Estimation du Rendement du Mil Perlé (*Pennisetum glaucum*) par Machine Learning à l'aide d'Images Satellites

A. Chemchem[1], L. Mohimont[2], F. Alin[2], L.A. Steffenel[2]

[1] ATOS - Pôle Data Driven Intelligence
Rue du Mas de Verchant, 34000 Montpellier, France
[2] Université de Reims Champagne-Ardenne,
Laboratoire LICIIS - LRC CEA DIGIT

lamine.chemchem@atos.net, lucas.mohimont@univ-reims.fr,
francois.alin@univ-reims.fr, luiz-angelo.steffenel@univ-reims.fr

Résumé

L'estimation du rendement agricole joue un rôle crucial dans la poursuite des objectifs de développement durable des Nations Unies, représentant ainsi un outil essentiel dans la prise de décisions concernant les systèmes d'approvisionnement. Dans ce travail, nous nous intéressons à la prédiction du rendement du Pennisetum glaucum, aussi connu comme "mil à chandelle" ou "mil perlé". Connaître le potentiel de production le plus tôt possible permet de prendre des mesures préventives et éviter des défauts d'approvisionnement pour la population. Pour ce faire, nous croisons les données historiques de rendement des parcelles au Sénégal avec des données satellitaires couvrant trois phases différentes du cycle de vie du mil, grâce à des méthodes d'apprentissage automatique. En comparant différentes méthodes, nous avons obtenu des estimations de rendement assez précises 1 mois avant la récolte, avec un taux d'erreur qui ne dépasse pas 140 kg/ha.

Mots-clés

Rendement agricole, Télédétection optique, Apprentissage automatique

Abstract

Agricultural yield estimation contributes to many of the United Nations' sustainable development goals, and can be considered as a decision-making tool for a supply system. In this work, we are interested in predicting the yield of Pennisetum glaucum, also known as "pearl millet". Knowing the production potential of this cereal as early as possible enables authorities to take preventive measures and avoid supply shortages for the population. In this work, we cross-reference historical plot yield data in Senegal with satellite data covering three different phases of the millet life cycle, using machine learning methods. By comparing different methods, we obtained fairly accurate yield estimates 1 month before harvest, with an error rate of no more than 140 kg/ha.

Keywords

Crop yield estimation, Optical remote sensing, Machine learning

1 Introduction

Le "mil à chandelle", "mil perlé" ou simplement **mil** (*Pennisetum glaucum*) est une espèce de plantes annuelles de la famille des Poaceae (Graminées). Elle est cultivée comme céréale pour ses graines comestibles et joue un rôle important en tant que culture vivrière en Inde et au Pakistan, ainsi que dans le Sahel africain et dans des zones semi-arides. Connaître le potentiel de production fait partie des mesures permettant de garantir la sécurité alimentaire de la population en cas de baisse de production.

Cette étude vise à estimer les rendements de mil par télédétection optique dans la petite agriculture familiale au Sénégal. Ce projet s'inscrit dans le cadre de la sécurité alimentaire et des moyens de subsistance des populations, et ses résultats pourraient représenter un outil d'aide à la décision pour les services d'approvisionnement de plusieurs ODD (Objectifs de Développement Durable) adaptés par les Nations Unies. En particulier pour l'ODD 2, qui vise à "Éliminer la faim, assurer la sécurité alimentaire et une meilleure nutrition et promouvoir l'agriculture durable".

Le site choisi pour cette étude est situé dans la zone de l'ancien bassin arachidier du Sénégal, comme le montre la figure 1. Il représente la principale zone de production agricole du pays. Le bassin est constitué de sols ferrugineux tropicaux permettant une production agricole principalement composée de céréales sèches (mil, sorgho, maïs) et de légumineuses (arachide, niébé) cultivées seules ou en association.

Le jeu de données de cette étude a été partagé dans le cadre d'un défi de science des données organisé par Acta [1]. Il se compose d'un total de 81 parcelles de mil réparties dans le site du bassin. Cette région a été choisie parce qu'il s'agit d'une zone d'intérêt de longue date pour plusieurs équipes de recherche, de sorte que nous disposons de connaissances

1. Instituts techniques agricoles `http://www.acta.asso.fr/`

sur le terrain, en plus de la base de données agronomique ou paysagère historique. Un autre avantage très important est la présence d'équipes de recherche sur le terrain pour assurer et coordonner la collecte de données.

Le climat de la région est unimodal avec une saison des pluies entre juillet et octobre. Cependant, la zone connaît une forte croissance démographique, à laquelle s'ajoutent une réduction des temps de jachère, un appauvrissement progressif des sols, ce qui conduit à une baisse des rendements observés en milieu rural, à une dégradation des ressources naturelles et à la perte de biodiversité [3].

Afin d'atteindre nos objectifs, nous mettons en oeuvre des méthodes d'apprentissage automatique avec deux axes : régression et classification. Le premier axe d'étude consiste à prédire le rendement quantitatif avec des méthodes de régression pour chaque parcelle. Dans le deuxième axe, nous adaptons les méthodes de classification supervisée afin d'affiner encore la précision des prédictions. Le résultat final de cette étude pourrait être un OAD (outil d'aide à la décision) de surveillance des risques d'approvisionnement.

Le reste du document est organisé comme suit : la section 2 passe en revue la littérature sur l'apprentissage automatique appliqués à l'agriculture intelligente. La section 3 décrit l'ensemble de données et le dispositif expérimental mis en place. La section 4 expose les résultats obtenus avec quelques analyses et discussions. Enfin, la section 5 conclut le document avec quelques perspectives.

2 État de l'Art

Les méthodes d'apprentissage automatique sont utilisées comme une nouvelle approche qui révolutionne les modèles classiques de prédiction du rendement. Toutefois, les approches basées sur l'apprentissage de données historiques nécessitent des données étiquetées pour fournir une modélisation précise. En effectuant une étape d'apprentissage sur les données étiquetées, ces méthodes sont capables de produire des modèles de prédiction quantitatifs appelés méthodes de régression. Par exemple, dans le cas de la prédiction du rendement, on veut prédire la quantité produite en tonnes par hectare, ou en kilogrammes par hectare. Il est aussi possible de construire un modèle de prédiction qualitatif via les méthodes de classification, permettant par exemple de prédire la classe de rendement attendue : faible, moyen ou élevé.

Dans la littérature, on peut citer le travail de [9], qui a comparé deux méthodes de régression : Perceptron multicouche (MLP) et Support Vector Regressor (SVR) pour la prédiction du rendement du maïs. En utilisant les données de l'indice de végétation amélioré (EVI) et des séries temporelles climatiques des dix dernières années, la méthode MLP mise en œuvre a atteint un score R^2 de 0,81. L'EVI et les données satellitaires en général sont utiles pour compenser le manque de données agricoles de détection sur le terrain. Malheureusement, nous n'avons pas pu accéder à la base de données de cette étude pour comparer nos méthodes, mais nous avons mis en œuvre les approches SVR et MLP sur notre ensemble de données.

Une autre recherche intéressante est celle de [14], dans laquelle les auteurs ont utilisé des séries temporelles de NDVI (Normalized Difference Vegetation Index) sur 10 jours avec des apports d'engrais chimiques comme caractéristiques d'apprentissage pour la prédiction du rendement du blé. Ils ont comparé différentes caractéristiques basées sur le NDVI, le NDVI cumulé et le NDVI cumulé à des dates significatives avec l'arbre de régression boosté (Boosted Regression Tree - BRT) et le SVR. Les deux techniques ont été utilisées pour la sélection des caractéristiques et la modélisation de la régression. Les meilleurs résultats ont été obtenus avec le BRT, avec une erreur calculée par RMSE (Root Means Square Error) inférieure à 0,2 tonne par hectare. Dans notre étude, nous avons utilisé l'indice NDVI avec cinq autres indices.

Dans [8], les auteurs ont utilisé des séries temporelles NDVI de 16 jours avec une résolution de 250 mètres et des observations climatiques SILO [2] pour la prévision du rendement du blé en Australie. Chaque échantillon de données est un pixel correspondant à un carré de 250m x 250m et différents modèles de base et d'ensemble ont été comparés sur la précision basée sur le pixel avec une validation croisée 5 fois. Les meilleurs résultats ont été obtenus par un SVR (Support vector Regressor) adapté utilisant le noyau RBF (radial basis function) qui atteint une erreur RMSE de 0,55t/ha et R^2 de 0,77.

Une autre étude de [12] se concentre sur l'axe de la classification supervisée afin de traiter les images satellites. Les auteurs ont utilisé les données NDVI, pédologiques et météorologiques pour prédire le rendement du maïs à l'intérieur d'un champ. Le champ a été divisé en 63 unités de traitement et la modélisation a été simplifiée à une classification binaire : classes de faible rendement et de rendement moyen à élevé. Cinq classificateurs supervisés différents ont été comparés avec une validation croisée 5 fois, et les meilleurs résultats ont été obtenus par le classificateur XG-Boost avec une précision de 95%. Cette étude nous a incités à réaliser une étude comparative des méthodes de classification les plus populaires.

Des méthodes basées uniquement sur l'historique de la production et les données météorologiques ont aussi montré des résultats prometteurs. Ainsi, [5] a pu obtenir des taux de précision supérieurs à 99% pour la culture du soja ou du riz, et de 98% pour la culture du maïs, en utilisant la méthode Random Forest sur des données issues de régions agricoles brésiliennes. Ce travail a toutefois bénéficié d'un historique sur plus de 20 ans de production agricole, ce qui n'est pas possible dans notre cas.

L'apprentissage profond peut également être utilisé pour prédire le rendement. Souvent, l'apprentissage profond nécessite de grands ensembles de données, mais dans certains cas, l'apprentissage par transfert peut être mis en œuvre pour compenser le manque de données en utilisant un modèle pré-entraîné. Les auteurs de [15] ont utilisé l'apprentissage par transfert pour former un modèle de mémoire à

2. SILO est une base de données du gouvernement du Queensland contenant des données climatiques quotidiennes continues pour l'Australie de 1889 à nos jours.

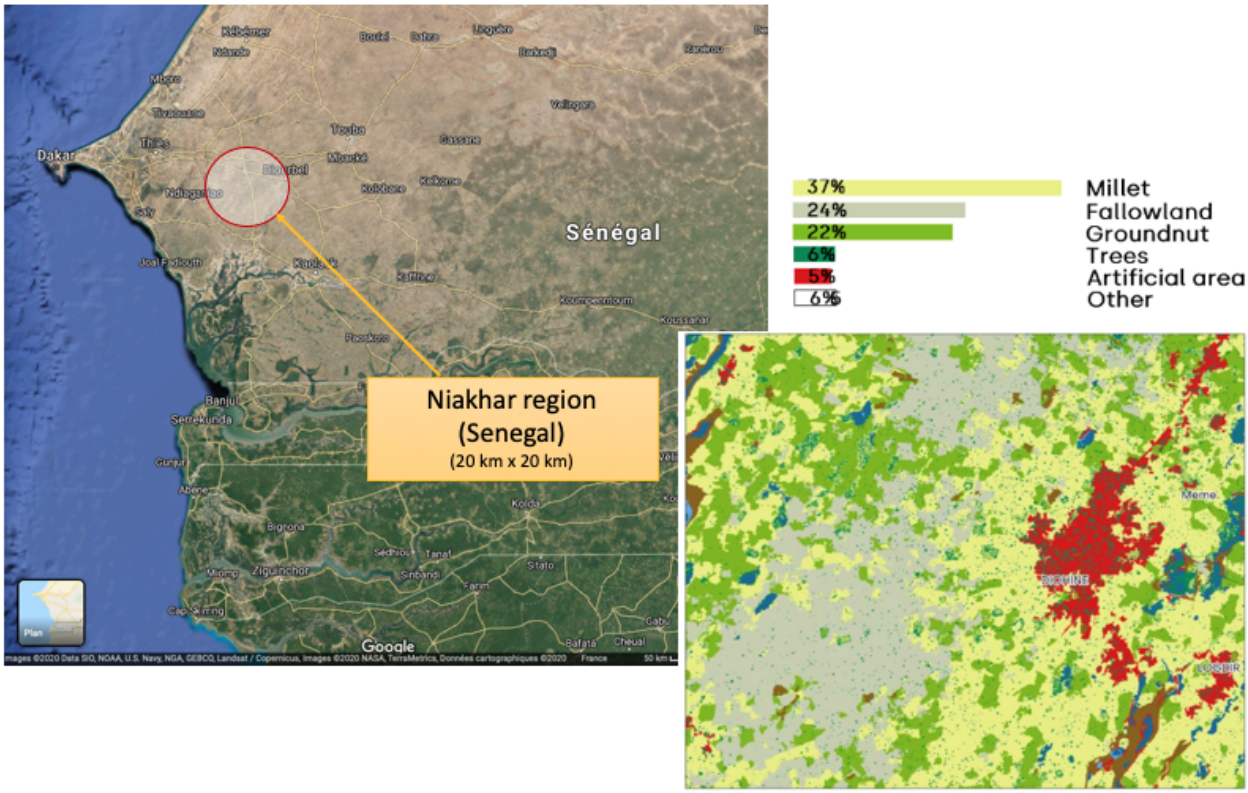

FIGURE 1 – Site étudié : région de Niakhar, ancien bassin producteur d'arachide, Sénégal [11].

long terme (LSTM) avec la réflectance MODIS [3] et les séries temporelles de températures pour la prédiction du rendement du maïs au Brésil. Le modèle a d'abord été entraîné sur un plus grand ensemble de données, avec 1837 échantillons de récoltes en Argentine, puis le modèle a été réentraîné sur 336 données de récoltes brésiliennes. La LSTM avec apprentissage par transfert a obtenu de meilleurs résultats moyens que la LSTM sans apprentissage par transfert. C'est pourquoi nous prévoyons d'explorer cette méthode sur notre ensemble de données dans un travail futur.

3 Matériaux & Méthodes

3.1 L'ensemble de données

L'ensemble de données utilisé dans cette étude est récolté à partir de 81 parcelles de mil situées dans la région de Niakhar, au Sénégal. Ce jeu de données comprend des données historiques collectées sur les années 2017 et 2018, représentant le rendement du mil en Kg/ha, ainsi que les données satellitaires des parcelles correspondantes, comme le montre la figure 2. Il faut noter que les parcelles étudiées sont trop petites, pour cette raison nous ne pouvons pas réaliser une étude individualisée (*intra-field*).

Les données satellitaires extraites contiennent les indices de végétation suivants : **NDVI, MSAVI2, NDWI, CIGreen, GDVI** et **PSRINIR**. Ces indices permettent d'estimer les paramètres biophysiques et sont calculés à partir de la réflectance de deux bandes spectrales, rouge (R) et proche infrarouge (NIR).

L'indice NDVI (Normalized Difference Vegetation Index) est le plus utilisé. Le calcul de cet indice est basé sur la réflectance de la chlorophylle dans le proche infrarouge et permet de suivre la biomasse intraparcellaire. L'indice SAVI (Soil-Adjusted Vegetation) est dérivé de cet indice et propose un ajustement avec une constante.

Plus tard, l'indice MSAVI2 (Modified Soil-Adjusted Vegetation Index) a été proposé par [13], il utilise une constante ajustée aux conditions locales.

Le NDWI (Normalized Difference Water Index) est basé sur le même principe que le NDVI et permet de surveiller l'état hydrique des cultures (Gao, 1996). Le NDWI est basé sur le pic d'absorption de l'eau dans une bande infrarouge de courte longueur d'onde.

Le CIGreen (Green Chlorophyll Index) est utilisé pour évaluer la teneur en chlorophylle des feuilles. Cet indice est sensible aux petites variations de chlorophylle.

Le GDVI (Generalized Difference Vegetation Index) est un indice dérivé du NDVI, particulièrement adapté aux zones arides où le couvert végétal est faible [16].

Finalement, le PSRINIR (Plant Senescence Reflectance Index Near Infra Red), proposé par [10], compare les caroténoïdes et les chlorophylles, identifiant ainsi la sénescence de la canopée (augmentation des caroténoïdes).

Les approches d'apprentissage automatique présentées dans cet article ont été mises en œuvre en utilisant le langage python avec la bibliothèque Scikit-Learn. L'optimi-

3. https://visibleearth.nasa.gov/images/54078/modis-surface-reflectance

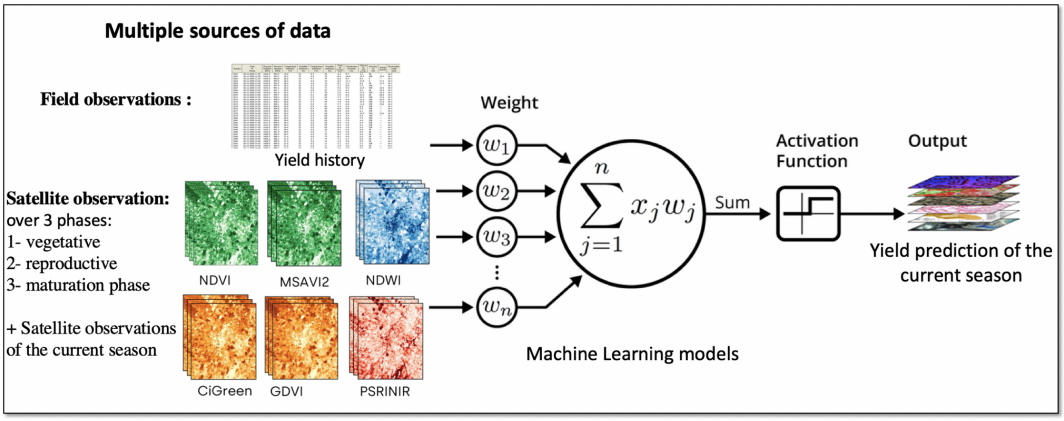

FIGURE 2 – Workflow général : De l'intégration des données à la prédiction du rendement.

sation des hyperparamètres a été accélérée grâce aux ressources du Centre de Calcul Régional ROMEO [4] de l'Université de Reims Champagne-Ardenne.

3.2 Méthodologie

Les observations par satellite sont collectées au cours de trois périodes de croissance différentes, conformément au calendrier de culture du mil à chandelle (*Pennisetum glaucum*) expliqué dans la figure 3.

FIGURE 3 – Calendrier cultural du mil à chandelle dans l'ancien bassin arachidier, Sénégal [11].

Dans un premier moment, nous avons choisi d'estimer les rendements le plus tôt possible avant la récolte. Cela se fait en estimant le rendement uniquement avec des données de la phase végétative (5 mois avant la récolte), sans prendre en compte les données des phases reproductive et de maturation. Une deuxième expérience consiste à prédire les rendements 3 mois avant la récolte, utilisant cette fois-ci les données des phases végétative et reproductive. Finalement, la troisième expérience consiste à utiliser toutes les données disponibles, c'est-à-dire les données des trois phases végétation, reproduction et maturation, afin de prédire le rendement du mil environ 1 mois avant la récolte.
En outre, et pour tirer parti des méthodes d'apprentissage automatique, nous avons mis en œuvre des algorithmes pour estimer le rendement quantitatif, par le biais de méthodes de régression. Puis, dans un second temps, nous avons exploré les méthodes de classification supervisée afin de valider les résultats et d'être le plus précis possible. Pour chaque type d'approche d'apprentissage automatique, nous avons mis en place une étude comparative des algorithmes les plus prometteurs tels que décrits dans la section littérature. Il est à noter que, pour chaque approche d'apprentissage, les meilleurs hyperparamètres sont sélectionnés par validation croisée et recherche en grille (*GridSearch*).

4. http://romeo.univ-reims.fr

3.3 Méthodes pour la régression et leur évaluation

Dans cette partie, nous expliquons les étapes que nous avons suivies pour réaliser l'étude comparative des algorithmes de régression.

Prétraitement des données par mise à l'échelle des caractéristiques : cette étape est appliquée pour normaliser la plage des variables indépendantes de l'ensemble de données. Étant donné que la plage de valeurs des données brutes varie considérablement, les fonctions objectives ne fonctionneront pas correctement sans normalisation. Une autre raison pour laquelle la mise à l'échelle des caractéristiques est appliquée est que la descente de gradient converge beaucoup plus rapidement avec la mise à l'échelle des caractéristiques [7].
Dans notre implémentation, nous avons appliqué la normalisation min-max, qui est la méthode la plus simple et qui consiste à remettre à l'échelle la plage de caractéristiques dans un intervalle [0, 1] ou [-1,+1]. La sélection de la plage cible dépend de la nature des données, et puisque dans notre ensemble de données il n'y a pas de données négatives, nous les avons mises à l'échelle entre [0, 1].

Validation par répartition *train/test* : Fondamentalement, l'évaluation des approches d'apprentissage automatique se fait par la division de l'ensemble de données en deux ensembles, l'un appelé ensemble d'entraînement (*train*) et l'autre ensemble de test. Le premier contient les données avec les étiquettes utilisées pour construire le modèle, tandis que le second est utilisé pour tester les performances de ce modèle. Dans notre cas, comme nous ne disposons que de deux années d'historique de données, nous avons pris les données de 2017 pour former les modèles et celles de 2018 pour les tester.

Évaluation de la régression par R2-score et RMSE : nous avons évalué nos modèles de régression par les deux formules d'évaluation les plus connues : R^2-score et RMSE. Le Score R^2 (appelé aussi R-carré) est une mesure statistique de la proximité des données par rapport à la

droite de régression ajustée. Il est également connu sous le nom de coefficient de détermination ou de coefficient de détermination multiple pour la régression multiple [4]. En général, plus le R au carré est élevé, mieux le modèle s'adapte à vos données.
Le RMSE (Root Mean Squared Error) est la racine carrée de l'erreur quadratique moyenne (Mean Squared Error - MSE), qui est une fonction de risque correspondant à la valeur attendue de la perte d'erreur quadratique.

3.4 Méthodes de classification supervisée et leur évaluation

Après la mise en œuvre de l'étude comparative de régression, nous avons exploré les méthodes d'apprentissage automatique pour la classification supervisée, en transformant le problème de régression en un problème de classification multi-classes. Cette approche vise à affiner la précision de notre prédiction de rendement, de sorte qu'au lieu d'essayer de prédire une valeur quantitative, il nous suffit de prédire une classe parmi trois classes possibles : rendement faible, rendement moyen ou rendement élevé.

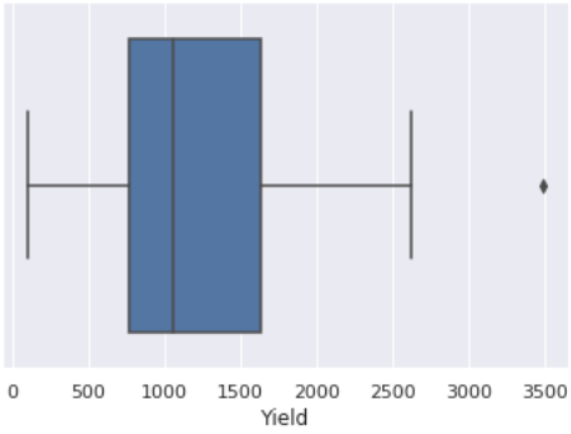

FIGURE 4 – Distribution des valeurs de rendement.

La distribution des valeurs de rendement se situe entre 107,9 kg/ha et 3488,9 kg/ha, comme le montre la figure 4. À partir de cette distribution, nous avons créé les trois classes de rendement suivantes :

— Classe "rendement bas" : si 700 kg/ha $>$ rendement ;
— Classe "rendement moyen" : si 700 kg/ha $\leq$ rendement $<$ 1600 kg/ha ;
— Classe "rendement haut" : si rendement $\geq$ 1600 kg/ha.

Le résultat de la distribution des classes obtenues est présenté dans la figure 5. Nous pouvons remarquer que les classes obtenues sont fortement déséquilibrées, ce qui nécessite une étape de prétraitement supplémentaire pour cet ensemble de données, avec la technique SMOTE (Synthetic Minority Over-sampling Technique). Dans SMOTE, les classes minoritaires sont suréchantillonnées en introduisant des instances synthétiques dans lesquelles chaque échantillon de classe minoritaire est prélevé. Les données générées sont insérées le long des segments de ligne reliant certains des k plus proches voisins de la classe minoritaire. Les voisins sont choisis au hasard parmi les k plus proches voisins en fonction de l'ampleur du suréchantillonnage nécessaire. Cinq voisins les plus proches sont actuellement utilisés dans la mise en œuvre de SMOTE [1] [2].

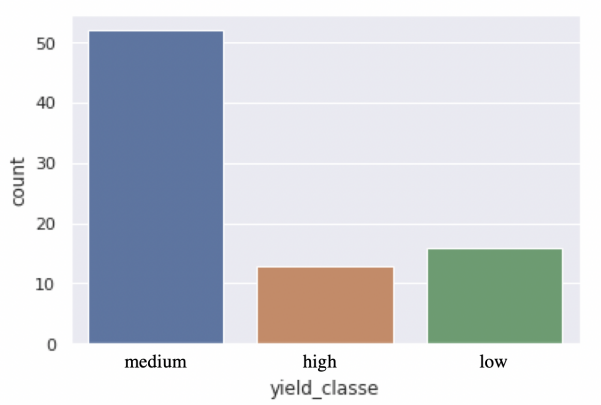

FIGURE 5 – Distribution des classes de rendement.

Afin de valider les résultats, nous avons utilisé les méthodes et métriques suivantes :

Validation des résultats par répartition *train/test* : De la même manière que pour l'étude comparative de régression, les données ont été réparties en deux groupes, l'un pour l'entraînement et l'autre pour le test. En raison de la faible quantité de données, l'ensemble d'entraînement couvre l'année 2017 tandis que l'ensemble de test correspond aux données de 2018.

Résultats Évaluation par F1-score & Accuracy : Nous avons évalué nos modèles de classification par les deux formules d'évaluation les plus connues : le score F et la précision (*accuracy*) de la classification. Le score F, également appelé mesure F, est basé sur les deux mesures principales : la précision (p*recision*) et le rappel (*recall*). La précision est la proportion de cas que le sujet a classés comme positifs et qui étaient vraiment positifs (TP - *true positive*). Elle est équivalente à la valeur prédictive positive. Le rappel est la proportion de cas vraiment positifs qui ont été classés comme positifs par le modèle. Il est équivalent à la sensibilité.

$$Precision = \frac{TP}{TP + FP}$$

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

Où TP est le nombre de vrais positifs, TN est le nombre de vrais négatifs, FP est le nombre de faux positifs et FN est le nombre de faux négatifs.
Les deux métriques sont souvent combinées sous la forme de leur moyenne harmonique [6] appelé F-Score. La métrique F-score peut être utilisée pour équilibrer la contribution des faux négatifs en pondérant le rappel par un paramètre $\beta \geq 0$. Dans notre cas, β est fixé à 1, le score F1 est alors égal à :

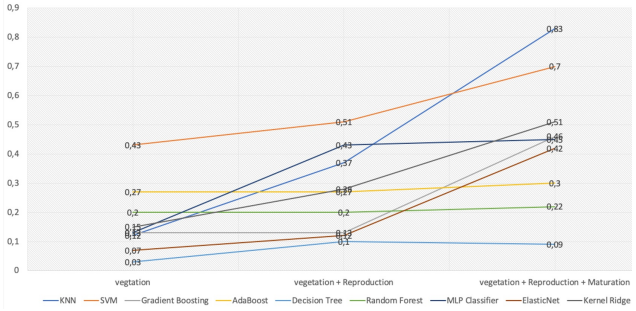

FIGURE 6 – Comparaison R^2-score.

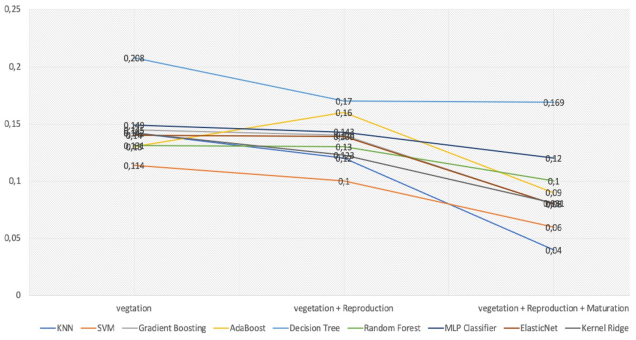

FIGURE 7 – Comparaison RMSE.

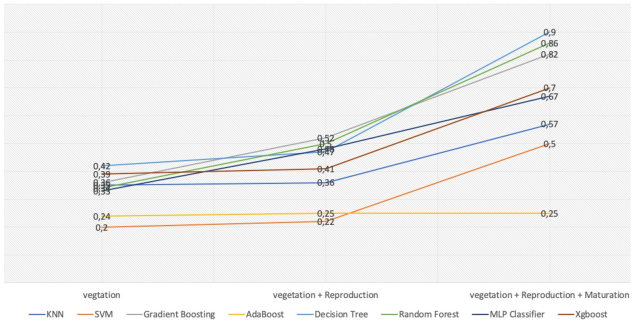

FIGURE 8 – Comparaison F-score.

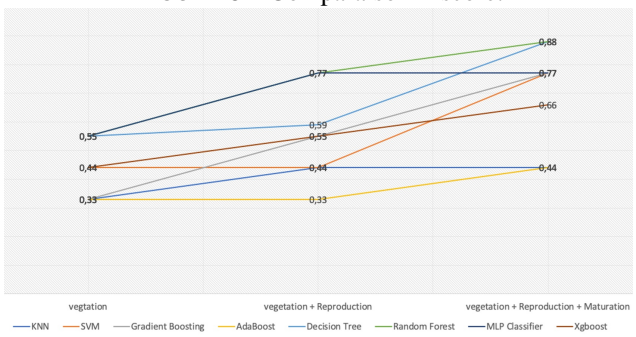

FIGURE 9 – Comparaison accuracy.

$$F1\_score = \frac{2 \times recall \times precision}{precision + recall}$$

Finalement, nous utilisons la métrique *accuracy* (traduite par exactitude ou justesse), l'un des critères les plus connus pour évaluer les modèles de classification. D'une manière non formelle, elle se réfère à la proportion de prédictions correctes faites par le modèle. Sa formule est la suivante :

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

4 Résultats et Discussion

4.1 Le cas de la Régression

Après avoir effectué l'étape de prétraitement, les résultats obtenus à partir des méthodes de régression sont résumés dans le tableau 1.

En synthétisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 6 la comparaison du score R^2 des méthodes de régression au cours des trois phases de maturation du mil : végétative, reproductive et de maturation. De même manière, la figure 7 compare ces méthodes selon la métrique RMSE.

A partir de ces résultats, nous pouvons voir que les meilleurs scores pour les prédictions de rendement dans les phases végétative et reproductive sont donnés par la méthode SVM, alors que dans la phase de maturation les meilleurs scores sont donnés par la méthode du régresseur K plus proche voisin. Cela est en partie dû à la plus grande quantité de données disponibles lorsqu'on réduit le temps avant la récolte : plus nous ajoutons de nouvelles données aux cycles de vie du mil, plus la précision des méthodes de régression augmente.

De ce fait, l'erreur de prédiction diminue de RMSE = 0,11, équivalent à une erreur de 400 kg/ha si la prédiction n'inclut que des données de la phase végétative, à un RMSE de 0,04, correspondant à une erreur de prédiction de 140 kg/ha lors de la phase de maturation.

4.2 Le cas de la Classification

Après l'étape de prétraitement, les résultats obtenus par les méthodes mises en œuvre en utilisant les méthodes de validation et d'évaluation expliquées précédemment sont résumés dans le tableau 2. En schématisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 8 la comparaison du score R2 des méthodes de classification au cours des trois phases du cycle de vie du mil : végétative, reproductive et de maturation. De même, la figure 9 compare ces méthodes en évaluant leur précision.

D'après ces résultats, nous pouvons voir que les meilleurs scores pour la prédiction de la classe de rendement en utilisant uniquement les données végétatives (c'est-à-dire 5 mois avant la récolte) sont donnés par la méthode de l'arbre de décision, qui atteint une précision de 0,56. De même, dans la phase de maturation (lorsque l'on utilise tous les stades du cycle de vie du mil), les meilleures prédictions sont faites par le modèle d'arbre de décision avec 90% du score F et 88% de l'accuracy, suivi de près par la méthode Random Forest.

Comme nous nous y attendions, la précision de prédiction augmente de F-score = 0,42 pendant la phase végétative, à F-score = 0,90, ce qui signifie une erreur de prédiction de

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	R2-score	RMSE	R2-score	RMSE	R2-score	RMSE
K-Nearest Neighbors	0.12	0.14	0.37	0.12	**0.83**	**0.04**
Support Vector Machine	**0.43**	**0.11**	**0.51**	**0.10**	0.70	0.06
Gradient Boosting	0.13	0.14	0.13	0.14	0.46	0.08
Ada Boosting	0.27	0.13	0.27	0.16	0.30	0.09
Decision Tree	0.03	0.20	0.10	0.17	0.09	0.16
Random Forest	0.20	0.13	0.20	0.13	0.22	0.10
ElasticNet	0.07	0.14	0.12	0.14	0.42	0.08
Kernel Ridge	0.15	0.14	0.28	0.12	0.51	0.08
Multi Layer Perceptron	0.13	0.19	0.43	0.11	0.45	0.09

TABLE 1 – Comparaison de performance pour les algorithmes de régression.

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
K-Nearest Neighbors	0.35	0.33	0.36	0.44	0.57	0.44
Support Vector Machine	0.20	0.44	0.22	0.44	0.50	0.77
Gradient Boosting	0.36	0.33	**0.52**	0.55	0.82	0.77
Ada Boosting	0.24	0.33	0.25	0.33	0.24	0.44
Decision Tree	**0.42**	**0.56**	0.47	0.59	**0.90**	**0.88**
Random Forest	0.34	0.55	0.50	**0.77**	0.86	**0.88**
Xgboost	0.39	0.44	0.41	0.55	0.70	0.66
Multi Layer Perceptron	0.20	0.55	0.50	**0.77**	0.72	0.77

TABLE 2 – Comparaison des performances des algorithmes de classification.

classe de 10% pendant la phase de maturation (c.-à-d. un mois avant la récolte).

4.3 Discussion

Les résultats obtenus dans cette étude se révèlent particulièrement satisfaisants, surtout lorsque l'on considère les limitations auxquelles nous avons dû faire face. Premièrement, notre ensemble de données d'entraînement était assez restreint, ne couvrant qu'une période de deux ans. Malgré cette contrainte temporelle, les performances de notre modèle ont été prometteuses, ce qui témoigne de son potentiel même avec des données limitées. De plus, un défi majeur auquel nous avons été confrontés était le manque de données météorologiques et de données de sol. Ces informations sont cruciales pour modéliser avec précision les rendements agricoles, mais malheureusement, leur disponibilité était limitée dans notre contexte. Malgré ces obstacles, les résultats que nous avons obtenus soulignent l'efficacité de notre approche méthodologique et suggèrent des possibilités futures pour améliorer encore davantage la prédiction des rendements agricoles dans des conditions de données similaires.

5 Conclusions et Perspectives

Cette étude montre l'application de méthodes d'apprentissage automatique en particulier afin d'améliorer l'estimation des rendements pour des paysages agricoles complexes, en utilisant des images satellites optiques à haute résolution spatiale et temporelle.

Dans un premier temps, nous avons exploré les méthodes de régression pour obtenir des estimations assez précises, avec un R2 score qui atteint 0.83 un mois avant la récolte. Dans un deuxième temps, et pour affiner encore la précision des prédictions, nous avons mis en œuvre des méthodes de classification supervisée. Grâce à cela, nous avons obtenu de bonnes prédictions de rendement des cultures avec une précision de 90% pour la classe de rendement un mois avant la récolte. Enfin, nous pouvons dire que les résultats obtenus par cette étude sont vraiment satisfaisants, car nous ne disposons pas d'un grand ensemble d'entraînement (seulement deux années de données).

Comme perspective, nous avons l'intention d'enrichir cette base de données par l'historique des années subséquentes, et aussi d'essayer de généraliser cette application pour d'autres cultures telles que le blé ou le maïs. Nous prévoyons également de croiser les cartes satellites avec les cartes météorologiques afin d'implémenter des réseaux de neurones profonds et d'étudier leurs comportements pour la prédiction des rendements.

Références

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[2] Amine Chemchem, François Alin, and Michaël Krajecki. Combining smote sampling and machine learning for forecasting wheat yields in france. In *2019 IEEE Second International Conference on Artificial Intelli-*

gence and Knowledge Engineering (AIKE), pages 9–14. IEEE, 2019.

[3] Agricultural Research for Development Cirad. Diversité paysagère et securite alimentaire en Afrique. `https://www.projects.igeo.fr/sites-d-etudes/`, 2018. [Online; accessed 26-February-2020].

[4] Manal Fawzy, Mahmoud Nasr, Samar Adel, and Shacker Helmi. Regression model, artificial neural network, and cost estimation for biosorption of ni (ii)-ions from aqueous solutions by potamogeton pectinatus. *International journal of phytoremediation*, 20(4) :321–329, 2018.

[5] Lilian Hollard, Angelica Durigon, and Luiz Angelo Steffenel. Machine learning forecast of soybean yields on south brazil. In *Workshop on Edge AI for Smart Agriculture (EAISA 2022)*, 2022.

[6] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3) :296–298, 2005.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv :1502.03167*, 2015.

[8] Elisa Kamir, François Waldner, and Zvi Hochman. Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160 :124 – 135, 2020.

[9] K. Kuwata and R. Shibasaki. Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 858–861, July 2015.

[10] Mark N Merzlyak, Anatoly A Gitelson, Olga B Chivkunova, and Victor YU Rakitin. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia plantarum*, 106(1) :135–141, 1999.

[11] Babacar Ndao, Louise Leroux, Abdoul Aziz Diouf, Valerie Soti, and Bienvenu Sambou. A remote sensing based approach for optimizing the sampling strategies in crop monitoring and crop yield estimation studies. In Souleye Wade, editor, *Earth Observations and Geospatial Science in Service of Sustainable Development Goals*, pages 25–36, Cham, 2019. Springer International Publishing.

[12] A. Nyéki, C. Kerepesi, B. Daróczy, A. Benczúr, G. Milics, A.J. Kovács, and M. Neményi1. *Maize yield prediction based on artificial intelligence using spatio-temporal data*, chapter 124, pages 1011–1017. Wageningen Academic Publishers, 2019.

[13] J. Qi, A. Chehbouni, A.R. Huete, Y.H. Kerr, and S. Sorooshian. A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2) :119–126, 1994.

[14] M. Stas, J. Van Orshoven, Q. Dong, S. Heremans, and B. Zhang. A comparison of machine learning algorithms for regional wheat yield prediction using ndvi time series of spot-vgt. In *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–5, July 2016.

[15] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Weicheng Wu. The generalized difference vegetation index (gdvi) for dryland characterization. *Remote Sensing*, 6(2) :1211–1233, 2014.