Rapport de stage – M1 AETPF

Modélisation de l'impact des ravageurs sur le rendement en grandes cultures en fonction des traitements

Encadrants : Corentin BARBU, chargé de recherches,
Clément CHEVALEYRE, ingénieur de recherches
- UMR AGRONOMIE

Adélaïde Subts

Stage réalisé du 3 juin au 16 août 2024

AgroParisTech - Université Paris-Saclay

M1 AETPF

Engagement de non-plagiat

Principes

- Le plagiat se définit comme l'action d'un individu qui présente comme sien ce qu'il a pris à autrui.
- Le plagiat de tout ou parties de documents existants constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée
- Le plagiat concerne entre autres : des phrases, une partie d'un document, des données, des tableaux, des graphiques, des images et illustrations.
- Le plagiat se situe plus particulièrement à deux niveaux : Ne pas citer la provenance du texte que l'on utilise, ce qui revient à le faire passer pour sien de manière passive ; recopier quasi intégralement un texte ou une partie de texte, sans véritable contribution personnelle, même si la source est citée.

Consignes

- Il est rappelé que la rédaction fait partie du travail de création d'un rapport ou d'un mémoire, en conséquence lorsque l'auteur s'appuie sur un document existant, il ne doit pas recopier les parties l'intéressant mais il doit les synthétiser, les rédiger à sa façon dans son propre texte.
- Vous devez systématiquement et correctement citer les sources des textes, parties de textes, images et autres informations reprises sur d'autres documents, trouvés sur quelque support que ce soit, papier ou numérique en particulier sur internet.
- Vous êtes autorisés à reprendre d'un autre document de très courts passages in extenso, mais à la stricte condition de les faire figurer entièrement entre guillemets et bien sûr d'en citer la source.

Sanction : En cas de manquement à ces consignes, le responsable du M1 AETPF se réserve le droit d'exiger la réécriture du document. Dans ce cas, la validation de l'Unité d'Enseignement ou du diplôme de fin d'études sera suspendue.

Engagement :

Je soussigné Adélaïde SUBTS

Reconnais avoir lu et m'engage à respecter les consignes de non-plagiat

À Vitry-sur-Seine, le 31/07/2024
Signature :

Cet engagement de non-plagiat doit être inséré en début de tous les rapports, dossiers et mémoires.

Remerciements

Je tiens à exprimer ma reconnaissance envers celles et ceux qui m’ont aidée au cours de ce stage de Master 1 et lors de la réalisation de ce rapport.

En premier lieu, je souhaite remercier Corentin Barbu, mon maître de stage à l’INRAE, pour son soutien constant, sa rigueur et son encadrement efficace et bienveillant. J’ai le sentiment grâce à lui d’avoir énormément progressé et d’avoir une meilleure compréhension du monde de la recherche. Sa patience face à mes capacités limitées en informatique n’a pas d’égale.

Je remercie également Clément Chevaleyre, mon deuxième encadrant, ingénieur de recherches pour le projet MoCoRiBA, grâce à qui R n’a plus (beaucoup) de secrets pour moi. Ses conseils et commentaires constructifs m’ont beaucoup aidée lors de l’écriture des modèles.

J’exprime également ma gratitude à Jean-François Castell, mon responsable universitaire, et à Erwan Personne, qui ont su tout au long de l’année mais spécialement au cours de ce stage m’accompagner et me soutenir dans les difficultés de parcours.

Merci à Sébastien Gervois, chargé de recherches à Terres Inovia, qui a accepté de relire et de corriger ce rapport au pied levé.

Merci à Michèle Fanucci-Malacrida, dont la sagacité et l’empathie n’ont pas de borne, et sans qui le chemin aurait été beaucoup plus difficile.

Enfin, toute ma reconnaissance aux membres de l’UMR Agronomie, pour leur accueil chaleureux et les discussions autour d’un bon thé.

Table des matières

Abstract

Understanding yield evolutions in crops, particularly the impact of bio aggressors, is paramount for supporting farmers, researchers, and policymakers. This is essential to anticipate the necessary changes to maintain productivity while reducing the use of phytosanitary products, in line with European objectives. This article presents a statistical model of yields based on pests and diseases and the Treatment Frequency Index (IFT) for herbicides, fungicides, and pesticides per plot, using data collected over 20 years across France on thirteen different crops. The combined use of a GAMSEL, Lasso and a Random Forest model, using R helps to mitigate the weaknesses of each approach to capture the non-linear effects of several variables on the target variable (yield), while maintaining a good predictive accuracy and model robustness. Cross-validation performed on multiple subsets (by year, by department or by random sampling) ensures a better generalization. The aim of this kind of model is to guide agricultural practices. Meteorological, soil, and epidemiological data account for most of the observed yield variations. The discrepancies between the theoretical yields produced by the model and the observed yields indicate areas for improvement to augment relevance and avoid overfitting. These differences also highlight the complexity of consistently measuring yields, given the many influencing factors. Notably, the interactions between crops and pests, and the resulting yield losses, are only partially captured by the model.

Résumé

Il est crucial de mieux comprendre l’évolution des rendements des grandes cultures, en particulier l’impact des bioagresseurs, pour aider les exploitants mais aussi les chercheurs et les décideurs à anticiper les changements nécessaires pour maintenir la productivité en réduisant l’utilisation de produits phytosanitaires selon les objectifs européens. Ici, nous développons une modélisation statistique des rendements en fonction des bioagresseurs et des indices de fréquence de traitement (IFT) herbicides, fongicides et insecticides par parcelle à partir de données sur les grandes cultures récoltées sur 13 ans et dans la France entière sur 13 cultures. L’usage combiné dans R d’un modèle GAMSEL, d’un modèle Lasso et d’un modèle Random Forest permet de limiter les faiblesses respectives de chaque approche en capturant les effets non linéaires de variables explicatives sur la variable cible (le rendement), tout en gardant une bonne précision de prédiction et une relative robustesse du modèle. La validation croisée, effectuée sur plusieurs sous-ensembles (par année, par département ou échantillonnage aléatoire) garantit une meilleure

généralisation. Ce type de modélisation peut permettre à terme d'orienter les pratiques des exploitants. Les données météorologiques, pédologiques et épidémiologiques expliquent la majeure partie des variations de rendement observé. Les différences entre les rendements théoriques produits par le modèle et les rendements observés montrent que le modèle peut être amélioré pour gagner en pertinence et éviter le surajustement, et témoignent de la difficulté de mesurer de manière cohérente des rendements dont la variabilité dépend de très nombreux facteurs. En particulier, les interactions entre les cultures et les ravageurs, ainsi que les pertes de rendement qu'ils causent, semble échapper en partie à la modélisation.

Introduction

L'Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) est un institut public de recherches. Issu de l'INRA (Institut National de Recherche Agronomique) et de l'IRSTEA (Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture), il se consacre à la recherche sur l'agriculture durable, la sécurité et la qualité de l'alimentation, la protection environnementale et l'impact du changement climatique. Il vise à développer et améliorer les techniques d'agriculture en France pour répondre aux défis de l'alimentation de demain, par exemple en déterminant les évolutions de rendement des cultures face à un changement des pratiques culturales.

En 2019, selon l'INSEE, la part des grandes cultures en France correspondait à 48% de la Surface agricole utilisée (SAU), qui couvre 52% de la métropole et 2% des DOM (Annexes 1 et 2). Cette répartition est inéquitable : 94% de la SAU en Ile-de-France est occupée par des grandes cultures, contre 2% en Corse. L'agriculture joue un rôle important dans les exportations françaises, lui assurant en 2019 une balance commerciale agricole excédentaire de 7,8 milliards d'euros. Cependant, ce rendement est soumis à des pressions grandissantes de bioagresseurs de plus en plus résistants aux phytosanitaires, dont l'utilisation est de plus fréquemment remise en cause en raison de leurs effets sur les sols, les eaux et la biodiversité (V. Langlois, 2019). L'impact des bioagresseurs sur les rendements reste difficile à évaluer (Devaud et Barbu, 2019) : la nocivité et la pression qu'ils exercent peut varier rapidement selon les régions ou les années, et l'augmentation des résistances complique la tâche aux exploitants. Il est en particulier difficile d'évaluer les pertes réelles de rendements provoquées par des maladies et ravageurs en co-occurrence.

Le projet MoCoRiBA-GC (Modélisation et Communication du Risque de BioAgresseurs en Grandes Cultures) est dirigé par l'INRAE depuis 2020 en lien avec différents partenaires (AWIUZ, Terres Inovia, ITB). Il visait initialement à étudier les possibilités, avancées par différentes études (Butault, 2010 ; Lechenet, 2017), de diminuer l'utilisation de produits phytosanitaires de 10 à 30% sans perte de marge pour les agriculteurs en enrichissant en temps réel l'information disponible pour les agriculteurs et les conseillers. Dans ce but, des modèles statistiques ont été produits pour mieux intégrer la pression des bioagresseurs sur le rendement dans le cadre des pratiques des agriculteurs. A cause de la difficulté d'obtention de données en temps réel, le projet s'est réorienté vers la production d'un

outil qui permette une réflexion des exploitants sur leurs pratiques par comparaison avec les exploitations du réseau DEPHY, engagé dans la réduction de l'utilisation des produits phytosanitaires grâce à des nouvelles pratiques et techniques culturales. Une version test de l'application est en développement.

Mon stage s'inscrit dans le projet MoCoRiBA sur la thématique de modélisation de l'impact des bioagresseurs sur le rendement en fonction des traitements phytosanitaires. L'objectif est d'améliorer la quantification de cet impact en appliquant dans R des modèles statistiques aux données de l'équipe. Dans un premier temps, il s'agit d'adapter et d'améliorer un modèle GAMSEL de rendement potentiel de 13 cultures parmi l'ensemble des cultures sur lesquelles le jeu de données est suffisamment conséquent. Par la suite, différents modèles (LASSO, Random Forest) sont testés et adaptés pour améliorer leur pertinence et la qualité des prédictions.

Données et méthode

Cette étude cherche à évaluer l'impact des bioagresseurs (maladies et ravageurs) sur le rendement des grandes cultures, en utilisant la modélisation statistique (forêts aléatoires, Lasso et Gamsel). Les données des du réseau d'épidémiosurveillance utilisées pour les BSV (Bulletin de Santé du Végétal), associées aux données de rendement utilisés, renforcent l'efficacité des modèles statistiques du fait de la taille de l'échantillon et des grandes échelles spatiales sur lesquelles il est distribué.

1. Bref état de l'art

La modélisation joue un rôle majeur dans l'analyse et la conception des systèmes de culture (Gonzalez-Sanchez *et al.*, 2014), en particulier dans l'estimation des rendements, auxquels participent de très nombreux facteurs environnementaux (climat, sols), facteurs économiques (marchés et filières) et agronomiques (irrigation, traitement, rotation culturale, travail du sol). Même si des estimateurs simples, comme la moyenne des rendements précédents, peuvent être utilisés, la variation des rendements n'est pas linéaire.

Les modèles mécanistes permettent de simuler directement le mécanisme responsable de la relation entre deux variables : ils sont basés sur la compréhension des interactions entre différentes parties d'un système, et sont construits sur des principes théoriques et des lois physiques ou biologiques. Ils permettent des prédictions détaillées et très précises. Si la plupart des modèles mécanistes de rendement sont

spécifiques à une culture, certains modèles comme STICS (Brisson et al., 1998) sont facilement adaptables grâce à l'ajustement des paramètres. Cependant, ces modèles sont coûteux et peu pratiques car gourmands en temps de développement et de calcul, ce qui les rend peu applicables à la planification agricole à grande échelle (Drummond *et al.,* 2003).

En revanche, la modélisation basée sur l'analyse des données permet de prédire ou de comprendre des relations entre variables, sans forcément connaître en profondeur les mécanismes sous-jacents. Cela leur apporte une grande flexibilité malgré leur forte dépendance à la qualité des données utilisées (Drummond *et al.,* 2003). L'un des risques majeurs associés à la modélisation statistique est le surajustement à des variables non déterminantes qui se trouvent par accident corrélées dans la prédiction à une caractéristique des données. La méthode de la validation croisée (cross validation) permet de comparer les prédictions avec les observations réalisées sur des données indépendantes, qui n'ont pas servi lors de l'ajustement du modèle. Si l'erreur observée est similaire à l'erreur des données d'ajustement, alors le modèle n'a pas réalisé de surajustement. Le *machine learning* offre des outils puissants pour la prédiction des rendements des cultures. Par exemple, parmi les modèles linéaires, un classifieur bayésien naïf peut être utilisé pour modéliser les rendements en utilisant des prédicteurs continus et discrets tels que la température, le CO2, le déficit de pression de vapeur et le rayonnement solaire (Qaddoum, 2014). Cette méthode classe les données en supposant les attributs indépendants les uns des autres.

Ont été utilisées également les méthodes de régression, telles que la régression linéaire multiple, les arbres de régression M5-Prime, les MLP (*multilayer perceptron*), qui utilisent plusieurs couches de neurones pour apprendre des représentations complexes de données et modéliser les relations non linéaires entre les variables d'entrée et les rendements, et la régression par Support Vector (SVR), qui minimise les erreurs de prédiction tout en maximisant la marge entre les données (González Sánchez, 2014).

Les modèles *random forest*, ou « forêt aléatoire », qui utilisent un ensemble d'arbres de décision pour améliorer la précision et combinent les résultats de plusieurs arbres, ont montré une grande précision (Priya, 2018), tandis que les SVM (*support vector machine*) ont également été efficaces, surtout pour gérer des jeux de données complexes (Bondre, 2019).

Les modèles additifs généralisés (GAM), qui permettent de modéliser les relations non linéaires entre des variables, combinés avec la régression LASSO, permettent de sélectionner les variables les plus pertinentes en simplifiant le modèle (Devaud et Barbu, 2019).

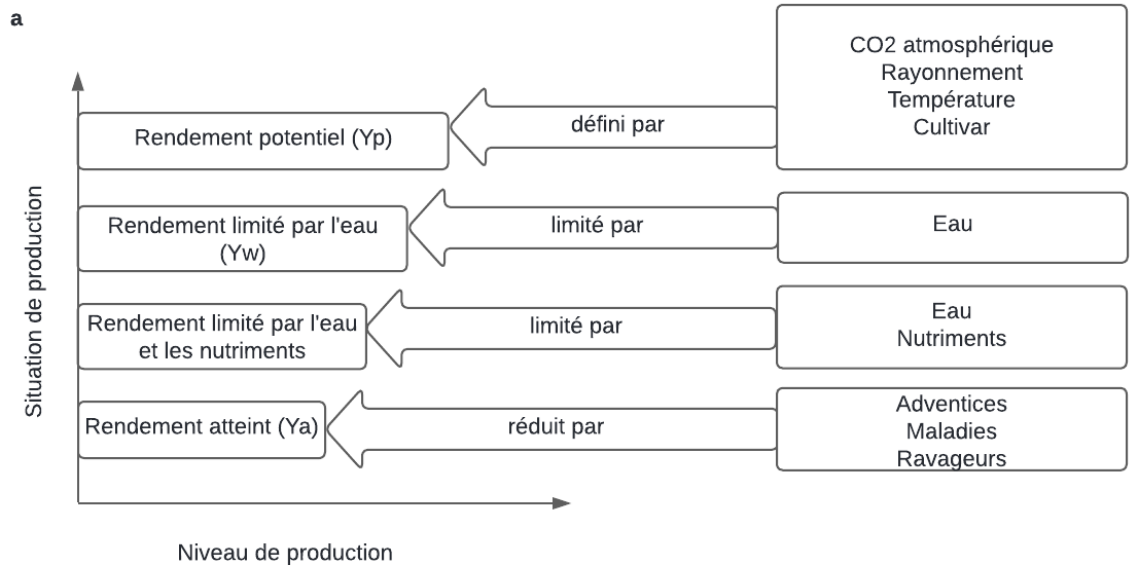
Dans le cadre de la réduction de l'utilisation des produits phytosanitaires, des expérimentations sur les systèmes de culture ont pu être menées pour mesurer la réduction du rendement atteignable. Selon une étude menée en France sur 946 fermes par Lechenet (2017), on pourrait réduire les herbicides de 37%, les fongicides de 47% et les insecticides de 60% sans avoir d'effets négatifs sur la production ou les revenus de l'exploitation, à condition d'adopter localement des itinéraires techniques semblables pour plusieurs exploitations et de privilégier l'utilisation de biopesticides, la rotation des cultures, et un travail du sol pertinent.

2. Concepts

Le rendement potentiel (Yp) est le rendement d'une culture lorsque les apports en eau et en nutriments sont non limitants et lorsque le stress biotique est efficacement contrôlé ((Evans, 1993). Lorsqu'elle est cultivée dans ces conditions, la production et la croissance sont déterminées uniquement par le rayonnement solaire, la température, le CO2 atmosphérique, l'interception de lumière par la canopée et les traits génétiques du cultivar. Le rendement potentiel, même s'il est théoriquement spécifique à un lieu précis (conditions climatiques et environnementales), ne dépend pas des autres propriétés du sol. Pour des cultures pluviales, on utilisera plutôt le rendement limité par l'eau (Yw), qui peut être également utile pour des cultures irriguées, et qui dépend aussi du type de sol (capacité de rétention d'eau, profondeur d'enracinement) et de la topographie de la parcelle (ruissellement). On calcule Yp et Yw pour des dates de semis recommandées, une densité de semis et un cultivar donné.

Le rendement moyen atteint (Ya) est le rendement effectivement obtenu au champ. Le contexte et les pratiques culturales sont des facteurs majeurs de la croissance : le rendement doit être maximisé pour l'ensemble du système cultural et non simplement pour les bénéfices d'une culture. Le Ya dépend effectivement des pratiques de gestion majoritaires dans une région donnée (date de semis, cultivar, densité de semis, gestion des nutriments, protection et traitement des cultures).

L’écart de rendement (Yg) est la différence entre le rendement potentiel (Yp ou Yw) et le rendement réel (Ya). Il est impossible pour la majorité des exploitants d’atteindre effectivement l’équilibre dans les pratiques culturales nécessaire pour atteindre le rendement potentiel, et il n’est pas généralement rentable de chercher à le faire à cause d’un effet plafond : la réponse du rendement aux intrants diminue lorsqu’on atteint certains niveaux d’IFT, indice de fréquence de traitement (Koning et al., 2008) ; l’efficacité de l’utilisation des ressources diminue également avec les facteurs de rendement, comme les températures élevées, les précipitations variables, les vents forts (risque de verse accru). Le changement climatique (température et disponibilité en eau) affecte directement et indirectement ces rendements par les adaptations qu’il



exige (date de semis, changements chez les ravageurs et maladies).

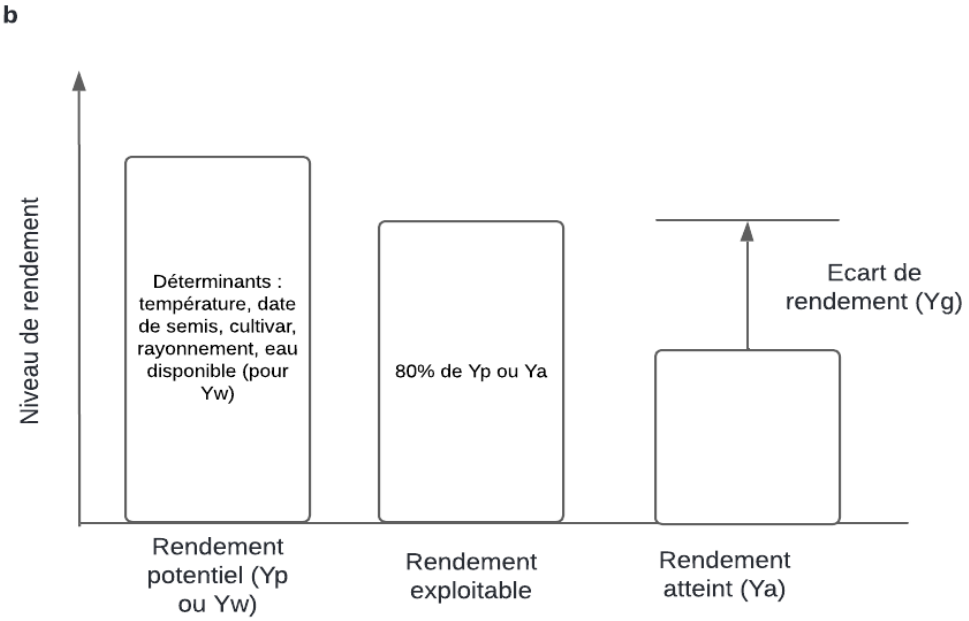


Figure 1 : Niveaux de production selon leurs facteurs de définition, limitants ou de réduction (a). L’écart de rendement (b) représente l’écart entre le rendement atteint et 80% du rendement potentiel (source : adapté de van Ittersum *et al.*, 2012).

3. Données

Les données utilisées au cours de cette étude proviennent de bases de données nationales, telles que le réseau DEPHY, du réseau d'épidémiosurveillance (Epiphyt et Vigicultures®), et de la base météorologique SAFRAN. Les régions sélectionnées sont représentatives des zones de culture majeures en France, permettant une bonne vue d'ensemble des conditions agricoles et environnementales.

Plusieurs filtres ont été appliqués pour garantir la qualité du jeu de données : les informations manquantes ont été supprimées, de même que les données incohérentes, comme un IFT fongicide supérieur à 30 ou des rendements de blé tendre d'hiver supérieurs à 200 $q.ha^{-1}$.

A. Les données Agrosyst

Elaboré pour le réseau DEPHY par l'INRAE dans le cadre du plan Ecophyto, le Système d'Information (SI) Agrosyst sert d'appui à la description et à l'évaluation des systèmes de culture (Ancelet *et al.,* 2015), en partenariat avec de nombreux partenaires et exploitants. L'un de ses objectifs revendiqués est la diminution de l'usage des produits phytosanitaires.

Dans le cadre de la préparation des données pour le projet MoCoRiBA-GC (Lay, 2020), sont notamment utilisées les données des itinéraires techniques (par exemple d'IFT, Indice de Fréquence de Traitement) des fermes du réseau DEPHY. Deux types d'informations ont été rassemblées :

- Réalisé : décrit les itinéraires techniques sur chaque parcelle pour chaque année de récolte (cultures, interventions, mesures et observations)
- Synthétisé : décrit les itinéraires relatifs à plusieurs parcelles au même stade de la rotation (même culture, même précédent cultural, même place dans la rotation), agrégées en supprimant la dimension spatiale de la parcelle. Il couvre une ou plusieurs années de récoltes.

Nous avons utilisé les données provenant de 13 cultures : betterave, blé tendre d'hiver, blé dur d'hiver, colza d'hiver, maïs ensilage, maïs grain, orge d'hiver, orge de printemps, pois d'hiver, pois de printemps, pomme de terre, tournesol et triticale. L'anonymisation a été faite préalablement à notre réception des données, en supprimant les noms des agriculteurs, des exploitations et des parcelles, mais le nom du département, de la commune et le numéro de ferme DEPHY ont été conservés. Nous n'avons accès qu'au code INSEE des exploitations.

Les données sur la réserve utile (RU) ont été obtenues en croisant les données Agrosyst avec les données de Gis Sol et en faisant la moyenne des données de chaque classe par commune.

B. Les données Safran

Le choix d'utiliser ces données climatiques de travaux antérieurs de l'équipe (Chevaleyre, 2023). Les données climatiques Safran sont les résultats d'un modèle mécaniste climatique dans lequel ont été assimilées les données d'observations collectées par Météo France depuis plusieurs décennies. Ces données sont disponibles à la résolution spatiale de 8km par 8km, et au pas de temps journalier. Ces données permettent ensuite d'établir des modèles de bioagresseurs sur une base climatique précis localement. Les variables extraites pour nos modèles (Lay, 2020) concernent les températures minimales, moyennes et maximales pour chaque année de récolte, l'évapotranspiration (ETP en mm), les précipitations (mm), le rayonnement. A partir de ces données les indicateurs suivants ont calculés : le nombre de jours de pluie, le nombre de jours où la température minimale était inférieure à -17°C, le nombre de jours où la température maximum était comprise entre 0 et 10°C, et le nombre de jours où la température maximale dépassait les 34°C. Chaque parcelle Agrosyst est associée aux mailles Safran les plus proches grâce au code INSEE de l'exploitation.

C. Les données d'épidémiosurveillance

Les données d'observations concernant les bioagresseurs (maladies et ravageurs) proviennent de la base de données Vigicultures®, qui centralise des données publiques et privées. Vigicultures® est administré par les structures représentatives de leurs cultures et filière (Arvalis, Terres Inovia, ITB, ASTREDHOR, IFV, CDAF, ACTA). Les données hebdomadaires ont été moyennées à l'année pour chaque parcelle (Arvalis, 2020), et traitées par interpolation et krigeage par des travaux antérieurs (Chevaleyre, 2023).

Culture	Type	Nom
Betterave	Maladie	Cercosporiose
		Rouille
	Ravageur	Puceron noir
		Puceron vert
		Pégomyie
		Teigne de la betterave
Blé dur d'hiver	Maladie	Fusariose
		Helminthosporiose
		Oïdium des céréales
		Piétin verse
		Rouille brune du blé
		Rouille jaune des céréales
		Septoriose des céréales
	Ravageur	Puceron
		Pucerons vecteurs de viroses
Blé tendre d'hiver	Maladie	Fusariose
		Helminthosporiose
		Oïdium des céréales
		Piétin verse
		Rouille brune du blé
		Rouille jaune des céréales
		Septoriose des céréales
	Ravageur	Puceron
		Pucerons vecteurs de viroses
Colza d'hiver	Maladie	Phoma
		Sclérotiniose
	Ravageur	Altise
		Altise Grosse d'hiver du Colza
		Altise petite des crucifères
		Charançon de la tige du chou
		Charançon de la tige du colza
		Charançon du bourgeon terminal
		Méligèthe du colza
		Puceron vert du pêcher
Maïs ensilage	Ravageur	Chrysomèle
		Foreurs
		Pyrales
		Sesamies
		Taupins
Maïs grain	Ravageur	Chrysomèle
		Foreurs
		Pyrales
		Sesamies
		Taupins
Orge d'hiver	Maladie	Helminthosporiose de l'orge
		Oïdium des céréales
		Rhynchosporiose
		Rouille jaune de l'orge
	Ravageur	Pucerons vecteurs de viroses
Orge de printemps	Maladie	Helminthosporiose de l'orge
		Oïdium des céréales
		Rhynchosporiose
	Ravageur	Pucerons vecteurs de viroses
Pois d'hiver	Maladie	Anthracnose
		Botrytis du pois
		Mildiou du pois
		Oïdium du pois
		Rouille du pois
	Ravageur	Puceron vert du pois
		Sitone du pois
		Tordeuse du pois
Pois de printemps	Maladie	Anthracnose
		Botrytis du pois
		Mildiou du pois
		Oïdium du pois
		Rouille du pois
	Ravageur	Puceron vert du pois
		Sitone du pois
		Tordeuse du pois
Pomme de terre	Maladie	Alternariose de la pomme de terre
		Mildiou de la pomme de terre
	Ravageur	Doryphores
		Puceron
Tournesol	Maladie	Phoma macdonaldi (Maladie des tâches noires)
		Phomopsis du tournesol
	Ravageur	Limace
		Puceron noir de la fève
		Puceron vert du prunier
Triticale	Maladie	Fusariose
		Oïdium des céréales
		Piétin verse
		Rouille brune du blé
		Rouille jaune des céréales
		Septoriose des céréales
	Ravageur	Pucerons vecteurs de viroses

Tableau 1 : Liste des ravageurs et maladies pris en compte dans les modèles pour les 13 cultures identifiées.

4. Prédiction du rendement potentiel

Nous avons implémenté un modèle de régression semi paramétrique (permettant de capturer des relations non linéaires dans les données sans

faire d'hypothèse sur leur forme) en utilisant le modèle additif généralisé (GAM). La relation entre la variable dépendante et chacune des variables explicatives est modélisée comme somme de fonctions lisses qui représentent chacune une partie non paramétrique de la relation :

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \epsilon$$

- β_0 est un paramètre intercept.
- $f_i(X_i)$ sont des fonctions lisses (non paramétriques) des variables explicatives X_i.
- ϵ est l'erreur aléatoire.

Ici, on utilise des splines pour modéliser les effets non linéaires des variables explicatives sur le rendement des cultures, en utilisant le package de R 'gamsel'. Pour chacune des 13 cultures d'intérêt, après avoir supprimé les valeurs manquantes, on extrait les variables climatiques et la RU, variables explicatives, et le « rendZone », rendement réel à prédire. Les données sont converties en matrices pour être compatibles avec les fonctions du package 'gamsel'. Pour l'ajustement des paramètres du modèle, on a fixé les degrés de liberté à 3 pour chaque variable explicative. Par validation croisée, on détermine le meilleur paramètre de régularisation (lambda.1se), qui garantit le meilleur équilibre entre le biais et la variance du modèle. Ce lambda contrôle la complexité du modèle pour éviter le surajustement en réduisant la variance. Plus sa valeur est élevée, plus le modèle est lissé, au risque de perdre en pertinence. Le lambda.1se est la valeur de lambda qui minimise l'erreur de validation moyenne, soit le plus grand lambda qui produise un modèle avec une erreur de validation proche du meilleur modèle, mais avec une meilleure régularisation. L'usage de lambda.1se permet d'obtenir un modèle plus robuste aux variations des données d'entraînement, ce qui permet une interprétation plus facile, quitte à sacrifier un petit peu de la performance optimale.

Grâce à la fonction « getActive » de 'gamsel', on sélectionne les variables explicatives actives, en prenant en compte à la fois les effets linéaires et non linéaires, puis les variables linéaires sont utilisées pour formuler un Gam sous forme de *spline*. On fait ensuite la prédiction ajustée aux données par culture.

L'évaluation des performances du modèle est faite à partir de l'erreur quadratique moyenne (RMSE), l'erreur moyenne (ME) et les coefficients de détermination (R^2).

5. L’impact des bioagresseurs sur le rendement

A. Le Modèle Random Forest

Pour analyser l’impact des ravageurs sur le rendement et la mitigation de cet impact par les traitements herbicides, fongicides et insecticides, un gros ensemble de données a été traité grâce au modèle Random Forest (RF), qui a permis d’en évaluer l’importance relative. Les packages R ‘dplyr’ et ‘randomForest’ ont été utilisés. Le modèle RF a été choisi pour sa robustesse de prédiction et sa capacité à traiter de grands ensembles de données en modélisant les relations complexes entre les caractéristiques des cultures et de l’environnement d’une part, et leur rendement d’autre part. Plusieurs arbres de décision (500) ont été combinés dans le modèle RF pour améliorer la précision des prédictions et réduire le surajustement. Chaque arbre est construit à partir d’un sous-ensemble aléatoire des données, et la prédiction finale repose sur l’agrégation des résultats issus de tous les arbres. Chaque *split* prend en compte un certain nombre de variables (par défaut, la racine carrée du nombre total de variables) pour éviter le surajustement.

Les données ont été préparées en sélectionnant les variables explicatives par la suppression de certains éléments du jeu de données pour alléger le modèle, comme les identifiants des parcelles. Elles ont ensuite été divisées selon chaque culture, pour laquelle un modèle Random Forest est ajusté. Lors de la construction de chaque arbre du modèle, à chaque nœud de décision, on sélectionne aléatoirement un certain nombre de caractéristiques parmi lesquelles on choisit le meilleur attribut pour diviser les données. La prédiction finale est obtenue par agrégation des prédictions et sélection de la classe majoritaire parmi tous les arbres. Une validation croisée interne est réalisée avec les données d’entraînement pour vérifier les performances du modèle, puis 4 validations croisées différentes ont été effectuées : par année, par département, par échantillonnage aléatoire et par année et département combinés. A chaque fois, les données ont été divisées en un ensemble d’entraînement, sur lequel ont été réalisées les prédictions, et un ensemble de test, pour évaluer les performances du modèle en termes de RMSE (erreur quadratique moyenne), erreur moyenne et R^2. Enfin, les résultats de tous les groupes de test sont agrégés pour fournir une évaluation globale de la performance du modèle, en utilisant la RMSE, l’erreur moyenne absolue, le pourcentage moyen d’erreur absolue et le R^2.

Pour analyser les résultats du modèle RF, on a extrait les 30 variables les plus influentes en se basant sur le pourcentage d’augmentation du MSE et on a sélectionné leur importance (fonction « importance » du package ‘randomForest’) dans la variabilité du modèle,

puis on a réalisé des graphiques pour visualiser cette importance et faciliter l'interprétation pour chaque culture.

Pour analyser les interactions, on a d'abord étudié les interactions dans le modèle entre l'indice de fréquence de traitement (IFT) fongicide et pesticide et les bioagresseurs correspondant à chaque type. Pour chaque culture, on a ajusté les données en fonction des niveaux de pression des maladies (de 0 à 1 par incrément de 0,2), puis on a prédit les rendements en faisant varier les IFT. Des graphes lissés et des boxplots ont été réalisés pour chaque culture et niveau de pression des maladies et ravageurs.

B. Le modèle Gamsel

Les données ont été traitées par un modèle additif généralisé avec sélection automatique des variables les plus pertinentes. Nous avons ajouté aux données les interactions spécifiques à chaque culture entre les bioagresseurs, classifiés en maladies et ravageurs, et les traitements insecticides et fongicides pour étudier leur impact sur le rendement, en multipliant les variables des bioagresseurs avec les variables de traitement correspondantes, pour modéliser les effets combinés sur le rendement. Le modèle a tourné avec et sans normalisation des données pour faciliter l'interprétation. Les colonnes ayant moins de trois valeurs uniques ont été supprimées pour permettre un ajustement efficace du modèle.

L'analyse effectuée sur le modèle Gamsel montre comment évoluent les coefficients non nuls dans le modèle, répartis en termes linéaires et non linéaires. La qualité d'ajustement du modèle est analysée grâce au pourcentage de déviance expliqué par le modèle. Au fur et à mesure que le lambda choisi pour la modélisation (paramètre de régularisation) varie, la complexité du modèle augmente : les pourcentages de déviance expliqués sont croissants, ce qui indique une meilleure explication de la variance du rendement (variable de réponse). Pour chaque culture, on a déterminé la valeur optimale du paramètre de régularisation « lambda » grâce à la validation croisée. Le lambda pour lequel les erreurs de validation croisée cessent de diminuer de manière monotone a été identifié, avec une marge d'un écart-type au-dessus de l'erreur minimale monotone.

La validation croisée est appliquée au modèle sur les mêmes critères de sélection que pour le modèle RF (par année, par département, par échantillonnage aléatoire et en combinant années et départements).

Les performances du modèle ont été évaluées grâce à la RMSE, pour mesurer l'erreur quadratique moyenne entre les valeurs prédites et les valeurs observées, grâce à l'erreur moyenne, et par R^2.

Par évaluation croisée des effets des bioagresseurs par culture, en tenant compte des variations de rendement selon les traitements, on a déterminé si l'effet des bioagresseurs sur le rendement était croissant, décroissant ou stable en évaluant les pentes : plus la pression des bioagresseurs est forte, plus le rendement diminue, et inversement. Cette analyse a permis de distinguer les cas de surcompensation du modèle, d'inefficacité du traitement, ou de l'absence d'effet. Enfin, par analyse statistique, on a évalué la significativité des résultats du modèle par rapport au hasard.

C. Le modèle GLM Lasso

Le même ensemble de données a enfin été traité par régression avec un modèle GLM LASSO (*Least Absolute Shrinkage and Selection Operator*), utilisé pour la sélection de variables en grand nombre. Il ajoute une pénalité à la somme des valeurs absolues des coefficients des variables explicatives, selon un paramètre de contrôle lambda. La fonction

$$\text{Minimiser}\left(\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

où :

- **RSS** (Residual Sum of Squares) est la somme des carrés des résidus (erreurs) du modèle.
- **λ** est le paramètre de régularisation, contrôlant la force de la pénalité.
- **β_j** représente les coefficients des variables explicatives.

de coût pour le Lasso est la suivante :

Le Lasso permet à certains coefficients de devenir exactement nuls, donc de sélectionner automatiquement les variables pertinentes pour le modèle. Cela permet de limiter *l'overfitting*, encore trop présent dans le Random Forest, et améliore la généralisation du modèle. Après avoir importé les données pertinentes, en supprimant certaines colonnes, comme les identifiants des parcelles, on a ajouté au modèle de rendement les interactions entre bioagresseurs (maladies et ravageurs) et les IFT correspondantes (pesticides et fongicides). Les variables des bioagresseurs ont été forcée négativement : l'activité des bioagresseurs sur une parcelle ne peut pas améliorer son rendement. Le modèle a tourné avec et sans l'inclusion du rendement potentiel (PredGamSel) obtenu précédemment avec le modèle de rendement, pour évaluer la validité des performances du modèle et obtenir une meilleure estimation des relations entre variables explicatives et rendement, et éviter le surajustement en

améliorant la généralisation. Les données ont été normalisées et standardisées pour faciliter l'analyse et l'interprétation, puis le modèle a été ajusté avec le package R 'glmnet'. Par validation croisée interne (fonction « cv.glmnet »), le meilleur paramètre de régularisation lambda a été choisi. Les résultats du modèle incluent les coefficients, les prédictions et le meilleur lambda pour chaque culture. La robustesse du modèle a ensuite été testée par validation croisée. La validation croisée est utilisée encore une fois par année, par département, par échantillonnage aléatoire et par année et département combinés.

Pour analyser les interactions spécifiques à chaque culture, une colonne d'interaction a été ajoutée en multipliant les variables de maladies et les fongicides et les ravageurs et insecticides.

Résultats

Les différentes modélisations menées sur les 13 cultures du jeu de données ont permis d'établir quelques résultats quant à l'impact des bioagresseurs sur le rendement et à la réduction de cet impact par les traitements pesticides et fongicides. Le jeu de données utilisé s'est révélé inégal selon les cultures, étant proportionnel aux surfaces nationales de chacune de ces cultures, ce qui a impacté la capacité du modèle à produire des résultats fiables. Ainsi, les données pour le blé tendre d'hiver sont les plus abondantes (Tableau 2).

Betterave	2432
Blé dur d'hiver	1376
Blé tendre d'hiver	28914
Colza d'hiver	7565
Maïs ensilage	7606
Maïs grain	8388
Orge d'hiver	7388
Orge de printemps	3970
Pois d'hiver	582
Pois de printemps	1702
Pomme de terre	1478
Tournesol	3538
Triticale	1755

Tableau 2 : Visualisation de la quantité de points de données pour

Fonctionnement global des modèles

Les différentes modélisations réalisées visaient à mieux connaître l'impact des bioagresseurs sur le rendement, et dans quelle mesure les traitements phytosanitaires permettent de réduire cet impact. Les

données traitées, issues du réseau DEPHY, reflètent déjà une utilisation optimisée des traitements phytosanitaires : dans la majeure partie des cas, les rendements modélisés ne témoignent pas d'un usage trop intensif de ces traitements.

Le Modèle Random Forest a livré des performances variables selon les cultures, ce qui s'explique en partie par les disparités dans le nombre de points de données. Ses performances globales sont les suivantes :

1. Performance par Culture

Crop	Mean Yield	RMSE	RRMSE	R-squared
Betterave	13.72145	2.285658	0.1665755	0.2260809
Blé dur d'hiver	52.57088	14.48337	0.2755018	0.4318006
Blé tendre d'hiver	69.42566	14.45683	0.2082347	0.5295118
Colza d'hiver	31.26417	7.72228	0.2470009	0.3894112
Maïs ensilage	13.2282	2.985043	0.2256576	0.371006
Maïs grain	92.45134	20.89219	0.2259804	0.5464258
Orge d'hiver	62.40712	12.56411	0.201325	0.4327094
Orge de printemps	55.74812	13.00184	0.2332246	0.4631815
Pois d'hiver	31.0467	12.6367	0.4070224	0.3546476
Pois de printemps	33.58583	11.7251	0.3491086	0.3582422
Pomme de terre	45.26446	8.629249	0.1906407	0.1451244
Tournesol	22.522	6.173994	0.2741318	0.4189145

Triticale	52.12668	12.42789	0.238417	0.4747247

2. Performance Globale des Modèles

- Modèle Moyen Toutes Années Confondues
  - RMSE : 16.82524
  - RRMSE : 0.3227761
  - R^2 : 0.5000489
- Modèle Moyen Par Année
  - RMSE moyen : 14.79302
  - RRMSE moyen: 0.2878096

La RMSE permet d'évaluer les biais systématiques : le choix de ce paramètre permet une bonne évaluation globale du modèle. Ici, elle n'est pas normalisée : elle est à rapporter aux valeurs de rendement de chaque culture. La RRMSE permet de normaliser la RMSE en fonction du rendement moyen.

Les résultats des modèles révèlent des différences significatives en termes de précision de prédiction entre les différentes cultures. Le blé tendre d'hiver et le maïs grain ont un R^2 assez élevé, ce qui indique une bonne explication par le modèle de la variabilité des rendements, et une RRMSE relativement basse, ce qui indique une précision assez bonne des prédictions. En revanche, certaines cultures comme la pomme de terre sont moins performantes, avec un R^2 inférieur à 0,2. Les RRMSE les plus élevées, pour le pois d'hiver et le pois de printemps, montrent d'importantes erreurs de prédictions. Globalement, avec un R^2 de 0,5, le modèle a une bonne performance pour expliquer la variabilité des rendements. Cependant, la RRMSE moyenne est plus faible pour le modèle moyen par année, ce qui suggère que les modèles s'ajustent mieux aux variations spécifiques d'année en année, montrant l'importance des facteurs annuels dans la précision des prédictions.

Le modèle Gamsel a également livré des performances variables :

1. Performances par culture :

Culture	R^2	RMSE	Rdt moyen	RRMSE
Betterave	0.4379	16.1546	81.9791	0.1971
Blé dur d'hiver	0.3775	15.4332	53.0875	0.2907
Blé tendre d'hiver	0.3421	17.9423	69.4382	0.2584
Colza d'hiver	0.3351	7.8233	32.3782	0.2416
Maïs ensilage	0.1898	4.1576	13.1643	0.3158

Maïs grain	0.3345	25.9026	91.6950	0.2825
Orge d'hiver	0.2652	15.5860	62.2750	0.2503
Orge de printemps	0.3006	15.6773	55.8673	0.2806
Pois d'hiver	0.3652	12.6703	29.0122	0.4367
Pois de printemps	0.2945	12.6282	33.8029	0.3736
Pomme de terre	0.4182	9.4882	43.1769	0.2198
Tournesol	0.2157	7.4913	22.5479	0.3322
Triticale	0.3234	14.6341	50.7399	0.2884

2. Performances globales :

- R^2 : 0.3748
- RRMSE : 0,2639

Les résultats révèlent des variations significatives de performances du modèle : il est plus efficace pour expliquer les variations de rendements pour la betterave ou le blé tendre d'hiver que pour le maïs ensilage. Des valeurs basses de RRMSE indiquent une meilleure précision des prédictions par rapport aux rendements attendus, tandis que des valeurs élevées, comme pour le pois de printemps, suggèrent des erreurs de prédictions plus grandes par rapport aux rendements moyens. Globalement, avec un R^2 de 0,37, le modèle a une performance correcte mais pas excellente, et une erreur relative modérée par rapport aux rendements moyens.

Les graphes d'importance des variables dans l'élaboration du rendement ont été établis pour le modèle RF (Annexe). Ils sont relativement consistants avec la littérature en ce qui concerne le rendement des grandes cultures. Ainsi, pour le blé tendre d'hiver par exemple, les variables les dont l'importance est la plus élevée sont, outre le rendement potentiel prédit par notre modèle et le rendement standard fourni par DEPHY, la fertilisation minérale en azote, l'IFT fongicide, le travail du sol avant semis, l'IFT herbicide et la date de semis. Viennent ensuite les premiers bioagresseurs et les variations climatiques.

Maladies

L'un des buts de ce stage était de quantifier les impacts sur le rendement des maladies, et la façon dont les traitements phytosanitaires peuvent mitiger cet impact, en utilisant les données d'un réseau d'exploitation qui vise à en réduire l'usage. Dans le cas des maladies, nous avons pu constater au cours des différentes modélisations que les modèles rendent bien compte d'un impact sur le rendement. La culture pour laquelle les modèles ont le mieux fonctionné est le blé tendre d'hiver, en

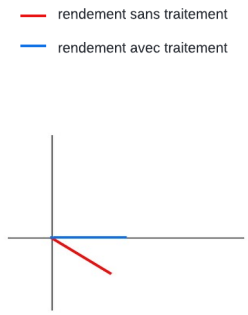
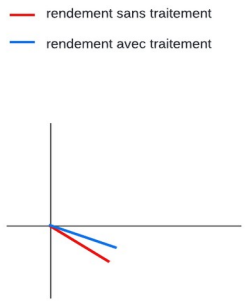
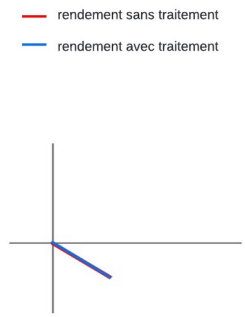
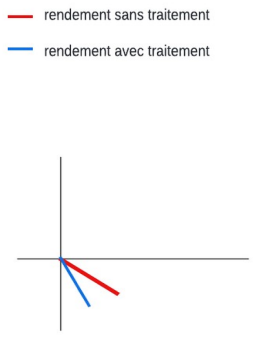
raison du grand nombre de points de données qui ont permis une bonne prédiction.

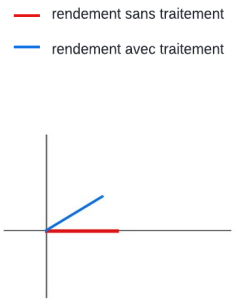
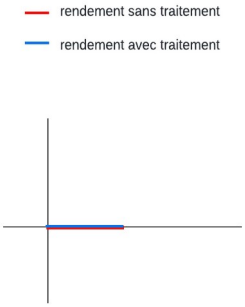
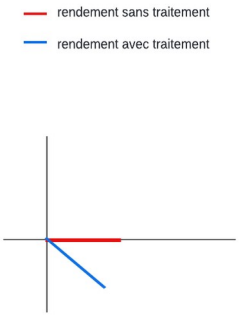
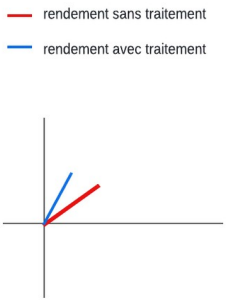
D’après le modèle Gamsel, pour les cultures de blé tendre d’hiver, de pois d’hiver et de triticale, le modèle montre que le traitement compense la perte de rendement occasionnée par les maladies (Tableau 3). Pour le colza d’hiver et l’orge de printemps, cette compensation n’est que partielle. Pour le tournesol, le traitement ne compense pas la perte de rendement liée aux maladies. Le modèle RF (Annexe 5) donne des résultats un peu différents : blé tendre d’hiver, colza d’hiver et pois d’hiver semblent montrer une bonne compensation de l’impact des maladies par les traitements phytosanitaires, tandis que pois de printemps, tournesol et triticale montrent une compensation partielle (Annexe 4).

Dans le tableau suivant, nous avons rassemblé les résultats donnés par le modèle RF et le modèle Gamsel pour les maladies, selon les critères que voici :

- Pression BA négative : les bioagresseurs exercent une pression négative sur les rendements, qui diminuent
- Pression BA neutre : les bioagresseurs n’exercent pas de pression significative sur le rendement d’après nos modélisations, ce qui peut révéler un problème dans le modèle
- Pression BA positive : les bioagresseurs exercent une pression positive sur le rendement, qui augmente : c’est très probablement un problème dans notre modèle

Impact des Bioagresseurs (BA)	**Impact des traitements (IFT)**	**Représentation schématique des évolutions de rendement**	**Cultures correspondantes d'après le modèle GAMSEL (nombre de maladies)**	**Cultures correspondantes d'après le modèle RF (nombre de maladies)**

Pression BA négative (le rendement diminue)	Le traitement compense la perte de rendement		Blé tendre d'hiver (7) Pois d'hiver (5) Triticale (6)	Blé tendre d'hiver (7) Colza d'hiver (2) Pois de printemps (4)
	Le traitement compense partiellement la perte de rendement		Colza d'hiver (2) Orge de printemps (3)	Pois d'hiver (5) Tournesol (2) Triticale (6)
	Le traitement ne compense pas la perte de rendement		Tournesol (2)	Aucun cas dans nos modélisations
	Le traitement aggrave la perte de rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations

Pression BA neutre (le rendement n'est pas impacté)	Le traitement surcompense la perte de rendement		Orge d'hiver (4) Pois de printemps (5) Pomme de terre (2)	Orge d’hiver (4)
	Le traitement n'a pas d'effet		Betterave (2) Maïs ensilage (0) Maïs grain (0)	Betterave (2) Maïs grain (0) Maïs ensilage (0)
	Le traitement a un effet négatif sur le rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations
Pression BA positive (le rendement augmente)	Le traitement augmente beaucoup le rendement		Blé dur d'hiver (7)	Blé dur d’hiver (7) Orge de printemps (3)

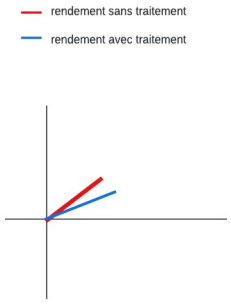
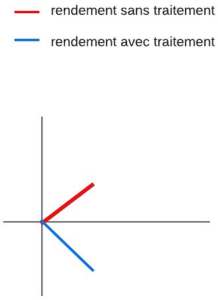
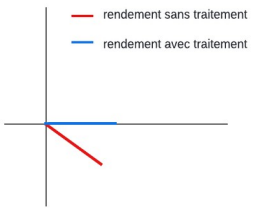
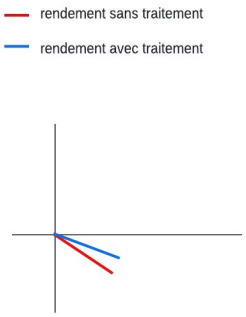
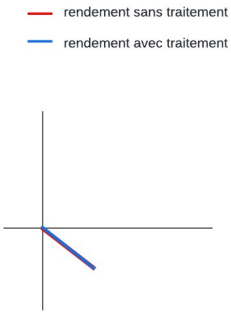
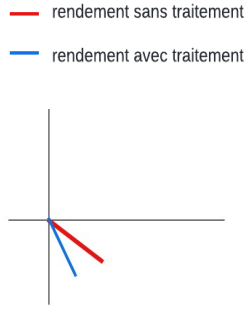
	Le traitement augmente légèrement le rendement		Aucun cas dans nos modélisations	Pomme de terre (2)
	Le traitement diminue le rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations

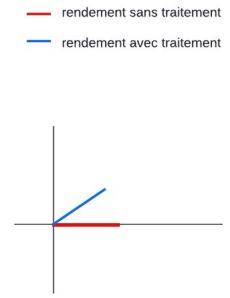
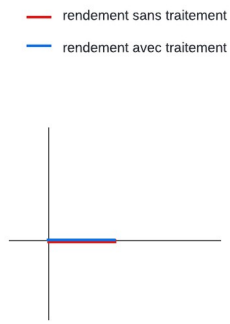
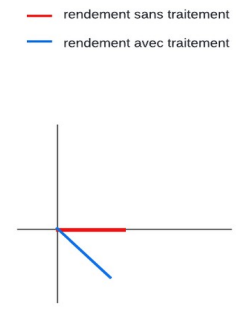
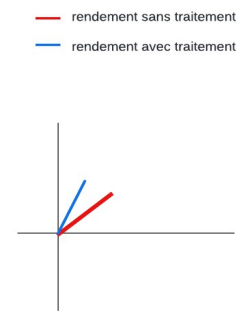
Tableau 3 : Résultats de la modélisation : impact des maladies et mitigation des traitements phytosanitaires. Le seuil d'acceptation a été fixé à 5%.

Ravageurs

Les modèles utilisés ont été moins bons pour prédire comment les ravageurs affectent les rendements et quel est l'impact des traitements phytosanitaires sur ce dernier en utilisant les données DEPHY. Ainsi, aucune des cultures ne montre à la fois un impact négatif des ravageurs sur le rendement et une compensation de cet impact par les traitements insecticides pour le modèle Gamsel (Annexe 4). Pour le modèle RF (Annexe 5), seuls l'orge de printemps, l'orge d'hiver et le pois de printemps montrent, assez faiblement, une pression des ravageurs sur le rendement.

Impact des Bioagress	Impact des traiteme	Représentation schématique des évolutions de	Cultures correspondantes d'après le	Cultures correspondantes d'après le

eurs (BA)	nts (IFT)	rendement	modèle GAMSEL (nombre de ravageurs)	modèle RF (nombre de ravageurs)
Pression BA négative (le rendement diminue)	Le traitement compense la perte de rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations
	Le traitement compense partiellement la perte de rendement		Aucun cas dans nos modélisations	Orge de printemps (1) Orge d'hiver (1) Pois de printemps (3)
	Le traitement ne compense pas la perte de rendement		Tournesol (3) Betterave (1)	Betterave (4) Maïs ensilage (1) Pomme de terre (2) Triticale (1)
	Le traitement aggrave la perte de rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations

Pression BA neutre (le rendement n'est pas impacté)	Le traitement surcompense la perte de rendement		Triticale (1) Pois de printemps (3)	Aucun cas dans nos modélisations
	Le traitement n'a pas d'effet		Maïs grain (0) Maïs ensilage (0) Orge d'hiver (1) Orge de printemps (1) Blé tendre d'hiver (2) Colza d'hiver (8)	Blé dur d'hiver (2) Blé tendre d'hiver (2) Colza d'hiver (8) Tournesol (3) Maïs grain (1)
	Le traitement a un effet négatif sur le rendement		Blé dur d'hiver (2)	Aucun cas dans nos modélisations
Pression BA positive (le rendement augmente)	Le traitement augmente beaucoup le rendement		Pomme de terre (2)	Aucun cas dans nos modélisations

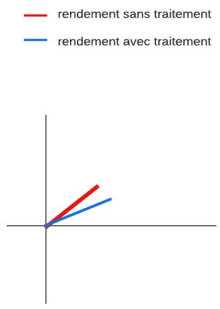
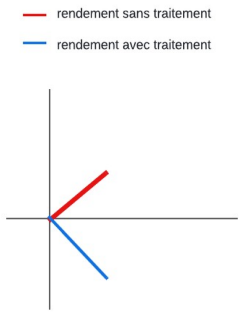
	Le traitement augmente légèrement le rendement		Pois d'hiver (3)	Pois d'hiver (3)
	Le traitement diminue le rendement		Aucun cas dans nos modélisations	Aucun cas dans nos modélisations

Tableau 4 : Résultats de la modélisation : impact des ravageurs et mitigation des traitements phytosanitaires. Le seuil d'acceptation a été fixé à 5%.

Discussions

L'une des difficultés de la modélisation est sa grande amplitude d'utilisation : beaucoup de modèles de rendement se concentrent sur une culture en particulier (van Ittersum, 2012) et ne cherchent pas forcément à s'appliquer à toutes les grandes cultures, contrairement aux modèles du projet MoCoRiBA-GC. La précision de nos modèles est parfois sacrifiée au profit de leur généricité et de leur capacité à traiter des cultures très diverses : ils représentent un compromis entre robustesse, précision, complexité et capacité à gérer les grands ensembles de données utilisés.

Les résultats du modèle RF montrent une bonne précision pour la betterave et le maïs ensilage, mais des erreurs moyennes importants pour le maïs grain et le pois de printemps. Au niveau du R^2, le blé tendre d'hiver et le maïs grain présentent une bonne adéquation du modèle aux données, mais celui-ci échoue à expliquer la variance de la pomme de terre et du pois d'hiver. Globalement, la performance du modèle est équilibrée, avec un R^2 satisfaisant mais qui appelle à davantage de perfectionnement du modèle, dont les performances montrent une grande variabilité.

Tous les résultats des graphes d'importance (Annexe) sont à discuter : le défaut du modèle RF c'est qu'il peut échouer à percevoir les interactions entre certaines variables (en particulier non linéaires) et donne ainsi une importance plus grande à une variable en interaction avec d'autres.

Pour ce qui est du modèle Gamsel, les variations importantes des performances des modèles ainsi que les erreurs élevées montrent que le modèle pourrait être amélioré. A ce titre, il serait pertinent par exemple de prendre en compte les traitements phytosanitaires mensuels plutôt que leur moyenne annuelle, qui varie beaucoup. De même, cela permettrait de mieux cibler les traitements spécifiques à différents bioagresseurs.

Pour les tableaux de résultats, le seuil d'acceptation des variations des courbes de rendement pour quantifier les impacts des bioagresseurs sur le rendement, et dans quelle mesure les traitements phytosanitaires sont efficaces (tableaux 2 et 3) a été fixé arbitrairement à 5%. Le choix de ce seuil est lié aux taux de diminutions de rendement jugés acceptables pour un exploitant. Nous l'avions dans un premier temps fixé à 3%, mais cela rendait les résultats moins apparents.

Un certain nombre de résultats montrent des effets positifs sur le rendement en cas de traitement phytosanitaire : ceux-ci peuvent en partie être expliqués par un « effet vert » (F. Vancutsem, 2006) des traitements sur les cultures, en particulier pour les céréales. L'absence d'effet des traitements insecticides témoigne, dans une certaine mesure, des effets de résistance (Siddiqui JA, 2023) observés sur le terrain. Enfin, une partie de la variabilité des rendements n'est pas expliquée par le modèle : par exemple, un IFT fongicide très haut sur le colza peut être expliqué par le choix des exploitants de traiter davantage pour des questions d'assurance. A niveaux d'herbicides égaux, un travail du sol plus important peut indiquer davantage de problèmes liés aux adventices ; ce que nos modèles peinent à prendre en compte. Il n'est enfin pas étonnant d'atteindre un certain plateau de rendement par rapport à la pression des bioagresseurs. L'un des problèmes de nos résultats est la difficulté à lier explicitement l'importance des bioagresseurs et des interactions entre bioagresseurs et traitements phytosanitaires sur le rendement.

Conclusion

L'étude des performances de nos modèles de prédiction des rendements agricoles en fonction de la pression des bioagresseurs et des effets des traitement phytosanitaires, appliqués aux 13 cultures sur des données issues du réseau DEPHY, a produit des résultats hétérogènes. Les modèles montrent une variabilité significative dans leur capacité de

prédiction, influencée par la grande disparité dans la disponibilité des données. En particulier, le blé tendre d'hiver se distingue par une meilleure qualité de prédiction. Les résultats montrent que le modèle fournit, malgré ses limites, des informations utiles. Il nécessite cependant des ajustements, notamment en affinant l'analyse de l'apport des traitements phytosanitaires. Ces modèles sont un travail en cours, qui n'est pas achevé.

Nos résultats mettent tout de même en lumière une efficacité variable des traitements phytosanitaires pour compenser l'impact des bioagresseurs sur les rendements. Certains traitements montrent une bonne compensation et limitent les pertes, d'autres semblent au contraire être inutiles voire les aggraver, ce qui révèle des problèmes dans la modélisation. Les modèles échouent cependant à capturer l'ensemble des effets des bioagresseurs et des traitements phytosanitaires, en partie en raison des limites inhérentes à la modélisation de rendements agricoles complexes et très variés, puisqu'appliqués à l'ensemble des grandes cultures.

Des ajustements aux modèles sont nécessaires pour mieux prendre en compte les interactions complexes entre bioagresseurs et traitements, améliorer la précision des prédictions et mieux comprendre les facteurs de variations des rendements afin de fournir des recommandations plus fiables pour la gestion des cultures. Il faudrait peut-être augmenter la sophistication du modèle et intégrer davantage de données sur certaines cultures pour renforcer les capacités prédictives et la robustesse de ces modèles.

Références

Agreste. (s.d.). Récupéré sur Agreste: https://agreste.agriculture.gouv.fr/agreste-web/download/publication/publie/IraGcu21053/2021_53inforapgdescultures.pdf

Alberto Gonzalez-Sanchez, J. F.-S.-B. (2014). Predictive ability of machine learning methods for massive. *Spanish Journal of Agricultural Research*.

Arvalis, I. d. (2020). *Arvalis - Institut du végétal*. Récupéré sur https://www.arvalis.fr/

Brisson N, M. B. (1998). STICS: a generic model for the simulation of crops and their water and nitrogen balance. *Agronomie 18*.

Butault, J.-P. D. (2010). Quelles voies pour réduire l'usage des pesticides ? Synthèse du rapport d'étude. *Ecophyto R&D*.

Chevaleyre, C. (2023). *Modélisation de la pression des bioagresseurs (insectes et maladies fongiques) des grandes cultures.*

D. Bondre, S. M. (2019). Prediction of crop yield and fertilizer recommandation using machine learning algorithms.

Estelle Ancelet, N. M.-J. (2015). Agrosyst, le système d'information au coeur du Plan Ecophyto de réduction d'usage des pesticides. . *Colloque "Données, Agriculture, Environnement.. Innovation", Agreenium-IAVFF.*

Evans, L. T. (1993). Crop evolution, adaptation and yield. *Cambridge University Press*.

F. Vancutsem, B. B.-M.-P. (2006). Les effets « extra-fongicides » des strobilurines en froment : mythe ou réalité ?

INSEE. (s.d.). *La France et ses territoires Édition 2021*. Récupéré sur https://www.insee.fr/fr/statistiques/5039859?sommaire=5040030

Lay, C. (2020). *Using DEPHY farm experience to advise farmer on pesticide use.*

Lechenet, M. D.-J. (2017). Reducing pesticide use while preserving crop productivity and profitability on arable farms. *Nature plants 3(3)*, 1-6.

M. van Ittersum, K. C. (2012). Yield gap analysis with local to global relevance - A review. *Elsevier*.

N. Devaud, C. B. (2019). Quantification of bioagressors induced yield gap for grain crops in France. *bioRxiv*.

N. Koning, M. v. (2008). Long-term global availability of food: continued abundance or new scarcity ? . *NJA 55*.

P. Priya, U. M. (2018). Predicting yield of the crop using machine learning algorithm.

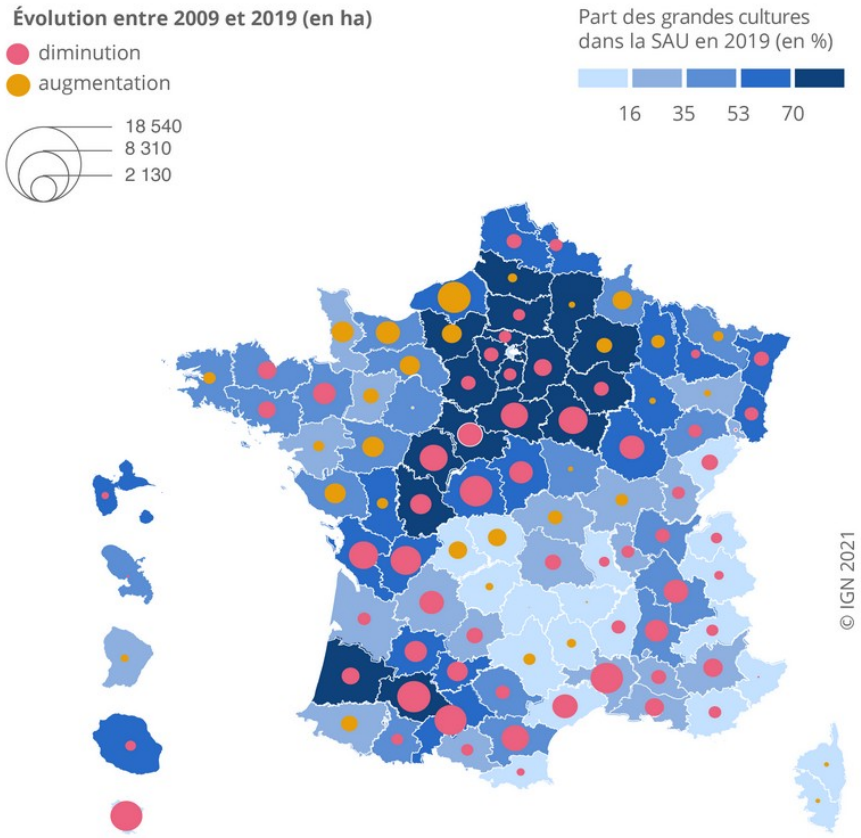
Qaddoum, K. (2014). Modified naive bayes based prediction modeling for crop yield prediction. *International Journal of biological, Biomolecular, Agricultural, Food and Biotechnological Engineering, vol. 8*.

R. Carew, E. G. (2015). *Factors influencing wheat yield and variability: evidence from Manitoba, Canada.*

S. T. Drummont, K. S. (2003). Statistical and neural methods for site-specific yield prediction. *T ASABE 46*.

Siddiqui JA, F. R. (2023). Insights into insecticide-resistance mechanisms in invasive species: Challenges and control strategies. *Frontiers in Physiology*.

V. Langlois, R. B. (2019). *mémoire sur l'impact des pesticides dans l'environnement au Québec.* Commission de l'agriculture, des pêchesries, de l'énergie et des ressources naturelles.


Évolution entre 2009 et 2019 (en ha)
diminution
augmentation
18 540
8 310
2 130
Part des grandes cultures
dans la SAU en 2019 (en %)
16
35
53
70
© IGN 2021

Annexes

Annexe 1 : Répartition de la SAU en France en 2019 (source : Agreste,

Unités : 1 000 ha, %.	2016 (1)	2017 (1)	2018 (1)	2019 (1)	2020 (2)	MOY. 16-20	2021 (3)	2021 /2020	2021 /MOY. 16-20
CEREALES (a)	**9 525**	**9 339**	**9 055**	**9 393**	**8 909**	**9 244**	**9 191**	+ 3,2	– 0,6
Blé tendre	5 132	4 962	4 880	4 999	4 262	4 847	4 891	+ 14,8	+ 0,9
hiver	5 120	4 948	4 866	4 983	4 222	4 828	4 873	+ 15,4	+ 0,9
printemps	12	14	14	16	40	19	18	– 56,2	– 8,0
Blé dur	394	370	354	246	252	323	267	+ 6,1	– 17,3
hiver	386	361	347	239	219	310	249	+ 14,1	– 19,6
printemps	8	9	7	7	33	13	18	– 46,8	+ 39,5
Orge, escourgeon	1 917	1 905	1 768	1 944	1 972	1 901	1 796	– 8,9	– 5,5
hiver	1 506	1 398	1 284	1 305	1 177	1 334	1 203	+ 2,2	– 9,8
printemps	411	507	484	639	795	567	593	– 25,4	+ 4,6
Avoine	85	113	92	87	100	96	99	– 0,4	+ 4,0
hiver	51	71	59	49	42	54	50	+ 20,0	– 7,6
printemps	35	42	32	38	58	41	49	– 15,1	+ 19,3
Seigle	25	24	24	29	32	27	32	+ 0,3	+ 19,0
Triticale	331	305	284	305	261	297	305	+ 17,1	+ 2,8
Autres (pures et mélanges)	120	153	153	180	208	163	172	– 17,4	+ 5,8
Riz	15	15	12	14	14	14	13	– 8,2	– 9,2
Céréales à paille	**8 019**	**7 847**	**7 567**	**7 804**	**7 100**	7 668	**7 576**	+ 6,7	– 1,2
Maïs (b)	1 458	1 436	1 426	1 506	1 692	1 503	1 521	– 10,1	+ 1,2
grain (b)	1 392	1 376	1 365	1 436	1 609	1 436	1 442	– 10,4	+ 0,4
semences	66	60	61	70	82	68	80	– 3,2	+ 17,4
Sorgho grain	48	56	61	83	117	73	94	– 19,7	+ 28,5
OLEAGINEUX (a)	**2 262**	**2 169**	**2 357**	**1 907**	**2 121**	**2 163**	**1 868**	– 11,9	– 13,6
Colza	1 549	1 401	1 617	1 107	1 114	1 358	989	– 11,2	– 27,1
hiver	1 548	1 399	1 615	1 105	1 112	1 356	987	– 11,2	– 27,2
printemps	1	2	2	2	3	2	2	– 14,2	+ 7,9
Tournesol	542	586	552	604	778	612	666	– 14,4	+ 8,7
Soja	137	142	154	164	187	157	172	– 8,1	+ 9,6
Autres oléagineux	34	39	35	32	42	36	42	+ 0,3	+ 15,7
PROTEAGINEUX (a)	**301**	**299**	**227**	**242**	**312**	**276**	**319**	+ 2,2	+ 15,5
Féveroles (et fèves)	78	77	57	63	77	70	78	+ 1,4	+ 10,3
Pois protéagineux	216	216	167	176	230	201	235	+ 2,5	+ 17,2
Pois protéagineux purs							201		
Mélange de pois							35		
Lupin doux	8	5	3	3	6	5	6	+ 0,7	+ 19,4
BETTERAVES (c)	**405**	**486**	**486**	**447**	**421**	**449**	**396**	– 5,9	– 11,8
POMMES DE TERRE (d)	**172**	**185**	**191**	**198**	**...**	**...**	...	...	...
Plants	19	21	22	23	...	...	...	...	...
Féculerie	23	23	24	22	23	23	23	– 0,9	– 0,3
Conservation et demi-saison	130	141	145	153	159	146	152	– 4,2	+ 4,7
MAIS FOURRAGE	**1 433**	**1 406**	**1 416**	**1 436**	**1 419**	**1 422**	**1 317**	– 7,2	– 7,4

Source : AGRESTE

(a) Y compris semences (b) Y compris maïs grain humide

(1) Statistique Agricole Annuelle - Agreste

(2) Statistique Agricole Annuelle Provisoire 2020 - Agreste

(3) Statistique Mensuelle au 1er mai 2021 - Agreste

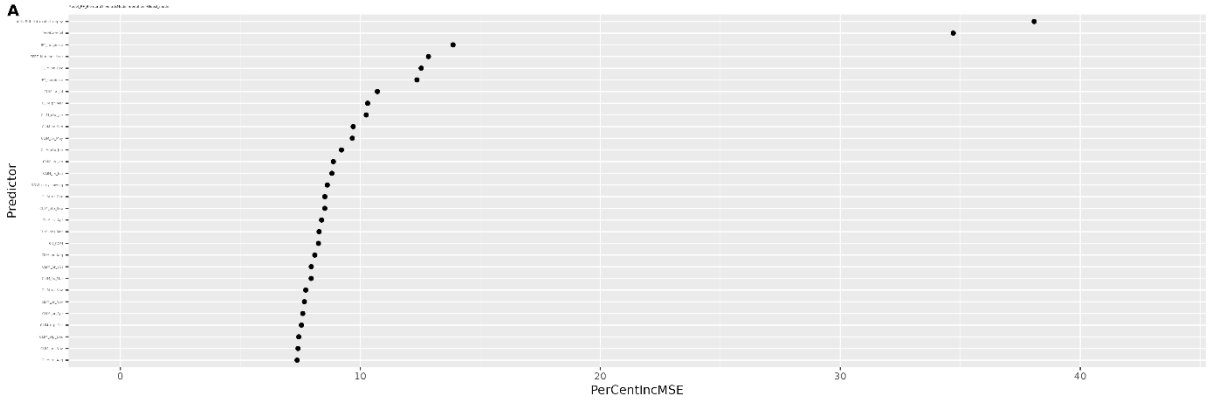
(c) Non compris semences

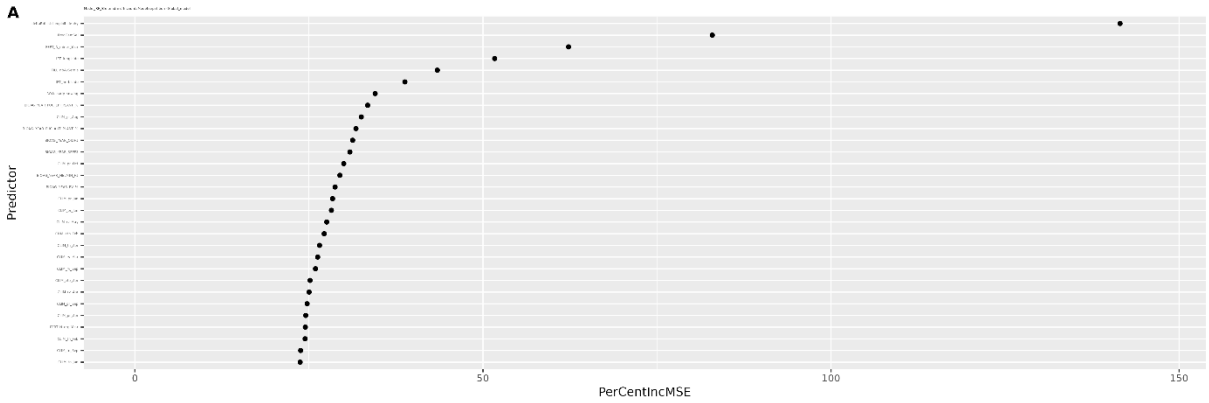
(d) Dessus de plants inclus dans la production, non compris dans les surfaces et rendements

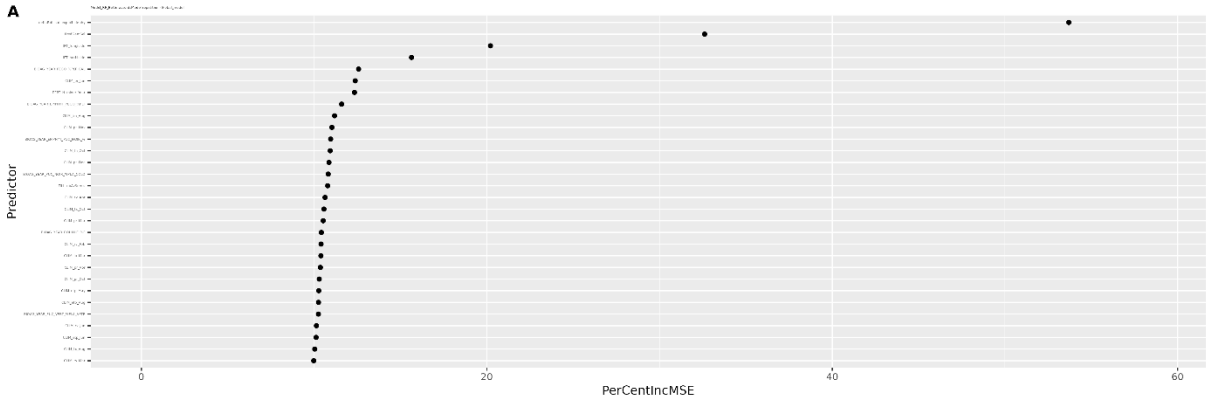
... données non disponibles

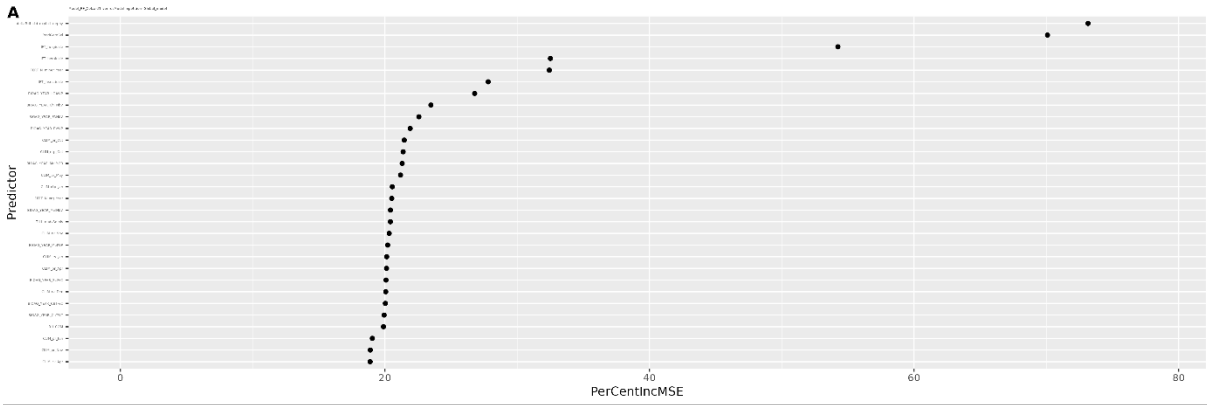
Variations positives

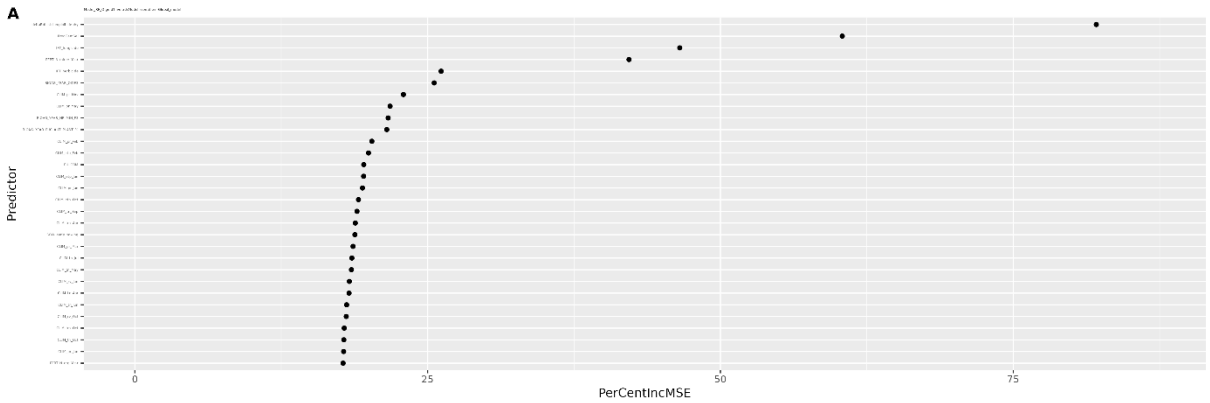
Annexe 2 : Superficies et évolutions des grandes cultures en France (données

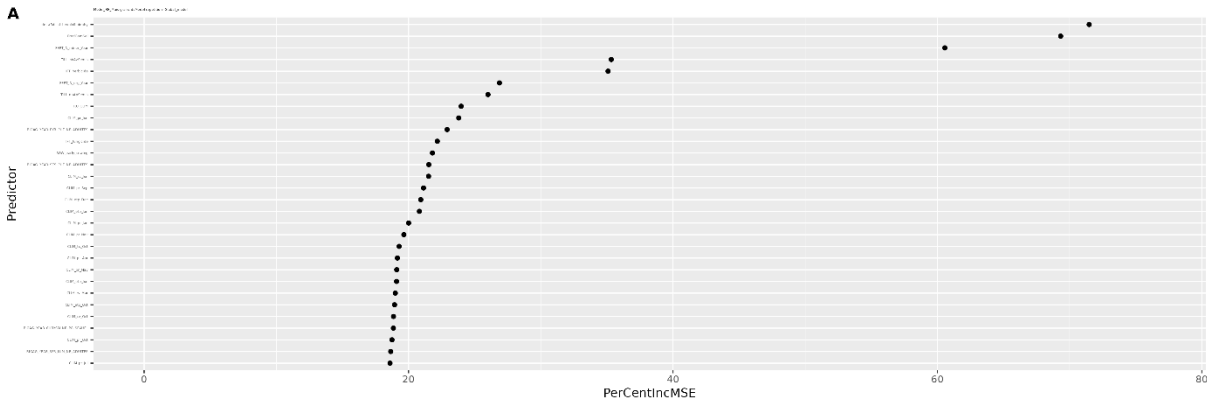

A
Predictor
0
10
20
30
40
PerCentIncMSE

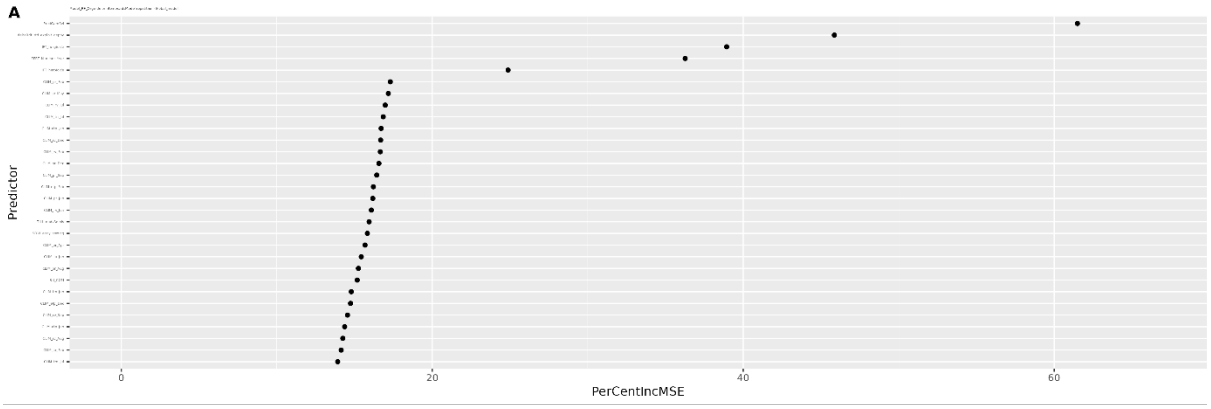

A
Predictor
0
50
100
150
PerCentIncMSE

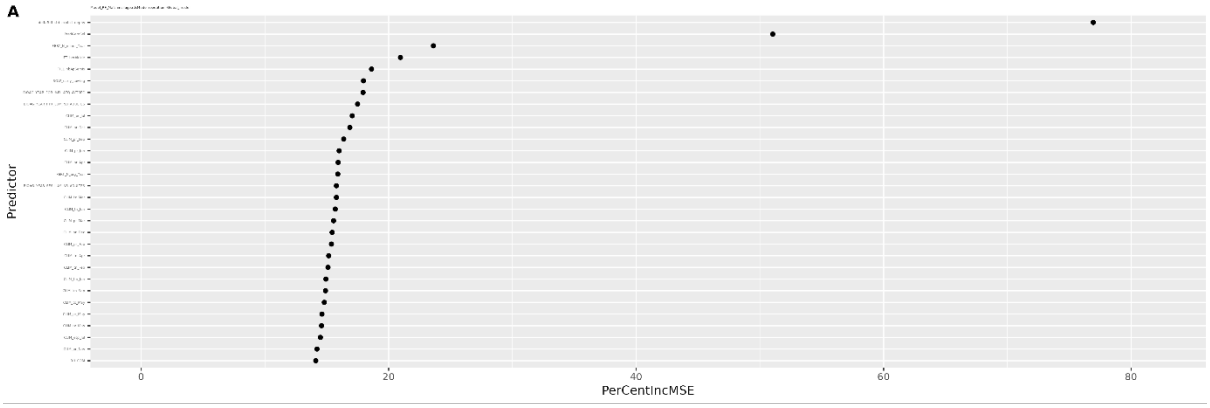

A
Predictor
0
20
40
60
PerCentIncMSE

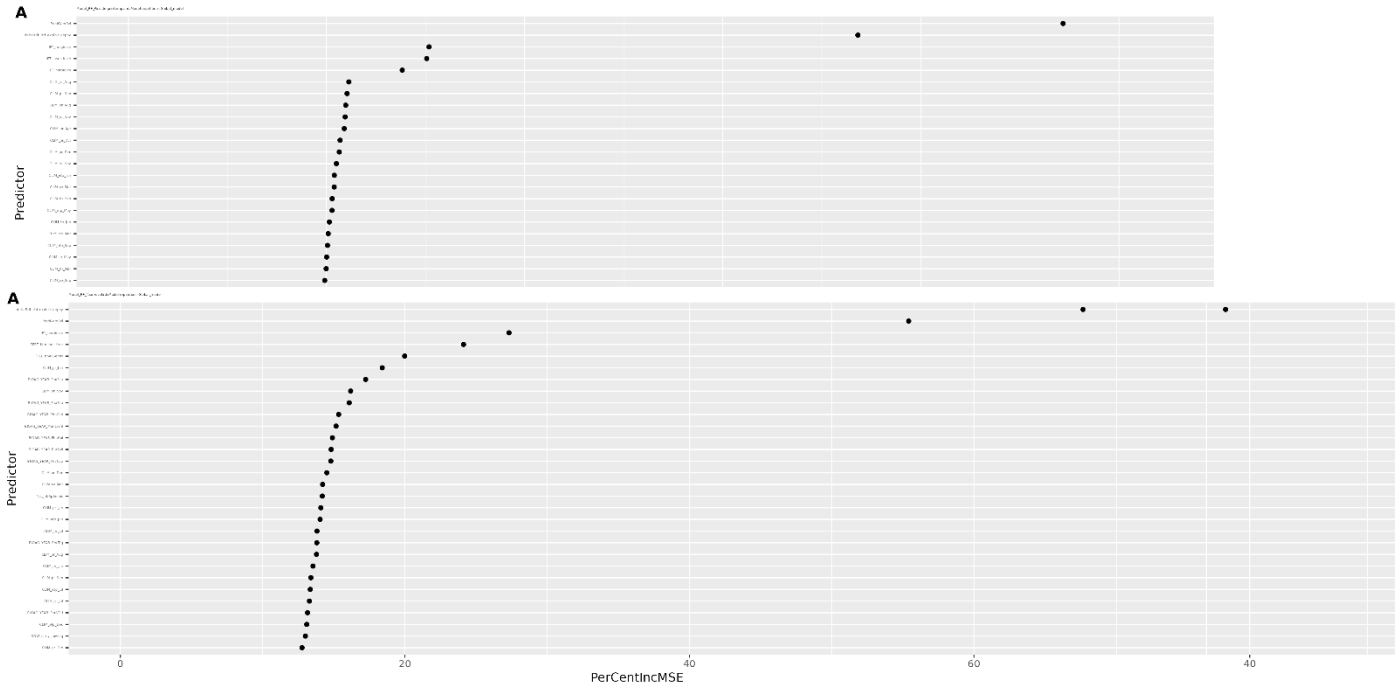

A
Predictor
0
20
40
60
80
PerCentIncMSE

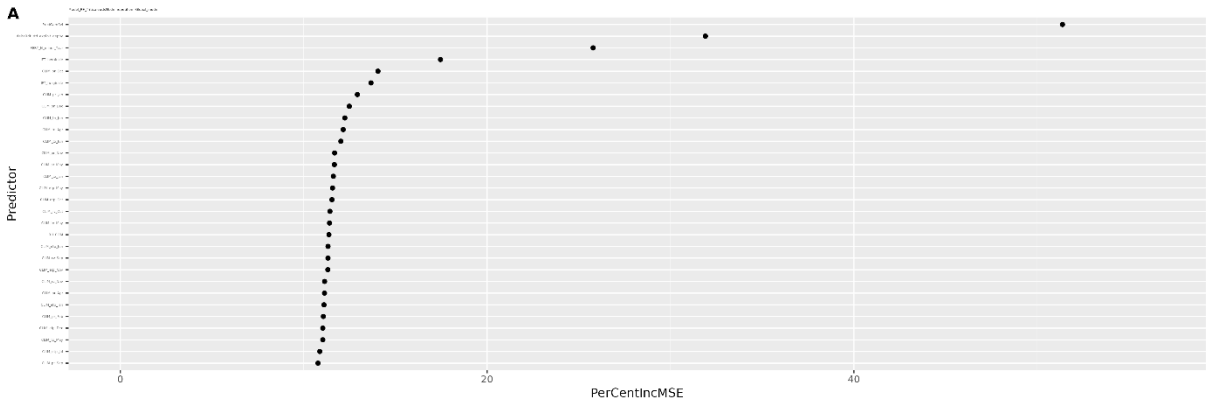

A
Predictor
0
25
50
75
PerCentIncMSE


A
Predictor
0
20
40
60
80
PerCentIncMSE

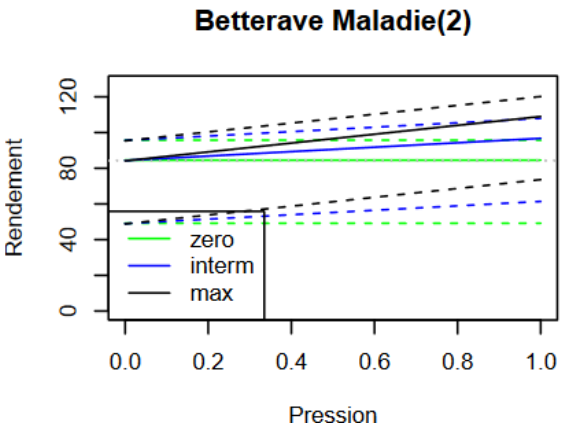

A
Predictor
0
20
40
60
PerCentIncMSE

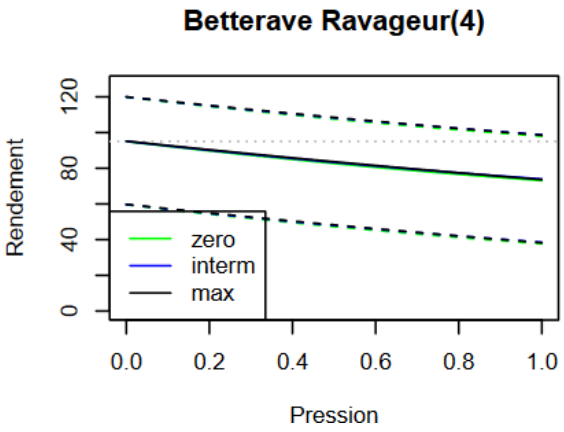

A
Predictor
0
20
40
60
80
PerCentIncMSE

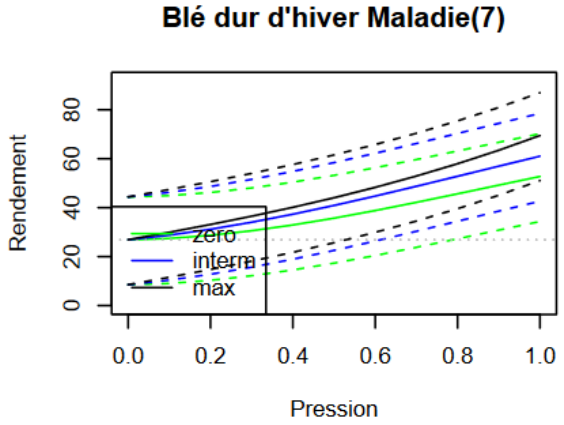

A
Predictor
A
Predictor
0
20
40
60
40
PerCentIncMSE

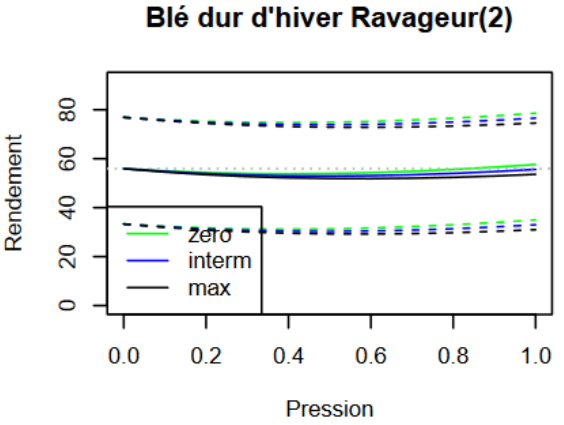


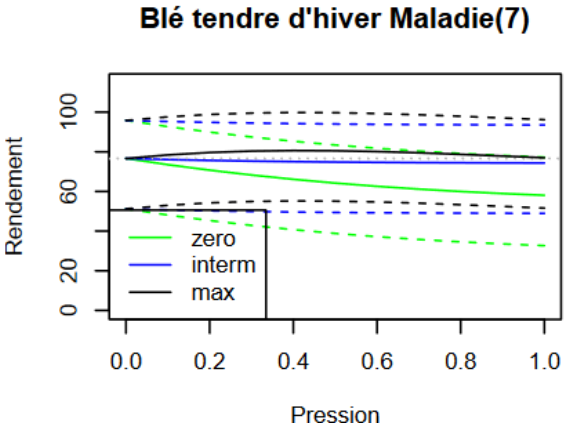
Annexe 3 : Graphes d’importance des variables par culture (Modèle RF)

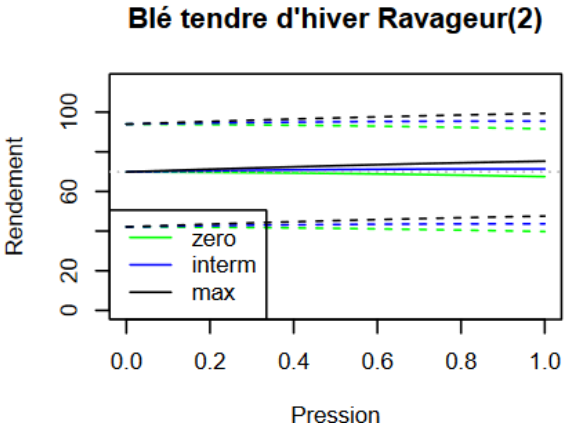

Betterave Maladie(2)
Rendement
0
40
80
120
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

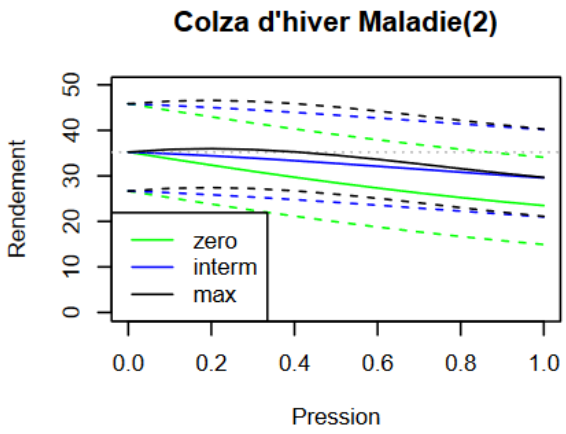

Betterave Ravageur(4)
Rendement
0
40
80
120
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

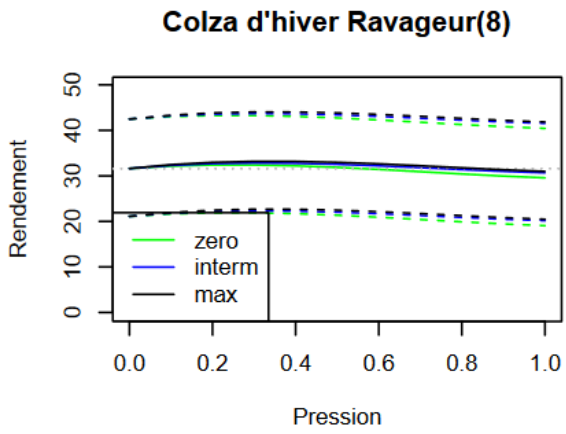

Blé dur d'hiver Maladie(7)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

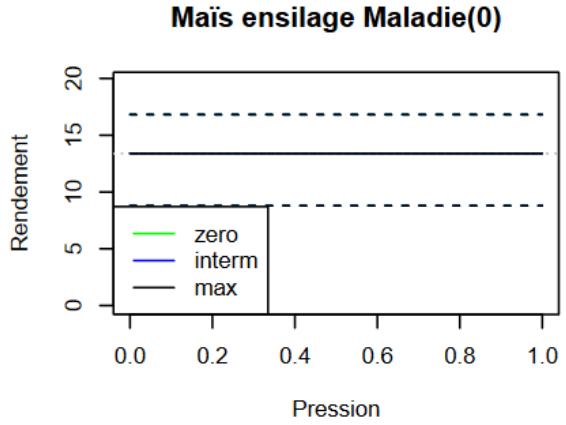

Blé dur d'hiver Ravageur(2)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

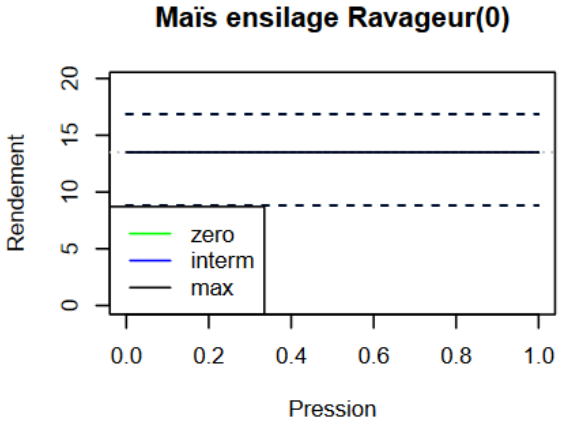

Blé tendre d'hiver Maladie(7)
Rendement
0
20
60
100
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

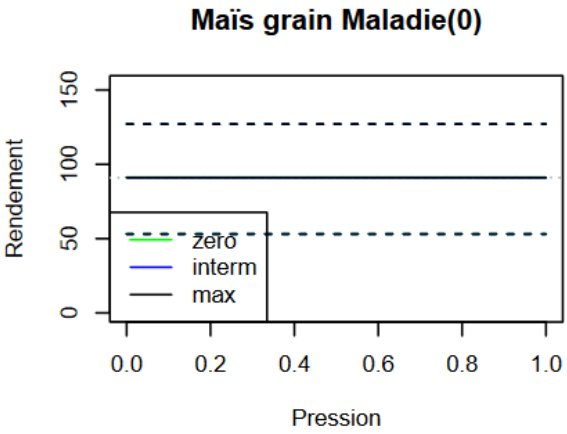

Blé tendre d'hiver Ravageur(2)
Rendement
0
20
60
100
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

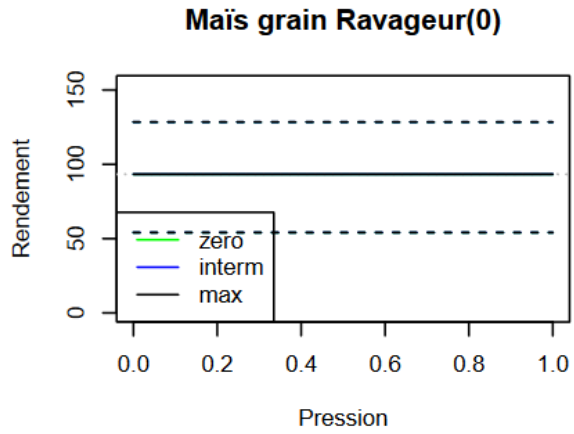

Colza d'hiver Maladie(2)
Rendement
0
10
20
30
40
50
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

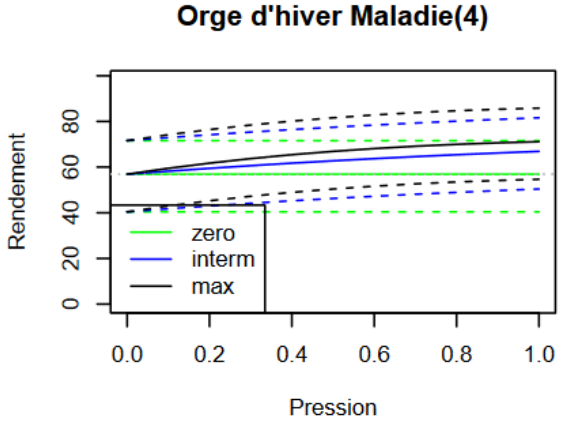

Colza d'hiver Ravageur(8)
Rendement
0
10
20
30
40
50
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

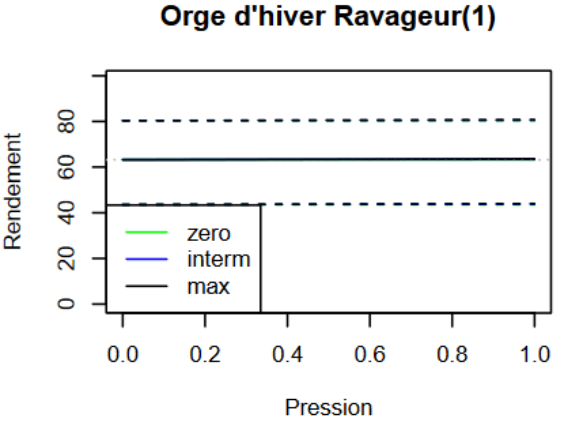

Maïs ensilage Maladie(0)
Rendement
0
5
10
15
20
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

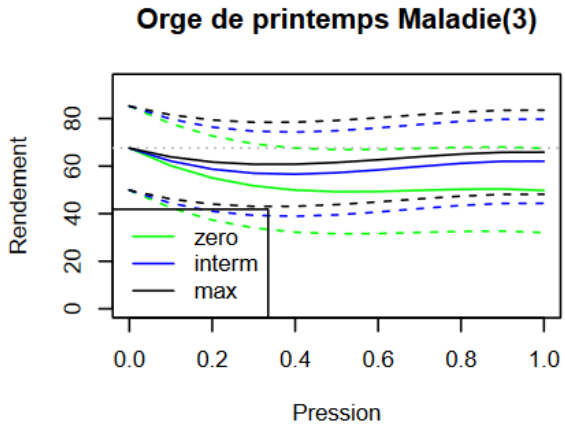

Maïs ensilage Ravageur(0)
Rendement
0
5
10
15
20
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

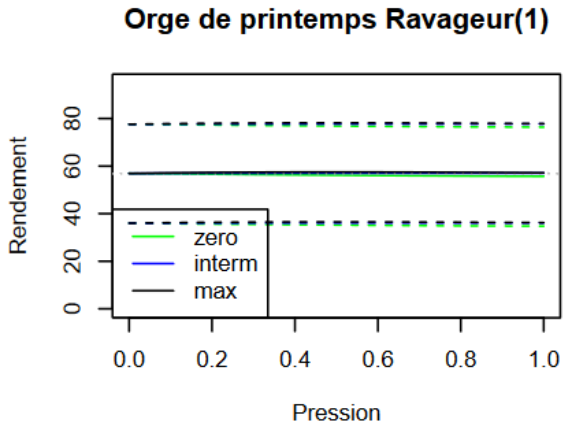

Maïs grain Maladie(0)
Rendement
0
50
100
150
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

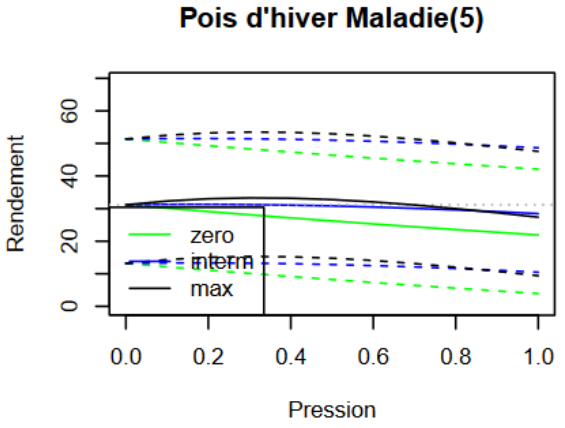

Maïs grain Ravageur(0)
Rendement
0
50
100
150
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

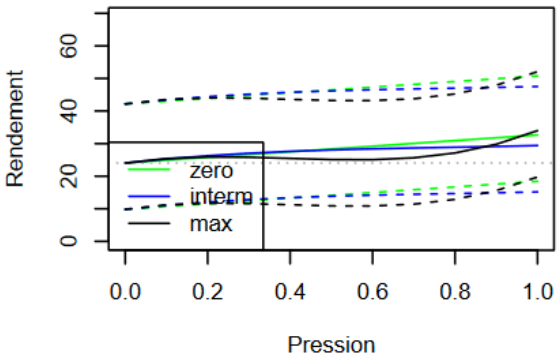

Orge d'hiver Maladie(4)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

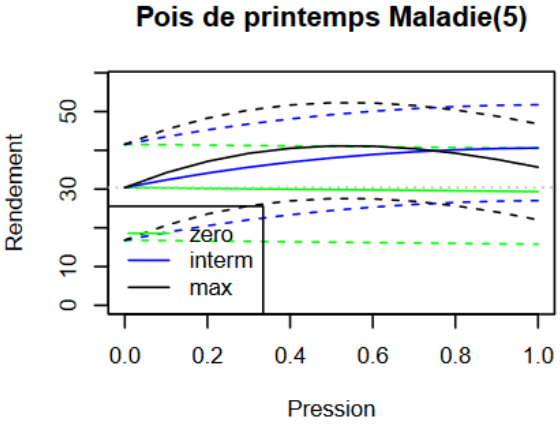

Orge d'hiver Ravageur(1)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

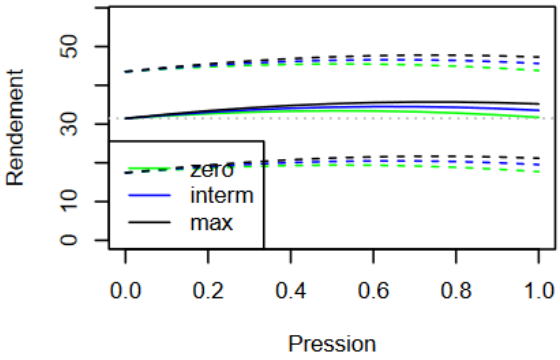

Orge de printemps Maladie(3)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

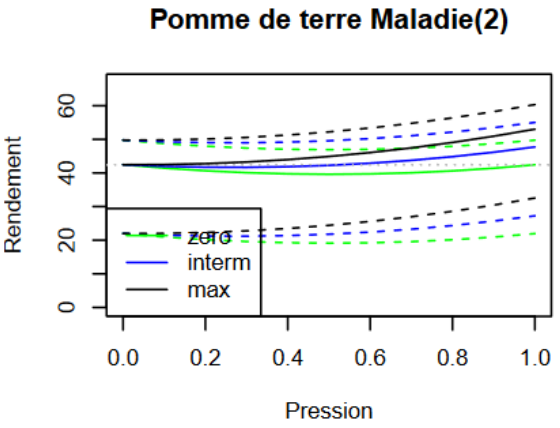

Orge de printemps Ravageur(1)
Rendement
0
20
40
60
80
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

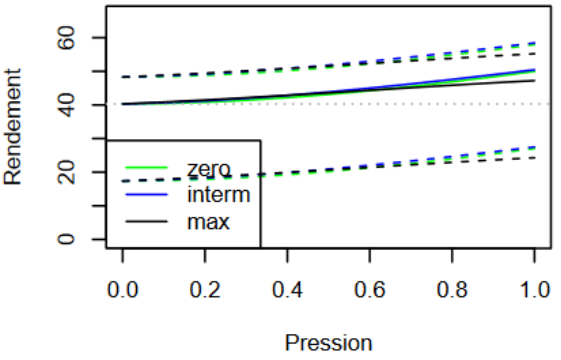

Pois d'hiver Maladie(5)
Rendement
0
20
40
60
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

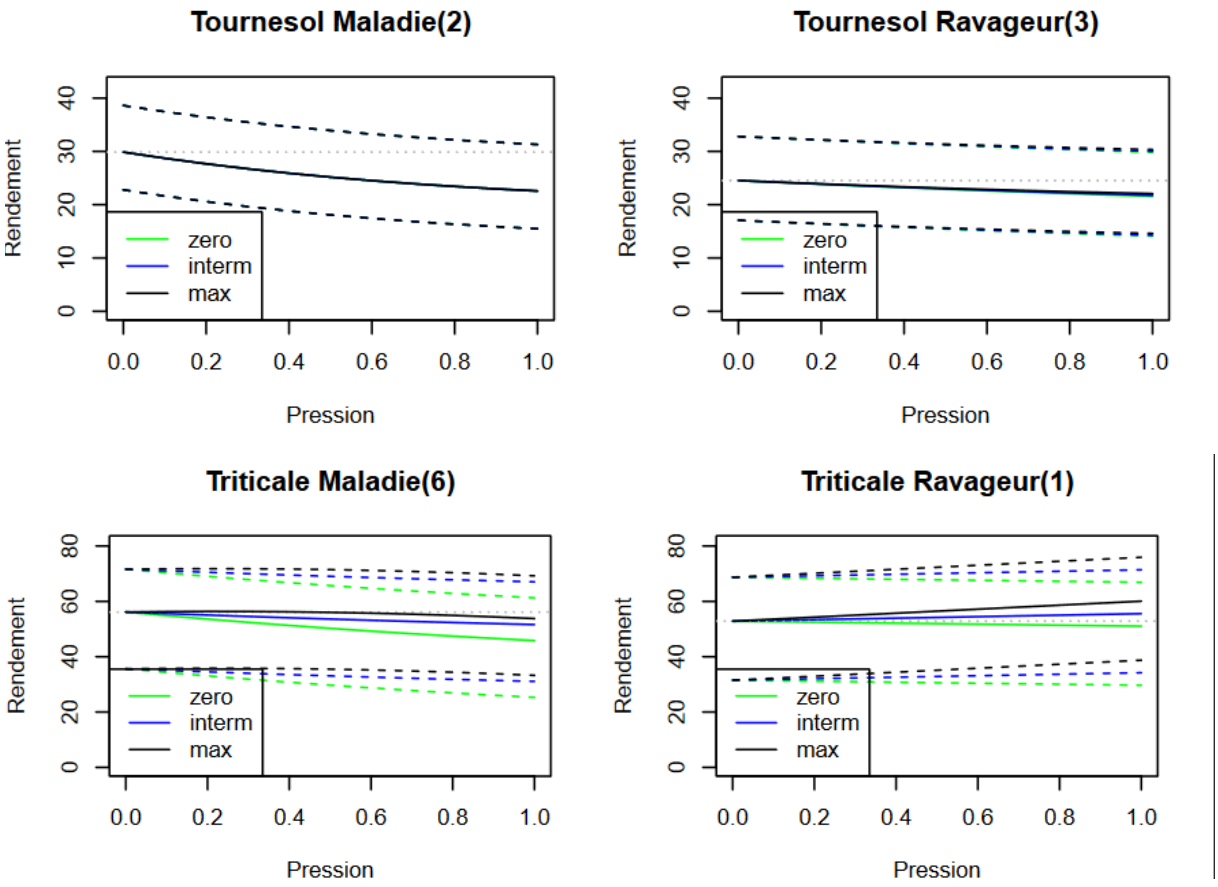

Pois d'hiver Ravageur(3)
Rendement
0
20
40
60
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression


Pois de printemps Maladie(5)
Rendement
0
10
30
50
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

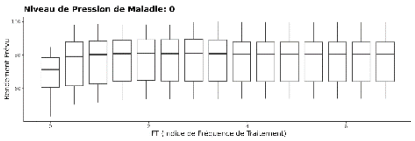

Pois de printemps Ravageur(3)
Rendement
0
10
30
50
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

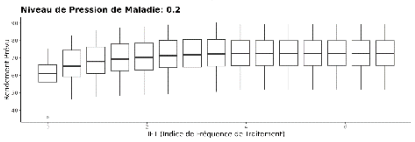

Pomme de terre Maladie(2)
Rendement
0
20
40
60
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

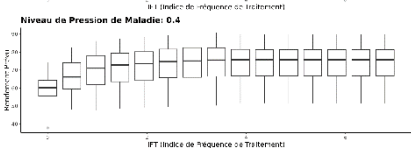

Pomme de terre Ravageur(2)
Rendement
0
20
40
60
zero
interm
max
0.0
0.2
0.4
0.6
0.8
1.0
Pression

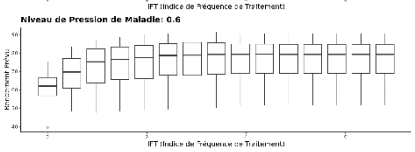


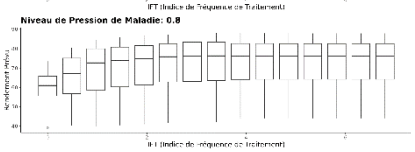
Annexe 4 : Graphes d'évolution du rendement en fonction des bioagresseurs et des traitements phytosanitaires (Modèle Gamsel).

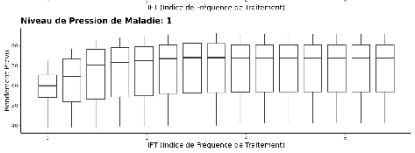

Niveau de Pression de Maladie: 0
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

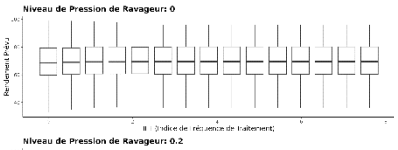

Niveau de Pression de Maladie: 0.2
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

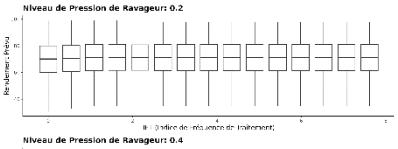

Niveau de Pression de Maladie: 0.4
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

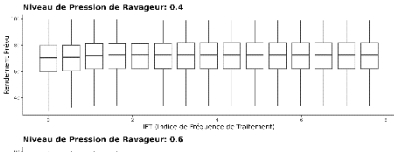

Niveau de Pression de Maladie: 0.6
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

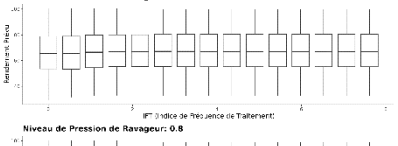

Niveau de Pression de Maladie: 0.8
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

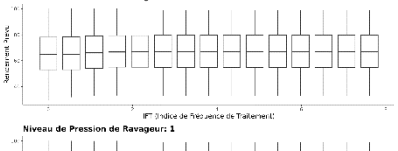

Niveau de Pression de Maladie: 1
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

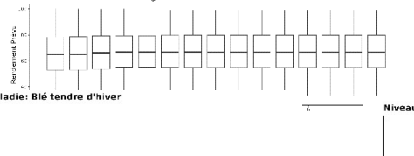

Niveau de Pression de Ravageur: 0
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

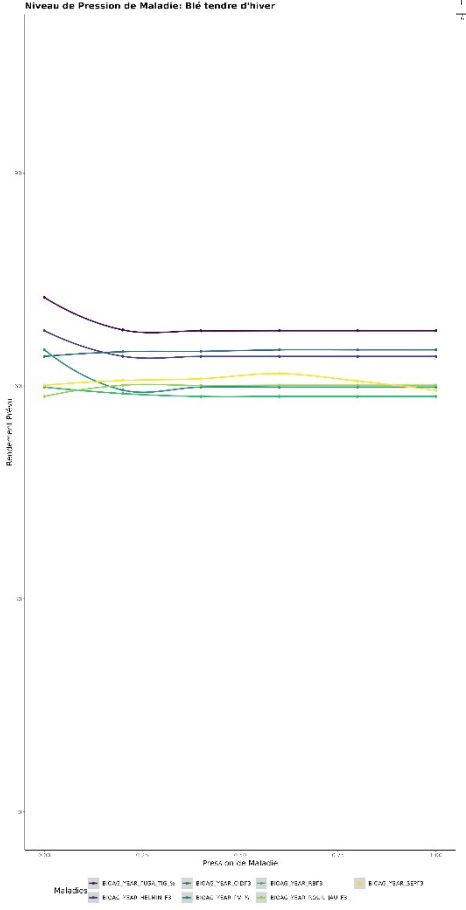

Niveau de Pression de Ravageur: 0.2
Rendement Prévu
IFT (Indice de Fréquence de Traitement)

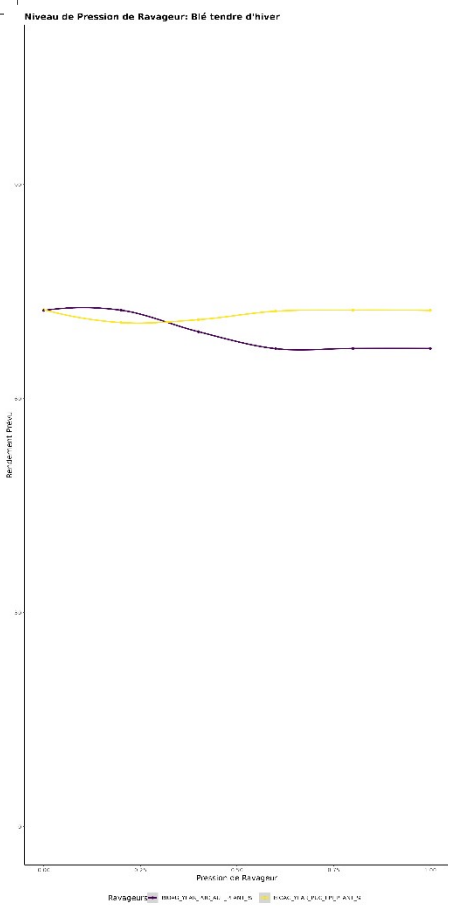

Niveau de Pression de Ravageur: 0.4
Rendement Prévu
IFT (Indice de Fréquence de Traitement)


Niveau de Pression de Ravageur: 0.6
Rendement Prévu
IFT (Indice de Fréquence de Traitement)


Niveau de Pression de Ravageur: 0.8
Rendement Prévu
IFT (Indice de Fréquence de Traitement)


Niveau de Pression de Ravageur: 1
Rendement Prévu


Niveau de Pression de Maladie: Blé tendre d'hiver
Rendement Prévu
Pression de Maladie
Maladies


Niveau de Pression de Ravageur: Blé tendre d'hiver
Rendement Prévu
Pression de Ravageur
Ravageurs

Annexe 5 : Exemple de graphes d’analyses d’une culture (ici blé tendre d’hiver) avec le modèle RF