MÉMOIRE DE FIN D'ÉTUDES

CHEVALEYRE Clément

Dans le cadre du stage de 3ème année

Stage effectué du 01/08/2023 au 31/12/2023

À l'Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)

Sur le thème de :

Modélisation de la pression des bioagresseurs (insectes et maladies fongiques) des grandes cultures

Rapport confidentiel : NON

Enseignant référent responsable: Jean-Marc GILLIOT

Maître de stage : Corentin BARBU

Engagement de non-plagiat

1 - Principes

- Le plagiat se définit comme l'action d'un individu qui présente comme sien ce qu'il a pris a 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.

2 - 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 sur d'en citer la source.

3 - Sanction

En cas de manquement à ces consignes, la DEVE/le correcteur se réservent le droit d'exiger la réécriture du document sans préjuger d'éventuelles sanctions disciplinaires.

4 - Engagement

Je soussigné(e) Clément CHEVALEYRE

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

A Voisins-le-Bretonneux le 05/11/2023

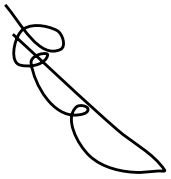
Signature : 

Table des matières

Liste des abréviations

BSV : Bulletin de Santé du Végétal

CART : Classification and Regression Trees

GAM-LASSO : Generalized Additive Model with Lasso

h : hyperparamètre *h* du modèle d'interpolation

IDW : Pondération inverse de la distance (Inverse Distance Weighting)

INRA : Institut National de la Recherche Agronomique

INRAE : Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement

IRSTEA : Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture

LASSO : Least Absolute Shrinkage and Selection Operator

MAE : Erreur absolue moyenne (Mean Absolute Error)

MARS : Multivariate Adaptive Regression Splines

NObsEff : Nombre d'Observations Efficaces

NPosEff : Nombre d'observation Positives Efficaces

nNeg : nombre d'observations négatives

nObs : nombre d'observations

nPos : nombre d'observations positives

RMSE : Erreur quadratique moyenne (Root-Mean-Square Error)

RPG : Registre Parcellaire Graphique

R^2 : Coefficient de détermination

R^2_{obs} : Coefficient de détermination observé

$R^2_{obs.w}$: Coefficient de détermination observé pondéré par le nombre d'observations

R^2_{opt} : Coefficient de détermination optimal

$R^2_{opt.w}$: Coefficient de détermination optimal pondéré par le nombre d'observations

R^2_{simu} : Coefficient de détermination simulé

$R^2_{simu.w}$: Coefficient de détermination simulé pondéré par le nombre d'observations

SAFRAN : Système d'Analyse Fournissant des Renseignements Adaptés à la Nivologie

Table des illustrations

Table des tableaux

Table des annexes

Remerciements

Je tiens à remercier en tout premier, Corentin BARBU, mon maître de stage qui m'a accordé sa confiance et m'a accompagné dans ce projet. Je remercie également Muriel VALANTIN-MORISON, directrice de l'UMR Agronomie, qui m'a permis de réaliser ce stage. J'adresse aussi de chaleureux remerciements à toute l'équipe de l'unité avec qui j'ai pu échanger et solliciter en cas de besoin. Pour finir, je remercie ma famille et mes amis qui ont pu m'aider dans la rédaction de mon mémoire.

Introduction

L'institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) est un institut public de recherche né de la fusion de l'INRA (Institut National de la Recherche Agronomique) et de l'IRSTEA (Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture) en janvier 2020 (INRAE, s.d.). Fondé après la Seconde Guerre mondiale, l'INRA avait pour objectif de répondre aux attentes de la population française qui fait alors face à une importante crise alimentaire. Sa mission était de développer et d'améliorer les techniques de l'agriculture afin que la France puisse devenir auto-suffisante. Au fil des années, les objectifs de l'INRA évoluent afin de s'adapter au conte-y xte socio-économique des différentes époques (crise énergétique de 1973, crise liée à la sécurité alimentaire dans la décennie 90 avec notamment la vache folle, changement climatique...). En 2006, l'INRA devient le premier organisme de recherche agronomique européen (INRAE, 2021).

Le projet MoCoRiBA-GC (Modélisation et Communication du Risque de BioAgresseurs en Grandes Cultures) est porté par l'INRAE et ses partenaires depuis 2020. Il visait initialement à produire des modèles statistiques pour prédire en temps réel la pression d'une quarantaine de bioagresseurs en grandes cultures sur le territoire français métropolitain (Tableau 1), et à intégrer cela dans un modèle de rendement prenant en compte les pratiques des agriculteurs. Plus largement, ce projet vise à éviter l'utilisation de produits phytosanitaires par les agriculteurs et ainsi de permettre une diminution de l'utilisation des produits phytosanitaires fongicides et insecticides en grandes cultures sans diminuer la marge économique pour les agriculteurs. En effet, des études montrent qu'une réduction jusqu'à 30% de l'utilisation des produits phytosanitaires serait possible sans perte de marge et sans modification importante des systèmes de culture (Butault, 2010 ; Lechenet, 2017).

Face à une réduction inattendue de la disponibilité en temps réel des données, le projet, tout en maintenant les volets de modélisation, s'est réorienté vers la production d'un outil à visée stratégique. Avec cet outil, les agriculteurs et leurs techniciens pourront juger de la pertinence de leurs décisions passées de traitement phytosanitaire en comparant leurs pratiques et résultats aux fermes du réseau DEPHY. Ce réseau s'engage dans la réduction de l'utilisation des produits phytosanitaires à travers de nouvelles pratiques et techniques culturales (EcophytoPIC, 2020). Pour favoriser ces comparaisons, l'application utilise des modèles de similarités entre parcelles et exploitations. Et de proximité entre parcelles et entre exploitations. Une version test de l'application est déjà en développement et en cours de présentation auprès d'acteurs intéressés pour l'utiliser.

Mon stage s'inscrit dans le projet MoCoRiBA sur la thématique de modélisation des bioagresseurs. L'objectif premier est dans un premier temps d'évaluer et d'améliorer le modèle de prédiction de la pression moyenne annuelle des bioagresseurs par interpolation utilisé jusqu'à présent. Dans un deuxième temps, des modèles statistiques prenant en compte des variables climatiques et/ou du paysage sont envisagés. Une combinaison des 2 catégories de modèles (interpolation et statistiques) est envisagée selon les résultats de chacun. Enfin, ces modèles pourront d'une part être implémentés dans l'application MoCoRiBA et d'autre part utilisés dans les modèles de prédiction de perte de rendement liés aux bioagresseurs. Tout le développement des modèles et leur évaluation est produit avec le langage informatique R.

1. Etat de l'art sur les méthodes de modélisation en biologie

Les bioagresseurs sont des déterminants importants du rendement des cultures. C'est pourquoi la connaissance de la pression des bioagresseurs est une information stratégique qui permet de piloter les traitements phytosanitaires des cultures (Devaud & Barbu, 2019).

Les réseaux de surveillance épidémiologique permettent de produire des bulletins d'informations, à destination des producteurs, sur les risques pour les cultures liés aux ravageurs et aux agents pathogènes. Ces informations sont utilisées par les agriculteurs et conseillers pour prendre les décisions de traitement. Cependant, la surveillance à grande échelle présente un défi logistique considérable avec des lieux d'échantillonnages qui ne sont pas uniformes sur un territoire pour de multiples raisons (distribution des cultures, paysages, main-d'œuvre,...). Pour la modélisation des pertes de rendement comme pour permettre des comparaisons de pratiques à pressions biologiques comparables, il faut donc, à partir de ces données, estimer de manière fiable la répartition des bioagresseurs quelle que soit la localisation géographique. Cet objectif de modélisation n'est pas spécifique aux bioagresseurs des cultures agricoles, mais est également utilisé pour suivre des espèces exotiques envahissantes (Venette et al., 2010), ou pour suivre des parasites responsables de maladies humaines, comme pour la maladie de Chagas responsable de plusieurs millions de morts, principalement en Amérique Latine (Barbu et al., 2013). Il existe 3 grandes catégories d'approches pour modéliser la présence d'une espèce biologique : les modèles corrélatifs, les modèles mécanistes et les modèles d'interpolation.

- Les modèles corrélatifs ou statistiques

Les modèles corrélatifs s'appuient sur des analyses statistiques qui cherchent la relation entre des variables et la répartition d'une espèce (Eyre et al., 2012). La construction d'un modèle se fait en plusieurs étapes : la sélection du modèle, l'ajustement du modèle puis la vérification et validation de celui-ci (Figure 1). Il existe une diversité de modèles possibles plus ou moins adaptés aux processus et aux observations que l'on souhaite modéliser. Il est donc nécessaire de choisir a priori un ou plusieurs pertinents puis d'évaluer la performance de ces modèles. Cette opération de vérification et validation du modèle permettra de définir un domaine de validité de celui-ci dans lequel il performe et ensuite de tenter de repousser ses limites en essayant de comprendre pourquoi le modèle ne se comporte pas comme attendu sur tel ou tel sous-ensemble (Quantmetry, 2023). Un risque majeur en modélisation statistique est le sur-ajustement du modèle à des caractéristiques non déterminantes du jeu de données, accidentellement corrélées dans ce jeu de données à une caractéristique des observations. La méthode de la validation croisée (*cross validation*), consiste en une comparaison des prédictions avec la réalité observée pour des données non utilisées lors de l'ajustement, donc indépendantes. Si l'erreur observée pour ces données est similaire à l'erreur observée pour les données d'ajustement, alors le modèle est vérifié, au sens qu'il ne comporte pas de sur-ajustement aux caractéristiques spécifiques du jeu de données d'ajustement. La méthode la plus rigoureuse de validation croisée est d'induire directement un biais lors de la sélection du jeu d'ajustement (spatialement ou climatiquement) pour avoir une mesure plus significative de sa robustesse et de la capacité à pouvoir appliquer le modèle à de nouveaux lieux ou périodes (Phillips, 2008).

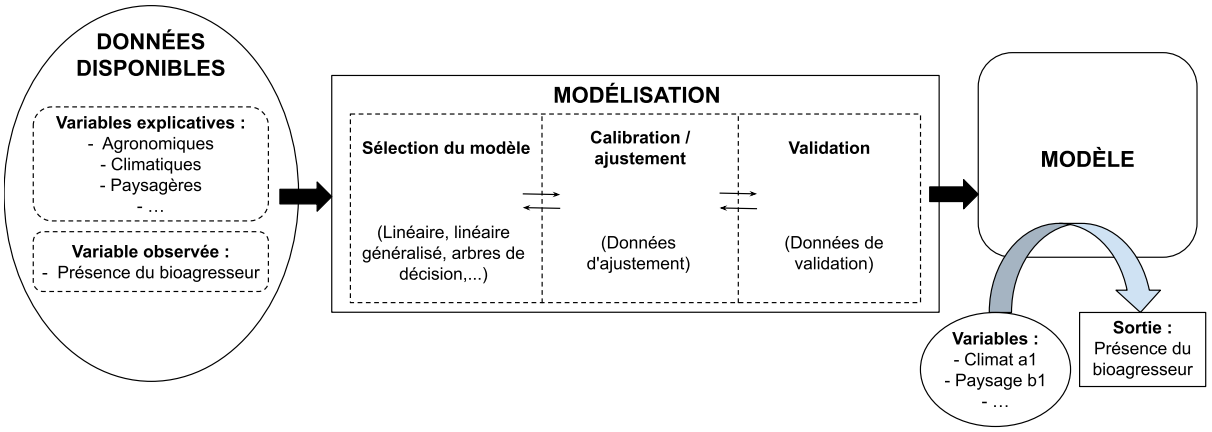


Figure 1 - Etapes de construction d'un modèle statistique pour prédire la présence d'un bioagresseur

Les facteurs climatiques sont couramment utilisés pour prédire les risques liés aux ravageurs et aux maladies puisqu'ils les déterminent dans une large mesure (Magarey et al., 2002 ; Koenig, 2002). D'autres variables peuvent être prises en compte dans la construction du modèle pour ajouter de la précision dans les prédictions. Celles-ci peuvent sembler, prisent seul dans le modèle, n'avoir aucun effet prédicteur pour les bioagresseurs mais, elles permettent un meilleur ajustement du modèle lorsqu'elles accompagnent une très bonne variable prédictive. Par exemple, l'impact du paysage sur la distribution des bioagresseurs est bien inférieur à celui induit par les conditions météorologiques, mais il n'est visible que si le climat (dans notre cas faisant partie des très bonnes variables prédictives) est pris en considération dans le modèle (Delaune et al., 2021).

Les modèles statistiques présentent l'avantage d'être disponibles dans des outils libres et gratuits tels que R ou Python. Par ailleurs, leurs résultats sont moins sujets à être affectés par le modélisateur puisque les paramètres du modèle sont ajustés indépendamment de celui-ci (Eyre et al., 2012). Ils permettent également de s'affranchir de toute connaissance sur la biologie des espèces étudiées (Venette et al., 2010). En contrepartie, ce type de modèle nécessite d'avoir accès à d'importantes bases de données sur la présence de l'espèce étudiée pour diverses situations climatiques (Eyre et al., 2012).

- Les modèles mécanistes

Les modèles mécanistes intègrent directement le mécanisme qui est responsable de la relation observée entre un caractère fonctionnel d'un organisme et son environnement. C'est-à-dire le lien direct qu'a l'organisme avec sa niche écologique dans son cycle de vie (Kearney & Porter, 2009 ; Eyre et al., 2012). De tels modèles se basent sur la dynamique spatio-temporelle des populations et simulent des processus biologiques et écologiques (ressources, hôtes, prédateurs, …).

La construction d'un modèle mécaniste ne nécessite pas de données pour son ajustement comme un modèle corrélatif, les différentes valeurs de paramètres pouvant être estimées à partir de la littérature ou en captant des connaissances spécifiques d'experts sur la biologie des bioagresseurs, en gestion des cultures ou encore en météorologie (Magarey et al., 2017). Les modèles mécanistes peuvent cependant aussi être entièrement ou partiellement ajustés à des observations, notamment dans un cadre bayésien ou les valeurs de paramètres estimées sont utilisées comme a priori. On peut alors parler de modèles mécano-statistiques.

- Les modèles d'interpolation

Bien que les modèles d'interpolation soient au sens strict des modèles corrélatifs, nous les traitons ici à part car la description de la structure d'auto-corrélation est en elle-même un problème complexe. Ici nous appellerons modèles d'interpolation les modèles visant à estimer des valeurs sur des sites non échantillonnés à l'aide de valeurs de la même variable provenant d'observations ponctuelles aux alentours des sites non échantillonnés (Li & Heap, 2011). Les modèles couramment utilisés sont les modèles de pondération inverse de la distance (IDW) et de krigeage. L'IDW est directement basé sur des valeurs de relevés avoisinants, pondérées inversement à la distance par rapport à l'emplacement de prévision. Le krigeage va plutôt se baser sur l'évaluation des corrélations entre les points de relevés en prenant en considération leur organisation spatiale (ArcGIS PRO,s.d.). L'interpolation repose sur l'hypothèse que les abondances d'un bioagresseur sur des sites proches devrait avoir une tendance similaire (Cohen et al., 2022). L'étape d'évaluation du modèle se fait via une multitude d'indices, mais le MAE (erreur absolue moyenne) et le RMSE (erreur quadratique moyenne) sont des estimations de l'erreur moyenne du modèle les plus fréquemment utilisées (Li & Heap, 2011).

L'interpolation peut être très utile dans les systèmes d'aide à la décision pour donner de l'information en temps réel sur la présence des bioagresseurs (Jones et al., 2010). C'est pourquoi c'est ce type de modèle qui a été utilisé jusqu'à maintenant dans le projet MoCoRiBA.

2. État des lieux de la modélisation des bioagresseurs dans MoCoRiBA

Le projet MoCoRiBA utilise de la modélisation statistique et des modèles d'interpolation pour prédire la pression des bioagresseurs. Initialement, il utilisait une moyenne pondérée par l'inverse de la distance au site d'intérêt comme modèle de pression des bioagresseurs.

Ces modèles sont développés à partir de bases de données nationales : données d'épidémiosurveillance, Epiphyt et Vigicultures®, et données météorologiques, SAFRAN. Epiphyt se veut être une base de données qui centralise l'ensemble des données d'épidémiosurveillance des partenaires publics et privés (Ministère de l'Agriculture et de la Souveraineté Alimentaire, 2011). Vigicultures® centralise les données collectées par les instituts techniques des grandes cultures pour l'épidémiosurveillance nationale et notamment l'élaboration des BSV (Bulletin de Santé du Végétal) (Arvalis, 2020). Un extrait de la base de données d'épidémiosurveillance est présenté en annexe 1.

La base météo SAFRAN (Système d'Analyse Fournissant des Renseignements Adaptés à la Nivologie), produite par Météo-France, regroupe des données climatiques pour une grille de points espacés de 8 km chacun sur le territoire français. Par commodité, nous utiliserons aussi dans la suite cette grille pour la cartographie des prédictions à l'échelle nationale.

2.1. Présentation des bioagresseurs et métriques utilisées pour illustrer ce mémoire

Pour évaluer la présence d'un bioagresseur dans une parcelle, il existe différents protocoles de mesure fournissant chacun une métrique. Les données d'observations récoltées pour l'ensemble des bioagresseurs suivent donc des protocoles assez diversifiés même pour un unique bioagresseur. Néanmoins, pour les insectes, ce seront principalement des dénombrements d'individus ou une proportion de dégâts sur la culture, et pour les maladies fongiques un pourcentage de feuilles atteintes.

Dans ce mémoire, toutes les analyses de modèles sont réalisées sur les métriques de références de chaque bioagresseur des cultures. Nous avons choisi pour métrique de référence celle qui, pour un bioagresseur donné, comptabilise le plus d'observations au cours des différentes campagnes de culture (Tableau 1).

Tableau 1 - Bioagresseurs étudiés dans le projet MoCoRiBA pour les différentes cultures avec l'identifiant Vigiculture® de la métrique de référence

Cultures	Bioagresseurs	Identifiant Vigicultures®
Betterave	Rouille	CERCO_%F
	Cercosporiose	PEGO_%PLT_GAL
	Pégomyie	PUC_NOIR_%PLA_APTE
	Puceron vert	PUC_VERT_%PLA_APTE
	Puceron noir	ROUILLE_%F
	Teigne de la betterave	TEIGNE_%DEG
Blé tendre d'hiver	Helminthosporiose	HELMIN_F3
	Oïdium des céréales	OIDF3
	Puceron	PV_%
	Piétin verse	PUC_EPI_PLANT_%
	Puceron vecteurs de viroses	PUC_AUT_PLANT_%
	Rouille brune du blé	RBF3
	Septoriose des céréales tritici	SEPF3
Colza d'hiver	Méligèthe du colza	A%M
	Puceron vert du pêcher	GANbPE
	Altise	PANbV
	Sclérotiniose	ChCNbV
	Altise petite des crucifères	ChTNbV
	Charançon du bourgeon terminal	CBTNbV
	Charançon de la tige du colza	Mel%P
	Altise Grosse d'hiver du Colza	PV%P
	Charançon de la tige du chou	EPIPHYT_Scl%PTE
Maïs	Sésamie	CHRYSM_NB_PG_SEXUEL
	Pyrale du maïs	FOR_%PL_ATQ_AVTREC
	Insectes foreurs de la tige	PYR_PHE_NB_ADULTES
	Chrysomèle des racines du maïs	SES_PHE_NB_ADULTES
	Taupin	TAUPINS
Orge d'hiver	Piétin verse	HELMIN_F3
	Puceron vecteurs de viroses	OIDF3
	Taupin	PV_%
	Rhynchosporiose	PUC_AUT_PLANT_%
	Oïdium des céréales	RHYNCF3
	Helminthosporiose de l'orge D teres	TAUPINS
Pomme de terre	Puceron	ALTERNA
	Doryphores	DORYPHORES_ADULTES
	Alternariose de la pomme de terre	MILDIOU
	Mildiou de la pomme de terre	NB_PUCERON_FOL
Tournesol	Puceron vert du prunier	ManLimB
	Puceron noir de la fève	PhaFeu
	Phoma macdonaldi Maladie des tâches noires	PhoFeu
	Phomopsis du tournesol	PucNoi
	Limace	PucV%PI

J'ai réalisé lors du stage la modélisation pour 40 ravageurs et agents pathogènes, cependant, pour illustrer l'évolution des modèles, des calculs et des visuels cartographiques, je prendrai ici comme exemple 2 bioagresseurs sur une campagne de suivi. J'ai choisi la septoriose du blé, une maladie fongique, avec le pourcentage de feuilles F3 atteintes (métrique de référence pour cette maladie), et la grosse altise d'hiver du colza, un insecte ravageur, avec le pourcentage de plantes avec un cœur détruit ou un port buissonnant. Ces 2 métriques diffèrent en abondance de données disponibles. Elles permettent de voir les différences et/ou limites qu'il peut y avoir pour des métriques avec beaucoup ou peu de données disponibles par campagne de culture.

Le boxplot suivant illustre la distribution du nombre moyen de parcelles observées annuellement pour l'ensemble des métriques du projet. Plus de la moitié des métriques présentent en moyenne moins de 200 observations par an. On peut expliquer cette distribution par les raisons suivantes :

- Une métrique peut être beaucoup observée lorsqu'elle correspond à une culture majeure produite en France (ex : blé, colza, orge).
- Une métrique peut être très peu regardée par les agriculteurs et les techniciens agricoles lorsque d'autres métriques sont plus utilisées (facilité de mesure, intérêt pour l'estimation de dégâts, …).
- Une culture peut être principalement localisée dans une aire géographique due à la spécialisation des territoires et les observations sont regroupées sur cette aire. On a donc très peu de points à l'échelle national mais une plus forte concentration dans une région donnée (exemple de la betterave pour le nord de la France)

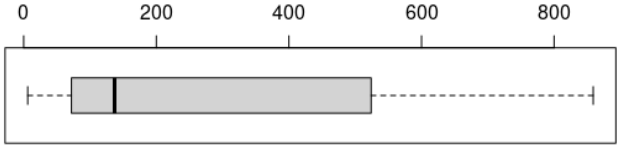


Figure 2 - Distribution du nombre moyen de parcelles suivies annuellement pour les différentes métriques des bioagresseurs suivies par le réseau d'épidémiosurveillance et considérées dans le projet MoCoRiBA

- Septoriose des céréales

La septoriose est une maladie fongique visible chaque année, d'intensité variable selon les conditions climatiques et présente tout au long du cycle des céréales. Elle est due au champignon *Zymoseptoria tritici* qui engendre des taches blanches et/ou brunes sur les feuilles nécrosant et dégradant l'activité photosynthétique des plantes. Sans traitement, les pertes de rendement peuvent être de l'ordre de 40% (Arvalis,s.d.). La méthode de suivi classique correspond à la notation des 3 feuilles les plus récentes (numérotées F1, F2 et F3, respectivement de la plus jeune à la plus ancienne) sur plusieurs plantes pour mesurer la fréquence d'apparition de la septoriose. On obtient donc une note de fréquence de présence de la maladie pour chaque feuille (Simonneau, 2015 ; Esquirol, 2012).

Figure 3 - Tâches brunes de septoriose (Septoria tritici) sur céréale à paille
(Crédit photo : ARVALIS - Institut du végétal)

La métrique utilisée pour illustrer ce mémoire correspond au pourcentage de feuilles F3 atteintes par la septoriose (code Vigicultures® : SEPF3) pour la campagne 2017. Elle est suivie dès la reprise de la croissance en sortie d'hiver jusqu'à la récolte sur de nombreuses parcelles chaque année (802 parcelles en 2017).

- Grosse altise d'hiver du colza (*Psylliodes chrysocephala*)

C'est un insecte de 3 à 5 cm de la famille des Coléoptères et ravageur du colza. A partir de septembre, lorsque les températures commencent à chuter mais restent supérieures à 20°C en journée, les adultes s'attaquent aux jeunes pousses du colza qui sont très sensibles du stade levée à 3 feuilles en se nourrissant et perforant les cotylédons (Figure 4). Les dégâts entraînent un affaiblissement général des plantes qui ne pourront pas compenser par la suite. Cependant la plus grande menace provient des larves issues des pontes de l'automne qui creusent des galeries dans les pétioles des tiges et peuvent remonter jusqu'au bourgeon terminal dans la tige. Cela peut engendrer des pertes importantes sur le rendement. Il est donc important de suivre l'arrivée et la présence des grosses altises sur sa parcelle au cours de la culture. Plusieurs méthodes sont disponibles : pièges en forme de cuvette jaune enterrée, dénombrement d'insectes sur les plantes ou encore suivi des dégâts sur la culture (Terre Inovia, 2023 ; Lieven, 2016).

Figure 4 - Grosse altise (A) ; dégât sur cotylédons du colza (B)
(Crédit photo : Terre Innovia)

La métrique utilisée pour illustrer ce mémoire correspond au pourcentage de plantes avec un cœur détruit ou un port buissonnant (code Vigicultures® : LGA%B), pour la campagne 2016. Elle est suivie pendant la période hivernale jusqu'en Avril et est donc plus tardive que l'observation des altises sur la culture. Ce n'est pas une métrique qui permet de suivre en temps réel la présence du ravageur, elle est plutôt utilisée pour estimer les pertes de rendement de la culture. Cette métrique à la particularité de faire partie des métriques ayant un nombre de parcelles d'observations présentes chaque année assez faible. (188 parcelles en 2016).

2.2. Etat d'avancement du projet sur la modélisation des bioagresseurs

- Simplification des données pour faciliter l'analyse statistique

Les protocoles d'évaluation des métriques peuvent aboutir à des quantifications très différentes comme des comptes d'individus, des pourcentages ou des indications de présence ou d'absence. A cela, s'ajoutent des distorsions des données dûes à l'hétérogénéité des observateurs, par exemple sur les approximations qui sont faites lors de la mesure. En regardant attentivement les données d'observations des bioagresseurs sur l'ensemble des parcelles, on s'aperçoit que leur distribution présente souvent des pics caractéristiques d'une variable discrète (Figure 5) bien que la variable soit continue en théorie. Cela indique des arrondis à la dizaine ou à 5% suivant les observateurs lors des mesures terrains.

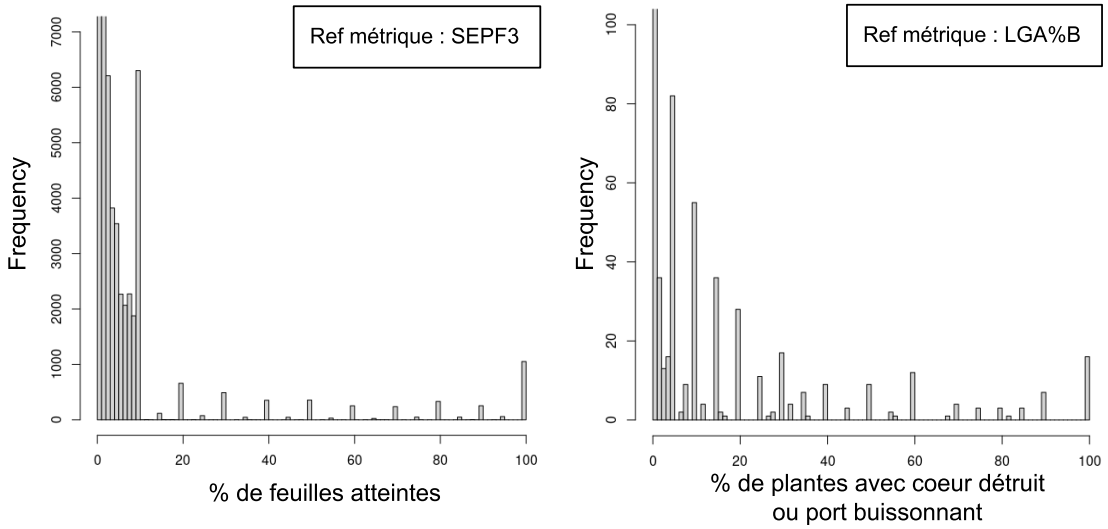


Figure 5 - Distribution des observations des métriques de la septoriose du blé (SEPF3) et de la grosse altise d'hiver du colza (LGA%B)

Pour une analyse statistique correcte, notamment pour inférer des intervalles de confiance, il est nécessaire de décrire correctement le processus d'échantillonnage qui génère la distribution des données. Afin de fiabiliser l'analyse des données, il peut être utile de simplifier les données. Par exemple dans le cas de *SEPF3* ou de *LGA%B* il serait possible d'arrondir toutes les données à 10% près pour s'assurer d'avoir des arrondis homogènes entre l'ensemble des observateurs et correspondant à un processus d'observation binomial classique. Pour permettre aussi d'analyser conjointement des données de présence/absence, nous simplifions ici les données en introduisant un seuil pour obtenir une nouvelle variable binaire prenant des valeurs d'observation 0 ou 1 suivant si l'on est au-dessus et en dessous du seuil (Delaune, 2021). Le seuil est défini par la médiane des valeurs de l'ensemble des observations présentes dans notre base de données par métrique (pour tous les champs et toutes les campagnes). L'utilisation de la médiane pour chaque bioagresseur permet de maximiser la puissance statistique de l'analyse car elle répartie les observations de manière équilibrée entre 0 et 1, à moins que les valeurs initiales soient trop déséquilibrées (sur-abondance de 0 par exemple). On fait donc l'hypothèse que la médiane reflète bien la valeur seuil à partir de laquelle le ravageur à un impact négatif sur la culture. Ensuite, nous observons le nombre d'observations dans l'année où l'observation de la métrique a dépassé le seuil choisi. Outre le fait de faciliter l'analyse des données, cela permet aussi d'avoir pour chacune des métriques, des variables qui suivent toute une même

loi binomiale (annexe 2). De ce fait, on peut s'affranchir de la spécificité de chaque échelle de mesure pour chaque métrique ainsi que de la définition à dire d'expert d'un seuil spécifique à chaque bioagresseur.

Les seuils pour les métriques étudiées sont les suivants :

- $Seuil_{SEPF3}$ = 2%, soit au-dessus de 2% de feuilles F3 atteintes par la septoriose, l'observation est positive.
- $Seuil_{LGA\%B}$ = 0%, soit à partir de 1% de plantes avec coeurs détruit et ou port buissonnant, l'observation est positive.

La somme des valeurs positives supérieures au seuil au cours d'une campagne est notée *nPos* et l'ensemble des observations *nObs.*

- Calcul de la pression des bioagresseurs dans une parcelle observée

La pression observée d'un bioagresseur sur une parcelle correspond mathématiquement pour une variable qui suit la loi binomiale à la probabilité d'obtenir un succès lors d'une observation. La pression du bioagresseur *i* dans une parcelle *p* est donnée par :

$$P_p^{obs}(bio\_i) = nPos_{p,bio\_i} / nObs_{p,bio\_i} \qquad (2.2.a)$$

où :
$nPos_{p,Bio\_i}$: nombre d'observations qui dépassent le seuil pour le bioagresseur de l'espèce *i* dans la parcelle *p*
$nObs_{p,Bio\_i}$: nombre total d'observations du bioagresseur de l'espèce *i* dans la parcelle *p*

- Prédiction de la pression moyenne annuelle des bioagresseurs dans une parcelle non-observée

Pour étudier l'impact des bioagresseurs sur le rendement, il faut pouvoir estimer la pression des bioagresseurs dans une parcelle d'intérêt. De précédents travaux réalisés par Cisse A. ont comparé 3 méthodes de prédiction de la pression annuelle des bioagresseurs dans une parcelle :

- méthode de la moyenne pondérée
- méthode des *k-NN* (*k* plus proches voisins)
- simulation binomiale

Les résultats ont montré que la méthode la plus pertinente à utiliser était celle de la moyenne pondérée. Il s'agit d'un modèle d'interpolation qui détermine la pression d'un bioagresseur au niveau d'une parcelle en faisant une moyenne des pressions observées des parcelles connues aux alentours, pondérées par un poids. Plusieurs structures de poids en fonction de la distance peuvent être utilisées. Utiliser l'inverse de la distance est une mesure classique mais inadaptée ici. A. Cissé a montré qu'une évolution exponentielle décroissante était plus pertinente. Associé à la construction binomiale des données, les travaux réalisés ont abouti au modèle d'interpolation qui suit :
Supposons que nous connaissions la pression du bioagresseur de l'espèce *i* pour 1 campagne donnée, dans un ensemble de *m* parcelles notée P_p^{obs} *(bio_i), pour p = 1,.....,m.*
La pression du bioagresseur d'espèce *i* prédite dans la parcelle *p'* est donnée par :

$$P_{p'}^{prédite}(bio\_i) = \frac{\sum_{p=1}^{m} e^{\frac{-d_{pp'}}{h}} P_p^{obs}(bio\_i)}{\sum_{p=1}^{m} e^{\frac{-d_{pp'}}{h}}} \quad (2.2.b)$$

Où :
$d_{pp'}$: distance qui sépare la parcelle *p'* et la parcelle *p* avec 1 ≤ *p* ≤ m
h : hyperparamètre

Ce ratio peut être interprété comme le ratio d'un nombre d'observations positives efficaces et du nombre d'observations efficaces en un point. Interprétation que nous utiliserons dans les chapitres suivants :

$$P_{p'}^{prédite}(bio\_i) = \frac{NPosEff_{p',bio\_i}}{NObsEff_{p',bio\_i}} \quad (2.2.c)$$

avec :
$$NPosEff_{p',bio\_i} = \sum_{p=1}^{m} nPos_{p,bio\_i} e^{\frac{-d_{pp'}}{h}} \quad (2.2.d)$$

$$NObsEff_{p',bio\_i} = \sum_{p=1}^{m} nObs_{p,bio\_i} e^{\frac{-d_{pp'}}{h}} \quad (2.2.e)$$

NObsEff : Nombre d'Observations Efficaces. C'est l'agrégation des informations d'observations apportant une information dans la prédiction de la pression. Chacune des observations étant pondérée suivant sa localisation et apportant plus ou moins d'informations. Sa valeur est rarement un entier et correspond à la taille de l'échantillon sur laquelle se base la prédiction.
NPosEff : Nombre d'observations Positives Efficaces. De la même manière que *NObsEff*, mais restreint aux observations positives.

- Calcul de l'hyperparamètre *h* optimal

L'hyperparamètre *h* joue sur l'allure de l'exponentielle en divisant la distance et donc modifiant le poids donné à une parcelle (Figure 6). Il est homogène à une distance, et indique où les poids correspondent à e^-1 soit environ ⅜ de leur valeur au lieu d'observation soit environ ⅜ de leur maximum.

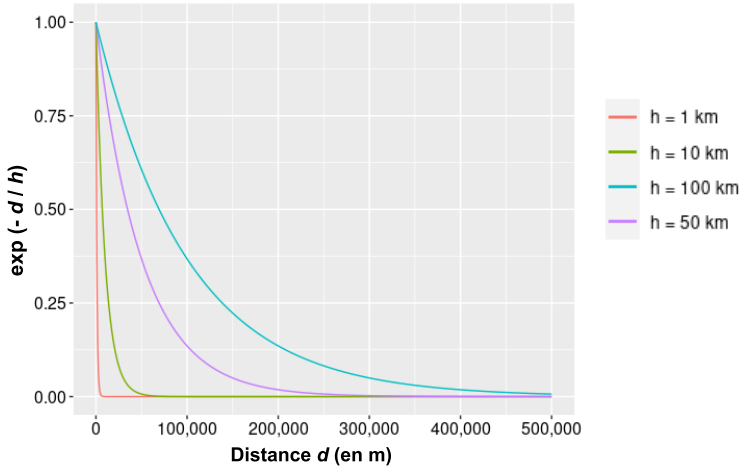


Figure 6 - Evolution de l'allure de la décroissance exponentielle selon la valeur prise par l'hyperparamètre h

Procédure de recherche du *h* optimal (Cisse, 2022) :

- Variation de l'hyperparamètre *h* dans un intervalle de distance de 0 à 250 km;
- Pour chaque valeur de *h*, prédiction de la pression des bioagresseurs au niveau de chacune des parcelles par campagne (cette dernière étant exclu à chaque fois) ;
- Calcul du coefficient de détermination R^2 entre les pressions réelles et les pressions prédites correspondant à un *h.* Le meilleur R^2 correspondant au meilleur *h*;

Les graphiques suivants représentent le coefficient de détermination R^2 entre les pressions réelles et les pressions prédites par le modèle (2.2.b), en fonction de la valeur que prend l'hyperparamètre *h* pour les 2 bioagresseurs étudiés. La valeur de l'hyperparamètre *h* qui donne la meilleure interpolation est de 19 km pour la métrique de la septoriose (*SEPF3*) et de 23 km pour la métrique de la grosse altise (*LGA%B*).

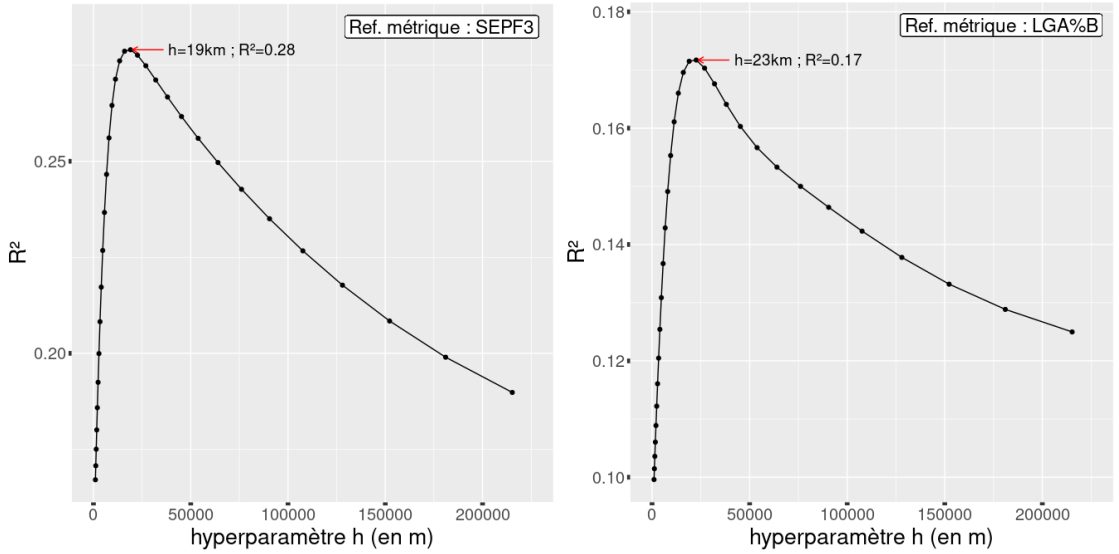


Figure 7 - R^2 de la pression prédite par interpolation en fonction de la valeur prise par l'hyperparamètre h pour la septoriose du blé (métrique SEPF3) et la grosse altise d'hiver du colza (métrique LGA%B)

Le paramètre *h* déterminé pour chaque bioagresseur permet de faire l'interpolation sur le territoire métropolitain (2.2.b). Le modèle d'interpolation est construit sur la loi binomiale et une moyenne pondérée inversement à la distance avec une évolution exponentielle décroissante. Il a été évalué sur sa capacité à prédire correctement les points au niveau de chacune des parcelles et, pour chaque culture à l'échelle de la France entière. C'est le modèle actuellement utilisé dans l'outil MoCoRiBA. Il est utilisable en l'état mais présente quelques limites que nous qu'il est intéressant de pallier. C'est ce que nous allons voir dans le chapitre suivant.

3. Apports du stage à la modélisation de la pression annuelle des bioagresseurs par interpolation

La pression des bioagresseurs est connue pour chaque campagne sur une multitude de parcelles réparties dans toute la France selon les bassins de production de chaque culture. Le premier modèle d'interpolation utilisé par le projet (2.2.b) est basé sur une loi binomiale des observations. Dans une première partie du stage, nous avons évalué la capacité du modèle d'interpolation à donner une estimation du risque de présence des bioagresseurs entre les parcelles d'observations en estimant l'incertitude théorique de l'interpolation. Ces analyses ont permis d'identifier certaines limites qui ont abouti à l'évolution du modèle.

Pour construire les visuels cartographiques, les interpolations et les calculs d'incertitude vont être réalisés pour un nombre important de points géographiques répartis sur tout le territoire. Pour cela, les points utilisés sont ceux de la grille SAFRAN. Des illustrations cartographiques réalisées avec la fonction mapview() sur R permettent ensuite de visualiser l'interpolation de la pression des bioagresseurs au niveau national. L'ensemble des cartes utilisées pour illustrer cette partie sont présentes en plus grand format en annexe 7 et 8.

3.1. Estimation de l'incertitude théorique pour une distribution binomiale des observations telle qu'utilisée pour l'interpolation par Cisse A.

L'interpolation avec le modèle basé sur la loi binomiale (2.2.b) est réalisée sur les points de la grille SAFRAN après avoir déterminé pour chaque métrique de chacun des bioagresseurs son hyperparamètre *h*.
Un exemple de l'interpolation au niveau d'une maille SAFRAN est détaillé pour la grosse altise du colza en annexe 3.

- Estimation de l'incertitude de l'interpolation

Pour représenter l'erreur théorique de l'interpolation on calcule l'intervalle de confiance à 90% de la loi de probabilité binomiale de l'interpolation de la manière suivante :

$$IC_{90\%} = q_{0,95} - q_{0,05} \qquad (3.1)$$

où :
$q_{0,95}$: représente le quantile à 95%
$q_{0,05}$: représente le quantile à 5%
$IC_{90\%}$: représente la différence des 2 quantiles 0,05 et 0,95. Plus la différence est petite, plus notre interpolation sera correcte et inversement.

La figure suivante illustre l'incertitude que l'on souhaite connaître :

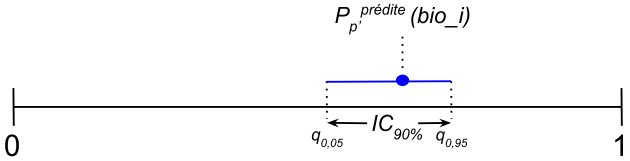


Figure 8 - Illustration de la signification de l'incertitude à 90% calculée

Le calcul des quantiles théoriques pour la loi binomiale se fait initialement sur le nombre d'observations positives avec la fonction R *qbinom* qui prend en entrée une taille d'échantillon *n,* avec une probabilité *p* de succès et un quantile *q* (dans notre cas 0,05 et 0,95). Le quantile est défini comme la plus petite valeur *x* telle que F(*x*) ≤ q, où F est la fonction de distribution. Pour avoir ensuite le quantile dans l'intervalle [0,1] on divise *x* par le nombre d'observations *n*.

Cependant, le modèle reposant sur une loi binomiale pose un problème pour le calcul des quantiles puisqu'il nécessite une taille d'échantillon entière sinon celle-ci est automatiquement arrondie à l'entier le plus proche par la fonction *qbinom*. Or, nous avons des tailles d'échantillons (*NObsEff)* qui ne sont jamais entières et sont souvent très petites (Figure 9). Lorsque la métrique est très bien représentée en nombre d'observations (SEPF3), l'interpolation est réalisée à partir d'un *NObsEff* important, alors l'arrondi reflète bien la réalité. Cependant, pour les métriques très peu représentées les *NObsEff* sont beaucoup plus faibles, et fréquemment inférieures à 1 comme pour l'observation des colzas buissonnants en lien avec la grosse altise (LGA%B). Pour être sûr de ne pas sous-estimer notre incertitude, on arrondit préalablement les tailles d'échantillons à l'entier inférieur. Pour les très faibles tailles d'échantillons, l'arrondi pose problème puisque toute les valeurs sont arrondies au même entier et donc il n'est pas possible d'avoir une incertitude graduelle. Les quantiles à 5% et 95% sont égaux à 0 et à la taille d'échantillon, ce qui donne une incertitude de 1.

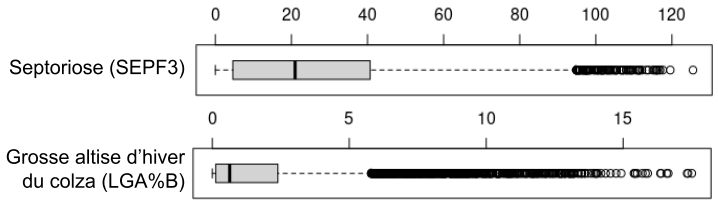


Figure 9 - Distribution des NObsEff lors du calcul d'interpolation pour les 8981 mailles SAFRAN pour la septoriose en 2017 et la grosse altise d'hiver du colza en 2016

- Etude statistique de la qualité du modèle

Commençons par visualiser les prédictions réalisées lors de la détermination de l'hyperparamètre *h* en fonction de la réalité observée sur le terrain pour toutes les campagnes confondues (Figure 10). Une tendance de corrélation se dégage pour les 2 métriques étudiées même si la prédiction est très dispersée. La corrélation est bien plus visible pour la septoriose puisqu'il y a beaucoup plus d'observations avec environ 9500 contre 1500 pour la grosse altise. On remarque aussi l'effet seuil de la pression réelle observée avec des paliers bien visibles pour 0.33, 0,5 et 0.66. C'est dû au faible nombre d'observations réalisées par parcelle qui produit une pression parcellaire discrète (2.1.a) dont une proportion importante de pression observé de valeur 0 ou 1.

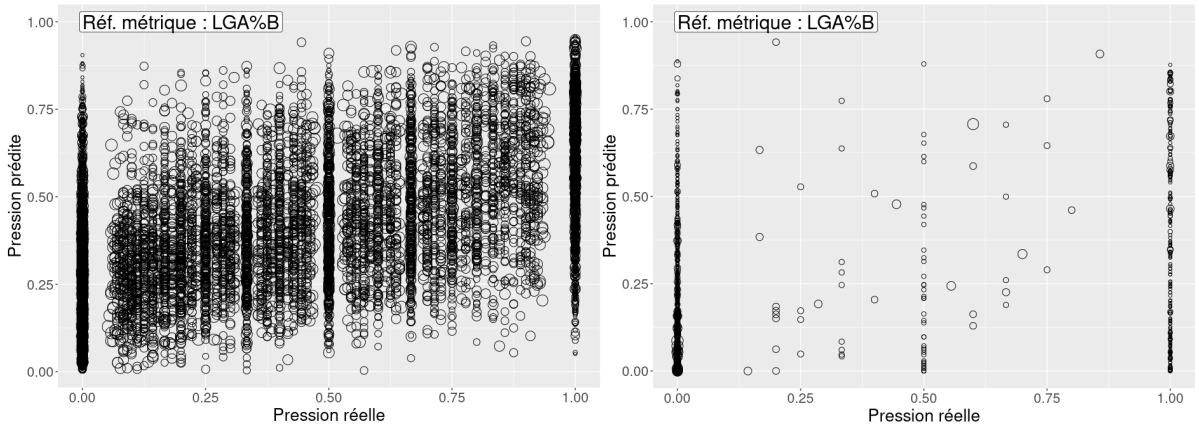


Figure 10 - Relation entre la pression prédite par interpolation et la pression réellement observée sur toutes les années pour la septoriose du blé (SEPF3) et la grosse altise du colza (LGA%B)

Il existe différentes manières d'analyser statistiquement la qualité du modèle d'interpolation basées sur des méthodes analytiques. Pour une régression linéaire, comme c'est le cas entre nos variables de pression réelle et pression prédite, le coefficient de détermination ou R² permet d'évaluer la qualité de l'ajustement d'un modèle de régression aux données, c'est-à-dire de la prédiction par rapport aux observations en indiquant la proportion de la variance expliquée (Wikipedia, s.d.).

Le MAE (erreur absolue moyenne) et le RMSE (erreur quadratique moyenne) sont aussi souvent utilisés pour estimer l'erreur moyenne des modèles (Li & Heap, 2011). Le MAE correspond à la moyenne de toutes les erreurs de prédiction en absolues (soit la moyenne de la différence entre les valeurs réellement observées et les valeurs prédites). Le RMSE correspond à la différence absolue moyenne entre une valeur prédite et la valeur observée. Il s'agit de l'écart type des résidus (la différence entre la valeur observée et la valeur prédite. Les valeurs du MAE et RMSE sont exprimées dans la même unité que celle de la valeur cible (Hodson, 2022).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

Dans le cas d'un modèle logistique où l'on essaye de prédire la probabilité d'une occurrence pour un résultat binaire (probabilité que la parcelle soit infestée par un bioagresseur ou non). Plusieurs mathématiciens ont développé des méthodes pour calculer des Pseudo-R² analogues dont les plus connus sont le R-carré de Cox et Snell, le R-carré de Nagelkerke ou le R-carré de McFadden plus connu comme la "déviance" d'un modèle (Smith & McKernna, 2013).

Nous avons fait le choix d'évaluer notre modèle en se basant sur le coefficient de détermination observé (R^2_{obs}) de la régression entre la pression prédite par interpolation et la pression réellement observée. Cependant, on intègre le fait que les données d'observations soient binomiales et limitées en nombre d'observations réalisées sur chaque parcelle, en calculant un coefficient de détermination pondéré par le nombre d'observations ($R^2_{obs.w}$).

$$R^2_{obs.w} = \frac{\sum_{i=1}^{n}\left(nObs_i * (P_i - P'_i)^2\right)}{\sum_{i=1}^{n}\left(nObs_i * (P_i - \overline{P})^2\right)}$$

avec:
P'_i : la pression prédite sur la parcelle *i*
P_i : la pression observée sur la parcelle *i*
$\overline{P}$: moyenne des pressions observées
$nObs_i$: nombre d'observations réalisé sur la parcelle *i*

Pour mieux cerner la justesse de notre R^2_{obs}, nous allons le comparer à un R^2optimal (R^2_{opt}) qui correspondrait aux prédictions que l'on pourrait faire avec notre modèle dans l'hypothèse où celui-ci serait parfait. Le R^2_{opt} est déterminé par simulation. On simule des observations, c'est-à-dire les valeurs de succès (*nPos)* ou d'échec, obtenues dans chacune des parcelles à partir de la pression prédite. A partir de cette simulation on peut recalculer une pression pseudo-observée *P"* que l'on va comparer à notre pression prédite par interpolation *P'* en calculant le R^2_{simu}. On répète cette opération 1000 fois pour obtenir une distribution des R^2_{simu}. Le R^2_{opt} correspond au R^2 médian de cette distribution. On peut donc regarder comment se place notre R^2_{obs} dans la distribution des R^2_{simu} obtenus avec la simulation et par rapport au R^2_{opt} en calculant le ratio R^2_{obs} / R^2_{opt} (Figure 11). Plus celui-ci sera proche de 1, mieux la prédiction de l'interpolation est jugée correcte. Notons qu'il est possible que le R^2_{obs} soit supérieur au R^2_{opt} *tant* qu'il appartient à la distribution des R^2_{simu}. . Le diagramme d'une simulation est présenté en annexe 4.

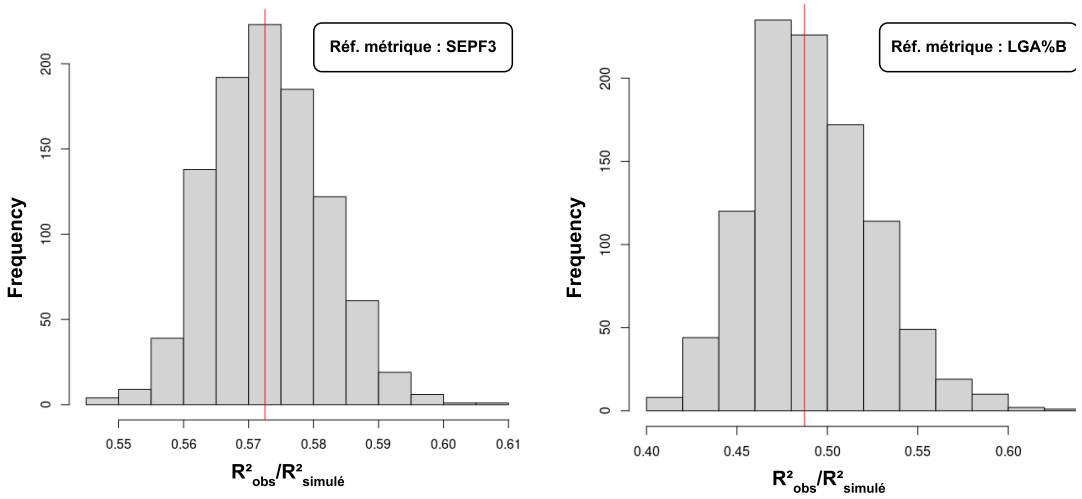


Figure 11 - Distribution du ratio R^2_{obs} sur la distribution de 1000 R^2_{simu}

Les résultats de l'analyse statistique du modèle d'interpolation sont présentés dans les figures 12 et 13. Pour rappel, l'analyse est réalisée par rapport aux métriques de référence de chaque bioagresseur ainsi que pour les 2 métriques permettant d'illustrer ce mémoire. Les R^2_{obs} sont pour 75% supérieurs à 0,10. On voit que la prise en compte du nombre d'observations dans le $R^2_{obs.w}$ améliore nettement la distribution avec plus de 50% au-dessus de 0,25. Pour un 1er modèle qui traite de biologie environnementale et généralisé pour une multitude de bioagresseurs cela est très convenable. La quantité de données d'observations semble également jouer sur la valeur de la régression, le $R^2_{obs.w}$ pour *SEPF3* étant de 0,8 point supérieurs à *LGA%B.* Mais le ratio avec le $R^2_{opt.w}$ n'est pas significativement différent pour ces 2 bioagresseurs (environ à 0,55). Par ailleurs, 50% de la distribution de ce ratio est supérieur à 0,5.

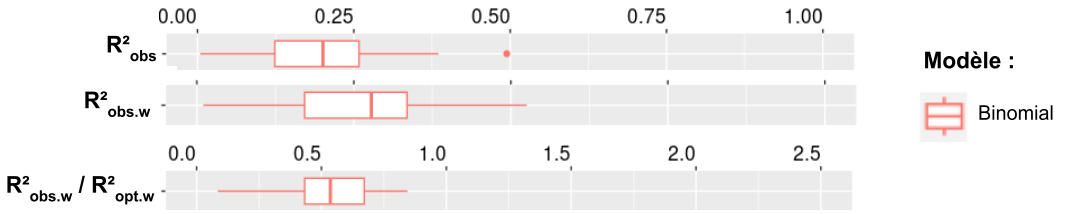


Figure 12 - Résultat de l'évaluation des interpolations produites à partir des modèles construit sur la loi binomiale agrégé pour l'ensemble des métriques de référence des bioagresseurs du projet MoCoRiBA

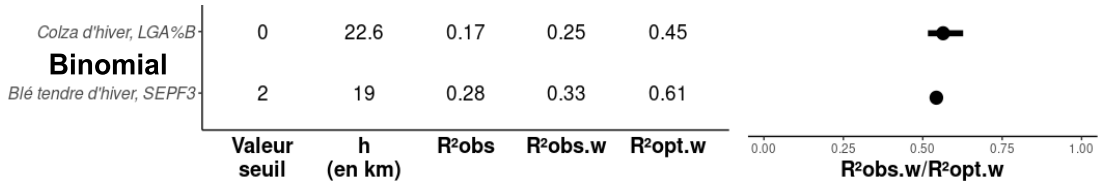


Figure 13 - Détail des résultats de l'évaluation des interpolations réalisées à partir du modèle construit sur la loi binomiale pour la septoriose du blé (SEPF3) et la grosse altise du colza (LGA%B). Définit ici les $R^2_{obs.w}$ et $R^2_{opt.w}$

- Création des visuels cartographiques avec incertitude théorique de l'interpolation

L'évaluation ci-dessus du modèle ne porte que sur la prédiction au niveau des parcelles observées Quand on s'éloigne des zones ayant de la donnée, l'incertitude doit augmenter théoriquement du fait du plus petit nombre d'observations. La représentation cartographique permet de visualiser les variations spatiales de cette incertitude théorique telle que définie plus haut (3.1). Une série de cartes est donc générée pour diverses cultures, bioagresseurs, métriques et campagnes (Figure 14).

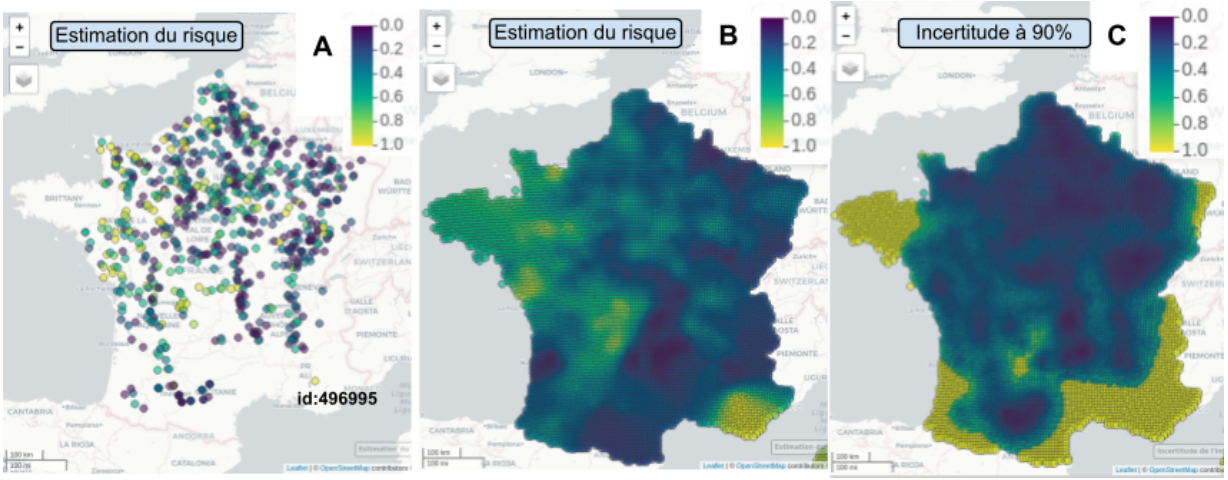


Figure 14 - Cartes de la pression de la septoriose du blé (SEPF3) pour l'année 2017 dans les parcelles suivies en 2017 (A) ; de l'interpolation à l'échelle nationale sur la grille SAFRAN avec le modèle binomial (B) et de l'incertitude (C)

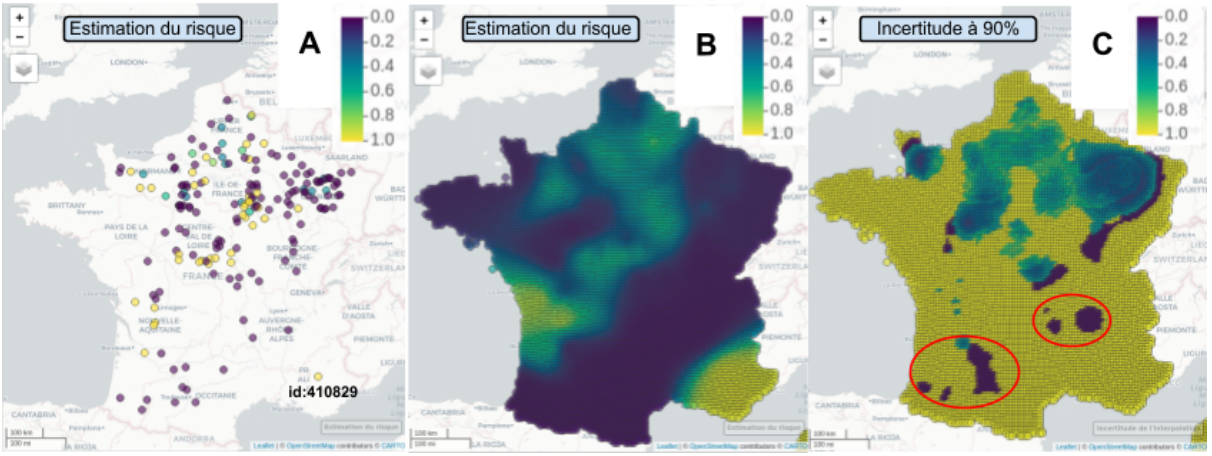


Figure 15 - Cartes de la pression de la grosse altise d'hiver du colza (LGA%B) pour l'année 2016 dans les parcelles suivies en 2016 (A) ; de l'interpolation à l'échelle nationale sur la grille SAFRAN avec le modèle binomial (B) et de la taille de l'intervalle de confiance à 90% (C)

L'analyse de ces cartes permet d'identifier plusieurs limites qui n'étaient pas visibles jusqu'à maintenant. Celles-ci sont détaillées dans la partie suivante.

- Limites du modèle

Les cartes d'interpolation permettent d'identifier des limites de notre modèle :

- Une parcelle isolée sur le territoire va imposer sur l'espace environnant sa valeur de pression du bioagresseur. Cette limite est d'autant plus problématique que la valeur de pression peut être à 0 ou à 1 s'il n'y a qu'une observation. Il est donc très exagéré de dire que des régions seraient totalement indemnes du bioagresseur ou inversement. Cette limite est bien illustrée sur les cartes de pression avec les parcelles *410829* (Figure 14) et *496995* (Figure 15);
- Une région dans laquelle aucune parcelle d'observation n'est présente se verra imposer la pression du bioagresseur mesurée sur les parcelles les plus proches puisque ce sont elles qui auront le poids le plus important dans le modèle. C'est le cas de la région Bretagne où quelques parcelles imposent une pression faible en bioagresseurs (Figure 14 et 15).
- Il est impossible de prendre correctement en compte les *NPosEff* et *NObsEff* dans le calcul d'incertitude et elle est rapidement maximale dès que l'on s'éloigne des lieux où se trouvent les parcelles d'observations. De ce fait, pour une très grande proportion du territoire métropolitain, dès lors que la métrique n'a que peu de données, l'interpolation perd toute sa pertinence. De plus, on voit sur la carte d'incertitude de la métrique *LGA%B* que quelques parcelles assez isolées peuvent facilement créer autour d'elle une ceinture où l'incertitude est minimale. Pour la métrique de la septoriose, une faible incertitude couvre une proportion beaucoup plus importante du territoire, et l'incertitude devient comme précédemment maximale quand il n'y a plus de parcelles suivies.

Ces limites identifiées poussent à faire évoluer le modèle initial afin de corriger l'interpolation dans les régions ayant peu de données mais aussi pour essayer d'améliorer les statistiques du modèle.

3.2. Utilisation des statistiques bayésiennes avec *a priori* national

- Statistique bayésienne

Pour corriger notre modèle, il faut utiliser un autre pan des statistiques qui sont les statistiques bayésiennes. L'utilisation de la loi binomiale faite jusqu'à maintenant s'intègrait dans les statistiques fréquentistes qui reposent sur une unique loi d'échantillonnage construite en s'appuyant sur des observations. A l'inverse, la statistique bayésienne repose sur l'inférence bayésienne (Dupuis, 2007). C'est-à-dire un degré de croyance basé sur des connaissances *a priori*, telles que des études antérieures.
Soit $\boldsymbol{\theta}$ un paramètre à estimer et $\boldsymbol{x}$ l'ensemble des observations traduisant l'état des connaissances *a priori* sur $\boldsymbol{\theta}$.
Les composantes des statistiques bayésiennes sont :

- l'*a priori* : il correspond à la croyance initiale ou aux connaissances antérieures sur un événement avant d'observer les données.
- La vraisemblance des observations (ou fonction de vraisemblance) : fonction du paramètre $\boldsymbol{\theta}$ contenant l'ensemble de l'information apportée par les données observées.
- La loi de probabilité *a posteriori* : loi de probabilité conditionnelle du paramètre $\boldsymbol{\theta}$ sachant $\boldsymbol{x}$. Elle combine la probabilité *a priori* et la vraisemblance en utilisant le théorème de Bayes.
- Le théorème de Bayes : c'est l'équation centrale pour calculer les probabilités *a posteriori* en fonction de la probabilité *a priori* et de la vraisemblance. Son expression est la suivante :

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{x} \cap \boldsymbol{\theta})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{p(\boldsymbol{x})} \qquad (3.2.a)$$

avec :
$\boldsymbol{\theta}$ le paramètre à estimer
$\mathbf{x} = (x_1, \ldots x_i, \ldots, x_n)$, les observations

Pour le modèle d'interpolation on utilisera la loi conjuguée de la loi binomiale, soit la loi bêta. En statistique bayésienne, cela signifie que si l'a priori suit une distribution donnée, l'*a posteriori* va suivre une distribution de même nature. Ici, soit $\boldsymbol{x}$ notre croyance sur l'évènement $\boldsymbol{\theta}$ (= pression du bioagresseur), on cherche à prédire la distribution *a postériori* de $\boldsymbol{\theta}$ conditionnellement à $\boldsymbol{x}$. La distribution *a priori* de $\boldsymbol{\theta}$ est une loi beta sur [0, 1] et $\boldsymbol{x} \sim B(n, p)$, variable aléatoire binomiale de paramètre *n* et *p*, soit respectivement la taille d'échantillon et la probabilité de succès. Dans ce cas là, la distribution de $\boldsymbol{\theta}$ conditionnellement à $\boldsymbol{x}$ est aussi une distribution bêta (Robert, 2006) :

$$\boldsymbol{\theta}|\boldsymbol{x} \sim Be(\boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad (3.2.b)$$

Où les deux nombres alpha et beta peut être interprétés comme :
$\boldsymbol{\alpha}$: nombre de succès
$\boldsymbol{\beta}$: nombre d'échecs

La moyenne et la variance *a posteriori* de la loi de probabilité sont les suivantes :

$$E[\theta|\boldsymbol{x}] = \frac{\boldsymbol{\alpha}}{\boldsymbol{\alpha}+\boldsymbol{\beta}} \quad \text{et} \quad V[\theta|\boldsymbol{x}] = \frac{\boldsymbol{\alpha\beta}}{(\boldsymbol{\alpha}+\boldsymbol{\beta}+1)(\boldsymbol{\alpha}+\boldsymbol{\beta})^2} \qquad (3.2.c)$$

Une particularité de la loi Bêta est que selon les valeurs des 2 paramètres **α** et ***β***, la forme de la fonction densité de probabilité sera différente, voici quelques exemples :

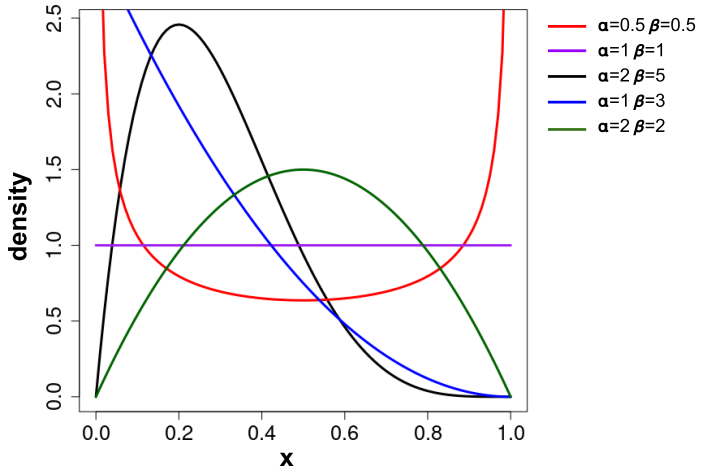


*Figure 16 - Graphique de la fonction densité de la loi bêta pour différents paramètres **α** et ***β***

Lorsque **α** et ***β*** sont tous les deux proches de 1, la courbe sera similaire à une distribution uniforme. À mesure que **α** et ***β*** augmentent, la distribution devient plus resserrée autour de la moyenne, caractérisée par une variance plus faible. Et plus **α** sera grand par rapport à ***β*** (resp. ***β*** par rapport à **α**), plus la distribution se décalera vers 0 (resp. vers 1).

L'*a priori* doit être décrit sous forme de loi beta de paramètre *a* et *b*, correspondant à de pseudo observations reflétant la confiance que l'on a dans l'a priori. L'intégration de l'*a priori* dans la distribution beta *a posteriori* se fait alors simplement par addition des pseudo observations et des observations binomiales :

$$\boldsymbol{\alpha} = a + s \quad \text{et} \quad \boldsymbol{\beta} = b + n - s \qquad (3.2.d)$$

où

a et ***b*** : correspondent respectivement au nombre de pseudo succès et pseudo échecs *a priori*

s : nombre de succès

n : nombre d'observations

Les 2 constituants de l'*a priori* ***a*** et ***b*** sont définis et fixés à partir des connaissances d'études précédentes. La somme de ***a*** + ***b*** correspond à un nombre d'observations dans la loi Béta, soit au poids de notre *a priori* dans la loi de probabilité. La valeur relative des 2 constituants fera pencher le paramètre ***θ*** du côté du succès ou de l'échec lorsqu'il y aura peu d'observations *n*.

- Modèle d'interpolation basé sur une construction bayésienne

Supposons que nous connaissions la pression des bioagresseurs de l'espèce *i* pour 1 campagne donnée, dans un ensemble de *m* parcelles notée P_p^{obs} *(bio_i)*, pour p = 1,.....,m. Le modèle initial (2.4.c) ajusté sur la base de l'inférence bayésienne définit la pression du bioagresseur d'espèce *i* prédite dans la parcelle *p'* par :

$$P_{p'}^{prédite}(bio\_i) = \frac{a_{bio\_i} + NPosEff_{p',bio\_i}}{a_{bio\_i} + b_{bio\_i} + NObsEf_{p',bio\_i}} \quad (3.2.e)$$

qui suit la loi : $Be(a + NPosEff\,,\ b + NObsEff - NPosEff)$
avec :
$\boldsymbol{a}_{bio\_i}$ et $\boldsymbol{b}_{bio\_i}$ notre *a priori* qui dépend du bioagresseur *i*
$NPosEff_{p',bio\_i}$: équation (2.4.d)
$NObsEff_{p',bio\_i}$: équation (2.4.e)

Il faut ensuite définir les valeurs de l'*a priori*. Celui-ci va être spécifique à chaque bioagresseur. Dans un premier temps, nous avons testé l'utilisation de la pression moyenne du bioagresseur au niveau national, calculée sur l'ensemble des données de toutes les années à notre disposition. On pose donc l'équation suivante :

$$\frac{a_{bio\_i}}{a_{bio\_i} + b_{bio\_i}} = \sum_{c=1}^{n} \sum_{p=1}^{m\_c} \frac{nPos_{p,c,bio\_i}}{nObs_{p,c,bio\_i}} \quad (3.2.f)$$

où : *n* est le nombre d'années d'étude et *m_c,* le nombre de parcelles présentes l'année *c*
avec :
$\boldsymbol{a}_{bio\_i}$, $\boldsymbol{b}_{bio\_i}$: correspondent aux paramètres ***a*** et ***b*** de l'*a priori* pour le bioagresseur de l'espèce *i* quelle que soit l'année
$nPos_{p,c,bio\_i}$: nombre d'observations positives pour la parcelle *p* de l'année *c*
$nObs_{p,c,bio\_i}$: nombre d'observations réalisées pour la parcelle *p* de l'année *c*

Pour donner un poids limité à l'*a priori*, nous choisissons de rendre l'*a priori* équivalent à 2 observations, soit que ***a*** + ***b*** = 2. Ce qui permet de déterminer a et b :

$$a_{bio\_i} = 2\hat{y}_{bio\_i} \quad (3.2.g1) \qquad b_{bio\_i} = 2 - 2\hat{y}_{bio\_i} \quad (3.2.g2)$$

avec : $\hat{y}_{bio\_i} = \sum_{c=1}^{n} \sum_{p=1}^{m\_c} \frac{nPos_{p,c,bio\_i}}{nObs_{p,c,bio\_i}}$

Ou : $\hat{y}_{bio\_i}$: estimation moyenne interannuelle nationale de la pression du bioagresseur d'espèce *i*

Enfin, on réalise l'interpolation bayésienne en réitérant les mêmes étapes de calcul que pour le modèle binomial, seulement, on utilise notre nouveau modèle intégrant l'*a priori* calculé préalablement pour chaque bioagresseur d'espèce i (annexe 5 : diagramme des étapes). On calcule l'incertitude à 90% de l'interpolation de manière à précédemment (3.1) en remplaçant l'usage de la fonction *qbinom()* par celui de la fonction *qbeta()* avec les paramètres alpha et beta tels que précédemment déterminés pour calculer les quantiles à 0,05 et 0,95.

- Analyse statistique du modèle

Comme pour le 1er modèle, on détermine des valeurs optimales de l'hyperparamètre *h* pour chacune des métriques. Celle-ci reste cohérente pour la majorité des valeurs avec les valeurs obtenues pour les modèles fréquentistes puisque l'on reste dans les mêmes ordres de grandeur de distance (Figure 17.A). Les valeurs de R^2 associées aux différents *h* sont légèrement dégradées en moyenne, mais pas de manière significative (Figure 17.B et 18).

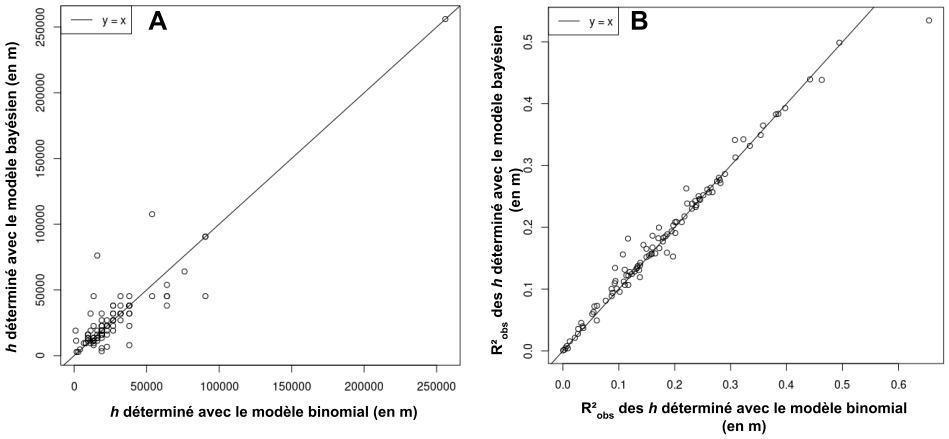


*Figure 17 - Graphiques d'analyse de l'évolution de l'hyperparamètre h (**A**) et des R² associés (**B**) entre le modèle binomial initial et le modèle bayésien*

La figure 18 présente les résultats agrégés de l'analyse statistique du modèle, mis en lien avec les résultats du précédent modèle pour l'ensemble des métriques de référence des bioagresseurs du projet. La médiane des R^2_{obs} reste très similaire à 0,16 mais diminue pour les $R^2_{obs.w}$. Les distributions des ratio R^2_{obs}/R^2_{opt} avec le modèle bayésien est décalée à la baisse avec la médiane de la distribution nettement diminuée de 0,13 point.

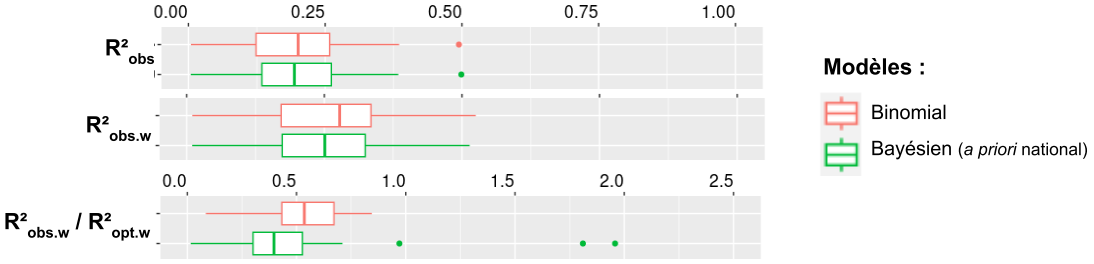


Figure 18 - Résultat de l'étude statistique des interpolations produites à partir des modèles construits sur la loi binomiale et la loi bêta des statistiques bayésiennes avec un a priori unique à l'échelle nationale par bioagresseur, agrégé sur l'ensemble des métriques de référence des bioagresseurs du projet MoCoRiBA

Le $R^2_{obs.w}$ de *SEPF3* n'est pas modifié mais celui de *LGA%B* augmente de 0,03 point (Figure 19). Ces éléments nous font penser que ce sont les $R^2_{obs.w}$ les plus faibles qui ont subi une légère amélioration. La métrique *LGA%B* a un ratio amélioré significativement de 0,38 point et la métrique *SEPF3 à* une tendance plutôt dégradée dont le ratio diminue significativement de 0,15 point.

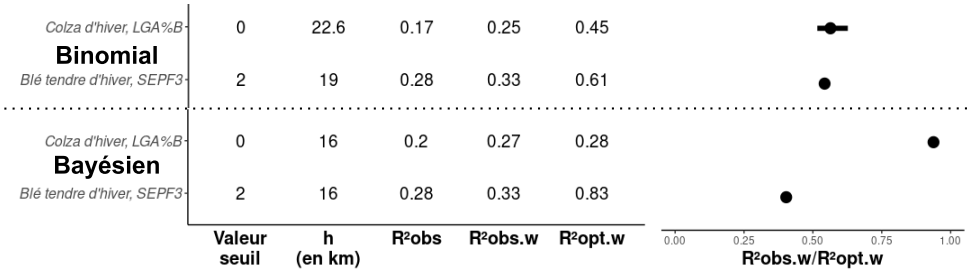


Figure 19 - Résultat de l'étude statistique des interpolations produites à partir des modèles construits sur la loi binomiale et la loi bêta des statistiques bayésiennes pour la septoriose du blé (SEPF3) et la grosse altise du colza (LGA%B)

Bien que dans l'ensemble, les R^2 soient dégradés par ce premier modèle bayésien, il peut y avoir pour certaines métriques comme le pourcentage de colzas buissonnants une amélioration notable.

- Création des visuels cartographiques

Les cartes d'interpolation des métriques de la grosse altises du colza (*LGA%B*) et de la septoriose du blé (*SEPF3*) associées avec leur carte d'incertitudes sont créées à partir du modèle d'interpolation construit sur les statistiques bayésiennes avec un *a priori* (3.2.e).

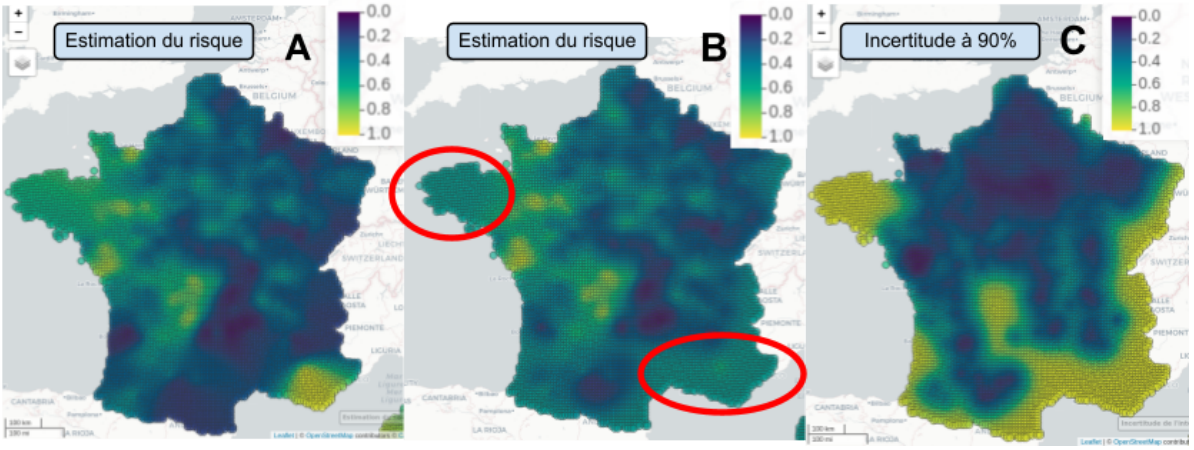


Figure 20 - Cartes de la pression de la septoriose du blé (SEPF3), de l'interpolation de la pression avec le modèle binomial (A) et bayésien avec un a priori (a = 0.82 ; b = 1.17) (B) et des incertitudes théoriques de l'interpolation avec le modèle bayésien (C)

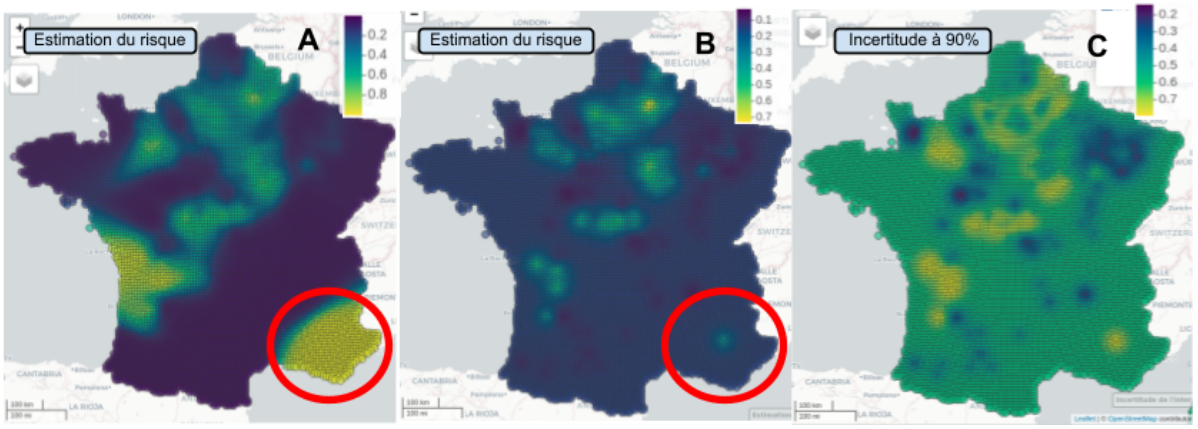


Figure 21 - Cartes de la pression de la grosse altise d'hiver du colza (LGA%B), de l'interpolation à l'échelle nationale du risque avec le modèle binomial (A) et bayésien avec un a priori (a = 0.27 ; b = 1.73) (B) ; des incertitudes théorique de l'interpolation bayésienne (C)

Les limites identifiées avec le modèle initial, construit sur la loi binomiale, ne sont plus observables. C'est très visible pour la région Bretagne (Figure 21) et pour la région PACA (Figure 20 et 21). L'interpolation est nuancée par l'*a priori* et il prend le dessus lorsqu'il y a peu ou pas de parcelles à proximité. On peut aussi voir que la nuance se fait également en limitant le contraste entre les zones ayant une pression forte ou faible.

La carte d'incertitudes de la métrique de la grosse altises du colza présente les incertitudes les plus importantes (proche de 0,8) et les plus faibles (proche de 0,2) dans les aires où les parcelles d'observations sont principalement localisées (Figure 20.C). Cela correspond également aux régions où la pression des ravageurs est respectivement la plus importante (incertitude élevée du fait d'un groupe de parcelles à la fois avec une forte et une

faible pression de bioagresseurs) ou la plus faible (zone où sont majoritairement présentes des parcelles avec une faible pression de bioagresseurs). L'incertitude sur le reste de la France est à environ 0,6. Elle correspond à toute la surface où l'*a priori* (***a***=0,27 et ***b***=1,73) prend le dessus lors de l'interpolation par manque de parcelles d'observation. L'*a priori* ne reflétant pas la réalité, mais plutôt une approximation par rapport à une moyenne nationale.

La métrique du blé *SEPF3*, qui a beaucoup de parcelles d'observation à l'échelle de la France, présente une incertitude très faible (inférieure à 0,3) pour une grande partie du territoire. Par contre, l'incertitude augmente fortement (environ 0,9) dans plusieurs endroits qui correspondent aux zones dans lesquelles il y a moins de parcelles suivies, et où notre *a priori* devient prépondérant.

- Limites du modèle

Le mode de calcul de l'incertitude présente une limite au niveau des territoires où la pression est uniquement définie par l'*a priori*, ce qui est systématiquement le cas en Bretagne et Pays de la Loire où les données des campagnes antérieures à 2019 n'ont pas été remontées au niveau national. Il est à noter par ailleurs que l'incertitude par défaut de l'estimation est dépendante de la valeur de l'estimation. On aurait pu s'attendre à ce que loin des points d'observation, l'incertitude des valeurs soit similaire pour les 2 métriques étudiées, mais celle-ci est en fait liée à la valeur de l'*a priori*. Plus la pression du bioagresseur se rapproche de 0.5, plus l'incertitude sera grande et inversement lorsque la pression par défaut tend vers 0 ou 1 (Figure 22). Cela s'explique par la forme de la fonction densité de la loi qui va être plus ou moins resserrée autour de la moyenne.

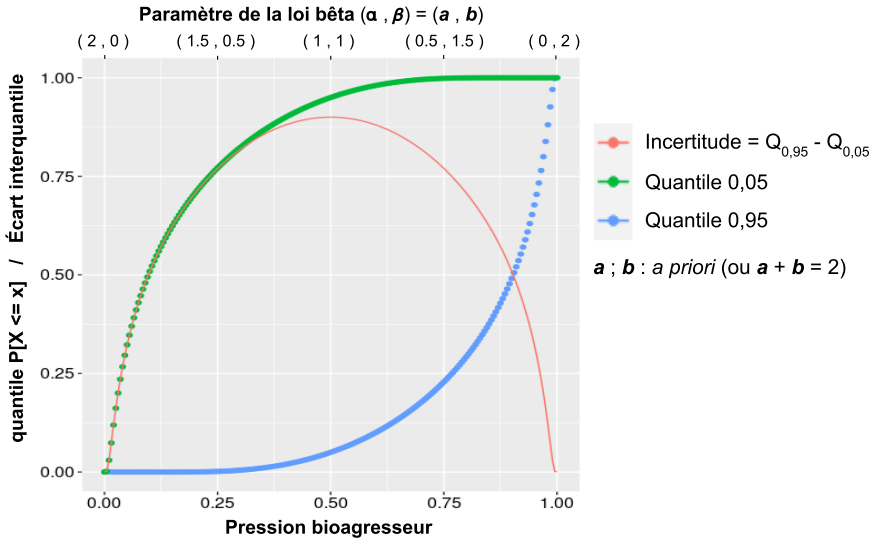


Figure 22 - Évolution des quantiles 0,05 et 0,95 et de l'écart interquantile suivant la valeur de l'a priori donnant la pression du bioagresseur, pour un poids de 2 pseudo observations

Par ailleurs, la construction d'un *a priori* unique quelle que soit la localisation en France est peut-être aberrante suivant la région dans laquelle l'on se situe. On pourrait se dire qu'on a dans le Massif Central une intensité d'attaque d'un insecte ravageur du colza plus faible que dans les grandes plaines céréalières de la Beauce.

3.3. Adoption d'un *a priori* local pour le modèle bayésien

Le modèle d'interpolation reposant sur les statistiques bayésiennes a permis d'améliorer les calculs d'incertitude théorique et dans certains cas d'améliorer la prédiction, afin de le rendre plus pertinent, on souhaite modifier l'*a priori*. Le nouvel *a priori* ne sera plus issu d'une moyenne interannuelle nationale de la pression des bioagresseurs, mais il deviendra local. Il correspondra à la moyenne interannuelle locale de la pression des bioagresseurs. Cela signifie que l'*a priori* défini par ***a*** et ***b*** n'est plus unique pour un même bioagresseur, mais il prendra plusieurs valeurs selon la position géographique.

Les cartes de la figure 23 illustrent la moyenne interannuelle de la pression des bioagresseurs des métriques *SEPF3* et *LGA%B*. On voit que certaines zones sont sujettes à une pression plus importante que d'autres pour les 2 bioagresseurs. Donc, il est bien intéressant d'implémenter dans notre modèle un *a priori* local, même si le contraste entre zones à plus faible ou plus haut risque est mieux marqué pour les métriques ayant beaucoup de données.

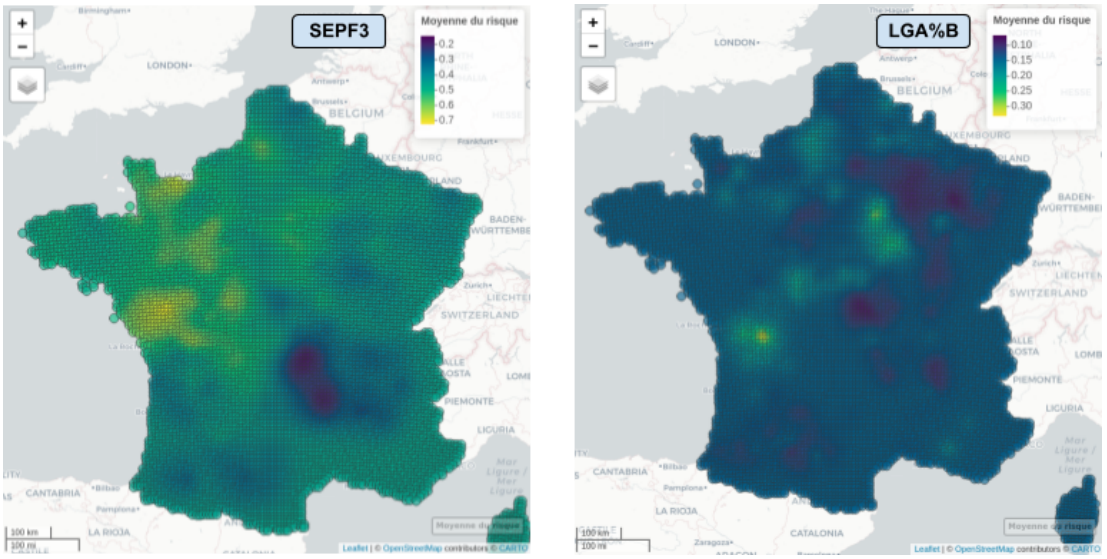


Figure 23 - Cartes de l'interpolation moyenne interannuelle de la pression des bioagresseurs de la septoriose du blé (SEPF3) et de la grosse altise d'hiver du colza (LGA%B) avec le modèle bayésien

- Modèle d'interpolation bayésien basé sur un *a priori* local

Supposons que nous connaissons la pression d'un bioagresseur de l'espèce *i* pour 1 campagne donnée, dans un ensemble de *m* parcelles notées $P_p^{réelle}(bio\_i)$, pour p = 1,.....,m. La pression du bioagresseur de l'espèce *i* pour l'année donnée prédite par notre modèle d'interpolation bayésien amélioré dans la parcelle *p'* est donnée par :

$$P_{p'}^{prédite}(bio\_i) = \frac{a_{bio\_i,p'} + NPosEff_{p',bio\_i}}{a_{bio\_i,p'} + b_{bio\_i,p'} + NObsEff_{p',bio\_i}} \qquad (3.3.a)$$

avec :

$\boldsymbol{a_{bio\_i,p'}}$ et $\boldsymbol{b_{bio\_i,p'}}$: forment l'*a priori* qui dépend du bioagresseur *i* et de la localisation de la parcelle *p'*

$NPosEff_{p',bio\_i}$: équation (2.4.d)

$NObsEff_{p',bio\_i}$: équation (2.4.e)

Il faut dans un premier temps calculer les nouvelles valeurs locales ***a*** et ***b*** de l'*a priori* pour le calcul de l'hyperparamètre *h* optimal du bioagresseur. Ces dernières sont spécifiques à chaque parcelle d'observation. Les parcelles ne sont présentes qu'une seule campagne *n* dans le jeu de données. Il est important d'exclure la campagne *n* de la moyenne interannuelle locale pour éviter de fournir, dans l'*a priori,* de l'information sur la pression réelle de la campagne *n,* que la méthode de recherche du *h* va essayer de prédire par interpolation ensuite. Pour ce faire, on se base uniquement sur une moyenne des interpolations faites à ce point pour les autres années d'étude.
Pour connaître les paramètres de l'*a priori* au niveau d'une parcelle *p* (présente lors de la campagne c') et pour le bioagresseur d'espèce *i*, on calcule l'interpolation du risque au niveau de cette parcelle pour chaque campagne de notre jeu de données (2009-2022).
Puis on calcule ***a*** et ***b*** sur la base des équations (3.2.f) comme définies par les formules suivantes :

$$a_{bio\_i,p} = 2\hat{y}_{bio\_i,p} \quad (3.3.b1) \qquad b_{bio\_i,p} = 2 - 2\hat{y}_{bio\_i,p} \quad (3.3.b2)$$

où :
$$\hat{y}_{bio\_i,p} = \sum_{c=1}^{n} P_p^{prédite}(bio\_i, c) \quad (3.3.c)$$

avec *c* qui parcourt les campagnes de 1 à n à l'exception de *c'*

avec :
$a_{bio\_i,p}$, $b_{bio\_i,p}$: correspondent aux paramètres ***a*** et ***b*** pour le bioagresseur de l'espèce *i* au niveau de la parcelle *p*
$P_p^{prédite}(bio\_i)$: correspond à la prédiction de la pression du bioagresseur d'espèce *i* pour la campagne *c*

Pour la visualisation en carte, il faut calculer les paramètres ***a*** et ***b*** pour chacune des mailles SAFRAN. Ils sont calculés de la même manière que précédemment, mais avec la moyenne interannuelle de la pression des bioagresseurs au niveau de chaque maille d'après l'interpolation avec le modèle bayésien initial (3.2.e).

On réitère ensuite les mêmes étapes du calcul d'interpolation et des incertitudes avec cette fois-ci des *a priori* locaux. La démarche des étapes de calcul est illustrée dans un diagramme en annexe 6.

- Analyse statistique du modèle

L'analyse de l'hyperparamètre *h* pour chacune des métriques montre qu'il reste similaire entre le modèle bayésien de base et celui avec l'*a priori* local pour la plupart des métriques. Toutefois, il y a quelques métriques qui voient leur *h* diminuer fortement avec ce changement de modèle, pouvant arriver à la valeur minimale, fixée à 1 km, pour *h*. Cela concerne des métriques pour lesquelles le R^2_{obs} était inférieur à 0,01 avec les modèles précédents. Dans ce cas, l'*a priori* local (seul) devient une meilleure estimation de la métrique que l'interpolation basée sur les autres parcelles. C'est pourquoi on observe une diminution du *h*, qui permet de réduire la distance jusqu'à laquelle on juge pertinent d'accorder un poids non nul à une parcelle dans l'interpolation.

Les R^2_{obs} associés au modèle bayésien basé sur l'*a priori* local ont une tendance globale à une amélioration avec un décalage de leur distribution de 0,02 point à la hausse par rapport aux 2 modèles précédents (Figure 25). C'est certainement dû au fait d'une

amélioration de plusieurs R^2_{obs} qui étaient jusqu'à présent assez faible pour plusieurs métriques (Figure 24.B).

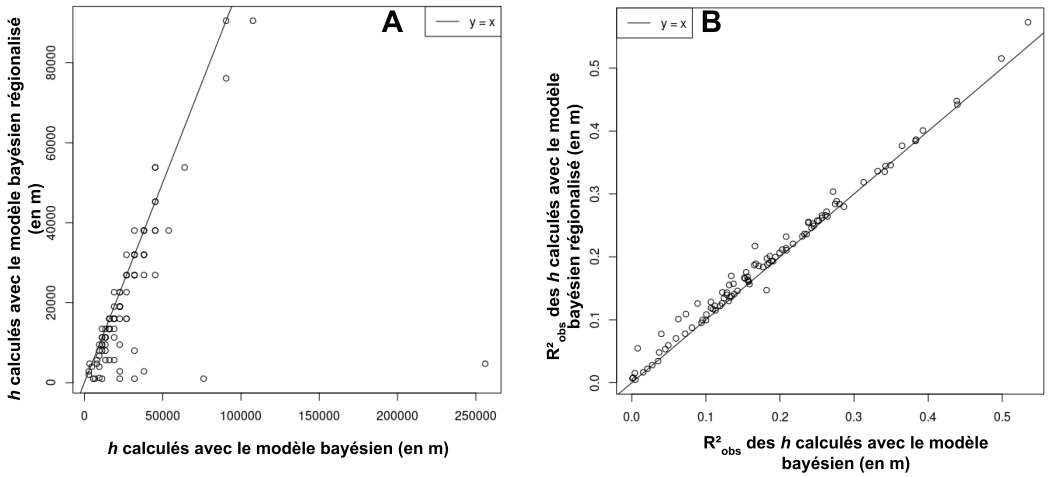


Figure 24 - Graphiques d'analyse de l'évolution de l'hyperparamètre h de chacunes des métriques (A) et des R² associés (B) entre le modèle bayésien initial et le modèle bayésien avec un a priori local

La distribution des $R^2_{obs,w}$ évolue peu avec le nouvelle *a priori* localisé du modèle bayésien. La distribution du ratio $R^2_{obs.w}/R^2_{opt.w}$ est décalée à la hausse de 0,06 point pour un même écart interquartile par rapport au modèle bayésien initial. Pour ce qui est de la comparaison avec le modèle binomial, la distribution est maintenant décalée à la baisse avec un écart interquartile similaire, une médiane diminué de 0,1 point et des valeurs extrêmes dont plus aucune n'est à 0 et certaines dépassant même 1.

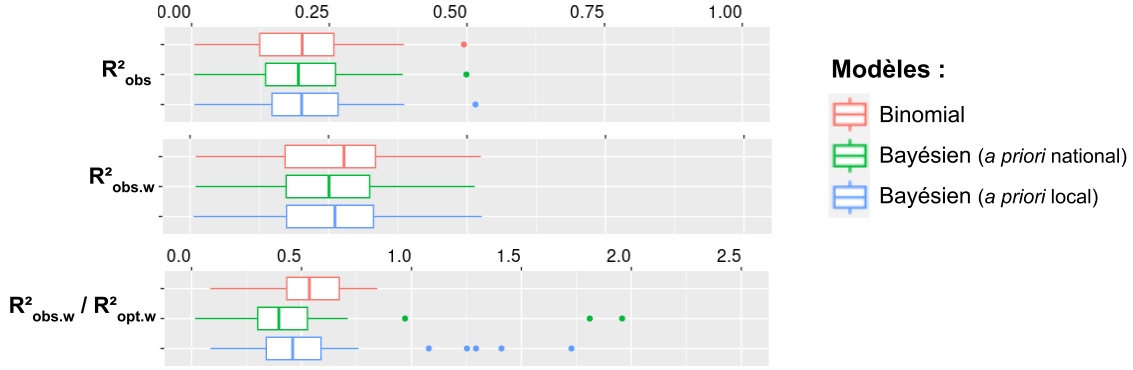


Figure 25 - Résultat de l'étude statistique des interpolations produites à partir des modèles construits sur les loi binomiale et loi bêta des statistiques bayésiennes avec un a priori unique à l'échelle nationale ou locale, agrégé pour l'ensemble des métriques de référence des bioagresseurs du projet MoCoRiBA

Les R^2_{obs} deviennent comparables avec une légère tendance à l'amélioration en bayésien localisé par rapport au binomial, cette tendance est cependant inversée pour le R^2_{obs} pondéré. Cela indique que le modèle bayésien localisé tend à être meilleur pour les points entourés de nombreuses observations que pour les points entourés de peu d'observations. C'est attendu vu que les modèles bayésiens sont essentiellement là pour corriger des situations avec peu d'observations mais ajoutent une information peu pertinente lorsqu'il y a beaucoup d'observations.

Les résultats pour les 2 métriques illustrées restent très similaires avec l'évolution de l'*a priori*. Le changement de modèle de base (loi binomiale ou loi bêta) a eu le plus gros effet

d'amélioration pour les prédictions de la grosse altise d'hiver du colza (+0,04 point pour le $R^2_{obs.w}$ et +0,27 point pour le ratio $R^2_{obs.w}/R^2_{opt.w}$) et une dégradation partielle pour la septoriose du blé ($R^2_{obs.w}$ similaire et -0,15 pour le ratio $R^2_{obs.w}/R^2_{opt.w}$).

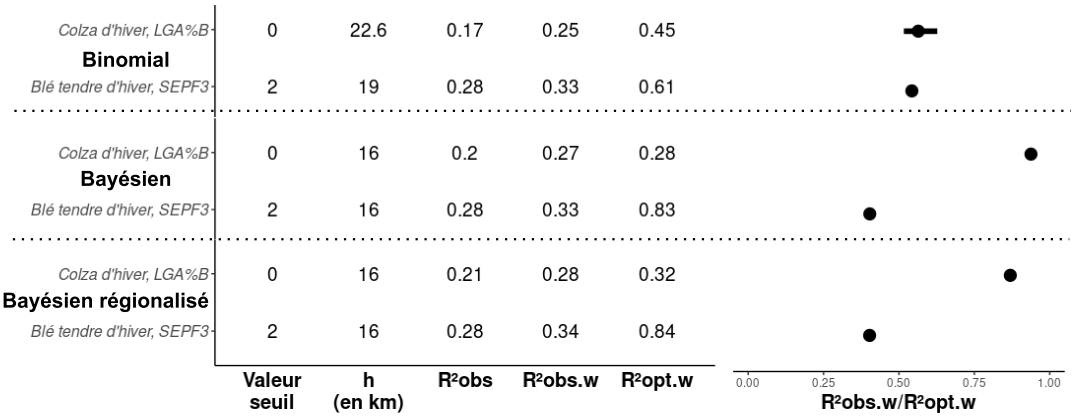


Figure 26 - Résultat de l'évaluation des interpolations produites à partir des modèles construits sur la loi binomiale et la loi bêta des statistiques bayésiennes pour un a priori unique à l'échelle nationale ou locale pour la septoriose du blé (SEPF3) et la grosse altise du colza (LGA%B)

- Création des visuels

Les cartes d'interpolation des métriques *LGA%B* et *SEPF3* associées avec leur carte d'incertitudes sont créées à partir du modèle bayésien à l'*a priori* localisé (3.3.a). Celles-ci comportent des évolutions mineures très peu visibles, c'est pourquoi elles ne sont pas présentées ici mais sont disponibles en annexes 7 et 8. Le léger changement, plus visible pour la carte d'interpolation de la septoriose, est le rajout de contraste entre les zones totalement indemnes (pression qui tend vers 0) et les zones fortement touchées (pression qui tend vers 1). L'incertitude est également très légèrement améliorée par endroits.

- Limites du modèle

La principale limite restante concerne la construction de l'*a priori* local. En effet, celui-ci est construit à partir de l'*a priori* unique du 1er modèle bayésien qui peut être potentiellement aberrant dans certaines zones sans ou avec très peu d'observations, mais conservé tout de même pour ces régions. Dans ces territoires, l'*a priori* local est donc similaire à l'*a priori* initial et unique à l'échelle nationale.

3.4. Discussion sur le modèle d'interpolation

La réflexion menée sur l'amélioration du modèle d'interpolation de la pression moyenne annuelle des bioagresseurs du projet MoCoRiBA a permis de bien avancer sur le projet. Le modèle construit sur les bases des statistiques bayésiennes intègre maintenant un *a priori* traduisant une première base de connaissances sur ce que l'on pense être la pression d'un bioagresseur pour un lieu donné. Cette connaissance pèse de moins en moins dans la prédiction finale avec l'augmentation du nombre de parcelles d'observations environnantes présentes dans notre base de données d'épidémiosurveillance. Inversement lorsque la quantité de données diminue, la croyance initiale a un poids plus important, mais l'incertitude en est également augmentée.

Cependant, l'estimation de l'incertitude de la pression des bioagresseurs prédite par interpolation n'est pas totalement convenable. En effet, on a vu avec le 1er modèle bayésien

que l'incertitude actuelle, calculée sur la base des quantiles des fonctions densité, dépendait autant du nombre d'observations prises en compte, que de la valeur de la pression estimée. Ce qui est problématique dans les régions pauvres en données où l'*a priori* est la seule donnée puisque suivant sa valeur, l'incertitude sera différente (Figure 23) .

On pourrait aussi se questionner sur le poids accordé à l'*a priori* dans le calcul. Celui-ci est actuellement fixé à 2 observations, mais peut-être que le diminuer ou l'augmenter pourrait être pertinent. Il faudrait mener une étude plus approfondie où l'on fait varier le poids accordé à l'*a priori* et regarder comment évolue la précision de l'interpolation,de la même manière que pour l'hyperparamètre *h* et être différent selon les bioagresseurs. De plus, la construction actuelle de l'*a priori* ne permet pas d'avoir une connaissance préalable locale dans toutes les régions de France lorsque peu de suivis y sont réalisés, comme c'est le cas pour de nombreuses métriques en Bretagne et Pays de la Loire. Le poids de l'*a priori* pourrait donc y être diminué en faveur des quelques observations potentiellement présentes. Par ailleurs, cette diminution permettrait aussi de résoudre en partie le problème lié aux valeurs d'incertitudes développé précédemment.

Une limite intrinsèque au modèle d'interpolation est qu'il est très dépendant de la quantité de données disponible pour une campagne. Dans le cadre d'une nouvelle campagne, les données issues de la base Vigicultures® (principale source de données), sont généralement transmises tardivement dans la campagne, voire après la campagne. Ce problème fait que seul un nombre restreint de parcelles provenant d'autres sources de données d'épidémio-surveillance sont disponibles. Les parcelles étant très clairsemées sur le territoire, il en ressort une interpolation basée uniquement sur l'*a priori*. D'où l'importance d'avoir un bon *a priori* le plus représentatif de sa localité pour une campagne donnée.

Une nouvelle méthode de construction de l'*a priori* serait envisageable, surtout si cela permet de mieux prédire ce qu'il se passe dans les territoires vides de données d'observations. C'est ce que nous allons étudier dans le chapitre 4 avec de la modélisation statistique en intégrant des données climatiques et paysagères mais avant cela, je vais indiquer comment j'ai intégré les résultats de cette modélisation dans l'application web MoCoRiBA.

3.5. Implémentation du nouveau modèle dans l'application web MoCoRiBA

L'application web MoCoRiBA proposait jusqu'à présent des indications de pressions calculées avec le modèle d'interpolation développé par Cisse A. Il a fallu mettre à jour l'application avec le nouveau modèle d'interpolation prenant en compte un *a priori* dans les calculs. Cependant, le fonctionnement d'une application a des spécificités propres qui peuvent être très différentes de ce qui est faisable en développement. Une application doit répondre aux exigences des utilisateurs, mais elle est aussi contrainte par les capacités du serveur web hébergeant l'application. En effet, un utilisateur va vouloir une réponse rapide de l'interface lors d'une quelconque action. C'est ce qu'on appelle fluidifier l'expérience utilisateur. Il faut donc que l'ensemble des processus qui fonctionnent en arrière-plan soit optimisés et efficaces.

Dans le cas des données d'interpolation, on pourrait se dire que le plus efficace serait d'aller chercher directement la donnée de la pression du bioagresseur souhaité dans une base de données où tout est pré-calculé à l'avance. Cependant, la taille de cette base de données serait assez énorme et trop volumineuse pour les capacités du serveur. L'autre option est de faire le calcul en temps réel mais de manière optimisée pour satisfaire l'expérience utilisateur. Le choix fait afin de répondre aux différentes exigences et contraintes techniques, a été de recourir aux 2 méthodes. Pour rappel l'application permet

une comparaison avec la base de données des fermes DEPHY. Pour cela, elle sélectionne les fermes du réseau se trouvant dans des contextes agro-climatiques similaires à la localisation de l'exploitation cible. Cet ensemble de fermes correspond à la base de référence comparative. Pour l'ensemble du réseau DEPHY, les données de pressions des bioagresseurs sont directement disponibles dans une base de données dans laquelle l'application pourra instantanément aller récupérer les informations. Mais pour l'exploitation ou la commune cible, le calcul se fait en temps réel en se référant à 3 bases de données contenant les informations nécessaires à l'interpolation :

- les observations d'épidémiosurveillance.
- la valeur des hyperparamètres *h* de chaque métrique.
- la valeur des *a priori* locaux pour chaque métrique et maille SAFRAN.

De plus, il a fallu implémenter la visualisation de l'incertitude théorique de notre interpolation caractérisée par les quantiles 5% et 95% qui n'avait jusque là pas d'équivalent dans l'application.

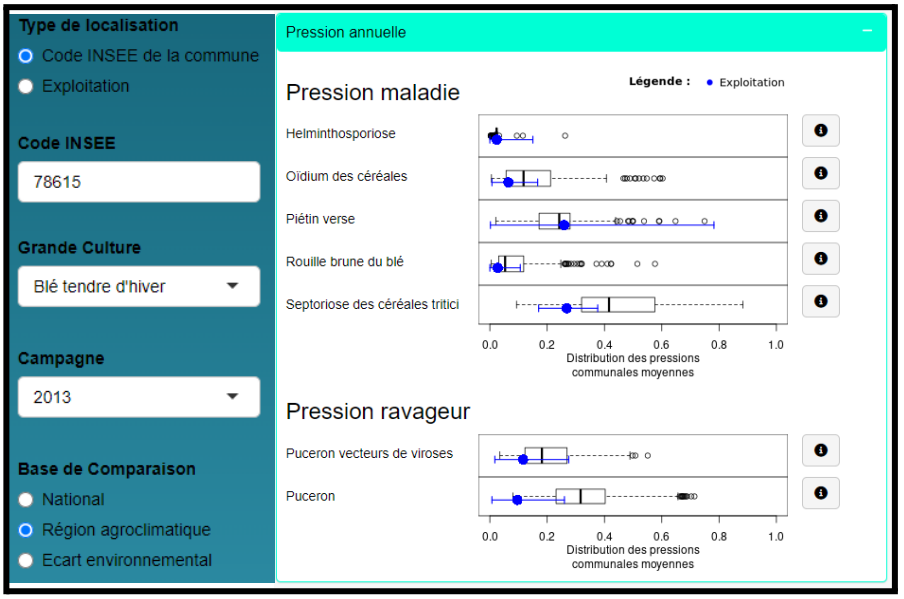


Figure 27 - Interface de l'application MoCoRiBA présentant les données de pression des bioagresseurs du blé tendre d'hiver pour l'année 2013 au niveau de la localisation de l'exploitation, points et intervales bleus (code INSEE 78615), et de la distribution des pressions des bioagresseurs relevée au niveau des fermes du réseau DEPHY étant dans un contexte agroclimatique similaire (414 exploitations, boxplot en noir). Les données de pression des bioagresseurs sont représentées pour l'exploitation cible en bleu et la distribution pour les fermes DEPHY en noire.

Pour réaliser ce travail, j'ai dû me familiariser avec le package R shiny qui facilite le développement d'applications web en R, j'ai aussi commencé à me familiariser avec l'utilisation de calculs en parallèle pour améliorer la réactivité de l'application.

4. Modélisation statistique avec des données climatiques et paysagères

Les modèles statistiques établissent un lien entre des variables explicatives et une variable observée. En fournissant ces informations au modèle, celui-ci va de lui-même essayer de trouver les liens entre les différentes variables et la pression des bioagresseurs observée dans les parcelles. Une fois le modèle construit, il suffira de lui fournir des données correspondant aux variables explicatives pour connaître la prédiction du modèle.

Ce type de modèle présente plusieurs intérêts puisque par ajustement sur les données collectées dans diverses régions, il pourrait être capable de prédire la présence d'un bioagresseur (a) dans les régions où aucune donnée n'est collectée (Bretagne, Pays de la Loire) et (b) pour une nouvelle campagne pour laquelle très peu de données sont disponibles. Les données Vigicultures® pouvant nous être transmises tard dans la campagne.

Les modèles statistiques pourraient donc permettre de pallier aux limites identifiées avec le modèle d'interpolation dans le chapitre précédent. Plusieurs modèles statistiques avec des caractéristiques différentes dans leur méthode d'analyse vont être étudiés dans cette partie afin de déterminer celui/ceux produisant la meilleure prédiction.

La démarche consiste dans un premier temps en une étape d'exploration pour examiner le comportement des différents modèles et des différentes variables explicatives utilisables en entrée. Cette étape permet de déterminer les modèles qui se dégagent comme étant les meilleurs et les variables explicatives qui semblent les plus pertinentes. La seconde étape consiste à construire la démarche pour intégrer le modèle statistique sélectionné avec le modèle d'interpolation afin de prédire la pression des bioagresseurs.

4.1. Présentation des modèles statistiques étudiés

Il existe une diversité de modèles statistiques basés sur des algorithmes d'analyse différents. Des modèles reposant sur 2 grandes catégories sont testés : des modèles de régression linéaire (LASSO, GAM-LASSO et MARS) et des modèles basés sur la construction d'arbres de décision (CART et Random Forest). Un dernier modèle alliant à la fois des arbres de décision et de la régression linéaire est testé (Cubist).

- LASSO

Le modèle LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1997) est une méthode de régression linéaire généralisée permettant une sélection de variables automatique par rapport à la régression linéaire classique. Une pénalité est introduite pour forcer certains coefficients de régression à devenir exactement égaux à zéro. La force de la pénalité est contrôlée par un paramètre appelé lambda. Plus lambda est élevé, plus la pénalité est forte, et plus le modèle Lasso aura tendance à réduire les coefficients vers zéro. Le choix optimal de l'hyperparamètre lambda peut être déterminé par des méthodes de validation croisée.
Le modèle LASSO est utilisé à l'aide du paquet *glmnet de* R (Friedman et al., 2010).

- GAM-LASSO

Le GAM-LASSO (Generalized Additive Model with Lasso) est une extension du modèle Lasso à des modèles additifs généralisés (GAM). Il combine les avantages des modèles additifs généralisés avec la pénalité du LASSO pour la sélection de variables

(Ghosal & Matthias, 2019). Cela permet de modéliser des relations complexes et non linéaires entre les variables explicatives et la variable réponse, tout en intégrant la parcimonie induite par le Lasso pour sélectionner automatiquement les variables importantes.
Le modèle GAM-LASSO est utilisé à l'aide du paquet *plsmselect* de R (Ghosal & Matthias, 2019).

- MARS

Le modèle MARS (Multivariate Adaptive Regression Splines)(Friedman, 1991) représente la relation entre les variables explicatives et la variable réponse en utilisant des morceaux de fonctions linéaires, appelés splines. Ces splines sont construites de manière adaptative en fonction de la structure des données (Figure 28).
Le modèle MARS est utilisé à l'aide du paquet *earth* de R (Milborrow, 2011).

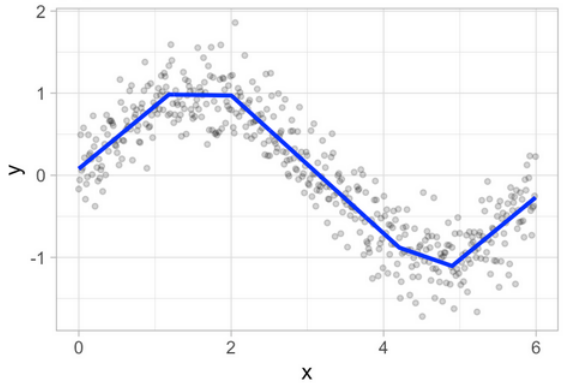


Figure 28 - Exemple de spline de régression ajustée à des données (source : https://bradleyboehmke.github.io/HOML/mars.html)

- CART

Le modèle CART (Classification and Regression Trees) (Breiman, 1984) est un modèle à règles de décision. Il crée une division des données en sous-ensembles en fonction des variables et de valeurs de seuil. Les divisions les plus performantes sont retenues et le processus se répète jusqu'à obtenir le résultat idéal. Il en résulte un arbre de décision (Figure 29) représenté par une série de divisions binaires débouchant sur des nœuds terminaux qui peuvent être décrits par un ensemble de règles spécifiques.
Le modèle CART est utilisé à l'aide du paquet *bagged CART* de R.

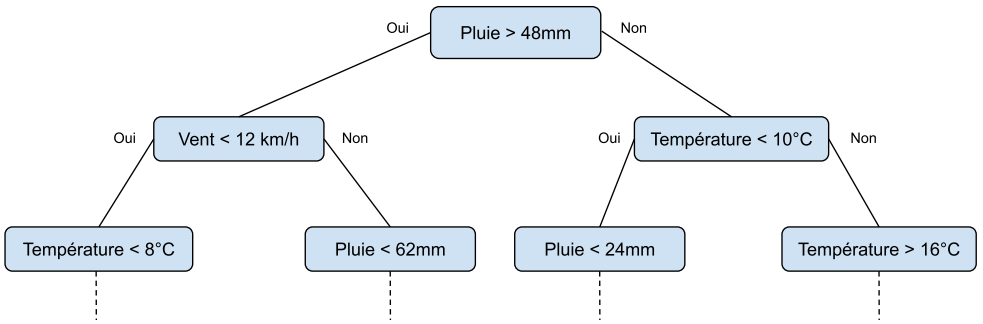


Figure 29 - Exemple d'arbre de décision basé sur des données météorologiques de précipitations, de températures et de vent

- Random Forest

Le modèle Random Forest (Breiman, 2001) est un algorithme basé sur un ensemble d'arbres de décision. Cela permet d'améliorer la précision des prédictions par rapport à un seul arbre de décision comme pour le modèle CART. Les prédictions de chaque arbre sont combinées pour produire une prédiction finale. Ce modèle a la capacité de gérer un grand nombre de variables et de trouver des relations non linéaires complexes entre-elles.
Le modèle Random Forest est utilisé à l'aide du paquet *randomForest* de R (RColorBrewer & Liam, 2018).

- Cubist

Le modèle Cubist (Quinlan, 1992) est un modèle qui combine des techniques d'arbres de décision avec des méthodes de régression linéaire (Figure 30).
Le modèle Cubiste est utilisé à l'aide du paquet *Cubist* de R (Kuhn, 2023).

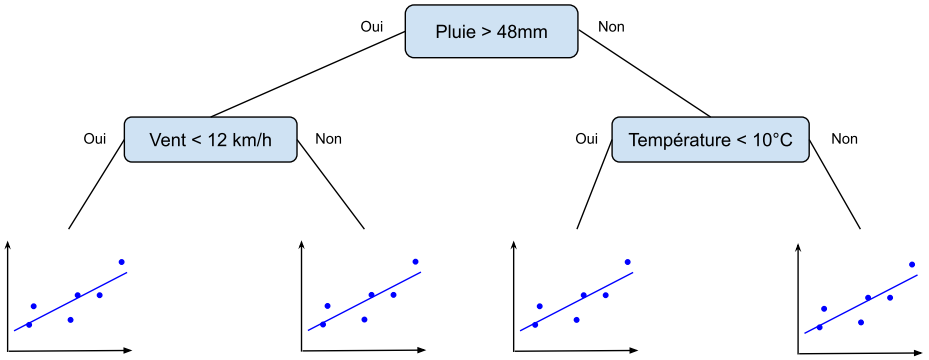


Figure 30 - Exemple d'une combinaison d'un arbre de décision et de régression linéaire basée sur des données météorologiques de précipitations, de températures et de vent

4.2. Présentation des sources de données à disposition

Pour entraîner les modèles statistiques, quatres catégories de données explicatives sont à notre disposition, deux possibilités de variables observées et une variable de pondération (Figure 31).

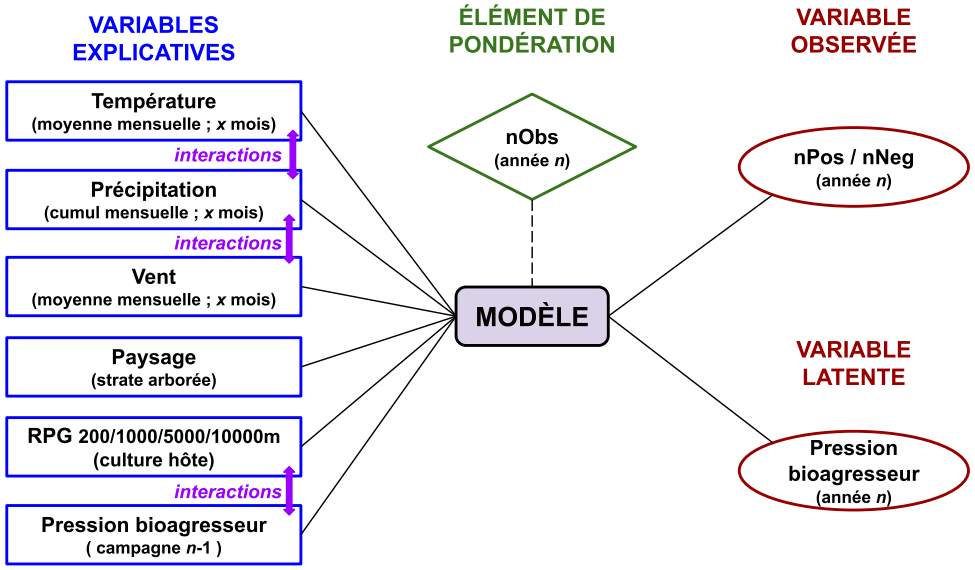


Figure 31 - Schéma explicatif de la construction des modèles statistiques et des variables impliquées

- Variables climatiques (précipitation, température, vent)

Les variables climatiques proviennent de la base de données SAFRAN. Les variables météos associées à chaque parcelle d'observation de la pression d'un bioagresseur sont issues de la maille SAFRAN la plus proche (distance < 4km).
Les données météorologiques avec lesquelles nous travaillons sont mensuelles (moyenne des températures et du vent, et cumul des précipitations) et sont disponibles pour plusieurs mois avant la dernière observation du bioagresseur sur la culture. Lors de l'évaluation des modèles, les conditions climatiques des 18 mois précédant la dernière observations des bioagresseurs dans la campagne sont utilisés. Cela permet d'avoir le climat au moment du pic des bioagresseurs lors de la campagne précédente.

- BD TOPO (couche "végétation" version 2017)

La BD TOPO est une carte vectorielle produite par l'Institut National de l'Information Géographique et Forestière (IGN). Avec sa précision métrique, elle fournit de l'information sur les composantes du paysage ; bois, haies, prairies, landes, garrigues,...
Les éléments du paysage utilisés sont la présence de bois et de haies car les autres éléments comportant une proportion trop importante de données manquantes.

- RPG (Registre Parcellaire Graphique)

Le RPG est un système d'information géographique (SIG) qui contient les données des parcelles et îlots culturaux basées sur les déclarations des exploitations pour obtenir les aides de la politique agricole commune (PAC). L'information fournie correspond à l'usage du sol et aux cultures présentes. Différentes tailles d'îlot sont possibles autour d'un point géographique et ils vont synthétiser la surface que représente chaque culture dans l'îlot.
Nous utilisons les données du RPG de l'année en cours et de l'année précédente à l'observation et des tailles d'îlots avec lesquelles nous travaillons sont de 200m, 1000m, 5000m et 10000m.

- Pression bioagresseurs (campagne n-1)

La pression des bioagresseurs est connue pour une multitude de parcelles de la campagne précédente avec les données issues des réseaux d'épidémiosurveillance. Pour avoir une estimation sur l'ensemble du territoire métropolitain, on utilise les prédictions d'un modèle d'interpolation bayésien avec comme *a priori* la moyenne nationale annuelle (et non pluri-annuelle comme en 2.2) de la pression des bioagresseurs. Les prédictions sont effectuées sur les mailles de la grille SAFRAN, puis, comme pour les données climatiques, les données de pression en bioagresseurs de l'année n-1 d'une parcelle d'observations seront issues de la maille la plus proche.

- Les intéractions entre variables

Les interactions entre les variables peuvent être prises en compte puisque, par exemple, la pression potentielle d'un bioagresseurs pourrait n'avoir un impact l'année suivante par la constitution d'un innoculum que proportionnellement à la présence de la culture hôte.

- Variables observées et latente

Une variable observée correspond à l'information que l'on souhaite voir prédire par le modèle. Deux variables sont envisagées : la variable binomiale observée constituée à la fois du nombre d'observations positives et du nombre d'observations négatives (*nPos* et *nNeg*) sous forme de matrice ou la variable latente de pression observée *P* (*nPos/nNeg*). Les modèles de régressions peuvent intégrer dans l'argument *family* le fait que ce soit du binomial sinon on utilise la loi gaussienne ou de poisson. Pour les modèles basés sur des arbres de décision (Random Forest, MARS et Cubist), il n'ont pas cette possibilité de leur fournir les données de *nPos* et *nNeg* dans une matrice en leur spécifiant que l'on travaille sur du binomial. On leur fournit donc la donnée sous la forme la plus basique : une seule variable qui pour chaque observation prendra la valeur 0 ou 1 (Figure 32). C'est une manière de les forcer à prédire une variable binomiale.

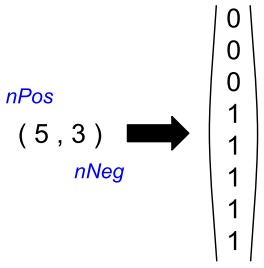


Figure 32 - Illustration de la transformation de la variable observée utilisée pour entraîner les modèles basés sur des arbres de décision afin de conserver le caractère binomial de la variable

Un inconvénient de cette transformation est cependant que l'on occulte la relation entre les différentes observations sur une même parcelle. Le parti est pris ici d'accepter que les différents modèles testés soient imparfaits d'un point de vue de la validité statistique de p-values (qui ne seront d'ailleurs pas utilisées) mais que la validation croisée permette de retenir les modèles les plus prédictifs.

La notation suivante sera utilisée pour spécifier un modèle en particulier :

- le nom du modèle seul correspondra au modèle dit "classique" ajusté sur la variable de la pression des bioagresseurs (ex : *Lasso*, *RandomForest*, *Cart*, *Mars*, *Cubist*, *GamLasso*)
- le nom du modèle plus "Binom" correspondra au modèle ajusté sur la variable binomiale *nPos* et *nNeg* (ex : *LassoBinom*, *MarsBinom*, *GamLassoBinom*)
- le nom du modèle plus "bin" correspondra au modèle ajusté sur la variable binaire 0 ou 1 (ex : *RandomForest_bin*, *Cart_bin*, *Cubist_bin*)

- Pondération lors de l'ajustement du modèle

Pour les modèles à l'exception du GAM-LASSO, il est possible d'utiliser un argument qui donne des poids différents aux différentes observations qu'on lui fournit. Cette information correspondrait à la fiabilité de chaque information que le modèle prend en considération. Pour nos données, il correspond au nombre d'observations (*nObs*) réalisées par parcelle. On donne cette information au modèle lorsqu'on utilise pas déjà un modèle binomial où l'on indique nPos et nNeg sinon l'information serait redondante.

4.3. Méthodologie de calibration et vérification des modèles statistiques

Chaque type de modèle statistique est calibré indépendamment pour chaque bioagresseur afin de produire un modèle ajusté pour chacun. Il est important de vérifier la validité des modèles en procédant à une étape de validation croisée. 3 hypothèses d'utilisation des modèles impliquant différents biais de sélection plus ou moins fort dans les données utilisés pour l'ajustement des modèles sont testées :

- Hypothèse 1 - disponibilité de données de pression la même année mais uniquement dans d'autres département et du même département mais uniquement d'autres années, biais temporel/climatique et spatial : le modèle permet de prédire des associations campagne de culture et départements non présents dans le jeu de calibration (Ex : le département 63 pour la campagne 2015)
- Hypothèse 2 - disponibilité uniquement de données de pression d'autres départements, biais spatial : le modèle permet de prédire des départements pour lesquels il n'y aucune donnée d'observation, toutes années confondues (Ex : le département 63)
- Hypothèse 3 - disponibilité uniquement de données de pression d'autres années, biais temporel/climatique : le modèle permet de prédire une nouvelle campagne de culture (Ex : campagne 2015)

Pour vérifier les hypothèses, on procède par étapes (Figure 33) :

1 - Ajustement du modèle sur l'ensemble des données et re-prédiction des données par le modèle.

2 - Analyse statistique des prédictions par rapport à la réalité observée (calcul du $R^2_{obs.w}$).

3.1 - Division du jeu de données en plusieurs groupes selon le test des hypothèses.

Hyp 1 - 5 groupes sélectionnant chacun 20% des associations campagne-département présentes dans les données.

Hyp 2 - 5 groupes sélectionnant chacun 20% des départements présents dans les données.

Hyp 3 - division en *n* campagnes de culture présent dans les données.

3.2 - Pour chaque groupe, on entraîne un modèle en excluant ce groupe. Avec les modèles créés, on prédit le groupe exclu du jeu de données d'ajustement. On obtient ainsi des prédictions pour l'ensemble des données sans jamais qu'une donnée ne soit utilisée dans la prédiction de sa propre valeur.

4 - Rassemblement de l'ensemble des prédictions par différents tests d'hypothèse pour réaliser l'analyse statistique entre les prédictions et la réalité observée (calcul des $R^2_{obs.w}$).

5 - Comparaison des résultats d'analyse entre le modèle complet et la validation croisée pour vérifier que les caractéristiques du modèle sont conservées.

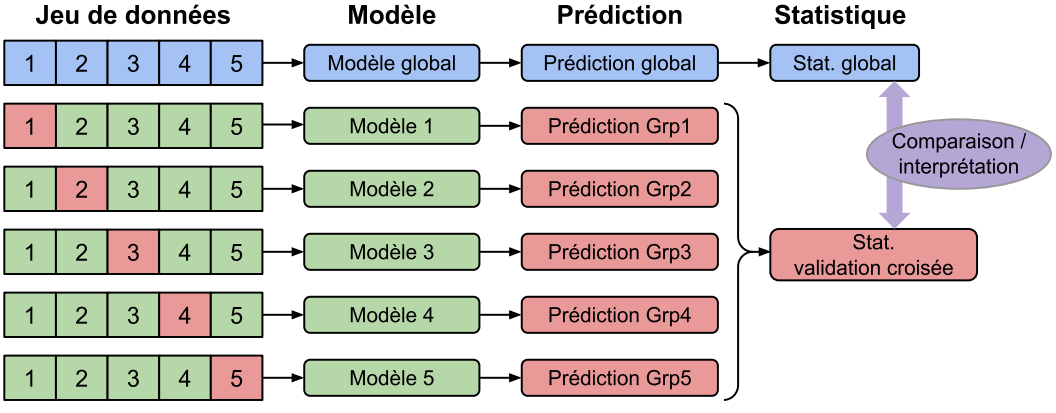


Figure 33 - Illustration de la méthode de validation des modèles statistiques par subdivision du jeu de données en sous-groupes. Un groupe peut être un ensemble de départements, de campagnes-départements ou une campagne.

4.4. Résultats

Une multitude de combinaisons de variables a été testée pour l'ensemble des modèles. Cette analyse a permis de tirer plusieurs informations sur les variables qui avaient un impact visible sur la qualité de la prédiction pour chacun des modèles, et sur les performances de chacun comparé aux autres. Les modèles GAM-LASSO et Cubist ont rapidement été abandonnés au vu de leurs mauvais résultats. Les résultats pour les autres modèles sont les suivants (Figure 35) :

- Quel que soit le modèle, la combinaison de 3 **variables** est nécessaire pour avoir un bon modèle.

  ➡ **Pression bioagresseur (campagne n-1)**, **Température** et **Précipitations**

  La dégradation de la qualité de prédiction est nettement visible sur la figure 34 sans ces éléments (météo d'une part et pression bioagresseurs d'autre part).

  L'utilisation de la variable vent, du paysage et du RPG de l'année précédente permet d'améliorer les modèles Random Forest et CART. L'ajout seul de l'une de ses variables n'a aucun effet d'amélioration par rapport aux variables déjà sélectionnées.

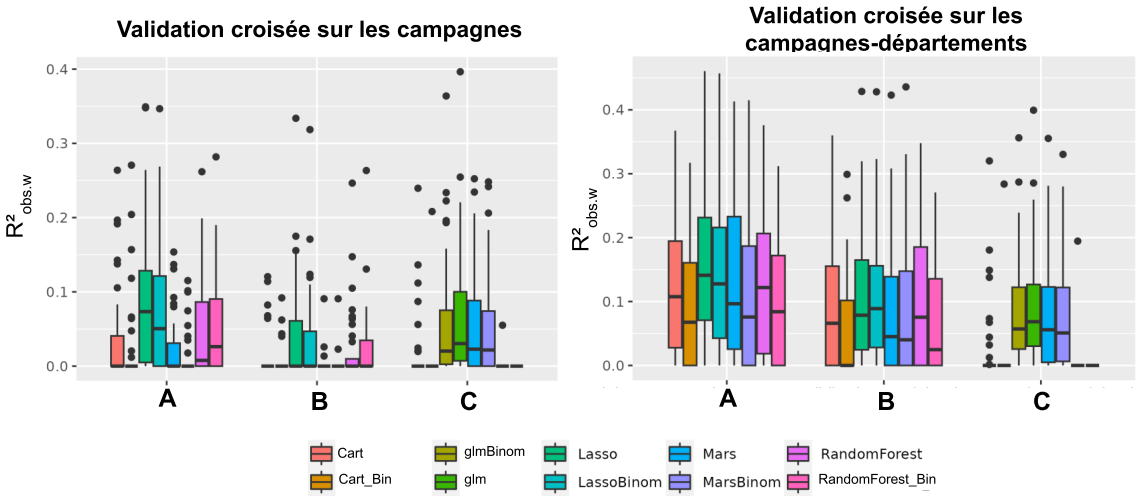


Figure 34 - Boxplots comparatifs des $R^2_{obs.w}$ des prédictions pour chaque modèle selon 2 types de validation croisée. Les modèles sont ajustés selon le nombre d'observations et selon 3 groupes de variables : (A) Précipitation - Température - Pression bioagresseur (année n-1) ; (B) Précipitation - Température ; (C) Pression bioagresseur (campagne n-1). Un modèle LASSO avec qu'une seul variable explicative revient à faire un modèle de régression linéaire généralisé (glm)

- Plusieurs modèles statistiques ont une tendance forte au **sur-apprentissage**. C'est-à-dire qu'il peuvent re-prédire les données de l'ajustement avec une bien meilleure précision que pour des nouvelles données. Cette différence est visible avec une distribution des R^2 nettement supérieurs pour le modèle global.

  ➡ **Random Forest** et **CART**

- Les modèles pour lesquels on ajuste sur la variable de la pression des bioagresseurs observée non binomiale, donnent de meilleurs résultats à l'exception de la Random Forest pour la validation croisée sur les campagnes.

  Quatre types de modèles permettent une prédiction semblable pour les **départements** ou les associations **campagnes-départements**.

➡ **Random Forest**, **CART, LASSO** et **MARS**

- La prédiction d'une nouvelle campagne est ce qu'il y a de plus difficile quel que soit le type de modèle statistique. Mais 2 types de modèles permettent une qualité de prédiction semblable pour une **nouvelle campagne**.

➡ **Random Forest** et **LASSO**

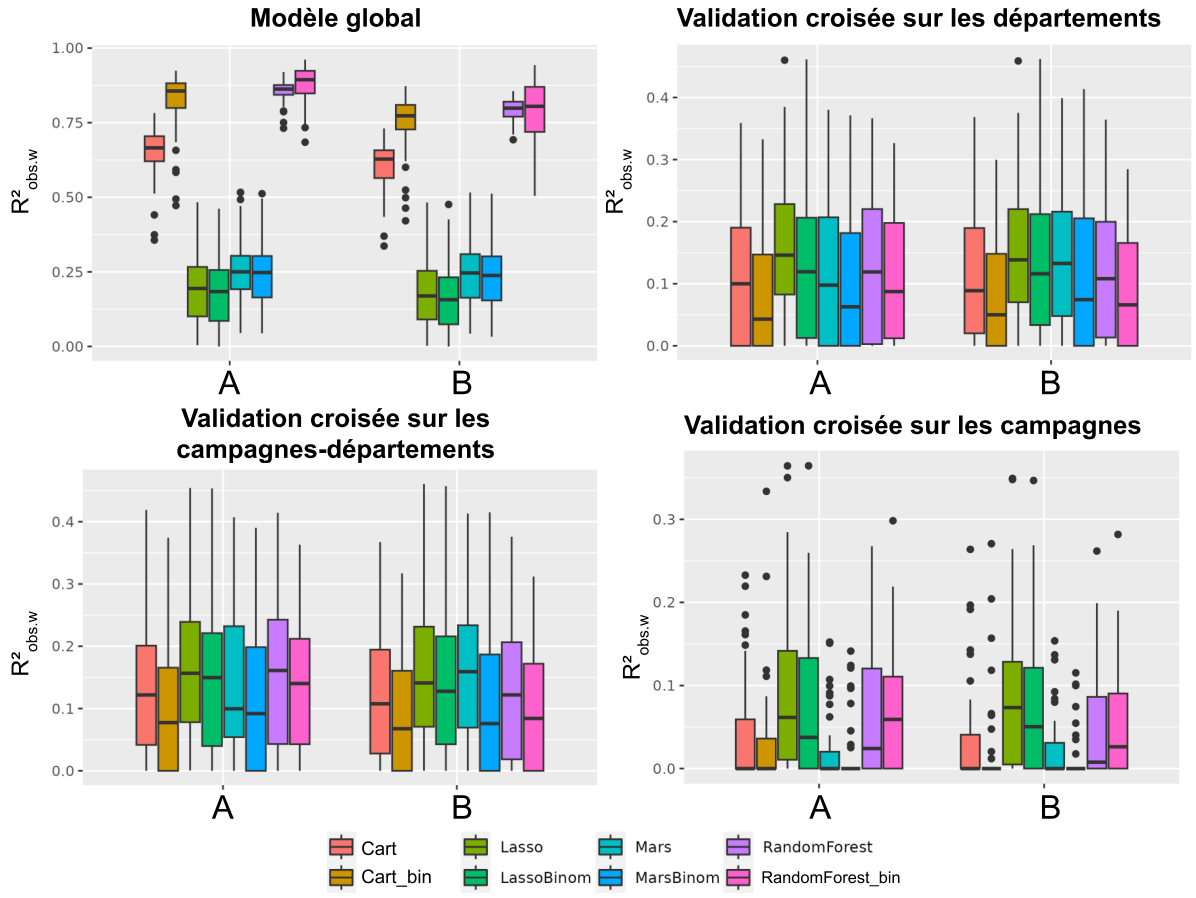


Figure 35 - Boxplots comparatifs des $R^2_{obs.w}$ des prédictions pour chaque modèle selon les types de validations croisées en lien avec le modèle global. Les modèles sont ajustés selon le nombre d'observations et selon 2 groupes de variables : (A) Paysage - Climat - Pression bioagresseur (campagne n-1) - RPG (année n-1) ; (B) Précipitation - Température - Pression bioagresseur (campagne n-1)

Trois modèles ressortent meilleurs que les autres, le modèle *Lasso* (gaussien), le *RandomForest* et le *RandomForest_bin* ajusté sur une variable binomiale. L'analyse du ratio $R^2_{obs.w}/R^2_{opt;w}$ montre que la variance de la qualité de prédiction du modèle Random Forest est bien plus faible que celle du *Lasso* (Figure 36).

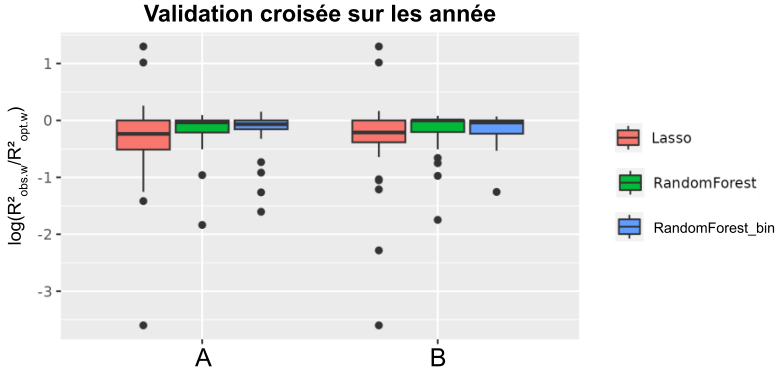


Figure 36 - Boxplots du log des ratios $R^2_{obs.w}/R^2_{opt.w}$ pour les modèles LASSO et Random Forest. Une valeur de 0 correspond à un ratio de 1. Les modèles sont ajustés selon le nombre d'observations et selon 2 groupes de variables : (A) Paysage - Climat - Pression bioagresseur (campagne n-1) - RPG (année n-1) ; (B) Précipitation - Température - Pression bioagresseur (campagne n-1)

La figure 37 permet de comparer les modèles entre eux pour identifier si certains bioagresseurs seraient mieux prédit par un modèle en particulier. Le LASSO est bien meilleur pour la prédiction que la Random Forest avec beaucoup plus de points au-dessous de la bissectrice, notamment pour la validation croisée sur la campagne, mais quelques bioagresseurs sont quand même mieux prédit par Random Forest.

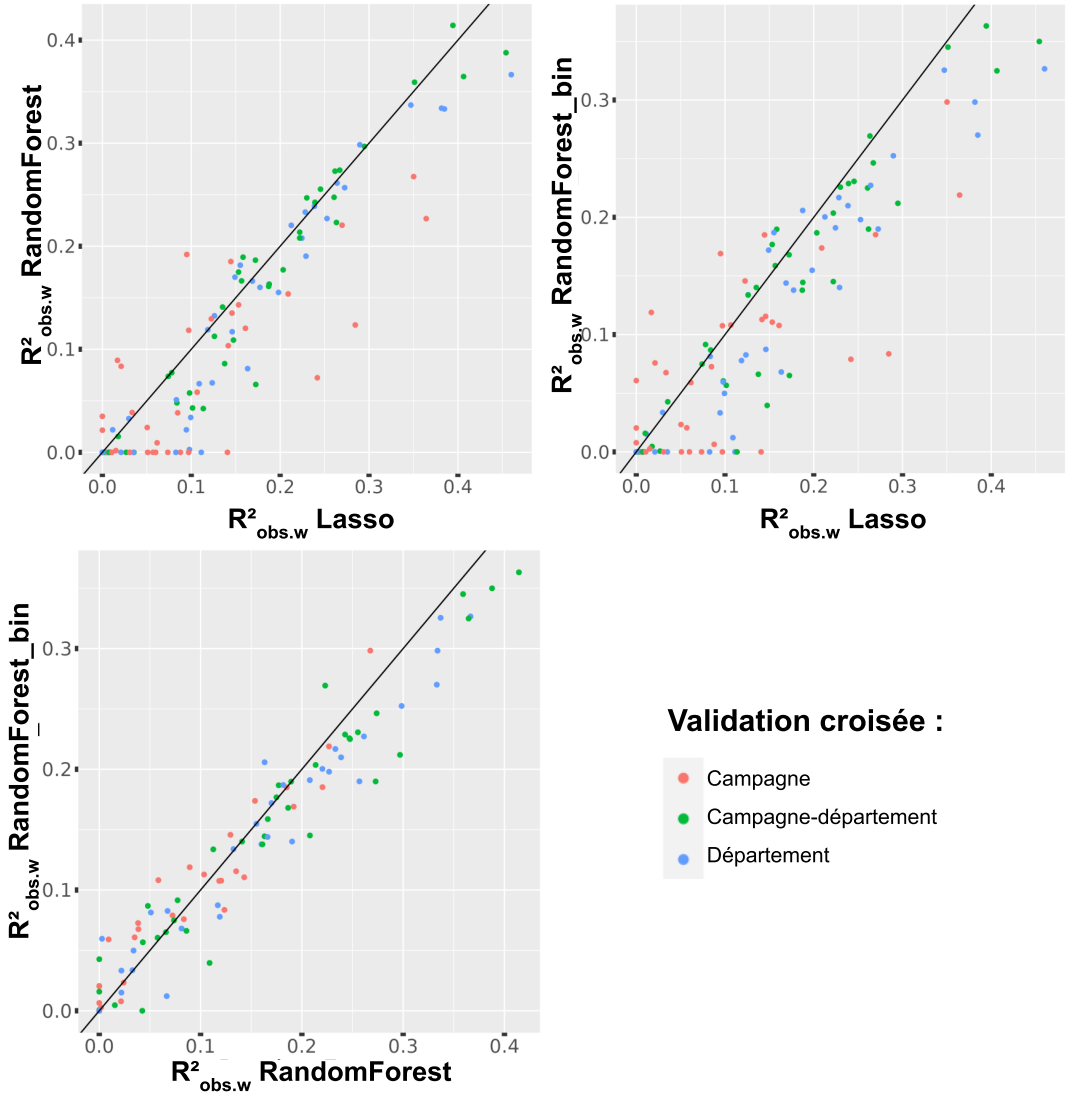


Figure 37 - Corrélations entre les $R^2_{obs.w}$ des prédictions des modèles Lasso et Random Forest par bioagresseurs (variables : Paysage - Climat - Pression bioagresseur (campgne n-1) - RPG (année n-1))

Ces informations permettent de réfléchir à une démarche qui pourrait être développée afin d'intégrer les modèles statistiques dans l'application MoCoRiBA et l'estimation de la pression d'un bioagresseur.

4.5. Démarche d'intégration des modèles statistiques à MoCoRiBA

Les modèles statistiques restent dans l'ensemble moins bons que le modèle d'interpolation (Figure 38) et ne peuvent donc pas le remplacer. La médiane des boxplots des modèles statistiques est inférieure au 1er quartile du modèle d'interpolation. Cependant, ils peuvent quand même être utilisés dans la construction de l'*a priori* local du modèle d'interpolation bayésien.

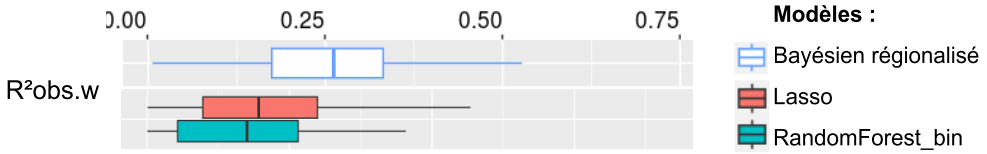


Figure 38 - Distribution des $R^2_{obs.w}$ pour les prédictions du modèle bayésien avec un a priori local, du modèle Lasso et du modèle RandomForest_bin. Les R^2 des modèles statistiques proviennent de la validation croisée sur l'association campagnes-départements

Les résultats sur les modèles statistiques avec le modèle LASSO, modèle le plus performant pour la plupart des bioagresseurs, et le Random Forest, prédisant mieux quelques bioagresseurs donnent la possibilité de faire du *Model Averaging*. C'est une méthode qui vise soit à choisir l'un des modèles pour chaque bioagresseur, soit à faire une moyenne pondérée des prédictions des modèles. Il faudrait appliquer cette méthode au modèle *Lasso* et au *RandomForest_bin* puisque c'est cette dernière qui présente le plus de bioagresseurs mieux prédit que le *Lasso* (par rapport au *RandomForest*).

Une recherche de la meilleure pondération applicable à la prédiction de chaque modèle statistique pourrait se faire avec une méthode similaire à la recherche de l'hyperparamètre *h* du modèle d'interpolation (partie 2.2). Pour une séquence possible de de coefficients de pondération des modèles, on calcule les R^2 de la moyenne pondérée et l'on sélectionne celle qui donne le R^2 le plus élevé (Figure 39).

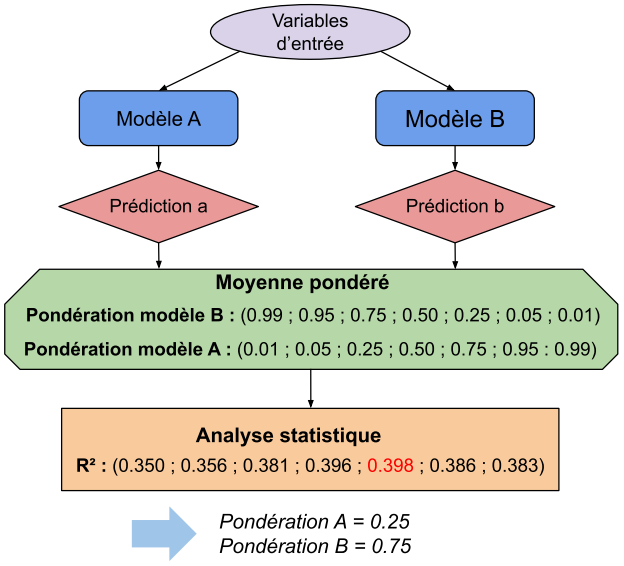


Figure 39 - Diagramme des étapes de détermination des coefficients de pondération pour du model averaging

La figure 40 présente la démarche d'intégration des modèles statistiques dans l'outil MoCoRiBA. Les 2 modèles statistiques sont produits et ajustés sur les données

d'épidémiosurveillance acquises depuis 2009 à partir de variables explicatives climatiques, paysagères, de RPG et de pression des bioagresseurs estimées de la campagne précédente. Cette dernière provient d'un modèle d'interpolation bayésien prenant en compte un *a priori* basé sur la moyenne annuelle nationale de la pression du bioagresseurs. Les modèles statistiques fournissent des prédictions locales pour la campagne de culture en cours avec la méthode de *model averaging*. Cette prédiction correspond à l'*a priori* local du modèle d'interpolation bayésien. La construction globale de ce modèle fait que la prédiction est dans un premier temps issue des données de la campagne en cours par interpolation, et dans un second temps des modèles statistiques.

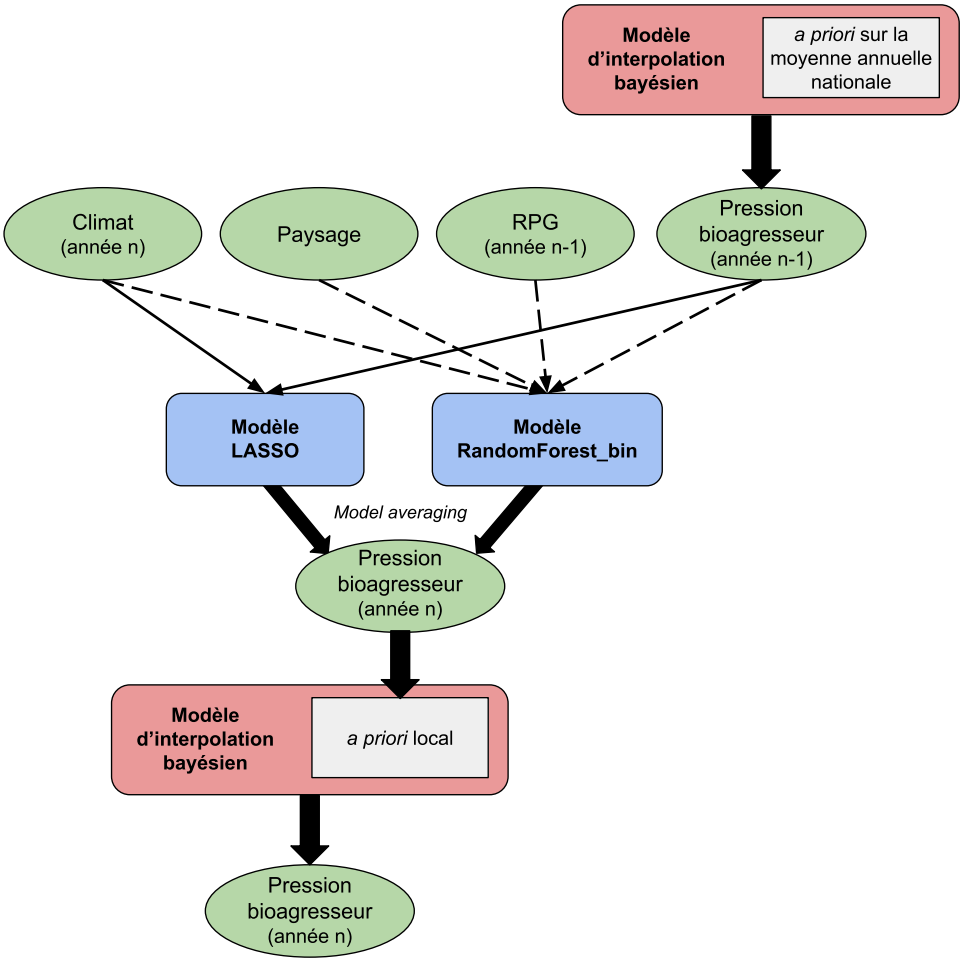


Figure 40 - Diagramme du modèle de prédiction de la pression des bioagresseurs pour le projet MoCoRiBA intégrant 2 modèles d'interpolation et du model averaging de 2 modèles statistiques

4.6. Discussion

Deux types de modèles statistiques pertinents sont ressortis de notre analyse : LASSO et Random Forest. Plusieurs méthodes de validations croisées ont été réalisées afin de tester la réponse de ces modèles en situation d'extrapolation (manque de données pour une nouvelle campagne et absence totale de données pour certains départements). L'étude démontre la difficulté de ces modèles à prédire une nouvelle campagne, bien plus que pour prédire la pression d'un département manquant une campagne donnée. Il serait intéressant de quantifier la proportion de données ou le nombre d'observations nécessaires lors de l'ajustement sur une nouvelle campagne, pour améliorer la prédiction au niveau de ce qui est obtenu en validation campagne-département. Il est possible qu'un nombre assez faible de

points d'observations permettent d'améliorer les prédictions au même niveau que la validation croisée sur les campagnes-départements.

Le parti pris au cours de ce stage a été de mettre en place des procédures de modélisation simples et facilement adaptables à des bioagresseurs aux types d'observations et d'infestations très variées. Cela nous a permis de modéliser très rapidement une quarantaine de bioagresseurs. Des adaptations de la procédure à certains bioagresseurs pourraient cependant bénéficier à la qualité de la modélisation. Par exemple, il n'est pas possible de faire de la validation croisée à l'échelle des régions pour les bioagresseurs de certaines cultures, telles que la betterave ou la pomme de terre, dont l'aire de culture ne couvre que deux ou trois régions. Pourtant, la validation croisée sur les départements présente plusieurs limites, surtout si l'on se place dans l'objectif de prédire une nouvelle région telle que la Bretagne. En effet, on ne se place pas à la même échelle géographique. Peut-être qu'une validation croisée à l'échelle des régions aurait été plus pertinente quant à l'évaluation de la robustesse du modèle dans les cas qui nous importent. Par ailleurs, actuellement les données d'observations d'un département sont utilisées par le modèle d'interpolation qui fournit la pression des bioagresseurs la campagne n-1. Comme c'est cette variable qui à généralement le plus de poids dans les modèles statistiques, la prédiction par le modèle statistique bénéficie grandement de ces données. Nous ne sommes donc pas vraiment dans le cas des régions Bretagne et Pays-de-la-Loire où les données ne sont présentes pour aucune campagne.

En se plaçant dans des conditions de cultures, de pratiques et de paysage similaires, on peut voir à quel point la pression des bioagresseurs dépend des conditions météorologiques. Cette capacité prédictive est cependant limitée, même quand on y ajoute des éléments de composition du paysage. Actuellement il n'est pas pris en compte que plusieurs cultures puissent servir d'hôte à un bioagresseur donné. Il serait possible de composer d'autres variables à partir du RPG pour y remédier. Par exemple, pour la septoriose, il serait raisonnable de considérer les parcelles d'orge et non seulement de blé. De même, pour les pucerons de la betterave, le colza peut aussi être une culture hôte.

Plus généralement, une meilleure prise en compte des fonctionnements biologiques pourrait permettre de combler le déficit de prédiction, à l'image de ce qui est obtenu en intégrant l'importance de l'inoculum par par la prise en compte de la pression de la campagne précédente. Dans ce sens, l'observation des populations d'auxiliaires pourrait améliorer les prédictions car les auxiliaires peuvent entraîner des cycles pluri-annuels faussant les liens directs entre inoculum une campagne donnée et pression la campagne suivante.

Une autre piste d'amélioration serait de faire de l'agrégation de métriques. Chaque métrique suit un protocole différent et peut apporter des informations complémentaires sur la présence des bioagresseurs, potentiellement à différents moments de leur cycle de vie. La prédiction pourrait être plus précise si on se basait sur l'ensemble des métriques pour un même bioagresseur au lieu de travailler sur les données d'une seule métrique.

Enfin, au lieu de n'utiliser pour chaque observation que le dépassement d'un seuil, il serait possible de décrire les observations de manière moins simplifiée, donc plus informative, augmentant ainsi la puissance statistique des modèles réalisés.

L'ensemble de ces pistes d'amélioration du modèle statistique induirait nécessairement la prise en compte des spécificités des bioagresseurs et des observations qui en sont faites. Une attention particulière devra être apportée à maintenir une certaine généricité dans la procédure de modélisation statistique. A cette condition, la performance des modèles pourra être améliorée tout en maintenant l'intégration comme *a priori* dans les

modèles d'interpolation. Une difficulté qui n'a pas été prise en compte à ce stade est la modulation du poids de ces modèles statistiques en tant qu'*a priori* pour l'interpolation en fonction de leur performance lors de la validation croisée.

Dans cette perspective d'intégration et d'évaluation des deux types de modèles, il serait bénéfique de réviser notre manière d'utiliser la pondération par le nombre d'observations des coefficients de détermination. Pour les prédictions finales du modèle il est nécessaire d'utiliser le $R^2_{obs.w}$ donnant plus de poids aux points mieux observés. Au contraire, pour les estimés qui fournissent l'*a priori* il faut particulièrement être bon là où il n'y a peu de points d'observations, donc le R^2_{obs} , non pondéré par le nombre d'observations pourrait être plus pertinent.

Un niveau supérieur d'amélioration des modèles serait l'utilisation de modèles hiérarchiques bayésiens. Ces modèles sont beaucoup plus coûteux tant en temps de développement qu'en temps de calcul. Cependant, ils permettent de prendre en compte les incertitudes dans les variables explicatives sous la forme de distribution de probabilités. Ainsi, ils refléteraient dans les pondérations comme dans l'incertitude de prédictions les incertitudes liées à l'éloignement des observations la campagne précédente où la campagne en cours.

Conclusion

Le projet MoCoRiBA a comme objectif de permettre aux agriculteurs de diminuer le recours aux produits phytosanitaires, notamment en facilitant une réflexion stratégique sur leur gestion des cultures. La construction de l'outil comme la modélisation des pertes de rendement nécessite des modèles de pression des bioagresseurs des grandes cultures et de rendement. Les objectifs du stage étaient d'évaluer et d'améliorer le modèle actuel de pression des ravageurs. Le travail réalisé sur les modèles de bioagresseurs a permis d'améliorer le modèle initial d'interpolation, basé sur la loi binomiale, en produisant un nouveau modèle résultant de l'assemblage de modèles d'interpolations, basés sur les statistiques bayésiennes, et de modèles statistiques. L'assez faible capacité prédictive des modèles statistiques dans leur ensemble fait que leurs prédictions sont utilisées que dans un second ordre, quand l'interpolation perd en fiabilité. Le codage du modèle global n'a cependant pas pu être réalisé donc ses résultats en termes de prédiction restent à évaluer. Des limites persistent dans ce nouveau modèle et nous avons proposé des étapes d'amélioration allant de la simple révision des procédures de validation jusqu'à l'utilisation de modèles hiérarchiques bayésiens.

L'amélioration du modèle de bioagresseurs va permettre de fournir une information plus fiable dans l'outil MoCoRiBA qui sera plus pertinente pour la réflexion des agriculteurs. Elle va aussi permettre de poursuivre les travaux sur l'évaluation des pertes de rendement en fonction des pratiques culturales qui est assez complexe et pour lesquelles il est nécessaire d'avoir des estimations de pression potentielle de bioagresseurs fiables.

Summary

My internship is part of the MoCoRiBA-GC project (Modelling and Communication of the Risk of BioAggressors in Field Crops) run by INRAE and its partners since 2020. The aim of this project is to avoid the unnecessary use of plant protection products by farmers and thus to reduce the use of fungicide and insecticide plant protection products in field crops, without reducing the economic margin for farmers. The initial aim was to produce statistical models to predict in real time the pressure of around forty pests and diseases on field crops in mainland France. Given the limited availability of real-time data, the project has shifted its focus to the production of a strategic tool. With this tool, farmers and advisors will be able to assess the relevance of their past plant protection treatment decisions by comparing their practices and results with those of farms in the DEPHY network (a network committed to reducing the use of plant protection products through new practices and cultivation techniques).

The work carried out during this internship is part of an effort begun at the start of the project to integrate interpolation and statistical models to estimate annual disease and pest pressures at all points in mainland France. Observations from plant epidemiological surveillance were used to produce these models.

This internship first proposed theoretical uncertainty estimates for interpolation. A first test was carried out on the interpolation models proposed in previous work. Secondly, we proposed a Bayesian interpolation model to improve the calculation of theoretical uncertainty, while also opening up the possibility of integrating statistical models as a useful a prior, particularly when observations are few or absent. Indeed, the project had already shown that statistical models based on meteorology are generally less efficient than interpolation models.

Finally, we have sought to improve the statistical models already produced, on the one hand by integrating new climatic, landscape and biological variables, and on the other by testing other statistical model formalisms. At the same time, Bayesian interpolation models of pest presence were integrated into the MoCoRiBA project's web application, adding comparisons of pest pressure to comparisons of practices and yields.

The work carried out during this internship means that we can now envisage the integration of interpolation models, with statistical models as a priori, into the strategic reflections of farmers and agricultural advisors on the use of phytosanitary products.

Résumé

Mon stage s'inscrit dans le projet MoCoRiBA-GC (Modélisation et Communication du Risque de BioAgresseurs en Grandes Cultures) porté par l'INRAE et ses partenaires depuis 2020. Ce projet vise à éviter l'utilisation inutile de produits phytosanitaires par les agriculteurs et ainsi de permettre une diminution de l'utilisation des produits phytosanitaires fongicides et insecticides en grandes cultures, sans diminuer la marge économique pour les agriculteurs. Il visait initialement à produire des modèles statistiques pour prédire en temps réel la pression d'une quarantaine de bioagresseurs en grandes cultures sur le territoire français métropolitain. Face à la faible disponibilité en temps réel des données, le projet s'est réorienté vers la production d'un outil à visée stratégique. Avec cet outil, agriculteurs et conseillers agricoles pourront juger de la pertinence de leurs décisions passées de traitement phytosanitaire en comparant leurs pratiques et résultats à ceux des fermes du réseau DEPHY (réseau qui s'engage dans la réduction de l'utilisation des produits phytosanitaires à travers de nouvelles pratiques et techniques culturales).

Le travail mené durant ce stage participe à une effort commencé dès le début du projet d'intégrer des modèles d'interpolation et des modèles statistiques pour estimer les pressions annuelles de maladies et de ravageurs en tout point du territoire français métropolitain. Les observations d'épidémiosurveillance végétale ont été utilisées pour réaliser ces modèles.

Ce stage a d'abord permis de proposer des estimations d'incertitude théorique de l'interpolation. Un premier essai a été réalisé sur les modèles d'interpolation proposés lors des travaux précédents. Ensuite nous avons proposé un modèle d'interpolation bayésien améliorant le calcul de l'incertitude théorique d'une part et ouvrant d'autre part la possibilité d'intégrer les modèles statistiques comme *a priori* utile notamment lorsque les observations sont peu nombreuses ou absentes. En effet, le projet avait déjà montré que les modèles statistiques basés sur la météorologie sont en général moins performants que les modèles d'interpolations.

Enfin, nous avons cherché à améliorer les modèles statistiques déjà réalisés, d'une part en intégrant de nouvelles variables, climatiques, paysagères et biologiques, et d'autre part en testant d'autres formalismes de modèles statistiques. En parallèle les modèles d'interpolation bayésiens de présence de bioagresseurs ont été intégrés dans l'application web du projet MoCoRiBA ajoutant les comparaisons de pressions de bioagresseurs aux comparaisons de pratiques et de rendement.

Le travail réalisé pendant ce stage permet ainsi d'envisager à court terme l'intégration des modèles d'interpolation avec comme *a priori* des modèles statistiques dans les réflexions stratégiques des agriculteurs et conseillers agricoles sur les pratiques d'utilisations de produits phytosanitaires.

Bibliographie

ArcGIS PRO (s.d.). *Fonctionnement du Krigeage*. URL : https://pro.arcgis.com/fr/pro-app/latest/tool-reference/3d-analyst/how-kriging-works.htm (Consulté me 14/11/2023)

Arvalis (2020, 30 mars). *Les BSV aident les agriculteurs à protéger leurs cultures.* URL : https://www.arvalis.fr/l-institut/nos-actualites/les-bsv-aident-les-agriculteurs-proteger-leurs-cultures (Consulté le 07/10/2023)

Arvalis (s.d.). *Septoriose Z. tritici, P. nodorum.* Arvalis-info. URL : http://www.fiches.arvalis-infos.fr/fiche_accident/fiches_accidents.php?mode=fa&type_cul=1&type_acc=4&id_acc=46 (Consulté le 07/10/2023

Barbu, C. M., Hong, A., Manne, J. M., Small, D. S., Quintanilla Calderón, J. E., Sethuraman, K., ... & Levy, M. Z. (2013). The effects of city streets on an urban disease vector. *PLoS computational biology*, *9*(1), e1002801.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Butault, J.P., Dedryver, C.A., Gary, C., Guichard, L., Jacquet, F., Meynard, J.M., Nicot, P., Reau, R., Sauphanor, B., Savini, I., Volay, T., 2010. Ecophyto R&D. *Quelles voies pour réduire l'usage des pesticides ?* Synthèse du rapport d'étude.

CISSE, A., 2022. Modélisation de la présence de bioagresseurs (insectes,maladies fongiques) des grandes cultures. Rapport de stage. SIGMA Clermont.

Cohen, A. L., Illan, J. G., Pfeiffer, V. W., Wohleb, C. H., & Crowder, D. W. (2022). Linking herbivore monitoring with interpolation to map regional risk of pest species. *Journal of Pest Science*, *95*(1), 315-325.

Delaune, T., Ouattara, M. S., Ballot, R., Sausse, C., Felix, I., Maupas, F., ... & Barbu, C. (2021). Landscape drivers of pests and pathogens abundance in arable crops. *Ecography*, *44*(10), 1429-1442.

Devaud, N. G., & Barbu, C. M. (2019). Quantification of bioagressors induced yield gap for grain crops in France. bioRxiv, 641563.

Dupuis, J. (2007). *Statistique bayésienne et algorithmes MCMC*. IMAT (Master 1). University of Mathematics of Toulouse, France.

EcophytoPIC (2020, mis à jour le 12 juil 2023). *DEPHY FERME : un réseau de fermes engagées dans la réduction des phytos*. URL : https://ecophytopic.fr/dephy/le-dispositif-dephy-ferme (Consulté le 07/10/2023)

Esquirol, L. (2012). Comment coupler observations et prédictions pour améliorer les prédictions d'épidémie de Septoriose sur le blé. *Rapport de Master. AgroCampusOuest*.

Eyre, D., Baker, R. H., Brunel, S., Dupin, M., Jarošík, V., Kriticos, D. J., ... & Worner, S. (2012). Rating and mapping the suitability of the climate for pest risk analysis. *EPPO Bulletin*, *42*(1), 48-55.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, *19*(1), 1-67.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Gouvernement (s.d.). *Registre parcellaire graphique (RPG)*. Portail de l'artificialisation. URL : https://artificialisation.developpement-durable.gouv.fr/bases-donnees/registre-parcellaire-graphique (consulté le 21/11/2023)

Ghosal, I. & Matthias, K. (2019). *The plsmselect package*.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481-5487.

INRAE (2021). *Histoire de l'INRA*. URL : https://www6.inrae.fr/comitedhistoire/Organisations-amies/Histoire-des-Instituts/INRA (Consulté le 28/10/2023)

INRAE (s.d.). *Nous connaître*. https://www.inrae.fr/nous-connaitre (Consulté le 28/10/2023)

Jones, V. P., Brunner, J. F., Grove, G. G., Petit, B., Tangren, G. V., & Jones, W. E. (2010). A web-based decision support system to enhance IPM programs in Washington tree fruit. *Pest Management Science: formerly Pesticide Science*, *66*(6), 587-595.

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters*, *12*(4), 334-350.

Koenig, W. D. (2002). Global patterns of environmental synchrony and the Moran effect. *Ecography*, *25*(3), 283-288.

Kuhn, M., Weston, S., Keefer, C., & Kuhn, M. M. (2023). Package 'Cubist'. *Rule-and Instance-Based Regression Modeling. R Package Version 0.4*, *1*.

Lechenet, M., Dessaint, F., Py, G., Makowski, D., & Munier-Jolain, N. (2017). Reducing pesticide use while preserving crop productivity and profitability on arable farms. *Nature plants*, *3*(3), 1-6.

Lieven, J. (2016). Protocole de suivi d'une parcelle fixe de colza d'un réseau d'épidemio-surveillance. Version 1.0. Vigicultures®.

Li, J., & Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, *6*(3-4), 228-241.

Magarey, R. D., Travis, J. W., Russo, J. M., Seem, R. C., & Magarey, P. A. (2002). Decision support systems: quenching the thirst. *Plant Disease*, *86*(1), 4-14.

Magarey, R. D., & Isard, S. A. (2017). A troubleshooting guide for mechanistic plant pest forecast models. Journal of Integrated Pest Management, 8(1), 3.

Ministère de l'Agriculture et de la Souveraineté Alimentaire (2011, 3 juin). *Épidémiosurveillance : le système d'information Epiphyt.* URL : https://agriculture.gouv.fr/epidemiosurveillance-le-systeme-dinformation-epiphyt (Consulté le 07/12/2023)

Milborrow, M. S. (2011). Package 'earth'. *R Software package*.

Phillips, S. J. (2008). Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al.(2007). *Ecography*, *31*(2), 272-278.

Quantmetry (2023). *La Robustesse, un impératif pour une intelligence artificielle de confiance.* URL : https://www.quantmetry.com/blog/ia-confiance-robustesse/ (Consulté le 16/11/2023)

Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343-348).

RColorBrewer, S., & Liaw, M. A. (2018). Package 'randomforest'. *University of California, Berkeley: Berkeley, CA, USA*.

Robert, C. (2006). *Le choix bayésien: Principes et pratique*. Springer Science & Business Media.

Simonneau, D. (2015). Mode opératoires observations Blés. Version n°13a. Vigicultures®.

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints*, *39*(2), 17-26.

Terre Inovia (2023, 28 septembre).Gestion en cours de campagne des grosses altises adultes (altises d'hiver). URL : https://www.terresinovia.fr/-/surveillance-et-lutte-contre-la-grosse-altise (Consulté le 07/12/2023)

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine*, *16*(4), 385-395.

Venette, R. C., Kriticos, D. J., Magarey, R. D., Koch, F. H., Baker, R. H., Worner, S. P., ... & Pedlar, J. (2010). Pest risk maps for invasive alien species: a roadmap for improvement. *BioScience*, *60*(5), 349-362.

Wikipédia (s.d.). *Loi binomiale.* URL : https://fr.wikipedia.org/wiki/Loi_binomiale (Consulté le 08/10/2023)

Wikipédia (s.d.). *Coefficient of determination*. URL : https://en.wikipedia.org/wiki/Coefficient_of_determination (Consulté le 15/10/2023)

Annexes

Annexe 1 - Présentation de la base de données d'épidémio-surveillance

Notre jeux de données est formé de l'ensemble des informations issues des bases de données d'épidémio surveillance de 2009 à 2022.

Il comprend notamment les informations suivantes :

- **ID_plot** : Numéro d'identification de la parcelle ;
- **Culture** : Type de culture de la parcelle ;
- **Code_obs_gen** : ID du type d'observation (métrique) ;
- **Campagne** : Année de récolte ;
- **database** : le nom de la base de données (Vigiculture ou Epiphyt) d'où provient l'observation ;
- **Obs_date** : Date d'observation ;
- **Stade_phenologique** : Stade phénologique de la culture au moment de l'observation ;
- **Valeur** : Valeur de l'observation (qui peut être un comptage ou un pourcentage) ;
- **Latitude** : Latitude de la parcelle ;
- **Longitude** : Longitude de la parcelle ;
- **Code_insee** : Code élaboré par l'INSEE, à 5 chiffres. Concaténation du code département (2 chiffres) et du code commune (3 chiffres) ;
- ……

Le tableau suivant illustre notre jeux de donnée pour quelques observation :

ID_plot	Culture	Code_obs_gen	Campagne	database	Obs_date	Stade_phenologique	Valeur
77810	Colza d'hiver	GANbPE	2012	Vigiculture	03/10/2011	B6 / 80%	2
160036	Maïs	SES_PHE_NB_ADULTES	2013	Vigiculture	21/05/2013	4F : 4 Feuil	0
281812	Colza d'hiver	LGA%P	2015	Vigiculture	03/03/2015	C1 / 100%	65
119877	Tournesol	ManLimB	2012	Vigiculture	05/06/2012	B3-B4 / 60%	2
502151	Blé tendre d'hiver	OIDF3	2017	Vigiculture	24/04/2017	Z33 : 3 N	0
83288	Blé tendre d'hiver	SEPF3	2012	Vigiculture	14/05/2012	Z57 : 3/4 epiaison	10
200772	Colza d'hiver	MelNbP	2014	Vigiculture	24/03/2014	F1 / 20%	0
391527	Colza d'hiver	LGA%B	2016	Vigiculture	14/03/2016	E / 15%	30
67256	Tournesol	PucVCri	2011	Vigiculture	09/05/2011	B5 / 40%	6
394348	Colza d'hiver	LGA%P	2016	Vigiculture	02/11/2015	B7 / 60%	0

…with 1 587 411 more rows, and 25 more variables

Annexe 2 - Transformation des données pour se mettre dans le cadre binomial

Pour une observation, sur une parcelle, on obtient une variable qui prend la valeur 1 si le seuil (médiane des observations) est dépassé et 0 sinon. Cette variable suit une loi de Bernoulli de paramètre *p* qui correspond à la probabilité d'avoir un succès. Sur une parcelle, l'expérience (= les observations) est répétée plusieurs fois afin d'obtenir X_1, X_2, … , X_n qui sont des variables aléatoires de Bernoulli de paramètre *p.* Leur somme N est une variable aléatoire, qui suit la loi binomiale (Wikipédia, s.d.) :

$$N = \sum_{k=1}^{n} X_k \sim B(n,p)$$

avec :
n le nombre d'observations
p la probabilité d'un succès

Et l'espérance : $E(X) = np$

On obtient des données à la parcelle du nombre d'observation total (*nObs*) et du nombre de positif total (*nPos*) sous la forme suivante :

ID_plot	Culture	Code_obs_gen	Campagne	nPos	nObs
203657	Colza d'hiver	MelNbPB	2014	1	10
391765	Colza d'hiver	MelNbP	2016	6	7
286588	Blé tendre d'hiver	SEPF3	2015	3	14
160036	Maïs	SES_PHE_NB_ADULTES	2013	0	9
391765	Colza d'hiver	ChTNbV	2016	4	9
394348	Colza d'hiver	LGA%P	2016	0	4

…with 289 392 more rows, and 9 more variables

Annexe 3 - Exemple des étapes de calcul d'une interpolation avec le modèle construit sur la base d'une loi binomiale

Pour bien comprendre les calculs réalisés par le modèle d'interpolation basé sur la loi binomial, prenons l'exemple de la maille SAFRAN *906,* située au sud des Hauts-de-France (*49.3297,3.05194*). Nous souhaitons connaître l'interpolation de la pression du bioagresseur de la grosse altise d'hiver du colza, mesuré selon le protocole de la métrique LGA%B (correspondant au pourcentage de plantes avec le cœur détruit ou un port buissonnant) pour la campagne 2016.

- **Étape 1** : Calcul de l'hyperparamètre *h* pour le bioagresseur selon la méthode développée par Cisse A. et développée dans la partie 2.2 sur la base du modèle d'interpolation (2.2.b). Pour plusieurs valeurs de *h* possibles (entre 1 km et 250 km), selection de celui permettant d'obtenir la meilleur interpolation, défini par le R^2 évaluant la relation entre la prédiction et la réalité, au niveau des parcelles d'observations (quel que soit l'année de notre jeu de données).

  Dans une boucle testant une à une des valeurs de *h,* par année d'observation *n* de la métrique. Pour toutes les parcelles *p* ∈ (1,...*p'*...,*m*):

  - **Étape 1.1** : calcul des distances qui sépare la parcelle *p'* de toutes les parcelles d'observations de l'année *n*.

Le tableau suivant est un extrait de la table de données des distances d'éloignement (en m) avec la parcelle 434445 en 2016 pour les observations de la métrique *LGA%B* de la grosse altise du colza :

ID_plot	distance
434445	0
392295	37968.40
395029	21751.41
388796	34544.05
391527	19116.60

… with 183 more rows

- **Étape 1.2** : Retrait de la parcelle *p'* des données
- **Étape 1.3** : calcul des *NPosEff* et *NObsEff* pour la parcelle *p'* définie par les équations (2.2.d) et (2.2.e) rappelées ci-dessous. C'est-à-dire la somme des résultats de(s) observation(s) réalisée(s) sur chaque parcelle, pondérée par le poids accordé à celle-ci selon sa distance d'éloignement et de l'hyperparamètre *h.*

$$NPosEff_{p',bio\_i} = \sum_{p=1}^{m} nPos_{p,bio\_i}\, e^{\frac{-d_{pp'}}{h}}$$

$$NObsEff_{p',bio\_i} = \sum_{p=1}^{m} nObs_{p,bio\_i}\, e^{\frac{-d_{pp'}}{h}}$$

Où :
dpp' : distance qui sépare la parcelle *p'* et la parcelle *p*
h : hyperparamètre

- **Étape 1.4** : Calcul de la pression prédite sur la base de l'équation (2.2.c) rappelées ci-dessous :

$$P_{p'}^{prédite}(bio\_i) = \frac{NPosEff_{p',bio\_i}}{NObsEf_{p',bio\_i}}$$

Le tableau suivant illustre les données de *NPosEff*, *NObsEff* et de la prédiction réalisées sur les parcelles de l'année 2016 des données de la métrique LGA%B de la grosse altise du colza pour une valeur de l'hyperparamètre *h* de 22,6 km.

ID_plot	nPos	nObs	ratPos_real	NPosEff	NObsEff	ratPos_predicted
434445	7	7	1.00	2.148142	4.623895	0.4645742
392295	1	1	1.00	2.886282	5.808522	0.4969047
395029	0	2	0.00	3.860547	4.863057	0.7938518
388796	0	1	0.00	4.311200	6.760120	0.6377402
391527	3	4	0.75	4.277606	6.625685	0.6456096

… with 183 more rows

- **Étape 1.5** : Etude de la relation entre la prédiction et la réalité observée sur chaque parcelle au travers du calcul du R^2

On répète les étapes 1.1 à 1.5 pour les différentes valeurs de *h* possibles puis on sélectionne celui donnant le meilleur R^2, et donc la meilleure interpolation, pour la suite des calculs de l'interpolation au niveau de la maille SAFRAN *906*.

Le tableau suivant illustre la valeur de l'hyperparamètre *h* (en m) associé à son R^2, identifié comme celui permettant de réaliser la meilleure interpolation pour la métrique de la grosse altise d'hiver du colza LGA%B.

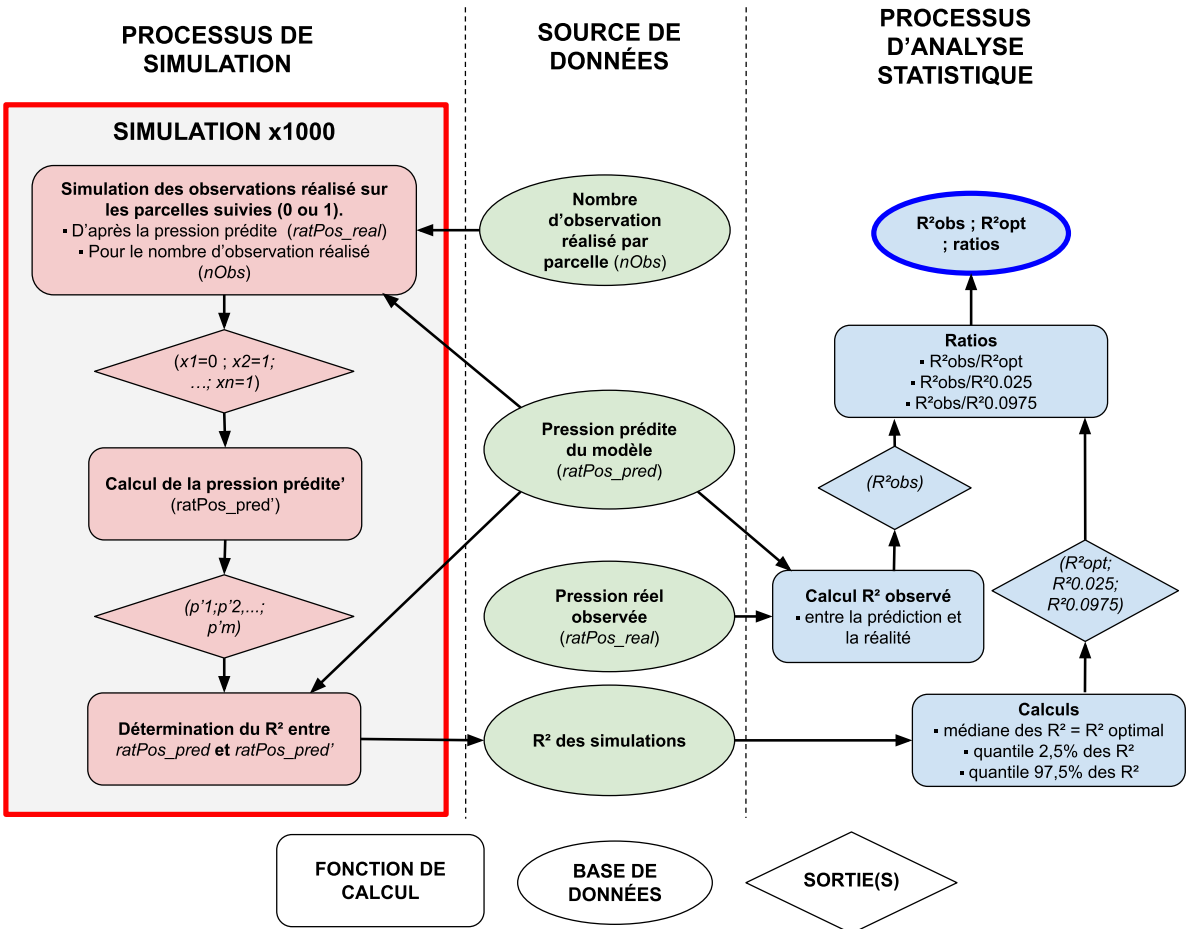
Culture	Code_obs_gen	h	R2
Colza d'hiver	LGA%B	22627.42	0.1717045

- **Étape 2** : Mesure la distance qui sépare la maille SAFRAN *906* de chacune des parcelles présentes l'année 2016 ;
- **Étape 3** : Calcul des *NPosEff* et *NObsEff* comme lors de l'étape 1.4 mais par rapport à la maille SAFRAN *906*
- **Étape 4** : Calcul de la pression du bioagresseur à la maille SAFRAN *906* par le ratio *NPosEff* sur *NObsEff*. Le tableau suivant donne les résultats de l'interpolation de la pression du bioagresseurs Altise Grosse d'hiver du Colza à la maille SAFRAN 906

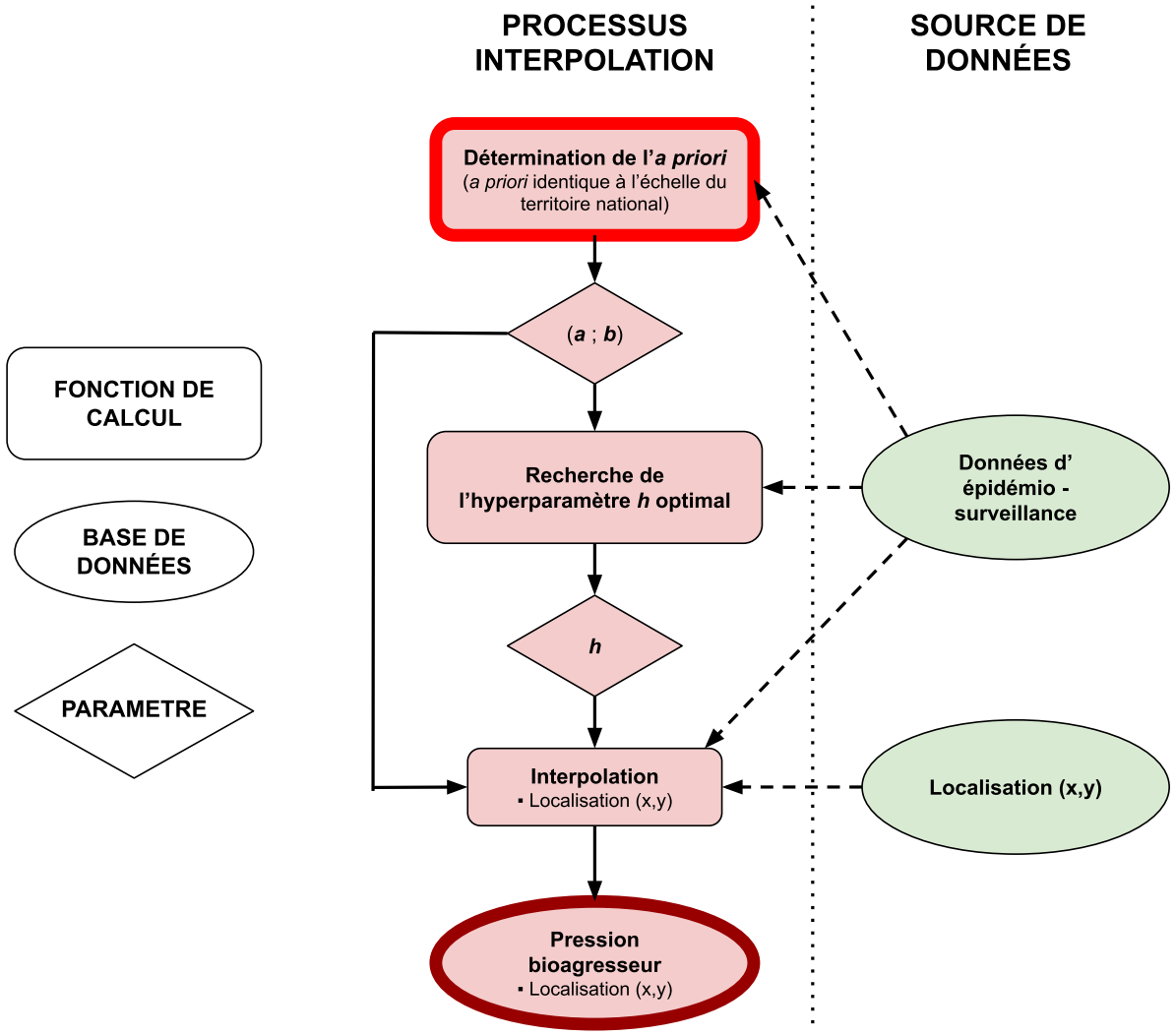
maille_safran	Culture	Code_obs_gen	Campagne	NPosEff	NObsEff	moy_risq
906	Colza d'hiver	LGA%B	2016	1,32465326	2,69324153	0,49184347

On obtient une estimation de la pression du bioagresseurs de la grosse altise du colza à la maille SAFRAN *906* qui est de 49%.

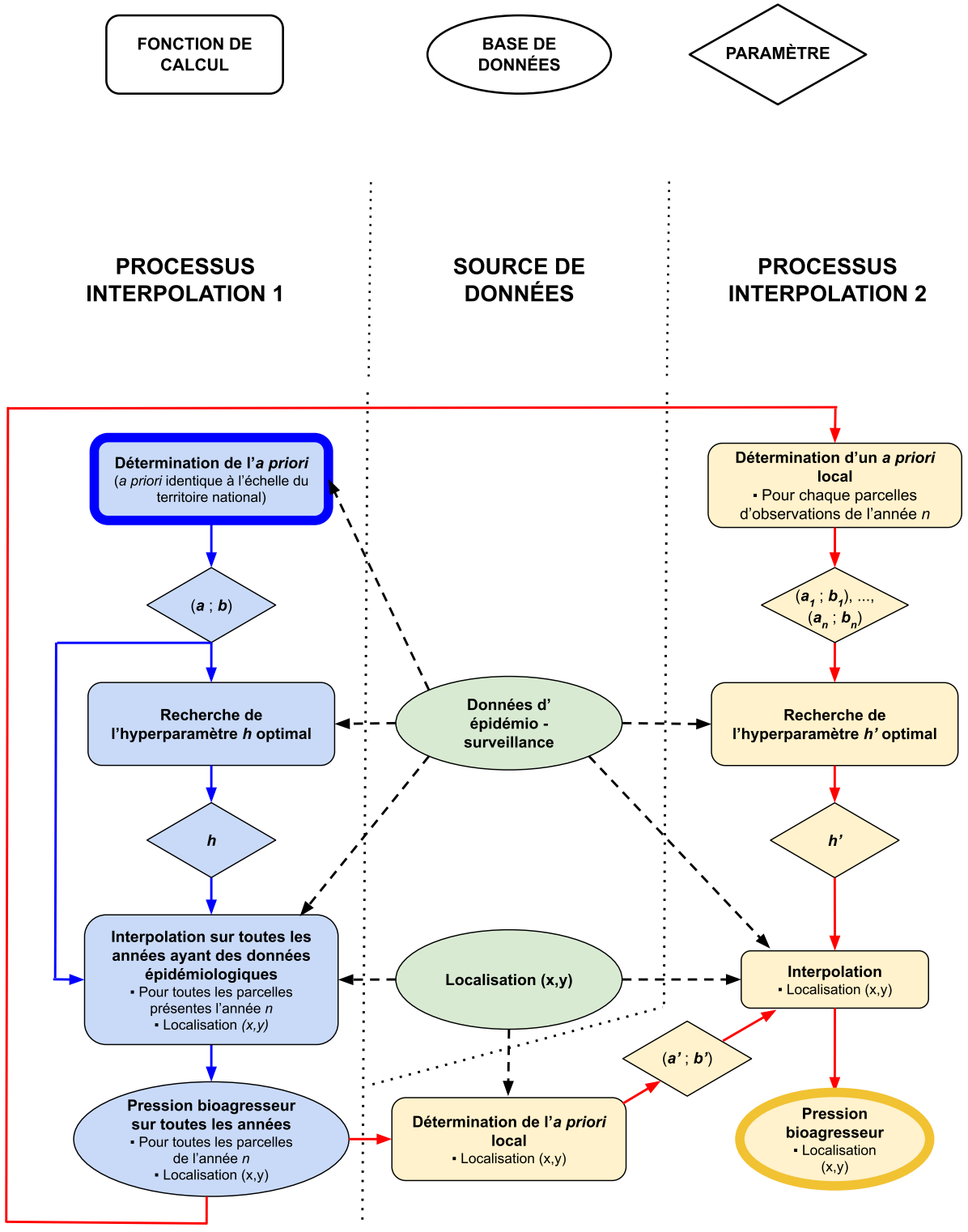
Annexe 4 - Diagramme du déroulement d'une estimation du R² optimal et des ratios



Annexe 5 - Diagramme des étapes d'une interpolation bayésienne avec un a priori unique au niveau national

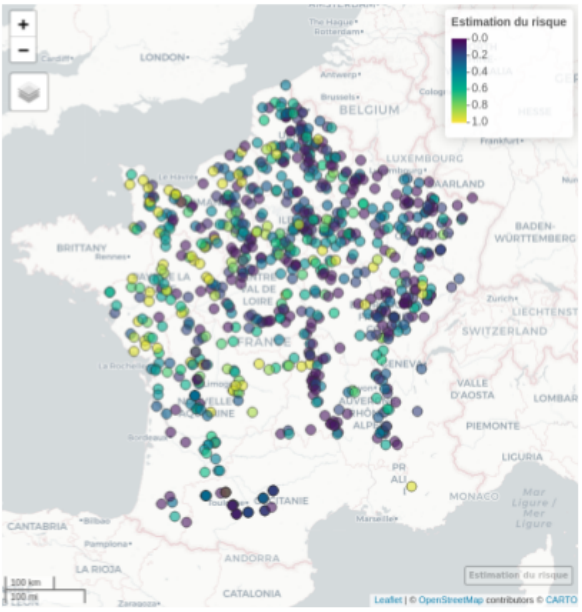


Annexe 6 - Diagramme des étapes d'une interpolation bayésienne avec un a priori local

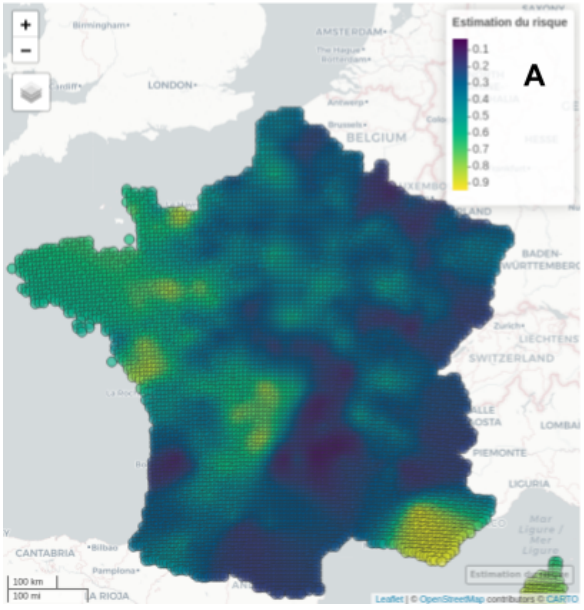


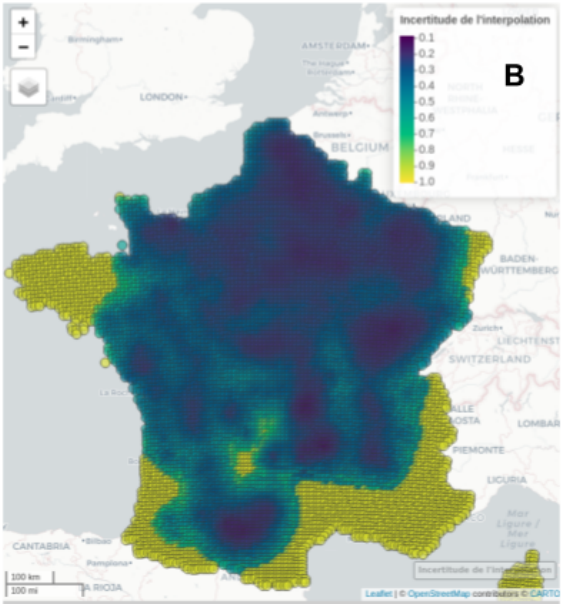
Annexe 7 - Cartes d'interpolation et d'incertitude théorique des modèles pour la septoriose du blé (SEPF3) pour l'année 2017

Parcelles suivies en 2017

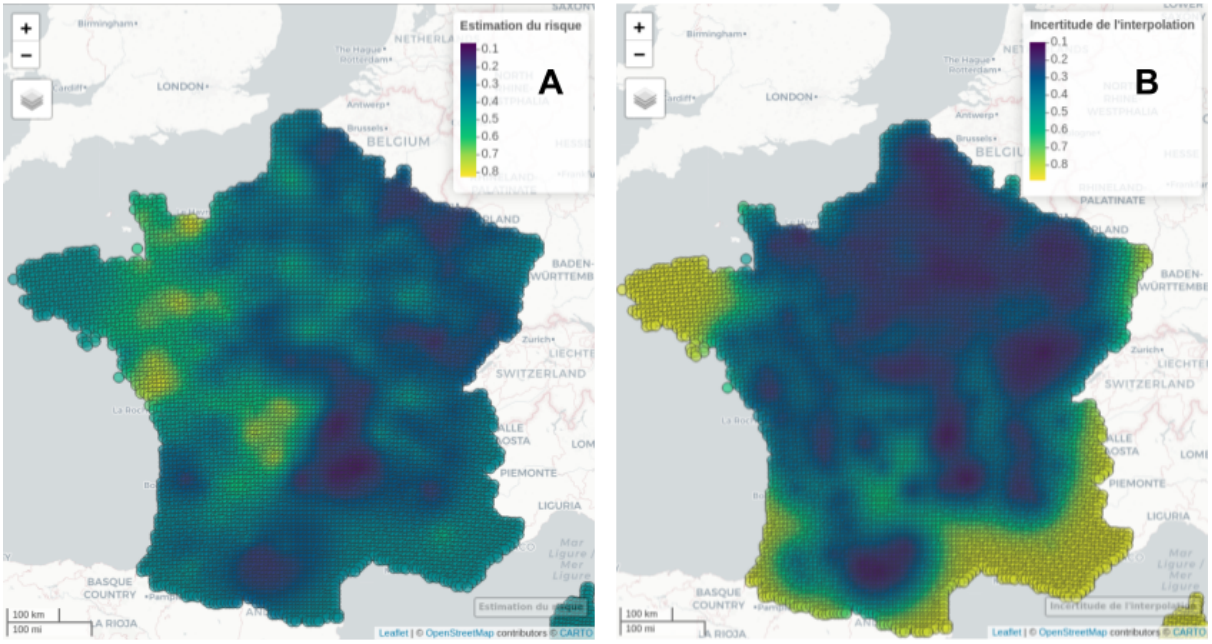


Interpolation de la pression avec le modèle binomial (A) et des incertitudes théoriques de l'interpolation (B)

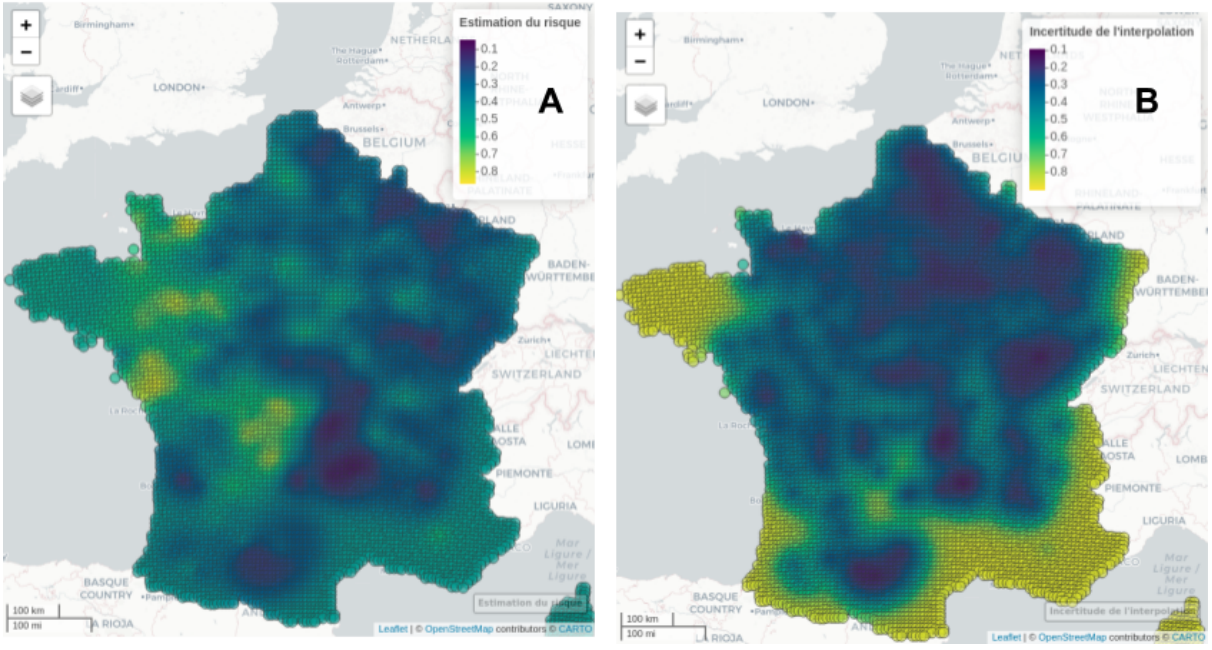




Interpolation de la pression du bioagresseur avec le modèle bayésien ayant un *a priori* national (A) et des incertitudes théoriques de l'interpolation (B)

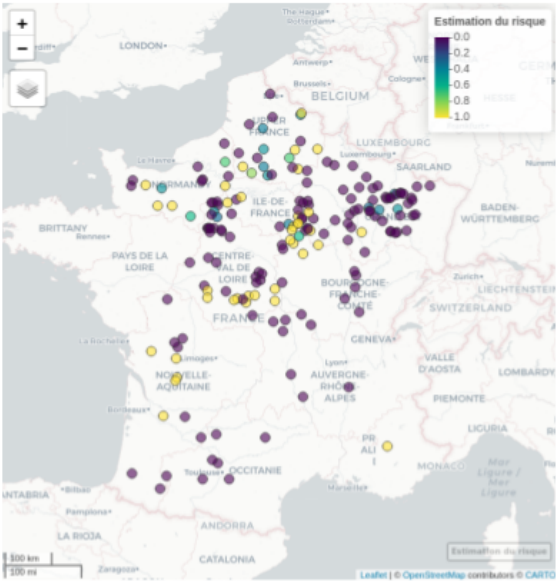


Interpolation de la pression du bioagresseur avec le modèle bayésien ayant un *a priori* local (A) et des incertitudes théoriques de l'interpolation (B)

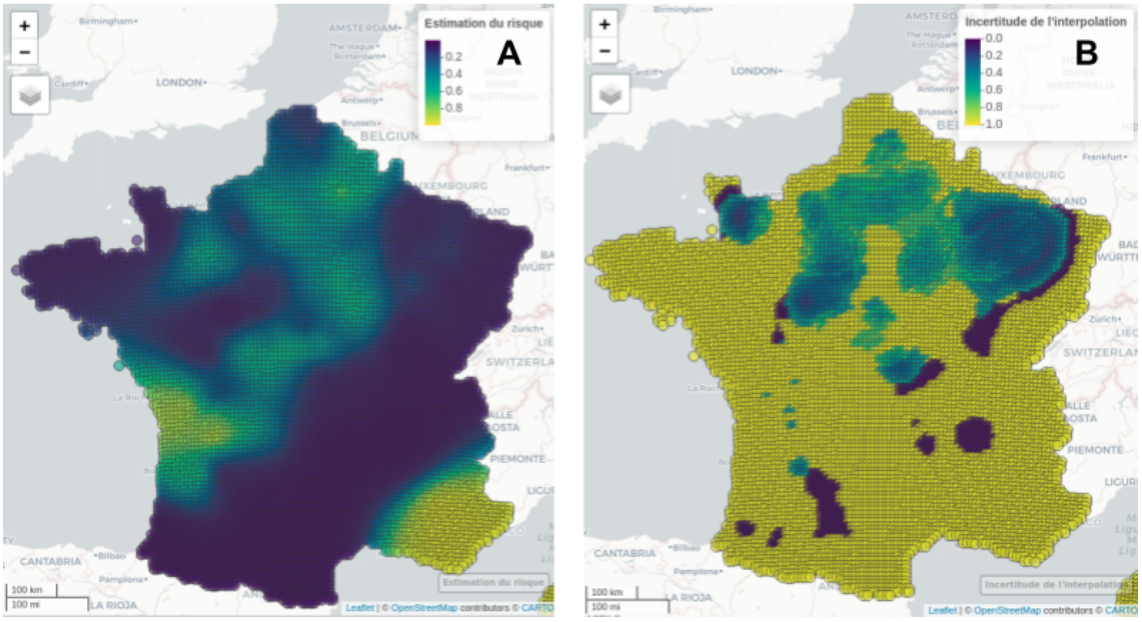


Annexe 8 - Cartes d'interpolation et d'incertitude théorique des modèles pour la grosse altise d'hiver du colza (LGA%B) pour l'année 2016

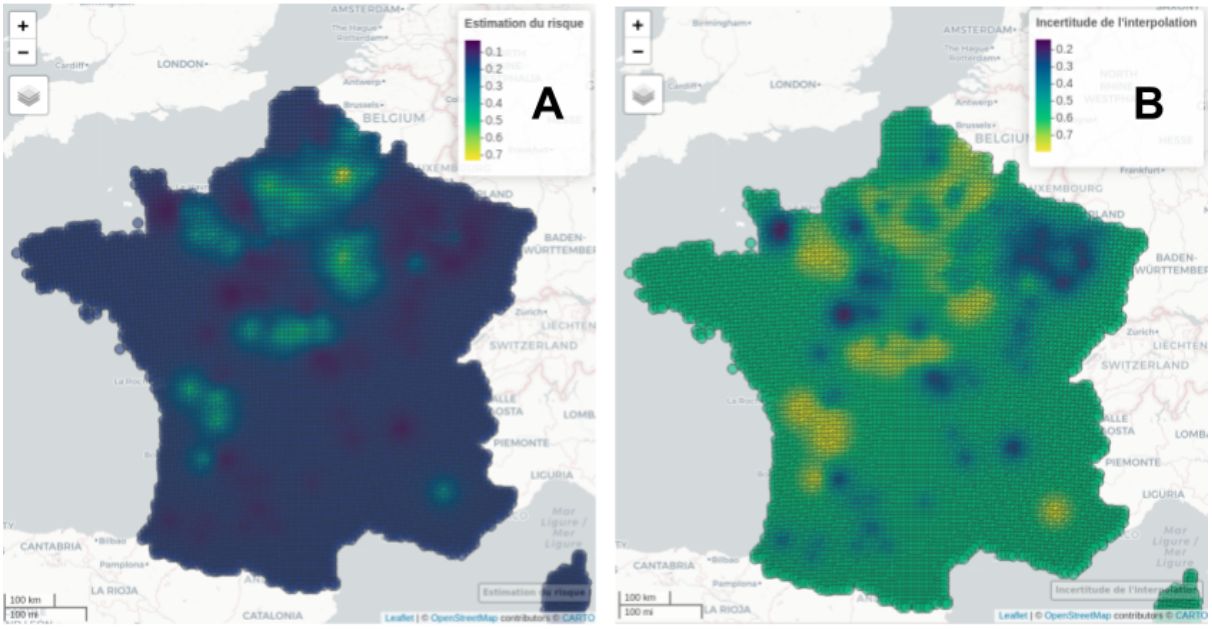
Parcelles suivies en 2016



Interpolation de la pression avec le modèle binomial (A) et des incertitudes théoriques de l'interpolation (B)



Interpolation de la pression du bioagresseur avec le modèle bayésien ayant un *a priori* national (A) et des incertitudes théoriques de l'interpolation (B)



Interpolation de la pression du bioagresseur avec le modèle bayésien ayant un *a priori* local (A) et des incertitudes théoriques de l'interpolation (B)