DÉPARTEMENT DE GÉOMATIQUE APPLIQUÉE
Faculté des lettres et sciences humaines
Université de Sherbrooke

Modèle intégré basé sur l'apprentissage automatique pour l'estimation de la production des cultures au Mali

Par

Lamine N'DAW

Mémoire de recherche - TEL804

Mémoire présenté pour l'obtention du grade de Maître ès sciences géographiques (M.Sc.), cheminement recherche en géomatique

Mai 2023

Approuvé par :

_____

Yacine Bouroubi Date

Professeur au département de géomatique appliquée,

Université de Sherbrooke

TABLE DES MATIERES

Liste des figures

Liste des tableaux

Liste des abbreviations

GEE : Google Earth Engine
GLIMMIX : generalized linear mixed models (modeles linéaires mixtes généralisé)
PQL : Penalized quasi-likelihood(quasi-vraisemblance pénalisée)
RF : forêt aléatoire
RVB : rouge, vert et bleu
NBR : Normalized Burn Ratio (Taux de combustion nromalisé)
ROC : Receiver Operating Characteristic (Caractéristiques de fonctionnement du recepteur)

Résumé

Au regard de la croissance actuelle de la population mondiale, qui atteindra selon l'Organisation des Nations Unies (ONU) environ 10 milliards de personnes d'ici 2050, il faudra une augmentation de 70% de la production alimentaire et de la superficie allouée aux cultures. Les pays de l'Afrique subsaharienne devront donc augmenter leur production alimentaire pour satisfaire la demande croissante en ce qui les concerne. Les pressions démographiques et le réchauffement climatique perturbent les moyens de subsistance et la sécurité alimentaire des populations. De plus, les systèmes agricoles des petits exploitants des pays africains font face à des problèmes de fertilité des sols. Cela est dû à de nombreux facteurs tels que : la taille des champs, l'hétérogénéité des pratiques de gestion, et la fragmentation des paysages qui en résulte et la présence généralisée d'arbres dans les champs. À cet égard, il est primordial de trouver une solution qui fournit des informations sur la productivité des cultures en milieu semi-aride. Ainsi, ce projet propose une approche basée sur la méthode d'échantillonnage aréolaire à partir des données satellitaires et a pour but de contribuer à l'amélioration de l'estimation de la production agricole en milieu semi-aride. Nous avons utilisé Google Earth Engine (GEE) pour faire d'abord une classification par forêt aléatoire (RF) supervisée, basée sur des pixels, pour discriminer les principaux types de cultures. Ensuite, nous avons déduit certains paramètres biophysiques, qui nous ont permis d'analyser les changements dynamiques des indices de végétation, dans le temps et dans l'espace. Enfin, la méthode d'analyse de régression logistique a été appliquée à ces données géospatiales pour estimer le rendement et la production des types de cultures par parcelle. Cette solution permet l'élaboration des cartes de production et de rendement spatialement explicites et résolues dans le temps, afin de servir d'aide en matière de prise de décision stratégique. Cela contribuera à la mise en place des scenarios futurs qui permettrons de mieux gérer la production agricole et de prévenir les pénuries.

Mots-clés : Agriculture de précision, variabilité spatio-temporelle, télédétection, Google Earth Engine (GEE), rendements agricoles.

1. Introduction

1.1 Mise en contexte

De nos jours, 7,6 milliards de personnes vivent sur Terre et ce nombre atteindra environ 10 milliards de d'ici 2050 (Mateo-Sanchis *et al.*, 2019). Cette croissance nécessitera une augmentation de 70%, voire de 100%, de la production alimentaire et de la superficie allouée aux cultures (Fieuzal *et al.*, 2017a). Ainsi, les pays de l'Afrique Subsaharienne devront augmenter leur production alimentaire pour satisfaire la demande croissante en ce qui les concerne, qui devrait doubler d'ici 2050 par rapport à 2011 au regard de l'évolution actuelle de leurs populations (Foley *et al.*, 2011). Au Mali, la part des aliments transformés est plus élevée dans les zones urbaines (60%) que dans les zones rurales (48%), mais la consommation de repas en dehors du domicile reste faible (Smale *et al.*, 2020). Fort de ce constat, différentes organisations institutionnelles et politiques dont l'objectif est d'améliorer la sécurité alimentaire ont été mises en place au Mali pour assurer les orientations stratégiques (Dury *et al.*, 2010). L'une de ces orientations consiste à faire des prévisions robustes de la production au niveau des cultures pour représenter la diversité des sources de nourriture et de revenus, et pour l'efficacité économique du pays (Lambert *et al.*, 2018).

Les pressions démographiques perturbent les moyens de subsistance et la sécurité alimentaire des populations, notamment dans les pays en développement (Karlson et Ostwald, 2016). Ce phénomène démographique est couplé au réchauffement climatique qui cause les changements brusques de saisons et de longues périodes de sècheresses et le Mali n'est pas en marge de cette situation. Ainsi, dans le centre du Mali, on constate que les systèmes agricoles des petits exploitants font face à des problèmes de fertilité des sols (Potts *et al.*, 2013). Par conséquent, la production agricole fortement perturbée dans le temps et dans l'espace pourrait entrainer une insécurité alimentaire chronique (Akponikpè *et al.*, 2011). À cet égard, il est primordial de trouver une solution qui fournit des informations sur l'état des cultures sur de grandes superficies (Atzberger, 2013). Cette solution devrait permettre l'élaboration des cartes de production et de rendement spatialement explicites pour servir d'aide en matière de prise de décision stratégique (Fieuzal *et al.*, 2017a). Les systèmes opérationnels de prévision du rendement des cultures ont une valeur très critique, avec une implication positive ou négative sur l'économie et directement sur le plan social (López-Lozano *et al.*, 2015). L'estimation d'un modèle intégré de manière précise de la production des cultures est actuellement l'un des principaux défis de la recherche agricole et comporte une

utilité primordiale pour les gouvernements (Fritz *et al.*, 2019). C'est dans ce cadre que s'inscrit notre projet de recherche.

Le présent projet propose donc une approche d'estimation des rendements basée sur la méthode d'échantillonnage aréolaire et sur les données satellitaires, afin de fournir une information concernant la productivité des cultures en milieu semi-aride. Cette approche permet d'extraire les statistiques agricoles à partir des images des satellites. C'est un outil efficace pour la production de statistiques sur l'utilisation des sols (types des cultures) et les rendements. Elle peut aussi être utile lorsqu'aucune liste d'agriculteurs n'est disponible, pour des raisons juridiques ou pratiques (ex. : pays connaissant d'importants changements agraires). Les enquêtes sur les superficies peuvent ensuite être réparties en deux catégories : les enquêtes sur l'utilisation des sols et les rendements et les enquêtes polyvalentes. La première catégorie ne fournit que des renseignements sur la superficie et les rendements, tandis que la seconde fournit des renseignements sur tous les aspects de l'exploitation. L'amélioration de l'estimation de la production agricole en milieu semi-aride permettra de mettre en place des scenarios futurs qui permettrons d'anticiper les pénuries de produits agricoles. Enfin, cette solution devrait permettre l'élaboration des cartes de production et de rendement spatialement explicites pour servir d'aide en matière de prise de décision stratégique (Fieuzal *et al.*, 2017b) .

1.2 Problématique

Les estimations des rendement de culture, comme le maïs, peuvent se baser sur des modèles agrométéorologiques appliqués à la saison de croissance (Camps-Valls et Bruzzone, 2009). Traditionnellement, les petites exploitations agricoles font face à des grands défis en raison de la taille des champs, l'hétérogénéité des pratiques de gestion, la fragmentation des paysages et la présence généralisée d'arbres dans les champs (Delrue *et al.*, 2013). L'application de ce type de modèle est compliquée, et souvent la combinaison avec les observations sur le terrain n'est pas possible (van Wart *et al.*, 2013). Parallèlement, la production des statistiques agricoles (superficie, rendement et production) au centre du Mali dépend des enquêtes de terrain à forte intensité de main-d'œuvre, ce qui les rend limitées en termes de représentativité et de fiabilité (Genesio *et al.*, 2011). De plus, ces enquêtes ne sont généralement pas adaptées à de plus grandes échelles et consomment du temps et des ressources considérables (Lambert *et al.*, 2018). Soulignons aussi que la consommation totale du pays en produits agricoles peut dépasser la production lorsque des

conditions météorologiques ne sont pas favorables (Zhao *et al.*, 2017). L'augmentation possible des conditions météorologiques défavorables pour la production agricole en raison des changements climatiques peut entraîner une baisse des rendements céréaliers (Zhao *et al.*, 2017).

Aujourd'hui, les outils de la géomatique sont utilisés pour aider à recueillir des données de qualité, surtout en agriculture de précision, et peuvent être exploitées grâce à des analyses appropriées afin de suivre la croissance au cours de la saison (Huuskonen et Oksanen, 2019). La précision dans l'observation de la croissance des cultures dans les petites exploitations agricoles avec l'abondance de la forêt nécessite une série chronologique d'images satellites à haute résolution (HR) (Schut *et al.*, 2018). La disponibilité de données d'imagerie satellitaire est de plus en plus grande (Bandrova *et al.*, 2016). Les satellites ont été largement utilisés pour surveiller différents types de cultures agricoles (Khot *et al.*, 2016) et étudier la phénologie des plantes (Sankaran *et al.*, 2015).

La question la plus importante est de trouver la meilleure approche pour extraire les statistiques agricoles à partir des images prises par des satellites, des avions ou des drones. La méthode *«* Échantillonnage aléatoire *» (Area Frame Sampling)* est l'une des approches le plus souvent utilisées. L'utilisation des images de télédétection et des données de terrain est problématique en appliquant les approches de régression traditionnelles selon Duniway *et al.* (2012), surtout si la taille de l'échantillon est très petite. Pour Lary *et al.* (2016), le dilemme lié à la non-linéarité des petites tailles de l'échantillon peut être surmonté si on applique les algorithmes d'apprentissage automatique.

Selon Partel *et al.* (2019), en télédétection, les applications agricoles utilisent de plus en plus l'apprentissage profond (ex. : les réseaux de neurones à convolution CNN). Pour avoir en sortie des informations sémantiques, d'après Krizhevsky *et al.* (2012), les CNN doivent nécessairement s'appuyer sur une quantité importante de données afin de créer des caractéristiques hiérarchiques significatives. Auparavant, Camps-Valls et Bruzzone (2009) ont introduit l'utilisation à la fois d'un modèle de régression linéaire et de son extension non linéaire. Les modèles sont développés et évalués pour l'estimation de la production agricole, où toutes les observations avant la récolte sont utilisées et dans les prévisions intra-saisonnières qui comprennent un nombre croissant d'observations temporelles (Mateo-Sanchis *et al.*, 2019).

La présente étude s'intéresse particulièrement à la question suivante : comment trouver la meilleure approche pour développer un modèle d'estimation robuste et fiable des rendements des cultures

par parcelle, à partir des données satellitaires et de moyennes de rendements obtenues par localité?

Cette étude présente une nouvelle application basée sur la méthode d'échantillonnage aréolaire (AFS) qui peut être utilisée pour traiter, analyser et visualiser de manière efficace les données collectées par satellites, et enfin d'étudier et évaluer le rendement des champs agricoles.

1.3 Objectifs et hypothèse

L'objectif principal de ce projet de maitrise est de contribuer à l'amélioration de l'estimation de la production agricole en milieu semi-aride à agriculture familiale, par une approche basée sur la méthode d'échantillonnage aréolaire et les données satellitaires. Pour atteindre cet objectif, les objectifs spécifiques suivants sont visés :

1) Le premier objectif spécifique est l'utilisation de deux méthodes pour l'identification des cultures, la première basée une classification par régression logistique incluant un réseau de neurones (CART) et la seconde basée sur les forêts aléatoire;
2) Le deuxième concerne l'estimation de la production totale des types de cultures dans la région tropicale semi-aride où l'agriculture familiale est pratiquée.

Notre hypothèse stipule est que le développement d'un modèle basé sur la combinaison de la méthode d'échantillonnage aréolaire (AFS) et des indices de végétation issus des images satellites permet d'obtenir une estimation de la production fiable et évolutive, à l'échelle de la parcelle.

2. Cadre théorique

2.1 Discrimination et stratégie de classification des cultures

L'élaboration d'une méthode originale pour détecter l'apparition de la sénescence végétale en temps quasi réel a été développée par (Renier *et al.*, 2015). Leur méthode de détection est fondée sur le comportement temporel de deux indices. Il s'agit du NDVI et de l'indice d'humidité du sol à différence normalisée NDMI. Ils sont sensibles à la végétation verte et sèche. Leur méthode a été appliquée en Mauritanie au moyen de composites MODIS de 10 jours pour les années 2010 et 2011. Les performances en matière de discrimination de trois classes (croissance, réduction de la densité et séchage) ont été analysées selon trois méthodes de classification. Il s'agit de la méthode de maximum de vraisemblance (exactitude globale de 61,4 %), de l'arbre décisionnel (avec une précision de 71,5 %) et les machines à vecteurs de support SVM (avec une précision de 72,3 %).

Les résultats obtenus ouvrent la voie à la première mise en œuvre opérationnelle de cartographies dynamiques de sénescence. Par conséquent, plusieurs facteurs ont eu une incidence sur l'exactitude du classement. Ce sont le gradient climatique nord-sud, la densité de la végétation, la période de cartographie de la végétation et le site. Chen *et al.* (2020) ont exploré dans leur recherche les caractéristiques de texture dérivées de Landsat-8 OLI et de la rétrodiffusion SAR de Sentinel-1 filtrée et non filtrée du speckle. Ils ont regroupé les résultats de la classification à l'aide de *Decision-level Fusion* (DLF) et évalué le rendement au niveau de la décision des cartes fusionnées. Ils ont également utilisé la *Grey Level Cooccurrence Matrix* (GLCM) pour obtenir des ensembles de sept caractéristiques de texture pour les huit bandes de Landsat-8. Chaque caractéristique de texture a été empilée avec une image source correspondante et déposée en utilisant les SVM. Les cartes classées des trois principales interprétations, filtrées et non filtrées, sont agrégées avec des cartes classées de Landsat-8. Leurs résultats indiquent une précision de classification globale de 96 % fondée sur les images DLF de Landsat-8 et Sentinel-1 non filtrées, alors que les filtres « Landsat-8 » et « Speckle » de GEE ont donné une précision de classification globale de près de 95%. Leur meilleure information de texture est dérivée de la bande bleue suivie par la bande rouge, alors que les textures multi-bandes non filtrées ont donné de meilleurs résultats. Les chercheurs ont conclu que l'intégration de Landsat-8 et Sentinel-1, filtrée ou non, améliore la classification des cultures. Kobayashi *et al.* (2020) ont utilisé 91 indices spectraux publiés dans leur étude, qui ont été dérivés de données Sentinel-2 MSI. Les algorithmes de classification ont été utilisés pour générer des cartes précises. Les auteurs ont utilisé le classificateur de forêt aléatoire et obtenu une précision globale de 93,1% sur la base de la réflectance de 4 bandes et 8 indices spectraux. Ainsi, la discrimination des cultures au niveau de la parcelle permet une gestion précise des cultures agricoles locales (Nidamanuri *et al.*, 2022).

2.2 Estimation du rendement des cultures par données satellitaires

Les données satellitaires englobent diverses plages spectrales susceptibles de contenir des renseignements supplémentaires sur la croissance et le rendement des cultures. Mais, celles-ci ont été dans une large mesure sous-étudiés et sous-utilisées (Duthoit, 2006). Surveiller les cultures à grande échelle et estimer la production sont importantes à la fois pour la recherche scientifique et pour les applications pratiques (Lobell *et al.*, 2009). Thorp *et al.* (2012) ont démontré que les données satellitaires constituent un moyen efficace de surveillance régionale et mondiale des terres cultivées, surtout dans les régions pauvres en données qui n'ont pas d'observations fiables ni de

rapports sur le terrain (ex. : le Mali). Pour van Ittersum *et al.* (2013), l'approche utilisée est basée sur les observations de l'indice de végétation NDVI qui sont disponibles depuis des décennies. Guan *et al.* (2017) ont mené l'une des premières tentatives de mise en synergie de plusieurs données satellitaires couvrant un spectre diversifié d'estimation du rendement des cultures pour la ceinture de maïs (*corn belt*) des États-Unis. Ils ont inclus dans leur étude MODIS *Enhanced VI* (EVI), la production primaire brute estimée basée sur la fluorescence induite par le soleil GOME-2 (SIF-GPP), l'évapotranspiration thermique ALEXI (ET), Quikscat Radar en bande 5 Ku, et AMSR-E micro-ondes passives (bande profondeur optique de végétation VOD). Leur modèle, qui comprend des données satellites et des variables climatiques entre 2007 et 2009, explique la variabilité du rendement des cultures avec une probabilité de 82 %. Ces résultats éclairent également sur l'utilisation synergique des diverses missions satellitaires en cours et celles de la génération suivante, en particulier pour les applications d'amélioration des rendements des récoltes et pour le système agricole en général.

White *et al.* (2020) ont étudié l'utilisation d'estimations par micro-ondes passives de la surface humide du sol (*soil moisture* SM) obtenues par le satellite Ocean Salinity Mission (SMOS). Il s'agissait de prédire les rendements du canola dans l'ensemble du Canada. Des prédicteurs clés ont été identifiés et des modèles de régression ont été construits à l'aide d'un solide système de régression au moindre angle. Ils ont découvert que l'humidité du sol donnait une meilleure description du stress du canola que le NDVI. Leurs conclusions suggèrent que les observations satellitaires du SMOS SM peuvent fournir un indicateur plus efficace du rendement du canola.

2.3 Méthode d'analyse des données d'échantillonnage aréolaire

La méthode d'analyse des données d'échantillonnage aréolaire, ou méthode d'enquête par secteur, est une méthode d'étude probabiliste où les unités sondées du dernier degré sont des secteurs appelés segments (Gallego, 1999). Gallego (1999) utilise des échantillonnages aréolaires définies à l'aide d'une grille carrée pour estimer les surfaces des parcelles agricoles dans les relevés sur le terrain et l'imagerie satellitaire à haute résolution. Cette étude probabiliste est proportionnelle à la surface mesurée de celle-ci. Pradhan (2001) adopte une autre approche pour élaborer un système d'information géographique (SIG) afin d'estimer les terres cultivées à l'appui des systèmes de prévision des cultures au niveau régional, en Iran. Grâce à l'échantillonnage sélectif, sa méthode consiste à associer l'échantillonnage aux données de télédétection. Ils ont élaboré un plan plus

détaillé visant à appuyer l'échantillonnage aréolaire. Les fonctions du système ont été mises à l'essai et évaluées à l'aide de données recueillies au cours de l'été 1997 dans la province de Hamadan, en Iran. Leurs résultats montrent un bon accord entre la superficie des principales cultures obtenues grâce au système qui a été développé et les superficies estimées par l'*Agricultural Statistical Information Division* (ASID) du ministère de l'Agriculture de la République islamique d'Iran.

Cependant, pour Gallego (2004), la méthode utilisée dépend de la qualité des documents disponibles et du type de référencement fourni par l'exploitant. D'après cet auteur, la comparaison entre les deux estimations tirées de l'enquête par segment peut être utile pour vérifier si un biais n'a pas été introduit dans l'estimation de la production. Toutefois, dans ce contexte, le coefficient de variation (CV) des estimations est le plus pertinent et le plus utile. Dans le cadre d'échantillonnage par zone, qui est le fondement de nombreux programmes d'apprentissage statistique dans le monde, l'objectif est d'améliorer l'exactitude, l'objectivité et l'efficacité des estimations d'enquête. C'est dans ce cadre que Boryan *et al.* (2014) ont utilisé les couches de données géospatiales de la fréquence des semis de maïs, de soya et de blé de 2008-2014 pour créer trois stratifications AFS particulières à une ou plusieurs cultures du Dakota du Sud, États-Unis. Le maïs, le soja et le blé sont les trois principales cultures du Dakota du Sud. Les *stratifications propres aux cultures* sont établies sur la base d'apprentissage statistique de fréquence des cultures obtenues au niveau de l'unité d'échantillonnage primaire (UEP) à partir des couches de données d'entrée et de sortie sur la fréquence des cultures. Outre le maïs, les fréquences moyennes d'ensemencement pour le soja et le blé *en monoculture sont stratifiées* en fonction du pourcentage de la récolte. Les trois stratifications de l'AFS proposées par les auteurs fournissent plus de renseignements propres aux cultures que l'actuel *national agricultural statistics Service* (NASSE). En outre, la *stratification multiculture* a été élaborée à partir de couches de données de fréquences d'ensemencement individuelles pour le maïs, le soya et le blé. Les auteurs ont constaté que les quatre stratifications de la AFS basées sur l'apprentissage statistique des cultures prédisent de façon cohérente les périodes d'ensemencement pour le maïs, le soya et le blé, qui sont vérifiées par la *Common Land Unit* (CLU) de 2014 du *Farm Service Agency* (FSA). Cela démontre que les nouvelles stratifications fondées à partir de l'apprentissage statistique sur la fréquence d'ensemencement et la culture sont indépendantes du type de culture et s'appliquent à toutes les principales cultures. De plus, ces résultats indiquent que les nouvelles stratifications de l'apprentissage statistique par culture ont un grand potentiel d'amélioration de l'exactitude, de l'efficacité et des estimations de la

production des cultures. L'utilisation de points seulement à la ferme peut augmenter le coût de l'enquête si le nombre de points par segment doit rester constant.

L'étude de Eltazarov *et al.* (2023) avait pour principal objectif d'examiner les gains possibles découlant de l'utilisation de la classification de l'utilisation des terres, et ce afin de mettre en œuvre une assurance indiciaire en comparant l'adéquation de plusieurs indices satellites, dans différents systèmes de production de blé en Asie centrale et en Mongolie. Les résultats de leur étude mettent en évidence l'importance des tests sur les cultures et les masques de blé lorsqu'ils conçoivent et élaborent une assurance-indice. En outre, les indices LAI et GCI se sont montrés légèrement supérieurs aux autres indices qui étaient bien connus dans la détection de la variation du rendement du blé et avaient donc un plus grand potentiel pour la conception de l'assurance indicielle. Le travail de Qader *et al.* (2023) est l'un des premiers à examiner l'utilisation des données tirées de Sentinelle-2 ainsi que des variables topographiques et climatiques qui peuvent être utilisées en tant que co-variables pour modéliser et prédire précisément le rendement des cultures. À l'échelle des exploitations, avec des données faibles pour les régions arides et semi-arides, leur travail montre qu'avec des données satellitaires librement accessibles des co-variables climatiques et topographiques, les rendements des cultures peuvent être estimés avec précision au moyen de modèles statistiques dans des systèmes agricoles complexes de petites exploitations arides et semi-arides.

D'après Dong *et al.* (2017), les enquêtes par zones semi-arides peuvent être soit des enquêtes sur l'utilisation des sols et les rendements, soit des enquêtes polyvalentes. La première catégorie fournit des informations sur la superficie et les rendements uniquement, tandis que la seconde fournit des informations sur tous les aspects de la ferme. En revanche, pour les limites des champs cultivées, les agriculteurs vont les numériser soit à partir de cartes cadastrales ou topographiques, soit à partir de photos aériennes. Dans le cas des cartes cadastrales, les limites des strates peuvent être des éléments physiques du paysage, comme les routes, les cours d'eau et les limites administratives. La qualité des frontières est le concept le plus important, puisque l'échantillonnage aréolaire doit être utilisée sur une période de plus de 10 ans. Par ailleurs, il n'existe aucune différence entre la nature des limites des strates et la nature des unités d'échantillonnage du dernier ordre (Mugabowindekwe et Rwanyiziri, 2020). Lorsque l'exploitant refuse de coopérer ou n'est pas identifiable, il faut remplacez les lignes de la base de données de cette personne dans la table par les valeurs agricoles moyennes du segment qui a répondu. Autrement, on le remplace par la moyenne des exploitations

qui ont répondu dans tous les segments de la strate traitée. Il faut être très vigilant dans la prise de décision sur les données très grandes. Ils sont souvent liés à la difficulté de repérer les opérateurs et de vérifier les cas où les instructions ont été mal comprises. La documentation présente un certain nombre de règles décisionnelles.

Mugabowindekwe et Rwanyiziri (2020) ont développé une méthode d'échantillonnage harmonisée, systématique et bidimensionnelle avec un seuil de distance dans un cadre de surface stratifié sur une grille carrée. Les résultats obtenus pour la détermination des superficies cultivées, des rendements et de la production ont été satisfaisants. Toutefois, les principaux problèmes étaient liés à la difficulté de repérer les opérateurs et de vérifier les cas où les instructions étaient mal comprises. La conception générale de leur système a été conçue pour faciliter l'estimation de la production agricole au moyen de l'apprentissage statistique.

3. Cadre expérimental du projet

3.1 Description du site d'étude

Le site d'étude est situé dans le bassin de production de coton du Mali, qui couvre une superficie 353 km^2 et est caractérisé par une densité de population relativement élevée. La région de Koutiala est la zone d'étude comportant les cultures de maïs, sorgho, mille et fonio (figure 1).

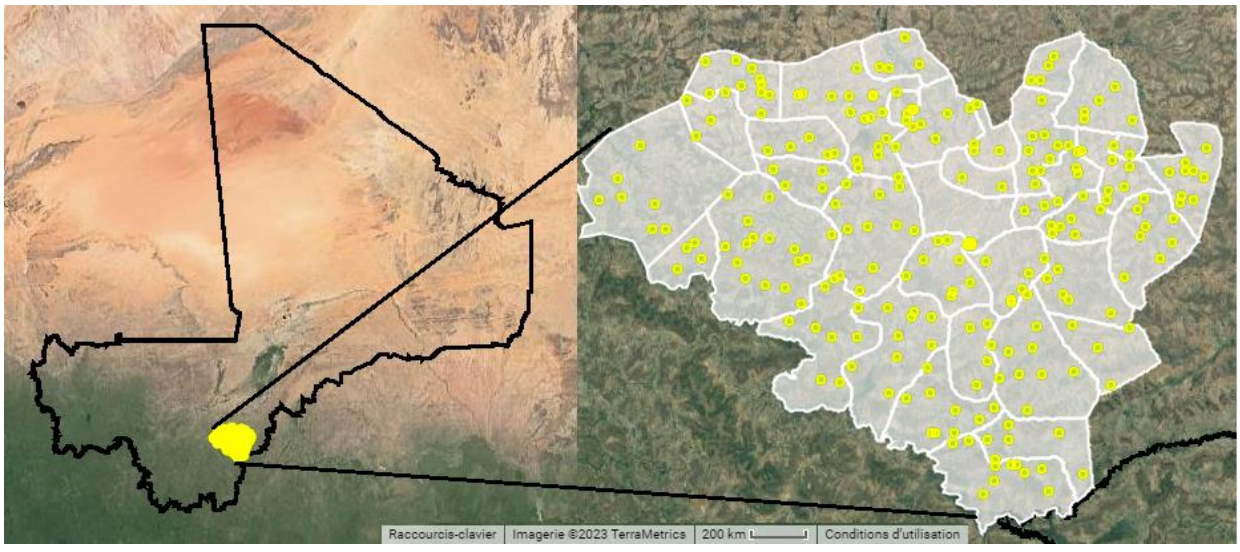


Figure 1 : La carte du Mali et la région de Koutiala (notre zone d'étude) avec la localisation des 205 villages enquêtés. Source graphe (https://code.earthengine.google.com/)

Avec une forte variabilité interannuelle et intra-saisonnière, les cumuls de précipitations annuelles atteignent plus 800 mm (Vischel *et al.*, 2017). L'occupation du territoire comporte des parcelles de cultures très diversifiées, avec une grande variété de types de sols (Blaes *et al.*, 2016). On y distingue :

- Sols rocheux situés sur les plateaux et les pentes;
- Sols graveleux rouges situés sur des plateaux;
- Sols sableux situés sur les pentes et les basses terres;
- Sols limoneux situés dans les basses terres;
- Sols argileux des basses terres.

Les principales cultures de la région sont le coton, le maïs, le sorgho et le millet, avec les premières pluies aux alentours du mois de mai, qui correspondent à la période de la semence des cultures, qui sont récoltées du début septembre (pour le maïs) à la fin novembre (pour le mil). La production de coton avec des organisations paysannes est gérée par la compagnie malienne de développement du textile (CMDT), ce qui a un impact important sur les moyens de subsistance locaux, car le coton est encore aujourd'hui la principale source végétale pour tout ce qui est tissu (Quazzo et Meunier, 2020). Le sorgho et le mil sont principalement cultivés pour la consommation des ménages.

3.2 Description des données

Choix des périodes pour les données images

La période d'intérêt pour le suivi des cultures s'étend de la senescence à la récolte des différents types de cultures et durant toute la période de l'activité agricole dans notre zone d'étude, c'est-à-dire d'avril à décembre des années 2013 à 2020, selon le calendrier indiqué sur la figure suivante :

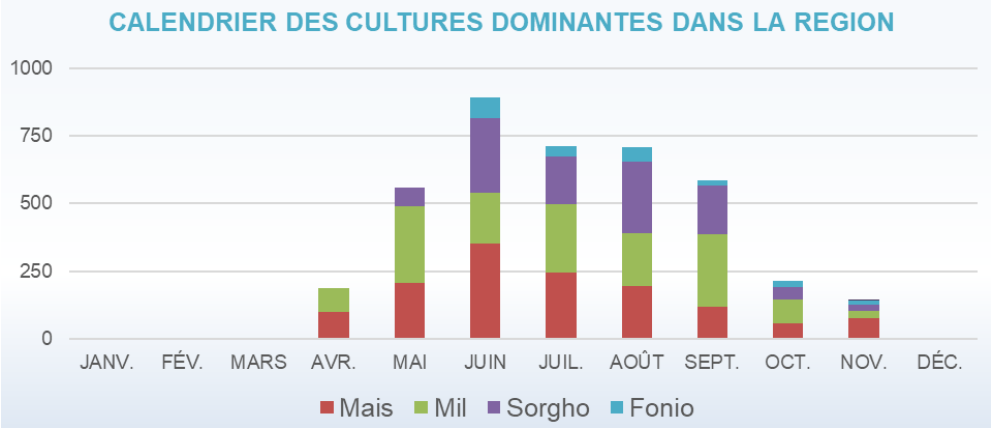


Figure 2 : Calendrier des types cultures dominantes dans la région en tonnes

Cette période a guidé le choix des capteurs et la recherche des images, notamment celles acquises par les capteurs Landsat, Sentinel-2 et Sentinel-1. La principale application de ces données sera la cartographie des types de cultures (Hagolle *et al.*, 2015).

Données de télédétections

Les données de télédétection (tableau 1) utilisées dans cette étude comprennent :

- Landsat TM/ETM+/OLI (résolution spatiale de 30 m);
- Sentinel-2 MSI Level-2A (bandes avec des résolutions spatiales de 10 et 20 m);
- Sentinel-1 SAR GRD (10 m résolution spatiale).

Les périodes d'acquisition sont les suivantes : Landsat 7 ETM+ de 2016 à 2020; Landsat-8 OLI de 2013 à 2018; Sentinel-2 MSI pour 2018; Sentinel-1 pour 2018.

Les données de télédétection et les autres données auxiliaires dont la référence dans GEE est « CGLS-LC100 collection 3 » ont été obtenues à partir de l'ensemble de données en ligne fourni par GEE var dataset = ee.FeatureCollection('USDOS/LSIB_SIMPLE/2017'). La synergie entre les données radar Sentinel-1 et les données optiques Sentinel-2 donne une classification plus stable pour les grands groupes de végétaux, avec une résolution spatiale et temporelle élevée. Ces données sont utilisées pour l'identification des différents types de culture en zones semi-arides, elles donnent une plus grande précision par rapport à la classification globale de la zone agricole de la région. Notons toutefois que ces capteurs ou la méthode de classification ne conviennent pas nécessairement pour distinguer tous les types de cultures, ce qui pourrait limiter l'approche au niveau certaines cultures.

- **Données Sentinel-1**

Lancés en avril 2014 sur Sentinel-1A et en avril 2016 sur Sentinel-1 B, ces capteurs actifs en bande C (5,6 cm) sont les premiers à être mis en orbite sous le programme Copernicus. Ils ont un cycle orbital de 12 jours et une orbite héliosynchrone presque polaire. Les quatre modes d'acquisitions du satellite Sentinel-1 sont : Interferometric Wide Swath, Wave mode, Strip Map mode et Extra Wide Swath (https://sentinel.esa.int).

- **Données Sentinel-2**

Lancés en juin 2015 pour Sentinel-2A et en mars 2017 pour Sentinel-2B, ces satellites sont dotés de l'instrument optique MSI (MultiSpectral Instrument) de 13 bandes spectrales. Le cycle orbital (orbite héliosynchrone polaire) de chaque satellite est de 10 jours. La résolution spatiale varie entre

10 et 60 mètres selon les bandes. Ce satellite vise à observer l'évolution de la végétation, l'utilisation des terres et l'impact du réchauffement climatique. Sentinel-2 est particulièrement adapté pour la cartographie de la végétation. Cela est dû à la présence de deux nouvelles bandes spectrales de dans la bordure rouge (red edge), entre 705 et 740 nm. La fourniture d'images « prêtes à l'emploi » constitue un autre aspect intéressant des produits Sentinel. Plusieurs niveaux de correction sont disponibles pour le téléchargement des images. Les niveaux offerts pour notre zone d'étude en 2017 étaient les niveaux 1C (réflectance apparente) et L2A (réflectance au sol). Les deux niveaux de traitement comportent une orthorectification.

Tableau 1 : Données de télédétection utilisées

Satellite	Capteur	Résolution spectrale		Résolution spatiale (m)	Autres propriétés et utilité pour les statistiques agricoles
		Bandes MS	Nb bandes		
Landsat-7	ETM+	VIS-NIR-SWIR	6	30	Revisite de 5 à 16 jours; utile à l'identification des cultures et à l'estimation des surfaces; présence et état des cultures à l'échelle régionale.
Landsat-8	OLI	VIS-NIR-SWIR	8		
Sentinel-1	C-SAR	Bande C, 5.6 cm	Dual-pol	≥ 5	Revisite de 6 à 12 jours; acquisition par toute condition d'ennuagement, de jour et de nuit.
Sentinel-2	MSI Level-2A	RGB-NIR	4	10	Revisite de 5 jours; contribuent à la délimitation des parcelles et à l'identification des cultures.
		Red-edge	4	20	
		SWIR	2	20	

Données météorologiques

Les données climatiques, plus précisément les données de pluviométrie, de température, les vents et l'ensoleillement sont basées sur les stations de mesure et les données estimées du site web https ://home.openweatermap.org/zip_code_data/new, dont la période varie entre 1991 et 2020. En tout, nous avons plus 307 500 observations.

Les diagrammes des paramètres climatiques (Figure 3) sont basés sur 29 années de simulations horaires de modèles météo et sont disponibles pour chaque emplacement dans notre zone d'étude. Ils fournissent une bonne indication sur les régimes climatiques typiques et les conditions attendues (température, précipitations, ensoleillement et vent). Les données météo simulées ont une résolution spatiale d'environ 30 km. Ils ne peuvent reproduire tous les effets météorologiques locaux, par exemple les orages, les vents locaux ou les tornades. La « moyenne journalière maximale » (ligne rouge continue) représente la température moyenne maximale d'un jour par mois

pour Koutiala. De même, la « moyenne journalière minimale » indique la moyenne de la température minimale. Les journées chaudes et les nuits froides sont la moyenne des journées les plus chaudes et les plus froides de chaque mois au cours des 29 dernières années.

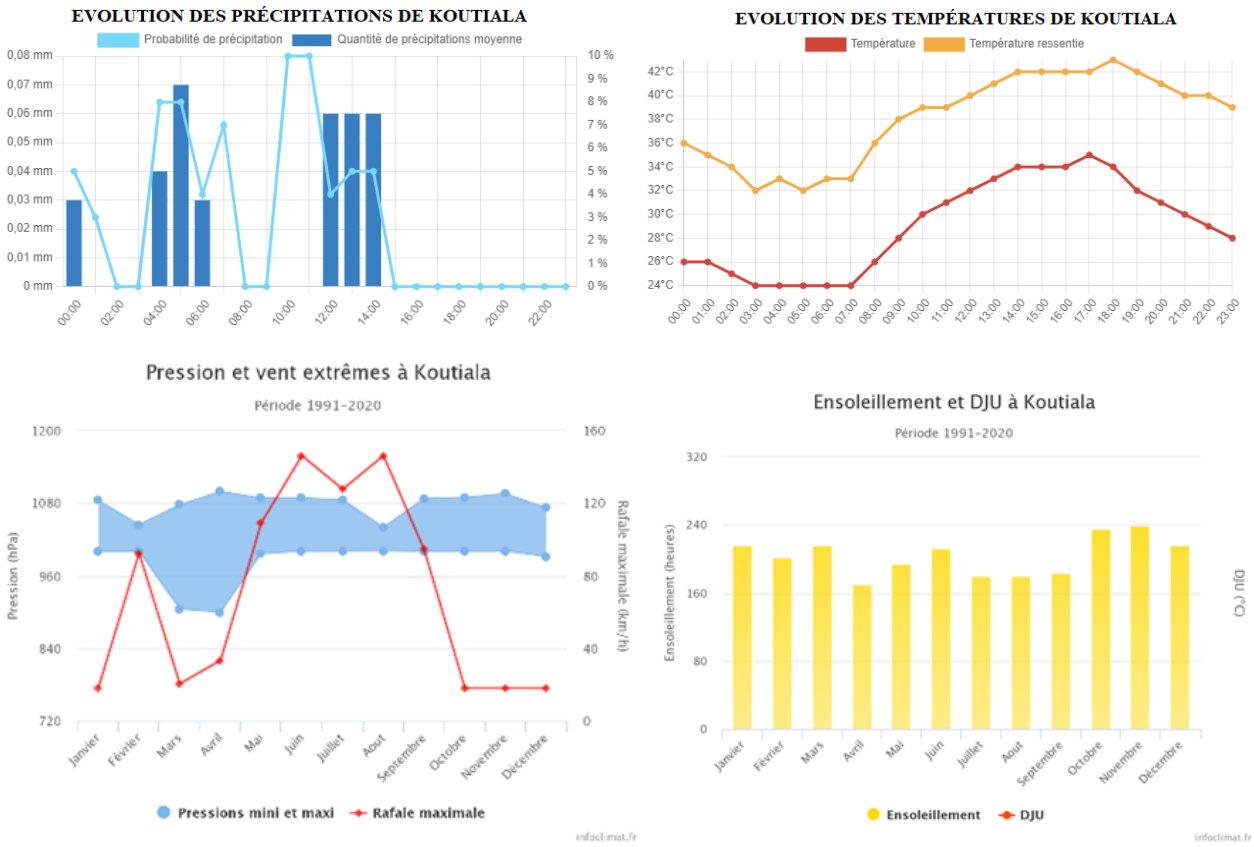


Figure 3: Moyennes mensuelles des données climatiques : températures (°C), précipitations (mm), ensoleillement (heures), vents (km/h) et la pression (hPa).

Données terrain et autres données auxiliaires

Un ensemble de points géoréférencés de types de cultures a été récolté au champ entre octobre et novembre 2019. Ces données incluent les classes suivantes, qui seront prises en considération dans cette étude : coton, maïs, millet et sorgho. Les limites des champs ont également été numérisées à partir de l'image du THR WorldView-3 pour chaque emplacement géoréférencé. Les données auxiliaires incluses sont :

- le SRTM (Shuttle Radar Topography Mission) : DEM (Digital Terrain Model) produit par la navette de la NASA à 30 m de résolution;
- les données vectorielles des limites administratives ainsi que les frontières du pays.

4. Méthodologie

Cette section présente l'approche méthodologique que nous avons suivie dans le cadre du projet. L'organigramme ci-dessous résume les principales étapes

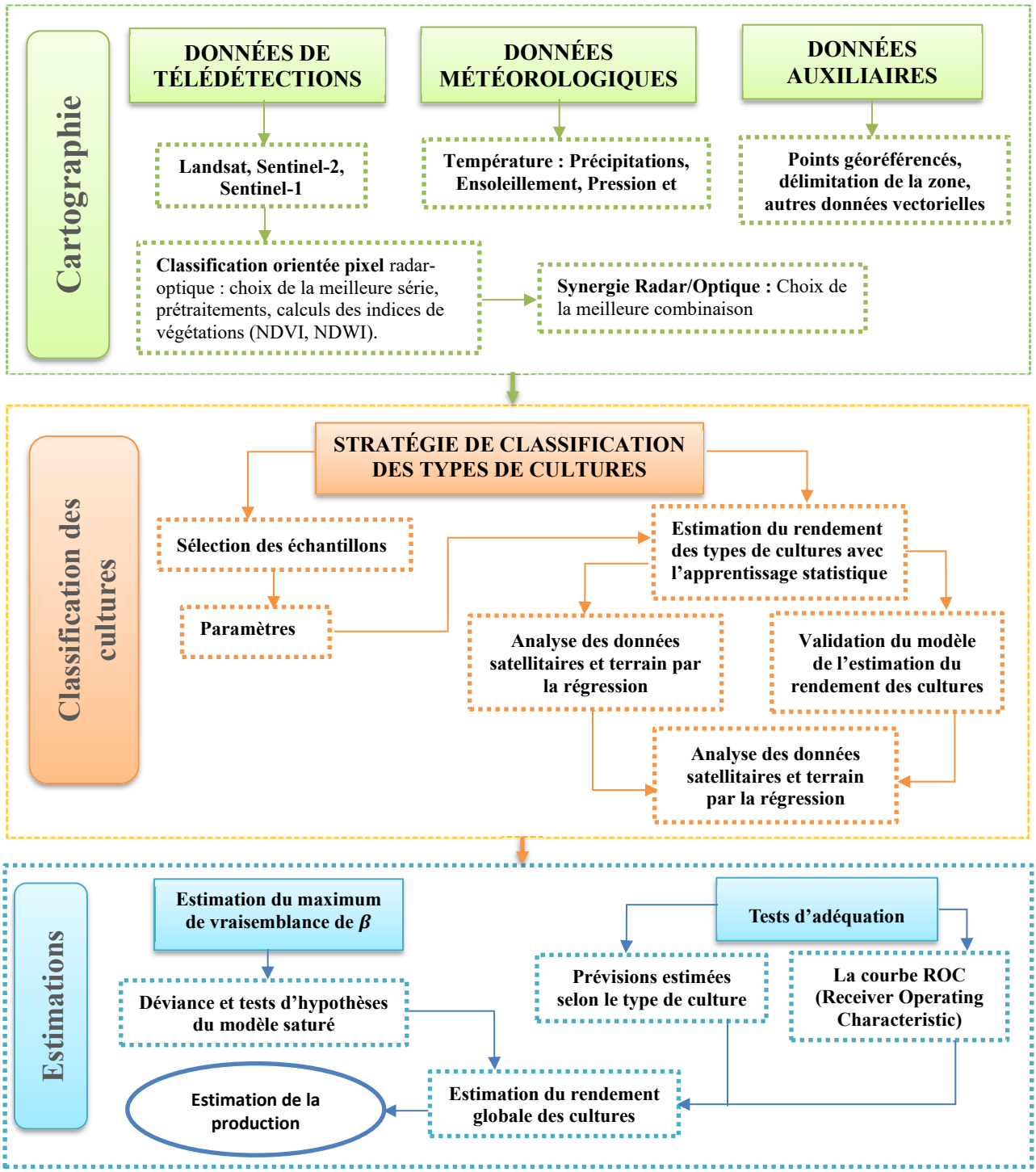


Figure 4 : Organigramme méthodologique

4.1 Étapes méthodologiques

Les étapes méthodologiques comprennent :

1) Proposer un algorithme implémenté dans *Google Earth Engine* (GEE) pour identifier les types de cultures à l'aide des images de télédétection (Sentinel-1, Sentinel-2 et DEM) et des indices dérivés de la saison de croissance dans la période [2016;2020] (Long *et al.*, 2021);

2) Déterminer les rendements des types de cultures à l'aide du changement de végétation et ce grâce à l'analyse superposée de la répartition des gains et pertes de végétation et de la carte de la couverture terrestre de 2020 (Abdulmana *et al.*, 2021).

Les traitements de données réalisés sont décrits dans les sections suivantes.

4.2 Logiciels utilisés

Pour cette étude, nous nous sommes contraints d'utiliser des logiciels Comme QGIS, ArcGIS et R.

ArcGIS est un logiciel utilisé pour le traitement des images issues des satellites Sentine-l Sentinel-2, Landsat-7 et Landsat-8. Les classifications sont réalisées à l'aide de **ArGIS**.

R est un environnement logiciel libre pour le calcul statistique et les graphiques.

QGIS est un logiciel SIG utilisé dans le traitement de fichiers d'images et de données auxiliaires.

Les estimations s'effectuent en 5 étapes avec ces logiciels.

Une première étape est de calculer la moyenne globale et l'écart-type de chaque bande contenue dans une image, et de les enregistrer dans un fichier XML (EXCEL). Celui-ci est alors utilisé dans la seconde étape, qui est une étape d'entrainement. Un modèle d'entraînement est créé à partir d'une image et des données vectorielles d'entraînement. La troisième étape effectue ensuite une classification par rapport au modèle de l'étape précédente. La quatrième a mesuré la qualité de la classification en créant une matrice de confusion à partir de données vectorielles de validation. La dernière étape consiste à évaluer l'estimation de la production de ces échantillons classifiés au moyen d'un modèle de régression logistique.

4.3 Stratégie de classification des types de cultures

La surveillance des changements dans la couverture terrestre dépend de la disponibilité de données de vérité terrain pertinentes pour calibrer et valider les algorithmes de classification. Nous avons

utilisé une classification de forêt aléatoire (RF) supervisée basée sur des pixels pour discriminer les principaux types de cultures dans la zone de terres cultivées comme l'ont proposé Mayer *et al.*, (2021). Malheureusement, les données de vérité au sol spatialement explicite, indispensables, ne sont souvent pas disponibles dans les systèmes agricoles d'Afrique subsaharienne particulièrement le Mali. Ce qui limite le développement d'outils analytiques pertinents pour surveiller la dynamique des terres cultivées en temps réel sur les systèmes agricoles.

Les images de la saison de croissance (mars à août) ont été sélectionnées en fonction de la date dans la base de données de l'échantillon. La bande d'évaluation de la qualité (QA) a été utilisée pour supprimer les nuages des images Sentinel-2 (couverture nuageuse inférieure à 10 %). Nous avons utilisé une fonction dans GEE pour calculer différents indices, y compris le NDVI (graphe en annexe 2), l' indice de végétation de différence normalisé à bord rouge (red-edge), l'indice d'humidité de différence normalisé (NDWI) et l'indice de végétation amélioré (EVI) (He *et al.*, 2021). Par la suite, nous avons extrait les coefficients de rétrodiffusion des données SAR Sentinel-1 en bande C (Syifa *et al.*, 2020). Nous avons également appliqué la fonction ee.Algorithms.Terrain () dans GEE pour calculer la pente à l'aide du SRTM. Les images Sentinel-2, des indices dérivés de Sentinel-2, du coefficient de rétrodiffusion Sentinel-1, du DEM et de la pente ont été fusionnées à l'aide des fonctions merge () et addBands () de GEE. Par la suite, les nouvelles images fusionnées ont été rééchantillonnées à une résolution de 10 × 10 m.

4.4 Sélection des échantillons

Pour donner suite à la définition de ces classes, les échantillons de formation et de validation ont été sélectionnés. Les échantillons sont des zones de contrôle, déterminées ou connues, qui permettent de classer une image. Les échantillons dits d'entraînement sont utilisés pour définir les classes basées sur les valeurs radiométriques des pixels sélectionnés. De plus, des validations sont effectuées pour évaluer la qualité de la classification. La qualité d'une classification supervisée repose dans une large mesure sur un nombre suffisant d'échantillons de pixels .

Pour une classification utilisant l'algorithme Random-Forest, il est recommandé de suivre une certaine méthodologie pour le choix des échantillons d'entraînement et de validation comme suit :

- ✓ ne pas prélever d'échantillons de dimensions trop importantes (relativement à la résolution spatiale).
- ✓ Prendre autant d'échantillons d'entraînement et de validation.

- ✓ Disposer d'une bonne proportion de pixels par classe afin d'éviter une classification biaisée.
- ✓ Choisir des échantillons présentant un minimum de corrélation spatiale.
- ✓ Répéter plusieurs classifications pour évaluer la stabilité dans la prévision de classe.

Grâce à ces mesures, nous avons choisi au hasard et de manière homogène nos 205 échantillons par rapport à l'étendue de notre zone semi-aride. Les données proviennent des enquêtes de conjoncture agricoles réalisées entre [2010 et 2020] par le service de planification et de statistiques du ministère de l'Agriculture du Mali. Afin de sélectionner ces échantillons, nous avons cherché à connaître leurs objets ainsi que leur répartition très pertinente sur l'image.

4.5 Les paramètres Random Forest

Même si les résultats varient peu. Une bonne sélection des paramètres de Random Forest permet de gagner en précision. Nous avons modifié les quatre paramètres de forêt aléatoire jugés importants et comparé les coefficients kappa sur la classification de la superficie de cultures. On a ainsi observé une hausse de 1,4% entre deux réglages. Il s'agit des quatre paramètres suivants:

- Le nombre maximal des arbres (n);
- Le nombre de variables choisies de façon aléatoire pour chaque nœud (m);
- Le nombre minimal d'échantillons utilisés par nœud de l'arbre (c);
- La profondeur de l'arbre (p)

4.6 Estimation du rendement des types de cultures avec l'apprentissage statistique:

La construction d'un modèle de régression logistique est très similaire à celle d'un modèle de régression linéaire. En général, on commence par choisir le lien canonique logit (si les méthodes de validation de modèle indiquent un problème, on pourra tenter de le régler en utilisant un lien différent) (Nattino *et al.*, 2020). Parmi les 205 données résultantes, nous avons déterminé le type de cultures de chaque point d'échantillonnage sur la base d'images haute résolution de *Google Earth*, et 2 300 points d'échantillonnage ont été obtenus avec 307 500 observations. Le modèle de régression logistique a ensuite été alimenté avec l'ensemble des données d'entraînement et l'ensemble des données de validation.

Le calcul des paramètres a été effectué dans SAS et R les valeurs calculées pour tous les paramètres étaient les valeurs moyennes dans la zone d'échantillonnage. Un réseau de neurones avec une régression logistiques (RL) se compose généralement d'une couche d'entrée (données d'entrainement), d'une couche cachée et d'une couche de sortie (données de validation) (Chang,

2020). Le réseau de neurones RL (figure 5) réalise la fonction de mappage de l'entrée à la sortie et peut approximer les fonctions continues non linéaires avec une précision suffisante de 95%. Le réseau à deux couches de neurones avec une fonction d'activation (ou de validation) et des neurones de sortie linéaires, avec des données cohérentes et suffisamment de neurones, peut bien s'adapter au problème de cartographie multidimensionnelle (Ortac et Ozcan, 2021).

Si nous augmentons le nombre de neurones et de couches, les performances du modèle augmentent aussi sur l'ensemble d'apprentissage, mais diminuent sur l'ensemble de test, ce qui indique que le modèle est sur-ajusté en raison d'un sur-paramétrage. L'algorithme d'entraînement de régularisation (RL) est utilisé pour entraîner le réseau de neurones. L'algorithme d'apprentissage de régularisation introduit une fonction de correction à la fonction de performance basée sur la fonction de déviance pendant le processus d'apprentissage du réseau de neurones (He *et al.*, 2021).

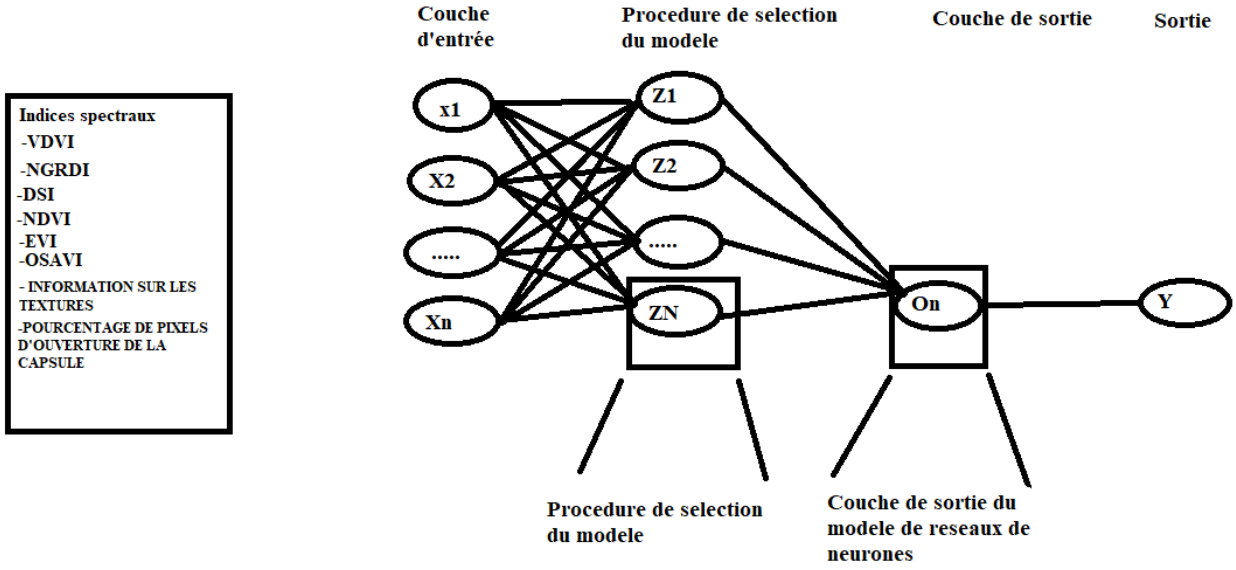


Figure 5 *:* Structure du modèle de réseau de neurones (RL)

4.7 Analyse des données satellitaires et terrain par la régression logistique

Nous considérons maintenant le cas où la variable endogène est $Y_i = {\hat{Y}_i}/{m_i}$, où $\hat{Y}_i$ suit une distribution binomiale (m_i, π_i) qui mesure la probabilité de succès des bandes de réflectance de surface pour des cas d'incident ou d'observations spécifiques. Dans le cas particulier fréquemment rencontré où m_i= 1 et donc Y_i=$\hat{Y}_i$, c-à-d que Y_i prend la valeur 1 avec probabilité π_i et 0 avec probabilité $1 - \pi_i$, on appelle ce type de régression logistique.

Le modèle la probabilité de succès pour les observations spécifiques (Kost *et al.*, 2021)

Dans ce cas, on a que $\mu_i = \mathbb{E}[Y_i] = \pi_i$. La fonction de lien canonique est le lien logit, c.-à-d. $g(u) = \ln\{u/(1-u)\}$, d'où le nom de ce type de régression. Sous le lien logit m_i= 1.

$$\pi_i = P[Yi = 1;\ xi] = g^{-1}(x_i'\beta) \tag{4-1}$$

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p'} x_{ip'})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p'} x_{ip'})} \tag{4-2}$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p'} x_{ip'} \tag{4-3}$$

$$\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p'} x_{ip'}) \tag{4-4}$$

Une première constatation est que l'interprétation de la valeur des coefficients de régression pour le modèle de régression logistique est très différente de celle pour le modèle de régression linéaire multiple.

- Si $\beta_j > 0$, alors une hausse de x_{ij} , avec la valeur de toutes les autres variables exogènes restant inchangée, augmentera la probabilité d'observer un succès (Yi = 1);
- Si $\beta_j < 0$, alors une hausse de x_{ij} , avec la valeur de toutes les autres variables exogènes restant inchangée, diminuera la probabilité d'observer un succès;
- Si $\beta_j = 0$, alors la valeur de x_{ij}, n'a aucun effet sur les chances de succès. Avec x_j nombre de parcelles et x_i de bandes de réflectance de surface de bandes de réflectance de surface.

Inférence de la fonction de log-vraisemblance pour les données du modèle

Supposons un échantillon d'observations indépendantes $(\hat{Y}_i, x_i)$, $i = 1, \ldots, n$ où $\hat{Y}_i \backsim binomiale\ (m_i, \pi_i)$, avec les valeurs de m_i connues. Soit une fonction de lien g(·) (en générale lien logit) et posons $g(\pi_i = x_i'\,\beta) \Leftrightarrow \pi_i = g^{-1}(x_i'\,\beta)$. La fonction de log-vraisemblance pour ces données et ce modèle est :

$$l(\boldsymbol{\beta}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{n} \left\{ ln\binom{m_i}{\hat{y}_i} + \hat{y}_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right) + m_i\, ln(1-\pi_i) \right\} \tag{4-5}$$

Puisque le terme $\sum \binom{m_i}{\hat{y}_i}$ ne dépend pas de π_i, il peut être omis de la fonction de log- vraisemblance sans conséquence.

Estimation du maximum de vraisemblance de $\boldsymbol{\beta}$

Comme π_i est en fait une fonction de β l'estimateur du maximum de vraisemblance de β est la valeur de $\hat{\beta}$ disons qui maximise la log-vraisemblance. La matrice de variance de $\hat{\beta}$ obtenu ci-dessus est estimée comme suit : on prend la matrice de l'opposé de toutes les dérivées mixtes de $l(\boldsymbol{\beta}, \hat{\boldsymbol{y}})$ (Glonek et McCullagh, 1995). Les intervalles de confiance sont ensuite obtenus par la méthode de *Wald*. Soit $\hat{\beta}_j$ l'estimateur de β_j $\widehat{Var}(\beta_j)$, l'estimateur de variance de $\hat{\beta}_j$ (dont la racine carrée est souvent appelée erreur type, ou directement de l'anglais "erreur standard"). L'intervalle de confiance de Wald de niveau $(1-\alpha)100\%$ pour β_j est donné par $\hat{\beta}_j \pm z_{\alpha/2}\sqrt{\widehat{Var}(\beta_j)}$, où $P\left[|N(0,1)| > z_{\alpha/2}\right] = \alpha$.

Déviance et tests d'hypothèses du modèle saturé

La valeur maximale de la fonction de log-vraisemblance du modèle saturé (avec n paramètres) est obtenue lorsque $\pi_i = y_i/m_i \equiv \hat{\pi}_i$. La fonction de déviance évaluée en π_i est donc obtenue ainsi :

$$\begin{aligned} D(\tilde{y};\pi) &= 2\{l(\tilde{\pi};\tilde{y}) - l(\pi;\tilde{y})\} \\ &= 2\sum_{i=1}^{n}\left\{\hat{y}\ln\left(\frac{\tilde{y}_i}{\pi_i}\right) + (m_i - \tilde{y}_i)\ln\left(\frac{m_i - \tilde{y}_i}{m_i - \pi_i}\right)\right\} \end{aligned} \tag{4-6}$$

Prévisions estimées selon le type de culture

Supposons que nous voulons avoir une idée de la probabilité de succès, disons π_0 pour un individu ayant x_0 comme vecteur de variables exogènes. La prévision ponctuelle pour cette probabilité est tout simplement son estimateur du maximum de vraisemblance, soit $\tilde{\pi}_i = g^{-1}(x_0'\beta)$. Pour avoir un intervalle de confiance, on peut utiliser la méthode de Fieller (Kost *et al.*, 2021).

Soit $v^2(x_0) = \widehat{Var}(x_0'\tilde{\beta}) = \widehat{Var}\left(\sum_{j=0}^{p'} x_{0j}\tilde{\beta}_j\right)$

$v^2(x_0) = \widehat{Var}(\tilde{\beta}_0) + x_{01}^2\widehat{Var}(\tilde{\beta}_1) + \ldots + \widehat{Var}(\tilde{\beta}_{p'}) + 2x_{01}\widehat{Cov}(\tilde{\beta}_0, \tilde{\beta}_1) + \ldots + 2x_{0p'-1}x_{0p'}\widehat{Cov}(\tilde{\beta}_{p'-1}, \tilde{\beta}_{p'})$

$$v^2(x_0) = \widehat{Var}(\hat{\beta})x_0 \tag{4-7}$$

Alors l'intervalle de confiance de niveau $(1-\alpha)100\%$ pour π_0 est l'ensemble des valeurs de π telles que : $\left|\frac{x_0'\hat{\beta} - g(\pi)}{v(x_0)}\right| \leq z_{\alpha/2}$.

En fait, cet intervalle est obtenu en construisant un intervalle de confiance de la forme $x_0'\hat{\beta} \pm z_{\alpha/2}\sqrt{v^2(x_0)}$ pour $x_0'\beta$, et ensuite en transformant l'intervalle en intervalle pour π en utilisant le fait que $\pi_i = g^{-1}(x_i'\,\beta)$.

Ensuite, on utilise une procédure de sélection de modèle algorithmique (inclusion, exclusion ou pas-à-pas) pour choisir les variables exogènes qui ont un effet sur la distribution de la variable endogène. Une fois ces variables choisies, on regarde si des termes de degré supérieur (interactions, polynômes, etc.) sont nécessaires en effectuant des tests d'hypothèses appropriés.

4.8 Validation du modèle de l'estimation du rendement des cultures

Tests d'adéquation (où de validation)

Il y a en général trois tests d'adéquation effectués par les logiciels lorsque l'on ajuste le modèle de régression logistique. Le premier est le test d'ajustement du khi-deux de Pearson. Pour calculer cette statistique de test, on doit tout d'abord obtenir les résidus de Pearson.

$$r_{p_i} = \frac{\tilde{y}_i - m_i\hat{\pi}_i}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{4-8}$$

Ou $\hat{\pi}_i = g^{-1}(x_i'\hat{\beta})$. On pose ensuite $X^2 = \sum_{i=1}^{n} r_{p_i}^2$.

- Pour le premier test si le modèle est bon, alors on s'attend à ce que X^2 suive, approximativement, une distribution du khi-deux avec $I - p$ degrés de liberté.
- Le second test est basé sur la statistique de déviance. On préfère donc tester l'ajustement du modèle avec la statistique du khi-deux de Pearson et on utilise la statistique de déviance pour les tests du rapport des vraisemblances.
- Le troisième test est le test d'ajustement d'Hosmer et Lemeshow. Il confronte l'hypothèse nulle que le modèle est adéquat à la contre-hypothèse que le modèle n'est pas adéquat. L'idée derrière ce test est de partitionner l'échantillon en quelques sous-échantillons sur la base des valeurs prédites $\hat{\pi}_i$, ensuite de comparer la moyenne des $m_i\hat{\pi}_i$ et la moyenne des y_i. Dans chaque sous-échantillon. Si la différence est faible, le modèle est approprié. Si la différence est forte, le modèle doit être amélioré.

La courbe ROC (Receiver Operating Characteristic)

Elle représente la sensibilité en fonction de 1-specificite pour toutes les valeurs seuils possibles. La sensibilité est la capacité du test à bien détecter les indices spectraux de végétation (IV) et la spécificité est la capacité du test à bien détecter les non indices spectraux.

Le fonctionnement de la courbe ROC est le suivant : pour chaque valeur de u dans [0, 1], effectuer les opérations suivantes :

1. Pour chaque individu, poser $Y_i^* = 1$ si $\hat{\pi}_i \geq u$ et $Y_i^* = 0$ si $\hat{\pi}_i < u$.
2. Calculer o_u, le taux de vrais positifs et a_u, le taux de faux positifs.
3. Sur un graphe, mettre un point ayant a_u comme abscisse et o_u comme ordonnée.

Lorsque tous les points sont placés, on obtient la courbe ROC pour le modèle en question. La courbe ROC part du point (0,0) et se termine au point (1,1). Elle se situe entre la courbe idéale qui fait l'angle droit de (0,0) à (0,1) à (1,1) et le scénario que l'on obtiendrait si on assignait les 0 et les 1 complètement au hasard, soit une droite à 45 degrés de (0,0) à (1,1) (Fanjul-Hevia *et al.*, 2021).

L'idée est que plus l'aire sous la courbe est grande, meilleur est le modèle. Règle du pouce : une aire sous la courbe ROC entre 1 et 0.9 veut dire un excellent modèle, une aire entre 0.9 et 0.8 signifie un bon modèle, une aire entre 0.8 et 0.7 est un modèle moyen, entre 0.7 et 0.6 est un modèle mauvais et entre 0.6 et 0.5 le modèle est très faible.

4.9 Estimation de la production globale des cultures

Les incertitudes liées au classement des types de cultures et à l'estimation du rendement sont combinées. Les estimations de production non biaisées à l'aide d'une approche de régression logistique avec 205 réalisations sont présentées dans la section « Résultats ». Pour chaque pixel classé comme type de culture, chaque réalisation décrit un type de culture aléatoire lié à la probabilité de succès de classification. La matrice de confusion qui en découle sert à appliquer le modèle de régression logistique du rendement correspondant, avec un terme d'erreur gaussienne à moyenne négligeable. Parmi les 205 données résultantes, nous avons déterminé le type d'occupation du sol de chaque point d'échantillonnage (voir Annexe 1) sur la base d'images haute résolution de Google Earth, et 2 300 points d'échantillonnage ont été obtenus avec 307 500 observations. Les indices de végétation du tableau 2 ont été également ajoutés dans l'analyse.

Tableau 2 *:* Indices de végétation utilisés

Indices utilisés	**Formules**
Indice de végétation par différence de bande visible (VDVI) (Wang *et al.*, 2018)	$\frac{2G - R - B}{2G + R + B}$
Indice de différence vert-rouge normalisé (NGRDI) (Barrero et Perdomo, 2018)	$\frac{G - R}{G + R}$
Indice spectral de différence (DSI) (Feng *et al.*, 2020)	$NIR - R$
Indice de végétation par différence normalisée (NDVI) (Ni *et al.*, 2019)	$\frac{NIR - R}{NIR + R}$
Indice de végétation amélioré (EVI) (Misra *et al.*, 2020)	$\frac{2{,}5(NIR - R)}{NIR + R + 1}$
Indice de végétation optimisé ajusté au sol (OSAVI) (Zou et Mõttus, 2017)	$\frac{1{,}16(NIR - R)}{NIR + R + 0{,}16}$

5. Résultats

Nous avons donc procédé à une classification des types de cultures à l'aide d'arbres de classification et de régression (CART), puis la construction d'un modèle de régression utilisant la stratégie d'apprentissage statistique. En dernier lieu, le LAI est utilisé comme prédicteur du rendement final.

5.1 Stratégie de classification des types de cultures

Pour la classification des types de cultures, nous avons appliqué deux classificateurs : les arbres de classification et de régression (CART) et les forêts aléatoires (RF), qui conviennent tous deux à la classification catégorielle, et qui ont été utilisés dans divers contextes pour la classification (Sahani et Ghosh, 2021). En comparant les sorties de CART et RF, nous avons fait une inférence avec le classificateur le plus précis, RF. Les paramètres par défaut sont inchangés, notamment les ajustements tels que l'optimisation du nombre d'arbres dans RF.

Dans le tableau 3, nous avons une très forte corrélation entre le stock de carbone organique du sol (SOC) et le réseau de transformation spatiale (STN). Ceci aide à l'identification de la culture et à la normalisation de l'échelle de la région concernée, ce qui peut simplifier la tâche de classification subséquente et améliorer l'estimation du rendement. Des données d'élévation provenant DEM-SRTM ont été importées de http://srtm.csi.cgiar.org/. Ce sont des données d'élévation avec une résolution de 90 m. Enfin, nous avons analysé les statistiques relatives à la conductivité électrique du sol (EC).

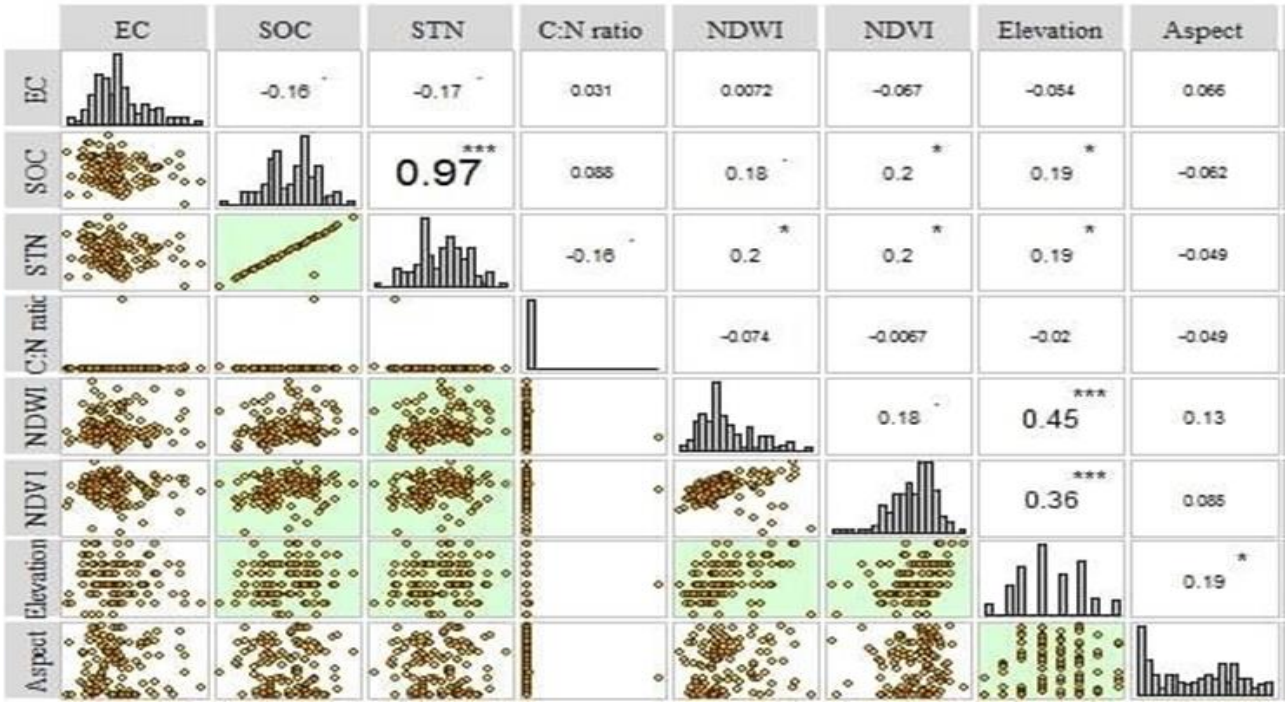


Figure 6 : Précision de classification des trois indices basés

Les caractéristiques de la figure 6 ont été examinées préalablement sur la base de la séparabilité des classes de cultures, mais il s'est avéré nécessaire de construire en plus un ensemble complémentaire de caractéristiques, qui ont été finalement utilisées pour la classification. Nous avons divisé l'ensemble des caractéristiques ajoutées pour la classification en quatre groupes : (a) bandes spectrales de Sentinel-2; (b) indices dérivés de Sentinel-2; (c) élévation et pente à partir des données DEM; et (d) coefficients de rétrodiffusion VV et VH de Sentinel-1.

La carte annuelle des types de cultures (maïs, millet, sorgho et fonio) couvrant plusieurs années a été utilisée pour créer les surfaces des principales cultures de la région de Koutiala (Figure 7). Il s'agit de la carte des couvertures propres aux cultures qui sont indépendantes de la saison agricole actuelle. Mais, les tendances générales des périodes de récolte des surfaces cultivées ne sont pas évidentes dans le masque global des cultures.

Les pixels à NDVI élevé dans la zone en croissance présentent une fréquence temporelle plus élevée et vice-versa. Par conséquent, les parcelles situées dans la zone primaire de la culture cible ont reçu des valeurs de densité céréalière supérieures, contrairement aux régions non agricoles telles que les grandes étendues d'eau, les forêts, les zones urbaines.

Le résultat final de la trame de densité des grains représente des parcelles avec l'agriculture familiale où, dans le temps et dans l'espace, on a observé la culture cible. La carte de densité spatiale représente la probabilité que les types de cultures aient lieu (Figure 7). Les cartes de densité des cultures peuvent être mises à jour annuellement avant la campagne agricole.

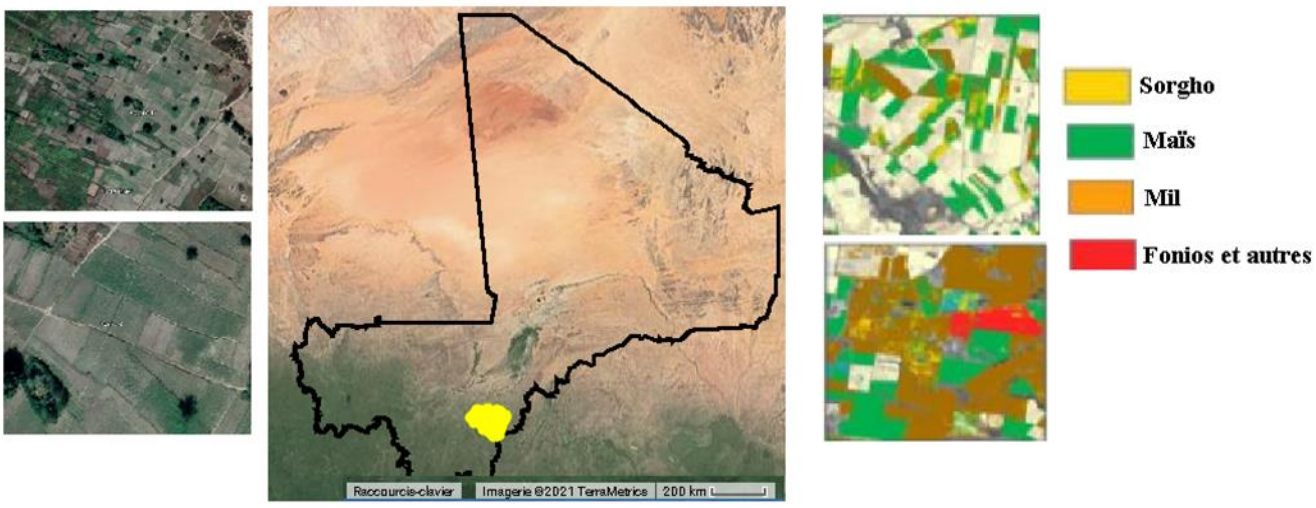


Figure 7 : Image de *Google Earth* sur les cartes de densité des cultures

5.2 Les paramètres Random Forest

Tout d'abord, tout en conservant les réglages définis par défaut pour les autres paramètres. Nous avons testé initialement le paramètre n sur le nombre maximum d'arbres. Nous avons effectué nos essais pour 10, 20, 30, 40 et 50. Nous avons constaté une hausse du kappa de 0,789 pour n=10 à 0,851 pour n=40. Deuxièmement, il n'y a pas de changement important dans le kappa des autres valeurs. Nous concluons que la valeur optimale est 40, les résultats se dégraderait et au-delà et la puissance de calcul de la machine serait nécessaire inutilement. Troisièmement, nous avons constaté que la variation des paramètres m et k n'a pas modifié de manière significative les résultats. C'est pourquoi nous avons décidé de conserver m=2 et k=10 par défaut. Enfin, dans un quatrième pas, on a changé la profondeur de l'arbre. On a vérifié les valeurs 10, 20, 30 et 40. Il est apparu que nous atteignons un kappa maximum de 0.891 avec p=30, pour 0.864 avec p=20. Les paramètres que nous avons choisi de maintenir sont par conséquent : n=40 ; m=2 ; k=20 ; p=30.

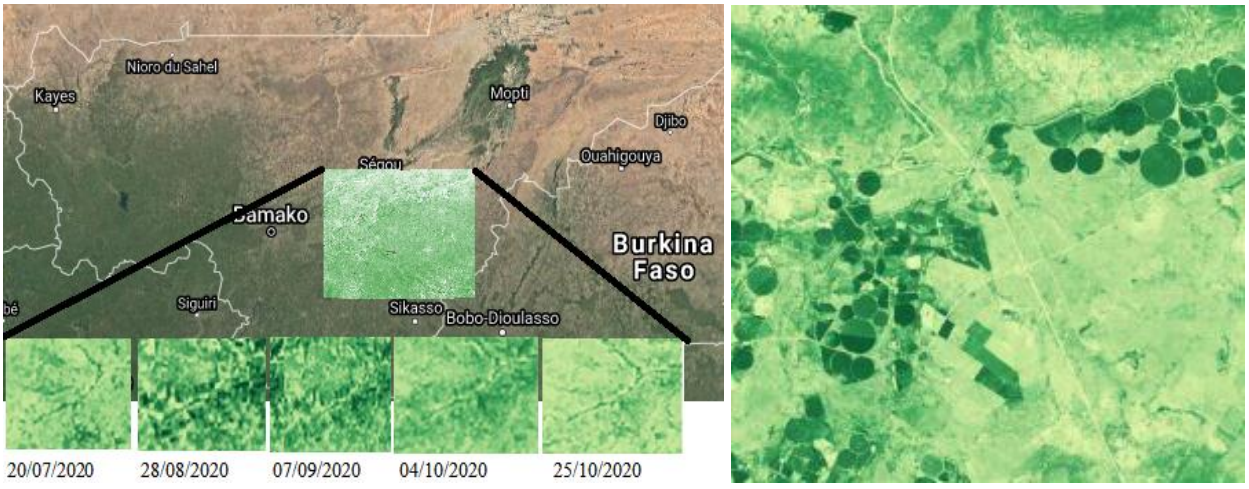
5.3 Estimation du rendement des types de cultures avec l'apprentissage statistique

La construction d'un modèle de régression logistique est très similaire à celle d'un modèle de régression linéaire. En général, on commence par choisir le lien canonique logit. Si les méthodes

de validation de modèle indiquent un problème, on pourra tenter de le régler en utilisant un lien différent comme probit (Nattino *et al.*, 2020). Le calcul des paramètres a été effectué dans SAS et R, les valeurs calculées pour tous les paramètres étaient les valeurs moyennes dans la zone d'échantillonnage. Un réseau de neurones avec une régression logistiques (RL) se compose généralement d'une couche d'entrée, d'une couche cachée et d'une couche de sortie (Chang, 2020). Le réseau de neurones RL réalise la fonction de *mappage* de l'entrée à la sortie et peut approximer les fonctions continues non linéaires avec une précision suffisante. Le réseau à deux couches de neurones avec une fonction d'activation et des neurones de sortie linéaires, avec des données cohérentes et suffisamment de neurones, peut bien s'adapter au problème de cartographie multidimensionnelle (Ortac et Ozcan, 2021). L'algorithme d'apprentissage de régularisation introduit une fonction de correction à la fonction de performance basée sur la fonction de déviance pendant le processus d'apprentissage du réseau de neurones (He *et al.*, 2021).

5.4 Estimation de l'indice de surface foliaire (*leaf area index*, LAI)

Divers indices spectraux de végétation et estimations de l'indice de surface foliaire (LAI) ont été utilisés dans cette étude comme prédicteurs du rendement final et de la production globale en grains, séparément pour chaque culture. Ceux-ci incluent les métriques de la chlorophylle de la canopée, qui se sont avérées plus sensibles aux valeurs élevées de LAI que le NDVI (figure 8).



(a) **(b)**

Figure 8 : ***(a)*** L'indice de surface foliaire (LAI) entre [2016;2020] et **(b)** la zone d'étude avec les champs des 4 types de cultures. Source (https://code.earthengine.google.com/).

Sur les 205 données résultantes, le type de culture à chaque lieu d'échantillonnage a été déterminé. Basé sur des images haute résolution de Google Earth, et 2300 points de prélèvement. Les 2/3 ont été utilisés pour l'entraînement et le 1/3 pour la validation ont été obtenus avec 307500 observations (figure 9). Les images de sortie prennent la valeur 1 pour les quatre types de cultures, affiché en jaune sur la carte, et 0 pour « autres », affiché en gris sur la carte.

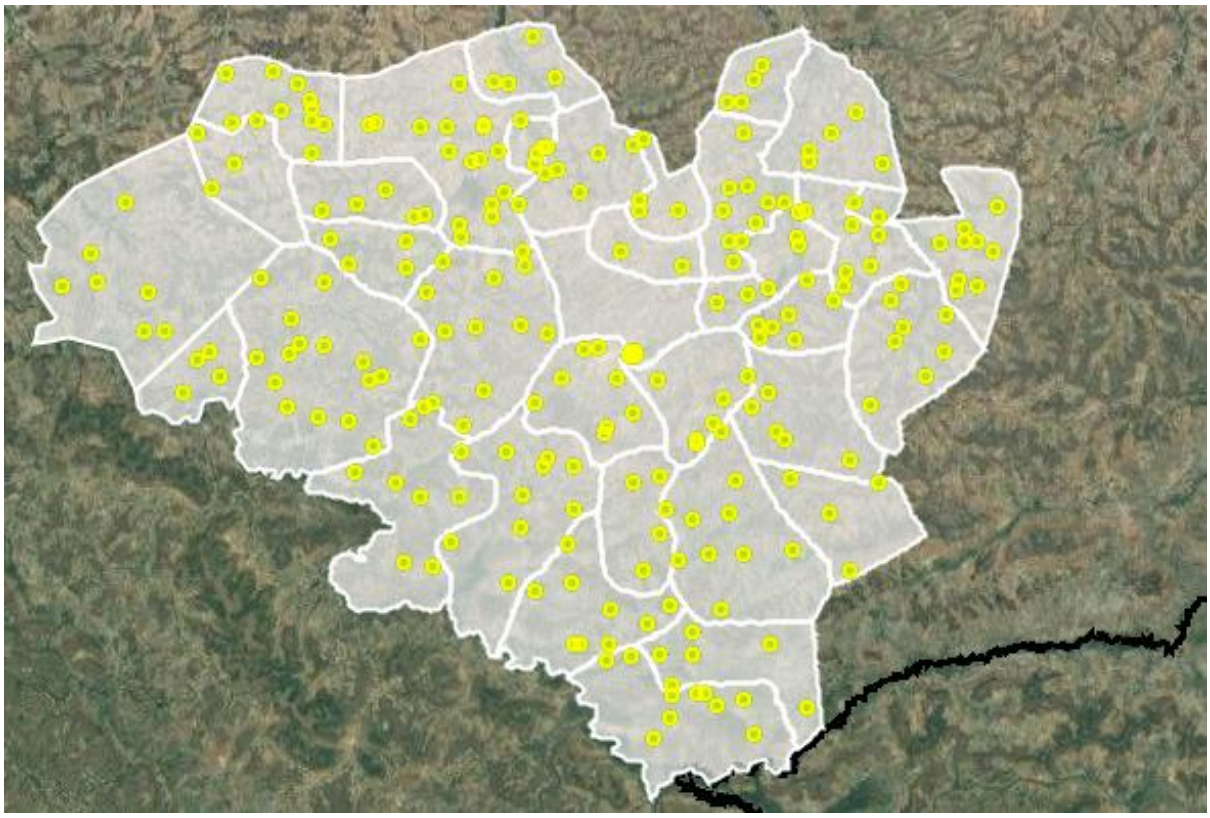


Figure 9 : Points d'échantillonnage [2016;2020]. Source : https://code.earthengine.google.com/

5.5 Analyse des données satellitaires et terrain par la régression logistique

Description des variables

Nous avons étudié la fréquence des indices spectraux sur la croissance de la végétation ainsi que le rendement dans une parcelle de culture cible ou non. Nous disposons de 205 données au sol de référence dans 30 villages. La variable dépendante (ou expliquée) Y prend la valeur 1 si la croissance de la végétation dans une parcelle fait référence aux divers indices spectraux de végétation et 0 sinon. Les variables exogènes (ou explicatives) sont x_1 (vaut 1 si la parcelle compte plusieurs indices spectraux de végétations élevés et 0 sinon) et x_2 (vaut 1 si la parcelle

compte un seul indice spectral élevé et 0 sinon). Dans le tableau 3, **i :** représente le village i, **n_i :** nombre de types de cultures par village i, **tot :** nombre total de types cultures, **x_1 :** nombre de bandes de réflectance de surface (par pixels) (1=plusieurs ; 0=sinon), **x_2 :** nombre de parcelles (1= plus de 2 ; 0=sinon), **y :** mesure de la bande de réflectance de surface (pixels d'une parcelle) du village i (1=croissance ; 0=sinon).

Tableau 3 : Description des données sortie SAS

Liste alphabétique des variables et des attributs			
#	**variable**	**Type**	**Long.**
1	i	Num	8
2	n_i	Num	8
3	tot	Num	8
4	X_1	Num	8
5	X_2	Num	8
6	y	Num	8

Pour rappel, notre objectif est d'abord de construire un modèle de régression logistique qui prédit Y à l'aide du nombre de bandes de réflectance de surface et le nombre de parcelles. Premièrement, nous avons analysé notre modèle par la régression logistique standard (tests au seuil 10%). Ensuite, nous avons ajouté une interaction entre le nombre de bandes de réflectance de surface et le nombre de parcelles, pour voir si le modèle s'ajuste mieux et pour le conserver. Enfin, nous avons calculé des intervalles de confiance à 90% avec l'écart de prédiction pour la production des types de cultures.

Ajustement du modèle de régression logistique avec des effets simples

Dans le tableau 4, les propriétés des estimateurs (rapport de vraisemblance, score et Wald) sont très proches pour les deux types de données le nombre de bandes de réflectance de surface et le nombre de parcelles. Néanmoins, certains concepts tels que la forme de la vraisemblance ou les tests d'adéquation par la déviance peuvent légèrement différer. Deux algorithmes sont généralement implémentés sur les logiciels de statistique (particulièrement SAS) pour calculer les estimateurs du maximum de vraisemblance : l'algorithme du score de Ficher et l'algorithme de Wald. Avec le test du score, on cherche ici à vérifier si la fonction de score (gradient de la log-vraisemblance) est proche de 0 sous H_0. Ces trois tests acceptent l'hypothèse nulle (Pr>khi-2 avec $p<0.001$).

Tableau 4 : Test de l'hypothèse nulle globale : Beta=0

Test de l'hypothèse nulle globale : Beta=0			
Test	**Khi-2**	**DDL**	**Pr>khi-2**
Rapport de vraisemblance	57,9422	3	p<0,001
Score de Fisher	54,4426	3	p<0,001
Wald	46,5899	3	p<0,001

Les variables explicatives du nombre de bandes de réflectance de surface et du nombre de parcelles sont significatives au seuil de 10% (tableau 5). Les tests globaux sont significatifs au seuil de 10%. Donc le nombre de pixels par parcelle est utile à la prédiction de la croissance de la végétation d'une parcelle.

Tableau 5 : Estimations par l'analyse du maximum de vraisemblance

Estimations par l'analyse du maximum de vraisemblance					
Paramètre	**DDL**	**Estimation**	**Erreur type**	**Khi-2 de Wald**	**Pr>Khi-2**
Intercept	1	-1,6033	0,2633	37,0854	<0,0001
X_1	1	0,6908	0,4131	2,7954	0,0945
X_2	1	2,2307	0,3334	44,772	<0,0001

Le modèle obtenu est donc :

$$P[Y_i = 1; x_1, x_2] = \frac{exp[-1.6033 + 0.6908x_1 + 2.2307x_2]}{1+exp[-1.6033 + 0.6908x_1 + 2.2307x_2]}$$

Interprétation : Le tableau 6 montre que le coefficient 0.6908 du nombre de bandes de réflectance de surface est positif, l'augmentation d'une unité d'indice spectral étant donné que le nombre de parcelles restent inchangées, multiplie la côte par 1.995311 (voir annexe), soit une augmentation de 99%. Le coefficient 2.2307 du nombre de parcelles est positif, l'augmentation d'une unité du nombre de parcelles, étant donné que divers indices spectraux sont inchangés, multiplie la côte par 1.887631, soit une augmentation de 88%. Il est facile de voir que les répartitions des parcelles selon la variable Y ne sont pas les mêmes dans les communes et dans l'échantillon. Ceci va entrainer un biais au niveau des estimateurs.

Qualité d'ajustement : Cette association de probabilités nous permet de disposer de statistiques sur l'exactitude des prédictions $\hat{Y}_i$. Le pourcentage de couples (du nombre de bandes de réflectance de surface, du nombre de parcelles) correspondants est de 65.5 et le pourcentage de non-concordances est de 11.1 (tableau 6). Nous avons le pourcentage relié 11.1 qui est le pourcentage

de paires non déterminées et le nombre de paires 10126. Les 3 indices (Somers' D, Gamma et de τ_a de Kendall) de la justesse des prédictions jouent le même rôle. A part le τ_a de Kendall, les autres mesures sont proches de 1. Donc il y'a une concordance entre les probabilités prédites et les réponses observées.

Tableau 6 : Association des probabilités prédites et des réponses observées

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	65.5	**D de Somers**	0.545
Pourcentage discordant	11.1	**Gamma**	0.711
Pourcentage lie	23.4	**Tau-a**	0.264
Paires	10126	**C**	0.772

5.6 Validation du modèle l'estimation du rendement des cultures sans interaction

L'aire sous la courbe ROC nous donne 0.7724, ce qui signifie que le modèle est moyen (figure 10). Le test d'ajustement de modèle de Hosmer-Lemeshow donne 0.1291 qui est supérieur à 0.1 (tableau 7), ce qui veut dire que le modèle semble bien s'ajuster aux données étudiées. Le test Hosmer & Lemeshow se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2 à $K-1$ degrés de liberté avec :

$$\pi_i = \frac{exp[-1.6033 + 0.6908x_1 + 2.2307x_2]}{1+exp[-1.6033 + 0.6908x_1 + 2.2307x_2]}.$$

Tableau 7 : *Test d'adéquation de Hosmer et de Lemeshow sans interaction*

Test d'adéquation de Hosmer et de Lemeshow		
Khi-2	**DDL**	**Pr>Khi-2**
4.0937	2	0.1291

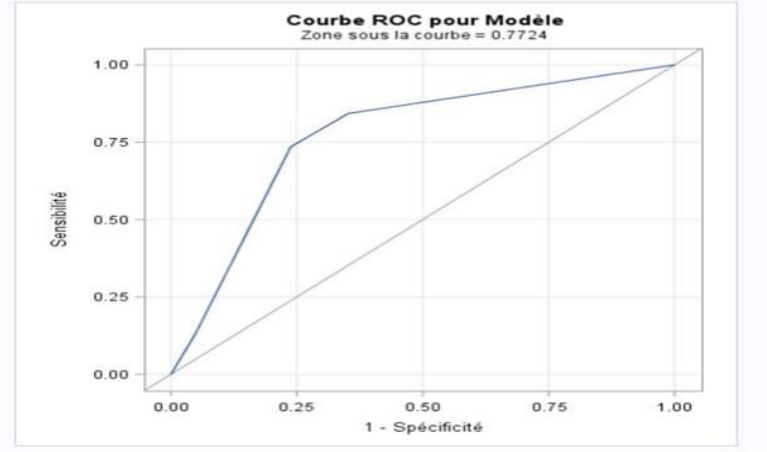


Figure 10 : Courbe ROC pour le modèle sans interaction des indices spectraux

La déviance et la statistique du khi-deux (X^2) de Pearson : Pour le modèle sans interaction, le calcul de la déviance et la statistique du χ^2 de Pearson donne avec les équations (4-5) et (4-6):

$$Deviance = 2\sum_{i=1}^{n}\left\{\hat{y}\ln\left(\frac{\tilde{y}_i}{\pi_i}\right) + (m_i - \tilde{y}_i)\ln\left(\frac{m_i - \tilde{y}_i}{m_i - \pi_i}\right)\right\}$$

$$r^2 = \left(\frac{\tilde{y}_i - m_i \times \pi_i}{\sqrt{m_i \times \pi_i(1-\pi_i)}}\right)^2.$$

La déviance constitue un écart en termes de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré (tableau 8-a). Elle compare la vraisemblance obtenue à celle d'un modèle de référence, c'est à dire, le modèle complet ou saturé (tableau 8-b). Ce modèle ne place pas de contrainte sur la forme du paramètre (du nombre de bandes de réflectance de surface, x_2). Nous savons que :

- Si la déviance est grande, alors le modèle considéré est loin du modèle saturé et que par conséquent il n'ajuste pas bien les données ;
- En revanche, si la déviance est proche de 0, le modèle considéré sera adéquat.

Tableau 8-a : Statistique d'adéquation de la déviance de Pearson du modèle sans interaction

X_1	X_2	$\hat{y}_i$	m_i	$m_i - \hat{y}_i$	π_i	$m_i \times \pi_i$	$m_i \times (1-\pi)_i$	Déviance	X^2 de Pearson
0	0	13	92	79	0, 167	15, 41	76, 58	0, 47	0, 453415778
0	1	50	73	23	0, 65	47, 58	25, 41	0, 35	0, 35099642
1	0	9	23	14	0, 28	6, 58	16, 41	1, 16	1, 236155968
1	1	11	17	6	0, 78	13, 41	3, 58	1, 80	2, 053083533
			205		1, 89			3, 80	4, 093651699

Tableau 8-b : Statistique d'adéquation de la déviance de Pearson du modèle sans interaction – sortie SAS

Statistique d'adéquation de la déviance de Pearson				
Critère	**Valeur**	**DDL**	**Valeur/DDL**	**Pr>Khi-2**
Ecart	3.8008	1	3.8008	p<0.0512
Pearson	4.0937	1	4.0937	p<0.0430

Donc on aura :

Déviance = 0, 474023618 + 0, 356585305 + 1, 163525106 + 1, 806662467 = 3, 800796496

r_{pi}^2= 0, 453415778 + 0, 35099642 + 1, 236155968 + 2, 053083533 = 4, 093651699

Il est difficile de se faire une idée sur l'ajustement en se basant sur la valeur de la vraisemblance puisqu'elle dépend (entre autres) de la taille de l'échantillon. La déviance est égale à 2 multiplié par une différence de log-vraisemblance. Le résultat avec la sortie de SAS :

- La déviance du modèle est 3.8008.
- La statistique de χ^2 de Pearson est 4.0937

Conclusion sur les effets simples : La déviance du modèle est 3.8008. La statistique de χ^2 de Pearson est 4.0937 et le rapport de déviance au nombre de degrés de liberté est 1.101901 qui est proche de 1 (Tableaux 9-a et -b). De plus le test de χ^2 de Pearson est significatif au seuil de 10%, ce qui indique un bon ajustement du modèle.

a. Ajustement du modèle avec une interaction entre x_1 et x_2.

On rappelle que deux variables interagissent si l'effet de l'une sur Y diffère suivant les valeurs de l'autre. Bien entendu, l'ajout d'une interaction augmente la dimension explicative du modèle. Avec les equations (4-5) et (4-6):

$$Deviance = 2\sum_{i=1}^{n}\left\{\hat{y}\ln\left(\frac{\tilde{y}_i}{\pi_i}\right) + (m_i - \tilde{y}_i)\ln\left(\frac{m_i - \tilde{y}_i}{m_i - \pi_i}\right)\right\}$$

$$r^2 = \left(\frac{\tilde{y}_i - m_i \times \pi_i}{\sqrt{m_i \times \pi_i(1-\pi_i)}}\right)^2$$

Pour mesurer l'ajustement d'un modèle sans interaction, nous l'avons comparé au modèle saturé (c'est-à-dire modèle avec interaction) en effectuant un test de rapport de vraisemblance. Ce test n'est valable qu'en présence de données répétées. Cependant, dans notre cas, nous avons écrit les hypothèses de ce test en termes de nullité de certains coefficients du modèle saturé (tableaux 9-a

et 9-b). La déviance du modèle est de 0.0000 ainsi que la statistique de $\chi 2$ de Pearson est 0.0000. Les variables explicatives x_1, x_2 et $x_1 \times x_2$(interaction entre le nombre de bandes de réflectance de surface et le nombre de parcelles) sont significatives au seuil de 10%. La déviance est proche de 0, le modèle considéré sera adéquat.

Tableau 9-a : Statistique d'adéquation de la déviance de Pearson du modèle avec interaction

X_1	X_2	$\hat{y}_i$	m_i	$m_i - \hat{y}_i$	π_i	$m_i \times \pi_i$	$m_i \times (1-\pi)_i$	Déviance	X^2 de pearson
0	0	13	92	79	0, 14	12, 99	79, 00	2, 52E − 11	2, 52E – 11
0	1	50	73	23	0, 68	49, 99	23, 00	1, 30E − 08	1, 30E – 08
1	0	9	23	14	0, 39	9, 00	13, 99	5, 87E − 09	5, 87E – 09
1	1	11	17	6	0, 64	10, 99	6, 00	4, 97E − 09	4, 97E – 09
			205		1, 86			2, 39E − 08	2, 39E – 08

Tableau 9-b : Statistique d'adéquation de la déviance de Pearson du modèle avec interaction – sortie SAS

Statistique d'adéquation de la déviance de Pearson				
Critère	**Valeur**	**DDL**	**Valeur/DDL**	**Pr>Khi-2**
Écart	0.0000	0	-	-
Pearson	0.0000	0	-	-

Donc on aura :

Déviance = 2, 5296E − 11 + 1, 3057E − 08 + 5, 87661E − 09 + 4, 97679E − 09 = 2, 39357E – 08

r_{pi}^2 = 2, 52844E − 11 + 1, 3057E − 08 + 5, 87659E − 09 + 4, 97677E − 09 = 2, 39356E – 08

Le modèle obtenu est donc (tableau 10) :

$$P[Y_i = 1; x_1, x_2] = \frac{exp[-1.8045 + 1.3627x_1 + 2.581x_2 - 1.5331x_1x_2]}{1 + exp[-1.8045 + 1.3627x_1 + 2.581x_2 - 1.5331x_1x_2]} \qquad (5\text{-}1)$$

Tableau 10 : Estimations par l'analyse du maximum de vraisemblance

Estimations par l'analyse du maximum de vraisemblance					
Paramètre	**DDL**	**Estimation**	**Erreur type**	**Khi-2 de Wald**	**Pr>Khi-2**
Intercept	1	-1,8045	0,2993	36,3493	p<0,0001
X1	1	1,3627	0,5217	6,8236	p<0,0090
X_2	1	2,5810	0,3912	43,5236	p<0,0001
X_1*X_2	1	-1,5331	0.7702	3,9622	p<0,0465

5.7 Validation du modèle l'estimation du rendement des cultures avec interaction

L'aire sous la courbe ROC nous donne 0.7771, ce qui signifie que le modèle est moyen (figure 11). Cependant, cette aire est plus petite que celle du modèle sans interaction. Le test d'ajustement du modèle de Hosmer-Lemeshow donne une p-value de 0.095 qui est inférieur à 0.05 (Tableau 11), ce qui veut dire que le modèle semble bien s'ajuster aux données étudiées.

Tableau 11 : *:* Test d'adéquation de Hosmer et de Lemeshow avec interaction.

Test d'adéquation de Hosmer et de Lemeshow		
Khi-2	**DDL**	**Pr>Khi-2**
2.706	2	p<0.095

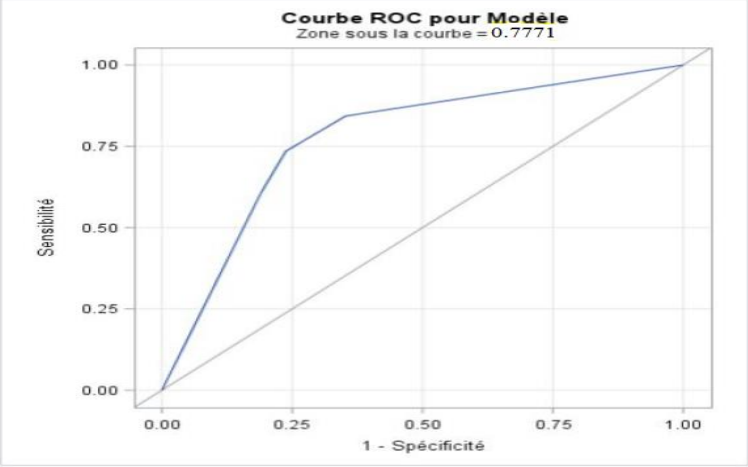


Figure 11 : Courbe ROC pour le modèle avec interaction des indices spectraux.

Conclusion sur effets avec interactions : Ce modèle revient à considérer l'interaction entre les variables parcelles et pixels. Bien entendu, l'ajout d'une interaction augmente la dimension explicative du modèle. Le test Hosmer-Lemeshow donne une p-value proche de 1, donc le modèle avec interaction s'ajuste mieux que celui sans interaction. Nous calculons la différence entre la déviance du modèle sans interaction et la déviance avec interaction. Nous obtenons 3.801 (Tableaux 8 vs 9). Le khi-deux à un degré de liberté au seuil 10% est de 2.706 (Tableau 11), qui est supérieur à la différence des déviances. La précision est utilisée pour évaluer la performance du modèle. Le modèle d'interaction a été le plus performant, avec une précision pour l'entraînement de 85.2%, et de 88.6% pour la validation. Quoi qu'il en soit, le modèle sans interaction fonctionne bien malgré cela. Cependant, on a par la suite choisi le modèle d'interaction pour l'analyse de l'estimation de la production. Donc le modèle avec interaction est accepté.

5.8 Estimation du rendement globale des cultures par le modèle avec interaction

Les régressions logistiques du rendement propres à chaque culture ont été étalonnées pour former un sous-ensemble homogène. Et des données sur le terrain pour limiter l'impact de l'hétérogénéité et des erreurs de mesure du sol. Régressions logistiques entre LAI maximale et rendement in situ moyen de la parcelle par type de culture. En replaçant les variables dans l'équation (5-1) on aura :

$$P[Y_i = 1; x_1, x_2] = \frac{exp[-1.8045 + 1.3627.\text{nombre de bandes} + 2.581.\text{nombre de parcelles} - 1.5331.\text{nombre de bandes} \times \text{nombre de parcelles}]}{1 + exp[-1.8045 + 1.3627.\text{nombre de bandes} + 2.581.\text{nombre de parcelles} - 1.5331.\text{nombre de bandes} \times \text{nombre de parcelles}]}$$

Selon la culture, ils représentaient entre 50 % et 85 % de la variabilité des rendements dans les champs d'étalonnage. L'EQM (écart carré moyen) se situait entre 0,6 et 1 t/ha pour les diverses cultures. Dans le cadre de grandes fourchettes de rendement de 1 à 5 t/ha. Comme le LAI maximal est atteint vers la fin août et que la récolte commence pour la plupart des cultures étudiées à la fin septembre. La LAI maximale indique un rendement estimatif au moins un mois avant la récolte.

Tableau 12 : Résultats d'estimation des rendements de toutes les cultures : EQM [kg.m^{-2}] et R^2

Mode	**N (nombre de champs)**	**Régressions Logistiques**	
T=205 (Localité)		EQM (écart carré moyen)	R^2
LAI+BANDES$_{Mil}$	218	8.25±0,45	0,84±0,23
LAI+BANDES$_{Maïs}$	205	7.04±0,63	0,8±0,64
LAI+BANDES$_{Sorgho}$	175	6.35±0,12	0,7±0,84
LAI+BANDES$_{Fonio}$	146	5.25±0,13	0,5±0,55

5.9 Estimation de la production globale des cultures par le modèle avec interaction

Dans chaque domaine, le rendement a été estimé sur la base d'un ratio de battage (poids du grain sur le poids de l'épi/panicule). Nous avons les estimations des écarts de prédictions des rendements par type de culture avec la sortie de SAS (voir le tableau 13 et l'annexe) :

Pour (nombre de parcelles, nombre de parcelles) = (1,1) :

L'intervalle de confiance à 90% pour le sorgho:

Nous obtenons ainsi l'intervalle de confiance $\left[\frac{0.9085}{1+0.9085}, \frac{0.01826}{1+0.01826}\right] = [0.4760, 0.0176] = 0.4584$

L'écart de prédiction pour le sorgho est 0.4584

L'intervalle de confiance à 90% pour le maïs:

Nous obtenons ainsi l'intervalle de confiance $\left[\frac{78.1889}{1+78.1889}, \frac{1.8991}{1+1.8991}\right] = [0.9873, 0.6550] = 0.3323$

L'écart de prédiction pour le maïs est 0.3323

L'intervalle de confiance à 90% pour le mil :

Nous obtenons ainsi l'intervalle de confiance $\left[\frac{0.03968}{1+0.03968}, \frac{0.000763}{1+0.000763}\right] = [0.0381, 7.624.\,\mathrm{E}-4] = 0.0373$

L'écart de prédiction pour le mil est 0.0373

L'intervalle de confiance à 90% pour le fonio :

Nous obtenons ainsi l'intervalle de confiance $\left[\frac{2.1586}{1+2.1586}, \frac{0.1256}{1+0.1256}\right] = [0.6834,\ 0.0880] = 0.5954$

L'écart de prédiction pour le fonio est 0.5954

Ainsi, nous obtenons l'intervalle de confiance a 90% avec l'écart de prédiction pour la production des types de cultures.

Tableau 13 : Écart-type et incertitude (%) du modèle pour estimer la production agricole en utilisant l'approche de régression logistique pour chaque type de culture en tonnes (voir l'annexe pour le tableau SAS)

Type de cultures	**Écart-type**	**Incertitude**
Mil	0.1194	0.0845
Mais	0.07878	0.0274
Sorgho	0.00650	<0.0001
Fonio	0.1936	0.4489

Le mil s'impose comme la culture dominante suivi du maïs et du sorgho, couvrant les écarts de prédictions respectivement 4%, 33%, 46% et reste 60% est consacré au fonio. La culture du mil, qui couvre plus de 218 champs des terres cultivées avec une écart de prédictions de 4% suivie du maïs qui couvre 205 champs avec un écart de prédiction de 33% et du sorgho pour 46% d'écart de prédictions couvrant 175 des champs. À l'instar des indicateurs au niveau des communes, les indicateurs ont été calculés à partir des rendements nationaux moyens (relevés), qui ont été obtenus par une moyenne pondérée de tous les rendements au niveau des régions, en utilisant la zone de récolte déclarée par « Statistiques Agricoles » du pays. Ceci nous donne après avoir applique le modèle au niveau des communes une estimation de la distribution de la production des types de cultures en tonnes au niveau des communes de la région de Koutiala (figure 12).

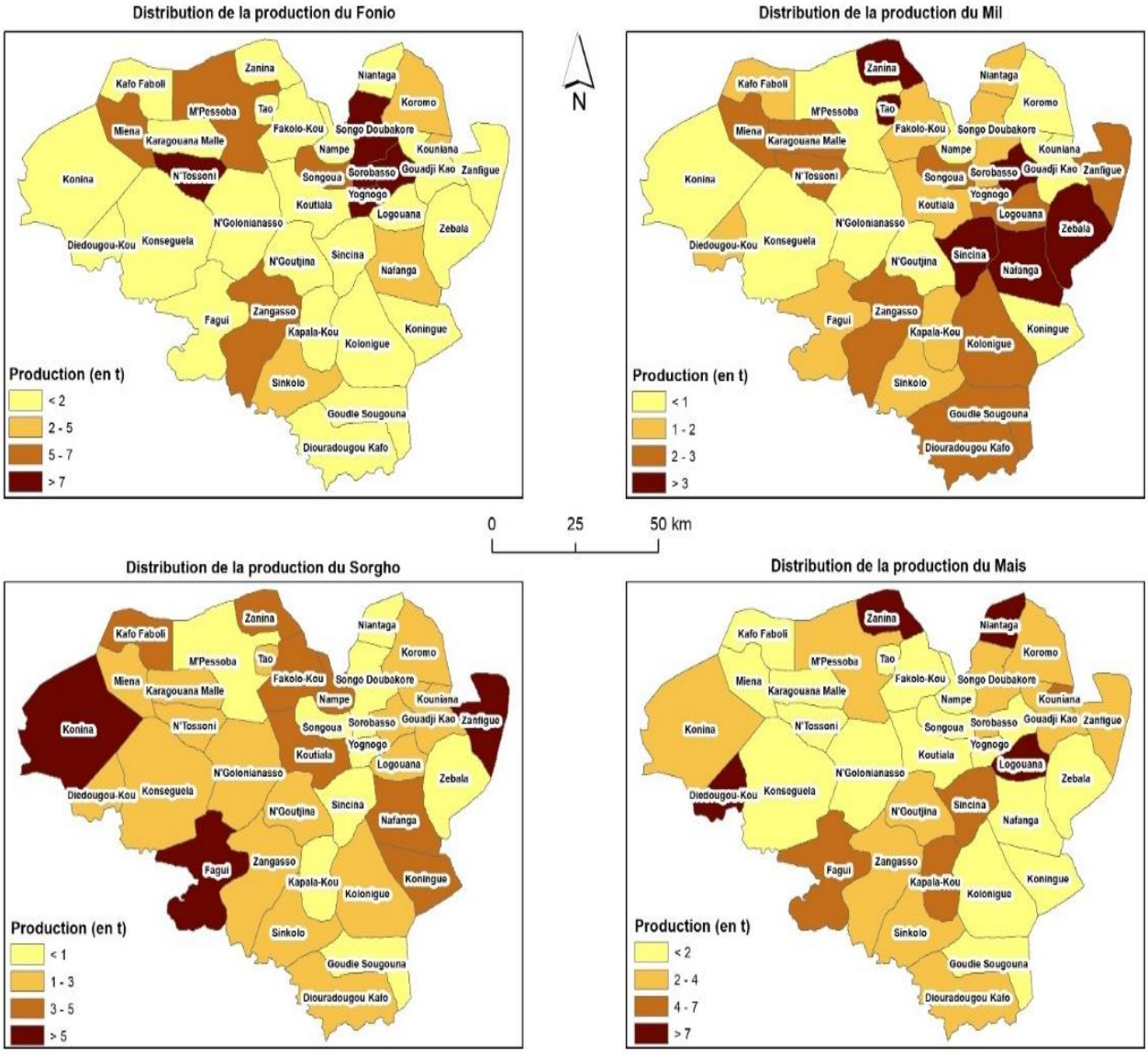


Figure 12 : Carte graphique sur l'estimation de la distribution de la production des types de cultures

6. Discussion et conclusion

6.1 Discussion

Notre projet peut servir de preuve de concept pour la mise en œuvre d'un modèle d'apprentissage automatique intégré pour l'estimation de la production agricole au moyen de la fusion d'images satellitaires multi-sources et des classificateurs tel que *ramdom forest*, et ce dans un cadre de stratégie d'apprentissage statistique. Par ailleurs, les résultats montrent le potentiel très prometteur de ce système et son utilité qui peut se justifier par : (1) la surveillance de la performance agricole dans les systèmes agricoles paysans; (2) une bonne évaluation de la sensibilité des phénotypes de

plantes dans l'estimation de la production; (3) l'acquisition des données « images » en temps réel et des opérations moins coûteuses (données gratuites et analyse automatisée).

Les régressions logistiques du rendement propres aux cultures ont été étalonnées en fonction d'un sous-groupe homogène. Les données terrain ont été utilisées pour limiter les répercussions des erreurs d'hétérogénéité et de mesure au sol. Pour l'estimation du rendement, le plus haut R^2 est obtenu pour le mille suivis du maïs. Le sorgho était souvent cultivé en plus faible quantité. La faible disponibilité de l'eau peut influer à la fois sur la santé des plantes et le rendement en grains dans différentes proportions. De plus, le sorgho est souvent fortement sensible à la photopériode, avec une quantité de biomasse arienne générée en fonction de sa date de semis très variable qui dure généralement jusqu'à 6 semaines dans un seul paysage. Un sorgho fortement photopériodique planté de façon précoce aurait non seulement un faible indice de végétation en raison de sa phase végétative allongée, qui pourrait également se traduire par une réduction du rendement en grains. Ceci peut être dû à la répartition concurrente de l'eau résiduelle du sol afin de maintenir la biomasse durant la phase de remplissage des grains.

Pour l'agrégation des valeurs de production estimées à l'aide de la méthode de régression logistique, l'incertitude dans les prévisions modélisées est faible pour tous les types de cultures. Des incertitudes plus importantes sont rencontrées dans certains champs en raison de la superficie cultivée plus petite, ce qui accroît l'erreur lors de l'agrégation au niveau du village. Le sorgho a des incertitudes de production plus grandes en raison de sa plus faible Fscore dans la classification. L'erreur réelle dans l'estimation de la production n'a pu être évaluée qu'à l'aide de comparaisons avec des statistiques agricoles officielles exactes disponibles seulement à un niveau d'agrégation supérieur à Koutiala.

L’approche traditionnelle couteuses de cartographie explicite des rendements au niveau des régions ne permettaient pas de réaliser assez de points de mesure afin de produire une information suffisante, afin d’estimer avec précision la production des cultures. Cette étude résout ce problème. En outre, cette étude nous a permis de présenter les performances de Google Earth Engine combiné aux données terrain pour la surveillance des cultures dans le cadre d'apprentissage automatique et de processus de traitement automatisé. Ainsi, cette étude a de nouveau confirmé qu'il est actuellement possible de développer un système d'acquisition, de traitement et d'analyse de données pour les systèmes agricoles multicultures en temps quasi-réel et à faible coût. En tant que preuve

de concept, cette étude présente certaines limites. Les tests ont été menés sous conditions particulières, c'est-à-dire une production agricole sérieusement perturbée avec beaucoup de manque de données. Ainsi, des études supplémentaires doivent être menées pour tester d'autres facteurs tels que la culture en contresaison par rapport à la saison principale de culture. De plus, cette étude a été faite sur la base de la classification de foret aléatoire supervisé basé sur des pixels, qui a comme limites des résultats peu lisibles, c'est-à-dire peu explicatifs. La théorie de la régression logistique nous a permis d'aller au-delà de ses limites. Des études futures pourraient établir un système de prévision des récoltes fondé sur cette théorie. De plus, pour ce qui est de l'estimation, des ajustements ont été apportés au modèle de régression logistique avec et sans interactions explicatives. Ce sont en effet une multitude de méthodes qui pourraient être testées dans le but de sélectionner les plus efficaces.

6.2 Conclusion

L'étude propose un modèle intégré d'estimation de la production agricole au niveau communautaire, basé sur des données satellitaires et quelques données terrain, et ce en combinant la classification des types de cultures, l'estimation de la production et les prédicteurs de rendement par télédétection stochastique. Nous avons trouvé que les estimations de production obtenues avec une régression logistique sont très fiables.

Ainsi, les données satellitaires, accessibles notamment sur Google Earth Engine (GEE), offrent de nouvelles perspectives pour surveiller le rendement agricole dans les petits systèmes agricoles, de type agriculture familiale. Avec la disponibilité d'un jeu de données de champ (polygones de type de culture pour former la classification), l'imagerie à haute résolution disponible dans GEE (ou d'autres plateformes géospatiales) semble suffisante pour classer avec précision le type de cultures dans les petits systèmes de ferme relativement « simples » (monocultures) de Koutiala.

Les estimations des pics dans les séries temporelles de LAI dépassent tous les autres indices de végétations testées pour l'estimation des taux de croissance de la végétation. L'avantage particulier d'un seul facteur de prévision basée sur le LAI de pointe est qu'il fournit une estimation du rendement un mois avant la récolte. De plus, les estimations par modèles de régression logistique améliorent la capacité à analyser avec précision la production céréalière au niveau communautaire. Plus important encore, la résolution fine de la variabilité intra-paysage des niveaux de production peut conduire à mieux cibler les interventions d'amélioration des pratiques agricoles. En effet, une

cartographie de base efficace et pédologique des niveaux de production céréalière des petits exploitants agricoles offre la possibilité de développer un environnement réellement transformateur pour l'intensification agricole nécessaire pour atteindre les objectifs d'amélioration des moyens de subsistance dans les pays en développement.

Il est à noter que ce projet a permis d'établir une nouvelle méthode d'estimation de la production des types de cultures au niveau communautaire. Cette approche peut être améliorée en utilisant un modèle d'apprentissage profond et la fusion de données satellitaires avec les données de terrain. Il s'agit d'une approche particulièrement novatrice, car elle repose sur la convergence de technologies alternatives aux mesures terrain qui se complètent pour la validation et l'exactitude. Néanmoins, cette étude a été réalisée à partir d'images satellitaires, mais les images de drones pourraient accroître l'exactitude de l'estimation et de la production des rendements. Les images de drones permettraient d'obtenir en temps réel les stades phénologiques des plantes ainsi que d'autres paramètres biophysiques (taux de chlorophylle, biomasse, etc.). Avec ces images plus précises, il sera possible de considérer les conditions phytosanitaires problématiques (maladies, présence de ravageurs, carences nutritionnelles et hydriques, etc.).

Cette étude s'était basée sur des données terrains recueillis entre 2016 et 2020 seulement. Des campagnes d'acquisition de données sur plusieurs années permettront à l'avenir de tenir compte des diverses variations météorologiques et évolutions climatiques à courte et à longue échelles de temps. De plus, cette étude s'appuyait sur la théorie de la régression logistique. Notre modèle fournit non seulement une mesure de la pertinence d'un prédicteur (taille du coefficient), mais aussi son sens d'association (positif ou négatif). Toutefois, il est limité par l'hypothèse de linéarité entre la variable dépendante et les variables indépendantes.

Puisque la théorie de la classification de l'entropie maximale (MaxEnt) nous permet de dépasser ses limites, nous pourrions développer le système pour les études à venir sur la base de cette théorie. Les prochaines études seraient en mesure de fonctionner avec une analyse complète avec une plus grande base de données. Enfin, la validation du modèle d'estimation des rendements a été effectuée uniquement sur la base de la méthode de la fonction d'efficacité du récepteur, plus couramment

dénommée la courbe ROC[1]. Mais, il y a une multitude de méthodes qui pourraient être testées pour garantir des estimations de confiance pour les hypothèses plus précises.

[1] Reaciver Operating Characteristic

7. Bibliographie

Abdulmana, S., Lim, A., Wongsai, S. et Wongsai, N. (2021) Land surface temperature and vegetation cover changes and their relationships in Taiwan from 2000 to 2020. Remote Sensing Applications: Society and Environment, vol. 24, p. 100636.

Akponikpè, P. B. I., Minet, J., Gérard, B., Defourny, P. et Bielders, C. L. (2011) Spatial fields' dispersion as a farmer strategy to reduce agro-climatic risk at the household level in pearl millet-based systems in the Sahel: A modeling perspective. Agricultural and Forest Meteorology, vol. 151, n°2, p. 215-227.

Atzberger, C. (2013) Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. Remote Sensing, vol. 5, n°2, p. 949-981.

Bandrova, T., Chen, M., Dukaczewski, D., Idrizi, B., Kone, M., Lapaine, M., Marinova, S., Ooms, K., Pashova, L., Pérez-Gómez, R., Nunez, J. J. R., Sanchez-Ortiz, P., Ulugtekin, N., Elzakker, C. V., Zentai, L. et Zlatanova, S. (2016) Editors: Temenoujka Bandrova, Milan Konečný, p. 870.

Barrero, O. et Perdomo, S. A. (2018) RGB and multispectral UAV image fusion for Gramineae weed detection in rice fields. Precision Agriculture, vol. 19, n°5, p. 809-822.

Blaes, X., Chomé, G., Lambert, M.-J., Traoré, P. S., Schut, A. G. T. et Defourny, P. (2016) Quantifying Fertilizer Application Response Variability with VHR Satellite NDVI Time Series in a Rainfed Smallholder Cropping System of Mali. Remote Sensing, vol. 8, n°6, p. 531.

Boryan, C. G., Yang, Z., Di, L. et Hunt, K. (2014) A New Automatic Stratification Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, n°11, p. 4317-4327.

Camps-Valls, G. et Bruzzone, L. (2009) Kernel Methods for Remote Sensing Data Analysis. John Wiley & Sons.

Catalogue de données central (s.d.).

Chang, A. C. (2020) Chapter 5 - Machine and Deep Learning. *In* A. C. Chang (dir.), Intelligence-Based Medicine (p. 67-140). Academic Press.

Chen, S., Useya, J. et Mugiyo, H. (2020) Decision-level fusion of Sentinel-1 SAR and Landsat 8 OLI texture features for crop discrimination and classification: case of Masvingo, Zimbabwe. Heliyon, vol. 6, n°11, p. e05358.

Delrue, J., Bydekerke, L., Eerens, H., Gilliams, S., Piccard, I. et Swinnen, E. (2013) Crop mapping in countries with small-scale farming: a case study for West Shewa, Ethiopia. International Journal of Remote Sensing, vol. 34, n°7, p. 2566-2582.

Dong, Q., Liu, J., Wang, L., Chen, Z. et Gallego, J. (2017) Estimating Crop Area at County Level on the North China Plain with an Indirect Sampling of Segments and an Adapted Regression Estimator. Sensors, vol. 17, n°11, p. 2638.

Duniway, M. C., Karl, J. W., Schrader, S., Baquera, N. et Herrick, J. E. (2012) Rangeland and pasture monitoring: an approach to interpretation of high-resolution imagery focused on observer calibration for repeatability. Environmental Monitoring and Assessment, vol. 184, n°6, p. 3789-3804.

Dury, S., Fouilleux, E. et Bricas, N. (2010) La production de statistiques pour les politiques de sécurité alimentaire : entre visions du monde et enjeux de pouvoir. Le cas du Mali, p. 12.

Duthoit, S. (2006) PRISE EN COMPTE DE L'AGREGATION DES CULTURES DANS LA SIMULATION DU TRANSFERT RADIATIF : IMPORTANCE POUR L'ESTIMATION DE L'INDICE FOLIAIRE (LAI), DE LA PARCELLE AU PAYSAGE (phdthesis). Thèse de doctorat, Université Paul Sabatier - Toulouse III.

Eltazarov, S., Bobojonov, I., Kuhn, L. et Glauben, T. (2023) The role of crop classification in detecting wheat yield variation for index-based agricultural insurance in arid and semiarid environments. Environmental and Sustainability Indicators, vol. 18, p. 100250.

Fanjul-Hevia, A., González-Manteiga, W. et Pardo-Fernández, J. C. (2021) A non-parametric test for comparing conditional ROC curves. Computational Statistics & Data Analysis, vol. 157, p. 107146.

Feng, S., Cao, Y.-L., Xu, T.-Y., Yu, F.-H., Chen, C.-L., Zhao, D.-X. et Jin, Y. (2020) Inversion Based on High Spectrum and NSGA2-ELM Algorithm for the Nitrogen Content of Japonica Rice Leaves. Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis, vol. 40, n°8, p. 2584-2591.

Fieuzal, R., Marais Sicre, C. et Baup, F. (2017a) Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. International Journal of Applied Earth Observation and Geoinformation, vol. 57, p. 14-23.

Fieuzal, R., Marais Sicre, C. et Baup, F. (2017b) Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. International Journal of Applied Earth Observation and Geoinformation, vol. 57, p. 14-23.

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., Mueller, N. D., O'Connell, C., Ray, D. K., West, P. C., Balzer, C., Bennett, E. M., Carpenter, S. R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., Tilman, D. et Zaks, D. P. M. (2011) Solutions for a cultivated planet. Nature, vol. 478, n°7369, p. 337-342.

Fritz, S., See, L., Bayas, J. C. L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I. et McCallum, I. (2019) A comparison of global agricultural monitoring systems and current gaps. Agricultural Systems, vol. 168, p. 258-272.

Gallego, F. J. (1999) CROP AREA ESTIMATION IN THE MARS PROJECT, p. 11.

Gallego, F. J. (2004) Remote sensing and land cover area estimation. International Journal of Remote Sensing, vol. 25, n°15, p. 3019-3047.

Genesio, L., Bacci, M., Baron, C., Diarra, B., Di Vecchia, A., Alhassane, A., Hassane, I., Ndiaye, M., Philippon, N., Tarchiani, V. et Traoré, S. (2011) Early warning systems for food

security in West Africa: evolution, achievements and challenges. Atmospheric Science Letters, vol. 12, n°1, p. 142-148.

Glonek, G. F. V. et McCullagh, P. (1995) Multivariate Logistic Models. Journal of the Royal Statistical Society: Series B (Methodological), vol. 57, n°3, p. 533-546.

Guan, K., Wu, J., Kimball, J. S., Anderson, M. C., Frolking, S., Li, B., Hain, C. R. et Lobell, D. B. (2017) The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. Remote Sensing of Environment, vol. 199, p. 333-349.

Hagolle, O., Huc, M., Villa Pascual, D. et Dedieu, G. (2015) A Multi-Temporal and Multi-Spectral Method to Estimate Aerosol Optical Thickness over Land, for the Atmospheric Correction of FormoSat-2, LandSat, VENμS and Sentinel-2 Images. Remote Sensing, vol. 7, n°3, p. 2668-2691.

He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y. et Zhou, C. (2021) Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. CATENA, vol. 205, p. 105442.

Huuskonen, J. et Oksanen, T. (2019) Augmented Reality for Supervising Multirobot System in Agricultural Field Operation. IFAC-PapersOnLine, 6th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL 2019, vol. 52, n°30, p. 367-372.

Karlson, M. et Ostwald, M. (2016) Remote sensing of vegetation in the Sudano-Sahelian zone: A literature review from 1975 to 2014. Journal of Arid Environments, vol. 124, p. 257-269.

Khot, L. R., Sankaran, S., Carter, A. H., Johnson, D. A. et Cummings, T. F. (2016) UAS imaging-based decision tools for arid winter wheat and irrigated potato production management. International Journal of Remote Sensing, vol. 37, n°1, p. 125-137.

Kobayashi, N., Tani, H., Wang, X. et Sonobe, R. (2020) Crop classification using spectral indices derived from Sentinel-2A imagery. Journal of Information and Telecommunication, vol. 4, n°1, p. 67-90.

Kost, S., Rheinbach, O. et Schaeben, H. (2021) Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling. Geochemistry, Mineral exploration: a journey from fieldwork, to laboratory work, computational modelling and mineral processing, vol. 81, n°4, p. 125826.

Krizhevsky, A., Sutskever, I. et Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *In* F. Pereira, C. J. C. Burges, L. Bottou, et K. Q. Weinberger (dir.), Advances in Neural Information Processing Systems 25 (p. 1097-1105). Curran Associates, Inc.

Lambert, M.-J., Traoré, P. C. S., Blaes, X., Baret, P. et Defourny, P. (2018) Estimating smallholder crops production at village level from Sentinel-2 time series in Mali's cotton belt. Remote Sensing of Environment, vol. 216, p. 647-657.

Lary, D. J., Alavi, A. H., Gandomi, A. H. et Walker, A. L. (2016) Machine learning in geosciences and remote sensing. Geoscience Frontiers, Special Issue: Progress of Machine Learning in Geosciences, vol. 7, n°1, p. 3-10.

Lobell, D. B., Cassman, K. G. et Field, C. B. (2009) Crop Yield Gaps: Their Importance, Magnitudes, and Causes. Annual Review of Environment and Resources, vol. 34, n°1, p. 179-204.

Long, X., Li, X., Lin, H. et Zhang, M. (2021) Mapping the vegetation distribution and dynamics of a wetland using adaptive-stacking and Google Earth Engine based on multi-source remote sensing data. International Journal of Applied Earth Observation and Geoinformation, vol. 102, p. 102453.

López-Lozano, R., Duveiller, G., Seguini, L., Meroni, M., García-Condado, S., Hooker, J., Leo, O. et Baruth, B. (2015) Towards regional grain yield forecasting with 1km-resolution EO biophysical products: Strengths and limitations at pan-European level. Agricultural and Forest Meteorology, vol. 206, p. 12-32.

Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J. E., Pérez-Suay, A. et Camps-Valls, G. (2019) Synergistic integration of optical and microwave satellite data for crop yield estimation. Remote Sensing of Environment, vol. 234, p. 111460.

Mayer, T., Poortinga, A., Bhandari, B., Nicolau, A. P., Markert, K., Thwal, N. S., Markert, A., Haag, A., Kilbride, J., Chishtie, F., Wadhwa, A., Clinton, N. et Saah, D. (2021) Deep learning approach for Sentinel-1 surface water mapping leveraging Google Earth Engine. ISPRS Open Journal of Photogrammetry and Remote Sensing, vol. 2, p. 100005.

Misra, G., Cawkwell, F. et Wingler, A. (2020) Status of Phenological Research Using Sentinel-2 Data: A Review. Remote Sensing, vol. 12, n°17, p. 2760.

Mugabowindekwe, M. et Rwanyiziri, G. (2020) Comparative Assessment of Homogeneity Differences in Multi-Temporal NDVI Strata and the Currently Used Agricultural Area Frames in Rwanda. South African Journal of Geomatics, vol. 9, n°1, p. 89-107-107.

Nattino, G., Pennell, M. L. et Lemeshow, S. (2020) Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. Biometrics, vol. 76, n°2, p. 549-560.

Ni, Z., Yang, Z., Li, W., Zhao, Y. et He, Z. (2019) Decreasing Trend of Geohazards Induced by the 2008 Wenchuan Earthquake Inferred from Time Series NDVI Data. Remote Sensing, vol. 11, n°19, p. 2192.

Nidamanuri, R. R., Jayakumari, R., Ramiya, A. M., Astor, T., Wachendorf, M. et Buerkert, A. (2022) High-resolution multispectral imagery and LiDAR point cloud fusion for the discrimination and biophysical characterisation of vegetable crops at different levels of nitrogen. Biosystems Engineering, vol. 222, p. 177-195.

Ortac, G. et Ozcan, G. (2021) Comparative study of hyperspectral image classification by multidimensional Convolutional Neural Network approaches to improve accuracy. Expert Systems with Applications, vol. 182, p. 115280.

Partel, V., Charan Kakarla, S. et Ampatzidis, Y. (2019) Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. Computers and Electronics in Agriculture, vol. 157, p. 339-350.

Potts, M., Zulu, E. M., Wehner, M., Castillo, F. et Henderson, C. (2013) Crisis In The Sahel – Possible Solutions and the Consequences of Inaction. Africa Portal. African Institute for Development Policy (AFIDEP).

Pradhan, S. (2001) Crop area estimation using GIS, remote sensing and area frame sampling. International Journal of Applied Earth Observation and Geoinformation, vol. 3, n°1, p. 86-92.

Qader, S. H., Utazi, C. E., Priyatikanto, R., Najmaddin, P., Hama-Ali, E. O., Khwarahm, N. R., Tatem, A. J. et Dash, J. (2023) Exploring the use of Sentinel-2 datasets and environmental variables to model wheat crop yield in smallholder arid and semi-arid farming systems. Science of The Total Environment, vol. 869, p. 161716.

QUAZZO, C. et MEUNIER, E. (2020) Des Etats-Unis à l'Inde : le coton transgénique tisse sa toile. Inf'OGM.

Renier, C., Waldner, F., Jacques, D. C., Babah Ebbe, M. A., Cressman, K. et Defourny, P. (2015) A Dynamic Vegetation Senescence Indicator for Near-Real-Time Desert Locust Habitat Monitoring with MODIS. Remote Sensing, vol. 7, n°6, p. 7545-7570.

Sahani, N. et Ghosh, T. (2021) GIS-based spatial prediction of recreational trail susceptibility in protected area of Sikkim Himalaya using logistic regression, decision tree and random forest model. Ecological Informatics, vol. 64, p. 101352.

Sankaran, S., Khot, L. R., Espinoza, C. Z., Jarolmasjed, S., Sathuvalli, V. R., Vandemark, G. J., Miklas, P. N., Carter, A. H., Pumphrey, M. O., Knowles, N. R. N. et Pavek, M. J. (2015) Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review. European Journal of Agronomy, vol. 70, p. 112-123.

Schut, A. G. T., Traore, P. C. S., Blaes, X. et de By, R. A. (2018) Assessing yield and fertilizer response in heterogeneous smallholder fields with UAVs and satellites. Field Crops Research, vol. 221, p. 98-107.

Smale, M., Thériault, V., Assima, A. et Kone, Y. (2020) Implications nutritionnelles du régime alimentaire au Mali. AgEcon Search.

Syifa, M., Park, S.-J. et Lee, C.-W. (2020) Detection of the Pine Wilt Disease Tree Candidates for Drone Remote Sensing Using Artificial Intelligence Techniques. Engineering.

Thorp, K. R., Wang, G., West, A. L., Moran, M. S., Bronson, K. F., White, J. W. et Mon, J. (2012) Estimating crop biophysical properties from remote sensing data by inverting linked radiative transfer and ecophysiological models. Remote Sensing of Environment, vol. 124, p. 224-233.

van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P. et Hochman, Z. (2013) Yield gap analysis with local to global relevance—A review. Field Crops Research, Crop Yield Gap Analysis – Rationale, Methods and Applications, vol. 143, p. 4-17.

van Wart, J., Kersebaum, K. C., Peng, S., Milner, M. et Cassman, K. G. (2013) Estimating crop yield potential at regional to national scales. Field Crops Research, Crop Yield Gap Analysis – Rationale, Methods and Applications, vol. 143, p. 34-43.

Vintrou, E. (2012) Cartographie et caractérisation des systèmes agricoles au Mali par télédétection à moyenne résolution spatiale (phdthesis). Thèse de doctorat, AgroParisTech.

Vischel, T., Lebel, T., Panthou, G., Quantin, G., Rossi, A. et Martinet, M. (2017) Chapitre 2. Le retour d'une période humide au Sahel ? : Observations et perspectives. *In* R. Lalou, A. Oumarou, M. A. Soumaré, B. Sultan, et M. Amadou Sanni (dir.), Les sociétés rurales face aux changements climatiques et environnementaux en Afrique de l'Ouest, Synthèses (p. 43-60). Marseille : IRD Éditions.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I. et Sides, T. (2018) Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecological Indicators, vol. 88, p. 425-438.

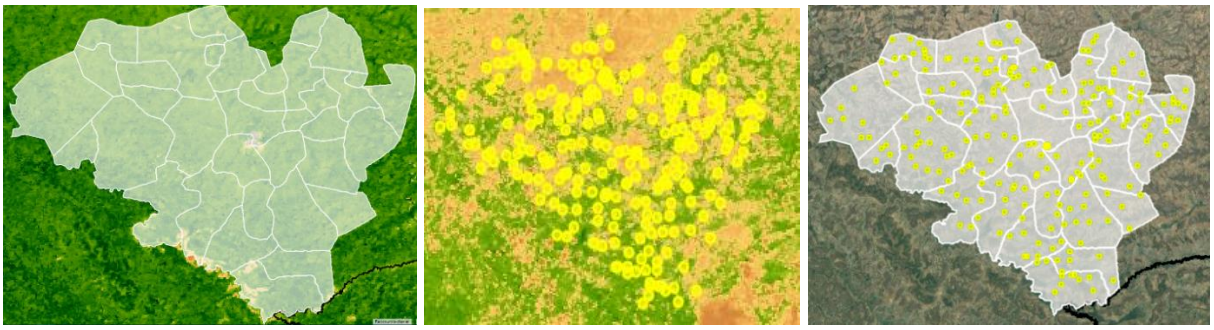
White, J., Berg, A. A., Champagne, C., Zhang, Y., Chipanshi, A. et Daneshfar, B. (2020) Improving crop yield forecasts with satellite-based soil moisture estimates: An example for township level canola yield forecasts over the Canadian Prairies. International Journal of Applied Earth Observation and Geoinformation, vol. 89, p. 102092.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J.-L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A. C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z. et Asseng, S. (2017) Temperature increase reduces global yields of major crops in four independent estimates. Proceedings of the National Academy of Sciences, vol. 114, n°35, p. 9326-9331.

Zou, X. et Mõttus, M. (2017) Sensitivity of Common Vegetation Indices to the Canopy Structure of Field Crops. Remote Sensing, vol. 9, n°10, p. 994.

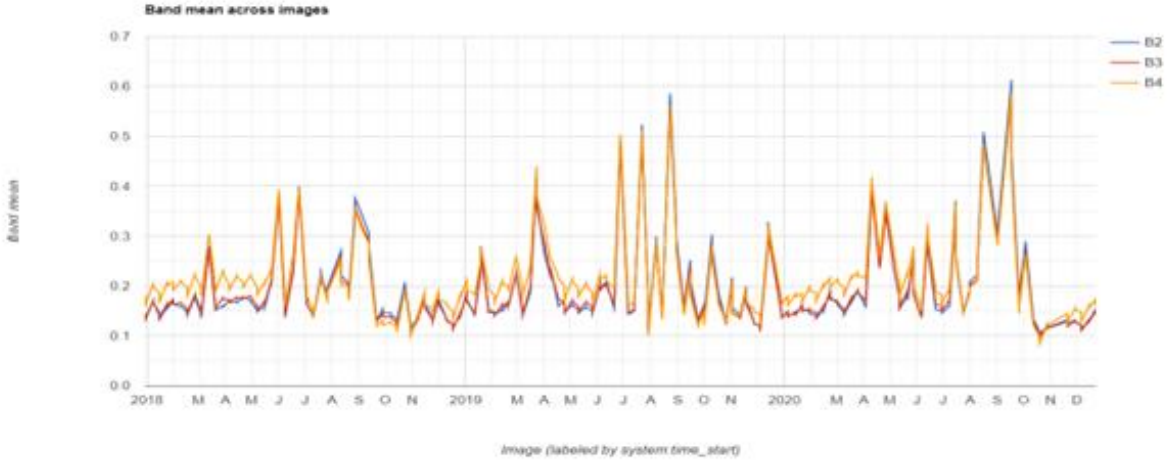
8. Annexes

1 Type d'occupation du sol de chaque point d'échantillonnage

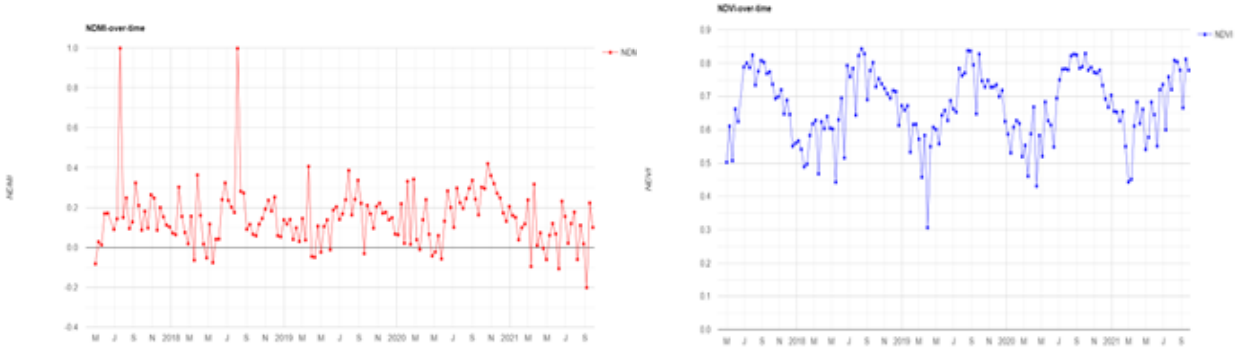


Source Google Earth Ingine

2 Les bandes de réflectance de surface



3 Le graphe des différents indices (NDVI) et (NDWI)



4 L'intervalle de confiance a 90% pour l'écart de prédiction des types de cultures (SAS)

The GLIMMIX Procedure

Estimates

Libellé	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|	Alpha	Inférieur	Supérieur	Moyenne	Erreur type de la moyenne	Moyenne inférieure	Moyenne supérieure	Estimation exponentielle	Exponentiel inférieur	Exponentiel supérieur
prob3	-2.0495	1.1813	172	-1.74	0.0845	0.1	-4.0030	-0.09597	0.1141	0.1194	0.01793	0.4760	0.1288	0.01826	0.9085
prob4	2.5003	1.1240	172	2.22	0.0274	0.1	0.6414	4.3591	0.9242	0.07878	0.6551	0.9874	12.1857	1.8991	78.1889
prob03	-5.2024	1.1945	172	-4.36	<.0001	0.1	-7.1778	-3.2269	0.005473	0.006502	0.000763	0.03816	0.005503	0.000763	0.03968
prob04	-0.6526	0.8599	172	-0.76	0.4489	0.1	-2.0747	0.7695	0.3424	0.1936	0.1116	0.6834	0.5207	0.1256	2.1586

5 Autres données auxiliaires.

- Élévation :

  Les données du Modèle Numérique de Terrain (MNT) mondial Shuttle Radar Topography Mission (SRTM) ont été importées depuis http://srtm.csi.cgiar.org/. Ce sont des données d'élévation à 90 m de précision au nadir.

- CN ration : C'est un rapport entre la masse de carbone et la masse d'azote dans une substance.
- STN : (Réseau de transformateurs spatiaux) permet de recadrer et de normaliser l'échelle de la région appropriée, ce qui peut simplifier la tâche de classification ultérieure et conduire à de meilleures performances de classification
- SOC : Stock de carbone organique du sol
- EC : Analyse statistique de la conductivité électrique.
- **Extraction des indicateurs de la phénologie simulés (LAI) :**

Les profils LAI ont été simulés à l'aide du modèle de plante. Le modèle de culture SARRA-HV32 est un modèle de culture « robuste », développé par le CIRAD[2] permet de simuler les dynamiques des biomasses des céréales et plus particulièrement de plusieurs variétés de mil, maïs et sorgho cultivées en milieu paysan (Vintrou, 2012). Les pratiques agricoles les plus courantes, choix des variétés et notamment stratégies de semis, peuvent être simulées dans le modèle, ce qui permet de calculer les dates de semis réussis en fonction de la mise en place de la saison des pluies. Les scénarios de simulation ont été faits par expertise selon deux critères importants de gestion des cultures au regard des pratiques paysannes : l'adaptation des espèces et variétés, et des pratiques de semis en fonction des zones climatiques.

- **Données sur le rendement :**

En plus des données terrain qui sont difficilement accessibles au Mali, nous avons comparé les

[2] ACRICUTURAL RESEARCH FOR DEVELOPMENT

phénologies des variables observées par télédétection à des variables simulées par le modèle de plante SARRA-H, à partir de profils de LAI[3]. Ce modèle a déjà été entrainé sur différentes espèces et variétés d'Afrique de l'Ouest, y compris les variétés de mil, de sorgho, et de maïs que l'on retrouve au Mali. Il est actuellement utilisé par Le (CILSS)[4] pour l'estimation des rendements des cultures.

[3] LAI (Leaf Area Index),

[4] Comité permanent Inter-Etats de Lutte contre la Sécheresse dans le Sahel