

Mémoire

Auteur : Snoeck, Lara

Promoteur(s) : Tychon, Bernard

Faculté : Faculté des Sciences

Diplôme : Master en sciences géographiques, orientation global change, à finalité approfondie

Année académique : 2023-2024

URI/URL : <http://hdl.handle.net/2268.2/21522>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



Faculté des sciences
Département de géographie

Prévision des rendements mondiaux du blé : évaluation du potentiel de l'agrométéorologie associée à la télédétection.

Mémoire présenté par : **Lara SNOECK**

pour l'obtention du titre de

**Master en sciences géographiques,
orientation Global Change, à finalité approfondie climatologie**

Année académique :

2023-2024

Date de défense :

Septembre 2024

Président de jury :

Pr. Xavier FETTWEIS

Promoteur :

Pr. Bernard TYCHON

Jury de lecture :

Pr. Sébastien DOUTRELOUP

Pr. Serge SCHMITZ



Faculté des sciences
Département de géographie

Prévision des rendements mondiaux du blé : évaluation du potentiel de l'agrométéorologie associée à la télédétection.

Mémoire présenté par : **Lara SNOECK**

pour l'obtention du titre de

**Master en sciences géographiques,
orientation Global Change, à finalité approfondie climatologie**

Année académique :

2023-2024

Date de défense :

Septembre 2024

Président de jury :

Pr. Xavier FETTWEIS

Promoteur :

Pr. Bernard TYCHON

Jury de lecture :

Pr. Sébastien DOUTRELOUP

Pr. Serge SCHMITZ

Remerciements

Ma gratitude va tout d'abord à l'ensemble du corps professoral de l'Université de Liège, qui a participé à ma formation tout au long de mon cursus académique. Grâce à eux, j'ai pu acquérir les compétences de bases nécessaires à la rédaction de ce mémoire.

Je tiens à remercier tout particulièrement Monsieur Bernard Tychon, promoteur de ce mémoire, pour son enthousiasme à l'égard du sujet traité, pour m'avoir accompagnée tout au long de ce travail et pour avoir pris le temps de répondre à mes questions.

Je remercie également Messieurs Serge Schmitz et Sébastien Doutreloup, lecteurs de mon mémoire, pour le temps consacré à la lecture de ce travail.

Je souhaite exprimer ma gratitude envers Monsieur Antoine Denis et Madame Marie Lang pour leur aide précieuse dans la compréhension et la mise en place des méthodes de modélisation utilisées dans le cadre de ce travail.

Je souhaite remercier l'université de Liège pour m'avoir permis d'accéder à de nombreuses ressources via les bibliothèques ULiège Library.

Je remercie Madame Irina Kovrova de la FAO ainsi que la team ASAP de la Commission européenne pour avoir répondu à mes questions relatives au téléchargement des données.

J'aimerais également exprimer ma reconnaissance à mes amis, pour leur soutien, leurs encouragements, et l'intérêt qu'ils ont porté à ce travail.

Enfin, je tiens à exprimer ma sincère gratitude à ma famille, et tout particulièrement à mes parents, qui m'ont permis d'entreprendre les études que je souhaitais et de les mener à bien dans les meilleures conditions possibles.

Résumé

Le blé est une culture essentielle car il s'agit de la céréale la plus consommée à travers le monde. L'amélioration de la prévision des rendements du blé à l'échelle mondiale pourrait permettre d'aider à la gestion des crises alimentaires en anticipant notamment la mobilisation de l'aide humanitaire ou encore à réduire la spéculation sur les marchés mondiaux. Ce travail a pour objectifs d'utiliser, de paramétrer et d'évaluer différents types de modèles. Un modèle de régression linéaire simple, considéré comme la référence, a été comparé avec des modèles plus complexes, dans lesquels ont été intégrées des variables agrométéorologiques ainsi que le prix du blé sur le marché mondial. Ces modèles sont basés sur les méthodes de régression linéaire multiple et sur les forêts aléatoires. L'analyse a été réalisée à l'échelle nationale des cinq plus grands pays producteurs de blé (Chine, Inde, Russie, Etats-Unis et France) et les résultats ont ensuite été associés pour former un modèle global. Bien que le modèle global n'ait pas surpassé la précision du modèle de référence (avec une racine carrée de l'erreur quadratique moyenne de 108,03 kg/ha contre 74,31 kg/ha), les résultats obtenus sont encourageants et offrent des perspectives d'améliorations dans le cadre de futures recherches.

Mots-clés : prévision des rendements, blé, agrométéorologie, télédétection, régression multiple, forêts aléatoires.

Abstract

Wheat is an essential crop as it is the most consumed cereal worldwide. Improving global wheat yield forecasting could help manage food crises by anticipating the mobilization of humanitarian aid and reducing speculation in global markets. The objectives of this work are to use, parameterize, and evaluate different types of models. A simple linear regression model, considered as the reference, was compared with more complex models that integrated agrometeorological variables as well as the price of wheat on the global market. These models are based on multiple linear regression methods and random forests. The analysis was conducted at the national scale for the five largest wheat-producing countries (China, India, Russia, the United States, and France), and the results were then combined to form a global model. Although the global model did not surpass the accuracy of the reference model (with a root mean square error of 108.03 kg/ha compared to 74.31 kg/ha), the results are encouraging and offer opportunities for improvement in future research.

Keywords: yield forecasting, wheat, agrometeorology, remote sensing, multiple regression, random forests.

Table des matières

Remerciements.....	4
Résumé.....	5
Abstract	6
1. Introduction	9
2. État de l'art.....	10
2.1 Importance de la prévision des rendements agricoles.....	10
2.2 Caractéristiques du blé	10
2.3 Programmes et systèmes de surveillance agricole	11
2.4 Choix des variables pour la modélisation	11
2.5 Méthodes de modélisation des rendements	14
3. Méthodologie.....	20
3.1 Choix de la culture	20
3.2 Choix des pays.....	21
3.3 Données brutes utilisées	22
3.3.1 NDVI	23
3.3.2 NDVI z-score	23
3.3.3 Rainfall.....	23
3.3.4 SPI-3 months (Standardized precipitation index)	23
3.3.5 Temperature.....	24
3.3.6 Water satisfaction index (WSI)	24
3.3.7 Solar radiation	25
3.3.8 Prix	25
3.3.9 Rendement et superficie.....	26
3.3.10 Obtention des données brutes agrométéorologiques.....	26
3.4 Définition de la durée de la saison de croissance du blé	27
3.5 Modélisation	30
3.5.1 Régression linéaire simple	30
3.5.2 Régression multiple.....	31
3.5.3 Random Forest	34
4. Résultats.....	37
4.1 Régression linéaire simple sur le rendement mondial	37
4.2 Chine	38

4.2.1 Régression multiple.....	38
4.2.2 Random forest.....	43
4.3 Inde	45
4.3.1 Régression multiple.....	46
4.3.2 Random forest.....	50
4.4 Russie	52
4.4.1 Régression multiple.....	52
4.4.2 Ranfom forest.....	57
4.5 États-Unis	59
4.5.1 Régression multiple.....	59
4.5.2 Random forest.....	64
4.6 France	66
4.6.1 Régression multiple.....	67
4.6.2 Random forest.....	67
4.7 Comparaison des résultats	69
5. Discussion.....	73
5.1 Discussion des résultats globaux	73
5.2 Discussion des résultats intermédiaires	74
5.3 Discussion de la qualité des données	76
6. Conclusion.....	79
7. Bibliographie	80
8. Annexes.....	84

1. Introduction

Ce travail porte sur une culture essentielle, à savoir, le blé. En effet, il s'agit de la céréale la plus consommée à travers le monde. Cette culture joue un rôle important dans le quotidien des européens et de la population mondiale en général et, d'un point de vue plus personnel, avait déjà été étudiée dans le cadre de mon cursus académique. Des détails plus précis justifiant le choix de cette culture sont fournis dans la partie « méthodologie » de ce travail.

La disponibilité de cette céréale indispensable à l'alimentation de millions d'êtres humains est parfois mise en péril pour diverses raisons (rendements insuffisants ou au contraire surabondants, conflits et tensions géopolitiques, instabilité des prix, ...).

Il existe déjà diverses recherches portant sur la prévision des rendements du blé. Cependant, les recherches à l'échelle mondiale n'intègrent pas ou peu de variables agrométéorologiques. Par ailleurs, il existe aussi des études plus précises qui pour certaines intègrent des variables agrométéorologiques issues de la télédétection mais qui se restreignent à des plus petites zones d'études (échelle nationale, régionale voire même limitée à une seule parcelle).

Dans le cadre de ce mémoire, il s'agira de déterminer si l'on peut améliorer la prévision des rendements du blé à **l'échelle mondiale** en paramétrant des modèles existants basés sur des données agrométéorologiques issues de la télédétection.

Les objectifs poursuivis lors de cette recherche consistent à utiliser et à paramétrer des modèles plus ou moins complexes et à évaluer leur capacité à prévoir le rendement du blé à l'échelle globale en les comparant. Autrement dit, il s'agit de comparer les résultats d'une approche de régression simple, considérée comme la référence, aux résultats de modèles plus sophistiqués afin d'évaluer leur efficacité et leur précision. En améliorant la qualité des prévisions, cette recherche pourrait favoriser la mise au point d'outils destinés aux décideurs politiques et aux agriculteurs, permettant d'anticiper d'éventuelles crises alimentaires ou encore de limiter le lobbying et la spéculation sur les marchés mondiaux relatifs au blé en rendant les prévisions de rendement plus transparentes et plus accessibles.

La suite de ce travail est organisée en différents chapitres. Dans un premier temps, l'état de l'art va être présenté. Ce-dernier passera en revue les recherches antérieures liées au sujet d'étude. Ensuite, la méthodologie utilisée dans le cadre de ce travail sera détaillée et suivie par les résultats obtenus grâce aux modèles exploités. Enfin, ce mémoire se termine par une discussion des résultats obtenus ainsi que des perspectives pour des futures recherches.

2. État de l'art

L'objectif de ce chapitre est de faire une revue des recherches déjà existantes en lien avec la prévision des rendements du blé à l'échelle globale. Plus concrètement, il s'agira de fournir des informations relatives :

- à l'importance des recherches sur la prévision des rendements,
- aux caractéristiques du blé,
- à la pertinence des différents types de données à utiliser,
- aux méthodes de modélisation existantes.

2.1 Importance de la prévision des rendements agricoles

Disposer d'informations précises quant aux prévisions des rendements du blé de manière régulière s'avère être une ressource d'information précieuse. En effet, cela permet de réduire les effets de surprise sur les marchés et donc de diminuer les fluctuations des prix. Ces effets de surprise peuvent notamment survenir suite à des conditions météorologiques particulières dans des régions où le blé est beaucoup cultivé ou encore lorsque les quantités de blé sont en flux tendu par rapport aux capacités de stockage et à la consommation de celui-ci. A titre d'exemple, durant les crises alimentaires de 2007-2008 et de 2011-2012, les prix des denrées alimentaires ont flambé suite à des mauvaises conditions météorologiques dans des régions de production agricole importante. Autre exemple, suite à d'importantes sécheresses en Afrique australe durant les années 2015 et 2016 puis dans l'Est de l'Afrique en 2017, des pénuries alimentaires conséquentes se sont faites ressentir. Ce type d'évènements entraîne de l'insécurité alimentaire ainsi que des tensions d'ordre socio-économique, c'est pourquoi les prévisions sont aussi importantes. Elles permettent d'anticiper des crises et éventuellement de mobiliser de l'aide humanitaire plus rapidement si nécessaire (Franch *et al.*, 2021).

2.2 Caractéristiques du blé

Le blé présente l'avantage d'être une culture polyvalente. En effet, il est possible de le cultiver à différentes altitudes, sur plusieurs types de sols et dans diverses conditions climatiques. De manière générale, les cultures de blé sont situées entre 30° et 60° de latitude nord et entre 27° et 40° de latitude sud. Le blé peut être cultivé jusqu'à 3000 mètres d'altitude. Les conditions idéales pour le développement du blé nécessitent des températures situées entre 3°C et 32°C et un taux de précipitations annuelles compris entre 375 mm et 875 mm. Cependant, il reste possible de cultiver le blé dans des régions recevant plus ou moins de pluie, la fourchette maximale allant de 250 mm à 1750 mm par an (Enghiad *et al.*, 2017).

A l'échelle mondiale, le blé est la culture qui occupe le plus de superficie pour sa production (plus de 240 millions d'hectares) et est aussi la céréale la plus consommée dans le monde. Le blé offre un apport représentant 20% voire 21% des calories et des protéines qui sont consommées à travers le monde. Cette céréale joue donc un rôle prépondérant dans la sécurité alimentaire et dans les systèmes agroalimentaires. Le blé est dès lors la céréale la plus exportée/importée à travers le monde. Pour la période 1980-2013, les pays en

développement, généralement des pays importateurs de blé, représentaient 77% de la consommation de blé à l'échelle globale. La demande mondiale de blé augmente chaque année, en grande partie en raison de l'augmentation démographique, mais aussi grâce à l'augmentation du pouvoir d'achat dans les pays en développement. L'augmentation constante de la production mondiale de blé depuis plusieurs décennies n'est toutefois pas due à une augmentation des terrains cultivés, mais à l'augmentation globale des rendements (Enghiad *et al.*, 2017 ;Erenstein *et al.*, 2022).

2.3 Programmes et systèmes de surveillance agricole

Il faut savoir qu'il existe des ressources qui n'ont pas pour objectif de modéliser ni de prédire les rendements du blé, mais qui fournissent des rapports intéressants sur le blé, contenant des analyses des conditions de production passées et présentes. Il y a entre autres le programme SMIAR (Système mondial d'Information et d'Alerte Rapide) qui appartient à la FAO. SMIAR publie environ 4 rapports par an qui contiennent des analyses relatives aux perspectives de récolte et à la situation alimentaire des derniers mois écoulés de différentes zones géographiques dans le monde. Les rapports se concentrent particulièrement sur « les perspectives de production céréalière, la situation du marché et les conditions de sécurité alimentaire, avec une attention particulière pour les pays à faible revenu et à déficit vivrier » (FAO,2024).

Parmi les systèmes de surveillance agricole, il existe CropWatch qui est un système chinois appartenant à l'Académie des Sciences Chinoises (CAS). Tout comme le SMIAR, CropWatch publie régulièrement des rapports faisant le bilan des conditions de production de diverses cultures dont le blé à travers le monde. Contrairement au SMIAR, CropWatch offre des informations relatives à la prévision des rendements des cultures à l'aide de la modélisation. Les données d'entrées du ou des modèles sont essentiellement d'ordre agrométéorologique puisqu'elles incluent par exemple les températures, les précipitations ou encore des indices relatifs à la santé de la végétation. Ces données sont soit issues de la télédétection, soit de mesures au sol. Néanmoins, aucun détail relatif aux modèles ou aux algorithmes utilisés pour obtenir les résultats présentés dans les rapports ne sont mis à disposition (Lab for Digital Agriculture, RADI, CAS, *n.d.*).

2.4 Choix des variables pour la modélisation

Bien que les programmes et systèmes présentés ci-dessus regorgent d'analyses intéressantes, l'objectif majeur de ce mémoire est d'utiliser des modèles en les paramétrant de manière à permettre de faire de la prévision de rendement et de les comparer entre eux. Avant de faire de la modélisation, il est nécessaire de s'intéresser aux variables potentiellement intéressantes à utiliser. Les études de Basso & Liu et de Hao *et al.*, présentées ci-dessous, apportent des informations intéressantes quant au choix des variables à utiliser pour la modélisation.

En 2019, Basso et Liu ont publié un article dressant un état de l'art des différents types de données à utiliser selon les méthodes de prévision des rendements saisonniers et discutant

notamment de leurs applications, de leur précision, de leurs avantages et inconvénients. (Basso & Liu, 2019).

La première méthode est l'obtention de **données via les enquêtes**. Les données sont généralement directement récoltées auprès des agriculteurs et ne sont pas toujours fiables. Il est possible de collecter les données de 2 manières différentes : soit via des entretiens par téléphone, soit via des enquêtes de terrain. Aux Etats-Unis, le NASS (National Agricultural Statistics Service) emploie ces 2 méthodes (Basso & Liu, 2019).

Pour prédire les rendements, il existe une méthode appelée la modélisation statistique. Présentée comme une méthode générale, elle se décline en sous-catégories. Pour réaliser ce type de modélisation, une régression statistique est faite au départ de données agrométéorologiques afin de prédire les rendements. Ces modèles sont souvent constitués de matrices qui reprennent les valeurs de rendements des années précédentes ainsi que les valeurs d'autres paramètres comme, par exemple, la température et les précipitations. L'avantage de la modélisation statistique est qu'elle est simple à appliquer. Cependant, au vu de sa limitation pour extrapoler les résultats, des changements climatiques de plus en plus conséquents ainsi que de l'augmentation du nombre d'événements extrêmes, il ne s'agit pas du modèle le plus adapté pour faire de la prévision. La modélisation statistique reste toutefois indispensable pour déterminer quelles sont les variables qui ont un rôle à jouer dans la prévision de rendements. (Basso & Liu, 2019).

- La première sous-catégorie de la modélisation statistique repose sur l'utilisation de **données agrométéorologiques**. Les variables indépendantes utilisées dans le modèle peuvent être des données météorologiques, agronomiques ou une combinaison des deux. Pour prédire les rendements d'une culture, plusieurs facteurs sont pris en considération. L'accumulation de la biomasse dépend notamment des conditions météorologiques en fonction de stades de croissance de la culture. Voici une liste assez complète du type de données qui peuvent être prises en compte : les précipitations sur une certaine période, la température, les minimas et maximas de température, les degrés-jours, l'humidité relative, l'épaisseur de neige, le rayonnement solaire, l'évapotranspiration, les heures d'ensoleillement, la vitesse du vent, la teneur en chlorophylle, le LAI (indice de surface foliaire), des variables biométriques (par exemple : la fraction de couvert forestier, la biomasse aérienne, le nombre de tiges ou encore la hauteur des plantes), des variables qui caractérisent le sol (exemples : la teneur en eau dans le sol, le teneur en carbone organique, l'azote, la densité apparente, la capacité au champ ou encore la conductivité électrique), la FAPAR (fraction du rayonnement photo synthétiquement actif). A noter qu'il est aussi possible de créer des indices (par exemple, un indice sur le stress des cultures) au départ des différentes variables déjà citées. Des variables relatives à la circulation atmosphérique peuvent également être intégrées comme par exemple la NAO (North Atlantic Oscillation) ou encore la SO (Southern Oscillation). D'autres paramètres sont parfois pris en compte comme les dégâts provoqués par les insectes, la mécanisation, la gestion, le capital, la main d'œuvre, ... Concernant les paramètres liés à l'agronomie, il y a également divers types de données qui peuvent être intégrés dans le modèle : l'historique sur le

rendement, la superficie cultivée ou encore l'utilisation d'engrais. Il y a des modèles déjà existants qui permettent de prédire le rendement des cultures parmi lesquels on peut citer le modèle WOFOST (World Food Studies) et le CGMS (Crop Growth Modelling System) (Basso & Liu, 2019).

- La seconde sous-catégorie est la modélisation statistique avec les **données de télédétection**. Ce type de modélisation comprend au minimum une variable mesurée par des capteurs éloignés. Il y a deux grands types de capteurs : les capteurs passifs qui dépendent d'une source lumineuse externe (comme le soleil) et les capteurs actifs qui possèdent leur propre source de lumière. Les images satellite les plus souvent utilisées afin de faire de la prévision de rendement sont : SPOT-Vegetation, AVHRR, Landsat et MODIS. Bien souvent, lorsque l'on étudie la végétation, la réflectance de plusieurs bandes spectrales est combinée pour créer des indices de végétation qui permettent de rendre compte de l'état de la végétation. L'indice le plus utilisé est le NDVI (Normalized Difference Vegetation Index). Par exemple, des prévisions du rendement du blé ont déjà pu être faites avec des modèles de régression linéaire reprenant le NDVI, lui-même déduit des images fournies par MODIS, 2 à 3 mois avant la récolte. D'autres indices sont aussi utilisés régulièrement tels que : l'INSEY qui associe le NDVI avec les degrés-jours afin d'estimer le rendement en saison ou encore le VCI (indice de l'état de la végétation) qui est fonction de la valeur du NDVI à une certaine date et des valeurs historiques du minimum et du maximum du NDVI (Basso & Liu, 2019).

Une autre méthode possible pour faire de la prévision de rendements est l'utilisation de **simulations des cultures**. Ces modèles permettent d'obtenir des informations sur la biomasse ainsi que sur le rendement au départ des données relatives à la météorologie, à la nature des sols, aux cultures et à leur gestion. Bien souvent, le cycle des nutriments ainsi que l'influence de l'eau sont également intégrés (Basso & Liu, 2019).

D'après cette étude, dans la littérature qui cherche à prédire le rendement, la télédétection est utilisée dans à peu près la moitié des recherches et l'agrométéorologie dans un tiers des cas. Il a été mis en évidence qu'intégrer des données de télédétection permet d'améliorer les prévisions des rendements. Il faut toutefois noter que lors de son développement, le blé a besoin de pluie or les nuages empêchent les observations par imagerie satellite. Il y a donc des limites à l'emploi de la télédétection. Les modèles statistiques présentent des limites dont le fait de ne généralement pas être transposables d'une région à une autre (Basso & Liu, 2019).

Il convient de préciser que dans le cadre de ce travail, les variables utilisées seront pratiquement toutes des données d'ordre agrométéorologiques obtenues grâce à la télédétection. Les modèles basés sur les simulations des cultures ne seront pas utilisés et les enquêtes non plus.

2.5 Méthodes de modélisation des rendements

En 2021, une étude visant à prédire le rendement du blé à l'échelle mondiale à l'aide du modèle « APSIM Classic Wheat » (Agricultural Production Systems SIMulator), un modèle de simulation agricole, a été publiée. Cette étude a pour objectif d'évaluer la qualité des prévisions ainsi que d'identifier quelles sont les variables qui influencent la qualité du modèle (Hao *et al.*, 2021).

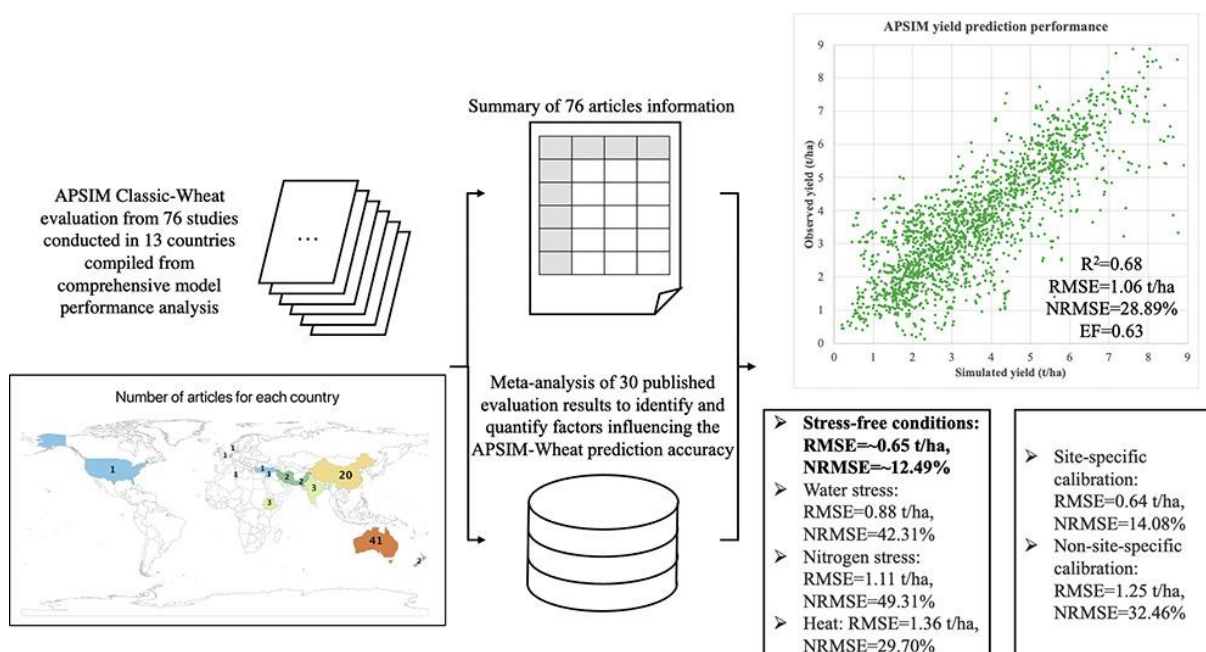


Figure 1. Résumé de l'étude de Hao et al. (Hao *et al.*, 2021).

Comme représenté sur la Figure 1, les analyses sont basées sur 76 études qui ont été publiées dans 13 pays différents au total. L'APSIM¹ est un modèle cultural qui repose sur divers processus biophysiques. Il permet de simuler la croissance du blé, les processus dans le sol, les pratiques de gestion des cultures ainsi que les conditions climatiques dans lesquelles elles sont cultivées. Il a été montré que pour certaines régions, le modèle est capable de prédire le rendement avec une erreur quadratique moyenne (RMSE) inférieure à 1 tonne par hectare. Cependant, dans les régions connaissant des conditions plus instables comme par exemple la présence de stress hydrique ou encore les limitations de la quantité d'azote disponible, les résultats deviennent moins précis. Dès lors, les variables ayant une influence dans la précision des prévisions ont été listées. Le premier facteur ayant une influence est le calibrage spécifique au site étudié. Les résultats donnés par le modèle sont de meilleure qualité lorsque des mesures directes existent pour calibrer le modèle. Ces mesures peuvent être relatives au sol ou bien à la culture. A titre d'exemple, il a pu être montré que l'intégration au modèle de mesures directes des paramètres hydrauliques du sol permettaient d'améliorer la précision des résultats de façon conséquente. Le second facteur est le stress hydrique. Les chercheurs ont démontré que le modèle avait tendance à faire des surestimations en cas de stress hydrique. Cela est dû au fait que le modèle n'intègre pas convenablement tous les effets qui découlent de la présence d'un déficit hydrique. La troisième variable ayant un impact sur la

¹ L'utilisation du modèle APSIM est uniquement possible sous licence, des frais d'utilisation peuvent s'appliquer selon les cas. (APSIM, 2024).

précision des prévisions est le stress thermique. L'impact des vagues de chaleur sur les cultures est généralement mal simulé et sous-estimé par le modèle rendant ainsi les prévisions incorrectes. Enfin le dernier facteur qui pose problème est le stress en azote. En l'absence d'une quantité suffisante de ce nutriment dans le sol pour répondre aux besoins de plantes, le modèle a tendance à sous-estimer les rendements car il surestime la durée des épisodes de stress (Hao *et al.*, 2021).

Les suggestions d'amélioration proposées à la fin de l'étude sont les suivantes :

- Intégration de données externes au modèle, comme, par exemple, des données de télédétection, afin d'améliorer les simulations dans des conditions instables,
- Amélioration des modules du modèle destinés à simuler des conditions de stress, qu'il soit hydrique, thermique ou azoté (Hao *et al.*, 2021).

Comme déjà mentionné, dans le cadre de ce mémoire, ce seront essentiellement des données issues de la télédétection qui seront intégrées dans la modélisation. En revanche, aucune variable relative aux techniques de gestion des cultures ni à la quantité de nutriments présents dans le sol ne seront intégrées.

Une autre étude, menée par Franch *et al.*, présentée ci-après, présente un algorithme permettant de faire des prévisions de rendement du blé. Cet algorithme présente des résultats intéressants, cependant, assez peu de variables ont été intégrées au modèle. L'article, paru en 2021, présente l'algorithme ARYA (Agriculture Remotely-sensed Yield Algorithm) qui permet de faire des prévisions des rendements du blé. Les données d'entrée du modèle sont l'indice de végétation (DVI) et les degrés-jours. Une évaluation de l'amélioration des prévisions en intégrant des données relatives à la température à la surface des terres (LST) a également été réalisée. Afin de tenir compte des conditions de stress telles que des températures extrêmes ou un déficit hydrique, une comparaison entre la LST et la température de l'air a été faite. Cette comparaison a permis de mieux mettre en évidence l'impact de la présence d'une sécheresse ou de conditions de gel. L'étude a été réalisée sur une plage de données allant de 2001 à 2019 sur les 7 plus gros pays exportateurs de blé, à savoir, les Etats-Unis, la Russie, l'Ukraine, la France, l'Allemagne, l'Australie et l'Argentine. A eux-seuls, ces pays représentent un peu plus de 70% des exportations mondiales de blé (Franch *et al.*, 2021).

Cette étude a été mise en place en réaction au manque d'informations disponibles sur les sujets suivants : manque d'informations quantitatives issues de l'observation de la terre, absence de prévisions détaillées à l'échelle globale et manque de données sur les erreurs potentielles des prévisions (Franch *et al.*, 2021).

Les prévisions rendues par ARYA 2 mois voire 2 mois et demi avant la récolte ont une RMSE de l'ordre de 200 à 400 kg/ha à l'échelle nationale et de l'ordre de 500 à 700 kg/ha à l'échelle infranationale. Il a pu être démontré que l'intégration de la LST aux données d'entrée du modèle permettait d'améliorer la précision des résultats, et ce, en particulier pour 3 des pays étudiés (l'Australie, l'Ukraine et l'Argentine). Suite à des conditions climatiques particulières

telles que les sécheresses ou encore des gelées tardives, la qualité des prévisions du modèle se retrouve significativement diminuée (Franch *et al.*, 2021).

Les propositions d'amélioration présentées à la fin de la recherche pré-décrite sont les suivantes :

- Intégration d'autres variables météorologiques, comme par exemple les précipitations, afin d'augmenter d'avantage la précision des résultats,
- Utilisation des réseaux neuronaux géographiquement pondérés afin d'améliorer le modèle (Franch *et al.*, 2021).

Comme suggéré dans les pistes d'améliorations ci-avant mentionnées, ce travail intégrera d'avantage des variables météorologiques.

Les résultats de deux autres études, centrées sur la modélisation utilisant des méthodes d'apprentissage automatiques, vont à présent être présentées.

- La première étude, menée par Paudel *et al.*, portant sur l'utilisation de « l'apprentissage automatique pour la prévision des rendements des cultures à grande échelle », est parue en 2021. La méthode d'apprentissage automatique présente de nombreux avantages comme la modélisation de relations non linéaires et la réutilisation de résultats d'autres méthodes. Lorsqu'un grand jeu de données est à disposition, les résultats obtenus sont généralement meilleurs. De plus, ils peuvent permettre de réduire le bruit dans les données (Paudel *et al.*, 2021).

L'objectif de cette étude est de trouver une méthodologie de prévision des rendements qui soit transposable à différentes cultures et à différentes localisations. La prévision des rendements a été déterminée à la fois en début et en fin de saison en se focalisant sur 3 points importants : « l'exactitude, la modularité et la réutilisabilité » (Paudel *et al.*, 2021).

L'étude porte sur 3 pays (l'Allemagne, les Pays-Bas et la France) et sur 5 céréales (le blé tendre, l'orge de printemps, le tournesol, la betterave sucrière et la pomme de terre). Les données utilisées sont les données MCYFS (MARS Crop Yield Forecasting System) et Eurostat. Le point de repère, appelé méthode nulle dans le cadre de l'étude, consiste à réaliser des prévisions avec une méthode très simple : soit avec une tendance linéaire, soit avec une moyenne d'ensemble d'entraînement (Paudel *et al.*, 2021).

D'après l'étude de Paudel *et al.*, il existe 4 techniques différentes afin de prédire le rendement du blé. Ces techniques, qui peuvent être combinées entre elles sont les suivantes : les enquêtes de terrain, les modèles de croissance du blé, la télédétection et les modèles statistiques. Bien que les enquêtes de terrain reflètent au mieux la situation réelle, les informations obtenues peuvent être limitées et contenir des erreurs. Les modèles de croissance ne tiennent pas compte de tous les facteurs de réduction des rendements et ne fonctionnent convenablement que si des données suffisamment précises leurs sont fournies. La télédétection donne quant à elle des mesures indirectes du rendement et doit souvent être associée à d'autres modèles pour pouvoir faire des prévisions de rendement. Enfin, les modèles statistiques

combinent les résultats obtenus par les méthodes citées ci-dessus avec des données météorologiques. Ces modèles sont souvent limités d'un point de vue spatio-temporel et ne sont donc pas transposables à toutes les situations. Dans le domaine agricole, diverses méthodes ont déjà été utilisées pour prédire le rendement des cultures. D'une part, les méthodes traditionnelles telles que Random Forest et d'autres modèles d'apprentissage automatique, et d'autre part, les méthodes d'apprentissage profond telles que les réseaux neuronaux, l'apprentissage par représentation et les cartes auto-organisées (Paudel *et al.*, 2021).

Le premier point sur lequel l'étude s'est concentrée est l'exactitude. Il faut veiller à éviter la fuite d'informations c'est-à-dire l'utilisation d'une partie des données test en tant que données d'entraînement ce qui va induire des erreurs. Pour ce faire, une attention particulière a été portée à la conception des données explicatives et à la manière de leur appliquer l'apprentissage automatique. A partir du modèle WOFOST, 6 étapes de la croissance des plantes ont été déterminées sur une période de 3 décades. Les données proviennent de différentes sources à savoir la télédétection, la météo et la croissance des plantes. En fonction des indicateurs, ont été analysées, soit la moyenne, soit la valeur cumulative totale, soit la durée (pour les variables relatives aux événements extrêmes) (Paudel *et al.*, 2021).

70 % des données sont utilisées pour l'entraînement du modèle et 30 % pour le test, en plaçant les années les plus récentes dans le groupe de test. Pour optimiser la sélection des variables et les algorithmes de prévision, des techniques adaptées ont été employées en fonction de la présence ou non d'une tendance dans les rendements. Un autre point clé est la modularité du modèle, ce qui facilite l'ajout ou la modification de fonctionnalités sans affecter les autres étapes du processus (Paudel *et al.*, 2021).

Pour la réutilisabilité, un travail d'homogénéisation des données a été effectué. Des tests ont été réalisés pour trois pays à deux niveaux spatiaux (appelés NUTS 2 et 3) et pour différentes cultures. Quatre algorithmes d'apprentissage automatique ont été évalués et se prénomment comme suit : régression de Ridge, k plus proches voisins, machines à vecteurs de support, et arbres de décision boostés par gradient. Les caractéristiques retenues varient selon le stade de croissance, avec une importance particulière pour la rétention d'eau, les températures, les précipitations, FAPAR, la biomasse et le LAI. La Figure 2 récapitule la méthodologie globale suivie dans le cadre de l'étude (Paudel *et al.*, 2021).

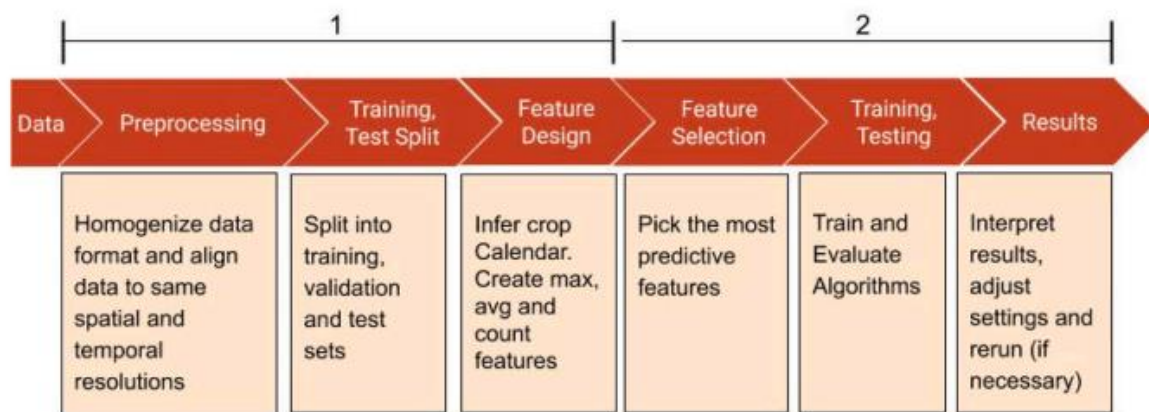


Figure 2. Méthodologie de l'étude de Paudel et al. (Paudel et al., 2021).

Les erreurs similaires entre l'utilisation ou non de la tendance du rendement montrent que l'apprentissage automatique fournit de meilleurs résultats que la méthode nulle quoi qu'il en soit. Les prévisions de MCYFS, qui intègrent des ressources supplémentaires comme des reportages agricoles, s'améliorent davantage en fin de saison (Paudel et al., 2021).

Dans l'étude, différents points d'amélioration sont suggérés tels que :

- Améliorer la qualité des données en trouvant les valeurs aberrantes ainsi que les doublons
 - Améliorer le modèle d'apprentissage automatique en optimisant tous ces constituants
 - Ajouter des nouvelles ressources
 - Préciser la conception des paramètres à l'aide de nouvelles données (par exemple : amélioration du calendrier des cultures, définition des seuils en fonction des cultures)
 - Créer des fonctionnalités plus poussées en intégrant des données météorologiques et pédologiques des années passées afin de mettre en évidence l'évolution des modèles de culture (Paudel et al., 2021).
- La seconde étude portant sur la prévision des rendements agricoles de diverses cultures en utilisant Random Forest est parue en 2016. Cette recherche, réalisée par Jeong et al. porte sur le blé mais également sur le maïs et la pomme de terre. Elle a pour but d'évaluer la capacité de la prévision des rendements avec Random Forest (méthode d'apprentissage automatique) vis-à-vis des régressions multiples, considérées comme des références dans le cadre de cette étude. Les résultats démontrent une meilleure performance des modèles construits avec les forêts aléatoires comparativement aux régressions linéaires multiples (Jeong et al.; 2016).

La modélisation pour la culture du blé a été réalisée pour l'échelle globale, contrairement aux autres cultures, et intègre différentes variables : la température, les précipitations, l'évapotranspiration annuelle, la durée du jour le plus long de l'année

(aussi appelée photopériode), et le taux d'engrais azotés utilisé. Les données utilisées pour le rendement du blé sont issues de méga-environnements définis par le CIMMYT (International Maize and Wheat Improvement Center). Ces vastes régions possèdent des caractéristiques qui leur sont propres telles que leurs conditions environnementales incluant le type de sol, le climat ainsi que différents types de stress qu'ils soient biotiques ou abiotiques. Les données utilisées ont été interpolées au départ des bases de données mises à disposition par WorldClim (Jeong *et al.*; 2016).

Le modèle Random Forest a un coefficient de détermination de 0.96. Autrement dit, 96% de la variance du rendement est expliquée par le modèle pour la culture du blé. La racine carrée de l'erreur quadratique moyenne vaut 320 kg/ha. Cette erreur est inférieure à celle du modèle de régression linéaire multiple puisque la RMSE obtenue pour ce modèle est de 1320 kg/ha. Les résultats obtenus grâce aux forêts aléatoires montrent que les variables les plus influentes sont : le taux d'application d'engrais azotés, l'évapotranspiration annuelle et la photopériode. Les auteurs mettent toutefois en évidence que, bien que Random Forest soit un modèle robuste et permette d'obtenir des résultats précis, sa performance diminue lorsque qu'il faut faire des prévisions en dehors de la plage de données utilisée pour l'entraînement du modèle (Jeong *et al.*; 2016).

Dans le cadre de ce travail, la modélisation réalisée va être basée sur les régressions linéaires multiples et sur les forêts aléatoires. Comme l'a démontré l'étude de Jeong *et al.*, ce sont des techniques de modélisation prometteuses qui laissent espérer l'obtention de résultats robustes et précis.

3. Méthodologie

Dans cette partie, divers éléments méthodologiques vont être successivement analysés :

- Le choix de la culture
- Le choix des pays étudiés
- Les données brutes utilisées
- La définition de la durée de la saison de croissance du blé
- La modélisation

3.1 Choix de la culture

Ce travail va porter sur une des cultures essentielles. Pour déterminer laquelle, il a fallu faire un choix entre le maïs, le blé et le riz.

C'est le blé qui a été choisi pour diverses raisons. Tout d'abord, d'un point de vue plus personnel, le blé est une céréale familière dans de nombreux pays, y compris en Belgique, pays de résidence de l'auteure, où le blé est cultivé et consommé couramment. De plus, cette céréale avait déjà été étudiée dans le cadre du cursus universitaire de l'auteure pour un cours relatif à l'agrométéorologie. De plus, ce mémoire étant rédigé en français, il se destine à un public également familier avec la culture de cette céréale. Récemment, la guerre en Ukraine a permis de mettre en lumière l'importance de cette céréale dans le quotidien de bon nombre de personnes. Il est dès lors d'autant plus pertinent de s'intéresser aux prévisions du rendement du blé. Il faut toutefois noter que ce mémoire vise à prédire les rendements à l'aide de l'agrométéorologie associée à la télédétection. Il ne se focalise donc pas sur l'impact des conflits géopolitiques sur la sécurité alimentaire dans le monde.

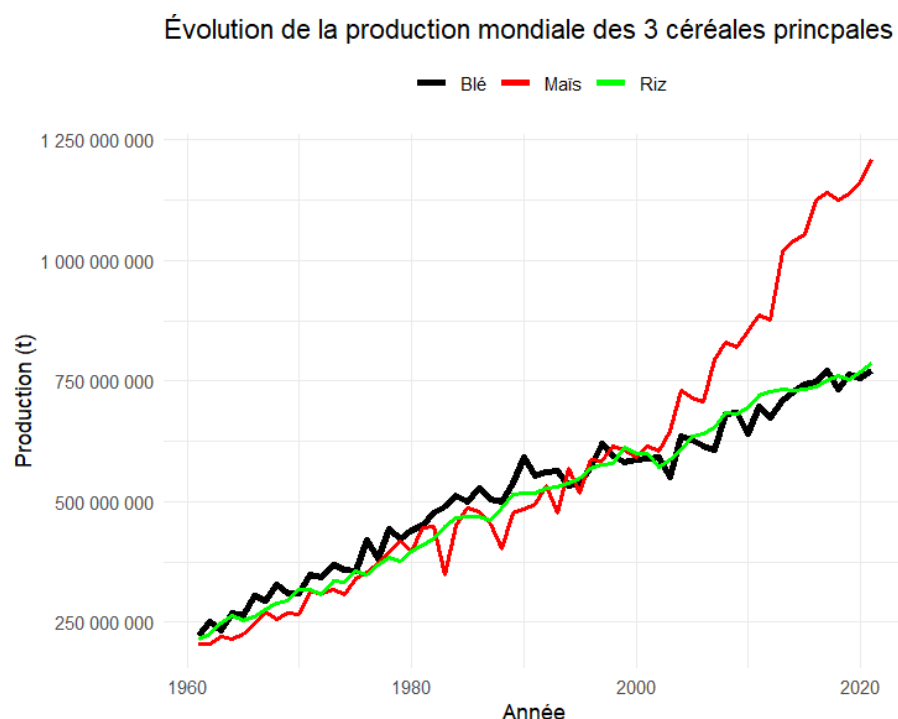


Figure 3. Évolution de la production mondiale des 3 céréales principales (source de données : FAO, 2023).

Le riz n'a pas été retenu car il est moins consommé en Europe et n'y est pas cultivé. Cela le rend moins impactant dans la vie quotidienne des potentiels lecteurs de ce travail. En ce qui concerne le maïs, il est certes cultivé en Belgique et en Europe, cependant, comme le montre la Figure 3, l'évolution de la production mondiale du blé au cours des dernières décennies est bien plus linéaire que celle du maïs. Cette stabilité plus importante pour le blé semble offrir de meilleures perspectives pour obtenir des prévisions de rendement, ce qui a motivé ce choix.

3.2 Choix des pays

Pour définir sur quels pays travailler, dans la mesure où les contraintes de temps ne permettent pas de travailler sur tous les pays du monde et où il n'est de toute façon pas nécessaire de tous les prendre en compte pour avoir des données représentatives de la situation mondiale, une première pré-sélection a été réalisée. Un cumul de la production de blé de chaque pays à l'échelle mondiale sur une période de 15 ans (2007 à 2021) a été réalisé. Les 20 plus gros pays producteurs suite à ce classement sont représentés à la Figure 4. Ce sont surtout les 4 premiers pays qui se démarquent à savoir : la Chine, l'Inde, la Russie et les Etats-Unis. Un tableau récapitulant la production cumulée de chacun des 20 pays sur 15 ans ainsi qu'un tableau montrant l'évolution du top 20 des pays les plus producteurs de blé par année entre 2007 et 2021 sont disponibles aux annexes 1 et 2.

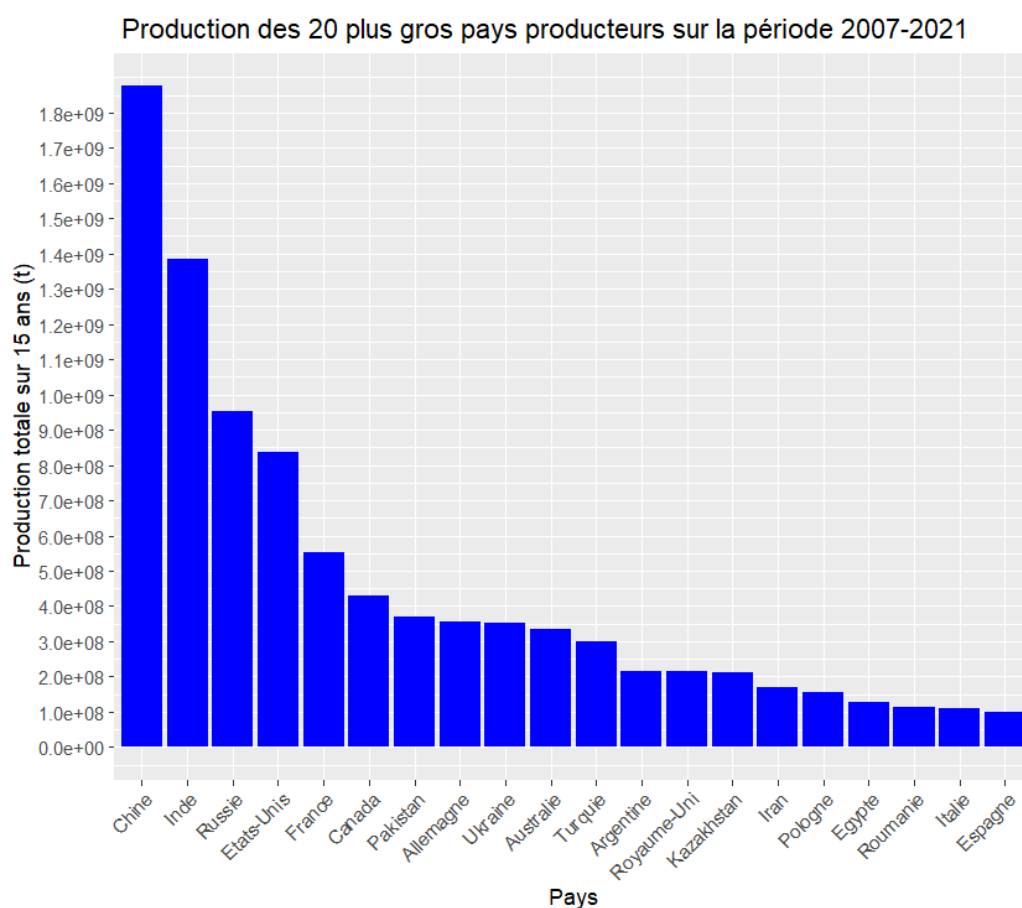


Figure 4. Production des 20 plus gros pays producteurs sur la période 2007-2021 (source de données : FAO, 2023).

Parmi ce classement, les 17 premiers pays ont la particularité d'être tous apparus chaque année dans le top 20 des plus gros pays producteurs de blé (cf. Annexe 2). Il pourrait donc être intéressant de restreindre la liste des pays à étudié à ces 17 pays, qui sont, dans l'ordre : la Chine, l'Inde, la Russie, les États-Unis, la France, le Canada, le Pakistan, l'Allemagne, l'Ukraine, l'Australie, la Turquie, l'Argentine, le Royaume-Uni, le Kazakhstan, l'Iran, la Pologne et l'Égypte.

Pour d'affiner encore davantage cette liste, des régressions linéaires ont été appliquées aux valeurs de rendements afin de déceler des tendances particulières, mais les résultats se sont avérés peu concluants en raison du risque de surajustement dû à la trop faible quantité de données. Une tentative de régression multiple prenant en compte le rendement, la production et la surface récoltée n'a également pas permis d'obtenir des résultats pertinents.

Finalement, la meilleure méthode pour sélectionner les pays s'est avérée être l'une des plus simple : examiner la part de la production mondiale représentée par différents groupes de pays. Ainsi en cumulant les valeurs de productions sur 15 ans (de 2007 à 2021), prendre au moins 50 % de la production mondiale revient à tenir compte des 5 premiers pays de la liste (Chine, Inde, Russie, États-Unis et France). Ces pays représentent très exactement 52,35 % de la production de blé mondiale pour la période 2007-2021. Les 17 premiers pays préalablement listés représentent quant à eux un peu plus de 80% de la production de blé mondiale, 82,62 % très exactement. Ce sont finalement les 5 premiers pays ayant produit le plus de blé sur la période de 15 ans considérée qui ont été retenus pour la suite de l'étude.

3.3 Données brutes utilisées

La plupart des données brutes utilisées pour la modélisation dans ce travail sont fournies par la Commission Européenne. Elles proviennent plus spécifiquement de la section ASAP (Anomaly Hotspots of Agricultural Production) qui regroupe divers jeux de données reprenant des informations relatives à l'agriculture ainsi qu'au climat (European Commission, 2024).

Sur le site d'ASAP, des jeux de données relativement complets sont fournis pour différents paramètres par région toutes les décades à partir de 2002 pour l'ensemble des pays du monde. Deux types de découpages spatiaux sont disponibles. Ils sont nommés « gaul1 » et « gaul2 ». Le premier découpage est celui qui a été utilisé et il consiste en un découpage par pays de grandes entités administratives. Par exemple, pour la Chine, le découpage est fait selon les provinces du pays tandis que pour l'Inde, il est réalisé selon les états. Différentes variables sont disponibles selon 4 classes possibles. Seuls deux sont pertinentes pour cette étude : « crop » et « crop during growing cycle ». Ainsi, les données font référence aux cultures de manière générale ou bien uniquement durant leur période de croissance. Dans le cas de la classe « crop during growing cycle », seuls les pixels avec des cultures en développement sont pris en compte alors que pour la classe « crop » tous les pixels faisant partie d'une zone agricole sont intégrés que des cultures y soient présentes ou non. Au vu de la faible différence de résultats obtenue entre les jeux de données des deux différentes classes, les différentes variables utilisées ont été enregistrées selon la classe « crop during growing cycle » qui semble être la plus pertinente des deux. Il est important de noter que depuis fin mars 2024, une nouvelle section appelée « yield forecast » a été créée sur le site d'ASAP. Cette section est toujours en

cours de développement et porte pour l'instant sur la prévision des rendements à l'aide de données agrométéorologiques dans les pays sujets à l'insécurité alimentaire. En raison de sa création récente et de sa mise en œuvre progressive, cette section n'a pas été prise en compte dans le cadre de ce travail (European Commission, 2024).

Concernant les variables en tant que telles, ci-dessous est présentée la liste des paramètres retenus pour la modélisation ainsi qu'une description de chacun d'eux.

3.3.1 NDVI

Cet indicateur, dont l'acronyme signifie « Normalized Difference Vegetation Index », reflète la santé de la végétation ainsi que sa présence. Il varie entre -1 et +1. Au plus il est élevé, au plus la végétation est luxuriante et verte.

Cet indice est calculé selon la formule suivante :

$$NDVI = \frac{PIR - VIS}{PIR + VIS}$$

PIR : proche infrarouge, aussi noté « NIR »

VIS : visible (plus précisément le rouge), aussi noté « rouge » ou « red » (Sergieieva, 2023).

3.3.2 NDVI z-score

Le NDVI z-score est l'anomalie standardisée du NDVI à une certaine décade (période de 10 jours) de l'année. Si la valeur de la variable est supérieure à 0, cela signifie que pour la date observée, la valeur du NDVI est supérieure à la moyenne. Si elle est négative alors la valeur de l'indice est en dessous de la moyenne (European Commission, 2024).

3.3.3 Rainfall

La variable rainfall indique le cumul de précipitations survenues sur une décade (European Commission, 2024).

3.3.4 SPI-3 months (Standardized precipitation index)

Le SPI-3 months est l'anomalie de précipitations qui sont cumulées sur les 3 mois précédant la décade observée. L'intérêt de cet indice est qu'il permet de refléter les conditions de précipitations saisonnières. Lorsque le SPI est positif, cela signifie que, comparativement à la médiane historique, les conditions sont plus humides. Par exemple, une valeur de +1 indique des conditions relativement humides tandis qu'une valeur de +2 est le reflet de conditions très humides. A l'inverse, pour la valeur de -1, les conditions sont relativement sèches et lorsque le SPI atteint une valeur de -2, elles sont très sèches (State of Indiana, 2024).

3.3.5 Temperature

La variable temperature indique la température moyenne de l'air pour la période de 10 jours (décade) (European Commission, 2024).

3.3.6 Water satisfaction index (WSI)

Le WSI est un indicateur qui renseigne sur la quantité d'eau disponible dans le sol pour le bon développement des cultures durant leur période de croissance. Cet indice se base sur un bilan hydrique qui tient compte des précipitations ainsi que de l'évapotranspiration. Le WSI s'exprime en pourcents. Lorsqu'il vaut 0 %, cela signifie qu'il n'y a pas eu du tout d'eau disponible pour la plante. Quand les 100 % sont atteints, la plante a alors reçu toute l'eau nécessaire pour vivre et bien se développer autrement dit, elle ne subit aucun stress hydrique. La valeur du WSI est calculée grâce au modèle Crop Specific Soil Water Balance (CSSWB) fournit par la FAO.

Cet indice est calculé de la manière suivante :

$$WSI = 100 \cdot \frac{\sum_{i=1}^{DOI} AETc_i}{\sum_{i=1}^{DOI} PETc_i}$$

$\sum_{i=1}^{DOI} AETc_i$: Somme de l'évapotranspiration réelle des cultures (c'est-à-dire l'eau qui est effectivement utilisée par les cultures) sur la période de croissance jusqu'au jour d'intérêt (Day Of Interest).

$\sum_{i=1}^{DOI} PETc_i$: Somme de l'évapotranspiration potentielle des cultures (c'est-à-dire l'eau qui est nécessaire pour répondre aux besoins des cultures) sur la période de croissance jusqu'au jour d'intérêt (Boogaard et al., 2019).

Informations complémentaires sur la formule du WSI :

L'AETc dépend de la quantité d'eau disponible dans le sol par rapport à la quantité d'eau dont les plantes ont besoin. Cette quantité d'eau disponible est déterminée à l'aide de 2 formules intermédiaires :

Le bilan hydrique : $W_i = \min(SWS, W_{i-1} + P_i - AETc_i)$

W_i : Eau disponible à la fin de la décade i

P_i : Précipitations cumulées durant la décade i

$AETc_i$: Evapotranspiration réelle pendant la décade i

La capacité de stockage d'eau dans le sol : $SWS = (FC - WP) \cdot RD$

SWS : Capacité de stockage d'eau dans le sol

FC : Capacité au champ

WP : Point de flétrissement

RD : Profondeur d'enracinement (Boogaard et al., 2019).

3.3.7 Solar radiation

La variable solar radiation représente la quantité de radiation solaire entrante cumulée sur une période de 10 jours (European Commission, 2024).

Remarque : contrairement à tous les indices détaillés précédemment, la variable solar radiation n'est pas disponible dans la classe « crop during growing season », mais uniquement dans la classe « crop » (European Commission, 2024)².

3.3.8 Prix

Après avoir visualisé l'évolution du prix du blé et du rendement entre 2007 et 2021 (cf. Annexes 3 et 4) il est apparu intuitivement qu'il pourrait exister un lien entre les deux variables. Cela sous-entend que le prix pourrait faire partie des variables ayant une influence sur l'évolution du rendement mondial. Afin d'évaluer ce lien statistiquement parlant, un coefficient de corrélation de Pearson³ a été calculé entre le prix et le rendement. Ce coefficient vaut -0,45, synonyme d'une corrélation négative modérée entre les 2 variables. La p-value de ce test est de 0,089. Elle est donc un peu supérieure au seuil conventionnel de 0,05. Cependant, en acceptant un seuil de 10 %, l'hypothèse nulle peut être rejetée ce qui rend le coefficient de corrélation significatif. L'hypothèse de l'existence d'un lien entre le prix et le rendement est donc confirmée avec une certitude modérée. Combiné à d'autres variables, le prix pourrait jouer un rôle dans la modélisation du rendement c'est pourquoi cette variable a été retenue pour la suite des analyses.

À cause du manque de données détaillées et gratuites disponibles sur le prix du blé, une seule source de données relatives au prix a été utilisée dans le cadre de la modélisation. Cette source est nommée MacroTrends et fournit des données historiques journalières du prix du blé sur le marché mondial. Cependant, il faut tout de même noter que ces données ne sont pas disponibles à toutes les dates. Il manque plusieurs dates qui varient d'une année à l'autre (MacroTrends, 2024).

Pour intégrer cette variable un peu à part dans la modélisation, c'est toujours le prix au moment de la récolte (par la suite ce moment sera aussi appelé la fin de la saison) qui a été considéré. Ce prix impacte directement les agriculteurs puisqu'il s'agit du moment où ils sont le plus susceptibles de vendre leur blé. Le prix enregistré pour une année est en fait le prix sur le marché mondial au moment de la récolte l'année précédente. A titre d'exemple, pour l'année 2010, le prix repris pour la modélisation est celui de la fin de saison de l'année 2009. Cette façon de faire permet de tenir compte des bénéfices ou des pertes qui ont été faits l'année précédente. Si un agriculteur réalise des bénéfices conséquents grâce à son blé une certaine année, l'année suivante, il disposera de davantage de ressources lui permettant d'améliorer la qualité de son blé. Il pourrait par exemple acheter plus d'engrais et de meilleure qualité ce qui contribuerait au rendement des cultures.

² Depuis juillet 2024, une nouvelle variable est disponible sur ASAP : la FAPAR (Fraction absorbée de rayonnement photo synthétiquement actif).

³ La formule permettant de calculer ce coefficient est disponible à l'annexe 5.

3.3.9 Rendement et superficie

Les valeurs historiques des rendements du blé ainsi que des superficies cultivées avec cette même céréale à l'échelle nationale ont été obtenues dans les bases de données de la FAO (FAO, 2023).

3.3.10 Obtention des données brutes agrométéorologiques

Les variables agrométéorologiques présentées précédemment ont pu être déterminées grâce à 3 capteurs différents : MODIS, ECMWF-CHIRPS et ECMWF. Le tableau ci-dessous présente les associations entre ces capteurs et leurs variables correspondantes :

Tableau 1. Répartition des variables en fonction des capteurs (Source de données : European Commission, 2024).

MODIS	ECMWF-CHIRPS	ECMWF
NDVI	Rainfall	Temperature
NDVI z-score	SPI-3 months	Solar Radiation
	WSI	

Une brève description de ces différents capteurs va être présentée ci-après.

1. MODIS

MODIS est l'acronyme de Moderate-Resolution Imaging Spectroradiometer. Ce capteur est un spectromètre et radiomètre imageur. Pour ce qui est de ses canaux spectraux, il possède 36 bandes spectrales dont les longueurs d'ondes varient entre 0,4 μm et 14,4 μm . La résolution spatiale de l'appareil varie entre 250 m et 5600 m. Concernant la résolution temporelle, elle se situe entre 1 et 2 jours. MODIS est embarqué à bord de 2 satellites : Terra, actif depuis 1999, et Aqua, opérationnel depuis 2002. Ces satellites font tous deux partie du programme Earth Observing System de la NASA. MODIS permet entre autres de contrôler l'état de santé de la végétation, mais également d'observer les changements d'occupation du sol ou encore de réaliser des analyses sur les nuages pour ne citer que quelques exemples (NASA, n.d.).

2. ECMWF

L'ECMWF (European Centre for Medium-Range Weather Forecasts) est un institut de recherche européen sur la météorologie. Il produit des prévisions météorologiques à l'échelle mondiale à l'aide de supercalculateurs et possède d'importantes archives de données météorologiques. Cet institut soutient diverses organisations telles que l'OMM (l'Organisation Météorologique Mondiale) ou encore Copernicus (programme européen destiné à l'observation de la Terre). L'ECMWF joue un rôle important au sein de Copernicus puisque l'institut gère 2 services importants de Copernicus. Ces services sont :

- **CAMS (Copernicus Atmosphere Monitoring Service)** qui comme son nom l'indique analyse la composition atmosphérique mais également la qualité de l'air et le rayonnement solaire.

- **C3S (Copernicus Climate Change Service)** qui fournit des données sur le climat passé et présent ainsi que des prévisions sur le futur dans le but d'aider les politiques à la prise de décisions en lien avec le réchauffement climatique.

Les principales missions de l'ECMWF sont : « fournir des prévisions météorologiques mondiales 4 fois par jour, analyser la qualité de l'air et la composition atmosphérique, surveiller le climat et analyser la circulation océanique, faire des prévisions hydrologiques et des risques d'incendies » (ECMWF, n.d.).

3. CHIRPS

CHIRPS, qui signifie Climate Hazards InfraRed Precipitation with Station data, est un ensemble de données sur les précipitations terrestres qui se produisent entre 50° de la latitude nord et 50° de latitude sud. La résolution spatiale de ces données varie entre 0,05° et 0,1° et la résolution temporelle varie entre 6 heures jusqu'à une saison complète. Ces variations de résolution dépendent à la fois des régions et de l'époque à laquelle les données ont été enregistrées. Les premières données remontent à 1981 et vont jusqu'au présent. L'ensemble des données comprend des mesures faites au sol via des stations et des données satellitaires. L'objectif principal des données CHIRPS est de pouvoir surveiller les sécheresses (Touma *et al.*, 2023).

3.4 Définition de la durée de la saison de croissance du blé

Dans l'objectif de pouvoir prédire le rendement du blé à l'aide de modèles, certaines variables présentées dans la partie des « données brutes utilisées » nécessitent des transformations car elles ne sont pas explicatives du rendement. À titre d'exemple, la valeur du NDVI prise pour une décade à elle seule ne semble, à priori, pas très pertinente pour expliquer le rendement. En revanche, cumuler les valeurs du NDVI sur une certaine période pourrait être un facteur intéressant pour expliquer le rendement du blé. Parmi les données détaillées précédemment, voici celles pour lesquelles un cumul sur une certaine période va être réalisé : le NDVI, le NDVI z-score, les précipitations, la température et le rayonnement solaire. Dès lors, il faut établir une méthodologie afin de définir sur quelle période le cumul va être réalisé. Il est à noter que comme la modélisation est, dans un premier temps, réalisée à l'échelle nationale, chaque pays étudié aura sa propre période puisque le blé n'est pas cultivé partout au même moment.

Le choix du début et de la fin de la saison pour la culture du blé se base sur l'étude suivante : « Global crop calendars of maize and wheat in the framework of the WorldCereal project » qui se traduit par « Calendriers de cultures mondiaux du maïs et du blé dans le cadre du projet WorldCereal » (Franch *et al.*, 2022).

Afin de pouvoir créer un nouveau calendrier pour la culture du blé permettant de connaître le début et la fin de la saison, plusieurs calendriers préexistants ont été intégrés. Une première carte a été réalisée en combinant les données de plusieurs organisations à savoir : le CM (Crop Monitor), l'USDA-FAS (United States Department of Agriculture – Foreign Agricultural Service), la FAO (Food and Agriculture Organization et ASAP (Anomaly hot Spots of Agricultural

Production). Cette carte va être utilisée comme base pour la création de modèles prédictifs et va également permettre de valider ces derniers. Comme l'état du blé ainsi que son développement sont fortement influencés par les conditions météorologiques, un modèle est destiné à rendre compte de la variabilité spatiale du calendrier des cultures du blé et est donc essentiellement constitué de données météorologiques. Il comprend entre autres des données sur les précipitations et sur la température du point de rosée. La performance des modèles a pu être évaluée grâce aux données de référence EO (Earth Observation). Ces données sont fournies par les satellites Sentinel-2 et Landsat 8. Les dates de début et de fin de la saison du blé sont définies grâce à des indices de végétation tels que le NDVI. Ces dates sont ensuite ajustées par rapport aux données de terrain et aux définitions des stades de développement du blé qui varient selon les sources. En effet, un décalage dans les calendriers va être induit si une des sources considère que le début de la saison est le moment de la plantation et qu'une autre considère que c'est le moment où les plantes émergent du sol. Dans le cadre de l'étude, le début de la saison a été défini comme ayant lieu 1 mois avant le début de la phase végétative. La fin de la saison est la date moyenne de la période de récolte (Franch *et al.*, 2022).

Cette étude a permis d'obtenir des cartes avec une résolution spatiale de 0,5° fournissant des résultats plus avancés que les calendriers agricoles déjà existants. L'évaluation du modèle montre de bons résultats avec un R^2 de 0.87 pour le début de la saison et 0.92 pour la fin de la saison. Pour ce qui est de la RMSE, elle vaut 27 jours pour le début de saison et 26 jours pour la fin de saison⁴ (Franch *et al.*, 2022).

La modélisation réalisée dans le cadre de l'étude repose sur un algorithme « Random Forest » qui sera détaillé plus tard. Cet algorithme d'apprentissage automatique basé sur le principe des arbres de décisions est performant, néanmoins il ne prend pas en compte l'autocorrélation spatiale. Pour pallier à cette problématique, les chercheurs ont fait tourner le modèle sur des aspects purement géographiques à savoir les distances euclidiennes.

⁴ : Un rappel des formules pour déterminer le R^2 et la RMSE est disponible à l'annexe 6.

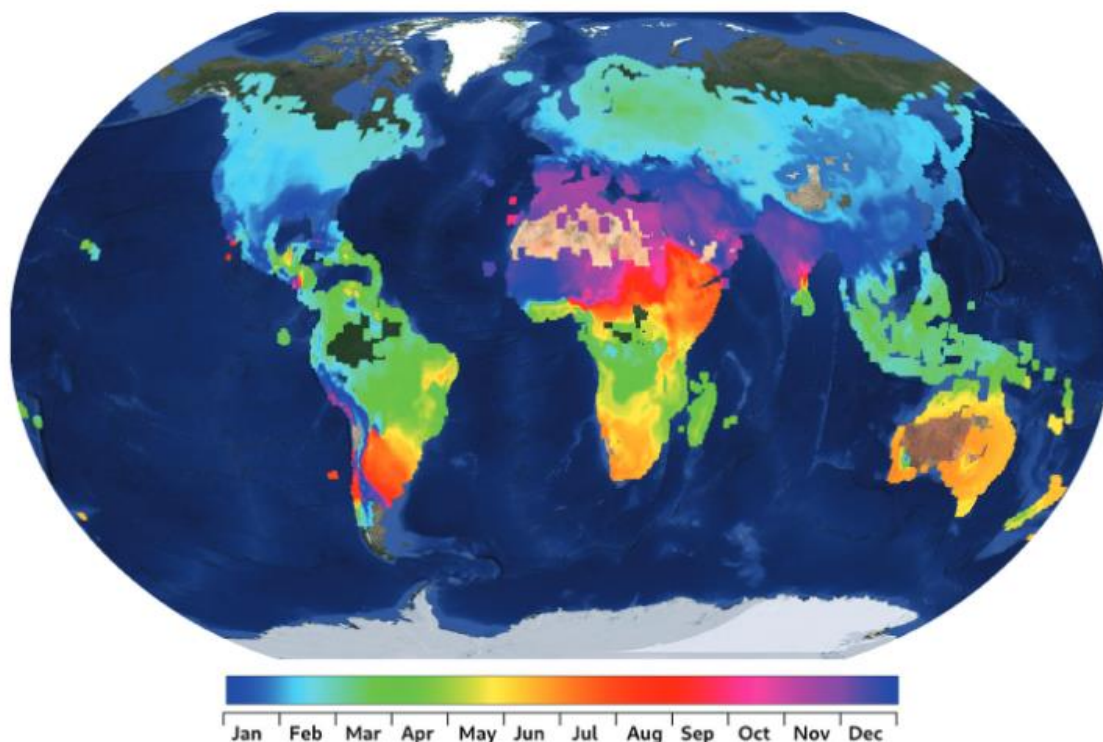


Figure 5. Début de la saison du blé (Franch et al., 2022).

La Figure 5 montre la variation du début de la saison du blé à travers l'ensemble du globe. Globalement, dans l'hémisphère nord, la saison début a lieu vers les mois de janvier, février et mars. Le Moyen-Orient ainsi que le bassin méditerranéen ont un début de saison différent qui s'étend entre le mois de novembre et de janvier. Dans les régions africaines situées entre le Sahara et l'équateur, la saison débute entre le mois de juillet et le mois de décembre. Dans la moitié Nord de l'Amérique du Sud ainsi qu'en Amérique centrale, le début de la saison du blé se produit vers mars-avril. Il en va de même pour les deux autres grandes régions tropicales. La moitié Sud de l'Amérique du Sud débute sa saison du blé entre le mois de janvier et le mois d'août. Enfin, le sud de l'Afrique ainsi que l'Australie ont un début de saison qui varient entre le mois d'avril et le mois de juillet (Franch et al., 2022).

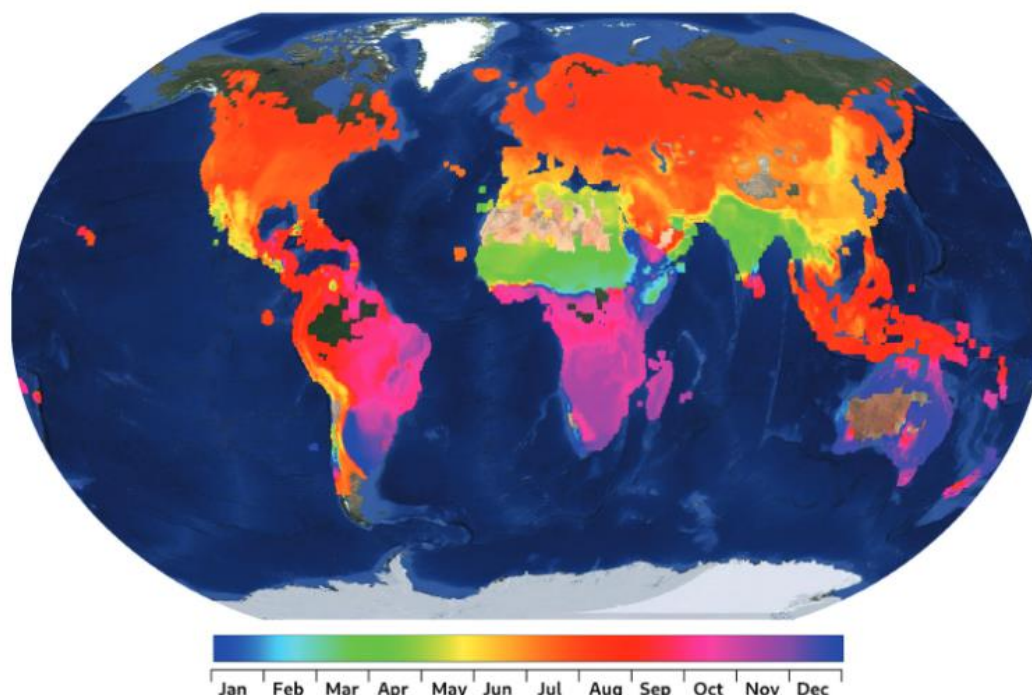


Figure 6. Fin de la saison du blé (Franch *et al.*, 2022).

La Figure 6 illustre la répartition spatiale à travers le monde du moment de la récolte du blé, autrement dit, de la fin de la saison de la céréale. Dans l'hémisphère nord, d'une manière générale, au plus la latitude est importante au plus la date de la fin de saison arrive tard. On va du mois d'avril pour les régions proches de l'équateur jusqu'au mois d'août pour les plus hautes latitudes. Dans l'hémisphère sud, la situation varie selon les continents. Dans le Nord de l'Amérique du Sud ainsi qu'en Amérique centrale, la fin de la saison a lieu entre le mois d'août et le mois de novembre. En Afrique, les dates tournent plutôt entre octobre et janvier. Dans le Sud-Est de l'Asie, les récoltes ont généralement lieu vers les mois de juillet-août. En Océanie, le blé est récolté entre les mois de septembre et de décembre (Franch *et al.*, 2022).

Concrètement, pour définir les débuts et fins de saison du blé des différents pays, des fichiers rasters disponibles en annexe de l'étude ont été utilisés pour plus de précision. Le choix des pixels retenus pour définir la saison de développement du blé a été réalisé de manière à englober la saison la plus longue que possible pour éviter de ne pas prendre en compte des données pertinentes pour les analyses à posteriori (Franch *et al.*, 2022, Annexes).

3.5 Modélisation

Dans cette section, les 3 types de modèles utilisés dans le cadre de ce travail vont être détaillés. Ces modèles sont : la régression linéaire simple, la régression linéaire multiple et Random Forest.

3.5.1 Régression linéaire simple

Dans un premier temps, une simple régression linéaire va être réalisée sur la valeur du rendement mondial. Les prévisions du rendement peuvent être obtenues en prolongeant la

droite de régression. Cette première modélisation servira de référence et sera utilisée pour faire des comparaisons avec les autres modèles. L'équation générique permettant d'obtenir une droite de régression linéaire est disponible à l'annexe 7.

3.5.2 Régression multiple

3.5.2.1 Description théorique

La régression multiple fonctionne de façon similaire à la régression linéaire simple, à la différence que dans ce cas-ci, plusieurs prédicteurs sont utilisés pour réaliser la modélisation. Ce type de modèle permet de déterminer quelles sont les variables/prédicteurs qui jouent un rôle important pour obtenir la valeur de la variable que l'on cherche à prédire (le rendement du blé) (Rousson, 2013).

L'équation générique d'une droite de régression linéaire multiple s'écrit comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

Avec y : la valeur prédite

β_0 : l'ordonnée à l'origine

$\beta_1, \beta_2, \dots, \beta_m$: les coefficients (resp. Pentes) associés à chaque variable prédictive (Rousson, 2013).

Ce type de modèle repose sur des hypothèses pour pouvoir être considéré comme valide. A travers une analyse des résidus obtenus suite à la modélisation, les hypothèses suivantes vont être vérifiées :

- Hypothèse de linéarité : Présence d'une relation linéaire entre les variables indépendantes et la variable dépendante (le rendement).
- Hypothèse d'homoscédasticité : la variance des résidus est constante.
- Hypothèse de normalité des erreurs : les résidus suivent une loi de distribution normale (Rousson, 2013).

3.5.2.2 Description de l'outil de modélisation

Les régressions linéaires multiples vont être réalisées au moyen du logiciel CGMS Statistical Tool (CST). Ce logiciel, mis à disposition sur le site de la Commission européenne, a été spécialement développé dans le but de pouvoir créer des modèles de prévision du rendement des cultures. Le logiciel est conçu de sorte à pouvoir utiliser facilement les variables fournies dans la section ASAP mentionnée précédemment. Il permet de faire des prévisions à l'échelle nationale, mais également sur des sous-divisions des territoires nationaux. Ce logiciel a initialement été créé dans le cadre d'un projet du programme de la commission européenne MARS JRC (Monitoring Agricultural Resources – Joint Research Center). Ce programme fournit régulièrement des informations sur l'état des cultures dans les pays de l'Union Européenne ainsi que dans quelques pays voisins. En plus de cela, les bulletins contiennent également des prévisions des rendements attendus dans ces mêmes pays pour diverses cultures. Il faut noter

qu'avant de pouvoir utiliser le logiciel, les données doivent être entrées dans une autre interface, appelée SQLiteStudio, qui permet d'organiser convenablement toutes les données afin qu'elles puissent ensuite être lues par CST (Commission européenne, 2024 ; JRC, 2024).

3.5.2.3 Description de la méthodologie

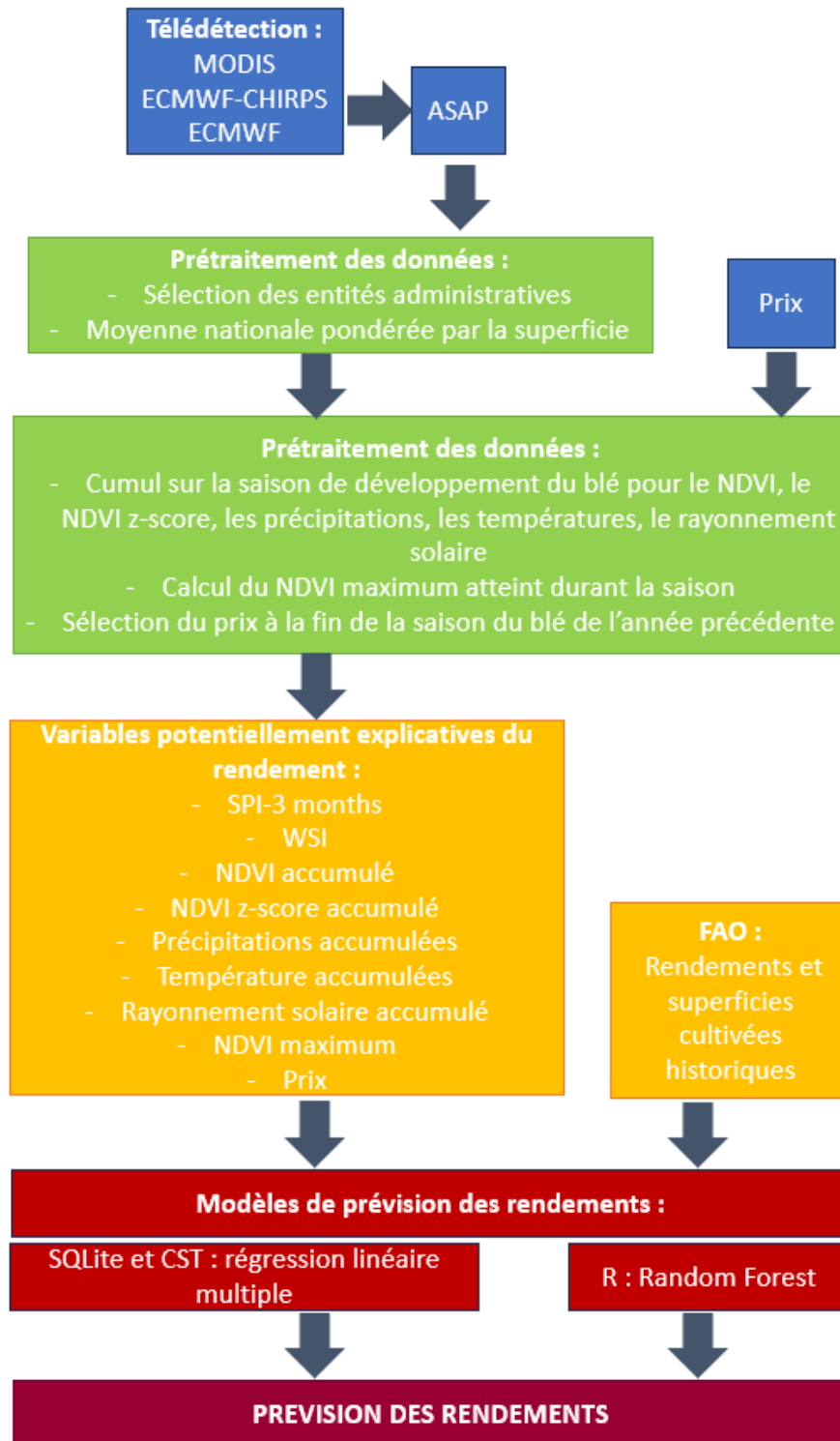


Figure 7. Résumé de la méthodologie.

La Figure 7 retrace l'ensemble des étapes qui ont permis d'obtenir des prévisions du rendement. Dans un premier temps, l'ensemble des données brutes, déjà présentées précédemment, ont été récoltées. Un premier traitement des données a été appliqué aux données ASAP. En effet, ces données sont disponibles pour des sous-entités du territoire national or l'objectif est d'établir un modèle par pays. Cela nécessite donc un prétraitement des données. Pour chacun des pays étudiés, seule une partie du territoire a été retenue et intégrée dans la modélisation du rendement du blé. En effet, comme l'étude est à l'échelle nationale, les territoires sont vastes et il n'y pas nécessairement des zones agricoles partout. Dès lors, seules les régions avec une quantité non négligeable de terres cultivées ont été intégrées au modèle. Les entités administratives avec une superficie cultivée nulle ou de moins de 1000 km² pour un territoire de plusieurs milliers de km² n'ont donc pas été sélectionnées. Une fois la sélection des entités administratives faite, afin d'obtenir des données à l'échelle nationale, des moyennes pondérées selon la superficie cultivée de chaque entité administrative ont été réalisées. D'autres traitements des données ont ensuite été réalisés sur certaines données issues d'ASAP, dont le NDVI, par exemple, ainsi que sur le prix. Certaines variables prises en tant que telles ne permettent pas d'expliquer le rendement. En revanche, le fait de les cumuler sur une certaine période est intéressant car cela permet de rendre compte des conditions dans lesquelles les cultures se sont développées. Après avoir défini ce qui sera considéré comme la saison du blé dans le pays, un cumul des variables suivantes a été réalisé sur cette période : NDVI, NDVI z-score, précipitations, températures et rayonnement solaire. La valeur du NDVI maximum atteinte durant la saison du blé chaque année a été enregistrée. Cette variable est un bon indicateur de la vigueur que les plantes sont parvenues à atteindre. Le prix à la toute fin de la saison du blé a également été enregistré avec chaque fois une année de décalage comme déjà expliqué. L'ensemble des données utilisées pour la modélisation comprend les rendements historiques du pays ainsi que les superficies où du blé était cultivé chaque année, le SPI-3 monts, le WSI, le NDVI accumulé, le NDVI z-score accumulé, les précipitations accumulées, les températures accumulées, le rayonnement solaire accumulé, le NDVI maximal et le prix. Le jeu de données débute en 2002 (première année pour laquelle toutes les données ASAP sont disponibles) et se termine en 2021 (dernière année pour laquelle les données de rendement et de superficie cultivées par du blé sont disponibles⁵). Une fois les données entrées via SQLiteStudio et CST, le rendement de la dernière année (2021) va tout d'abord être simulé en disposant de toutes les données, mis à part le rendement, jusqu'à la fin de la saison de la dernière année. Les 3 meilleurs modèles de régression multiples, classés selon la RMSE du rendement, vont être enregistrés. Une moyenne des 3 valeurs de rendement simulées par les modèles va être faite afin de pouvoir être comparée à la prévision. Les prévisions vont être réalisées 2 décades avant la fin de la saison. L'idée est donc de connaître le rendement 20 jours avant que le blé ne soit récolté. Les modèles nécessitent des plages de données complètes jusqu'à la fin de la saison, c'est pourquoi les données des 2 dernières décades considérées comme n'étant pas encore connues auront exactement les mêmes valeurs de données de la dernière décade connue, soit l'avant-avant dernière décade avant la fin de la saison. De nouveau, une moyenne des prévisions fournies

⁵ Mise à jour : les données de la FAO sont à présent disponibles pour l'année 2022.

par les 3 meilleurs modèles de régression linéaire multiple va être réalisée afin de pouvoir être comparée à la moyenne des simulations.

3.5.3 Random Forest

3.5.3.1 Description théorique

La méthode Random Forests ou Random forest (forêts aléatoires en français) a été inventée par Leo Breiman et publiée en 2001. Cette méthode, basée sur le « machine learning » (apprentissage automatique en français), est un modèle d'ensemble basé sur les arbres de décisions. Les forêts aléatoires permettent de faire de la régression et de la classification. Breiman a nommé cette approche « CART » (Classification & Regression Trees). La Figure 8 illustre le principe de ce type de modélisation : combiner les prévisions des différents arbres en une seule prévision (Genuer, 2010 ; M. Lang, comm. pers., 2024).

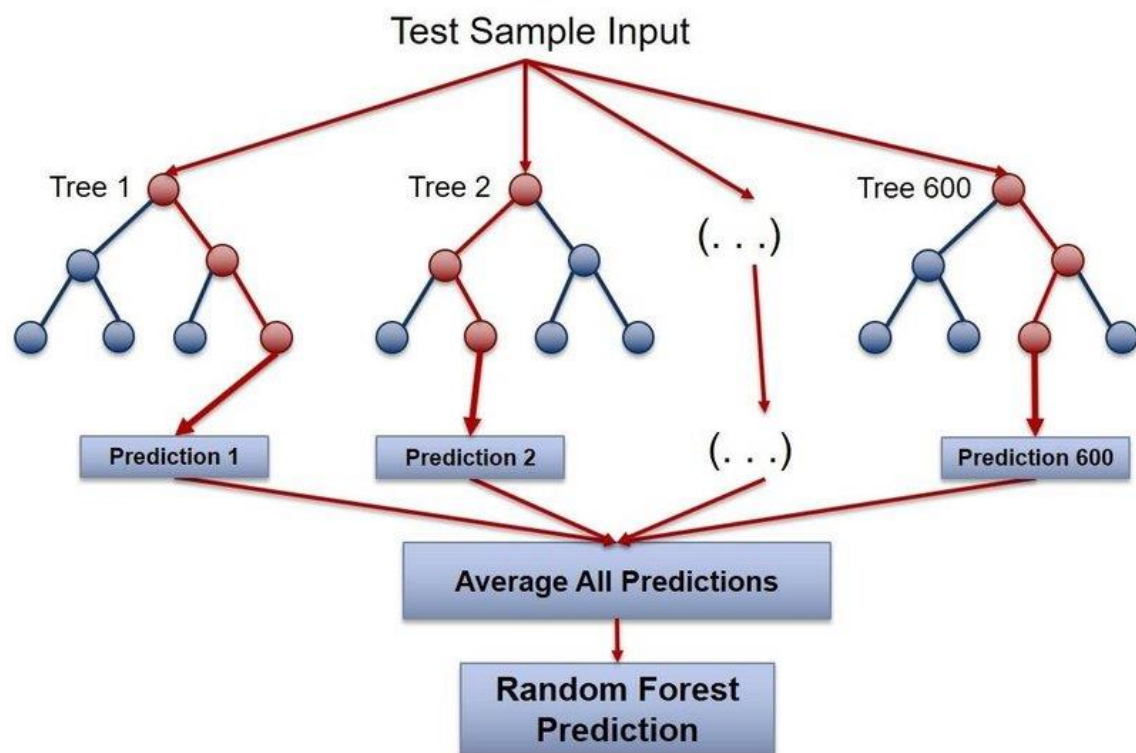


Figure 8. Visualisation de la méthode de création de forêts aléatoires (Blakely et al., 2018).

L'utilisation des forêts aléatoires présente plusieurs avantages. Tout d'abord, regrouper les prévisions obtenues au bout de chaque arbre pour obtenir une prévision globale pour toute la forêt assure une certaine robustesse du modèle ainsi que de la précision. En effet, le fait d'associer les résultats de différents arbres de décision permet d'éviter un surajustement. Ensuite, comme déjà mentionné, les forêts aléatoires ont la particularité de pouvoir être appliquées à des jeux de données de type catégoriel et de type continu. Les « randoms forests » renseignent sur l'importance plus ou moins grande des différentes variables intégrées au modèle de prévision. Contrairement aux régressions multiples, il n'y a pas de distribution particulière des données attendue et les forêts aléatoires ont la capacité de gérer des interactions complexes qui ne sont pas nécessairement linéaires. Enfin, la sélection aléatoire

de l'échantillon de données utilisé ainsi que la sélection aléatoire des variables prises en compte permettent d'éviter d'avoir des corrélations non désirées et donc de réduire la variance des prévisions du modèle (Genuer, 2010 ; M. Lang, comm. pers., 2024).

Comme mentionné ci-dessus, les forêts aléatoires rendent compte de l'importance des variables. Cette importance est déterminée grâce à une méthode de permutation. Si une variable prédictive joue effectivement un rôle important, alors, la permutation de ses valeurs dans l'ensemble de données aura un impact significatif sur les résultats du modèle. Il suffit donc de comparer les performances du modèle avant et après avoir permuté les valeurs de la variable prédictive pour évaluer son importance (Genuer, 2010 ; M. Lang, comm. pers., 2024).

Après avoir cité les avantages, il faut toutefois noter que cette méthode présente également des inconvénients. Les modèles ayant recourt aux forêts aléatoires sont complexes et sont souvent qualifiés de boîtes noires. En effet, bien que les résultats finaux fournis par le modèle soient robustes et précis, il est difficile de connaître et de visualiser les étapes intermédiaires qui ont permis d'obtenir les résultats finaux. Ainsi, il est, par exemple, difficile de connaître les résultats d'un seul arbre pris individuellement au sein de toute une forêt. De plus, quand bien même il serait possible de visualiser un arbre isolé, ce ne serait pas forcément pertinent de l'analyser seul. De fait, c'est uniquement la combinaison de tous les arbres qui permet d'obtenir des résultats pertinents et significatifs. Un autre point négatif est que les prévisions peuvent uniquement être faites au sein de la plage de données fournies par l'utilisateur (Genuer, 2010 ; M. Lang, comm. pers., 2024).

La première étape pour construire les arbres s'appelle le bagging ou bien l'échantillonnage bootstrap. Parmi l'ensemble des données d'entraînement, un échantillon bootstrap va être prélevé pour chaque arbre. Les mêmes données peuvent être sélectionnées plusieurs fois pour différents arbres c'est pourquoi on dit que les échantillons sont « avec remplacement ». Les données reprises dans un échantillon vont être divisées en 2 parties. Il y a, tout d'abord, l'ensemble de construction : $2/3$ des données sont utilisées pour construire l'arbre à proprement parler. Ensuite, le tiers restant, appelé « Out-Of-Bag » (OOB) est utilisé pour valider l'arbre. La deuxième étape consiste à sélectionner aléatoirement les variables. Cette sélection a lieu à chacun des nœuds de l'arbre. Parmi les variables sélectionnées, l'algorithme va définir la meilleure division des variables à faire, c'est-à-dire celle qui offre la meilleure séparation entre les classes dans le cas de la classification et celle qui minimise les erreurs dans le cas de la régression. Ce processus permet d'éviter d'avoir des corrélations entre les arbres et améliore donc la qualité du modèle. La division en nœuds successifs sera arrêtée lorsque la quantité de données restantes passera sous un certain seuil. A ce moment-là, les nœuds en terminaison de l'arbre seront appelés les feuilles. Enfin, la dernière étape qui permet d'obtenir la prévision finale implique de regrouper l'ensemble des prévisions des différents arbres, autrement dit les informations situées dans les feuilles. Dans le cas d'une classification, la prévision finale est obtenue par vote majoritaire tandis que pour une régression elle est obtenue en moyennant toutes les prévisions (Genuer, 2010 ; M. Lang, comm. pers., 2024).

3.5.3.2 Description de l'outil de modélisation

La modélisation des forêts aléatoires a été réalisée dans le logiciel R qui est à la fois un langage de programmation et un environnement permettant essentiellement de faire de l'analyse statistique ainsi que des représentations graphiques. L'un des principaux avantages de R est qu'il comprend toute une série de packages regroupant eux-mêmes un vaste ensemble de fonctions utiles pour des applications diverses et variées. Parmi les librairies, « caret » et « partykit » comprennent les fonctions nécessaires à la création d'un modèle qui utilise les forêts aléatoires (R Foundation, *n.d.*).

La fonction « train » du package « caret » permet de créer le modèle et de l'entraîner sur la base de données. Elle permet d'appliquer différentes techniques de modélisation aux données dont les forêts aléatoires. La fonction « cforest » du package « partykit » est utilisée afin de créer des arbres d'inférence conditionnelle (Intercondition inference trees). Cette fonction possède des améliorations par rapport aux forêts aléatoires traditionnelles présentées par Brieman en 2001 en y apportant des corrections. Tout d'abord, cette fonction permet de corriger le biais qui était présent lors de l'utilisation simultanée de variables continues et catégorielles. Dans les forêts aléatoires classiques, les variables continues étaient préférentiellement choisies au détriment des indicateurs binaires car il est plus facile de trouver le point de séparation le plus optimal dans un ensemble de données continues. Ce biais potentiel a été corrigé pour éviter une sous-représentation des variables non-continues. Ensuite, la fonction « cforest » présente une autre correction intéressante. Cette correction permet d'éviter qu'une variable qui semble corrélée directement au rendement, alors qu'elle est en réalité corrélée à une autre variable explicative, ne possède pas une importance artificiellement trop élevée (M. Lang, comm. pers., 2024).

3.5.3.3 Description de la méthodologie

Comme schématisé à la Figure 7 toutes les étapes préalables à l'intégration des variables dans la modélisation sont identiques pour la régression linéaire multiple et pour les forêts aléatoires. Tout comme pour les régressions multiples, le rendement a tout d'abord été simulé avec les données des variables disponibles jusqu'à la fin de la saison. Puis, le rendement a été prédit avec les données disponibles jusqu'à 2 décades avant la fin de la saison.

4. Résultats

Ce chapitre est dédié aux résultats. Dans un premier point, les résultats du modèle de base, un modèle de régression linéaire simple sur les rendements mondiaux va être présenté. Ensuite les résultats des modèles de régression multiple et des modèles avec les forêts aléatoires vont être analysés pour les 5 pays retenus dans le cadre de cette étude (Chine, Inde, Russie, États-Unis et France). Enfin, un comparatif des résultats obtenus va être réalisé.

4.1 Régression linéaire simple sur le rendement mondial

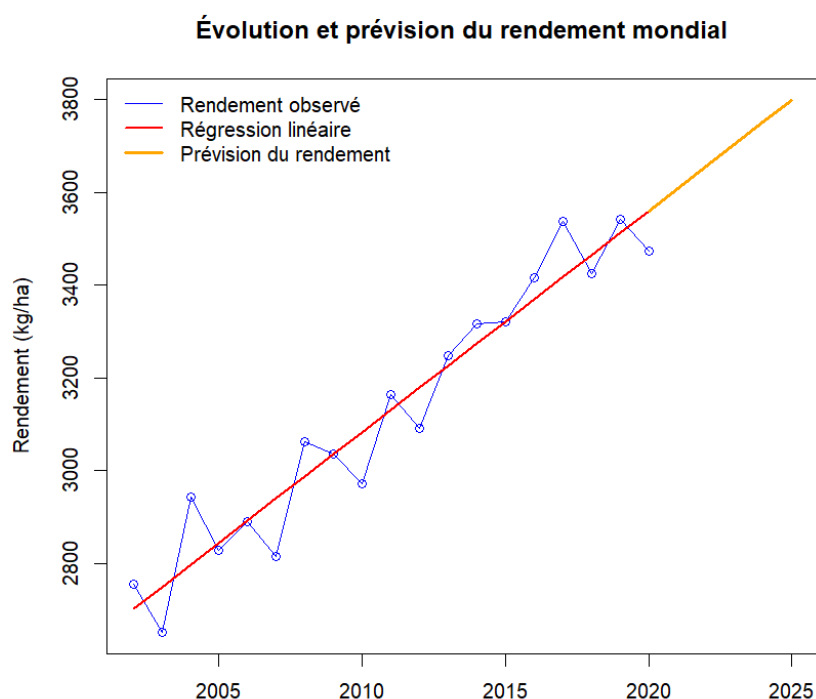


Figure 9. Évolution et prévision du rendement mondial du blé (source de données : FAO, 2023).

Un modèle de régression linéaire simple a été établi sur les données du rendement mondial de 2002 à 2020 (cf. Figure 9). L'équation de régression de ce modèle est la suivante :

$$y = -92752,44 + 47,68 \cdot x$$

Le coefficient de détermination (R^2) de ce modèle vaut 0,9251 et la racine carrée de l'erreur quadratique moyenne (RMSE) est de 74,31 kg/ha. En prolongeant la droite de régression linéaire simple, une prévision du rendement selon le modèle est obtenue. Ainsi, d'après la droite de régression, le rendement mondial prévu en 2021 est de 3607,77 kg/ha. Pour l'année 2021, la prévision du modèle surestime de 115,87 kg/ha la valeur de rendement recensée par la FAO puisque cette valeur vaut 3491,9 kg/ha (FAO, 2023).

4.2 Chine

Pour la Chine, comme pour chaque pays qui va suivre, une liste des entités administratives qui ont été retenues pour la modélisation est disponible à l'annexe 8.

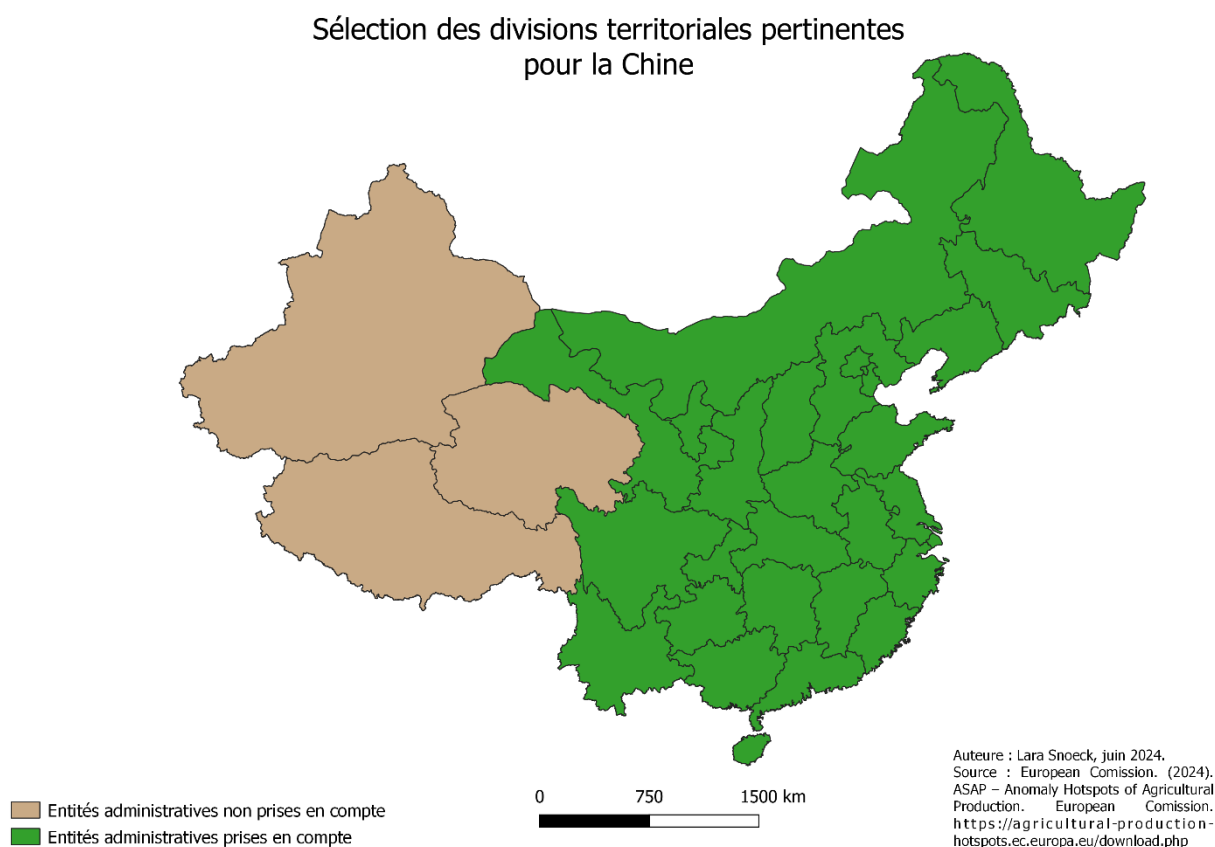


Figure 10. Sélection des divisions territoriales pertinentes pour la Chine.

La plupart des provinces chinoises ont été reprises dans le modèle car elles comportent des terres cultivées (cf. Figure 10). Toute la partie ouest du pays n'a pas été prise en compte car on n'y retrouve pas ou peu de cultures et elle n'apportera donc rien pour l'étude (European Commission, 2024 ; FAO, 2022).

Sur base de l'étude de Franch *et al.*, la saison de croissance du blé en Chine a été définie comme débutant à la 2^{ème} décade de mois de janvier et se terminant à la 2^{ème} décade du mois de juin. Par conséquent, le prix intégré au modèle était le prix du blé sur le marché mondial à la fin de la deuxième décade de juin (Franch *et al.*, 2022, Annexes).

4.2.1 Régression multiple

La Figure 11 montre l'évolution du rendement en Chine durant les années d'entraînement du modèle à savoir de 2002 à 2020. Les valeurs de rendement, représentées par des croix bleues sur le graphique, sont exprimées en kg/ha et sont fournies par la FAO. La ligne verte représente l'ajustement de la régression linéaire aux données et la courbe bleue représente la régression

quadratique⁶. Au vu des valeurs de p-value obtenues pour la régression linéaire et pour la régression quadratique, 0,0000 et 0,0210 respectivement, les résultats du modèle analysé seront ceux obtenus en suivant une tendance linéaire. Ce sera d'ailleurs le cas aussi pour les prochains pays qui seront analysés.

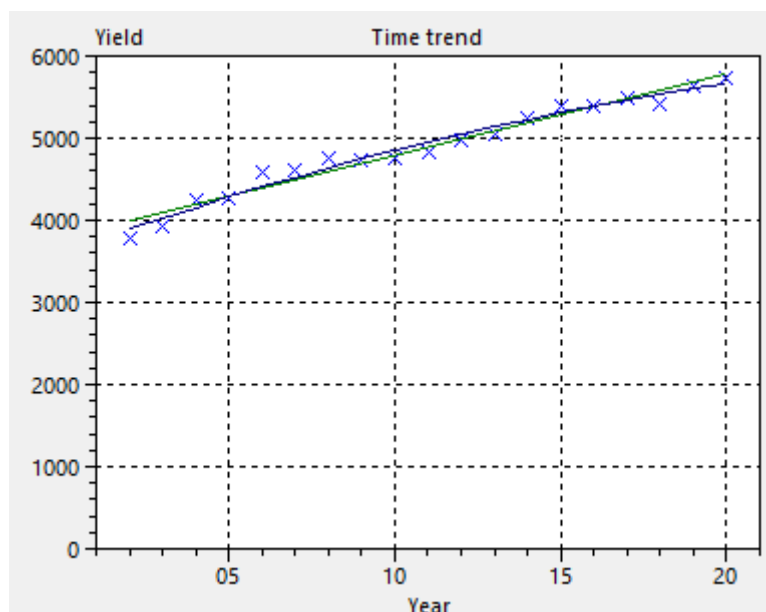


Figure 11. Évolution du rendement en Chine (en kg/ha) entre 2002 et 2021 (source de données : FAO, 2023).

Dans un premier temps, vont être analysés les résultats obtenus lorsque les variables, mis à part le rendement, sont disponibles jusqu'à la fin de la plage de données.

Le meilleur modèle obtenu, c'est-à-dire celui qui possède la plus faible valeur de RMSE pour la prévision du rendement comprend 3 variables prédictives à savoir : les précipitations cumulées durant la période préalablement définie comme étant la saison du blé en Chine, le rayonnement solaire cumulé durant la même période et le prix du blé sur le marché mondial de l'année précédente à la fin de la saison. Pour plus de facilité, par la suite, ce modèle sera désigné comme le modèle 1. Le tableau suivant récapitule des statistiques intéressantes propres au modèle1 :

Tableau 2. Statistiques du modèle 1 de la Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	106.875
R ² ajusté	0.9744
Valeur de la simulation (rendement de 2021)	5647.516 kg/ha

L'équation du modèle 1 peut s'écrire comme suit :

$$\text{Rdt} = -4002,63^7 + 91,453 \cdot \text{TL} + 1.339 \text{ CumRain} + 1.68 \cdot 10^{-3} \text{ CumSun} + 0.784 \cdot \text{Price}$$

⁶ Des graphiques identiques seront présentés pour d'autres pays. Les explications relatives à la légende du graphique de l'évolution du rendement pour la Chine sont également valables pour les autres pays.

⁷ Les constantes des équations de régression rendues par le logiciel étant aberrantes pour une raison indéterminée (probablement à cause d'un dysfonctionnement dans la programmation du logiciel), elles ont chaque fois été recalculées manuellement à partir des données de la dernière décade de la saison 2021. Lors ...

Avec Rdt : le rendement

TL : la tendance linéaire temporelle (Année – 1965)

CumRain : le cumul des précipitations

CumSun : le cumul des radiations solaires

Price : le prix

Il faut tout de même noter que les coefficients pour les précipitations et pour le prix sont indiqués comme étant non significatifs par le logiciel (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

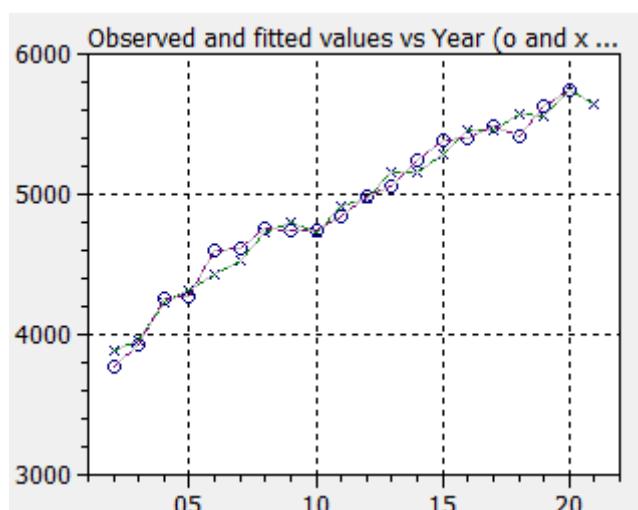


Figure 12. Évolution des rendements (en kg/ha) observés et simulés en Chine pour le modèle 1 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

La Figure 12 représente l'évolution du rendement au cours des années (cercles) ainsi que l'évolution du rendement simulé par le modèle. Pour pratiquement toutes les années de la période d'étude, la valeur réelle du rendement est très proche de la valeur simulée par le modèle. Afin de s'assurer de la qualité du modèle, les résidus vont également être analysés. Ils vérifient bien les conditions d'un modèle de régression linéaire multiple à savoir : les résidus suivent une distribution normale et sont indépendants les uns des autres (ce qui se traduit par un nuage de points aléatoires). Des graphiques appuyant ces commentaires sont disponibles à l'annexe 9.

Le second meilleur modèle obtenu (modèle 2) est un modèle à 4 variables. Il reprend les 3 mêmes prédicteurs que dans le modèle 1 (cumul des précipitations, cumul du rayonnement solaire et prix). La nouvelle variable incluse est le WSI. Le tableau suivant récapitule les statistiques du modèle 2 :

... de vérifications effectuées avec les données des années précédentes, des variations mineures, de l'ordre de quelques décimales, ont été observées dans la valeur de la constante. Ces variations sont probablement dues aux différences d'arrondi, le logiciel ne spécifiant pas le nombre de décimales exactes utilisées dans ses calculs.

Tableau 3. Statistiques du modèle 2 de la Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	109.36
R ² ajusté	0.97588
Valeur de la simulation (rendement de 2021)	5623.739 kg/ha

L'équation du modèle 2 peut s'écrire comme suit :

$$\text{Rdt} = -2236.689 + 93,188 \cdot \text{TL} + 2.695 \text{ CumRain} + 1.3 \cdot 10^{-3} \text{ CumSun} + 0.851 \cdot \text{Price} - 18.314 \cdot \text{WSI}$$

Les coefficients du prix et du WSI sont toutefois renseignés comme étant non-significatifs (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

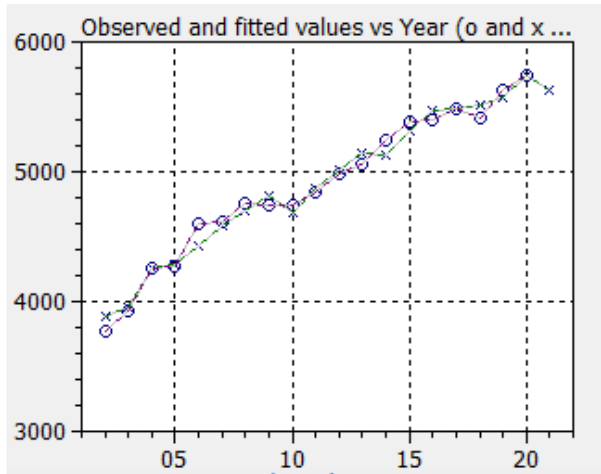


Figure 13. Évolution des rendements (en kg/ha) observés et simulés en Chine pour le modèle 2 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

L'ajustement de la courbe du rendement prédit (cf. Figure 13) par rapport à celle du rendement réel est très similaire au modèle 1. La différence entre les 2 rendements prédits pour 2021 est d'ailleurs inférieure à 25 kg/ha. Les résidus semblent également correspondre aux critères des régressions multiples. Les graphiques associés sont disponibles à l'annexe 10.

Enfin, le troisième meilleur modèle (modèle 3) comprend uniquement 2 variables à savoir l'accumulation du rayonnement solaire et le prix. Les statistiques du modèle 3 sont indiquées dans le tableau suivant :

Tableau 4. Statistiques du modèle 3 de la Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	110
R ² ajusté	0.97055
Valeur de la simulation (rendement de 2021)	5738.63 kg/ha

L'équation du modèle 3 peut s'écrire comme suit :

$$\text{Rdt} = -2605.909 + 94.679 \cdot \text{TL} + 1.24 \cdot 10^{-3} \cdot \text{CumSun} + 0.701 \cdot \text{Price}$$

Le coefficient du prix est renseigné comme non-significatif (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

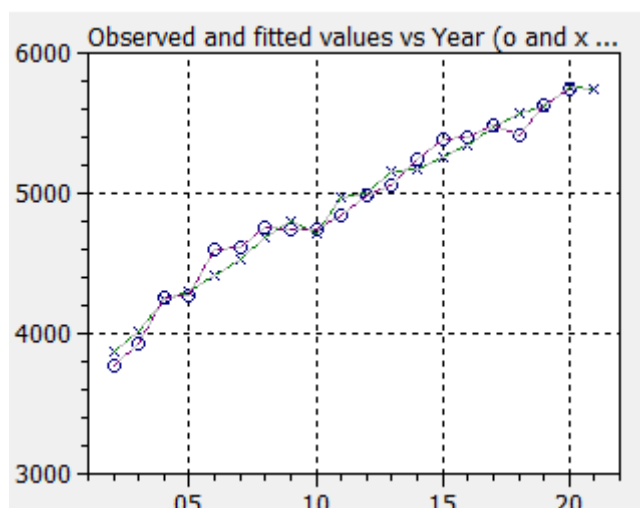


Figure 14. Évolution des rendements (en kg/ha) observés et simulés en Chine pour le modèle 3 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Comme le montre la Figure 14, la courbe du rendement simulé s'adapte un peu moins bien à celle du rendement réel comparativement aux deux premiers modèles. Autre différence, le rendement prédit pour 2021 est pratiquement équivalent à celui de 2020 alors que les autres modèles prévoient plutôt une baisse du rendement pour 2021. En ce qui concerne les résidus, ils sont bien indépendants les uns des autres et sont distribués selon une loi normale (cf. Annexe 11).

Sur base de ces 3 modèles et donc de ces 3 prévisions de rendement du blé pour 2021, une moyenne va être calculée afin de définir la valeur du rendement prédite lorsque l'on est à la fin de la saison (et donc que l'on connaît les valeurs de toutes les variables jusqu'à la fin).

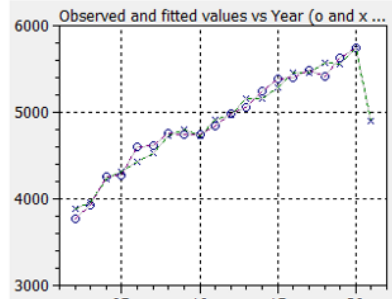
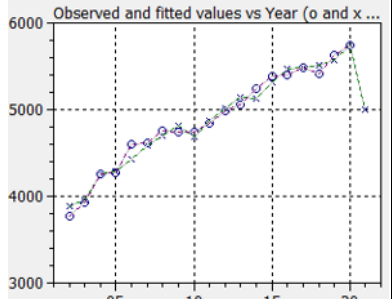
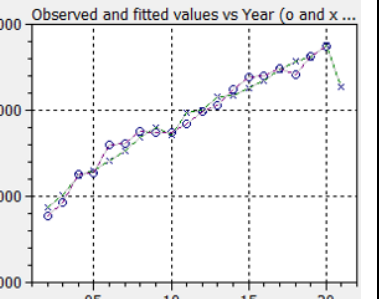
Valeur moyenne du rendement simulé pour 2021 en fin de saison : 5623,739 kg/ha.

Le vrai rendement du blé en 2021 était de 5810 kg/ha. La moyenne des 3 modèles sous-estime le rendement de 186,261 kg/ha. L'estimation est donc relativement bonne.

La partie qui va suivre va cette fois porter sur l'analyse des résultats obtenus lorsque l'on se situe 2 décades avant la fin de la saison. L'objectif ici est de réaliser de vraies prévisions en ne connaissant pas l'évolution des variables prédictives dans les 20 jours qui vont suivre. Comme expliqué précédemment, pour pouvoir obtenir des résultats, il est nécessaire de ne pas laisser de cases vides dans toute la plage de données. Par conséquent, pour les deux dernières décades de la saison de 2021, les variables prendront les valeurs de la dernière décade pour laquelle les données sont connues. Une fois de plus, les 3 meilleurs modèles vont être analysés et leurs prévisions seront ensuite moyennées. Ces modèles seront appelés modèle 1 bis, modèle 2 bis et modèle 3 bis.

Les modèles bis sont identiques aux premiers modèles, la seule variation est la valeur du rendement prédit. Le tableau ci-dessous reprend les nouvelles valeurs de rendement prédites ainsi que leur visualisation sur le graphique d'évolution du rendement :

Tableau 5 Prévisions et graphiques du rendement (en kg/ha) des versions bis des modèles en Chine (source de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Modèle 1 bis : prévision	Modèle 2 bis : prévision	Modèle 3 bis : prévision
4904.313 kg/ha	5002.611 kg/ha	5272.54 kg/ha
		

Les valeurs de rendements prédites pour l'année 2021 avec les modèles bis sont bien en dessous de celles obtenues avec les 3 premiers modèles. Les graphiques affichés dans le Tableau 5 montrent d'ailleurs que la prévision s'éloigne nettement de la tendance formée par les rendements des années précédentes. Une moyenne des rendements prédits par les 3 meilleurs modèles va être calculée comme précédemment.

Valeur moyenne du rendement prédit pour 2021, 2 décades avant la fin de saison : 5125,821 kg/ha.

La différence entre le rendement réel et le rendement moyen prédit est, cette fois, plus conséquente puisque la moyenne des 3 modèles bis sous-estime le rendement de 684,179 kg/ha par rapport à la réalité (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

4.2.2 Random forest

La Figure 15 montre l'importance relative des différentes variables pour les années d'entraînement du modèle de la Chine qui vont de 2002 à 2021. Si le prédicteur ayant le plus d'impact dans le modèle a une importance cruciale comparé aux autres, il aura une importance de 100 %. C'est le cas en Chine, avec NDVI maximal atteint durant la saison du blé qui domine toutes les autres variables. Deux autres prédicteurs ont un rôle important : le NDVIz-score cumulé et le prix qui, par rapport à la variable la plus importante, ont une importance de 46,92 % et 42,41 % respectivement. Les autres prédicteurs ont des importances inférieures à 5 % et ont donc peu ou pas d'influence pour établir le modèle. Pour ce qui est de la performance du modèle sur les données d'entraînement, la RMSE de l'ensemble de calibration vaut 133,184 kg/ha et la RMSE de l'ensemble de validation vaut 187,825 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

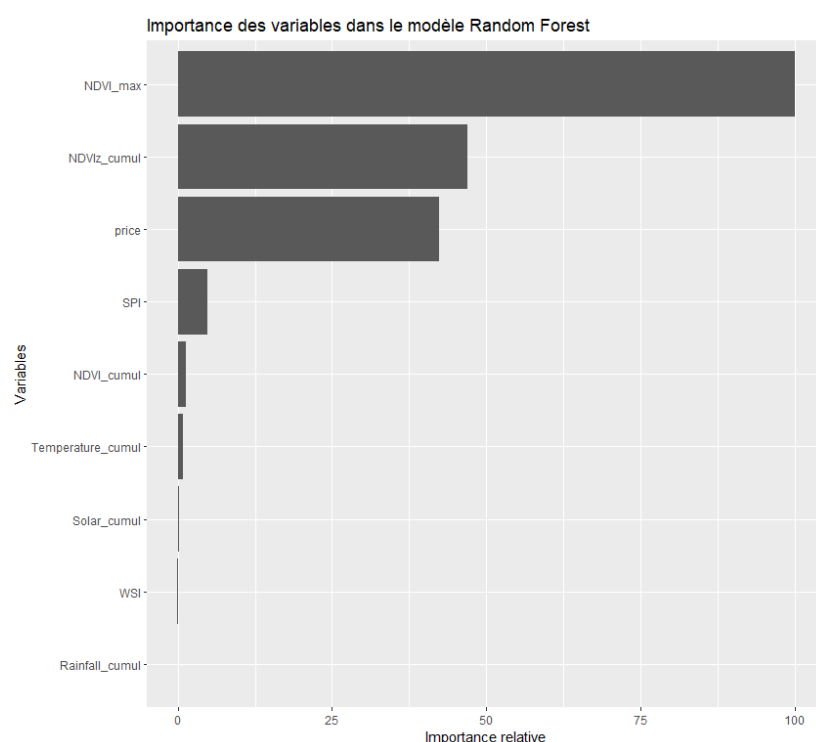


Figure 15. Importance des variables dans le modèle de forêts aléatoires en Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Lorsque le modèle a tourné sur l'ensemble des données jusqu'à la fin de l'année 2021, l'ordre d'importance des variables est resté inchangé. Le NDVI maximal a donc toujours une importance relative de 100 %. Le NDVI-z score cumulé possède à présent une importance de 51,26 % et le prix de 41,12 %. Les informations relatives à ce modèle sont reprises dans le tableau suivant :

Tableau 6. Statistiques du modèle avec les forêts aléatoires de la Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	187.825
R ² ajusté	0.93251
Valeur de la simulation (rendement de 2021)	5517kg/ha

Les résidus d'un modèle créé avec l'algorithme Random Forest ne doivent pas respecter certaines conditions comme pour les régressions multiples linéaires. Cependant, des graphiques similaires à ceux des résidus des modèles de régression multiple sont disponibles à l'annexe 12 à titre informatif.

La véritable valeur du rendement du blé, en Chine, en 2021, selon la FAO, était de 5810 kg/ha. Le modèle fait donc une sous-estimation de 293 kg/ha. Selon la même méthodologie que celle employée pour la partie sur la régression multiple, les résultats du modèle vont une nouvelle fois être analysés après avoir recopié les valeurs des données de l'avant-dernière décade dans les deux dernières. Cette façon de faire permet ainsi de faire une prévision du rendement

en Chine deux décades avant que le blé ne soit récolté. Le rendement prédit en 2021 à 2 décades de la fin de la saison vaut : 5519,83 kg/ha. Cette valeur est très proche du rendement simulé avec les données jusqu'à la fin de la saison. La différence entre les 2 est d'un peu moins de 3 kg/ha. La sous-estimation par rapport à la valeur réelle du rendement vaut 290,17 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

4.3 Inde

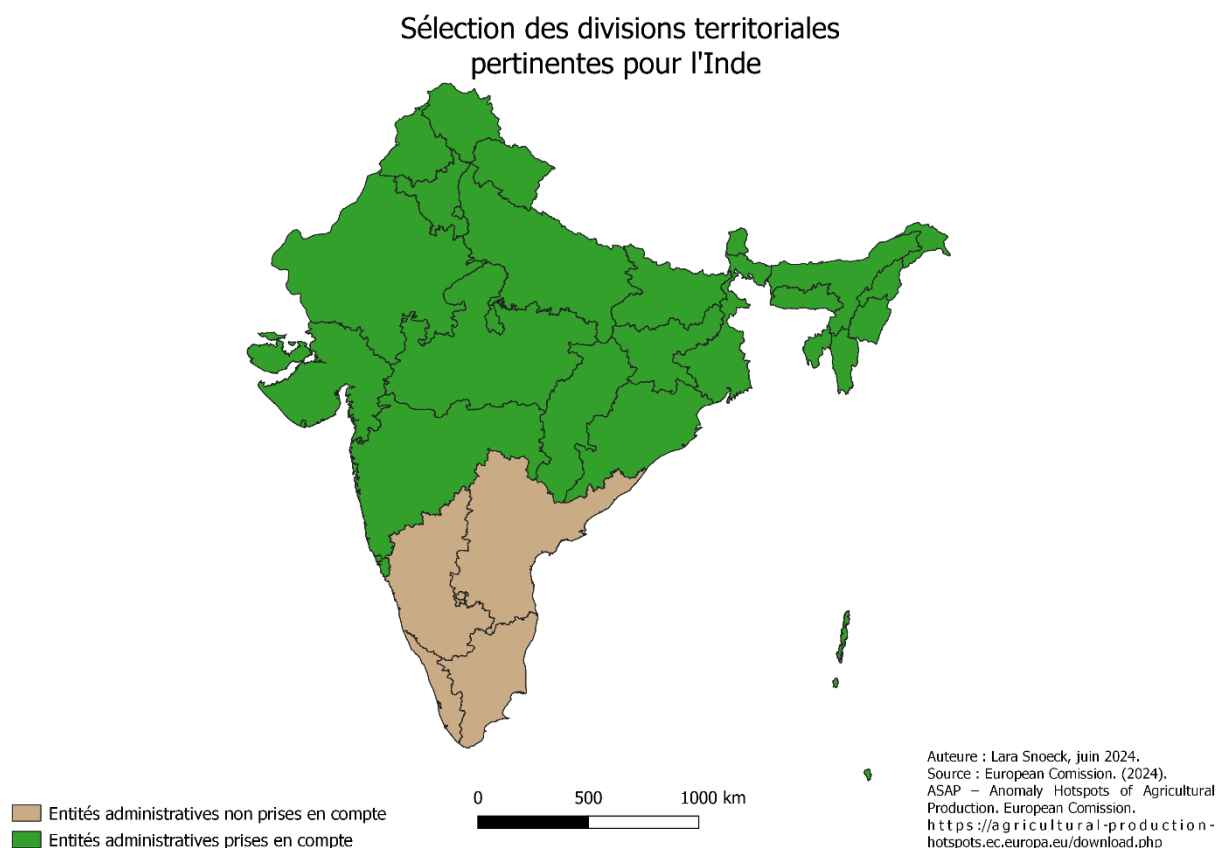


Figure 16. Sélection des divisions territoriales pertinentes pour l'Inde.

L'Inde est le deuxième plus grand pays producteur de blé. Les 4 États situés le plus au sud du pays n'ont pas été retenus pour la modélisation car le blé n'est pas cultivé (cf. Figure 16). La liste des États pris en compte est disponible à l'annexe 8 (European Commission, 2024 ; FAO, 2022).

Pour ce qui est de la définition de la saison du blé en Inde, elle débute à la première décade du mois de novembre et se termine à la deuxième décade du mois d'avril. Le prix renseigné pour une année est donc celui de l'année précédente au moment de la récolte soit à la fin de la 2^{ème} décade du mois d'avril (Franch *et al.*, 2022, Annexes).

4.3.1 Régression multiple

La Figure 17 montre l'évolution du rendement du blé en Inde sur les années d'entraînement du modèle, autrement dit, de 2002 à 2021. La p-value pour ces données est inférieure à 0,0000 pour une régression multiple linéaire.

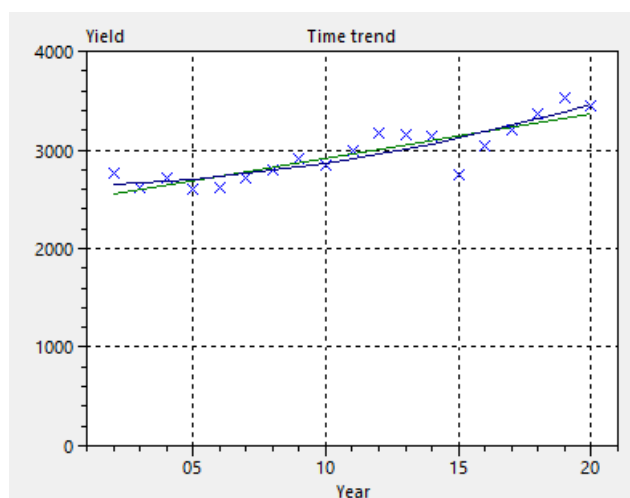


Figure 17. Évolution du rendement en Inde (en kg/ha) entre 2002 et 2021 (source de données : FAO, 2023).

Pour commencer, une description des résultats des 3 meilleurs modèles (modèle 1, modèle 2 et modèle 3) au moment de la fin de saison, avec pour seule inconnue le rendement de l'année pour laquelle on souhaite faire la prévision va être présentée. Par la suite, une analyse des 3 modèles ayant la plus faible RMSE pour la prévision du rendement, 2 décades avant la fin de la saison, sera détaillée (modèle 1 bis, modèle 2 bis et modèle 3 bis).

Parmi les modèles générés, le meilleur modèle (modèle 1) est celui qui n'intègre aucune variable prédictive. En d'autres mots, le modèle qui présente la plus petite valeur de RMSE pour la prévision du rendement est une régression linéaire simple sur les valeurs du rendement. Le tableau suivant regroupe les statistiques relatives à ce modèle :

Tableau 7. Statistiques du modèle 1 de l'Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	145.553
R ² ajusté	0.7781
Valeur de la simulation (rendement de 2021)	3444.595 kg/ha

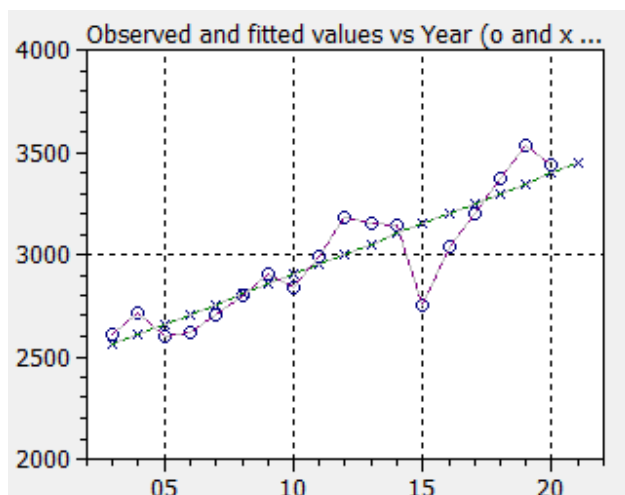


Figure 18. Évolution des rendements (en kg/ha) observés et simulés en Inde pour le modèle 1 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

La Figure 18 montre que l'évolution du rendement en Inde est plus variable comparativement à la Chine. Une baisse importante du rendement a eu lieu en 2015 et a été suivie par une nette ré-augmentation durant les 4 années qui ont suivi. Les écarts entre les valeurs réelles du rendement et les valeurs prédites sont plus importants que pour la Chine ce qui se traduit par une valeur de RMSE un peu plus élevée et un moins bon R^2 ajusté. Conformément aux hypothèses des régressions linéaires, les résidus semblent être indépendants et leur distribution s'approche bien d'une distribution selon une loi normale (cf. Annexe 13).

Le second modèle comporte 2 prédicteurs : le SPI et le cumul des température sur la saison du blé en Inde. Le tableau ci-dessous reprend les statistiques du modèle 2 :

Tableau 8. Statistiques du modèle 2 de l'Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	150
R^2 ajusté	0.81783
Valeur de la simulation (rendement de 2021)	3558.154 kg/ha

L'équation du modèle 2 s'écrit comme suit :

$$\text{Rdt} = 3739.418 + 50.801 \cdot \text{TL} - 155.322 \cdot \text{SPI} - 8.463 \cdot \text{TempCum}$$

Avec Rdt : le rendement

TL : la tendance linéaire temporelle (Année – 1965)

SPI : le SPI – 3 months

TempCum : le cumul des températures

Il faut souligner que la variable relative à la température ne possède pas un coefficient significatif (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

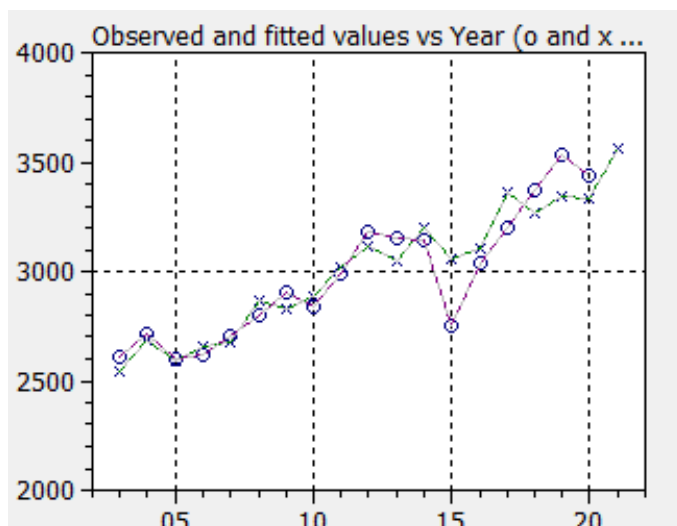


Figure 19. Évolution des rendements (en kg/ha) observés et simulés en Inde pour le modèle 2 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Le modèle 2 s'adapte bien à la première moitié des années prises en compte dans le modèle. Pour la seconde moitié, comme le montre la Figure 19, de plus grands écarts apparaissent entre la véritable valeur du rendement et celle simulée par le modèle. Une fois de plus, il n'y a rien de particulier à signaler pour les résidus du modèle, il semble donc pertinent et fiable (cf. Annexe 14).

La sélection du modèle 3 a été faite un peu différemment des fois précédentes. En se basant uniquement sur les RMSE pour la prévision, le meilleur choix aurait été un modèle avec une variable prédictrice, à savoir, l'accumulation de température. Toutefois, le coefficient de cette variable n'est pas significatif (pour un rejet de l'hypothèse nulle fixé à 0,05). Ce modèle est donc peu fiable d'un point de vue statistique. Le modèle 3 choisi est un modèle qui a certes une valeur de RMSE plus élevée, mais il possède au moins un coefficient significatif. Ce modèle est composé de 3 prédicteurs : le SPI- 3 months, le NDVI cumulé et la température cumulée. Les statistiques de ce modèle sont reprises dans le tableau suivant :

Tableau 9. Statistiques du modèle 3 de l'Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	161.925
R ² ajusté	0.8163
Valeur de la simulation (rendement de 2021)	3565.948 kg/ha

L'équation du modèle 3 s'écrit :

$$Rdt = 5223.019 + 60.199 \cdot TL - 152.269 \cdot SPI - 11.178 \cdot TempCum - 117.719 \cdot NDVICum$$

Avec NDVICum : le NDVI cumulé durant la saison du blé en Inde⁸

⁸ Les autres annotations sont identiques à celles de modèle 2 de l'Inde.

Les coefficients du NDVI cumulé et de la température cumulée ne sont pas significatifs (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

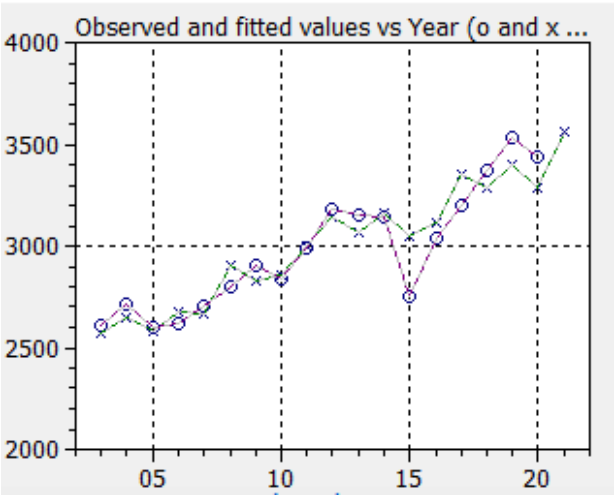


Figure 20. Évolution des rendements (en kg/ha) observés et simulés en Inde pour le modèle 3 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Le rendement prédit pour ce modèle en 2021 est très proche du rendement prédit dans le modèle précédent, la différence entre les 2 prévisions est inférieure à 10 kg/ha. Cela explique pourquoi la Figure 20 et la Figure 19 sont très semblables. Les graphiques relatifs aux résidus sont disponibles à l’annexe 15. Les résidus ne présentent pas d’anomalie.

La moyenne des 3 rendements simulés en fin de saison 2021 vaut : 3522,899 kg/ha.

Le vrai rendement du blé en Inde pour l’année 2021 s’élevait à 3521 kg/ha. La valeur prédite du rendement en moyennant les 3 meilleurs modèles est donc excellente car pratiquement égale.

Une analyse des versions bis des résultats des 3 modèles présentés va à présent être faite. Pour rappel, cette version bis fait une prévision du rendement 2 décades avant la fin de la saison du blé.

Tableau 10. Prévisions et graphiques du rendement (en kg/ha) des versions bis des modèles en Inde (source de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Modèle 1 bis : prévision	Modèle 2 bis : prévision	Modèle 3 bis : prévision
3444.595 kg/ha	4043.959 kg/ha	4297.641 kg/ha

Comme le modèle 1 est une régression linéaire simple, il reste inchangé dans sa version « bis ». Les modèles 2 bis et 3 bis, quant à eux, prévoient un rendement nettement supérieur (de l'ordre de plusieurs centaines de kilogrammes par hectares) par rapport à celui prédit dans les modèles 2 et 3 initiaux. Une moyenne des rendements des modèles bis va être réalisée afin de pouvoir la comparer avec celle des 3 premiers modèles.

Valeur moyenne du rendement prédit pour 2021 2 décades avant la fin de saison : 3928,832 kg/ha.

Comparativement au rendement réel en 2021 qui était de 3521 kg/ha, la valeur moyenne des rendements prédits surestime le rendement de 407,832 kg/ha.

4.3.2 Random forest

En analysant l'importance relative des facteurs agrométéorologiques, 3 facteurs ressortent lorsque l'on travaille avec les données d'entraînement qui, pour rappel, vont de 2002 à 2021 (cf. Figure 21). La variable la plus importante pour construire le modèle, qui domine largement toutes les autres, est le NDVI maximal. La deuxième variable la plus importante est le prix du blé avec une importance relative de 28,17 %. Enfin, la troisième variable ayant un impact significatif est le cumul du NDVIz-score sur la saison du blé avec une importance de 14,54 %. L'ensemble de calibration du modèle a une RMSE de 63,698 kg/ha et l'ensemble de validation a une RMSE de 70,296 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

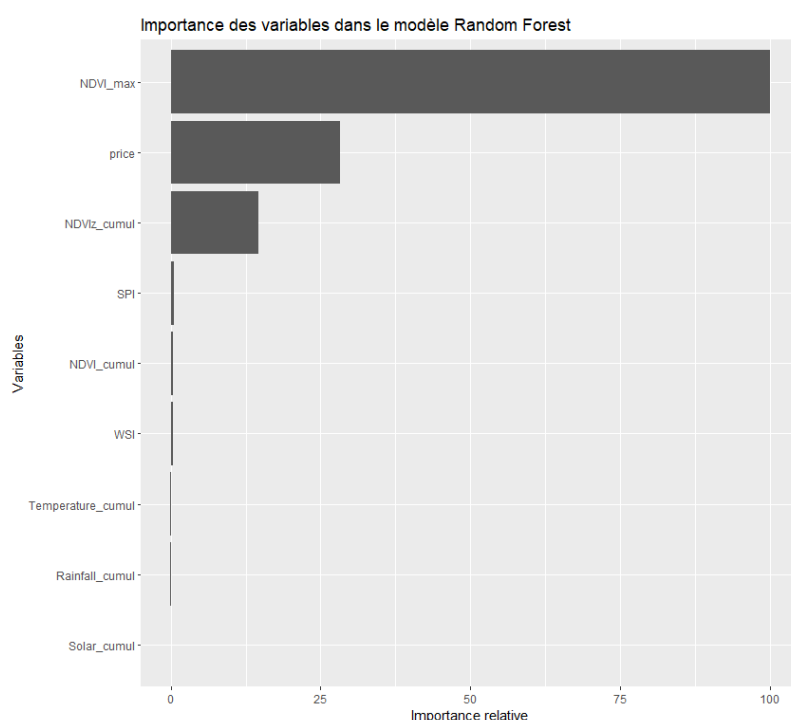


Figure 21. Importance des variables dans le modèle de forêts aléatoires en Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Après avoir fait tourner le modèle sur toute la plage de données, ce sont toujours les mêmes 3 variables qui ressortent comme les plus importantes dans le modèle. Cependant, l'importance du NDVIz-score cumulé descend en dessous de la barre de 5 % avec une

importance de 3,13 %. L'importance du NDVI maximal reste de 100 %. L'importance du prix vaut à présent 30,09 %. Les informations relatives au modèle sont :

Tableau 11. Statistiques du modèle avec les forêts aléatoires de l'Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	70.296
R ² ajusté	0.95404
Valeur de la simulation (rendement de 2021)	3260.29 kg/ha

Ainsi, la simulation du rendement sous-estime le véritable rendement en 2021 de 260,71 kg/ha puisque le rendement recensé par la FAO en 2021 est de 3521 kg/ha⁹. Le rendement a ensuite été ré-évalué en ne connaissant pas les valeurs des variables pour les 2 dernières décades de la saison du blé en Inde. La prévision du rendement vaut : 3260,59 kg/ha. Cette prévision est pratiquement identique à la valeur du rendement simulé avec la plage de données complètes jusqu'à la fin de la saison puisqu'il ne le dépasse que de 0,3 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

⁹ Les graphiques relatifs aux résidus de ce modèle sont disponibles à l'annexe 16 à titre informatif.

4.4 Russie

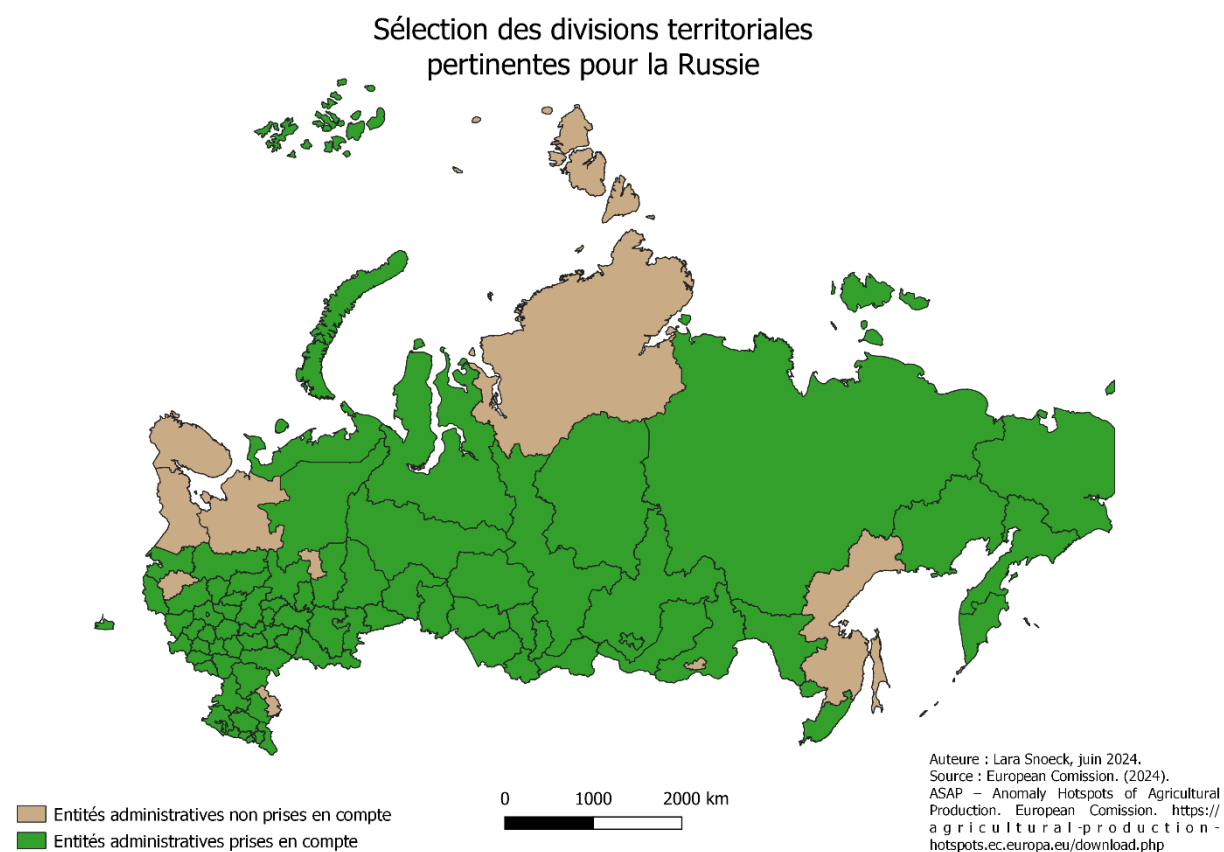


Figure 22. Sélection des divisions territoriales pertinentes pour la Russie.

La majorité des entités administratives russes ont été intégrées à la modélisation. Les entités non sélectionnées ne possèdent pas de cultures de blé sur leurs territoires ou très peu. Les entités ayant une surface cultivée inférieure à 1000 km² n'ont pas été reprises à moins que la superficie de l'entité elle-même ne soit du même ordre. Les entités qui ont été sélectionnées pour la modélisation sont répertoriées à l'annexe 8 (European Commission, 2024 ; FAO, 2022).

La saison du blé en Russie a été définie comme débutant à la 3^{ème} décennie du mois de janvier et se terminant à la première décennie du mois d'août (Franch *et al.*, 2022, Annexes).

4.4.1 Régression multiple

L'évolution du rendement du blé cultivé en Russie entre 2002 et 2020 est représentée à la Figure 23. La p-value est inférieure à 0,0000, ce qui indique que la tendance linéaire du modèle est statistiquement significative.

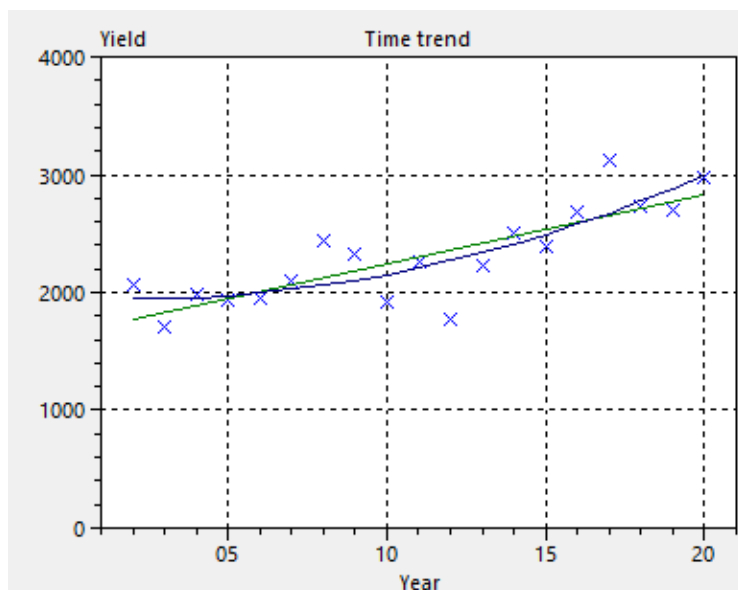


Figure 23. Évolution du rendement en Russie (en kg/ha) entre 2002 et 2021 (source de données : FAO, 2023).

Les 3 meilleurs modèles lorsque les données des variables prédictrices sont intégrées jusqu'à la toute fin de la saison vont une nouvelle fois être nommés « modèle 1 », « modèle 2 » et « modèle 3 », en commençant par le meilleur. Ces modèles vont être détaillés ci-dessous.

Le modèle 1 reprend une seule variable explicative du rendement : le NDVIz-score. Les statistiques relatives à ce modèle sont les suivantes :

Tableau 12. Statistiques du modèle 1 de la Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	209.412
R ² ajusté	0.76835
Valeur de la simulation (rendement de 2021)	2739.572 kg/ha

L'équation du modèle 1 se note :

$$\text{Rdt} = 25,478 + 49,576 \cdot \text{TL} + 29,133 \cdot \text{CumNDVIz}$$

Avec Rdt : le rendement

TL : la tendance linéaire temporelle (Année – 1965)

CumNDVIz : le cumul du NDVIz-score (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

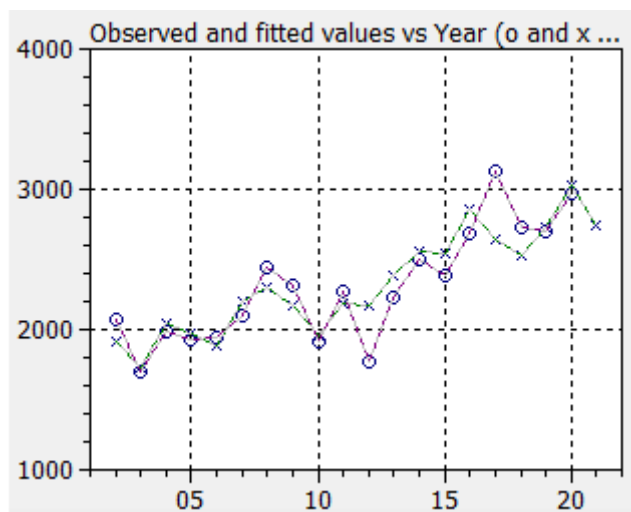


Figure 24. Évolution des rendements (en kg/ha) observés et simulés en Russie pour le modèle 1 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

La Figure 24 montre un bon ajustement des rendements simulés par le modèle comparativement aux vraies valeurs de rendements jusqu'en 2011. Puis, quelques écarts sont observables. Par exemple, c'est en 2012 et en 2017 que les plus grands écarts entre valeurs observées et ajustées ont lieu. Ces différences sont de l'ordre de 400-600 kg/ha. Les résidus sont indépendants les uns des autres et sont distribués selon une loi normale (cf. Annexe 17).

Le modèle 2 regroupe deux prédicteurs qui sont : la valeur maximale du NDVI atteinte durant la saison et la température cumulée sur la saison. Les statistiques en lien avec ce modèle sont présentées dans le tableau ci-dessous :

Tableau 13. Statistiques du modèle 2 de la Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	210.832
R ² ajusté	0.79575
Valeur de la simulation (rendement de 2021)	2986.935 kg/ha

L'équation du modèle 2 s'écrit de la façon suivante :

$$\text{Rdt} = -4754,663 + 60,966 \cdot \text{TL} + 5,527 \cdot \text{CumTemp} + 4979,09 \cdot \text{NDVIMax}$$

Avec CumTemp : la température cumulée

NDVIMax : la valeur du NDVI maximale atteinte durant la saison

Le coefficient de la température cumulée dans l'équation n'est pas significatif (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

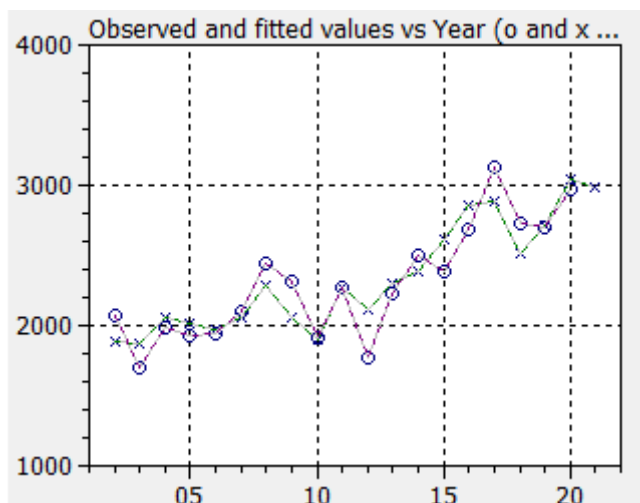


Figure 25. Évolution des rendements (en kg/ha) observés et simulés en Russie pour le modèle 2 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Contrairement au modèle 1, qui prévoit un rendement inférieur en 2021 par rapport à l'année 2020, pour le modèle 2, le rendement prédit est plutôt similaire au rendement de l'année précédente (cf. Figure 25). Comme pour le modèle 1, il n'y a rien de particulier à signaler pour les résidus du modèle 2 (cf. Annexe 18).

Le 3^{ième} meilleur modèle (modèle 3) est une fois de plus composé d'une variable relative à l'état de santé de la végétation. Cette fois, il s'agit du NDVI cumulé (NDVICum) sur la saison du blé en Russie. Les informations relatives à ce modèle sont :

Tableau 14. Statistiques du modèle 3 de la Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	214.909
R ² ajusté	0.76069
Valeur de la simulation (rendement de 2021)	2722.513 kg/ha

L'équation du modèle 3 est :

$$\text{Rdt} = -2219,232 + 49,122 \cdot \text{TL} + 260,395 \cdot \text{NDVICum}$$

(European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

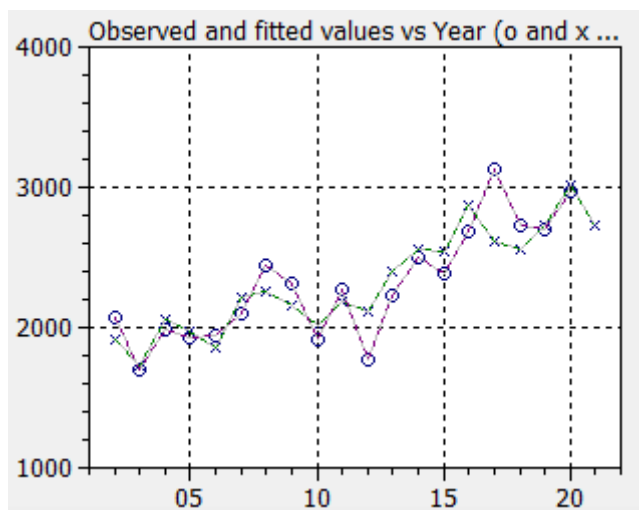


Figure 26. Évolution des rendements (en kg/ha) observés et simulés en Russie pour le modèle 3 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Comme le montre la Figure 26, le modèle 3 est très similaire au modèle 1. Il prévoit un rendement en 2021 inférieur d'environ 17 kg/ha par rapport à la prévision du modèle 1. Les résidus sont en accord avec les hypothèses posées pour les modèles de régression linéaire multiple (cf. Annexe 19).

La moyenne des rendements estimés par les 3 modèles vaut : 2816,34 kg/ha.

Cette moyenne surestime le rendement réel de 91,94 kg/ha puisque le rendement réel en Russie en 2021 était de 2724,4 kg/ha.

Les prévisions faites par les mêmes modèles quand les données sont disponibles jusqu'à 2 décades avant la fin de la saison vont à présent être données.

Tableau 15. Prévisions et graphiques du rendement (en kg/ha) des versions bis des modèles en Russie (source de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Modèle 1 bis : prévision	Modèle 2 bis : prévision	Modèle 3 bis : prévision
2750.35 kg/ha	2734.537 kg/ha	2429.728 kg/ha

Le modèle 1 bis prédit un rendement légèrement supérieur à celui simulé par le modèle 1. Cette petite surestimation vaut à peu près 10 kg/ha. Le modèle 2 bis prévoit un rendement inférieur de l'ordre de 250 kg/ha en comparaison avec le modèle 2. Enfin, le modèle 3 bis

prévoit un rendement d'une valeur de l'ordre d'un peu moins de 300 kg/ha en moins vis-à-vis de la valeur simulée par le modèle 3.

La moyenne des prévisions de rendement des 3 modèles bis équivaut à : 2638,205 kg/ha.

Cette fois, la moyenne sous-estime le rendement réel de 86,195 kg/ha.

4.4.2 Ranfom forest

Comme le montre la Figure 27, contrairement aux pays analysés précédemment, il n'y a pas de variable qui domine avec une importance relative de 100 % pour le modèle construit sur les années d'entraînement de la Russie. Toutefois, il y a tout de même 3 variables avec une importance plus conséquente qui se démarquent : le prix avec une importance de 32,15 %, le NDVIz-score cumulé avec une importance de 19,49 % et NDVI maximal avec une importance de 16,59 %. La RMSE de l'ensemble de calibration vaut 105,474 kg/ha tandis que celle de l'ensemble de validation s'élève à 121,833 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

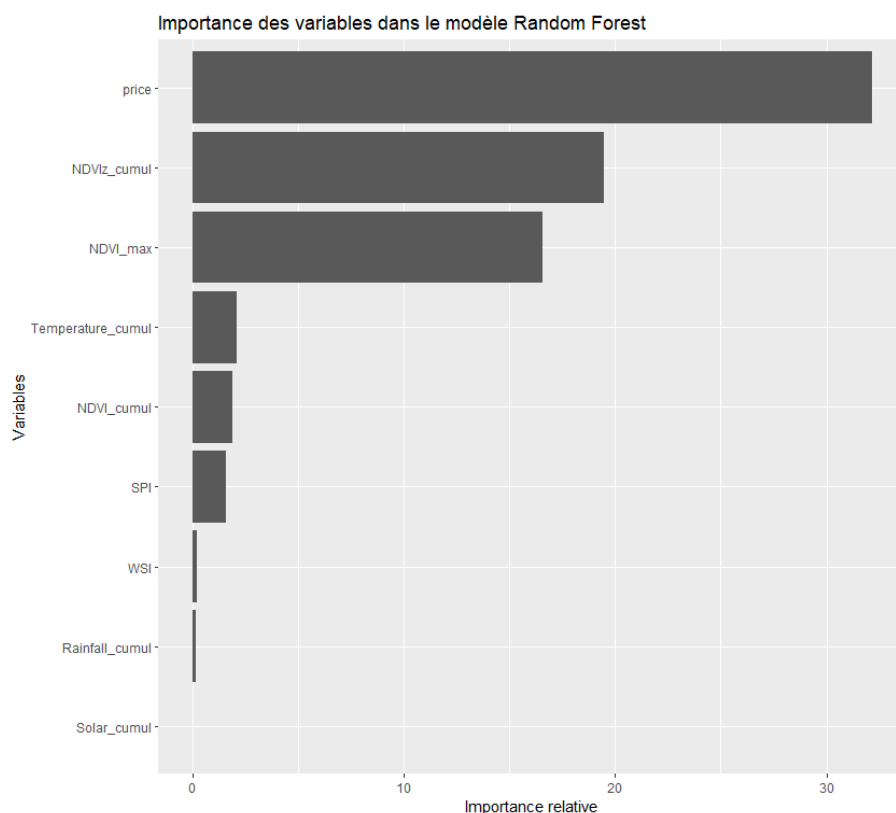


Figure 27. Importance des variables dans le modèle de forêts aléatoires en Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Après intégration des données de l'année 2021 au modèle, l'ordre des variables ayant un impact conséquent dans le modèle a changé comme le montre la Figure 28. Le prix a toujours la plus grande importance relative qui vaut 24,75 %. La seconde variable la plus importante est maintenant le NDVI maximal qui a une importance de 9,25 % suivi par le NDVI-z score cumulé avec une importance relative de 8,75 %.

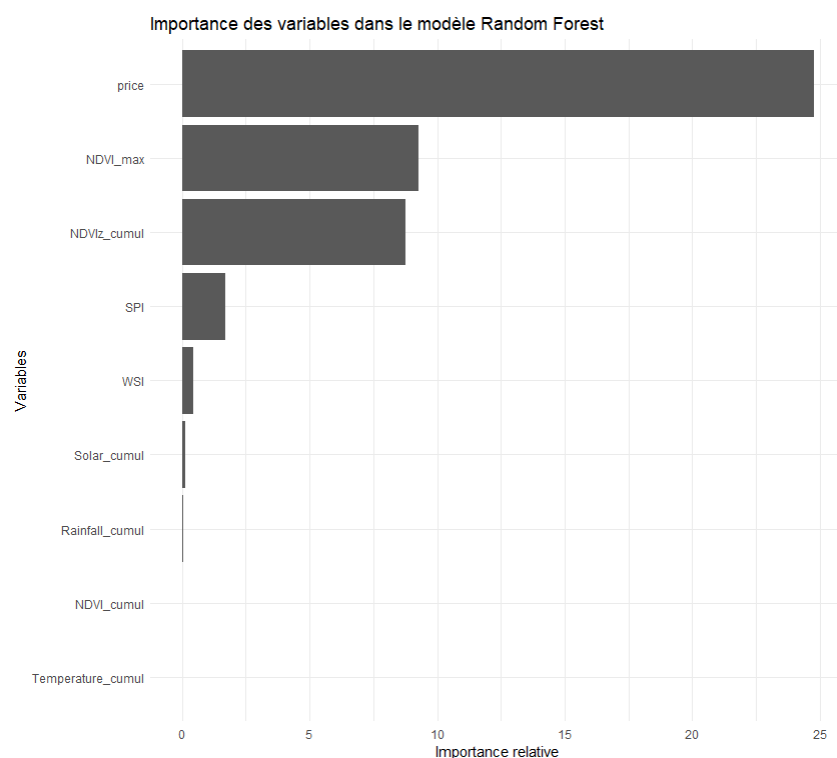


Figure 28. Importance des variables dans le modèle de forêts aléatoires en Russie après intégration des données de 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Les informations relatives au modèle sont reprises ci-après :

Tableau 16. Statistiques du modèle avec les forêts aléatoires de la Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	121.833
R ² ajusté	0.92478
Valeur de la simulation (rendement de 2021)	2799.82 kg/ha

La valeur simulée du rendement sur-estime de 75,42 kg/ha par rapport à la véritable valeur du rendement en Russie en 2021 qui vaut 2724,4 kg/ha¹⁰. La valeur de rendement que prévoit le modèle 2 décades avant la fin de saison est identique à celle simulée à la fin de la saison (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

¹⁰Les graphiques relatifs aux résidus de ce modèle sont disponibles à l'annexe 20 à titre informatif.

4.5 États-Unis

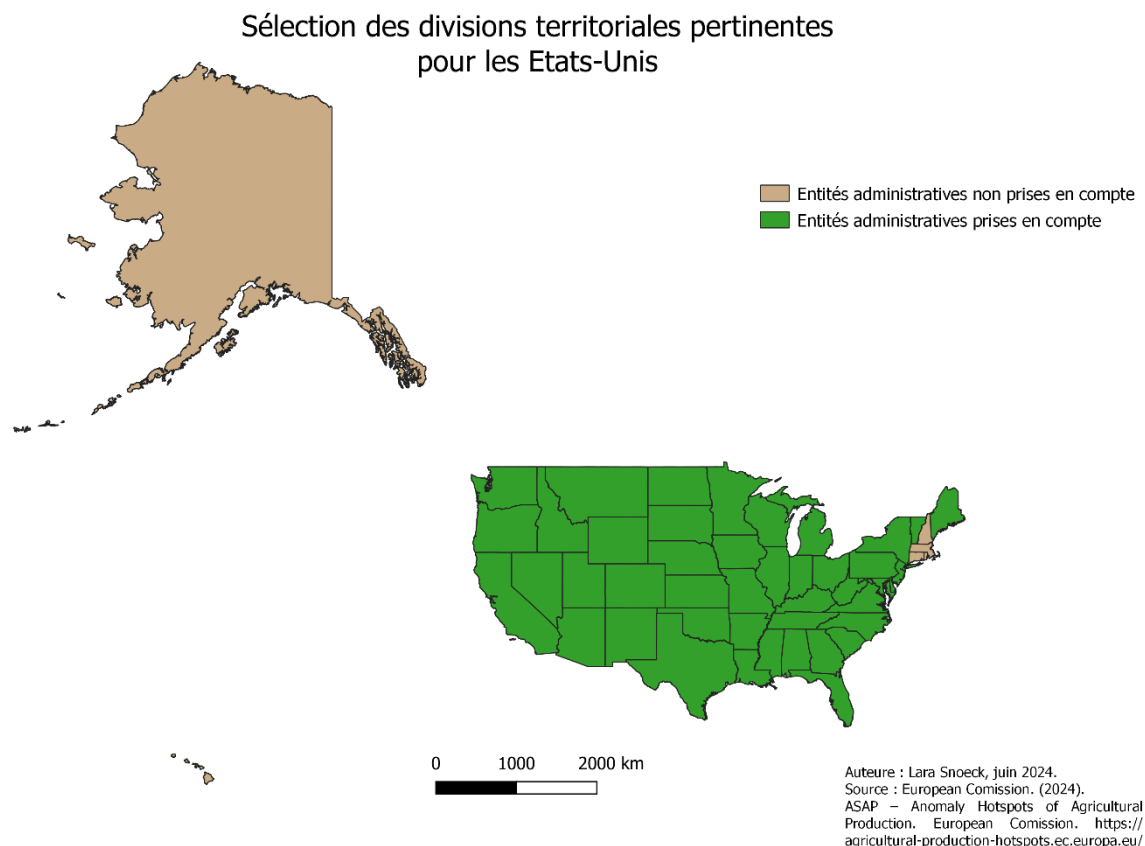


Figure 29. Sélection des divisions territoriales pertinentes pour les États-Unis.

Aux États-Unis, pratiquement tous les États sont sélectionnés car le blé est cultivé un peu partout dans le pays. L'Alaska n'a pas été repris puisqu'on n'y cultive pas de blé au même titre que dans quelques petits États situés dans le nord-est du pays. La liste des États intégrés pour la modélisation est disponible à l'annexe 8 (European Commission, 2024 ; FAO, 2022).

Le début de la saison du blé aux États-Unis a lieu au tout début de l'année, soit à la première décennie du mois de janvier. La saison se termine à la deuxième décennie du mois d'août (Franch *et al.*, 2022, Annexes).

4.5.1 Régression multiple

La Figure 30 illustre l'évolution du rendement aux États-Unis sur les années d'entraînement du modèle qui, pour rappel, vont de 2002 à 2020. La p-value de la régression linéaire s'ajustant au mieux aux données du rendement vaut 0,0001.

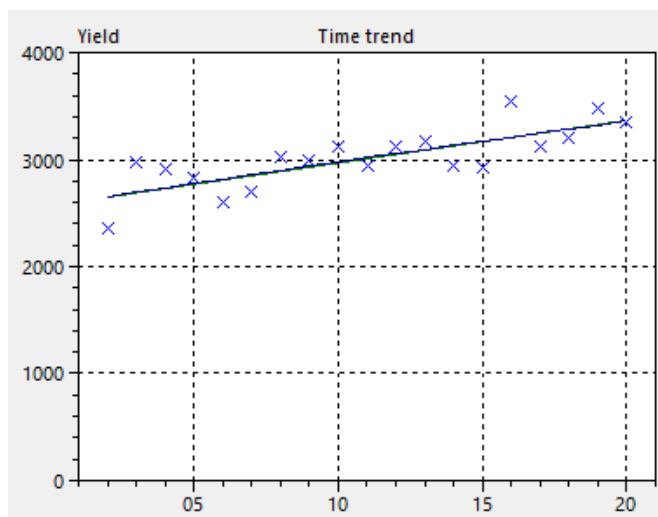


Figure 30. Évolution du rendement aux États-Unis (en kg/ha) entre 2002 et 2021 (source de données : FAO, 2023).

Comme pour les autres pays, les 3 meilleurs modèles lorsque toutes les données (mis à part le rendement de l'année à prédire) sont intégrées dans la modélisation vont être présentés.

Le meilleur modèle (modèle 1) est formé par 3 prédicteurs : le prix du blé sur le marché mondial au moment de la récolte l'année précédente, le NDVIz-score cumulé sur la saison définie précédemment et le rayonnement solaire également cumulé sur la saison. Le tableau suivant récapitule des statistiques intéressantes propres au modèle1 :

Tableau 17. Statistiques du modèle 1 des États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	191.912
R ² ajusté	0.64808
Valeur de la simulation (rendement de 2021)	3346.861 kg/ha

L'équation du modèle 1 est la suivante :

$$\text{Rdt} = 6686,426 + 30,097 \cdot \text{TL} + 16,032 \text{ CumNDVIz} - 1.27 \cdot 10^{-3} \text{ CumSun} + 0,95 \cdot \text{Price}$$

Avec Rdt : le rendement

TL : la tendance linéaire temporelle (Année – 1965)

CumNDVIz : le cumul du NDVIz-score

CumSun : le cumul des radiations solaires

Price : le prix

Il est important de souligner que pour ce modèle tout comme pour les 2 qui vont suivre, les coefficients de régression ne sont pas significatifs car aucun des modèles créés par le logiciel ne possède de coefficient significatif (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

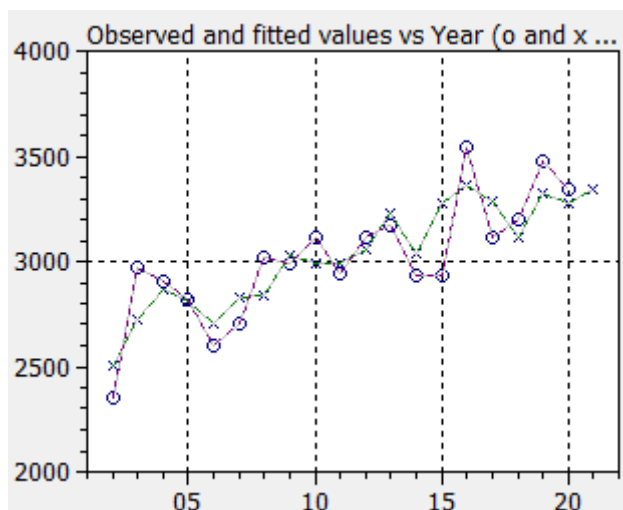


Figure 31. Évolution des rendements (en kg/ha) observés et simulés aux États-Unis pour le modèle 1 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

La Figure 31 montre que les simulations du rendement par le modèle 1 (représentées par des croix) suivent approximativement bien le rendement réel enregistré au cours des différentes années. Pour certaines années, comme par exemple les années 2003 et 2015, un écart notable de l'ordre 300-400 kg/ha peut être observé. Les résidus de ce modèle respectent bien les 2 hypothèses vérifiées à savoir l'absence de corrélation entre les résidus, caractérisée par une distribution aléatoire de la valeur des résidus autour de 0 ainsi qu'une distribution de résidus suivant une loi normale. Les graphiques relatifs aux résidus du modèle 1 sont disponibles à l'annexe 21.

Le second meilleur modèle, qui possède donc la seconde plus petite valeur pour la RMSE de la prévision du rendement va être nommé « modèle 2 ». Tout comme le modèle 1, il est composé de 3 prédicteurs : le prix, les précipitations cumulées (CumRain) et le NDVIz-score cumulé. Le tableau ci-après reprend les principales informations du modèle 2 :

Tableau 18. Statistiques du modèle 2 des États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	192.799
R ² ajusté	0.6393
Valeur de la simulation (rendement de 2021)	3213.261 kg/ha

Le modèle 2 peut être mis en équation sous la forme suivante :

$$\text{Rdt} = 717,14 + 22,809 \cdot \text{TL} + 18,173 \cdot \text{CumNDVIz} + 1,869 \cdot \text{CumRain} + 0,989 \cdot \text{Price}$$

Pour rappel, les coefficients des prédicteurs de cette équation ne sont pas significatifs (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

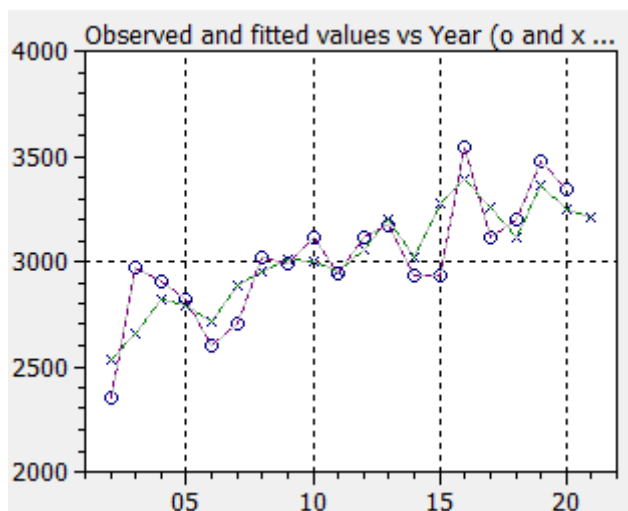


Figure 32. Évolution des rendements (en kg/ha) observés et simulés aux États-Unis pour le modèle 2 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

La Figure 32 permet de visualiser les rendements simulés par le modèle 2. Mis à part pour la prévision en 2021 où le rendement est un peu plus faible (de l'ordre de 100 kg/ha) par rapport au modèle 1, les rendements simulés sont assez similaires entre les deux modèles. Il n'y a rien de particulier à signaler pour les résidus de ce modèle (cf. Annexe 22).

Le troisième meilleur modèle (modèle 3) est aussi constitué de 3 variables : le prix, le NDVI cumulé (CumNDVI) et le rayonnement solaire. Voici les caractéristiques du modèle 3 :

Tableau 19 Statistiques du modèle 3 des États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	194.546
R ² ajusté	0.64113
Valeur de la simulation (rendement de 2021)	3340.337 kg/ha

L'équation du modèle 3 est :

$$\text{Rdt} = 3747,895 + 30,813 \cdot \text{TL} + 196,53 \cdot \text{CumNDVI} - 1.34 \cdot 10^{-3} \cdot \text{CumSun} + 0,876 \cdot \text{Price}$$

Une fois de plus, il faut rappeler que les coefficients des variables de cette équation ne sont pas significatifs (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

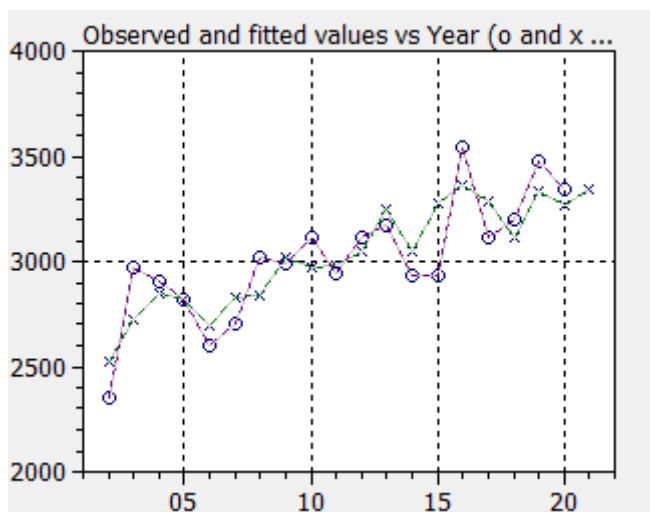


Figure 33. Évolution des rendements (en kg/ha) observés et simulés aux États-Unis pour le modèle 3 de 2002 à 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

L'évolution du rendement réel et du rendement simulé par le modèle 3 est présentée à la Figure 33. L'allure de la courbe du rendement simulé par le modèle est très proche de celle simulée par le modèle 1. De plus, les rendements prédits pour ces 2 modèles en 2021 sont très proches puisque la différence entre les 2 est à peine plus élevée que 5 kg/ha. Les résidus du modèle 3 confortent la robustesse du modèle. Les graphiques qui leur sont associés sont disponibles à l'annexe 23.

La moyenne des rendements des 3 modèles au moment de la fin de saison 2021 est de : 3300,153 kg/ha.

Cette moyenne est supérieure à la vraie valeur du rendement aux États-Unis en 2021 qui était de 2980,5 kg/ha. L'estimation du rendement grâce aux régressions linéaires multiples lorsque les valeurs de toutes les variables sont connues jusqu'à la fin de la saison montre, en effet, une surestimation de 319,653 kg/ha par rapport à la valeur réelle.

L'analyse des résultats des 3 mêmes modèles lorsque les données sont connues jusqu'à 2 décades avant la fin de la saison va maintenant être réalisée.

Tableau 20. Prévisions et graphiques du rendement (en kg/ha) des versions bis des modèles aux États-Unis (source de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Modèle 1 bis : prévision	Modèle 2 bis : prévision	Modèle 3 bis : prévision
3892.495 kg/ha	3111.525 kg/ha	3646.34 kg/ha

La version bis du modèle 1 prévoit un rendement du blé en 2021 largement supérieur à celle du modèle 1 puisque la prévision a augmenté d'environ 550 kg/ha. À l'inverse, le modèle 2 bis prédit un rendement plus faible que le modèle 2 avec une réduction du rendement d'environ 100kg/ha. Enfin, le modèle 3 bis prévoit un rendement plus élevé de plus ou moins 300 kg/ha par rapport au modèle 3.

Le rendement moyen prévu par les 3 modèles bis vaut : 3550,12 kg/ha.

Comparativement à la vraie valeur du rendement, l'écart se creuse encore un peu plus avec cette fois une surestimation de 569,62 kg/ha.

4.5.2 Random forest

Le modèle basé sur les données d'entraînement des États-Unis dénombre 4 variables explicatives principales (cf. Figure 34). La première, avec une importance relative de 100 % est le NDVI maximal. La seconde est le prix, qui a une importance de 78,06 %. La troisième est le SPI-3 months avec une importance de 18,71 %. Enfin, la quatrième est le NDVIz-score cumulé avec une importance de 14,04 %. La RMSE de l'ensemble de calibration du modèle vaut 78,676 kg/ha. Celle de l'ensemble de validation a une valeur de 87,467 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

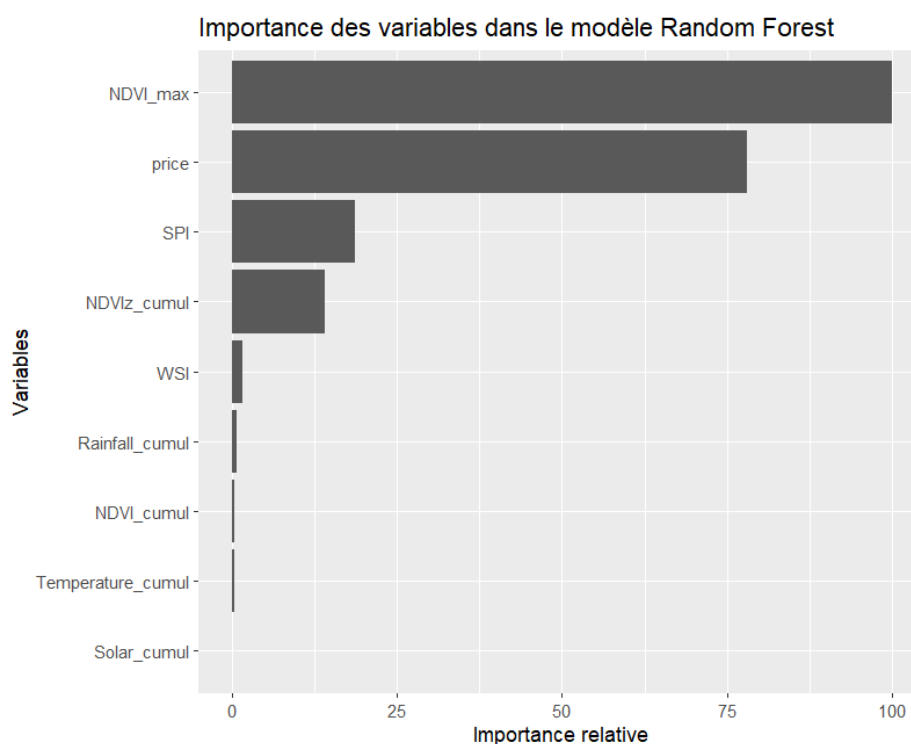


Figure 34. Importance des variables dans le modèle de forêts aléatoires aux États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Les résultats du modèle sur les années 2002 à 2021 montre une inversion dans l'ordre d'importance des 2 variables ayant le plus de poids dans la composition du modèle (cf. Figure 35). La variable la plus influente, avec une importance de 95,88 % est le prix. La seconde variable est le NDVI maximal avec 62,52 %. La troisième et la quatrième variables, à savoir le

SPI-3 months et le NDVIz cumulé, restent inchangées avec des importances de 19,87 % et 16,37 % respectivement.

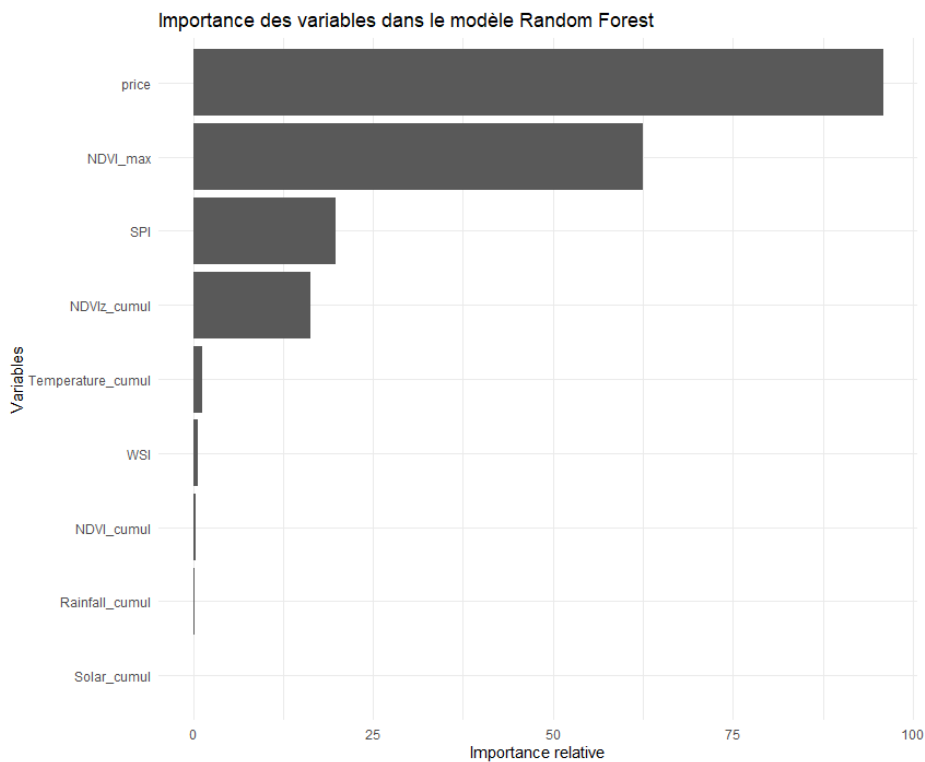


Figure 35. Importance des variables dans le modèle de forêts aléatoires aux États-Unis après intégration des données de 2021 (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Les statistiques relatives à ce modèle sont :

Tableau 21. Statistiques du modèle avec les forêts aléatoires des États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	87.467
R² ajusté	0.90068
Valeur de la simulation (rendement de 2021)	3211.74 kg/ha

Cette valeur de rendement simulée pour 2021 surestime de 231,24 kg/ha le rendement réellement recensé par la FOA pour 2021 puisqu’il était de 2980,5 kg/ha aux Etats-Unis¹¹. Le rendement prédit par le modèle, 2 décades avant la fin de la saison est de 3214,2 kg/ha. Autrement dit, il est supérieur de 2,46kg/ha par rapport au rendement simulé par le modèle avec les données jusqu’à la fin de la saison et il est supérieur de 233,7 kg/ha par rapport à la valeur réelle (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

¹¹ Les graphiques relatifs aux résidus de ce modèle sont disponibles à l’annexe 24 à titre informatif.

4.6 France

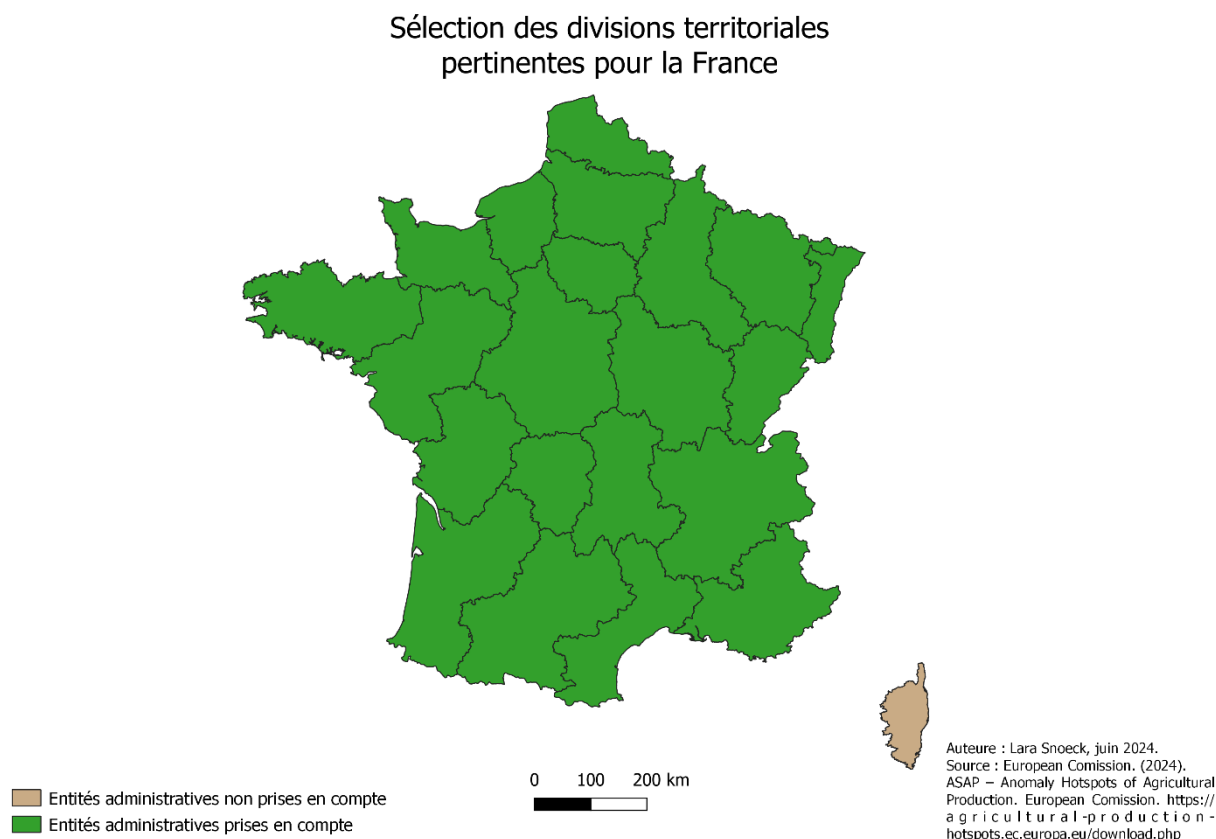


Figure 36. Sélection des divisions territoriales pertinentes pour la France.

En France, le blé est cultivé partout dans la zone métropolitaine (*cf.* Figure 36). C'est pourquoi toutes les régions françaises ont été prises en compte dans la modélisation mis à part la Corse. L'annexe 8 reprend le nom de toutes les entités administratives retenues (European Commission, 2024 ; FAO, 2022).

La saison du blé en France a été définie comme débutant tout début d'année, soit à la première décade de mois de janvier et se finissant à la seconde décade du mois d'août (Franch *et al.*, 2022, Annexes).

4.6.1 Régression multiple

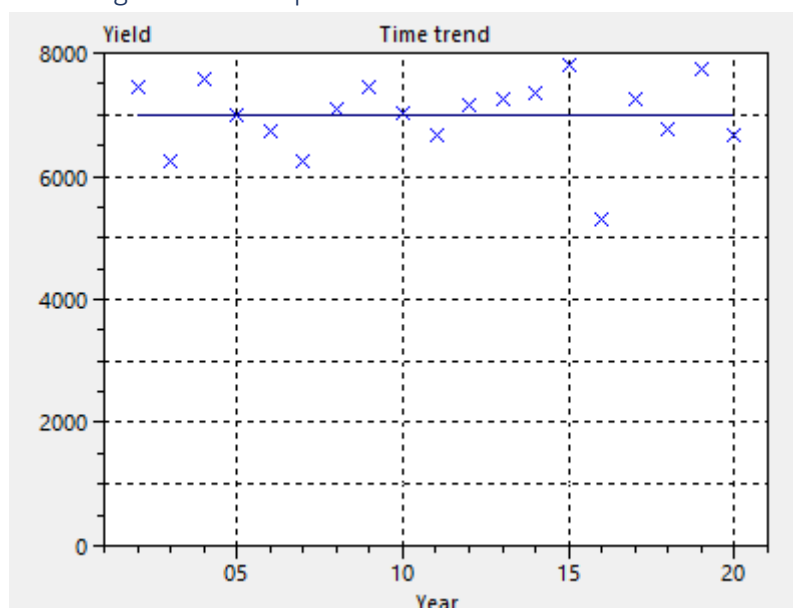


Figure 37. Évolution du rendement en France (en kg/ha) entre 2002 et 2021 (source de données : FAO, 2023).

L'évolution du rendement en France sur les années d'entraînement (2002 à 2020) ne présente aucune tendance linéaire (cf. Figure 37). La p-value est très proche de 1, elle vaut 0,9803. Par conséquent, un modèle de régression linéaire multiple n'est, dans ce cas-ci, pas pertinent car il n'est pas significatif.

4.6.2 Random forest

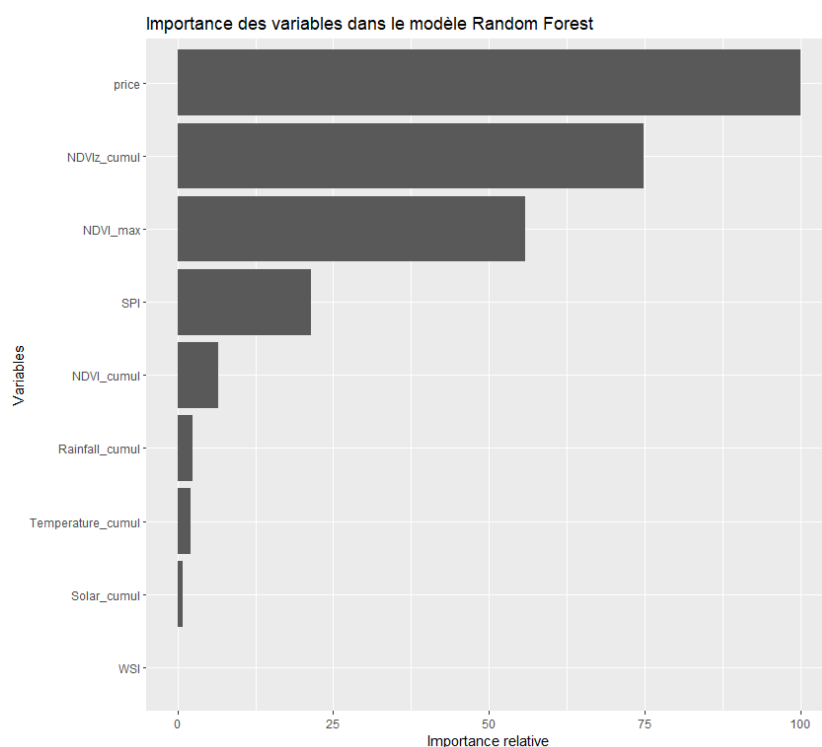


Figure 38. Importance des variables dans le modèle de forêts aléatoires en France (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

La Figure 38 montre l'importance des différents facteurs dans le modèle formé sur base des données d'entraînement qui vont de 2002 à 2020. La variable ayant la plus grande importance dans le modèle (100 %) est le prix du blé. La seconde variable est le NDVIz-score cumulé avec une importance de 74,92 %. C'est la valeur maximum du NDVI qui arrive en troisième position avec une importance relative de 55,83 %. On retrouve ensuite le SPI-3 months avec une importance de 21,51 % suivi par le NDVI cumulé dont l'importance vaut 6,46 %. Les autres facteurs ont une importance inférieure à 5 %. L'ensemble de calibration de ce modèle a une RMSE de 275,018 kg/ha et l'ensemble de validation a une RMSE de 312,949 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

Après avoir intégré toutes les données de 2021, mis à part le rendement, au modèle, l'ordre d'importance des 5 variables principales reste inchangé. Le prix a toujours une importance relative de 100 %. Le NDVIz-score a une importance de 85,48 %. L'importance du NDVI maximum vaut 63,62 %, celle du SPI-3 months vaut 17,23 % et enfin celle du NDVI cumulé vaut 6,28 %. Les statistiques de ce modèle sont reprises dans le tableau ci-dessous :

Tableau 22. Statistiques du modèle avec les forêts aléatoires de la France (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

RMSE de la simulation	312.95
R ² ajusté	0.83856
Valeur de la simulation (rendement de 2021)	6295.21 kg/ha

Le rendement réel en France en 2021 équivaut à 6928,4 kg/ha. Le rendement simulé par le modèle sous-estime donc le rendement de 633,19 kg/ha¹². Le rendement prédit pour 2021 deux décades avant la fin de la saison vaut 6300,89 kg/ha. Il sous-estime donc le rendement réel de 627,51 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

¹²Les graphiques relatifs aux résidus de ce modèle sont disponibles à l'annexe 25 à titre informatif.

4.7 Comparaison des résultats

Tableau 23. Récapitulatif des RMSE, R^2 et des variables (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024)

Modèle	Pays	RMSE	R^2	Variables principales
Régression linéaire simple	Monde	74.31	0.9251	/
Régression linéaire multiple	Chine	108.745	0.97361	Rayonnement cumulé, précipitations cumulées, (prix), (WSI)
Random Forest	Chine	187.825	0.93251	NDVI maximum, NDVIz cumulé, prix
Régression linéaire multiple	Inde	152.493	0.8039	SPI, (température cumulée), (NDVI cumulé)
Random Forest	Inde	70.296	0.95404	NDVI maximum, prix, NDVI cumulé
Régression linéaire multiple	Russie	211.718	0.77493	NDVIz cumulé, NDVI maximum, NDVI cumulé, (température cumulée)
Random Forest	Russie	121.833	0.92478	Prix, NDVIz cumulé, NDVI maximum
Régression linéaire multiple	Etats-Unis	193.086	0.64284	(NDVIz cumulé), (rayonnement cumulé), (prix), (précipitations cumulées), (NDVI cumulé)
Random Forest	Etats-Unis	87.467	0.90068	NDVI maximum, prix, SPI, NDVIz cumulé
Régression linéaire multiple	France	/	/	/
Random Forest	France	312.95	0.83856	prix, NDVIz cumulé, NDVI maximum, SPI

Le Tableau 23 regroupe les principales informations présentées dans les résultats jusqu'à présent à savoir les RMSE des différents modèles ainsi que les R^2 (les RMSE pour les régressions multiples sont la moyenne des RMSE des 3 meilleurs modèles, il en va de même pour les R^2) et les variables qui composent ces modèles (pour la régression multiple, les variables avec un coefficient non significatif ont été indiquées entre parenthèses).

Dans un premier temps, une comparaison entre la modélisation par régression multiple et par les forêts aléatoires a été réalisée. En ce qui concerne les variables, dans le cas des modèles utilisant Random Forest, une certaine récurrence dans la sélection des indicateurs est observée. En effet, pour chacun des modèles, le prix est toujours une variable ayant une importance majeure au même titre que les indicateurs relatifs à l'état de santé de la végétation (le NDVI maximal, le NDVI cumulé et le NDVI z-score cumulé). Le SPI 3-months est présent dans 2 dans 5 modèles Random Forest. En revanche, pour la régression linéaire multiple, il n'y a pas de tendance particulière. Les variables sélectionnées varient d'un modèle à l'autre. En termes de précision, les modèles utilisant les forêts aléatoires surpassent en général ceux basés sur des régressions multiples comme en témoigne la valeur de la RMSE, qui, pour les modèles Random Forest vaut approximativement la moitié de celle des modèles de régression multiple. Cette tendance ne se vérifie cependant pas pour la Chine puisque la RMSE de la régression multiple vaut 108,745 kg/ha tandis que celle du modèle avec les forêts aléatoires est de 187,825 kg/ha.

Dans un deuxième temps, avant d'assembler les résultats des meilleurs modèles des 5 pays étudiés, une comparaison des racines carrées des erreurs quadratiques moyennes va d'abord

être détaillée. C'est sur base de ce critère que la performance des modèles a été évaluée. Ce choix repose sur une volonté de faire primer la précision des modèles plutôt que leur capacité à utiliser les variables pour expliquer la variabilité du rendement. En se concentrant uniquement sur le meilleur modèle pour chaque pays, c'est-à-dire la régression multiple pour la Chine et Random Forest pour les 4 autres, le modèle offrant le plus de précision est le modèle Random Forest pour l'Inde avec une RMSE de 70,296 kg/ha. Cette valeur est inférieure à la RMSE du modèle de régression linéaire simple qui vaut 74,31 kg/ha. A contrario, les valeurs des RMSE des modèles des autres pays sont supérieures à celle du modèle de référence. Pour la Chine avec une RMSE de 108,745 kg/ha, la Russie avec une RMSE de 121,833 kg/ha et les Etats-Unis avec une RMSE de 87,467 kg/ha, la précision des modèles reste comparable à celle du modèle de référence puisque les valeurs oscillent autour de 100 kg/ha. Par contre, la précision de la modélisation du rendement en France est nettement inférieure avec une RMSE de 312,95 kg/ha.

Pour pouvoir former le modèle à l'échelle globale en combinant les résultats de modèles des 5 pays étudiés, il ne faut pas oublier que contrairement au modèle de régression linéaire simple qui permet de faire une prévision dès que le rendement de l'année précédente (2020 dans le cas étudié) est connu, le modèle combinant plusieurs pays se construit au fur et à mesure de l'avancement dans l'année. En effet, comme récapitulé à la Figure 39, le moment de la récolte ou encore des prévisions (2 décades avant la récolte) se fait à des moments différents en fonction du pays. Les valeurs de rendement prises en compte sont celles du modèle ayant fourni les résultats les plus précis : la régression multiple pour la Chine et les forêts aléatoires pour les autres pays.

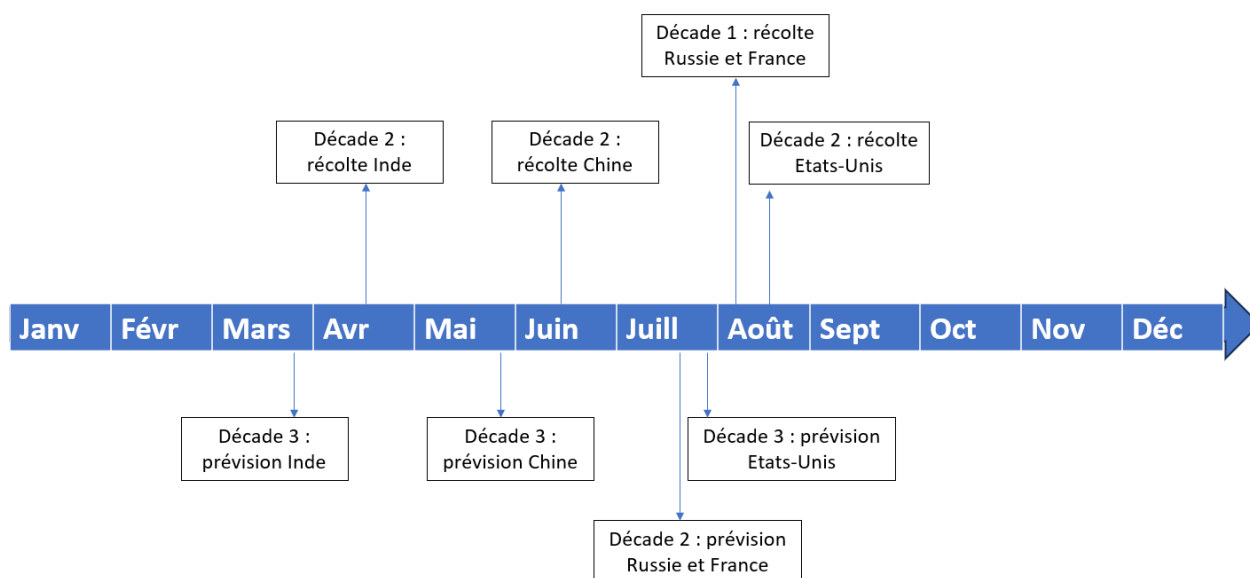


Figure 39. Ligne du temps des prévisions et des récoltes (Franch et al., 2022, Annexes).

La détermination du rendement global va donc pouvoir être affinée à mesure que l'année défile. La prévision finale du rendement global sur base des 5 pays étudiés va ainsi pouvoir être réalisée 2 décades avant la récolte du blé aux États-Unis puisqu'il s'agit du pays où la récolte se fait le plus tardivement. Le cheminement de construction de la prévision du rendement à l'échelle mondiale pour l'année 2021 va être présenté ci-dessous.

Pour commencer, le premier pays dans lequel le blé est récolté est l'Inde (la prévision du rendement pour ce pays a déjà été donnée précédemment). Le second pays pour lequel le blé est récolté est la Chine. Dès lors, une prévision du rendement mondial basée sur les modèles de 2 pays peut être réalisée à la fin du mois de mai puisque la prévision pour le rendement en Chine a été réalisée à la fin de la 3^{ième} décade du mois de mai. Pour l'Inde, comme la récolte aura déjà eu lieu à ce moment de l'année, ce n'est pas la prévision pour le rendement en Inde qui va être prise en compte, mais la valeur de la simulation du rendement (obtenu avec les données complètes jusqu'à la fin de la saison) puisque cette valeur est plus précise que la prévision. Le rendement global va être déterminé en pondérant selon la superficie du pays où du blé a été cultivé en 2021 par rapport à la superficie cultivée cumulée des pays pris en compte :

$$\text{Rendement}_{\text{global}} = \sum_{i=1}^n \left(\frac{\text{Superficie}_i}{\text{Superficie}_{\text{globale}}} \cdot \text{Rendement}_i \right)$$

Avec i : les nombres de pays pris en compte

Superficie globale : cumul de la superficie où le blé a été cultivé dans les pays pris en compte en 2021

Superficie _{i} : la superficie où du blé a été cultivé dans le $i^{\text{ième}}$ pays en 2021

Rendement _{i} : le rendement du blé dans le $i^{\text{ième}}$ pays en 2021.

Le rendement du blé mondial estimé grâce à la modélisation complexe à la fin du mois de mai vaut 4062,91 kg/ha. Une seconde estimation peut être réalisée à la fin de la 2^{ième} décade de juillet, soit vers le 21 juillet. Cette seconde estimation va prendre en compte le rendement simulé en Inde et en Chine ainsi que la prévision du rendement en Russie et en France. La valeur alors obtenue est de 3930,27 kg/ha. Un affinage de la valeur du rendement global est possible à la fin du mois de juillet en intégrant la prévision du rendement des États-Unis aux 4 rendements cités pour la deuxième estimation. En effet, la récolte en France et en Russie n'ayant pas encore eu lieu, ce sont toujours les prévisions qui sont prises en compte pour ces 2 pays car les simulations n'existent pas encore. Le rendement alors estimé comment étant le rendement mondial du blé pour l'année 2021 sur base des 5 pays étudiés est de 3826,2 kg/ha. Le rendement global, une fois le blé récolté dans les 5 pays, autrement dit à la fin de la deuxième décade du mois d'août, a également été déterminé pour pouvoir faire une comparaison. Le rendement global obtenu en utilisant les valeurs de rendement simulé pour les 5 pays est de 3825,8 kg. A mesure que l'année 2021 avance, la valeur du rendement global est affinée et s'approche de plus en plus de la valeur fournie par la FAO qui est de 3491,9 kg/ha. La dernière estimation du rendement global utilisant des prévisions ayant une valeur de 3826,6

kg/ha, il y a donc une surestimation de 334,7 kg/ha. Le modèle de régression linéaire simple prévoit également un rendement plus élevé que celui réellement enregistré, néanmoins, la surestimation de ce modèle est moins conséquente puisqu'elle est de 115,87 kg/ha. Afin d'estimer la RMSE du modèle global intégrant les modèles des 5 pays étudiés, tout comme pour le rendement global, une pondération par la superficie cultivée par du blé en 2021 a été appliquée aux valeurs des RMSE des modèles nationaux ayant la RMSE la plus faible (régression multiple pour la Chine et Random Forest pour les autres pays). Ainsi la RMSE du modèle global reprenant les modèles des 5 pays vaut 108,03 kg/ha (European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

5. Discussion

La discussion de ce travail va être scindée en 3 parties. La première sera relative aux résultats finaux qui viennent d'être présentés. Elle permettra de répondre à la question de recherche de ce travail. La seconde partie traitera des résultats intermédiaires. Autrement dit, elle se focalisera sur les modèles par pays qui ont permis de conduire aux résultats finaux. Enfin, la troisième partie discutera de la qualité des données utilisées comme entrées dans les différents modèles exploités.

5.1 Discussion des résultats globaux

Avant d'analyser les résultats obtenus plus en profondeur, il est pertinent de rappeler que l'objectif majeur de cette recherche est d'évaluer différents types de modélisation du rendement afin de savoir lesquels sont les meilleurs à utiliser pour pouvoir faire des prévisions du rendement du blé mondial.

Pour évaluer la capacité de modèles complexes, intégrant des données agrométéorologiques, à simuler et à prévoir le rendement, un modèle de régression linéaire simple a dans un premier temps été établi, afin de servir de référence. Ce modèle de base s'est avéré être très performant comme en atteste la valeur élevée de son coefficient de détermination (0,9251) et la faible valeur de la racine de l'erreur quadratique moyenne (74,31 kg/ha). Ces résultats montrent qu'à malgré la simplicité du modèle, il reflète très bien l'évolution des rendements mondiaux, ce qui laisse déjà penser que dépasser ces performances à l'aide de modèles plus complexes risquerait d'être difficile.

Comme présenté dans la comparaison des résultats, les modèles complexes ne permettent pas d'améliorer la précision de la prévision des rendements du modèle de référence. Pour rappel, seul le modèle des forêts aléatoires de l'Inde possède une RMSE inférieure à celle du modèle de référence (70,296 kg/ha pour le modèle de l'Inde et 74,31 kg/ha pour le modèle de base). La construction du modèle mondial intégrant les résultats des modèles complexes a permis de montrer, à travers l'illustration de la situation en 2021, qu'au plus les modèles de nouveaux pays sont ajoutés, au plus la prévision du rendement pour l'année en cours s'affine. Cependant, cette prévision reste moins bonne que celle du modèle de référence. Le rendement prévu pour l'année 2021 obtenu à la fin du mois de juillet avec les modèles complexes dépasse de 334,7 kg/ha le rendement réel recensé par la FAO tandis que celui du modèle de référence ne le surestime que de 115,87 kg/ha. L'estimation de la RMSE du modèle intégrant les 5 pays étudiés est de 108,03 kg/ha. Cette valeur est plus élevée que celle du modèle de base (74,31 kg/ha). Néanmoins, elle est moins importante que celle obtenue dans l'étude de Jeong *et al.*, qui est de 320 kg/ha. Comme expliqué précédemment, cette étude avait également recouru aux forêts aléatoires, mais utilisait un autre découpage spatial et une sélection de variables différentes. Il est important de souligner que le modèle global complexe intègre seulement les modèles de 5 pays. Dès lors, l'hypothèse peut être émise que l'ajout de modèles pour d'autres pays permettrait d'améliorer davantage la précision du modèle global complexe jusqu'à éventuellement dépasser la précision du modèle de référence. Cette hypothèse est d'autant plus soutenue par le fait que les valeurs de R^2 des meilleurs modèles de 2 des 5 pays étudiés sont plus élevées que celle du modèle de référence et qu'un des 5 pays possède un R^2 équivalent à celui du modèle de base. En effet, le coefficient de détermination de référence

vaut 0,9251, celui de la moyenne des 3 meilleurs modèles de régression multiple en Chine vaut 0,97361, celui du modèle Random Forest en Inde vaut 0,95404 et celui du modèle Random Forest en Russie vaut 0,92478 (pratiquement égal au R^2 de la régression linéaire simple). Il serait donc intéressant, dans le cadre d'une recherche future d'intégrer les modèles d'autres pays. Il pourrait par exemple être pertinent d'étudier les 17 plus gros pays producteurs de blé, sur la période 2007-2021, présentés dans la section méthodologie puisqu'ils représentent à eux seuls plus de 80 % de la production mondiale de blé.

5.2 Discussion des résultats intermédiaires

Cette section a pour but de discuter des résultats obtenus lors de la modélisation à l'échelle nationale. Ces modèles nationaux constituent la base du modèle global complexe qui a été analysé ci-dessus. Dès lors, seuls les modèles retenus pour la modélisation à l'échelle globale vont être analysés. Chaque modèle est spécialement adapté aux particularités de son pays, ce qui permet de représenter au mieux les conditions dans lequel le blé évolue. L'analyse des résultats à l'échelle nationale permet de mieux comprendre la contribution et l'impact de chacun des pays dans le modèle global.

Contrairement aux autres pays et aux résultats de l'étude de Jeong *et al.*, la précision du modèle de régression linéaire multiple en Chine, ou plutôt, de la moyenne des 3 meilleurs modèles de régression multiple, est meilleure que celle du modèle basé sur les forêts aléatoires avec une RMSE de 108,745 kg/ha contre 187,825 kg/ha. Cette exception s'explique vraisemblablement par l'excellente performance du modèle de régression linéaire multiple. Avec un excellent coefficient de détermination de 0,97361, le meilleur R^2 de tous les modèles créés dans ce travail, le modèle explique pratiquement l'entièreté de la variance du rendement. En Chine, l'évolution du rendement au fil des années suit très bien la droite de régression linéaire comme montré à la Figure 11. Cette stabilité dans les données est sans doute ce qui a permis au modèle de régression linéaire multiple de mieux performer. La modélisation par Random Forest étant surtout performante pour modéliser des relations complexes et non linéaires est un argument supplémentaire pour soutenir cette explication.

Les résultats de la modélisation avec Random Forest en Inde sont excellents. Parmi tous les modèles construits, c'est le plus précis avec une RMSE de 70,296 kg/ha. Le coefficient de détermination du modèle est également très bon avec une valeur de 0,95404. Ces résultats montrent que le modèle est robuste et parvient très bien à identifier les relations entre les variables indépendantes et le rendement.

En s'intéressant à l'importance des variables dans la modélisation avec Random Forest pour le territoire russe, une différence par rapport aux autres pays apparaît. Contrairement aux autres pays, l'importance de la variable la plus influente dans les données d'entraînement du modèle pour la Russie, le prix dans ce cas-ci, n'atteint que 32,15 % au lieu des 100 % observés dans les 4 autres pays étudiés. Dès lors, il n'y a pas une variable qui prévaut sur toutes les autres dans ce modèle. Cette particularité pourrait signifier que plusieurs variables ont des importances similaires dans le modèle. De plus, il est également possible que la complexité des interactions entre les données de la Russie fasse en sorte que les données des variables influent beaucoup les unes sur les autres. De par la difficulté d'interprétation des résultats de la modélisation avec les forêts aléatoires, les hypothèses qui viennent d'être citées ne peuvent être affirmées.

avec certitude. Cette problématique met en évidence une limite importante de la modélisation Random Forest : bien que ce type de modèle soit très performant pour rendre compte de relations complexes entre les variables, comme son fonctionnement s'apparente à une boîte noire, il est compliqué d'interpréter de façon détaillée les résultats obtenus. Malgré cela, les résultats du modèle russe restent largement acceptables avec une RMSE de 121,833 kg/ha et un R^2 de 0,92478.

Les résultats obtenus par la modélisation basée sur l'algorithme Random Forest aux Etats-Unis ne sont pas aussi excellents qu'en Inde, mais ils restent très bons. La racine carrée de l'erreur quadratique moyenne est inférieure à 100 kg/ha puisqu'elle est de 87,467 kg/ha, ce qui indique une très bonne précision du modèle. Le modèle s'ajuste plutôt bien aux variables puisqu'il permet d'expliquer 90,068 % de la variance du rendement.

En France, la modélisation par régression linéaire multiple ne fournissant pas des résultats concluants, c'est donc la modélisation avec Random Forest qui a été choisie par défaut comme étant le type de modélisation le plus adapté à ce pays. En effet, pour cause d'une p-value beaucoup trop élevée (0,9803) de la régression linéaire basée sur les valeurs de rendement, le modèle de régression multiple n'a pas pu être appliqué pour la France. L'absence d'une relation linéaire indique une forte complexité dans les facteurs influençant le rendement. Les forêts aléatoires ayant justement les capacités de modéliser des relations complexes entre les variables, des résultats d'aussi bonne qualité que les précédents modèles étaient attendus. Pourtant, le modèle français s'avère être moins qualitatif que les autres. En effet, il possède la RMSE la plus élevée de tous les modèles nationaux (312,95 kg/ha) et le R^2 le plus faible (0,83856). Non seulement ces indicateurs de la qualité du modèle sont les moins bons, mais en plus de cela ils montrent que le modèle est significativement moins performant comparé aux autres. Les RMSE des autres modèles nationaux oscillent entre 70 et 122 kg/ha alors qu'en France la valeur est de l'ordre du double. Pour ce qui est du coefficient de détermination, au minimum 90% de la variance du rendement est expliquée par les différents modèles nationaux, sauf en France où le R^2 est de seulement 0,83856. Ces moins bons résultats pourraient être dû à une plus grande variabilité au sein du jeu de données. Mais, l'hypothèse la plus plausible est la non prise en compte de facteurs ayant une influence non négligeable sur le rendement du blé en France. L'ajout d'autres variables au modèle pourrait permettre une amélioration des performances de ce dernier. Comme le montre la Figure 37, le rendement du blé français en 2016 a été catastrophique. D'après l'article d'European Scientist, paru en 2018, le rendement du blé enregistré en France en 2016 d'à peine plus de 5000 kg/ha, pour des rendements normalement compris entre 6000 et 8000 kg/ha, n'avait plus été aussi bas depuis 1983. Cela serait dû aux conditions climatiques extrêmes de la saison du blé cette année-là. Les températures ont été anormalement chaudes durant l'automne 2015 et les précipitations anormalement élevées durant le printemps 2016 (European Scientist, 2018). Ainsi, pour améliorer le modèle dans le cadre de futures recherches, il serait pertinent d'intégrer de nouvelles variables relatives au stress thermique ainsi qu'au stress hydrique. De plus, la saison du blé en France considérée dans le cadre de cette étude n'intégrant pas l'automne, puisque la saison a été définie comme débutant 1 mois avant le début de la phase végétative, il pourrait également être intéressant d'envisager de travailler sur des périodes plus longues.

Comme mentionné dans la partie comparaison des résultats, à l'inverse des modèles de régression linéaire multiple pour lesquels les variables explicatives du rendement varient d'un modèle à l'autre, les modèles avec les forêts aléatoires montrent une certaine récurrence dans les variables ayant les importances les plus grandes dans les modèles. Dans chaque modèle Random Forest, les variables principales contiennent toujours le prix et une combinaison de 2 variables relatives au NDVI (parmi le NDVI maximal, le NDVI cumulé et le NDVI z-score cumulé). Ces variables jouent donc un rôle déterminant dans la modélisation et la prévision du rendement. Dans le cadre de futures recherches, il pourrait être intéressant de chercher à comprendre pourquoi les modèles suggèrent que ces variables en particulier sont importantes.

5.3 Discussion de la qualité des données

Après avoir mis en évidence que l'intégration d'autres variables que celles retenues serait intéressante afin d'améliorer la qualité des résultats de modélisation, cette section va quant à elle s'intéresser aux données qui ont été utilisées dans le cadre de ce travail. Bien que les données soient issues de sources reconnues telles que la FOA ou encore la Commission européenne, il faut tout de même considérer la présence d'éventuelles limitations et biais.

Les données de la FAO, disponibles à l'échelle nationale et internationale, sont collectées au moyen de questionnaires annuels relatifs à la production agricole envoyés aux différents pays. Généralement, les données collectées sont fournies par des instituts nationaux de statistique. La qualité des réponses au questionnaire peut varier d'un pays à l'autre en fonction de ses compétences en statistique, en collecte de données, ... Si un pays ne fournit pas les informations demandées, ses dernières seront alors complétées au moyen de sources officielles, comme par exemple Eurostat, et de sources non officielles, comme par exemple les rapports de l'USDA. En cas de données manquantes malgré ces ressources supplémentaires, une estimation sera alors faite par des experts. A titre d'exemple, la FOA n'a pas obtenu les données de rendement en Chine pour les années 2020 et 2021 et a donc proposé des estimations pour compléter le jeu de données (I. Kovrova, comm. pers., 2023).

Les données agrométéorologiques d'ASAP rendues disponibles par la Commission européenne présentent un avantage considérable : elles permettent l'utilisation d'un jeu de données homogènes. Grâce à cette base de données, il a été possible de télécharger l'ensemble des données agrométéorologiques utilisées dans ce travail pour les différents pays étudiés au départ d'une unique source. Cette uniformité des données permet donc de comparer les résultats des pays analysés sans risquer d'inclure des biais dus à l'utilisation de sources multiples. En revanche, les données ASAP ne sont pas propres au blé, mais aux cultures en général. Certaines variables, telles que la température ou les précipitations sont indépendantes de la culture. En revanche, les variables relatives au NDVI sont directement liées au type de culture. Par conséquent, les données relatives à l'état de santé des cultures présentent un biais puisqu'elles ne sont pas spécifiques au blé, mais sont relatives aux cultures en général. Il existe un autre problème relatif à ces données qu'il faut souligner. Elles ne sont pas tout à fait disponibles en temps réel, mais avec environ une décade de décalage par rapport au temps présent. Pour appliquer les méthodes de modélisation présentées dans ce

travail au temps présent, cela poserait dès lors problème. Pour le solutionner, il faudrait soit trouver un moyen d'accéder aux données en temps réel, soit faire des prévisions sur une plus longue période que deux décades avant la récolte en partant du principe qu'une des décades considérée comme inconnue, avec des données à prédire, a en réalité déjà eu lieu. Cette problématique pourrait éventuellement être étudiée dans le cadre de futurs travaux afin de développer des solutions permettant une mise en application concrète des outils présentés dans ce travail en temps réel. Comme expliqué dans la méthodologie de ce travail, les données ASAP sont disponibles à l'échelle d'entités administratives des différents pays. Pour obtenir des données à l'échelle nationale, des moyennes pondérées par la superficie cultivée par du blé ont été réalisées. Ces moyennes sont nécessaires pour pouvoir travailler à l'échelle souhaitée, mais ont pour effet de lisser les données et induisent donc une perte de précision. La précision des données pour la Russie à l'échelle nationale est de moins bonne qualité que celle des autres pays. De fait, les données téléchargeables relatives à la Russie ne sont pas complètes. Selon les années et les entités administratives, les données des différentes variables sont manquantes pour certaines décades. Dès lors, les moyennes pondérées ont été réalisées uniquement sur base des données disponibles. Une fois les variables obtenues à l'échelle nationale, un cumul de certaines variables a été réalisé sur ce qui a été défini comme la saison du blé d'un pays. Comme mentionné dans la section méthodologie, la durée de la saison a été définie de telle sorte à englober le début de la saison jusqu'au moment où la récolte a eu lieu partout dans le pays de sorte à ne pas perdre d'information. Il faut donc souligner que cette approche implique de prendre en compte des données parfois sur une période plus longue que nécessaire sur des zones bien spécifiques du territoire national.

Comme déjà brièvement abordé dans la partie méthodologie, il aurait été intéressant de comparer les données de prix utilisées pour la modélisation à d'autres ressources. Il existe des rapports détaillés sur le marché du blé. Par exemple, le « wheat market report » fourni des nombreuses informations relatives au marché mondial du blé mais il coûte plusieurs centaines voire milliers de dollars (EMR,2024). Il n'a donc pas été possible de faire des comparaisons par manque de disponibilité de données détaillées accessibles gratuitement. De plus, la date à laquelle le prix devait être enregistré en fonction des pays n'est pas forcément disponible pour toutes les années prises en compte dans le jeu de données de MacroTrends. Lorsque cette date n'était pas disponible, le prix du blé enregistré pour le modèle était alors celui de la date la proche, généralement le jour avant ou après la date de la fin de décade synonyme de fin de saison. Sachant que le moment de la récolte déterminé reste une approximation puisqu'il a été déterminé à l'échelle nationale, le prix retenu est également une approximation. En effet, au vu de l'étendue des territoires étudiés, le blé n'est pas récolté exactement au même moment partout sur le territoire. Cela implique donc nécessairement de faire des approximations pour retenir une valeur du prix à l'échelle nationale. Il est important de noter que le prix du blé est une variable complexe qui ne dépend pas simplement des lois de l'offre et de la demande. Cette complexité dépend de nombreux facteurs. La gestion des stocks de blé pourrait permettre une certaine régulation dans les prix. Cependant, seule la Chine possède un stock suffisant (plus de la moitié du stock mondial) que pour pouvoir avoir un réel impact. Or, les prix du blé étant plus élevés sur le marché chinois que sur le marché mondial, les stocks ne sont pas nécessairement rendus disponibles sur le marché mondial. Cette

différence de prix s'explique par l'octroi de subventions du gouvernement chinois à ses agriculteurs ainsi qu'à la maintenance de prix plancher. L'Inde subventionne aussi les producteurs de blé modifiant donc artificiellement les lois de l'offre et de la demande que ce soit sur le marché national ou international. En effet, ce type de politique gouvernementale peut causer une diminution de l'offre disponible pour l'export ou encore causer un excédant à l'échelle nationale. Le prix est également influencé par les variations des prix des engrais tels que l'urée (un engrais azoté). Le prix des carburants a également une influence conséquente sur le prix du blé, et ce, de différentes manières. La première est une influence directe des prix des carburants sur les coûts de production. De fait, le fonctionnement des machines agricoles est directement lié au prix du pétrole. Ensuite, la demande en biocarburant a un impact, qui, cette fois, est indirect. Les biocarburants tels que l'éthanol sont créés à partir de cultures. La culture généralement utilisée est le maïs. Lorsque la demande en biocarburant augmente, par exemple en raison de nouvelles politiques énergétiques, d'avantages de terres vont être consacrées à la culture du maïs au détriment du blé. Le prix mondial du blé dépend également des crises économiques. Par exemple, suite à la crise économique de 2008-2009, les prix du blé ont été extrêmement hauts. Les crises politiques peuvent également faire fluctuer les prix sur le marché. Cela s'est notamment produit lors de l'épisode du Printemps arabe de 2010 à 2012 (Enghiad *et al.*, 2017 ; Erenstein *et al.*, 2022). Plus récemment, la guerre entre la Russie et l'Ukraine, qui a débuté en février 2022, a eu un impact immédiat sur le marché mondial du blé. En effet, le blé est une des cultures les plus importantes en Ukraine et le pays est un producteur important sur le marché mondial. L'invasion de la Russie a provoqué une diminution de la quantité de blé disponible sur le marché international ce qui a entraîné une hausse des prix d'environ 2 % sur le marché mondial. La guerre a mis à mal le secteur du blé ukrainien de différentes façons. Tout d'abord en abimant les cultures avec les bombardements, mais aussi en détruisant des barrages qui permettaient de réguler l'apport d'eau aux cultures ou encore en endommageant les ports depuis lesquels l'exportation de la céréale avait lieu (Devadoss & Ridley, 2024).

6. Conclusion

L'objectif principal de ce mémoire était de réaliser une évaluation de la modélisation du rendement du blé à l'échelle mondiale en utilisant des données agrométéorologiques obtenues grâce à la télédétection. En analysant et en comparant les résultats d'un modèle simple avec un modèle complexe intégrant des données agrométéorologiques, le but de ce travail était de vérifier dans quelle mesure ces données permettent ou non d'améliorer la prévision des rendements à l'échelle globale. Les résultats de ces recherches pourraient alors être exploités comme base pour construire un outil offrant un accès libre aux prévisions, pouvant notamment servir à des prises de décisions pour les politiques afin de gérer au mieux les crises alimentaires ou encore d'éviter la spéculation sur les marchés mondiaux.

Les résultats ont permis de montrer que, pour les variables prises en compte et pour les pays étudiés, la précision du modèle de régression linéaire simple avec une racine carrée de l'erreur quadratique moyenne de 74,31 kg/ha surpasse celle du modèle complexe intégrant les modèles de plusieurs pays (modèle de régression linéaire multiple pour la Chine, modèle de forêts aléatoires pour l'Inde, la Russie, les États-Unis et la France) dont la RMSE s'élève à 108,03 kg/ha.

Toutefois, il faut souligner qu'il existe des limites à cette étude qui ont une influence sur les résultats obtenus. La limitation la plus importante étant le nombre de modèles nationaux créés. Seuls les 5 plus gros pays producteurs sur la période 2007-2021 ont été étudiés. Les résultats obtenus ne doivent dès lors pas être vus comme une fin en soi, mais plutôt comme les prémisses d'une recherche plus globale car l'approche utilisée dans ce travail reste prometteuse et nécessite des recherches plus approfondies. De fait, une recherche postérieure intégrant d'avantage de modèles nationaux pourrait très probablement permettre une amélioration de la précision du modèle globale de ce travail. Une autre limite de ce travail est le choix des variables utilisées. Il serait intéressant d'étudier l'impact d'autres variables dans la modélisation telles que des indicateurs sur le stress hydrique, sur le stress thermique, sur les nutriments présents dans le sol, sur les prix des engrais, sur les prix du pétrole ou encore sur les conflits, qu'ils soient d'ordre politique, économique ou social. Afin d'améliorer les recherches, il pourrait également être utile de réaliser une analyse plus approfondie des résidus des modèles. Enfin, une autre piste intéressante à explorer serait l'utilisation d'autres modèles basés sur le « machine learning » tels que les réseaux neuronaux.

Pour conclure, ce travail établit les bases de la prévision des rendements mondiaux à l'aide de données agrométéorologiques issues de la télédétection qui pourront être utilisées pour des futurs travaux et contribuer à la création d'outils décisionnels plus robustes.

7. Bibliographie

APSIM. (2023). *General Use Licence Summary*. APSIM. <https://www.apsim.info/download-apsim/general-use-licence-summary/> Consulté le 29 novembre 2023.

Basso, B. & Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. In D. L. Sparks (ed., *Advances in Agronomy*, 154. 1^{ère} éd. Academic Press Cambridge, MA : Elsevier, 201–255. <https://doi.org/10.1016/bs.agron.2018.11.002>

Blakely, L., Reno, M. J. & Broderick, R. J. (2018). *Evaluation and Comparison of Machine Learning Techniques for Rapid QSTS Simulations (SAND2018-8018)*. Sandia National Laboratories. https://www.researchgate.net/publication/326560291_Evaluation_and_Comparison_of_Machine_Learning_Techniques_for_Rapid_QSTS_Simulations Consulté le 12 juillet 2024.

Bobbit, Z. (2021). *RMSE vs. R-Squared: Which Metric Should You Use?* Statology. <https://www.statology.org/rmse-vs-r-squared/> Consulté le 2 juillet 2024.

Boogaard, H., Van Der Wijngaart, R., Van Kraalingen, D., Meroni, M. & Rembold, F. (2019). ASAP Water Satisfaction Index. *Publications Office of the European Union*. <https://doi.org/10.2760/478822>

Devadoss, S., & Ridley, W. (2024). Impacts of the Russian invasion of Ukraine on the global wheat market. *World Development*, 173, 106396. <https://doi.org/10.1016/j.worlddev.2023.106396>

Enghiad, A., Ufer, D., Countryman, A. M., & Thilmany, D. D. (2017). An Overview of Global Wheat Market Fundamentals in an Era of Climate Concerns. *International Journal of Agronomy*, 2017(1), 1–15. <https://doi.org/10.1155/2017/3931897>

Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., & Braun, H.-J. (2022). Global Trends in Wheat Production, Consumption and Trade. In Reynolds, M. P. & Braun, H.-J. (eds), *Wheat Improvement*. 1^{ère} éd. Cham : Springer, 47–66. https://doi.org/10.1007/978-3-030-90673-3_4

European Centre for Medium-Range Weather Forecasts. (n.d.). *About us*. ECMWF. <https://www.ecmwf.int/en/about> Consulté le 11 mai 2024.

European Commission. (2024). *ASAP – Anomaly Hotspots of Agricultural Production*. European Commission. <https://agricultural-production-hotspots.ec.europa.eu/download.php> Consulté le 25 avril 2024.

European Scientist. (2018). *Comment expliquer la diminution de rendement des cultures de blé en France en 2016 ?* European Scientist. <https://www.europeanscientist.com/fr/agriculture-fr/comment-expliquer-la-diminution-de-rendement-des-cultures-de-ble-en-france-en-2016/> Consulté le 5 août 2024.

Expert Market Research. (2024). *Global Wheat Market Size, Share, Growth, Trends, Forecast: By Type: Whole/Raw, Flour, Others; By Application: Feed, Food, Biofuel, Others; Regional Analysis; Market Dynamics: SWOT Analysis: Porter's Five Forces Analysis, Key Indicators for Demand, Key Indicators for Price; Value Chain Analysis; Price Analysis; Competitive Landscape; 2024-2032*. GMR. [Wheat Market Size, Share, Industry Trend & Report | 2032 \(expertmarketresearch.com\)](https://www.expertmarketresearch.com/wheat-market-size-share-industry-trend-report-2032) Consulté le 2 août 2024.

Food and Agriculture Organization of the United Nations (FAO). (2022). *GAEZ Data Portal*. FAO. [Country-Specific Data | GAEZ v4 Data Portal \(fao.org\)](https://www.fao.org/gaez/data-portal/) Consulté le 25 avril 2024.

Food and Agriculture Organization of the United Nations (FAO). (2024). *GIEWS – Global Information and Early Warning System on Food and Agriculture*. FAO. <https://www.fao.org/giews/en/> Consulté le 4 mars 2024.

Food and Agriculture Organization of the United Nations (FAO). (2023). *FAOSTAT Crops and livestock products*. FAO. <https://www.fao.org/faostat/en/#data/QCL> Consulté le 15 novembre 2023.

Franch, B., Cintas, J., Becker-Reshef, I., Sanchez-Torres, M. J., Roger, J., Skakun, S., Sobrino, J. A., Van Tricht, K., Degerickx, J., Gilliams, S., Koetz, B., Szantoi, Z., & Whitcraft, A. (2022). Global crop calendars of maize and wheat in the framework of the WorldCereal project. *GIScience & Remote Sensing*, 59(1), 885–913. <https://doi.org/10.1080/15481603.2022.2079273>

Franch, B., Cintas, J., Becker-Reshef, I., Sanchez-Torres, M. J., Roger, J., Skakun, S., Sobrino, J. A., Van Tricht, K., Degerickx, J., Gilliams, S., Koetz, B., Szantoi, Z., & Whitcraft, A. (2022). Annexes à *Global crop calendars of maize and wheat in the framework of the WorldCereal project*. GitHub. https://github.com/ucg-uv/research_products/tree/main/cropcalendars Consulté le 17 mars 2024.

Franch, B., Vermote, E., Skakun, S., Santamaria-Artigas, A., Kalecinski, N., Roger, J.-C., Becker-Reshef, I., Barker, B., Justice, C., & Sobrino, J. A. (2021). The ARYA crop yield forecasting algorithm: Application to the main wheat exporting countries. *International Journal of Applied Earth Observation and Geoinformation*, 104, 102552. <https://doi.org/10.1016/j.jag.2021.102552>

Genuer, R. (2010). *Forêts aléatoires : aspects théoriques, sélection de variables et applications* [Thèse de doctorat, mathématiques, Université Paris Sud - Paris XI]. HAL Archives Ouvertes. <https://theses.hal.science/tel-00550989>

Hao, S., Ryu, D., Western, A., Perry, E., Bogen, H., & Franssen, H. J. H. (2021). Performance of a wheat yield prediction model and factors influencing the performance: A review and meta-analysis. *Agricultural Systems*, 194, 103278. <https://doi.org/10.1016/j.agsy.2021.103278>

Indiana Department of Natural Resources (DNR). (2024). *Explanation of Standard Precipitation Index (SPI)*. Indiana State Government. [DNR: Water: Explanation of Standard Precipitation Index \(SPI\)](#) Consulté le 10 mars 2024.

International Monetary Fund. (n.d.). *Primary Commodity Price System*. International Monetary Fund. <https://data.imf.org/?sk=471dddf8-d8a7-499a-81ba-5b332c01f8b9> Consulté le 3 avril 2024.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, 11(6), e0156571. <https://doi.org/10.1371/journal.pone.0156571>

Joint Research Center. (2024). *Monitoring Agricultural ResourceS (MARS)*. European Commission. https://joint-research-centre.ec.europa.eu/monitoring-agricultural-resources-mars_en Consulté le 25 avril 2024.

Lab for Digital Agriculture, RADl, CAS. (n.d.). *Cropwatch*. Cropwatch. <http://www.cropwatch.com.cn/htm/en/index.shtml> Consulté le 4 novembre 2023.

MacroTrends. (2024). *Wheat Prices – 40 Year Historical Chart*. MacroTrends. [Wheat Prices - 40 Year Historical Chart | MacroTrends](#) Consulté le 27 avril 2024.

NASA Earth Data. (n.d.). *Moderate-Resolution Imaging Spectroradiometer (MODIS)*. NASA Earth Data. <https://www.earthdata.nasa.gov/sensors/modis> Consulté le 10 mai 2024.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylaniadis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187, 103016. <https://doi.org/10.1016/j.agsy.2020.103016>

R Foundation. (n.d.). *What is R ?* R Foundation. <https://www.r-project.org/about.html> Consulté le 1 août 2024.

Rousson, V. (2013). *Statistiques appliquées aux sciences de la vie*. 1^{ère} éd. Paris : Springer, Statistiques et probabilités appliquées 318 p. <https://doi.org/10.1007/978-2-8178-0394-4>

Schiefer, H., & Schiefer, F. (2021). Correlation. In *Statistics for Engineers: An Introduction with Examples from Practice*, 1^{ère} éd. .Wiesbaden : Springer. 95–98. https://doi.org/10.1007/978-3-658-32397-4_6.

Sergieieva, K. (2023). *Indices De Végétation Pour L'Agriculture Numérique*. EOS Data Analytics. <https://eos.com/fr/blog/indices-de-vegetation/> Consulté le 29 juin 2024.

Touma, D., Martinez, C. & National Center for Atmospheric Research Staf. (2023). *CHIRPS: Climate Hazards InfraRed Precipitation with Station data*. NCAR. <https://climatedataguide.ucar.edu/climate-data/chirps-climate-hazards-infrared-precipitation-station-data> Consulté le 30 juin 2024.

8. Annexes

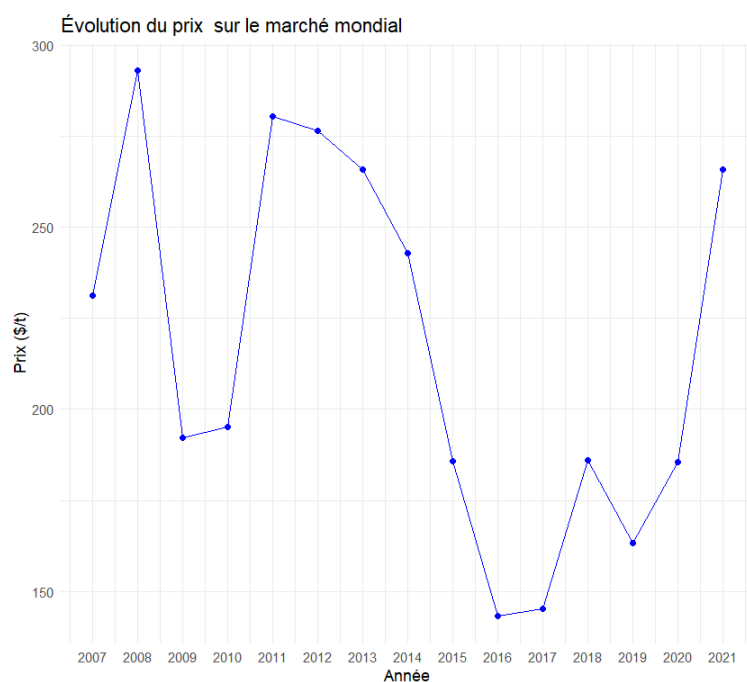
Annexe 1. Liste de la production de blé sur 15 ans dans 20 plus grands producteurs (Source de données : FAO, 2023).

Pays	Production totale sur 15 ans (t)
Chine	1875077340
Inde	1385220384
Russie	951818703
Etats-Unis	838091341
France	553271238
Canada	431443835
Pakistan	370004788
Allemagne	356240046
Ukraine	350998950
Australie	335896175
Turquie	300090000
Argentine	215787285
Royaume-Uni	215536498
Kazakhstan	210209390
Iran	171340723
Pologne	154155645
Egypte	128899820
Roumanie	112245717
Italie	108413933
Espagne	100363019

Annexe 2. Tableau du top 20 des plus grands pays producteurs de blé entre 2007 et 2021. (Source de données : FAO, 2023).

2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine	Chine
Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde	Inde
Etats-Unis	Etats-Unis	Russie	Etats-Unis	Russie	Etats-Unis	Etats-Unis	Russie	Russie	Russie	Russie	Russie	Russie	Russie	Russie
Russie	Russie	Etats-Unis	Russie	Etats-Unis	France	Russie	Etats-Unis	Etats-Unis	Etats-Unis	Etats-Unis	Etats-Unis	Etats-Unis	Etats-Unis	Etats-Unis
France	France	France	France	France	Russie	France	France	France	Canada	France	France	France	Canada	France
Pakistan	Canada	Canada	Allemagne	Australie	Australie	Canada	Canada	Canada	France	Australie	Canada	Canada	France	Ukraine
Allemagne	Allemagne	Allemagne	Pakistan	Canada	Canada	Allemagne	Allemagne	Allemagne	Ukraine	Canada	Pakistan	Ukraine	Pakistan	Australie
Canada	Ukraine	Pakistan	Canada	Pakistan	Pakistan	Pakistan	Pakistan	Ukraine	Pakistan	Pakistan	Ukraine	Pakistan	Ukraine	Pakistan
Turquie	Pakistan	Australie	Australie	Allemagne	Allemagne	Australie	Australie	Pakistan	Allemagne	Ukraine	Australie	Allemagne	Allemagne	Canada
Kazakhstan	Turquie	Ukraine	Turquie	Kazakhstan	Turquie	Ukraine	Ukraine	Australie	Australie	Allemagne	Allemagne	Argentine	Turquie	Allemagne
Iran	Royaume-Uni	Turquie	Ukraine	Ukraine	Ukraine	Turquie	Turquie	Turquie	Turquie	Turquie	Turquie	Turquie	Argentine	Turquie
Argentine	Argentine	Kazakhstan	Royaume-Uni	Turquie	Argentine	Kazakhstan	Royaume-Uni	Royaume-Uni	Kazakhstan	Argentine	Argentine	Australie	Australie	Argentine
Ukraine	Australie	Royaume-Uni	Iran	Argentine	Royaume-Uni	Royaume-Uni	Kazakhstan	Argentine	Iran	Royaume-Uni	Kazakhstan	Royaume-Uni	Kazakhstan	Royaume-Uni
Royaume-Uni	Kazakhstan	Iran	Kazakhstan	Royaume-Uni	Kazakhstan	Pologne	Pologne	Kazakhstan	Royaume-Uni	Kazakhstan	Royaume-Uni	Iran	Pologne	Pologne
Australie	Pologne	Pologne	Pologne	Pologne	Iran	Egypte	Iran	Iran	Argentine	Iran	Iran	Kazakhstan	Iran	Kazakhstan
Pologne	Italie	Egypte	Argentine	Iran	Egypte	Iran	Argentine	Pologne	Pologne	Pologne	Roumanie	Pologne	Royaume-Uni	Roumanie
Egypte	Egypte	Argentine	Egypte	Egypte	Pologne	Argentine	Egypte	Egypte	Egypte	Roumanie	Pologne	Roumanie	Egypte	Iran
Italie	Roumanie	Ouzbékistan	Italie	Roumanie	Italie	Espagne	Roumanie	Maroc	Roumanie	Egypte	Egypte	Egypte	Espagne	Egypte
Espagne	Iran	Italie	Ouzbékistan	Espagne	Ouzbékistan	Italie	Italie	Roumanie	Italie	Maroc	Espagne	Italie	Italie	Espagne
Ouzbékistan	Espagne	Maroc	Brésil	Italie	Roumanie	Roumanie	Ouzbékistan	Italie	Espagne	Italie	Maroc	Bulgarie	Roumanie	Brésil

Annexe 3. Evolution du prix du blé sur le marché mondial (Source de données : International Monetary Fund, *n.d.*).



Annexe 4. Évolution du rendement mondial du blé (Source de données : FAO, 2023).



Annexe 5. Formule de coefficient de corrélation de Pearson (Schiefer & Schiefer, 2021).

$$r = \frac{\sum_{i=1}^n (x_i - x_{moyen}) \cdot (y_i - y_{moyen})}{\sqrt{\sum_{i=1}^n (x_i - x_{moyen})^2 \cdot \sum_{i=1}^n (y_i - y_{moyen})^2}}$$

Avec x_i et y_i : les valeurs individuelles des 2 variables

x_{moyen} et y_{moyen} : les moyennes des 2 variables

n : le nombre d'observations

Annexe 6. Formules du R^2 et de la RMSE (Bobbit, 2021).

Formule du R^2 (coefficient de détermination) :

$$R^2 = 1 - \left(\frac{RSS}{TSS} \right)$$

Avec RSS : la somme du carré des résidus ($\sum_{i=1}^n (O_i - P_i)^2$)

TSS : la somme totale des carrés ($\sum_{i=1}^n (O_i - O_{i,moyen})^2$)

Formule de la RMSE (racine carrée de l'erreur quadratique moyenne) :

$$RMSE = \sqrt{\frac{\sum (P_i - O_i)^2}{n}}$$

Avec P_i : la valeur prédite à la $i^{\text{ème}}$ observation

O_i : la valeur observée à la $i^{\text{ème}}$ observation

n : le nombre d'observations

Annexe 7. (Rousson, 2013).

Equation générique d'une droite de régression linéaire :

$$y = \beta_0 + \beta_1 x$$

Avec y : la valeur prédite

β_0 et β_1 les coefficients qui minimisent la somme des carrés des résidus tels que :

$$\text{L'ordonnée à l'origine : } \beta_0 = y_{\text{moyen}} - \beta_1 \cdot x_{\text{moyen}}$$

$$\text{La pente : } \beta_1 = \frac{\sum (x_i - x_{\text{moyen}}) \cdot (y_i - y_{\text{moyen}})}{\sum (x_i - x_{\text{moyen}})^2}$$

Annexe 8. Entités administratives retenues par pays (Source de données : European Commission, 2024).

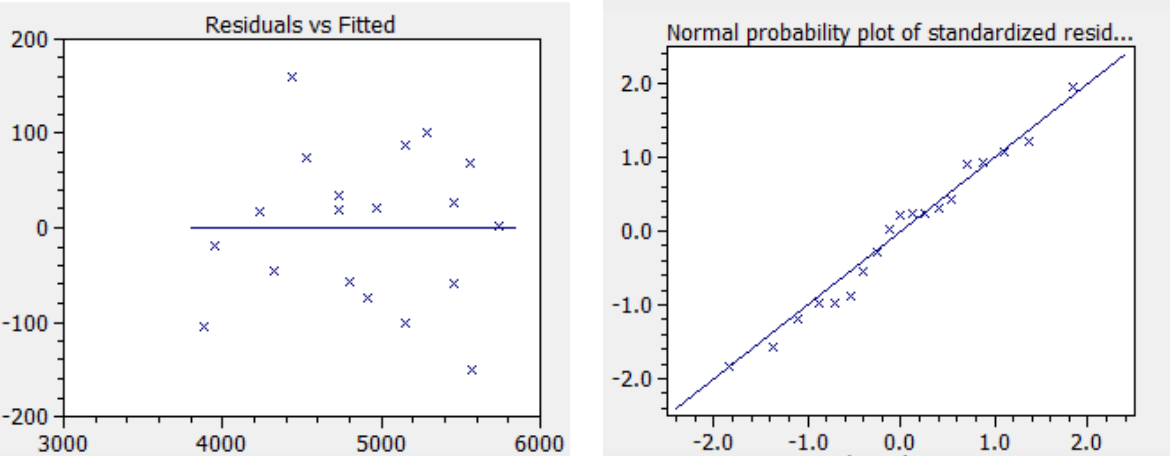
CHINE	INDE	RUSSIE	ETATS-UNIS	FRANCE
Anhui Sheng	Andaman	Adygeya Rep.	Alabama	Alsace
Beijing Shi	Arunachal P.	Altayskiy Kray	Arizona	Aquitaine
Chongqing Shi	Assam	Amurskaya	Arkansas	Auvergne
Fujan Sheng	Bihar	Bashkortostan	California	Basse-Normandie
Gansu Sheng	Chhattisgarh	Belgorodskaya	Colorado	Bourgogne
Guangdong Sheng	Dadra	Bryanskaya	Delaware	Bretagne
Guangxi Z.Z.	Daman and Diu	Buryatiya Rep.	Florida	Centre
Guizhou Sheng	Delhi	Chechnya Rep.	Georgia	Champagne-Ardenne
Hainan Sheng	Goa	Chelyabinskaya	Idaho	Franche-Comte
Hebei Sheng	Gujarat	Chitinskaya	Illinois	Haute-Normandie
Heilongjiang S.	Haryana	Chuvashiya Rep.	Indiana	Ile-de-France
Henan Sheng	Himachal Pradesh	Dagestan Rep.	Iowa	Languedoc-Rousillon
Hubei Sheng	Jharkhand	Ingushetiya Rep.	Kansas	Limousin

Hunan Sheng	Madhya Pradesh	Irkutskaya	Kentucky	Lorraine
Jiangsu Sheng	Maharashtra	Ivanovskaya	Louisiana	Midi-Pyrenees
Jiangxi Sheng	Manipur	Kabardino B.	Maine	Nord-Pas-de-Calais
Jilin Sheng	Meghalaya	Kaliningradskaya	Maryland	Pays-de-la-Loire
Liaoning Sheng	Mizoram	Kalmykiya Rep.	Michigan	Picardie
Nei Mongol Z.	Nagaland	Kaluzhskaya	Minnesota	Poitou-Charentes
Ningxia Huizu Z.	Orissa	Karatchayevo	Mississippi	Provence-Alpes-Cote-d'Azur
Shaanxi Sheng	Punjab	Kemerovskaya	Missouri	Rhone-Alpes
Shandong Sheng	Rajasthan	Khakasiya Rep.	Montana	
Shanghai Shi	Sikkim	Kirovskaya	Nebraska	
Shanxi Sheng	Tripura	Kostromskaya	Nevada	
Sichuan Sheng	Uttar Pradesh	Krasnodarskiy	New Jersey	
Tianjin Shi	Uttarakhand	Krasnoyarskiy	New Mexico	
Yunnan Sheng	West Bengal	Kurganskaya	New York	
Zhejiang Sheng		Kurskaya	North Carolina	
		Leningradskaya	North Dakota	
		Lipetskaya	Ohio	
		Mariy-el Rep.	Oklahoma	
		Mokovskaya	Oregon	
		Moskva	Pennsylvania	
		Name Unknown	South Carolina	
		Nizhegorodskaya	South Dakota	
		Novosibirskaya	Tennessee	
		Omskaya	Texas	

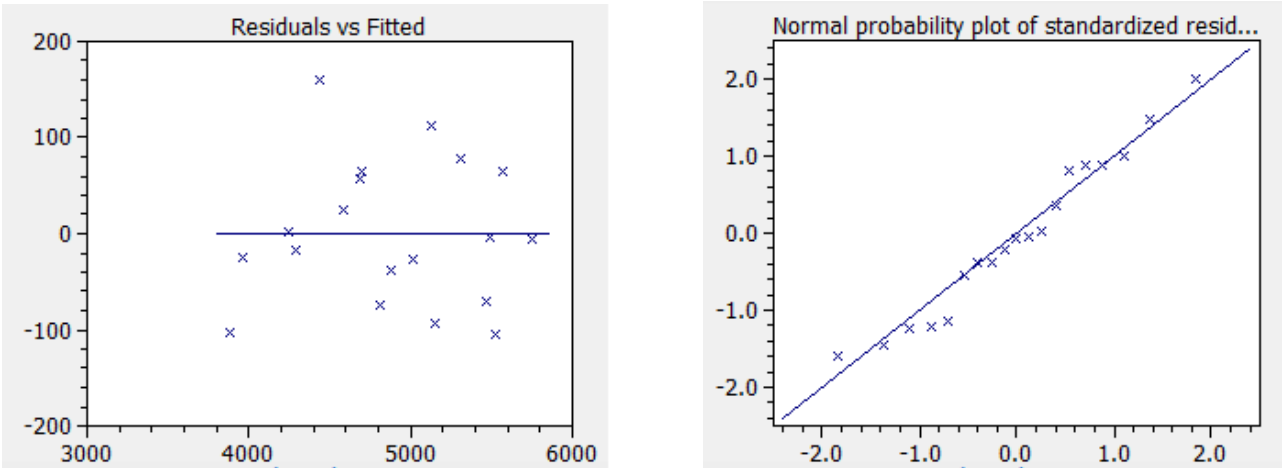
		Orenburgskaya	Utah	
		Orlovskaya	Vermont	
		Penzenskaya	Virginia	
		Permskaya	Washington	
		Primorskiy Kray	West Virginia	
		Pskovskaya	Wisconsin	
		Rostovskaya	Wyoming	
		Ryazanskaya		
		Samarskaya		
		Saratovskaya		
		Severnaya		
		Smolenskaya		
		Stavropolskiy		
		Sverdlovskaya		
		Tambovskaya		
		Tatarstan Rep.		
		Tomskaya		
		Tulskaya		
		Tverskaya		
		Tyumenskaya		
		Tyva Rep.		
		Udmurtiya Rep.		
		Ulyanovskaya		
		Ustordynskiy		
		Vladimirskaya		
		Volgogradskaya		
		Vologodskaya		
		Voronezhskaya		

		Yaroslavskaya		
		Yevreyskaya A.		

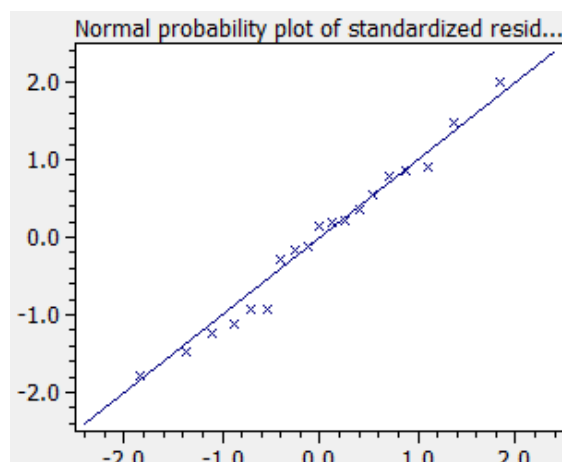
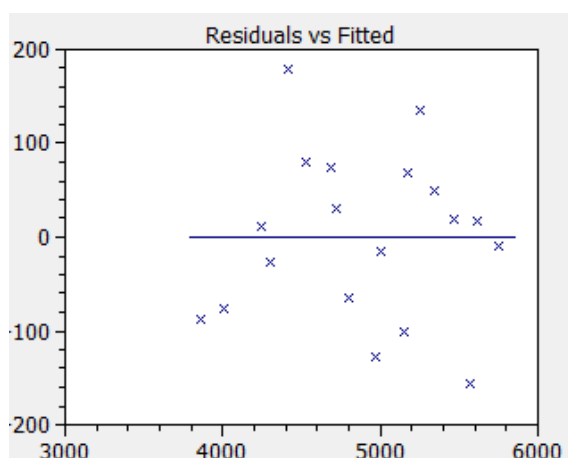
Annexe 9. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 1 en Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



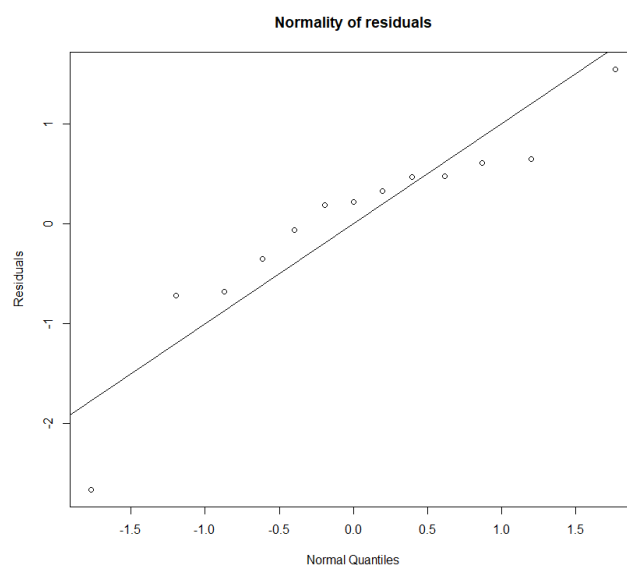
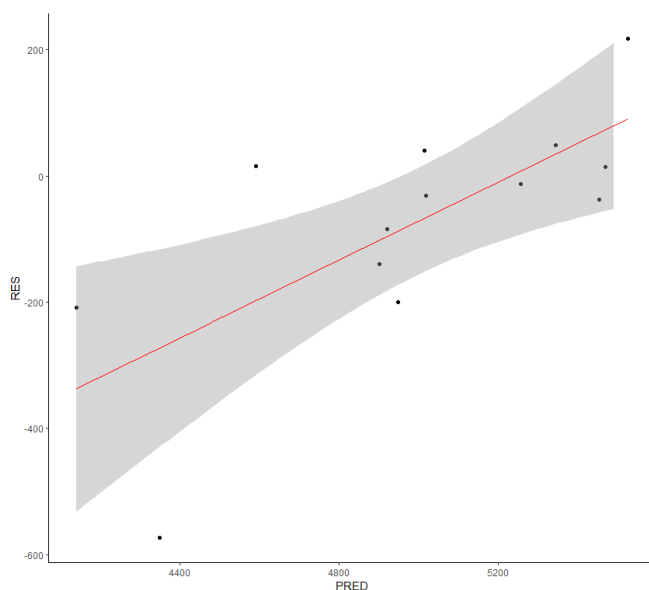
Annexe 10. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 2 en Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



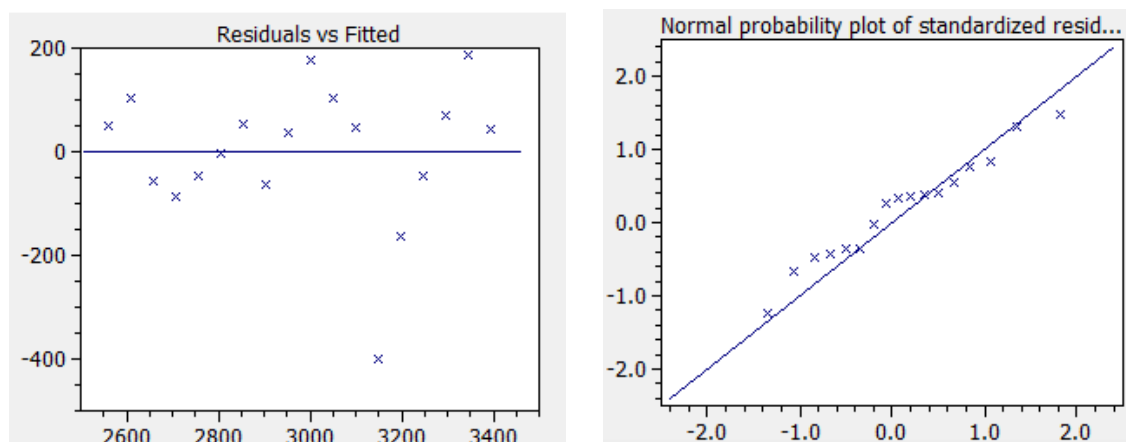
Annexe 11. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 3 en Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



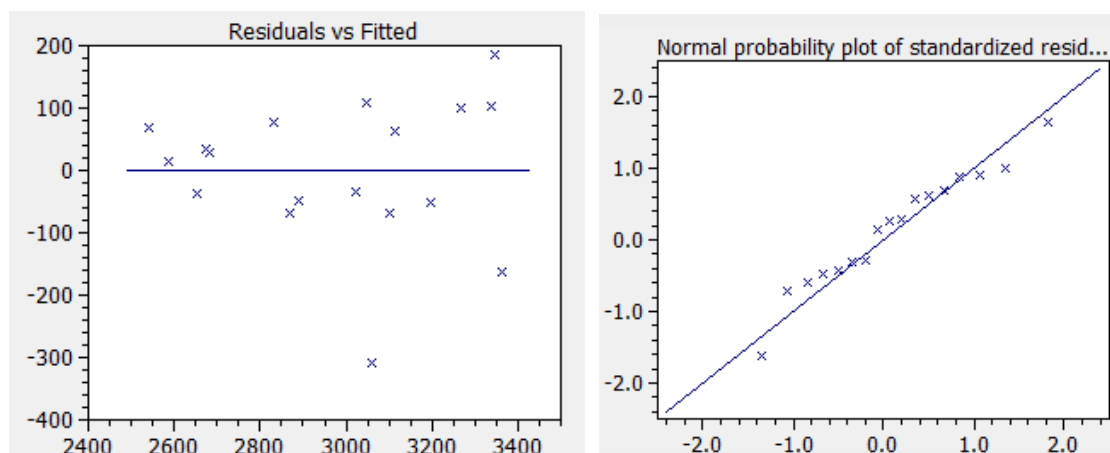
Annexe 12. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle avec les forêts aléatoires en Chine (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



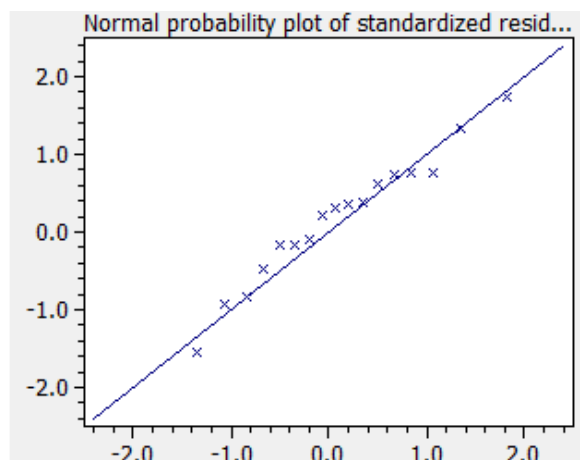
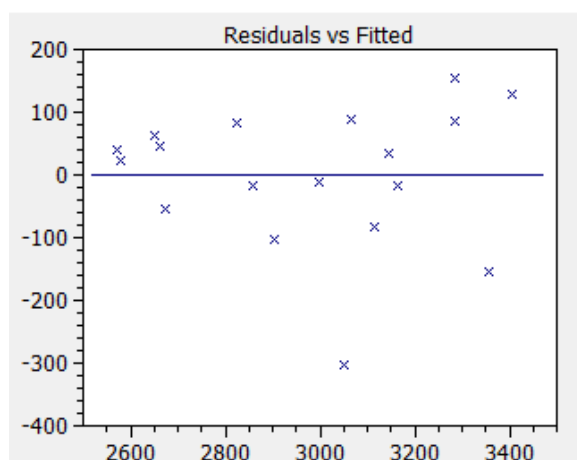
Annexe 13. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 1 en Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



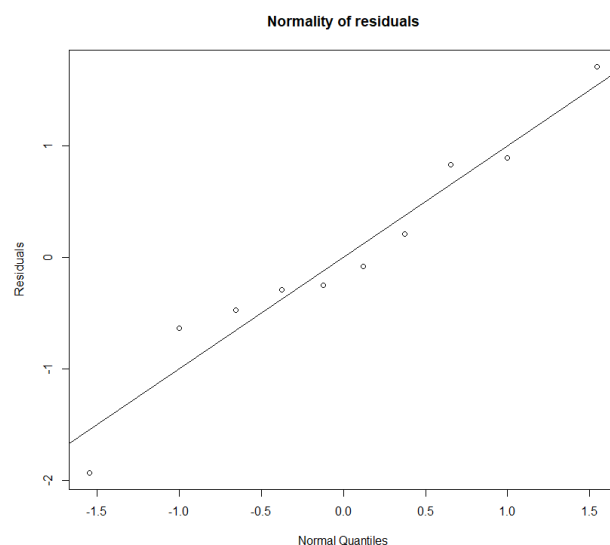
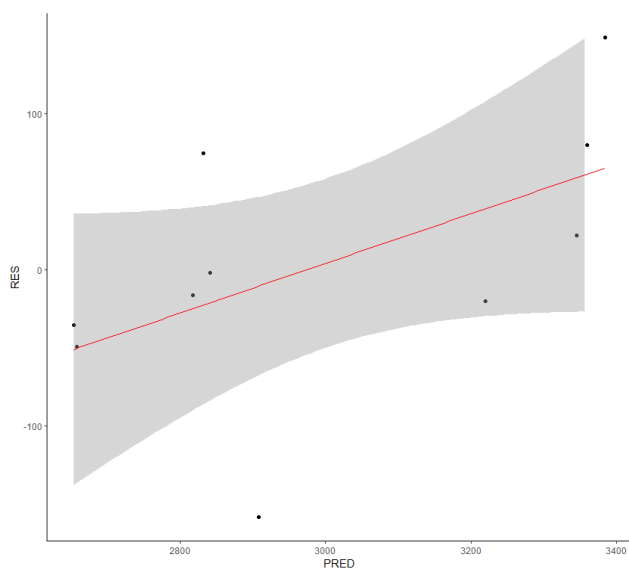
Annexe 14. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 2 en Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



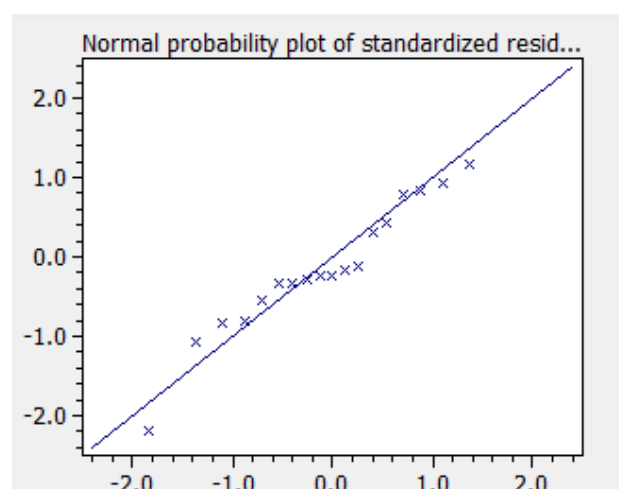
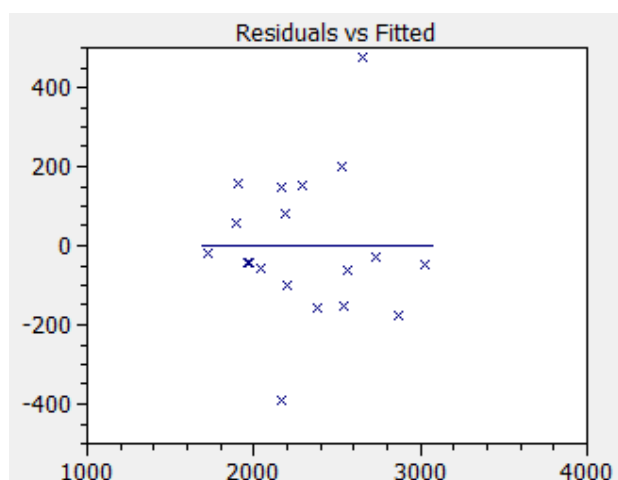
Annexe 15. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 3 en Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



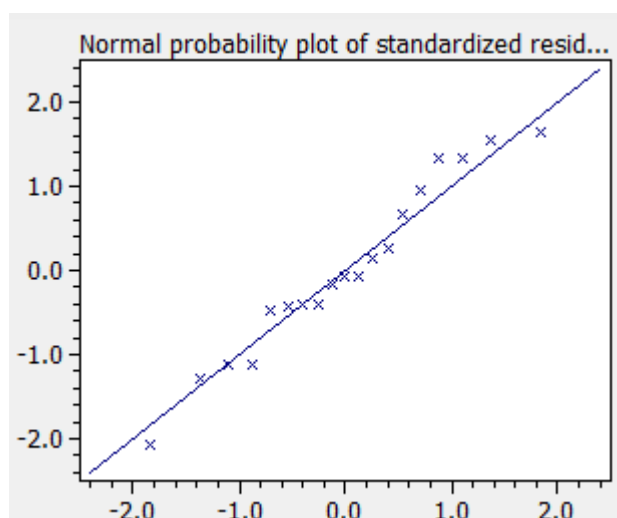
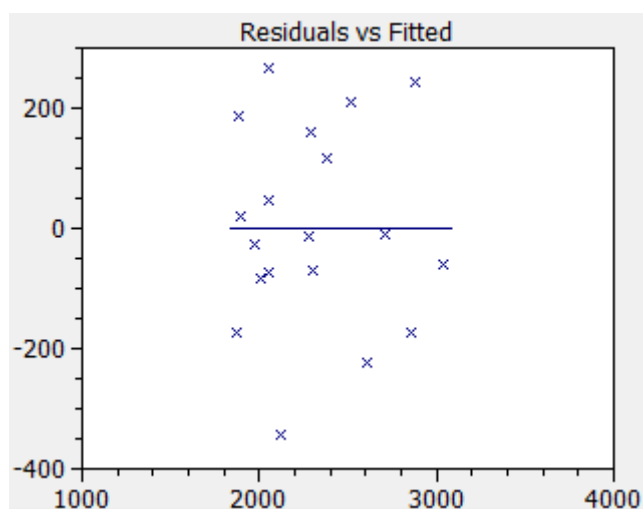
Annexe 16. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle avec les forêts aléatoires en Inde (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



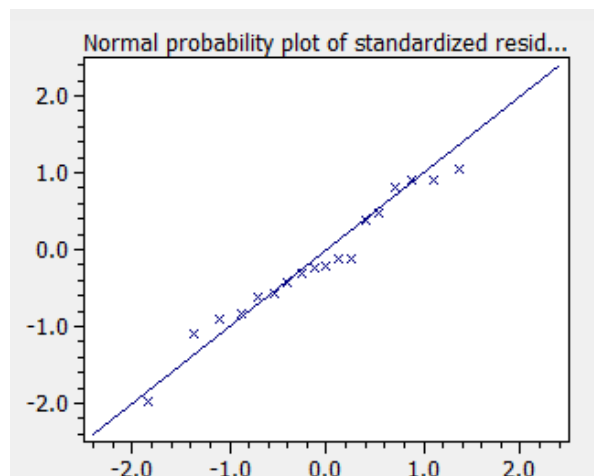
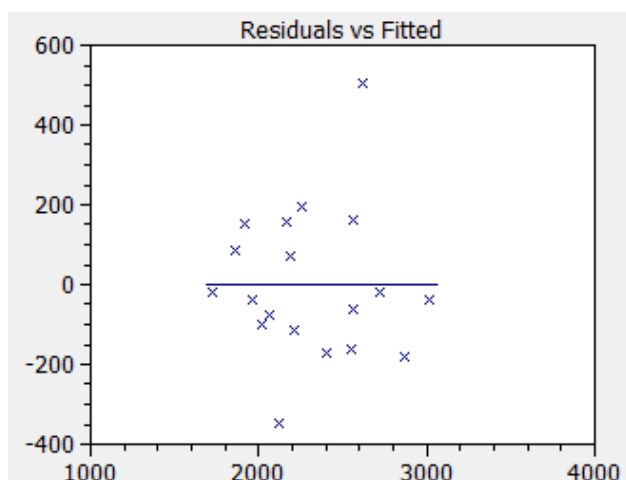
Annexe 17. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 1 en Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



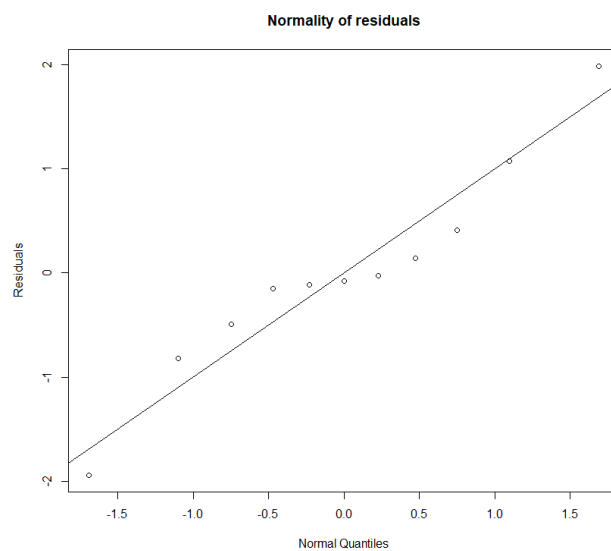
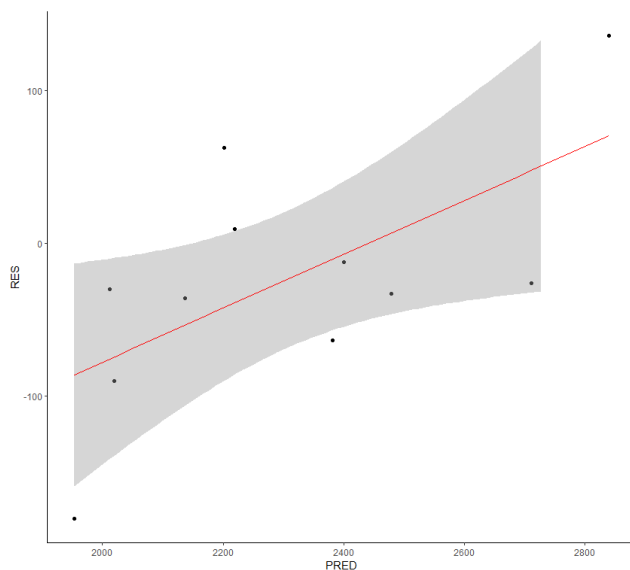
Annexe 18. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 2 en Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



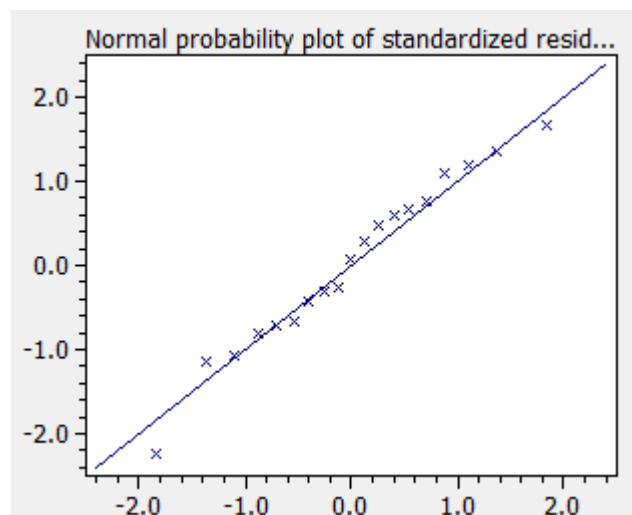
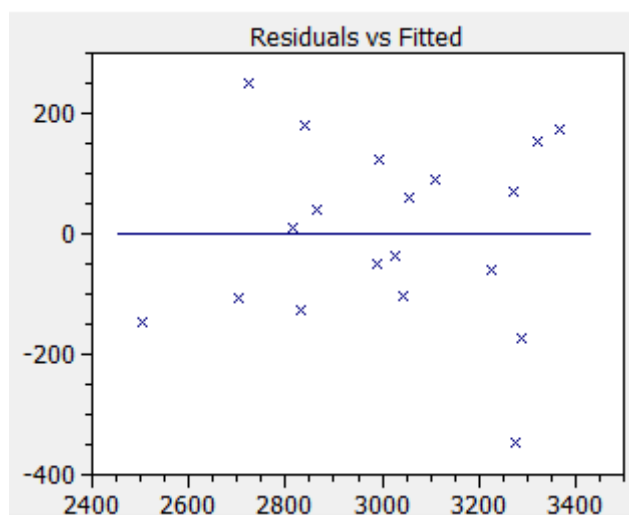
Annexe 19. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 3 en Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



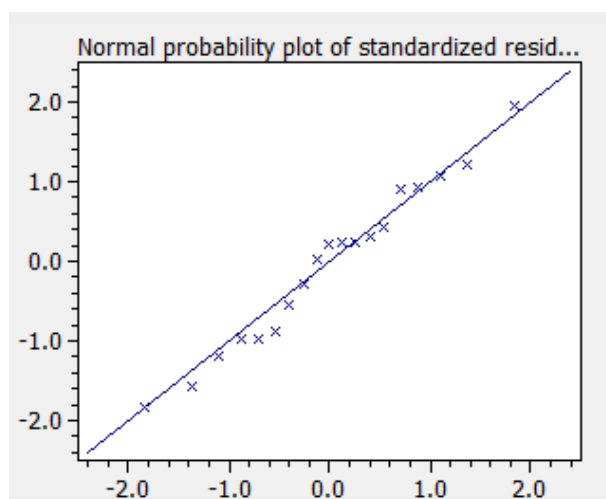
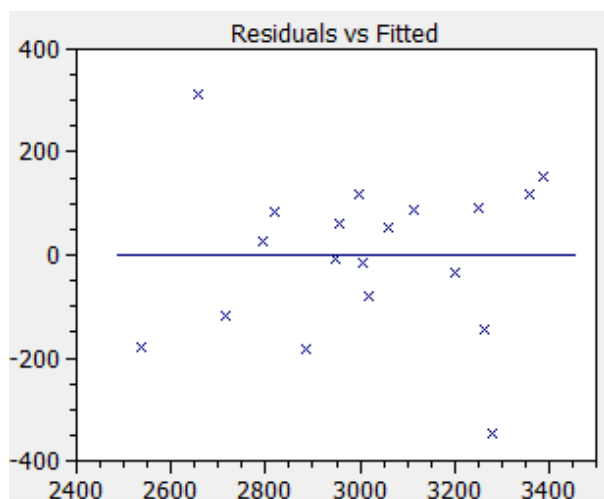
Annexe 20. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle avec les forêts aléatoires en Russie (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



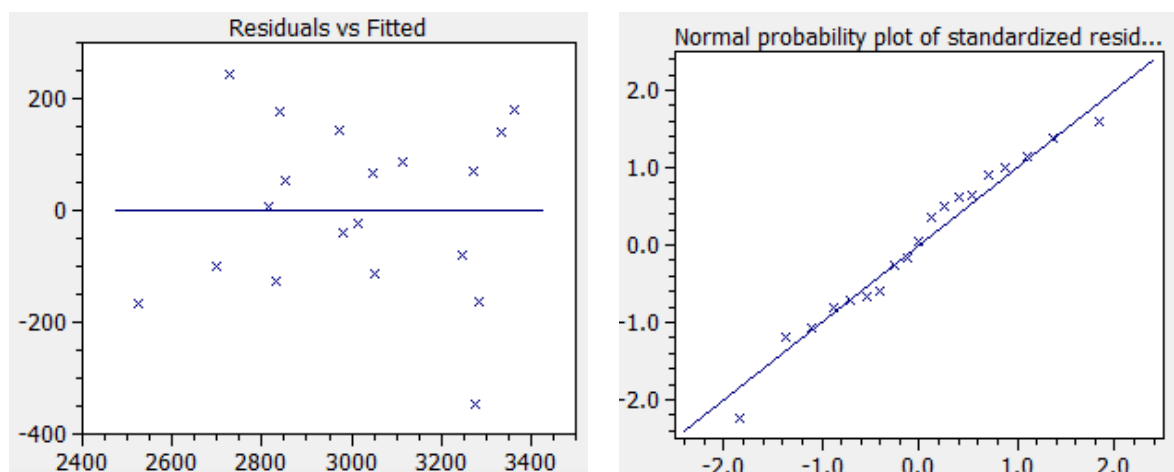
Annexe 21. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 1 aux États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



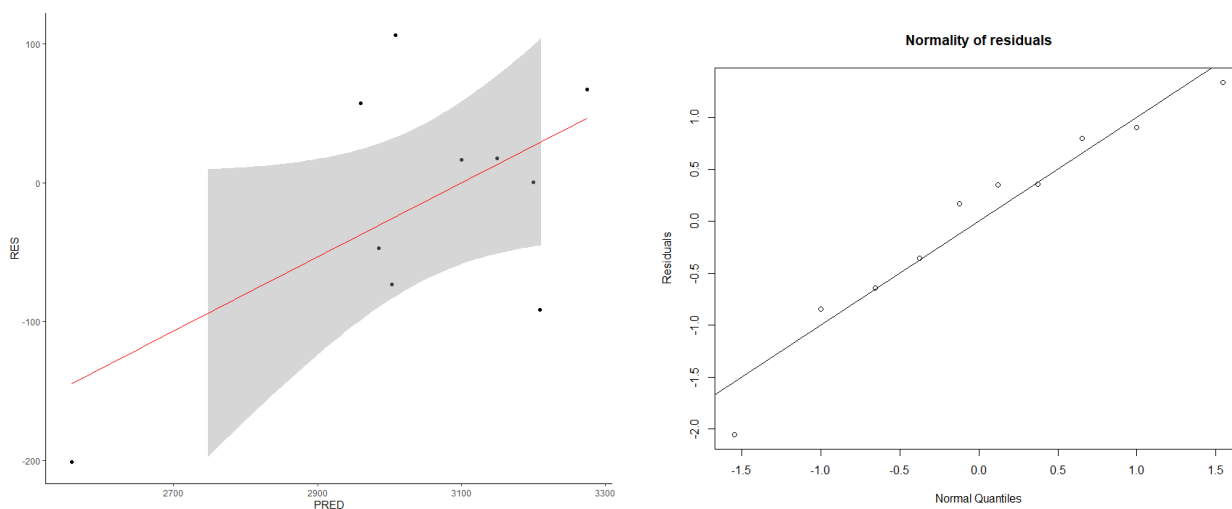
Annexe 22. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 2 aux États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



Annexe 23. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle 3 aux États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



Annexe 24. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle avec les forêts aléatoires aux États-Unis (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).



Annexe 25. Résidus (kg/ha) vs. Valeurs prédites (kg/ha) ; Probabilité normale des résidus standardisés pour le modèle avec les forêts aléatoires en France (sources de données : European Commission, 2024 ; FAO, 2023 ; MacroTrends, 2024).

