances comparées des modèles et contrainte n = 10

L'interprétation de la performance des modèles se concentre sur les quatre modèles les plus robustes au regard de la taille d'échantillon (MLR,

Ridge, MARS, RF) ; les cinq autres algorithmes (Lasso, PLSR, Cubist, XGBoost, SVM) sont reportés dans les tableaux à titre comparatif pour les études future. Les performances modérées des neuf modèles reflètent une contrainte fondamentale : la limitation à 10 classes pédologiques (clusters) comme unités d'entraînement. Avec $n = 10$, les algorithmes ensemblistes complexes (RF, XGBoost, SVM) souffrent d'un ratio signal/bruit défavorable. Ce constat est en accord avec les recommandations de Kuhn et Johnson (2013) qui préconisent $n > 30p$ pour une estimation fiable des hyperparamètres. La supériorité de MARS pour N s'explique par sa capacité à modéliser des ruptures de pente dans la relation sol-covariable avec peu de paramètres effectifs. La régression régularisée Ridge, en réduisant la variance des estimateurs au détriment d'un biais contrôlé, offre également un compromis biais-variance avantageux (Grimm et al., 2008 ; Viscarra Rossel & Behrens, 2010).

L'échec systématique des modèles pour P ($R^2 < 0$ pour tous) illustre les limites de la CNS pour les propriétés à forte dépendance stochastique locale (Pouladi et al., 2019). Le phosphore, très immobile et fortement contrôlé par des processus locaux (adsorption, pH, minéralogie), présente une structure spatiale à courte portée que 10 clusters ne peuvent caractériser. Une stratégie d'échantillonnage adaptatif guidée par l'incertitude cartographique (Brus & Heuvelink, 2007) permettrait d'allouer des observations supplémentaires dans les zones à forte variance de krigeage, améliorant sensiblement la précision du modèle P. Une transformation logarithmique préalable du P serait également recommandée étant donné son asymétrie forte ($\gamma = 2,60$). Sur le plan pédogénétique, cette imprévisibilité s'explique par la forte réactivité du P avec les oxydes de fer et d'aluminium et par sa sensibilité au pH local, deux facteurs à variabilité spatiale fine que les covariables SCORPAN actuelles ne capturent pas, un constat cohérent avec la diversité des mécanismes de carence phosphatée documentée à l'échelle de l'Afrique subsaharienne (Nziguheba et al., 2016). Des stratégies complémentaires pourraient être explorées : un fractionnement du phosphore (méthode de Hedley) permettant de distinguer les formes labiles et non labiles et d'identifier celle qui est effectivement prédictible ; l'ajout de covariables plus directement liées à la dynamique du P (pH, teneur en argile, indices minéralogiques d'oxydes de fer/aluminium, mesurables par exemple par spectroscopie ou fluorescence X, cf. Towett et al., 2015) ; et un échantillonnage ciblé à plus haute densité dans les zones à forte hétérogénéité présumée.

Krigeage de régression : apport et limites

Le krigeage de régression adopté dans cette étude s'inscrit dans le cadre des travaux de Hengl et al. (2007). Par rapport au krigeage ordinaire

seul, le KR incorpore les covariables environnementales via la tendance RF, améliorant la prédiction dans les zones non échantillonnées. Par rapport à la prédiction RF seule, l'ajout du krigeage des résidus réduit l'erreur locale en captant la dépendance spatiale à courte portée. La portée variographique des résidus (1 000-2 000 m) est cohérente avec l'échelle des unités cartographiques. Des approches alternatives - krigeage à transformée gaussienne (trans-Gaussian kriging, adapté à la distribution log-normale de P), simulations séquentielles gaussiennes pour la quantification de l'incertitude, ou krigeage bayésien (Lark & Papritz, 2003), constituent des perspectives d'amélioration méthodologique pour des études ultérieures.

Comparativement aux travaux récents de Hengl et al. (2021) sur SoilGrids250m et de Padarian et al. (2020) sur les réseaux de neurones profonds, notre approche se distingue par une résolution spatiale bien supérieure (polygones < 1 ha) et une adaptation explicite au contexte pédologique local via les clusters. Cela vient au prix d'une transférabilité réduite mais offre une précision locale accrue, pertinente pour des décisions de gestion à l'échelle de l'exploitation agricole.

Implications agronomiques

L'interprétation agronomique des cartes RK doit être conduite avec prudence, compte tenu des limitations méthodologiques identifiées. En particulier, les cartes du phosphore (R^2 LOO-CV < 0 pour tous les modèles) doivent être considérées comme purement exploratoires : elles reflètent la structure spatiale interpolée depuis les 10 clusters, sans capacité prédictive démontrée, et ne sont pas exploitables pour des recommandations de fertilisation phosphatée. Par contre cette carte du phosphore peut servir d'une base d'échantillonnage pour une densification des points de prélèvement. Pour N, les valeurs médianes prédites ($\approx 0,054$ - $0,065$ %) sont inférieures au seuil de suffisance ($N \geq 0,10$ %), suggérant un déficit azoté généralisé cohérent avec les données de littérature sur les sols sahéliens (BUNASAOL, 1990) ; ces observations appellent des études de fertilisation à confirmer par validation de terrain. Il convient de souligner que les seuils de carence mobilisés dans cette étude ($N \geq 0,10$ % ; $K \geq 0,12$ meq/100 g) sont des références génériques issues de la littérature internationale et n'ont pas été calibrées pour les systèmes de culture locaux (arachide, mil) ; une validation agronomique locale de ces seuils est nécessaire avant toute application directe aux recommandations de fertilisation. Pour K, seul le cluster 4 ($K = 0,168$ meq/100 g) approche le seuil de suffisance ; les neuf autres clusters apparaissent déficitaires ($K < 0,15$ meq/100 g pour CEC < 10 meq/100 g). Le cluster 9 constitue un signal de teneur élevée en phosphore ($P = 31$ ppm) qui justifie une investigation géologique indépendante. Ces cartes constituent une première hypothèse spatiale à valider sur des échantillons indépendants avant

toute utilisation opérationnelle dans un système d'aide à la décision de type ISFM (Vanlauwe et al., 2010).

Limites et perspectives

Plusieurs limites méthodologiques majeures doivent être explicitées. Premièrement, l'effectif d'entraînement ($n = 10$ clusters) est statistiquement insuffisant pour la plupart des algorithmes utilisés. L'application simultanée de neuf méthodes d'apprentissage automatique dont RF, XGBoost, SVM et SHAP sur $n = 10$ observations expose au risque de surajustement et limite la fiabilité des hyperparamètres estimés. Les R^2 in-sample élevés (0,941- 0,996 pour la tendance RF) confirment cet ajustement excessif et ne doivent pas être interprétés comme une mesure de performance prédictive. Deuxièmement, le faible effectif d'entraînement ($n = 10$) constitue la principale contrainte et limite la portée statistique des conclusions. Troisièmement, les covariables spectrales à résolution 10 m peuvent introduire des effets de mélange spectral dans les polygones hétérogènes. Quatrièmement, les données sont issues d'une campagne unique, sans considération de la variabilité saisonnière des nutriments. Cinquièmement, et de manière critique, les cartes du phosphore dont tous les modèles donnent un R^2 LOO-CV négatif sont dépourvues de capacité prédictive démontrée et ne doivent en aucun cas être utilisées pour des recommandations de fertilisation phosphatée ou pour quantifier des risques de pollution ; elles illustrent uniquement la structure d'interpolation spatiale entre les 10 clusters. Les perspectives de recherche incluent : (1) augmenter l'effectif d'entraînement par échantillonnage adaptatif guidé par l'incertitude RK ; (2) tester des approches de fusion multi-source (Sentinel-1/2, MODIS) pour P ; (3) évaluer la dynamique temporelle de la fertilité ; (4) intégrer les pratiques agricoles pour distinguer les effets naturels et anthropiques ; (5) valider les prédictions sur un jeu de test indépendant pour confirmer la transférabilité spatiale et, pour P, sur de nouvelles observations terrain avant toute utilisation opérationnelle.

Conclusion

Cette étude a mis en œuvre une démarche de cartographie numérique des sols pour prédire et spatialiser les teneurs en N, P et K du bassin arachidier sénégalais, à partir de dix clusters pédologiques ainsi quinze covariables spectrales et topographiques. Les résultats obtenus sont : (1) la variabilité spatiale des nutriments est élevée avec CV = 43 % pour N, 150 % pour P, 62 % pour K, reflétant la complexité géologique et pédologique de la zone ; (2) en LOO-CV, MARS offre les meilleures performances pour N ($R^2 = 0,457$; RPD = 1,43) et MLR pour K ($R^2 = 0,386$; RPD = 1,35), aucun modèle ne parvenant à prédire P ($R^2 < 0$ en LOO-CV, modèles à considérer comme non prédictifs, les cartes de P relevant uniquement d'une structure d'interpolation spatiale

exploratoire); (3) le krigeage de régression produit des cartes choroplèthes continues, avec des valeurs médianes prédites de $N = 0,054 \%$, $P = 5,17 \text{ ppm}$ et $K = 0,085 \text{ meq/100 g}$; (4) le cluster 9 constitue un signal de phosphore élevé ($P = 31 \text{ ppm}$) méritant une investigation géologique approfondie; (5) les cartes d'incertitude guident les efforts d'échantillonnage futurs. Ces résultats constituent une première base cartographique exploratoire pour le bassin arachidier sénégalais. Leur utilisation opérationnelle pour la gestion différenciée des intrants à l'échelle parcellaire nécessite: (i) l'augmentation de l'effectif d'entraînement au-delà de $n = 10$; (ii) la validation sur un jeu de données terrain indépendant; (iii) pour P, le développement d'une stratégie de modélisation alternative ou d'un échantillonnage adapté avant toute application agronomique.

Remerciements

L'auteur remercie l'Institut National de la Pédologie (INP) du Sénégal pour les analyses pédologiques, ainsi que le Laboratoire de Géomatique et d'Environnement de l'Université Assane Seck de Ziguinchor pour le cadre de travail et la mise à disposition des ressources scientifiques nécessaires à cette étude.

Conflit d'intérêts : Les auteurs n'ont signalé aucun conflit d'intérêts.

Disponibilité des données : Toutes les données sont incluses dans le contenu de l'article.

Déclaration de financement : Les auteurs n'ont obtenu aucun financement pour cette recherche.

References:

1. Banque Mondiale. (2025, May 15). *Revitaliser les sols d'Afrique de l'Ouest*. World Bank Blogs. <https://blogs.worldbank.org/fr/africacan/revitaliser-les-sols-afrique-de-ouest>
2. Bationo, A., Vanlauwe, B., Kihara, J., & Kimetu, J. (2004). Soil organic carbon dynamics, functions and management in West African agro-ecosystems. CIAT-TSBF, Nairobi, 51 p.
3. Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155(3-4), 195-205. <https://doi.org/10.1016/j.geoderma.2009.07.010>

4. Brus, D. J., & Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138(1-2), 86-95. <https://doi.org/10.1016/j.geoderma.2006.10.016>
5. Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., & Smaling, E. M. A. (2012). Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma*, 183-184, 41-48. <https://doi.org/10.1016/j.geoderma.2012.03.011>
6. Diallo, M. D., Ngamb, T., Tine, A. K., Guissé, M., Ndiaye, O., Mahamat Saleh, M., Diallo, A., Seck, S., Diop, A., & Guissé, A. (2015). Caractérisation agropédologique des sols de Mboltime dans la zone des Niayes (Sénégal). *Agronomie Africaine*, 27(1), 57-67.
7. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Márquez, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
8. Feller, C., & Beare, M. H. (1997). Physical control of soil organic matter dynamics in the tropics. *Geoderma*, 79(1-4), 69-116. [https://doi.org/10.1016/S0016-7061\(97\)00039-6](https://doi.org/10.1016/S0016-7061(97)00039-6)
9. Frossard, E., Condon, L. M., Oberson, A., Sinaj, S., & Fardeau, J. C. (2000). Processes governing phosphorus availability in temperate soils. *Journal of Environmental Quality*, 29(1), 15-23. <https://doi.org/10.2134/jeq2000.00472425002900010003x>
10. Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press, New York, 483 p.
11. Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. *Geoderma*, 146(1-2), 102-113. <https://doi.org/10.1016/j.geoderma.2008.05.008>
12. Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10), 1301-1315. <https://doi.org/10.1016/j.cageo.2007.05.001>
13. Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., et al. (2014). SoilGrids1km - Global soil information based on automated mapping. *PLoS ONE*, 9(8), e105992. <https://doi.org/10.1371/journal.pone.0105992>
14. Hengl, T., Miller, M. A. E., Križan, J., et al. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-

- scale ensemble machine learning. *Scientific Reports*, 11, 6130. <https://doi.org/10.1038/s41598-021-85639-y>
15. Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology*. McGraw-Hill, New York, 281 p.
 16. Keskin, H., & Grunwald, S. (2018). Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma*, 326, 22-41. <https://doi.org/10.1016/j.geoderma.2018.04.004>
 17. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer, 625 p.
 18. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). McGraw-Hill, 1396 p.
 19. Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*, 352, 395-413. <https://doi.org/10.1016/j.geoderma.2019.05.031>
 20. Lark, R. M., & Papritz, A. (2003). Fitting a linear model of coregionalization for soil properties using simulated annealing. *Geoderma*, 115(3-4), 245-260. [https://doi.org/10.1016/S0016-7061\(03\)00065-X](https://doi.org/10.1016/S0016-7061(03)00065-X)
 21. McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1-2), 3-52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
 22. Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301-311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
 23. Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping: A review. *Geoderma*, 162(1-2), 1-19. <https://doi.org/10.1016/j.geoderma.2010.12.018>
 24. Nziguheba, G., Zingore, S., Kihara, J., Merckx, R., Njoroge, S., Otinga, A., Vandamme, E., & Vanlauwe, B. (2016). Phosphorus in smallholder farming systems of sub-Saharan Africa: implications for agricultural intensification. *Nutrient Cycling in Agroecosystems*, 104(3), 321-340. <https://doi.org/10.1007/s10705-015-9729-y>
 25. Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, 6(1), 35-52. <https://doi.org/10.5194/soil-6-35-2020>
 26. Pouladi, N., Møller, A. B., Tabatabai, S., & Greve, M. H. (2019). Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma*, 342, 85-92. <https://doi.org/10.1016/j.geoderma.2019.02.019>

27. R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
28. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913-929. <https://doi.org/10.1111/ecog.02881>
29. Tittone, P., & Giller, K. E. (2013). When yield gaps are poverty traps: The paradigm of ecological intensification in African smallholder agriculture. *Field Crops Research*, 143, 76-90.
30. Towett, E. K., Shepherd, K. D., Tondoh, J. E., Winowiecki, L. A., Lulseged, T., Nyambura, M., Sila, A., Vågen, T.-G., & Cadisch, G. (2015). Total elemental composition of soils in Sub-Saharan Africa and relationship with soil forming factors. *Geoderma Regional*, 5, 157-168. <https://doi.org/10.1016/j.geodrs.2015.06.002>
31. Vågen, T.-G., Winowiecki, L. A., Abegaz, A., & Hadgu, K. M. (2013). Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. *Remote Sensing of Environment*, 134, 266-275. <https://doi.org/10.1016/j.rse.2013.03.006>
32. Vanlauwe, B., Bationo, A., Chianu, J., Giller, K. E., Merckx, R., Mokwunye, U., Ohiokpehai, O., Pypers, P., Tabo, R., & Sanginga, N. (2010). Integrated soil fertility management: Operational definition and consequences for implementation and dissemination. *Outlook on Agriculture*, 39(1), 17-24.
33. Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1-2), 46-54. <https://doi.org/10.1016/j.geoderma.2009.12.025>
34. Vohland, M., Besold, B., Bouma, J., & Schwinning, S. (2011). Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, 166(1-2), 198-205. <https://doi.org/10.1016/j.geoderma.2011.08.001>
35. Wadoux, A. M. J. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, Article 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
36. Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/9780470517277>