Parcours Dynamique et modélisation de la biodiversité

MÉMOIRE DE FIN D'ÉTUDES

Master mention Biodiversité Écologie Évolution

Formation Initiale

Approche machine learning pour la prédiction du vol du charançon du bourgeon terminal (*C.picitarsis)* sur les parcelles de colza

Stage réalisé du 01/02 au 31/07/2022

Renan LEHUÉDÉ

Enseignant référent :
François Munoz
Enseignant-Chercheur
Université Grenoble Alpes

Maître de stage :
Quentin Legros
Ingénieur
Terres Inovia

Nombre de pages : 21

Année de soutenance : 2022

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004 ». "This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004".

Résumé

Ceutorhynchus picitarsis, ou charançon du bourgeon terminal (CBT), est un ravageur d'automne des parcelles de colza face auquel les stratégies de luttes sont de plus en plus contraintes. L'historique d'usage systématique des solutions insecticides et la diversité progressivement réduites des produits phytosanitaires disponibles, ont multiplié les résistances. Améliorer les stratégies de lutte constitue alors un enjeu stratégique. Comprendre les conditions qui favorisent la présence du CBT est dans ce contexte, fondamental. Pourtant son écologie est encore mal connue. À partir des bases de données d'épidémiosurveillance agricole, nous proposons d'explorer plusieurs pistes de modélisation pour anticiper l'arrivée du ravageur au champ ainsi que le niveau de risque associé à la parcelle au cours de l'année.

Les variables choisies pour prédire la probabilité d'arrivée du charançon sont principalement météorologiques et paysagères, associées aux coordonnées spatiales et temporelles des points de captures. Le jeu de données, assez volumineux, est très représentatif de ce type d'exercice, c'est-à-dire spatialement hétérogène avec de nombreux prédicteurs corrélés entre eux. C'est sur cette base que nous avons procédé à une comparaison des performances, avant tout prédictives, de quatre types de modèles : fréquentiel, régression pénalisée, *random forest* et *gradient boosting*. Ce sont finalement les modèles basés sur les arbres de décisions, plus complexes qui ont pu fournir les prédictions les plus précises et généralisables. Ils se saisissent plus volontiers que les autres modèles des relations non-linéaires, permettant ainsi d'extrapoler les résultats aux zones géographiques peu couvertes par les captures.

Le modèle de type *gradient boostin*g, du fait d'une plus grande tolérance à la multicolinéarité que le *random forest*, permet des interprétations plus fiables quant à l'influence individuelle des variables sur les prédictions. En effet, un autre objectif de ce projet est de gagner en connaissances sur l'écologie de *C.picitarsis*. Or, les modèles les plus performants sont aussi connus comme « boites noires » pour leur manque d'interprétabilité. Le développement récent de techniques d'exploration agnostiques en termes de méthodes, permet d'estimer l'influence des variables. Ainsi, nous avons effectivement pu mettre en évidence des effets seuils, en particulier liés à la photopériode, qui augmentent la probabilité d'arrivée du charançon. Pour autant, de nombreuses pistes reste à explorer, et d'autres variables que celles étudiées, exercent probablement une influence déterminante sur le risque associé au ravageur.

Abstract

Ceutorhynchus picitarsis, or rape winter stem weevil (RWSW) is a winter oilseed rape's autumn pest with a increasing lack of control methods. The history of insecticide systematical use and the decreasing amount of available plant protection products have multiplied resistances. Enhancing pest control strategies then constitutes a strategic challenge. To understand, which conditions favour RWSW's presence, is in this context, fundamental. Yet, its ecology is still poorly known. From agricultural epidemiological surveillance data bases, we propose to explore several modelling leads to anticipate RWSW's field presence as well as the risk level associated with parcels along the year.

Selected variables to predict the weevil arrival probability are mainly meteorological and landscaped, associated with spatial and temporal capture points coordinates. The dataset, quite voluminous, is very representative of this kind of exercises, meaning spatially heterogenous with many correlated predictors. It's on this basis that we have proceeded to a performance's comparison, above all predictive, of four model types: frequential, penalized regression, random forest and gradient boosting. Decision trees-based models, more complex, have finally gave the most accurate and generalizable predictions. More than other models, they are likely to handle non-linear relations, allowing results extrapolation to geographical areas poorly covered by captures.

Gradient booting model type, with a better multicollinearity resistance than random forest, allow more reliable interpretation of variable individual influence on predictions. Indeed, another objective is to gain knowledge about *C.picitarsis* ecology. Yet, the most performant models are also known as « black box » for their lack of interpretability. Recent development of method agnostic exploratory technics allows to estimate variables influence. Thus, we have highlighted threshold effects, especially linked to photoperiod, that increase the weevil arrival probability. However, several leads remain to be explored, and others than studied variables may have a determining influence on pest associated risks.

Remerciements

Je souhaite en premier lieu remercier Terres Inovia pour m'avoir accueillie dans sa structure. Les conditions matérielles qui ont été mises à ma disposition m'ont permis de réaliser ce stage passionnant dans des conditions de travail confortables.

Je suis particulièrement reconnaissant à Quentin Legros pour son encadrement et son soutien tout au long du stage. Il m'a accordé confiance et autonomie tout en suivant avec curiosité les avancées parfois laborieuses de mon travail. C'est grâce à ces conditions que j'ai pu explorer à loisir les possibilités étonnantes de l'apprentissage machine.

Merci à l'ACTA et à François Brun pour son accompagnement et ses conseils avisés, qui ont permis à plusieurs reprises de recentrer les objectifs de ce projet.

Je remercie François Munoz pour m'avoir fourni les bases pédagogiques nécessaires pour mener ce stage à son terme.

Enfin, je remercie l'ensemble de l'équipe du Rheu pour leur soutien et l'intérêt qu'ils ont montré pour ce projet. Leur accueil chaleureux et les échanges informels ont favorisé, pour le déroulement du stage, un environnement détendu.

Table des matières

Liste des figures

Liste des annexes

Liste des abréviations

ALE : Accumulated local effects

AUC : Area under curve

CART : Classification and regression tree

CBT : Charançon du bourgeon terminal

ETP : Évapotranspiration potentielle

M-plots : Marginal plots

PDP : Partial dependance plot

ROC : Receiver operating characteristic

VIF : Variance inflation factor

1 - Introduction

1.1 - Contexte

Le charançon du bourgeon terminal (CBT, *C.picitarsis*) est longtemps resté un ravageur occasionnel du colza, provoquant des dégâts limités. À partir des années 1970, sa présence s'est généralisée à la plupart des régions de production sur le territoire national (Pilorgué, Maisonneuve, et Ballanger 1997). Bien qu'il soit aujourd'hui considéré comme un des principaux ravageurs du colza, sa biologie reste encore mal connue, tout comme les facteurs déterminants les étapes de son cycle de vie. En effet, son aire de répartition reste limitée à la France et ses pays proches ce qui explique que les efforts de recherche dont il a bénéficié soient particulièrement limités et localisés.

Les adultes arrivent à partir de septembre sur les parcelles de colza pour s'alimenter (Figure 1), se reproduire et pondre une dizaine de jours plus tard dans une cavité sous-épidermique située à la base des pétioles. Au début de l'hiver avec l'arrivée du froid, les larves vont migrer vers le collet des jeunes plants de colza alors au stade rosette. Chaque larve va progressivement creuser une cavité au-dessus du collet jusqu'à se rejoindre (Balachowsky 1963). Le bourgeon terminal a alors de fortes chances d'être détruit, entrainant soit la mort du pied, soit le développement d'un port buissonnant. Dans cette dernière situation, il s'agit d'une levée d'inhibition de bourgeons axillaires produisant des rejets qui resteront mal alimentés, sensibles au gel et peu productifs (Hebinger 2013).

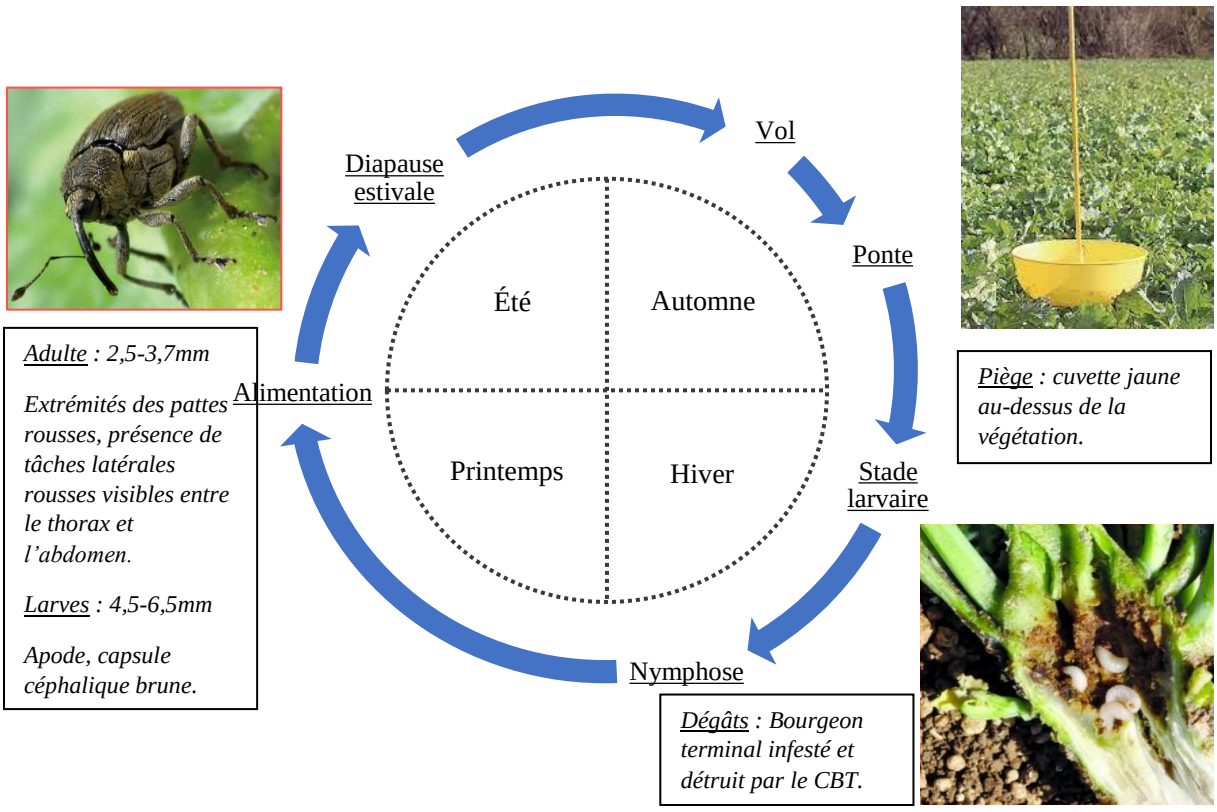


Figure 1 : Cycle de vie de Ceutorhynchus picitarsis. Illustrations et descriptions de l'espèce, des pièges et des dégâts occasionnés.

Au début du printemps, les larves quittent la plante hôte pour se nymphoser dans le sol de la parcelle. C'est au début de l'été que les adultes émergent pour s'alimenter brièvement. Ils sont alors visibles quelques semaines sur les hampes florales, puis entrent en diapause estivale, à proximité des parcelles, probablement dans la litière des sols forestiers jusqu'à l'automne suivant (Pilorgué, Maisonneuve, et Ballanger 1997).

Ce sont donc uniquement les larves qui sont responsables des pertes de rendement. Si elles peuvent être repérées à l'intérieur des plants, l'indicateur le plus efficace reste la capture des adultes en vol lors de leur arrivée sur la parcelle. Aucune relation satisfaisante n'a pu être établie entre le nombre d'individus capturés et les dégâts occasionnés. Aussi la méthode principale de protection des culture, reste l'utilisation d'insecticides, en général de la famille des pyréthrinoïdes (Roâ et Pastre 1990), bien que différents mécanismes de résistance se développent sur le territoire (Robert et al. 2017). Actuellement, la seule alternative est l'emploi du *phosmet*, un organo-phosphoré qui sera prochainement retiré du marché. Il faudra alors se tourner vers d'autres leviers agronomiques : rotations, variétés, auxiliaires, … (Robert et al. 2019; Terres Inovia 2022).

1.2 - Objectifs opérationnels

Terres Inovia, en tant qu'institut technique agricole en charge des filières oléoprotéagineuses, assure le développement des méthodes de protection des cultures de colza. Les contraintes de plus en plus fortes sur l'utilisation des insecticides imposent une forte adaptation des pratiques culturales. Par ailleurs, l'efficacité des mesures contre le CBT étant conditionnées par l'identification de leur date d'arrivée sur les parcelles, la possibilité de l'anticiper et de mieux évaluer le niveau de risque constitue un enjeu stratégique pour la filière. Si la dynamique des insectes une fois dans la parcelle peut être étudiée avec précision, leur comportement à l'extérieur reste peu connu. Or, la diapause estivale se déroule dans d'autres milieux jusqu'à plusieurs kilomètres autour de la parcelle. La sortie de dormance, les facteurs qui la déclenche et par conséquent les évènements qui provoquent le vol sur les parcelles, demeurent inconnus.

On peut considérer la prospection des variables impliquées comme une base de travail pour de futurs projets mieux ciblés. Néanmoins on gardera comme objectif opérationnel, et donc prioritaire la prédiction d'un niveau de « risque charançon » comme outil d'aide à la décision visant à raisonner la protection des cultures. C'est l'apprentissage machine qui semble la méthode la plus indiquée pour prédire au mieux ce phénomène quitte à se confronter au compromis entre capacité prédictive et interprétabilité (Linardatos, Papastefanopoulos, et Kotsiantis 2020) caractéristique des choix entre différentes méthodes de modélisation.

En termes de pistes de travail, bien que la bibliographie soit pauvre, il semble que les variables paysagères soit impliquées dans les variations d'abondance du charançon, en particulier la quantité de forêt dans l'environnement de la parcelle et la proportion de colza, dans l'assolement et au cours de la campagne précédente (Delaune et al. 2021). On sait également que les variables météorologiques sont souvent impliquées dans l'écophysiologie des insectes et notamment d'autres espèces de *Ceutorhynchus,* comme *C.napi* (Debouzie et Wimmer 1992), également ravageur du colza. Enfin, l'historique des pratiques culturales (utilisation d'insecticides, variétés, …) sur la parcelle peut être déterminant, bien que ces informations, souvent confidentielles soient peu accessibles. Nous chercherons à mobiliser des variables explicatives sur ces trois composantes.

Dans ce contexte plusieurs objectifs de modélisation ont été discutés. En cohérence avec le besoin de s'informer sur les pertes de rendement possibles, prédire le nombre d'insectes arrivés sur la parcelle semble pertinent. Bien qu'on ne connaisse pas la relation exacte entre les dégâts et la densité de ravageurs, on pourrait imaginer que de futures études sur le sujet permettraient de relier ces deux éléments. Cela dit, au vu du protocole actuel de capture (Partie 2), on ne peut être certains que le nombre d'individus capturés corresponde avec la densité sur la parcelle, notamment du fait de l'hétérogénéité spatiale intra-parcellaire caractéristique de cette espèce et d'un probable effet dilution avec d'autres parcelles de colza voisines et non suivies. Une autre option de modélisation, plus modeste, mais moins risquée, serait de chercher à prédire uniquement la probabilité quotidienne de présence sur les parcelles. Si on ne résout pas complètement le besoin d'information pour raisonner les interventions, on s'assure néanmoins d'une base de travail robuste, constituant un premier filtre pour identifier des parcelles où le risque serait suffisamment faible pour ne pas être inquiétant. Pour les risques de présence forts, la mise en place de méthodes de lutte dépendra toujours d'une appréciation à dire d'expert des risques agronomiques : densité de culture, historique de la parcelle, stade de développement du colza, … Au regard des données accessibles, du temps disponible et du besoin d'un support fiable, c'est ce dernier objectif de modélisation qui a été adopté, menant à l'élaboration d'une stratégie par étapes (Figure 2). Il permet également de mettre en évidence les variables les plus déterminantes parmi celles disponibles, et de comparer différents types de modèles afin de gagner en expérience, sur ce ravageur mais aussi pour de futurs projets que Terres Inovia serait susceptible de mener.

Étape	Objectif	Description
1	Récolte des données	Extraction des données de différentes sources en fonction des variables d'intérêt.
2	Préparation du jeu de données	Nettoyage de chaque jeu de données (valeurs aberrantes, manquantes, correction des erreurs de remplissage), calcul d'indicateurs supplémentaires, harmonisation dans un format commun. Prétraitement et partitionnement des données.
3	Choix des modèles	Choix d'une gamme de modèles du plus simple au plus complexe.
4	Optimisation des modèles	Choix des hyperparamètres, des variables et optimisation du temps de calcul.
5	Apprentissage/Test	Entrainement des modèles sur un jeu de données d'apprentissage puis calcul de performances sur un jeu de données test.
6	Comparaison des modèles	Calcul d'indicateurs de performance complémentaires et choix du modèle ayant la meilleure capacité prédictive.
7	Interprétation	Analyse de l'influence des variables sur la prédiction.

Figure 2 : Etapes méthodologiques du projet

2 - Matériels & méthodes

2.1 - Protocole de capture

Développée dans les années 50 en Allemagne, la cuvette jaune, facile d'utilisation et efficace, est sur colza la principale méthode de capture des insectes volants. Elle doit être placée au-dessus de la végétation, à plus de 10 mètres d'une bordure sous le vent dominant, et remplie d'eau savonneuse. Les relevés sont à effectuer toutes les semaines à partir de septembre. C'est l'occasion d'ajuster la hauteur de cuvette pour qu'elles restent bien visibles en fonction du niveau de végétation. Les insectes sont filtrés, séchés, identifiés et comptés. Dans le même temps, la cuvette est de nouveau remplie pour poursuivre les captures. Ces informations sont finalement enregistrées en ligne sur une base de données partagée. Ce protocole est principalement réalisé par les experts et techniciens des réseaux agricoles : chambres d'agriculture, coopératives et instituts techniques.

2.2 - Bases de données mobilisées

Vigiculture est un outil mis en place en 2008 par les instituts techniques agricoles comme réseau d'épidémiosurveillance des grandes cultures. L'objectif étant de fournir un état sanitaire permanent du territoire à l'échelle régionale et nationale, en diffusant des Bulletins de santé du végétal à destination des agriculteurs. C'est également une source précieuse de données pluriannuelles qui rassemble l'essentiel des informations sur les captures du CBT, mais aussi quelques informations sur les conditions de la culture au moment des relevés : stade de développement, variété, traitement de semence et éventuels dégâts observés.

À partir de 2009, les chambres d'agriculture et le réseau FREDON (Fédération Nationale de Lutte contre les Organismes Nuisibles), ont conjointement mis en place la plateforme VGObs, également destinée au recensement des bioagresseurs. Pour le CBT, seules les régions Bretagne et Pays de la Loire collectent l'ensemble des données de capture sur cette plateforme.

La jointure de ces bases de données a nécessité un travail d'harmonisation important. En particulier, certaines variables sur l'état des parcelles au moment du relevé étant peu remplies, elles ont été écartées. De plus, la localisation des pièges est définie par le nom de commune sur la base de données VGObs, ce qui a impliqué la recherche de coordonnées GPS associées, nécessaire à l'extraction des données météorologiques. Sur l'ensemble des deux bases de données, environ un quart des observations ont été éliminées lors de la phase de nettoyage.

Les variables météorologiques sont issues des stations Météo France et des instituts techniques Arvalis et Terres Inovia, soit 1497 stations et 779 encore actives. Les données sont extractibles et centralisées via l'outil climbox développé par Arvalis. Il permet également d'accéder à des indicateurs calculés dont l'évapotranspiration potentielle (ETP), les degrés jours à différentes bases, ou le bilan hydrique simplifié (pluie moins ETP). Les données de vent n'étant pas disponibles dans le cadre du contrat d'accès à climbox, elles ont été extraites séparément à partir des données spatialisées SAFRAN (Durand et al. 1993) (maille de 8km) par Météo-France, au point le plus proche du piège. En plus des extrêmes et moyennes sur la semaine de captures, a été comptabilisé le nombre de jours où chaque quantile de la vitesse des vents sur 10 ans ont été dépassés, soit une variable par quantile (Annexe 1).

Connaissant l'importance d'explorer, en phytopathologie, d'éventuels effets retards provoqués par des évènements météorologiques en amont des stades phénologiques étudiés, nous avons ajouté un certain nombre de variables décalées par rapport au jour de capture. Il s'agit en fait d'une adaptation de l'algorithme « Window Pane » (Coakley, Line, et McDaniel 1988; Gouache et al. 2015), où les valeurs, notamment moyennes et extrêmes de septembre sont soustraites à celles de la semaine précédant la capture. D'autres échelles temporelles sont prises en compte, comme la différence entre le début et la fin de semaine (Annexe 1). Au total, 55 variables climatiques ont été exploitées.

Au vu des suggestion de la littérature sur l'importance des variables paysagères (Delaune et al. 2021) les surfaces de colza de l'année n et n-1, ainsi que les surfaces de forêt autour des pièges ont été prises en compte à différents rayons : 100, 200, 300, 500, 1000, 2000, 3000 et 8000 mètres. Prendre plusieurs rayons permettra éventuellement de gagner en connaissances sur la capacité de déplacement ou migration du CBT. Ces valeurs ont été calculées à partir du registre parcellaire graphique et de la BD TOPO (https://geoservices.ign.fr/bdtopo) accessibles publiquement, avec une valeur commune de surface de forêt pour toutes les années confondues.

2.3 - Caractéristiques du jeu de données

À l'amont de l'exercice de modélisation, un certain nombre d'informations peuvent être examinées à partir des caractéristiques du jeu de données. Il est composé de 62 854 observations pour 9029 exercices d'observation, c'est-à-dire un même lieu et une même série de captures, généralement répartie de la fin de l'été à la fin de l'hiver. La plupart des relevés mettent en évidence des pièges vides menant à un déséquilibre de classes absences/présences. Dans certains cas, on ne peut être certains que le premier jour d'arrivée ait bien été identifié si le premier relevé de piège de l'exercice d'observation révèle la présence du ravageur (Figure 3).

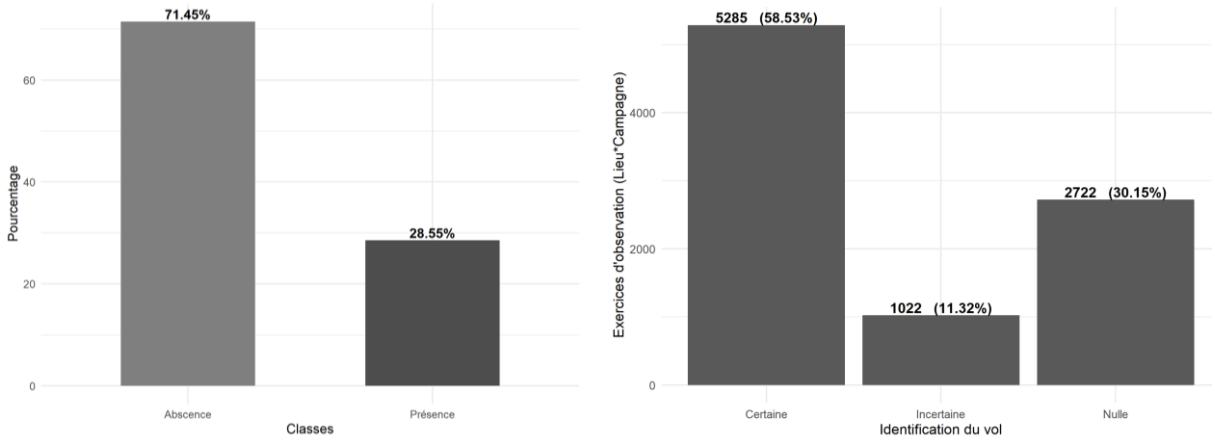


Figure 3 : Description des données de captures. À gauche, la répartition des classes absences et présences. À droite, capacité des exercices de suivi à identifier le premier jour d'arrivée (« Nulle » : aucun insecte capturé).

En prenant en compte les jours de l'année avec plus de 40 observations, on remarque que les suivis sont concentrés majoritairement sur la période de vol du ravageur, soit en automne, et s'étendent jusqu'au début de l'hiver, prenant en compte de rares vols tardifs. Sur ce même intervalle, on peut remarquer une diminution cyclique de l'effort de capture. En effet, les pièges sont relevés une fois par semaine, et souvent peu le weekend (Figure 5). Outre 2008 la première année de mise en place d'un réseau national, et le début d'année 2022, le nombre total d'observations reste assez homogène sur l'ensemble des années disponibles (Figure 4).

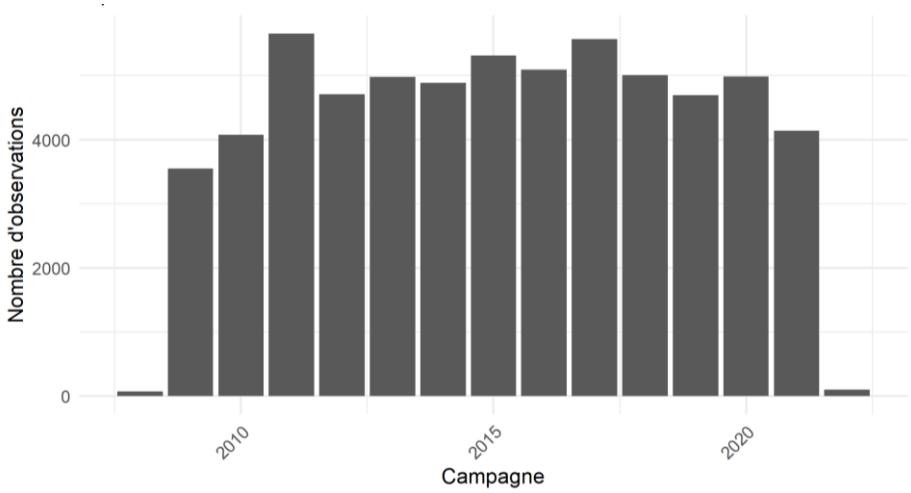


Figure 4 : Effort de capture par campagne de culture

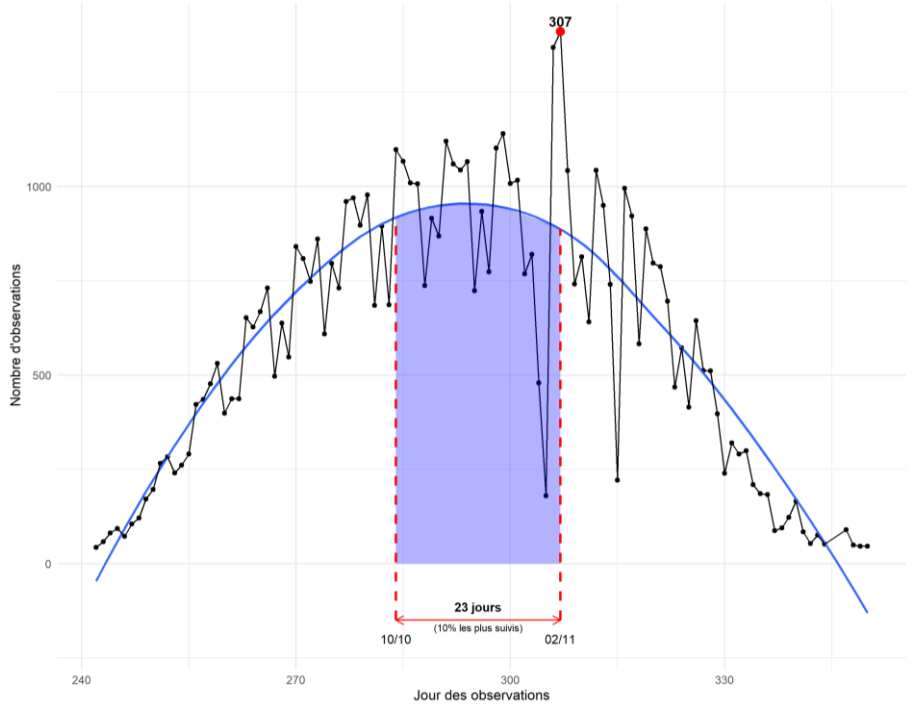


Figure 5 : Principale période de suivi (10% en bleu). Seuls les jours de l'année avec un minimum de 40 observations sur toutes les années ont été pris en compte.

L'effort de capture est spatialement très hétérogène. Non seulement, certaines zones du territoire comptent peu de surface de colza, notamment en montagne, mais d'autres sont encore peu inquiétées par ce ravageur, comme en Bretagne, et participent donc moins à la mise en place de suivis (Figure 6). Cela dit, quelques soit les années la proportion d'observation communiquées par chaque région reste stable (Annexe 2).

Enfin, la date moyenne d'arrivée varie légèrement selon les années. Néanmoins, on retrouve des schémas communs sur l'ensemble du jeu de données quant à la répartition spatiales des dates d'arrivées (Figure 6). Nous avons schématisé ce phénomène en procédant à une modélisation linéaire des dates d'arrivée en fonction des coordonnées géographiques des point de capture qui a mis en évidence un

gradient Sud-Ouest / Nord-Est très net. On peut supposer que les climats plus continentaux, avec une chute des températures plus précoce au cours de l'automne, favorise l'arrivée anticipée du CBT. Cet argument serait en faveur d'une implication forte des variables climatique dans l'enchainement des stades de vie du charançon.

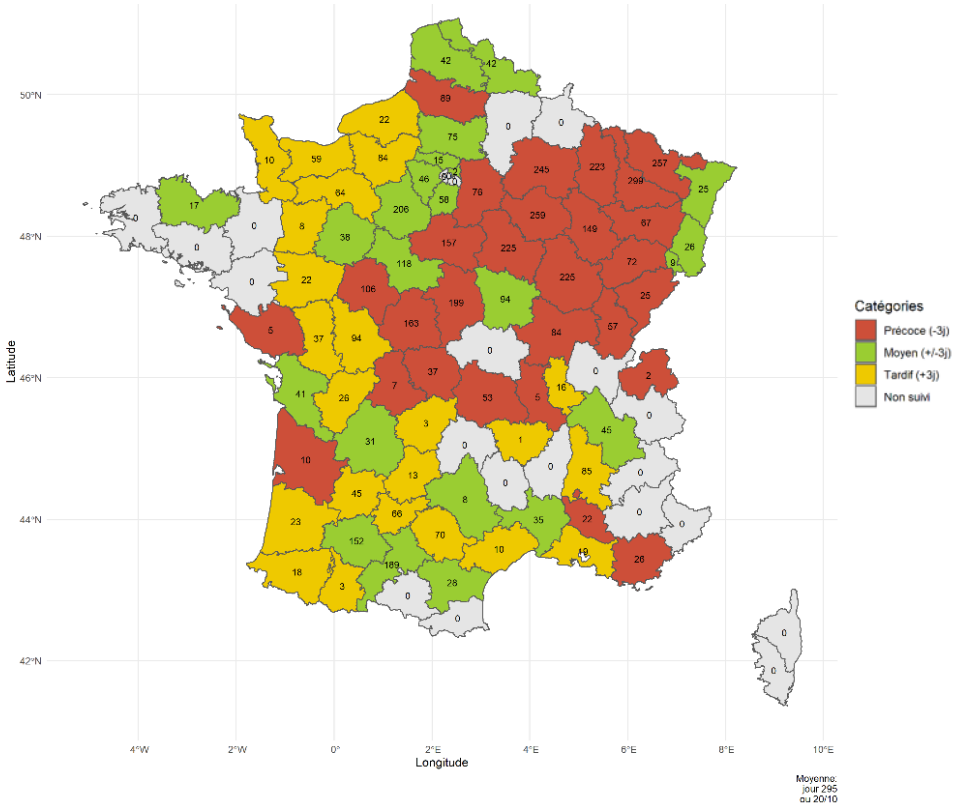


Figure 6 : Caractéristiques locales du vol (précocité) et nombre d'exercices de suivi par département.

2.4 - Types de modélisation à comparer

Quatre types de modèles d'apprentissage machine ont été comparés, notamment pour évaluer si le gain attendu en capacités prédictives avec des modèles plus complexes est suffisamment intéressant pour accepter de perdre en interprétabilité. Outre la structure propre des modèles plus ou moins hermétique à l'interprétation, la gestion des corrélations fortes inhérentes au jeu de données constitue également un enjeu pour expliquer le modèle qui sera sélectionné.

Le premier modèle testé est le plus simple possible et servira donc de référence. Il s'agit d'un modèle fréquentiel, c'est-à-dire que la probabilité, par département, que l'insecte soit présent sur la parcelle à une certaine date est calculée comme un pourcentage des observations de présence sur toutes les années confondues. Comme aucune autre variable que l'information de capture et le jour ne sont implémentées dans ce modèle naïf, certains indicateurs de performance, ou d'importance relative des variables ne sont pas disponibles. Il ne sert donc qu'à évaluer le potentiel gain en capacité prédictive.

Le second modèle testé est une régression logistique pénalisée de type ElasticNet. Cette méthode combine les deux termes de pénalisation des régressions Ridge et Lasso (Zou et Hastie 2005). La pénalisation permet de gérer les corrélations fortes entre les variables d'entrée tout en gardant la possibilité d'appliquer des méthodes d'amélioration et d'interprétation du modèle (Makowski et al. 2021). Les hyperparamètres à déterminer sont les coefficients de pénalisation pour chacun des termes (Ridge et Lasso), et le rapport entre les deux. L'importance relative des variables dépend de la valeur des coefficients estimés.

Deux modèles plus complexes, *random forest* et *gradient boosting*, réputés performants et résistants aux colinéarités entre variables, ont également été sélectionnés comme modèles candidats. Ils sont tous deux basés sur l'assemblage d'arbres de décisions. Cette structure particulière les rend difficiles à interpréter.

La technique des forêts aléatoire ou *random forest* (Breiman 2001), consiste à sélectionner une partie des observations et des variables, par un tirage aléatoire avec remise (*bootstrap*). Un arbre de décision de type CART est ensuite construit sur ce jeu de données synthétique. L'opération est répétée un grand nombre de fois avant d'agréger l'ensemble des arbres pour repérer les règles de décisions les plus représentées (*bagging*). Le g*radient boosting* (Friedman 2002) s'appuie sur le même fonctionnement mais les arbres n'ont qu'un niveau de profondeur et sont construits en série pour que chaque nouvel arbre tente d'expliquer les résidus des arbres précédents, améliorant ainsi les performances globales. Il n'est donc pas possible dans ce cas de paralléliser l'élaboration des arbres. Plusieurs hyperparamètres sont disponibles, notamment le nombre d'arbres, le nombre de variables tirées aléatoirement pour chaque arbre, et le critère déterminant quelle variable à chaque niveau de l'arbre, sera sélectionnée pour discriminer au mieux les observations. Ils n'ont pas été choisis a priori mais explorés afin de trouver la meilleure combinaison pour maximiser la capacité prédictive. L'évaluation de l'importance relative des variables dépend d'une gamme de méthodes propres à ce type de modèles. Dans notre cas, elle a été effectuée par permutation (Molnar 2022). Par randomisation des valeurs, on considère l'importance de la variable selon l'effet sur l'erreur de prédiction du modèle. Cette méthode, parmi les autres disponibles, est réputée plus robuste mais plus exigeante en temps de calcul.

2.5 - Indicateurs de performances & interprétabilité

L'objectif principal étant l'élaboration du modèle ayant la meilleure capacité prédictive avec une variable réponse binaire, le choix du modèle s'est avant tout basé sur l'AUC. Il s'agit de l'aire sous la courbe de ROC, calculée sur les sensibilités et spécificités du modèle lorsque le seuil de probabilité définissant la présence ou l'absence varie. C'est donc le critère de sélection, du modèle le plus performant, à chaque étape de modélisation.

Avec un effort de capture spatialement très hétérogène, on peut raisonnablement s'attendre à une capacité de prédiction également différenciée par aire géographique. Aussi, il est apparu pertinent d'analyser la performance par département, qui est l'échelle territoriale la plus précise pour qu'un tel partitionnement laisse des jeux de données suffisamment volumineux à comparer.

En termes d'interprétabilité, trois méthodes ont été déployées. En premier lieu, l'importance relative des variables a été calculée à l'aide des méthodes appropriées pour chaque type de modèle (coefficients ou permutation). Elle sert avant tout à alléger les modèles en ne sélectionnant que le nombre de variables, dans leur ordre d'importance, permettant de maximiser l'AUC. Cependant, nous avons pu remarquer que l'ordre d'importance des variables est particulièrement instable, notamment du fait des corrélations parfois extrêmement fortes entre certains prédicteurs formant des groupes difficiles à dissocier. En acceptant une moindre finesse d'analyse, on peut observer l'importance relative des groupes plutôt que des variables isolées.

Mesurer l'impact de la multicolinéarité des prédicteurs sélectionnés a donc été nécessaire, au moins pour quantifier la fiabilité des mesures d'importances. Si plusieurs méthodes existent, nous nous sommes tournés vers un indice de multicolinéarité, le VIF ou *Variance Inflation Factor*, calculé après l'étape de sélection de variables. Il s'agit en réalité d'un score attribué à chacun des prédicteurs et correspondant à la diagonale de la matrice de corrélation inversée.

Enfin, pour accéder à l'effet de chaque variable sur les prédictions, nous avons mobilisé la méthode *Accumulated Local Effects* (ALE) (Apley et Zhu 2020), adapté à tout type de modèle et plus résistante aux problèmes de multicolinéarité que les graphiques de dépendance partielles (PDP) ou les

graphiques marginaux (M-plots). Elle repose sur le même principe, c'est-à-dire qu'il s'agit de tester la réaction du modèle en changeant les valeurs de la variable d'intérêt. Contrairement au PDP, seules des valeurs réalistes au regard du jeu de données sont testées, ce qui respecte la distribution conditionnelle des prédicteurs. C'est également le cas du M-plot mais l'ALE mesure l'effet sur les prédictions ce qui évite de prendre en compte l'effet d'autres prédicteurs qui seraient corrélées. Il s'agit donc de combiner les avantages des deux méthodes tout en évitant d'être affecté par la multicolinéarité des variables.

2.6 - Protocole de modélisation

Préparation des données : Dans un premier temps les données ont été centrées et réduites pour faciliter leur utilisation ultérieure. Les informations encore manquantes ont été complétées par la méthode des K plus proches voisins. Puis, les données ont été partitionnées en jeu d'apprentissage, de validation et de test (méthode holdout), pour éviter les risques de surapprentissage et de surestimation des performances au cours des différentes étapes de modélisation. Les jeux de validation et de test représentent chacun les données de deux années distinctes.

Estimation de l'importance des variables : Un premier modèle de référence est produit, entrainé sur le jeu d'apprentissage avec une graine aléatoire fixée à 42. Pour estimer l'importance relative des variables, 14 autres modèles ont été produits avec une graine aléatoire non fixée afin de produire un ordre moyen d'importance des prédicteurs. En effet, la structure inhérente aux modèles implique généralement une fluctuation aléatoire des variables les plus influentes.

Optimisation du modèle : Il s'agit de comparer différents sous-modèles candidats et de sélectionner le plus performant en fonction de son AUC. De cette manière, différents sous-ensembles des n plus importantes variables sont testés, et pour chacun, les hyperparamètres les plus adaptés sont sélectionnés. Cette méthode permet de filtrer les variables qui n'apportent pas d'information pertinente, rendant le modèle inutilement complexe. Cette étape se décompose en deux parties. Dans un premier temps chacun des 17 jeux de variables sont modélisés pour de multiples combinaisons où 3 valeurs sont possibles pour chaque hyperparamètre. Les performances sont évaluées en testant chacun de ces sous-modèles sur le jeu de données de validation. Dans un second temps, les 17 modèles candidats avec leurs hyperparamètres sélectionnés sont de nouveau entrainés grâce au jeu d'apprentissage fusionné avec le jeu de validation, en validation croisée par lot de deux années. Le modèle final prend ainsi en compte la partie des variables ayant le plus d'importance et les meilleurs hyperparamètres associés.

Evaluation des performances : Le modèle candidat est confronté au jeu de données test pour vérifier sa capacité prédictive en calculant l'AUC qui en résulte. Les autres tests sont ensuite effectués sur ce modèle : AUC par département, VIF et ALE. Ces étapes sont répétées pour chaque type de modèle à comparer.

L'ensemble des interventions sur le jeu de données a été réalisée sur RStudio 2021.09.2 comme interface du logiciel R version 3.6.3 (2020-02-29) : contrainte de version imposée par ClimBox, package utilisé pour l'extraction des données météorologiques. Le package *caret* a été majoritairement mobilisé pour les étapes de modélisation.

3 - Résultats

3.1 - Capacité prédictive

L'aire sous la courbe de ROC (AUC) est le principal critère d'évaluation des modèles. Il n'y a pas de règles exactes pour interpréter ces valeurs, mais on peut estimer la qualité des prédictions à partir des seuils proposés en Figure 7. Dans notre cas, on considèrera surtout la qualité du modèle en fonction du différentiel d'AUC avec le modèle naïf.

Modèles	**AUC**
Fréquentiel	0,737
Régression *elastic net*	0,800
Random forest	0,840
Gradient boosting	0,845

Figure 7 : Aire sous la courbe de ROC des modèles comparés

Les modèles basés sur les arbres de décisions, considérés comme des modèles « boites noires » sont les plus adaptés sur ce type de modélisation, c'est-à-dire pour prédire des relations réponses-prédicteurs non-linéaires avec un choix de variables naïf. Leur capacité prédictive globale est supérieure aux autres modèles. Bien que les modèles fréquentiel et *elastic net* aient une AUC assez satisfaisante, leurs performances par département (Figure 8) est plus hétérogène avec quelques valeurs inquiétantes, inférieures à 0,5. Comme pour l'AUC globale, l'hétérogénéité spatiale des prédictions du *random forest* et *gradient boosting* ne montre pas d'écart significatif.

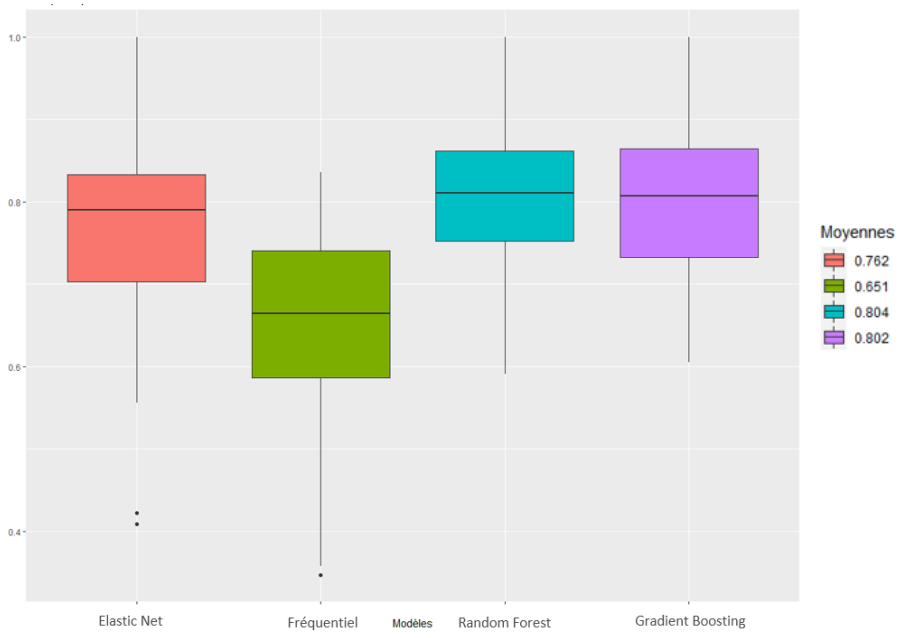


Figure 8 : Comparaison de l'hétérogénéité spatiale des modèles (boxplot)

En observant, la spatialisation des AUC par département et notamment leurs différences entre celles du *gradient boosting* et du modèle fréquentiel (Figure 9), on retrouve effectivement cet écart de performances. Toutefois, on peut remarquer que la qualité des prédictions peut aussi être associé à la répartition de l'effort de capture (Figure 6).

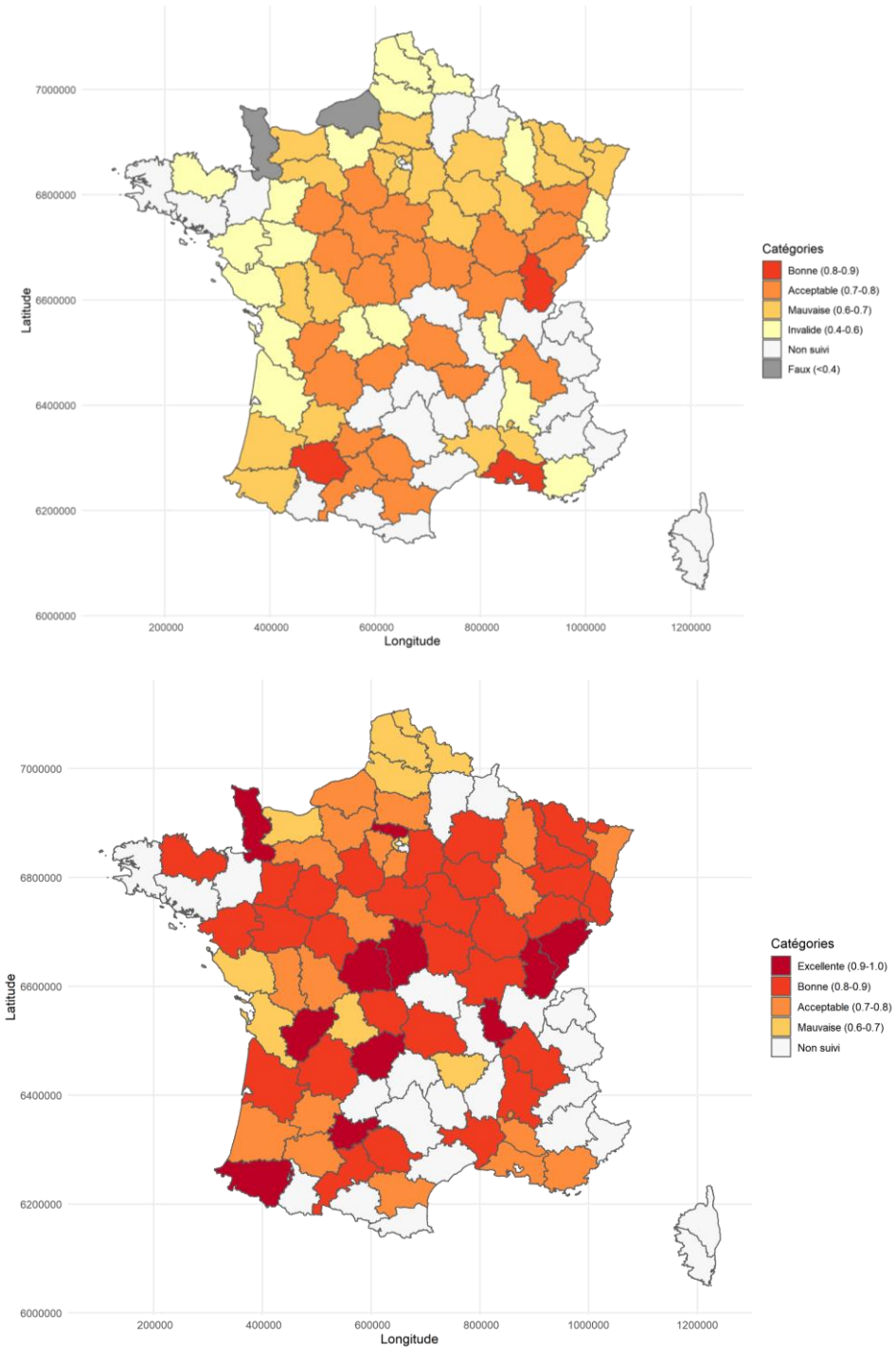


Figure 9 : Répartition départementale de l'AUC pour les modèles fréquentiel et gradient boosting

On retiendra le *gradient boosting* comme meilleur modèle prédictif bien que ses capacités soient à peine supérieures à celles du *random forest*, voire légèrement inférieures dans le cas de la variabilité spatiale. Un autre argument en faveur de ce choix est la différence non négligeable en temps de calcul lors de la modélisation.

3.2 - Importance des variables et multicolinéarité

Les étapes de sélection de variables permettent de réduire la complexité des modèles et d'éliminer les variables qui n'améliorent pas l'AUC. Cette méthode est fortement dépendante de la manière dont chaque type de modèle estime l'importance relative des prédicteurs. Certaines variables étant très corrélées entre elles (Figure 10), leurs effets sur les prédictions peuvent être difficilement dissociables. Plus le modèle arrive à gérer la multicolinéarité, moins l'ordre d'importance relative des prédicteurs est aléatoirement déterminé. Dans le cas d'une méthode peu adaptée à cette situation, le processus de modélisation aura tendance à surestimer une ou quelques variables de chaque groupe et sous-estimer les autres. Il n'y a aucun impact sur la prédiction mais cela pose un problème majeur pour l'interprétation puisque la fiabilité de l'ordre d'importance des variables peut devenir particulièrement incertain. Il est utile de préciser que cette instabilité a été en partie limitée puisque l'ordre des importances a été estimé en moyennant celui de 15 modèles avec prédicteurs et hyperparamètres identiques. Le modèle fréquentiel ne prenant pas en compte d'autres prédicteurs que le jour et la probabilité historique de présence, il n'est pas pris en compte dans ces analyses.

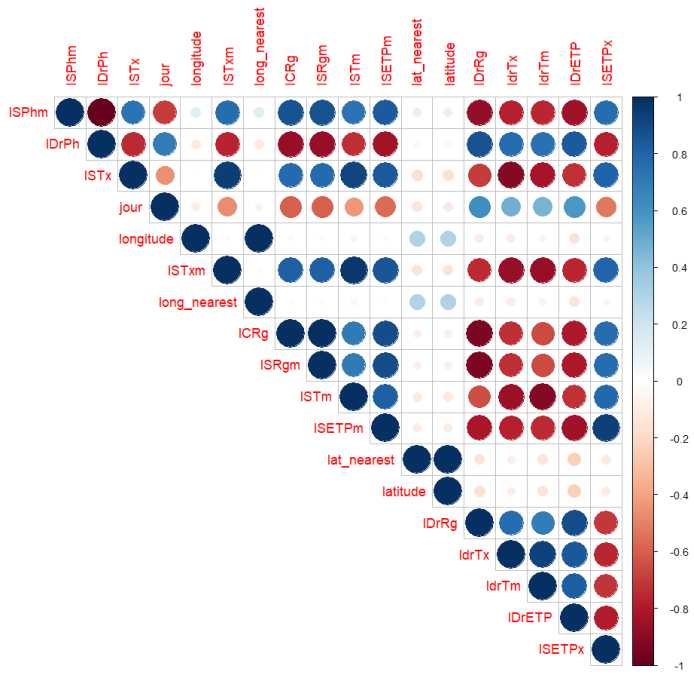


Figure 10 : Matrice de corrélation des variables sélectionnées pour le modèle gradient boosting

On peut observer que sur la matrice de corrélation des variables sélectionnées lors de la modélisation de type *gradient boosting*, températures, photopériodes, rayonnement, ensoleillement et évapotranspiration potentielle sont toutes très corrélées entre elles. On retrouve ce phénomène pour tous les modèles. Cela dit, il faut aller plus loin dans l'analyse pour estimer à quel point l'estimation de l'importance des prédicteurs finalement utilisés est biaisée par l'effet des multicolinéarités. Le VIF permet d'estimer ce phénomène. Si le modèle gère bien les multicolinéarités il ne sélectionnera qu'un prédicteur par groupe de variables. À l'inverse, certains prédicteurs absorberont l'importances des autres. On peut aussi appeler cet effet, facteur d'inflation de la variance ou *Variance Inflation Factor* (VIF). Il n'y a pas de consensus clair dans la littérature mais selon les sources, on considère qu'à partir d'un VIF de 10, l'importance de la variable n'est pas fiable.

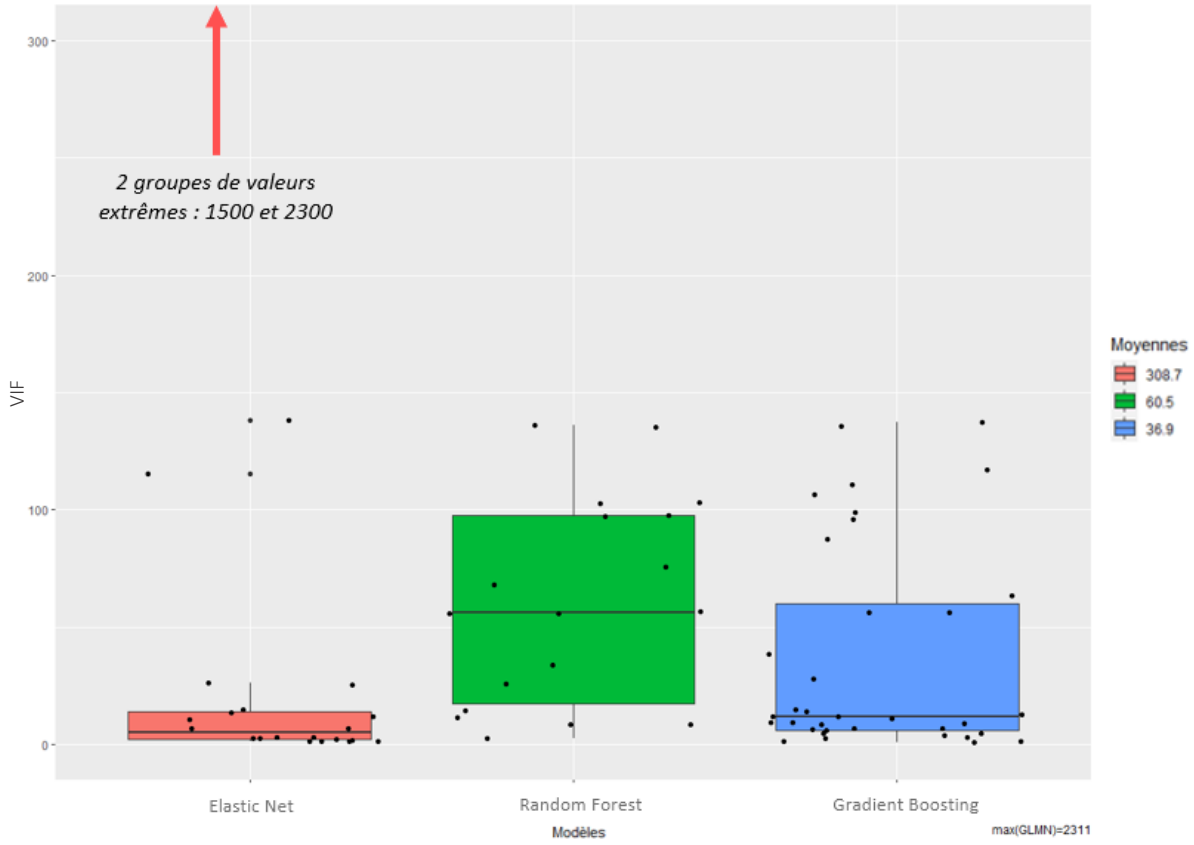


Figure 11 : Variance Inflation Factor (VIF) associés aux prédicteurs sélectionnés pour chaque modèle (boxplot)

Sur la Figure 11, pour des raisons d'échelle les valeurs extrêmes autour de 1500 et 3000 de quelques prédicteurs sélectionnés dans la procédure de modélisation *elastic net* ne sont pas représentés sur le graphique. Bien que la littérature suggère souvent que les *random forest* et *gradient boosting* sont particulièrement résistants voire immunisés contre les problèmes de multicolinéarité, ce n'est pas le cas. La moyenne des VIF étant nettement au-dessus du seuil de vigilance (10), on peut effectivement en déduire qu'ils n'ont pas pu complètement gérer la multicolinéarité. Le *gradient boosting* étant dans ce cas plus adapté que le *random forest*. Toutefois, la régression, même pénalisée, est complètement dépassée, quelques variables ayant absorbé de manière impressionnante l'importance des autres. Et bien que l'écart interquartile soit faible, la moyenne reste largement supérieure aux autres modèles.

Prendre en compte non pas l'importance individuelle des variables mais plutôt celle des différents groupes de prédicteurs corrélés entre eux permet de limiter le biais d'interprétation. Le principal groupe est constitué de plusieurs types de variables météorologiques. Son nom est ici simplifié par « ETP » (Figure 12). Quels que soit les modèles il apparait très largement majoritaire. On remarquera toutefois que les différentes méthodes de modélisation ne sont pas toutes aussi conservatrices, ce qui a un impact sur la diversité des groupes représentés. Dans le cas du *random forest* avec seulement 18 variables, seuls le groupe ETP, jour et coordonnées sont présents. Les variables paysagères et liées au vent apparaissent avec l'*elastic net* qui a conservé 25 variables. Enfin, le *gradient boosting*, avec 35 variables semble d'avantage valoriser l'effet marginal des groupes habituellement peu représentés au détriment du groupe ETP. Il reste malgré tout majoritaire.

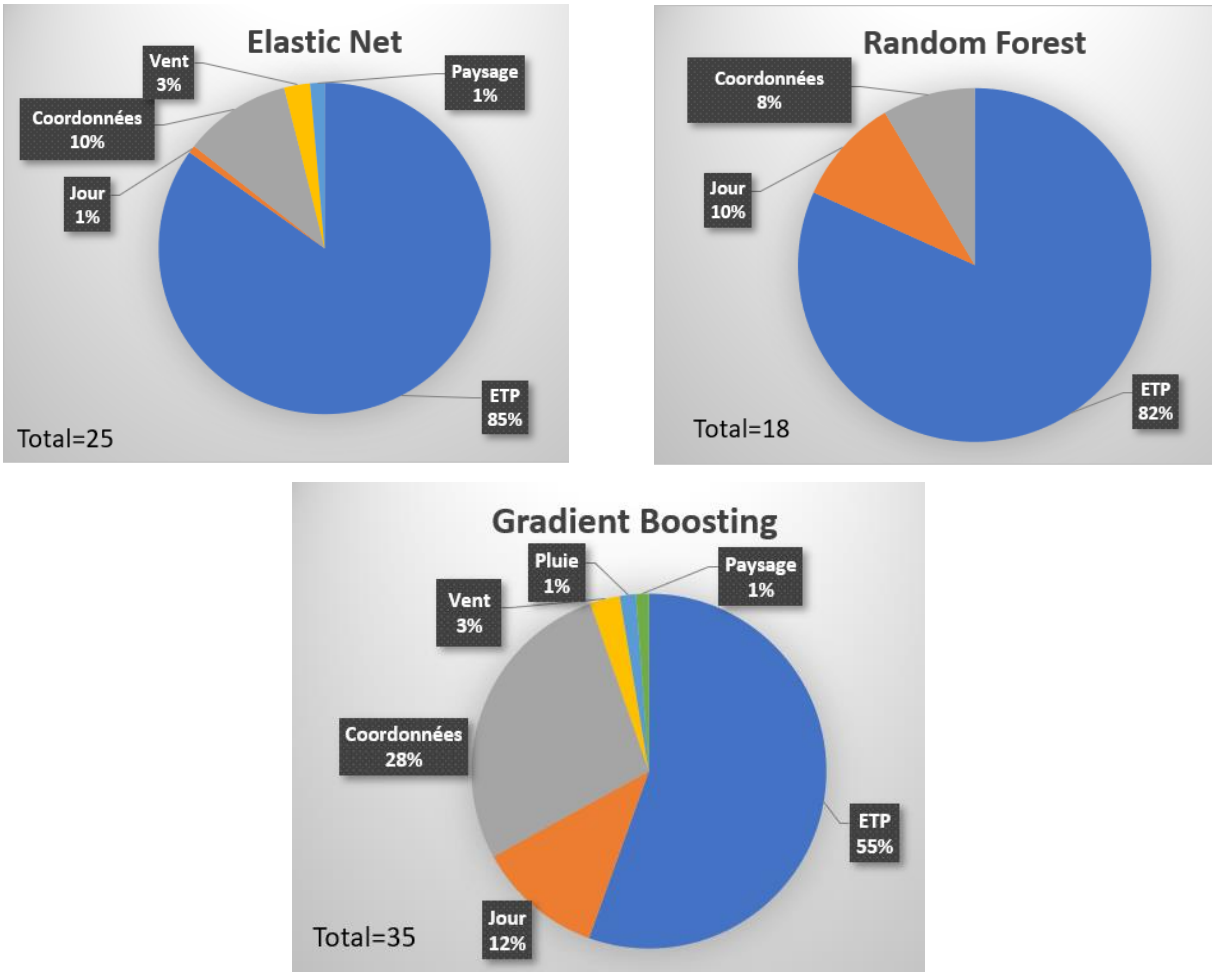


Figure 12 : Estimation de l'importance des variables par groupe

3.3 - Influence individuelle des prédicteurs

La méthode d'*accumulated local effects* (ALE) offre un aperçu de l'influence individuelle des variables sur les prédictions. Tout en restant biaisée par les caractéristiques du jeu de données et du fonctionnement des différents modèles, elle apporte des informations sur l'effet réel des variables utilisées. Les graphiques pour chaque prédicteur sont accompagnés d'une courbe de densité représentant la répartition des observations sur l'axe des valeurs de la variable. Tous les graphiques sont disponibles en annexe 3 à 5, seules les quatre plus importantes sont représentées ci-dessous pour les modèles *random forest* et *gradient boosting*, plus fiables que la régression *elastic net*.

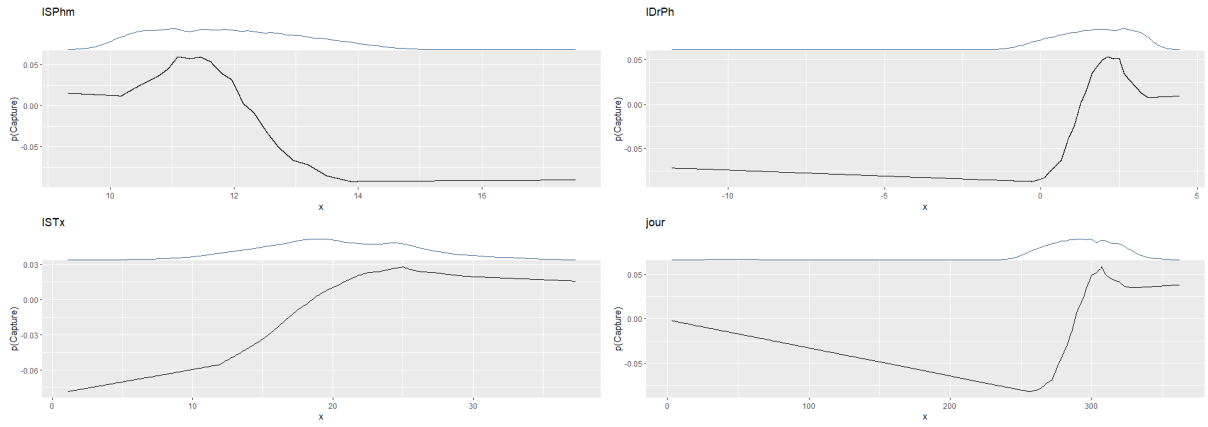


Figure 13 : ALE des variables les plus importantes du modèle random forest (en noir) associée à la courbe de densité des observations (en bleu).

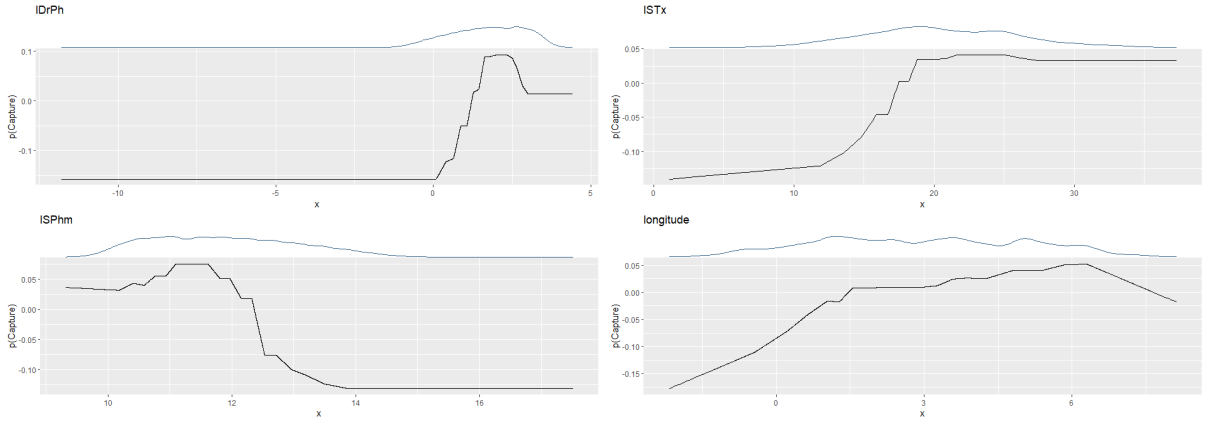


Figure 14 : ALE des variables les plus importantes du modèle gradient boosting (en noir) associée à la courbe de densité des observations (en bleu).

Les courbes de densités permettent d'évaluer les gammes de valeurs dont l'influence sur les prédictions sont les plus fiables. En effet, on peut considérer qu'aux valeurs extrêmes, le peu d'observations existantes sont insuffisantes pour bien représenter ces situations. On constate que la photopériode arrive en premier ordre d'importance pour les deux modèles, ainsi que la température maximale sur la semaine précédant la capture. La température semble plutôt agir comme un facteur limitant la probabilité de capture jusqu'à un certain seuil d'environ 17°C. La variable IDrPh est une différence de photopériode par rapport à celle de septembre. Elle peut donc prendre des valeurs négatives et se comporte inversement à ISPhm, la photopériode moyenne sur la semaine précédant la capture. Elle ne favorise les vols que sur une courte période de l'année, entre 10 et 12h. Considérant que la photopériode est d'environ 12h en septembre, c'est sur les mois qui suivent que les vols se déclenchent. On retrouve cet effet avec IDrPh, qui favorise les vols seulement lorsque l'écart de photopériode avec la moyenne de septembre se creuse. La photopériode suit un cycle annuel, on retrouve donc la même tendance sur les jours de l'année.

On constate également que l'influence des variables est ici interprétée comme beaucoup moins linéaire qu'avec la régression *elastic net* (Annexe 3). Bien que les comportements apparaissent comme similaire entre le modèle *random forest* et *gradient boosting*, ce dernier met davantage en évidence d'éventuels effets palier avec des courbes moins lissées.

D'autres variables environnementales suivent un cycle annuel. Soit, elles sont effectivement dans le jeu de données et leur influence est faible, soit elles n'ont pas été testées au cours de ce projet. Dans tous les cas, il pourrait être hasardeux de complètement dissocier l'effet de la photopériode et du jour de l'année. Ce dernier prédicteur peut en cacher d'autres. De la même manière que la longitude (Figure 14), il est évident que le CBT n'est pas sensible à ses coordonnées géographiques mais bien à des variables spatialisées qui n'ont pas été modélisées ici. On considèrera plutôt ces variables comme des pistes de réflexions, que comme des preuves de leur influence sur le cycle de vie du charançon.

4 - Discussion

4.1 - Résultats

Les modèles les plus simples ont l'avantage d'être peu consommateur en données, facilement manipulables et adaptés à une large gamme de méthodes d'analyse. Cependant, nous avons pu montrer que les modèles plus complexes basés sur l'agrégation d'arbres de décision sont nettement plus performants pour prédire la probabilité de capture de *C.picitarsis*, ce qui constitue l'objectif premier du projet de stage. Ils sont plus polyvalents, notamment parce qu'ils s'adaptent mieux aux relations non - linéaires entre les variables prédictives et de réponse. Même dans les départements où peu de données sont disponibles, les prédictions restent acceptables, ce qui suggère que ces méthodes permettent d'appréhender des relations prédicteurs-réponse plus généralisables qu'un simple modèle fréquentiel, plus sensible au sur-apprentissage.

Les méthodes de régression sont réputées pour offrir des possibilités d'interprétations bien supérieures aux modèles plus complexes d'apprentissage machine. Pourtant, nous avons pu constater que leur incapacité à gérer la multicolinéarité peut constituer un véritable obstacle dans l'évaluation de l'importance relative et de l'influence individuelle des variables sur les prédictions. C'est pourtant une caractéristique généralement inhérente aux jeux de données mobilisés pour analyser les dynamiques liées à des variables environnementales. En effet, les méthodes supposant une indépendance des variables sont en pratique peu applicables à des contextes réels. Les éventuelles transformations, alors nécessaires, comme la création de variables synthétiques (PCR, PLSR, …) (Makowski et al. 2021) ou les sélections basées sur des critères de corrélation contraignent quoi qu'il en soit l'exploration des relations prédicteurs-réponses. C'est d'autant plus le cas, lorsqu'il n'y a pas lieu d'effectuer des choix a priori sur l'importance des variables.

Les résultats de l'ALE, sont globalement cohérents. Si l'évolution de la photopériode est en pratique très proche de celle des jours de l'année sur la période la plus suivie, celle de la température est plus relative, ce qui suggère qu'il existe bien une différenciation d'effets avec le cycle annuel. La prise de position récurrente des variables de coordonnées spatiales et temporelles parmi les plus importantes est à interpréter avec prudence mais suggère qu'il existe une partie de la variance des captures qui n'a pas été ici expliquée. L'effet de ces prédicteurs ne permet pas d'apporter des conclusions robustes sur les dynamiques sous-jacentes mais invite à l'exploration de variables qui n'auraient pas été prises en compte. Leur identification pourrait s'avérer précieuse pour améliorer les méthodes de luttes sur le terrain. Il convient néanmoins de rappeler que si le modèle fréquentiel (naïf) donne généralement des résultats acceptables, cela prouve que l'association des variables jour/capture apporte en soit une partie de l'information qui n'est donc pas nécessairement imputable aux autres variables utilisées.

Bien que le *random forest* affiche de très bonnes performances et une tolérance satisfaisante par rapport à l'hétérogénéité des données, il est nettement moins résistant à la multicolinéarité que le *gradient boosting* et de fait, moins adapté à l'analyse exploratoire. Le gain en temps de calcul est également déterminant dans le choix final de conserver le modèle de *gradient boosting*. Les effets paliers sont plus volontiers mis en évidence. S'ils ne sont pas nécessairement représentatifs des effets réels, ils favorisent l'interprétation visuelle et la conversion en outil d'aide à la décision pour des applications très concrètes. Effectivement, pour les acteurs de terrains, il s'agit avant tout d'être capable de définir des situations à risques variables. Dans ce contexte, les effets paliers peuvent être compris comme tels. Lorsqu'une variable particulièrement déterminante atteint une gamme de valeurs, la probabilité de vol augmente, invitant ainsi les utilisateurs à orienter leurs choix techniques.

4.2 - Limites

Cet exercice, comme tout travail de modélisation est tributaire des hypothèses de départ et des méthodes mobilisées, ainsi que de la qualité des jeux de données disponibles. En effet, les relevés de captures ne sont pas suffisamment associés à des évaluations de pertes de rendements sur la parcelle. Partant du principe que les dégâts occasionnés par le CBT sont a priori peu corrélés au nombre d'individus capturés, le principal critère de mise en place des méthodes de lutte reste la présence ou non du ravageur au champ. En se concentrant sur cet évènement, on limite les ambitions des exercices de modélisation. S'il n'est pas garanti qu'ils puissent apporter une solution robuste, ils pourraient certainement consolider les connaissances actuelles. Un effort de capitalisation des connaissances empiriques serait ainsi susceptible d'alimenter les choix stratégiques pour consolider l'offre d'outils d'aide à la décision.

Les informations sur les pratiques agronomiques sont peu disponibles, et n'ont pas permis de se saisir de la réalité de terrain qui leur est associée. Certaines informations sur la culture en place, notamment la variété, le stade de développement ou l'éventuel traitement des semences, sont en principe disponibles sur la base de données vigiculture mais en pratique particulièrement incomplètes et donc peu exploitables. Par ailleurs, il aurait été intéressant de prendre en compte l'historique récent d'utilisation d'insecticides mais ces informations, considérées comme sensibles, demeurent difficilement accessibles. Au-delà d'une potentielle amélioration des prédictions, s'informer sur l'influence des pratiques agricoles sur la probabilité de présence du charançon, serait un complément intéressant pour l'évaluation des risques au champ.

La stratégie mise en place pour manipuler des groupes de variables fortement corrélées entre elles pourrait être étoffée. L'élimination des prédicteurs n'améliorant pas la qualité des prédictions, a effectivement permis d'alléger les modèles mais sans clarifier l'ambiguïté dans l'importance relative des variables. Les plus indépendantes, comme la pluie, ou le vent, n'ont que rarement été sélectionnées alors que leurs influences sur les prédictions n'auraient pas été gênée par d'autres corrélations. En somme, la prise de position face au compromis complexité-interprétabilité ne permet d'obtenir qu'un aperçu limité, de l'influence réelle des variables sur la probabilité d'arrivée du charançon.

Le choix de comparer des modèles en privilégiant l'objectif de prédiction plutôt que d'interprétation, avec l'AUC pour critère principal, a favorisé les modèles basés sur les arbres de décision. En plus des limites d'interprétation, s'assurer que chaque modèle soit élaboré de la même manière via le package *caret*, n'a pas permis d'exploiter leur plein potentiel. C'est d'autant plus vrai pour les régressions dont la structure est particulièrement manipulable. Notamment, avec la possibilité d'expliquer la structure des résidus en ajoutant des termes complémentaires, comme les effets aléatoires ou ceux des modèles ARIMA (Hyndman et Athanasopoulos 2021). Nous aurions ainsi pu dissocier les cycles temporels avec l'influence des variables météorologique. Bien que l'utilisation d'une régression se passe difficilement d'une stratégie complémentaire pour gérer les multicolinéarité, l'ajout d'effets supplémentaires aurait permis de gagner en capacité prédictive, sans compromettre l'interprétabilité.

4.3 - Perspectives

On peut compter sur le développement des bases de données, l'amélioration des protocoles de suivis et l'intégration de nouvelles variables pour améliorer nos connaissances sur l'écologie de *C.picitarsis*. Le modèle actuellement développé n'a pas pour vocation de rester figé, mais d'intégrer les connaissances qui pourront l'enrichir. Il serait intéressant, en termes de perspectives, de comparer ce travail à une prospective purement exploratoire des variables impliquées. L'objectif principal, d'avancer sur la création d'un outil d'aide à la décision fonctionnel est atteint. L'élaboration des méthodes de luttes contre le ravageur pourrait toutefois bénéficier d'une meilleure compréhension de son activité une fois au champ. Pour cela, d'autres stratégies de modélisation n'ont pas encore été exploitées.

La première possibilité serait de miser sur un modèle de comptage des individus capturés, comme une régression de Poisson, qui constituerait un pas de plus vers la modélisation des dégâts sur la parcelle. Cet objectif final, essentiel pour orienter les pratiques agricoles, dépend également de la mise en place de protocole expérimentaux pour mieux identifier les facteurs déterminants la relation densité de population/pertes de rendement.

Une autre stratégie serait de miser sur un modèle uniquement exploratoire, potentiellement support pour le développement de nouvelles méthodes de luttes alternatives. Au vu du développement des résistances aux pyréthrinoïdes, prochainement seule famille d'insecticides efficace encore disponible sur le marché, le développement de nouvelles tactiques de protection des cultures de colza est un enjeu stratégique. Si l'expérimentation est une étape essentielle, il serait dommageable de se passer de la modélisation, particulièrement adaptée pour capitaliser l'information contenue dans les quantités massives de données collectées. Dans ce contexte, en plus d'intégrer de nouvelles catégories de variables, la méthode « Window Pane » reste à exploiter (Gouache et al. 2015). Elle permet d'explorer l'effet des variables prisent en compte sur des périodes antérieures à la capture. Bien qu'elle ne puisse se passer d'un protocole de sélection de variables robuste mais lourd, il n'aurait pas vocation à produire directement un outil d'aide à la décision fonctionnel. Il est tout à fait possible que cet effort soit nécessaire pour comprendre quelles variables sous-jacentes donne tant d'importance aux prédicteurs spatiaux et temporels.

Ce projet a également des perspectives plus concrètes. À ce stade, le modèle élaboré n'est pas un outil d'aide à la décision fonctionnel. Le produit final sera une carte régulièrement mise à jour en fonction de l'évolution des variables sélectionnées. La probabilité d'arrivée du charançon sera convertie en niveaux de risque. Pour garantir sa précision, les informations seront fournies à une échelle départementale à destination des agriculteurs, mais également des experts et techniciens fournissant des conseils techniques agricoles. Associé aux caractéristiques agronomiques de la parcelle, cet outil permettra de mieux raisonner l'utilisation de produits phytosanitaires. Toutefois, avant d'être disponible, il sera d'abord confronté à des conditions réelles, pour s'assurer de sa fiabilité sur le terrain.

Conclusion

Dans un contexte où le maintien des surfaces de colza est compormis par le manque de solutions efficaces, développer des outils robustes d'aide à la décision est un enjeu stratégique. Anticiper l'arrivée du charançon du bourgeon terminal participe à l'élaboration de méthodes de lutte adaptées en raisonnant l'utilisation d'insecticides. L'effort de capture a permis de constituer une base de données suffisamment volumineuse pour construire un modèle précis. Néanmoins, l'état des connaissances sur l'écologie du ravageur, ne fournit pas d'indications précises sur les variables à prendre en compte. Dans ces conditions, la mise en place d'une procédure de sélection de variable basée sur leur importance relative a permis de mieux identifier les facteurs influençant le réveil du CBT. De plus, les nombreuses techniques disponibles en apprentissage machine requiert la mise en place d'une méthodologie prudente. La comparaison de différentes méthodes nous a ainsi permis de mieux appréhender les avantages des modèles testés et d'identifier le plus adapté pour assurer des prédictions fiables.

L'utilisation d'un modèle fréquentiel comme référence met en évidence l'intérêt de mobiliser des méthodes plus complexes pour comprendre la dynamique de vol du charançon. Les variables introduites, bien qu'en partie redondantes, apporte effectivement une information précieuse au regard des prédictions. Pour prendre en compte leur relation avec la probabilité de capture, l'utilisation de modèles plus flexibles que la régression semble nécessaire. Bien que plus exigeants en temps de calcul, les méthodes basées sur les arbres de décisions se sont montrées nettement plus performantes. En appréhendant avec plus de précisions les relations prédicteurs-réponse, elles se saisissent de règles plus générales à partir d'un jeu de données pourtant très localisé à certaines aire géographiques. Au regard de la qualité des prédictions, il n'y a pas d'argument fort en faveur du modèle *gradient boosting*. Cela dit, sa tolérance à la multicolinéarité offre de meilleures possibilités d'interprétation de l'influence des variables.

Nous avons pu montrer que le réveil de *C.picitarsis* était fortement influencé par des variables météorologiques en particulier liées au changements de températures et de photopériodes avec de nets effets seuil. Les variables paysagères, comme le vent et les précipitations semblent effectivement impliquées mais de manière marginale. Au contraire, le jour de l'année et les coordonnées géographiques du point de capture semblent très impliquées dans le réveil du charançon. Si c'est dernières variables apportent une information précieuse au modèle, elles ne permettent pas de tirer des conclusions sur la dynamique réelle du charançon. On peut toutefois espérer que de futurs efforts de recherche, permettront de déterminer quelles variables, corrélées aux coordonnées spatio-temporelles déterminent effectivement l'arrivée du ravageur.

Bibliographie

Apley, Daniel, et Jingyu Zhu. 2020. « Visualizing the effects of predictor variables in black box supervised learning models ». *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (juin). https://doi.org/10.1111/rssb.12377.

Balachowsky, A.S. 1963. *Coléoptères second volume*. Masson et Cie. Vol. 1. 8 vol. Entomologie appliquée à l'agriculture.

Breiman, L. 2001. « Random Forests ». *Machine Learning* 45 (octobre): 5-32. https://doi.org/10.1023/A:1010950718922.

Coakley, S. M., R. F. Line, et L. R. McDaniel. 1988. « Predicting Stripe Rust Severity on Winter Wheat Using an Improved Method for Analyzing Meteorological and Rust Data ». *Phytopathology (USA)*. https://scholar.google.com/scholar_lookup?title=Predicting+stripe+rust+severity+on+winter+wheat+using+an+improved+method+for+analyzing+meteorological+and+rust+data&author=Coakley%2C+S.M.&publication_year=1988.

Debouzie, D., et F. Wimmer. 1992. « Models for winter rape crop invasion by the stem weevil Ceuthorrhynchus napi Gyll. (Col., Curculionadae) ». *Journal of applied entomology*, nº 114: 298-304.

Delaune, Thomas, Malick Ouattara, Rémy Ballot, Christophe Sausse, Irène Felix, Fabienne Maupas, Mathilde Chen, Muriel Morison, David Makowski, et Corentin Barbu. 2021. « Landscape drivers of pests and pathogens abundance in arable crops (Supporting information) ». *Ecography* 44 (10): 1429-42. https://doi.org/10.1111/ecog.05433.

Durand, Y, E.L.M. Brun, Laurent Mérindol, Gilbert Guyomarc'h, Bernard Lesaffre, et Eric Martin. 1993. « A meteorogical estimation of relevant parameters for snow models ». *Annals of Glaciology* 18 (janvier): 65. https://doi.org/10.1017/S0260305500011277.

Friedman, Jerome. 2002. « Stochastic Gradient Boosting ». *Computational Statistics & Data Analysis* 38 (février): 367-78. https://doi.org/10.1016/S0167-9473(01)00065-2.

Gouache, David, Marie Sandrine Léon, Florent Duyme, et Phillipe Braun. 2015. « A novel solution to the variable selection problem in Window Pane approaches of plant pathogen – Climate models: Development, evaluation and application of a climatological model for brown rust of wheat ». *Agricultural and Forest Meteorology*, nº 205: 51-59. https://doi.org/10.1016/j.agrformet.2015.02.013.

Hebinger, Hubert. 2013. « Partie 3 : La conduite de la culture - La protection de la culture ». In *Le colza*, Agriproduction, 525. Productions végétales et Grandes cultures. France Agricole.

Hyndman, Rob J, et George Athanasopoulos. 2021. *Chapter 9 ARIMA models | Forecasting: Principles and Practice (3rd ed)*. https://otexts.com/fpp3/arima.html.

Linardatos, Pantelis, Vasilis Papastefanopoulos, et Sotiris Kotsiantis. 2020. « Explainable AI: A Review of Machine Learning Interpretability Methods ». *Entropy 2021* 23 (1): 18. https://doi.org/10.3390/e23010018.

Makowski, David, François Brun, Elodie Doutart, Florent Duyme, Mohammed El Jabri, Kevin Fauvel, Maxime Legris, Aurore Philibert, François Piraux, et Alexandre Termier. 2021. *Data science pour l'agriculture et l'environnement : Méthodes et applications avec R et Python*. Ellipses. Formations & Techniques.

Molnar, Christoph. 2022. *Interpretable Machine Learning : A Guide for Making Black Box Models Explainable*. 2e éd. https://christophm.github.io/interpretable-ml-book.

Pilorgué, Etienne, Catherine Maisonneuve, et Yannick Ballanger. 1997. *Les ravageurs du colza d'hiver*. Les points techniques du CETIOM. PROLEA CETIOM.

Roâ, L., et P. Pastre. 1990. « Chapitre 3 : La protection contre les insectes ravageurs du colza doit se rainsonner ». In *La lutte contre les ravageurs du colza : dossier deltaméthrine*, Agrovet, 179. Roussel Uclaf.

Robert, C, L Ruck, J Carpezat, A Lauvernay, M Siegwart, et M Leflon. 2017. « Suivi des résistances des populations d'altises d'hiver (Psylliodes Chrysocephala) et du charançon du bourgeon terminal (Ceutorhynchus picitarsis) aux pyréthrinoïdes en France en culture de colza ». *11e conférence internationale sur les ravageurs et auxiliaires des cultures*, AFPP, .

Robert, C, L Ruck, V Lecomte, C Pontet, et S Cadoux. 2019. « Réduire la pression charançon du bourgeon terminal et altise d'hiver ». *Phytoma*, nº 724: 25-29.
Terres Inovia. 2022. *Guide de culture : colza 2022*. Terres Inovia.
Zou, Hui, et Trevor Hastie. 2005. « Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005) ». *Journal of the Royal Statistical Society Series B* 67 (février): 768-768. https://doi.org/10.1111/j.1467-9868.2005.00527.x.

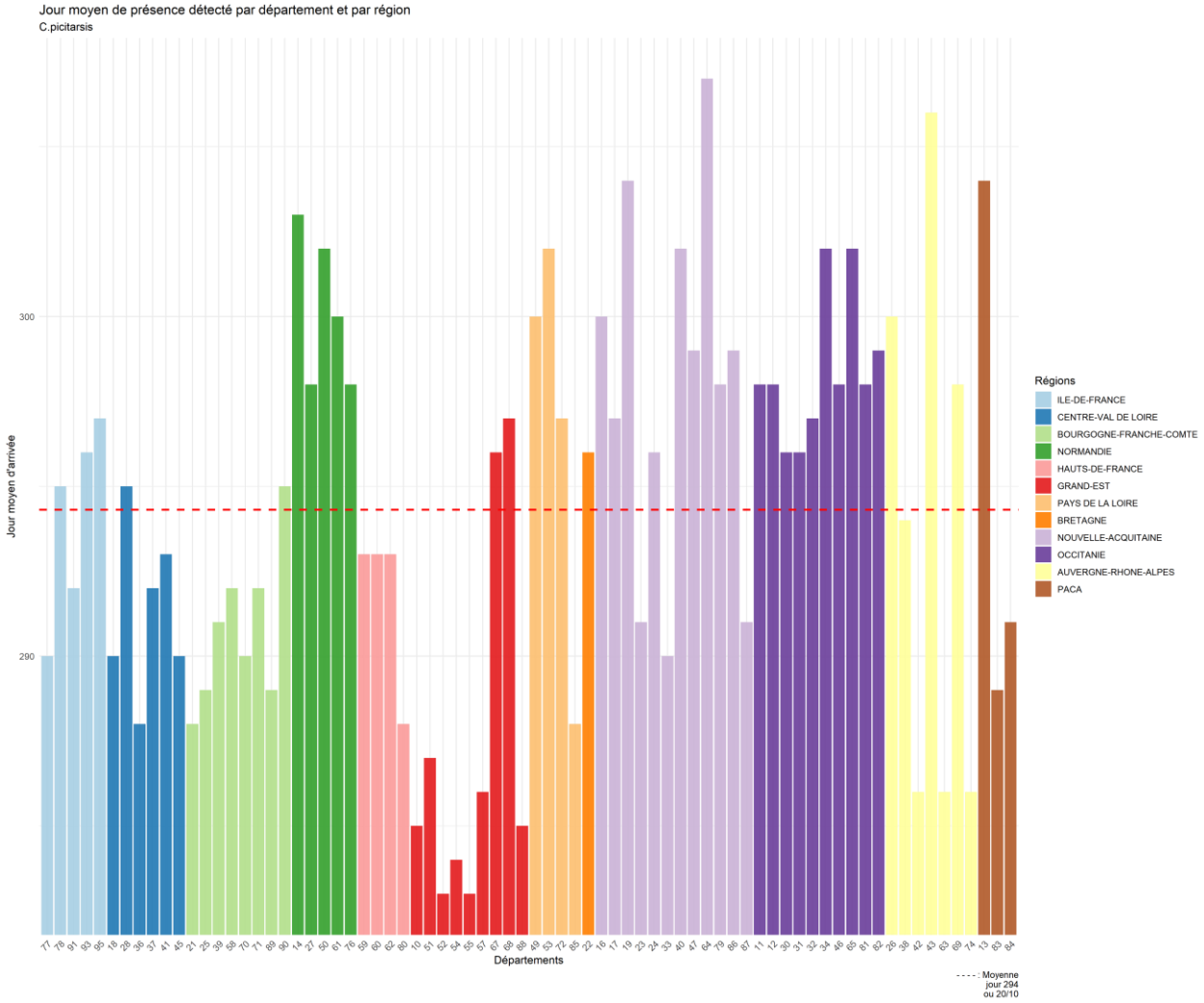
Annexes

Annexe 1 : Liste des variables modélisées

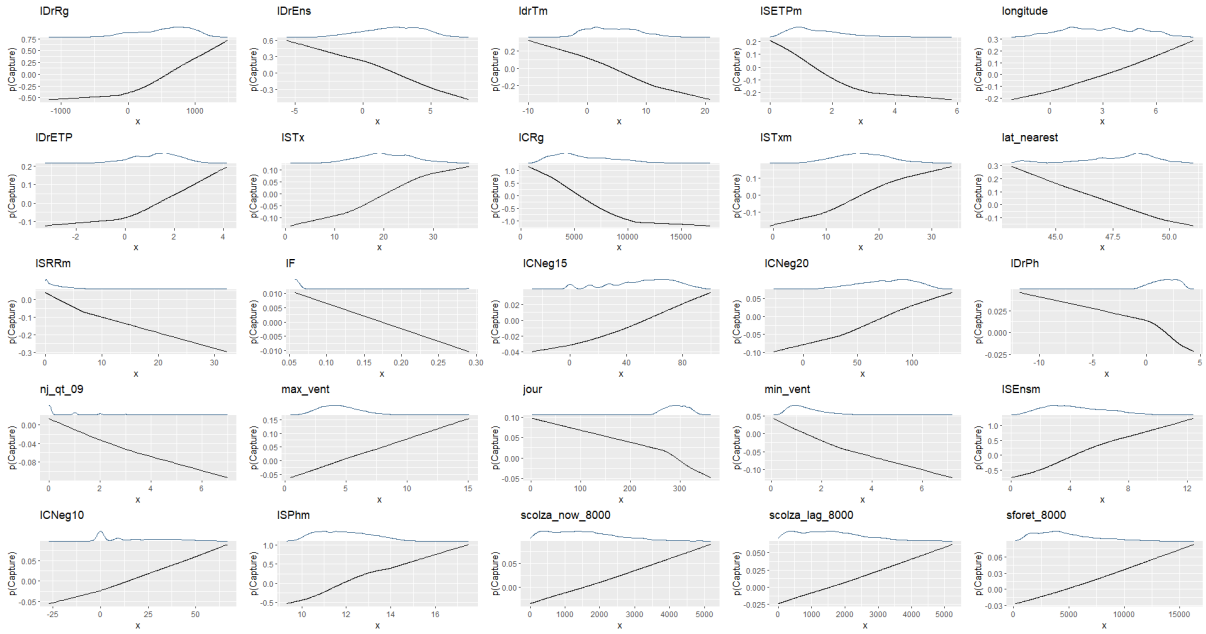
Code	Description
Variables de capture	
Jour	Date de capture en jour de l'année (1 à 365)
Capture	0/1 pour absence/présence de l'insecte lors du relevé de piège
Longitude	Longitude du piège
Latitude	Latitude du piège
Campagne_obs	Année au cours de laquelle débute la saison de capture (automne-hiver)
Ecart	Durée du suivi en nombre de jours au cours de la campagne
Distance	Distance en km entre la station météo, et le piège
Variable météorologiques	
Indicateurs simples sur les 7 jours avant la capture	
ISTn	Température minimale (°C)
ISTnm	Température minimale moyenne journalière (°C)
ISTx	Température maximale (°C)
ISTxm	Température maximale moyenne journalière (°C)
ISTm	Température moyenne (°C)
ISETPm	Evapotranspiration potentielle moyenne (mm)
ISETPx	Evapotranspiration potentielle maximale (mm)
ISEnsm	Ensoleillement moyen (heures)
ISEnsn	Ensoleillement minimum (heures)
ISRgm	Rayonnement global moyen (joules)
ISPhm	Photopériode moyenne (heures)
ISRRm	Pluviométrie moyenne (mm)
ISRRx	Pluviométrie maximale (mm)
ISBHS	Bilan hydrique simplifié moyen (mm)
Indicateurs de différence simple entre le jour de capture et le 7ème jour avant	
IDsTx	Température maximale (°C)
IDsTn	Température minimale (°C)
IDsTm	Température moyenne (°C)
IDsETP	Evapotranspiration potentielle (mm)
IDsEns	Ensoleillement (heures)
IDsRg	Rayonnement global (joules)
Indicateur de différence avec le début du réveil (septembre)	
IDrTx	Température maximale (°C)
IDrTn	Température minimale (°C)
IDrTm	Température moyenne (°C)
IDrETP	Evapotranspiration potentielle (mm)
IDrEns	Ensoleillement (heures)
IDrRg	Rayonnement global (joules)

IDrPh	Photopériode (heures)
IDrRR	Pluviométrie (mm)
IDrBHS	Bilan hydrique simplifié (mm)
Indicateurs de cumul sur la semaine précédant la capture	
ICRg	Rayonnement global (joules)
ICRR	Pluviométrie (mm)
ICBHS	Bilan hydrique simplifié (mm)
ICPos0	Degré jours positifs base 0 (°C)
ICPos4	Degré jours positifs base 4 (°C)
ICPos6	Degré jours positifs base 6 (°C)
ICPos8	Degré jours positifs base 8 (°C)
ICPosSeptm	Degrés jours positifs base température moyenne de septembre (°C)
ICNeg7	Degré jours négatifs base 7 (°C)
ICNeg10	Degré jours négatifs base 10 (°C)
ICNeg15	Degré jours négatifs base 15 (°C)
ICNegSeptm	Degrés jours négatifs base température moyenne de septembre (°C)
ICNeg20	Degré jours négatifs base 20 (°C)
Indicateur de fluctuation	
IF	0/1 si oui ou non la température journalière moyenne passe en-dessous puis au-dessus du seuil de 20°C sur 2 jours consécutifs au cours de la semaine précédant la capture.
Vent	
Long_nearest	Longitude du point SAFRAN
Lat_nearest	Latitude du point SAFRAN
Min_vent	Vent minimum sur la semaine (m/s)
Moy_vent	Vent moyen sur la semaine (m/s)
Max_vent	Vent maximum sur la semaine (m/s)
Nj_qt_01 … Nj_qt_09	Sur la base du vent enregistré sur toutes les années disponible, nombre de jour où le premier quantile est dépassé au cours de la semaine. Idem pour les quantiles 1 à 9. 9 variables
Variables paysagères	
Scolza_now_100 … Scolza_now_8000	Surface de colza autour du point de capture la même année, dans les rayons suivants : 100m, 200m, 300m, 500m, 1000m, 2000m, 3000m et 8000m. 8 variables (m²)
Scolza_lag_100 … Scolza_lag_8000	Surface de colza autour du point de capture l'année précédente dans les rayons suivants : 100m, 200m, 300m, 500m, 1000m, 2000m, 3000m et 8000m. 8 variables (m²)
Sforet_100 … Sforet_8000	Surface de colza autour du point de capture, dans les rayons suivants : 100m, 200m, 300m, 500m, 1000m, 2000m, 3000m et 8000m. La valeur est identique quelque soit les années. 8 variables (m²)

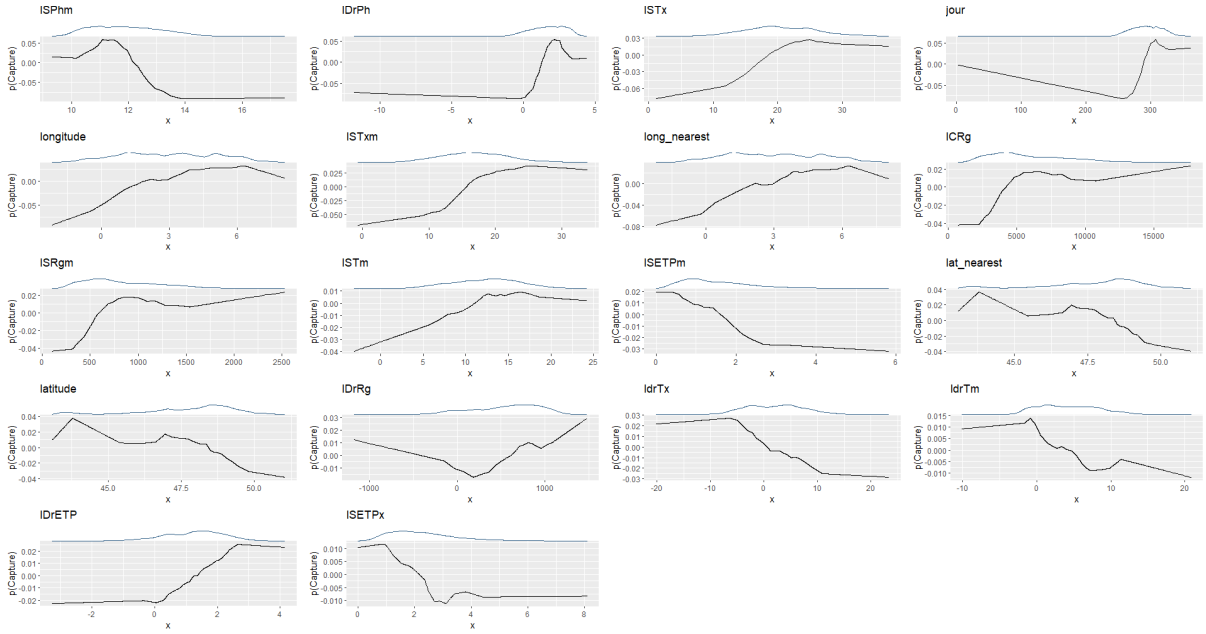
Annexe 2 : Jour moyen d'arrivée par département et par région sur l'ensemble des années.



Annexe 3 : Graphiques ALE du modèle *elastic net*



Annexe 4 : Graphiques ALE du modèle *random forest*



Annexe 5 : Graphiques ALE du modèle *gradient boosting*