Modéliser la performance de cultures associées céréale-légumineuse annuelles : une approche combinant écologie des communautés et science des données

Rémi Mahmoud

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistiques

École doctorale I2S – Information, Structures, Systèmes

UMR Mathématiques, Informatique et STatistique pour l'Environnement et l'Agronomie

Modéliser la performance de cultures associées céréale-légumineuse annuelles : une approche combinant écologie des communautés et science des données

Présentée par Rémi MAHMOUD

Le 4 avril 2023

Sous la direction de Nadine HILGERT, Pierre CASADEBAIG et Noémie GAUDIO

Devant le jury composé de

Muriel VALANTIN-MORISON, Directrice de Recherche INRAE, Unité Agronomie	Rapporteure
Nathalie VIALANEIX, Directrice de Recherche INRAE, Unité MIAT	Rapporteuse
Éric GARNIER, Directeur de Recherche CNRS, UMR CEFE	Examinateur
Xavier GENDRE, Chercheur associé, Institut de Mathématiques de Toulouse	Examinateur
Catherine TROTTIER, Maître de conférences, Université Paul Valéry Montpellier 3, UMR IMAG	Examinatrice
Nadine HILGERT, Directrice de Recherche INRAE, UMR MISTEA	Directrice
Pierre CASADEBAIG, chercheur INRAE, UMR AGIR	Encadrant
Noémie GAUDIO, chercheure INRAE, UMR AGIR	Encadrante

UNIVERSITÉ DE MONTPELLIER

Remerciements

L'écriture des remerciements est pour moi une étape ambivalente. Elle implique un soulagement, car liée à la fin de rédaction, mais également une certaine tristesse, car fin d'une belle aventure. C'est néanmoins l'occasion de se remémorer ces 3 ($+ \varepsilon$) belles années.

Bien évidemment, mes remerciements chaleureux vont d'abord vers Nadine, Noémie et Pierre. Il y a eu une complémentarité entre nous 4 au cours de cette thèse, vous avez toujours su me guider et me rassurer. Vous êtes en possession de la pierre philosophale, nos collaborations ne s'arrêteront pas en si bon chemin. Nadine, tu as accepté de dirigé ma thèse sans m'avoir rencontré, je t'en remercie, nous avons parcouru un joli chemin ensemble, tu as toujours été de bon conseil et je suis ravi d'avoir travaillé avec toi. Noémie, j'ai toujours apprécié travailler avec toi depuis le début (stage en 2018 !). J'ai énormément appris à ton contact, que ce soit sur le plan professionnel ou personnel. Ton rire communicatif résonnera un bout de temps dans mes oreilles. Pierre, j'ai beaucoup aimé échanger avec toi, sur les modèles, les cultures associées, l'agriculture, les limites planétaires, (etc. $\times 100$).

Mes remerciements vont également aux membres de mon jury de thèse : Muriel Valantin-Morison (Rapporteure), Nathalie Vialaneix (Rapporteuse), Éric Garnier (Examinateur), Xavier Gendre (Examinateur) et Catherine Trottier (Examinatrice). Il n'est jamais facile de lire une thèse interdisciplinaire. J'espère que vous apprendrez des choses en la lisant, et qu'elle suscitera votre intérêt. J'ai hâte d'échanger avec vous lors de la soutenance.

J'ai beaucoup de reconnaissance pour les membres de mon comité de suivi individuel Bénédicte Fontez, Gilles Le Moguédec et Joseph Salmon qui ont su, lors d'entretiens annuels courts me donner de bons conseils pour le déroulé de ma thèse.

De même, les deux comités de pilotage réalisés avec des scientifiques de disciplines différentes m'ont permis d'explorer certaines hypothèses et de recadrer ma manière de voir mon sujet de thèse. Merci à Safia, Xavier, Laurent, Florian, Cyrille pour ces recommandations.

Je n'aurais pas pu réaliser cette thèse sans le financement de INRAE et de Digitag, merci pour cela. Je n'ai pas réussi à développer toutes les interactions que j'aurais souhaitées avec la communauté Digitag en raison de la pandémie et ma présence à Toulouse malgré le travail formidable réalisé par l'équipe de gestion de Digitag : Véronique Bellon-Maurel, Élodie Merlier, Gabrielle Lartia, Martha Lucia Enriquez

etc. . Les moments de retrouvailles (notamment la Digitagora) ont été pour moi formateurs et très plaisants.

La recherche est le fruit de collaborations et d'interactions multiples. Ainsi, je remercie toutes les personnes avec qui j'ai pu échanger aux cours de colloques, congrès, réunions de projets. Et surtout, la recherche est le travail collectif et cumulatif d'individus, qui ne se connaissent pas et qui ne vivent parfois pas à la même époque. Merci à Camille Noûs. Tu es le géant sur lequel, moi nain, j'ai tenté de me hisser. Tu seras un jour la personne la plus citée, je l'espère.

De même, merci au milliers d'anonymes sur Stackoverflow (et sur Stackexchange !), qui ont posé les questions 10 ans avant que je me les pose, et à celles et ceux leur ayant répondu, à la communauté R, toujours prête à aider.

Une thèse est un chemin rempli d'échanges, de rencontres, de bons moments ; je remercie tous mes collègues d'AGIR. Dans mon bureau, je tiens à remercier Neïla et son rire faisant vibrer les murs, Alexandre et son esprit facétieux, Lucie toujours prête à aider. Au-delà de ces 4 murs, passer des moments de commensalité avec Maxime, Franck, Lin, Anastasia, Solène, Julien et al. a toujours été un plaisir.

Merci à l'équipe de gestion d'AGIR, vous m'avez protégé de ma phobie administrative. Votre maîtrise de l'appareil administratif est pour moi une énigme permanente. Je suis immensément reconnaissant envers Agnès, grâce à qui j'ai pu travailler dans un environnement propre et sain (malgré les !?*! travaux pendant plus d'un an). J'ai beaucoup ri à l'écoute de tes questions brutes de décoffrage.

Aller prendre l'air pour jardiner chaque midi était toujours un moment particulier pour moi. Merci aux membres du potager d'INRAE avec qui j'ai beaucoup appris !

En dehors de la thèse, je tiens à remercier tou.te.s mes ami.e.s grâce à qui j'ai pu bien manger, décompresser, jouer, rire etc., Alex, Romane, Capucine, Gabriel, Amandine, Hugo, Maria, Vincent, Léonard, Lucas, Guglielmo, Camilito ! Aux anciennes de l'INRA, avec qui je suis toujours en contact : Marine et Augustine, vous m'avez dès le début permis d'y voir plus clair sur l'agriculture.

De même, à mes amis d'études, qui m'ont durement manqué pendant tout ce temps à Toulouse, Constip, Demeuré, Guigui, Fripouille, Nounours, Tourette : je remonte dans le Nord taper du pied !

Mes remerciements vont à ma famille, notamment à mes parents qui m'ont toujours soutenu (financièrement, moralement) et écouté depuis le début. C'est grâce à eux que j'en suis arrivé là.

Last but not least : Camille. Tu es ma collègue, amie, confidente, copine (et tortue 🐢) depuis le début (trois mois avant le début de ma thèse !). Pendant ces trois ans, tu m'as supporté (dans les 2 sens du terme) pendant ces 3 ans, tu as ri (ou pas) à mes blagues pas drôles, tu m'as questionné, tu m'as fait déguster tes merveilleuses pâtisseries (mes kilos additionnels te remercient !) et surtout, tu t'es moquée de moi. L'amour est beau quand il est vache : nous nous sommes bien trouvés.

Résumé

L'utilisation de la diversité végétale cultivée est l'un des leviers pour s'orienter vers une agriculture plus durable. Dans ce contexte, les cultures associées céréale-légumineuse sont des mélanges prometteurs, notamment en conditions bas-intrants. L'utilisation de modèles statistiques peut améliorer notre compréhension du fonctionnement de ces cultures. Pour construire ces modèles, une stratégie est d'utiliser les données issues de l'agrégation d'expérimentations mettant en jeu ces cultures. Dans mon travail de thèse, nous partons d'un jeu de données de plusieurs variables mesurées sur 8 espèces (3 céréales et 5 légumineuses) en culture pure et en culture associée, dans 35 expérimentations.

Ce type de jeu de données, peu utilisé dans la littérature, soulève des questions méthodologiques de par son hétérogénéité. Dans le Chapitre II, nous discutons du rôle de ces jeux des données dans la recherche agronomique, mettons au point une méthode utilisant la théorie des graphes pour identifier des sous-jeux de données induisant des plans factoriels complets au sein de ces jeux de données globaux et illustrons l'utilisation de splines de lissage comme méthode de réduction de dimension de variables temporelles.

Nous utilisons ensuite, dans le Chapitre III, ce jeu de données pour évaluer deux processus clés d'interactions entre plantes (complémentarité et dominance), à l'oeuvre dans les cultures associées céréale-légumineuse sur l'ensemble du jeu de données. Puis nous discutons de l'effet de deux pratiques agronomiques, le choix des espèces associées et la fertilisation, sur ces deux processus, en montrant la perturbation des interactions plante-plante induite par ces pratiques.

Enfin, dans le Chapitre IV, le jeu de données est utilisé pour développer des modèles statistiques dont les variables explicatives sont construites à partir de théories issues de l'écologie des communautés et permettant de comprendre la performance de chacune des composantes du mélange sur un sous-ensemble du jeu de données. Notre procédure de modélisation inclut réduction de dimension, imputation des données, calcul des variables explicatives, sélection de variables (dans un objectif de parcimonie) et ajustement des modèles. Les modèles utilisés combinent les capacités de prédiction des forêts aléatoires à une prise en compte de la dépendance intra-expérimentation des observations via un facteur aléatoire. Nos résultats soulignent i) le rôle des interactions positives au sein de ces mélanges en conditions bas-intrants ainsi que ii) la place prépondérante des interactions plante-plante, notamment celles liées à l'architecture du couvert dans la performance de ces mélanges.

Mon travail de thèse souligne l'utilité de l'écologie des communautés dans la compréhension du fonctionnement d'agroécosystèmes complexes et le potentiel prometteur de l'utilisation de jeux de données globaux en agronomie analysés avec des méthodes statistiques avancées.

Abstract

The use of crop diversity is one of the levers for moving towards a more sustainable agriculture. In this context, cereal-legume intercrops are promising crop mixtures, especially in low-input conditions. The use of statistical models can improve our understanding of the functioning of intercrops. In order to build these models, a strategy is to use data coming from the gathering of existing agronomic experiments involving intercrops. In my PhD, we used a dataset including several variables measured on 8 crop species (3 cereals and 5 legumes) growing in sole crops and intercrops, in 35 experiments.

This type of dataset, rarely used in the literature, raises methodological issues due to its heterogeneity. In ChapterII, we discuss the role of these datasets in agronomic research, develop a method using graph theory to identify sub-datasets inducing complete factorial plans within these global datasets, and illustrate the use of smoothing splines as a dimension reduction method for dynamic variables.

We then use the dataset in Chapter III to evaluate two key plant-plant interaction processes (complementarity and dominance) in cereal-legume intercrops across the whole dataset. We then discuss the effect of two agronomic practices, species choice and fertilization, on these two processes, showing how they impact plant-plant interactions.

Finally, in Chapter IV, the dataset is used to develop statistical models whose explanatory variables are built based on theories coming from community ecology. The goal of these models is to understand the performance of each component of the mixture on a subset of the dataset. Our modeling procedure includes dimension reduction, data imputation, computation of explanatory features, variable selection (with a parsimony objective), and model fitting. The models used combine the prediction abilities of random forests with a consideration of intra-experimentation dependence of observations via a random factor. Our results highlight i) the role of positive interactions within these low-input mixtures and ii) the predominant role of plant-plant interactions, especially those related to cover architecture, in the performance of these mixtures.

My PhD work highlights the usefulness of community ecology in understanding the functioning of complex agroecosystems and the promising potential of using global datasets in agronomy analyzed with advanced statistical methods.

Table des matières

Ma thèse, intitulée *Modéliser la performance de cultures associées céréale-légumineuse annuelles : une approche combinant écologie des communautés et science des données*, est à l'interface de la science des données, de l'agronomie et de l'écologie des communautés.

Ce travail étant interdisciplinaire, les notions abordées sont variées. Je tenterai, tout au long du manuscrit, d'être le plus clair possible pour les membres du jury amené.e.s à le lire, quelque soit leur domaine de prédilection. Cela pourra cependant constituer quelques longueurs pour les spécialistes, et je m'en excuse à l'avance.

Terme	**Définition**
Agroécosystème	Écosystème cultivé
Cultivar / Variété	Termes utilisés indistinctement pour parler de génotype
Phénotype	Ensemble des caractères apparents des plantes (mesures de hauteurs, de surfaces foliaires, etc.)
Service écosystémique	Bénéfices que les humains retirent des écosystèmes.
Stress abiotique	Stress résultant de l'action néfaste de facteurs environnementaux (sécheresse, gel, les déficiences en nutriments, etc.)
Stress biotique	Stress résultant de l'action néfaste d'un organisme vivant sur une plante (champignons, ravageurs, etc.)

Table 1 – Définitions utiles (adaptées au contexte de la thèse)

Chapitre I

Introduction

1 Statistique et science des données, disciplines clés en agronomie

1.1 Les sciences du vivant, domaine d'application pertinent pour les mathématiques appliquées

De nombreux domaines des mathématiques appliquées (équations différentielles, théorie des jeux, probabilités, statistiques, etc.) permettent des avancées significatives en sciences du vivant. Par exemple, la théorie des jeux peut être utilisée pour étudier les interactions entre les différentes espèces d'un écosystème (Han et al., 2019), tandis que les équations différentielles permettent de modéliser les mouvements et les évolutions de ces espèces dans le temps (Preisler et al., 2004). La statistique tient également une place importante puisqu'elle permet d'analyser et de modéliser les données issues de l'observation de phénomènes biologiques. Ainsi, les applications des mathématiques dans les sciences du vivant sont historiquement nombreuses et sont amenées à le rester (Cohen, 2004).

Au sein des sciences du vivant, l'agronomie est un domaine où l'utilisation des mathématiques appliquées, notamment de la statistique, est cruciale pour la compréhension des systèmes étudiés.

1.2 La statistique : outil essentiel à l'analyse d'expérimentations agronomiques

La statistique est un inventaire de techniques et de procédures qui permettent d'organiser et de faire la synthèse d'une grande quantité d'informations afin d'en dégager des conclusions utiles à la compréhension d'un phénomène (Haccoun et Cousineau, 2009). L'analyse statistique de résultats d'expérimentations agronomiques est une étape clé dans le processus de recherche (Huck et al., 1974). Les travaux de Ronald Fisher dans la première moitié du XX^{e} siècle ont fourni un cadre théorique indispensable à la réalisation de celle-ci. Parmi ses contributions notables, on peut noter l'analyse de la variance (ANOVA), le principe de randomisation des expérimentations ainsi que les expérimentations factorielles (Fisher, 1935, Street, 1990). Les analyses effectuées en agronomie ont pour but de distinguer l'effet de différents facteurs (environnement, pratiques agronomiques, etc.) du hasard lié à l'échantillonnage sur une variable quantitative observée, comme le rendement ou la quantité d'azote absorbée par la culture.

Les analyses de type ANOVA, analyse de covariance et modèles mixtes sont souvent utilisées puisqu'elles permettent, conditionnellement à certaines hypothèses (ex. normalité des données, homogénéité des variances, indépendance des observations), de fournir des conclusions fiables sur le fonctionnement du système étudié. Certains travaux montrent cependant que ces analyses sont parfois réalisées avec un certain manque de rigueur pouvant causer des erreurs quant à leur interprétation (Acutis et al., 2012).

Des comparaisons de paramètres statistiques (moyenne, médiane, variance) de populations issues de groupes différents sont également souvent effectuées via des tests statistiques. Ces tests sont réalisés avant ou après l'utilisation des méthodes citées plus haut, par exemple pour vérifier des hypothèses ou investiguer des différences observées entre traitements. Parmi les tests les plus connus, on peut citer le t-test de Student pour comparer deux moyennes ou une moyenne à une valeur théorique, le test de Bartlett pour comparer les variances de deux échantillons, ainsi que les tests de comparaisons multiples de moyenne (comme le test HSD de Tukey à la suite d'une ANOVA).

Toutes ces analyses appartiennent au domaine de la statistique inférentielle, c'est-à-dire l'inventaire des techniques permettant d'inférer certaines caractéristiques d'une population à partir d'un échantillon de cette population.

Les analyses multivariées de type analyses en composantes principales (Pearson, 1901) et dérivées (Benzécri et al., 1973, analyses en composantes multiples par ex.) font également partie des méthodes couramment utilisées en agronomie, cette fois dans un objectif d'explication de corrélations (ou liens dans le cas de variables qualitatives) observées dans un ensemble de variables mesurées. Ces analyses sortent du cadre de la statistique inférentielle, puisqu'elles ne reposent pas sur un cadre probabiliste sous-tendu par des hypothèses sur la loi statistique des phénomènes étudiés, mais sont réalisées uniquement sur les données observées.

1.3 L'augmentation de la quantité de données disponibles rend la science des données indispensable

En agronomie, la multiplication des expérimentations, publications et l'utilisation grandissante des capteurs augmentent la quantité de données disponibles pour l'analyse. Ces données proviennent de sources hétérogènes, sont de natures diverses (e.g. quantitative, qualitative, images) et peuvent être produites à un rythme élevé, notamment via les plateformes de phénotypage haut-débit (Gosseau et al., 2019). Cette

profusion de données permet d'avoir des résultats plus généralisables (Makowski et al., 2014) mais nécessite de développer des méthodes adaptées pour leur collecte, traitement et analyse afin de les valoriser autant que possible en améliorant la compréhension du système étudié. La littérature scientifique caractérise parfois ces ensembles de données par une combinaison de propriétés identifiées par les 5 V (Volume, Vélocité, Variété, Valeur, Véracité) qui définissent la notion de Big Data, dont les applications en agronomie sont multiples (Kamilaris et al., 2017).

La gestion de ces données plus complexes et hétérogènes demande des compétences qui vont au-delà des compétences statistiques classiques. Ainsi, des compétences de gestion de données, d'informatique (automatiser des processus de collecte de données par exemple) combinées à une expertise liée au domaine d'application sont indispensables pour valoriser au mieux les données recueillies. C'est dans ce contexte qu'émerge de plus en plus le recours à la science des données. La science des données est un domaine interdisciplinaire qui s'appuie sur les statistiques, l'informatique et la communication pour étudier les données et leurs propriétés afin de les transformer en connaissances et en décisions en suivant un raisonnement et une méthodologie "data-knowledge-wisdom" (Cao, 2017).

Par exemple, l'augmentation du volume de données disponibles a rendu nécessaire le développement de technologies propres à la gestion et à l'analyse de données massives comme le calcul distribué (Spark) ou le stockage massif. En termes d'algorithmes, l'augmentation des capacités de calcul a popularisé l'usage de l'apprentissage profond (les réseaux de neurones) et de l'apprentissage statistique plus généralement. L'apprentissage statistique a pour but la conception et l'utilisation d'algorithmes permettant à un ordinateur d'apprendre "automatiquement" les relations existant entre différentes variables d'un jeu de données dans un objectif de prédiction d'une variable de sortie. Parmi les algorithmes utilisés en apprentissage statistique, on peut mentionner les forêts aléatoires (Breiman, 2001) ou l'algorithme de boosting de gradient (Friedman, 2001). L'apprentissage profond est une sous-branche de l'apprentissage. Les réseaux de neurones artificiels ont été théorisés dans les années 60 (Tappert, 2019) mais peu utilisés, par limites technologiques (manque de données et faible puissance de calcul) ainsi que théoriques (algorithme de rétropropagation du gradient), avant de connaître le succès à partir des années 2000, où les capacités de calcul étaient suffisamment élevées. Une critique faite à l'apprentissage statistique est son manque d'interprétabilité. Si les prédictions réalisées sont souvent bonnes, il est difficile de savoir pourquoi, ni même d'obtenir des relations analytiques reliant variables de sortie et variables explicatives. Pour remédier à cela, des métriques d'importance des variables explicatives existent. Ces métriques permettent de quantifier la dégradation

de la qualité de prédiction à la suite de la perturbation (permutation par exemple) de la variable explicative en question (Archer et Kimes, 2008).

2 Approches prédictives en agronomie

2.1 Les approches prédictives sont essentielles en sciences du vivant

Houlahan et al., 2017 donnent une définition assez large de la prédiction : une prédiction est une déclaration sur ce qu'était, est ou sera une quantité ou un état inconnu, qui est basée uniquement sur une compréhension supposée du fonctionnement des systèmes. Ainsi, on peut mesurer le degré de compréhension d'un système par notre capacité à en prédire son comportement dans différents environnements. L'utilisation d'approches prédictives basées sur les données ("data-driven") implique donc un rôle central de la science des données en sciences du vivant (Caranta et al., 2019). Parallèlement, en écologie, certains auteurs pointent le manque d'utilisation d'approches prédictives malgré leur capacité à quantifier le degré de compréhension des écosystèmes (Houlahan et al., 2017).

Les prédictions sont réalisées via des modèles. Un modèle peut être défini comme une représentation formelle et simplifiée de la réalité, ayant pour but de reproduire au mieux le comportement du système étudié (Coquillard et al., 1997). La modélisation est un outil fréquemment utilisé en agronomie. À la convergence de l'agronomie et de l'écologie, la modélisation peut donc permettre de mieux comprendre et concevoir les agroécosystèmes (à savoir un écosystème cultivé, Neyton et al., 2018), dont l'étude se situe entre ces deux disciplines.

Deux approches de modélisation sont décrites dans la littérature scientifique selon que l'approche est pilotée par les données ou par la connaissance des processus (Gunawardena, 2014) :

- la modélisation *guidée par les données* (*data-driven*) part des données expérimentales et recherche de potentielles causalités suggérées par les corrélations observées dans les données. On parle aussi de modèles *phénoménologiques*,
- la modélisation *guidée par les concepts* (*concept-driven*) part de processus et causalités connus ou suspectés pour effectuer des prédictions. On parle aussi de modèles *mécanistes*.

En pratique, la frontière entre ces deux approches de modélisation n'est pas aussi

marquée (Ellis et al., 2020). Généralement, les modèles sont classés dans l'une ou l'autre catégorie en fonction de la prépondérance des composantes mécanistes ou phénoménologiques. Les modèles mécanistes peuvent être plus complexes, car ils rendent les mécanismes de causalité explicites, ce qui implique plus de paramètres à calibrer et de données à renseigner en entrée. Les modèles phénoménologiques sont généralement plus simples, mais leurs champ d'application et domaine de validité restent limités par la disponibilité de données statistiques (Bellon Maurel et al., 2022).

Les prédictions générées peuvent être regroupées en deux grands types selon leurs objectifs (Maris et al., 2018) :

1. les prédictions *corroboratives* ont pour but de comprendre les phénomènes observés, dans un objectif de création de connaissances scientifiques ainsi que de validation/réfutation d'hypothèses. Ce type de prédictions correspond plus aux modèles *phénoménologiques*,
2. les prédictions *anticipatives* ont pour objectif d'aiguiller les décisions (politiques publiques, choix de gestion, etc.), via une anticipation de l'avenir, liée à des théories considérées comme justes (rapports du GIEC, IPCC, 2013). Ce type de prédictions correspond, selon les situations, aux modèles *mécanistes* ou *phénoménologiques*.

Les deux types de prédiction diffèrent sur trois aspects : les observations, le temps et le domaine de validité. La confrontation des prédictions corroboratives à la théorie permet de plus (confirmer) ou moins (infirmer) cette théorie (Popper, 1959), là où les prédictions anticipatives considèrent ces théories comme vraies. De plus, les prédictions corroboratives sont indissociables des observations, puisque ces dernières ont permis de les construire. Inversement, les prédictions anticipatives n'ont pas pour objectif direct d'être validées par les observations (elles en sont tout de même indissociables, dans la mesure où les observations servent à calibrer les modèles les générant). Enfin, ces approches ont un rapport au temps qui diffère puisque les prédictions anticipatives décrivent un futur probable alors que les prédictions corroboratives sont indépendantes du temps : une valeur prédite en fonction d'une nouvelle observation ne dépend pas du "moment" où cette observation est réalisée (Maris et al., 2018).

2.2 La modélisation en agronomie, un outil puissant et complémentaire à l'expérimentation

Les nombreuses interactions plante/environnement/pratiques agronomiques ayant lieu au sein d'un agroécosystème ne peuvent pas être appréhendées uniquement par le biais de l'expérimentation. De nombreux facteurs rendent les expérimentations agronomiques coûteuses (humainement, matériellement, financièrement) de par la nature des mesures à effectuer, ainsi que le nombre d'environnements dans lesquels il faudrait les effectuer. À titre d'exemple, supposons qu'on veuille étudier, en plan factoriel complet, avec 3 répétitions par traitement, l'effet de 3 niveaux de fertilisation azotée sur le rendement de 6 génotypes de blé dur en culture pure. Cela fait alors $(3 \times 3 \times 6) = 54$ parcelles à gérer sur le cycle de culture, alors même que seules deux pratiques agronomiques ont été évaluées (la fertilisation azotée et le choix variétal), sur une espèce, et pour une combinaison de site-année donnée.

Pour monter en généricité, la profusion et la diversité des données en agronomie permettent le développement de modèles agronomiques. En agronomie, plusieurs types de modèles existent (mécanistes, phénoménologiques, à dire d'experts) avec des variables différentes à modéliser (occurrences de maladies, présence de ravageurs, simulation du rendement sous diverses conditions pédo-climatiques, etc.). Les modèles agronomiques sont développés à différentes échelles spatiales et temporelles (Figure I.1) qui impliquent des méthodes, considérations, utilisations et limitations différentes (Pearson et Dawson, 2003). En effet, les facteurs affectant les processus biologiques sont multiples et de nature différente, *e.g.* facteurs climatiques, types de sol, pratiques agricoles. Cependant, ces différents facteurs n'influencent pas le système de la même manière selon les échelles considérées (région, territoire, parcelle agricole, plante) si bien que les facteurs pris en compte dans les modèles dépendent des objectifs de modélisation.

La modélisation des systèmes de culture a pour but d'explorer les interactions plante/environnement/pratiques pour comprendre le fonctionnement des systèmes de culture et adapter des choix sur leur gestion (Chenu et al., 2017). L'utilisation de modèles a permis de nombreuses avancées en agronomie, par exemple pour identifier des facteurs limitant la croissance des plantes (Brisson et al., 2010), évaluer l'impact de pratiques agricoles (Huth et al., 2010) ou adapter des pratiques au changement climatique (Lamichhane et al., 2020).

Les modèles de culture sont élaborés, calibrés et validés à partir de données issues d'expérimentations agronomiques. Ces expérimentations sont réalisées dans des conditions pédo-climatiques (température, précipitations, sol) et des pratiques agronomiques

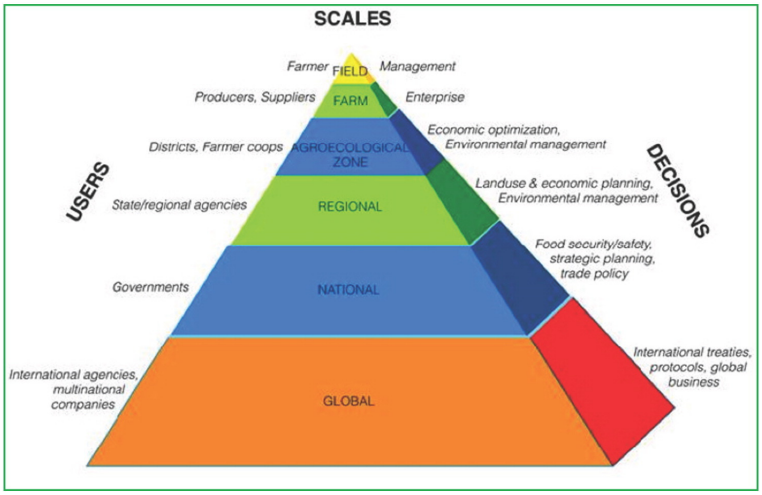


Figure I.1 – Différentes échelles de modélisation en agronomie (tirée de Jones et al., 2017)

données. L'expérimentation permet également de comprendre le fonctionnement de la culture étudiée, et de décrire quantitativement les processus étudiés (e.g. réponse à des stress abiotiques, architecture de la plante, phénologie). Ainsi, l'expérimentation agronomique est nécessaire à la construction de modèles de culture robustes.

La modélisation permet donc la généralisation et la compréhension de l'effet de pratiques agricoles sur le rendement (par exemple) dans un contexte environnemental changeant. Ainsi, expérimentation et modélisation sont des approches complémentaires (Chenu et al., 2017).

Les premiers modèles agronomiques datent des années 50 (Keating et Thorburn, 2018). Au cours de la seconde moitié du XX^{e} siècle, la puissance des ordinateurs s'est accrue et les systèmes et processus modélisés ont été de plus en plus nombreux, avec une résolution plus fine (Jones et al., 2017).

2.3 Cadre d'application : les cultures associées céréale-légumineuse

La culture associée est une pratique agricole qui consiste à implanter dans une même parcelle au moins deux espèces pendant une période significative de leur croissance (Willey, 1979). Ces cultures constituant le système que je modélise dans ma thèse, je présente ici succinctement i) certains problèmes liés à l'agriculture "conventionnelle" et ii) les raisons pour lesquelles les cultures associées, en tant que levier de diversification, constituent une pratique prometteuse.

2.3.1 La diversification en agriculture, pourquoi ?

Retour sur l'agriculture conventionnelle

Depuis la fin de la seconde guerre mondiale, l'agriculture dans les pays occidentaux a connu des changements majeurs. La mécanisation grandissante des exploitations agricoles a entraîné une augmentation de la taille moyenne des exploitations conjointement à une réduction du nombre d'agriculteurs et d'agricultrices dans la population. L'utilisation massive de fertilisants de synthèse et de l'irrigation a permis de réduire l'influence des stress abiotiques subis par les plantes. Enfin, l'utilisation de produits phytosanitaires a permis de mieux maîtriser les stress biotiques (ravageurs / adventices (mauvaises herbes) / maladies). En parallèle de l'augmentation des intrants, l'agriculture s'est simplifiée à différentes échelles. À l'échelle des territoires, l'expansion des parcelles agricoles a conduit a une simplification des paysages (Figueroa et al., 2020). À l'échelle des parcelles, la diversité génétique et spécifique ont également été réduites. Ainsi, en France, plus de 80 % de la surface cultivée implique seulement 5 cultures (blé, orge, maïs, colza, tournesol (Agreste, 2022)). Ces changements majeurs ont permis ce que certains auteurs appellent la Révolution Verte (Pingali, 2012), qui se traduit par une forte augmentation des rendements au cours du XX^{e} siècle, ce qui a permis de réduire la sous-nutrition et la malnutrition (Tilman et al., 2002).

Cependant, ces améliorations ont été accompagnées d'impacts environnementaux problématiques i) globaux comme l'érosion des sols (Pimentel, 2006), les fortes émissions de gaz à effet de serre (Vergé et al., 2007) ou l'érosion de la biodiversité (Gonthier et al., 2014) et ii) locaux comme la lixiviation (De Notaris et al., 2018) ou la pollution de l'eau (Parris, 2011).

Dès lors, un des défis auquel fait face l'agriculture est de continuer à produire suffisamment pour nourrir une population grandissante tout en réduisant les impacts environnementaux négatifs. Certains auteurs parlent d'intensification soutenable de

l'agriculture (Martin-Guay et al., 2018). Ce défi apparaît d'autant plus complexe que l'agriculture fait déjà face à certains impacts du changement climatique qui n'auront de cesse d'augmenter au cours du XXI^{e} siècle (Praveen et Sharma, 2019).

Diversification en agriculture

L'un des leviers utilisables pour une agriculture plus durable est la mobilisation de la diversité, qui peut prendre plusieurs formes, à plusieurs échelles (Figure I.2, Duru et al., 2015b).

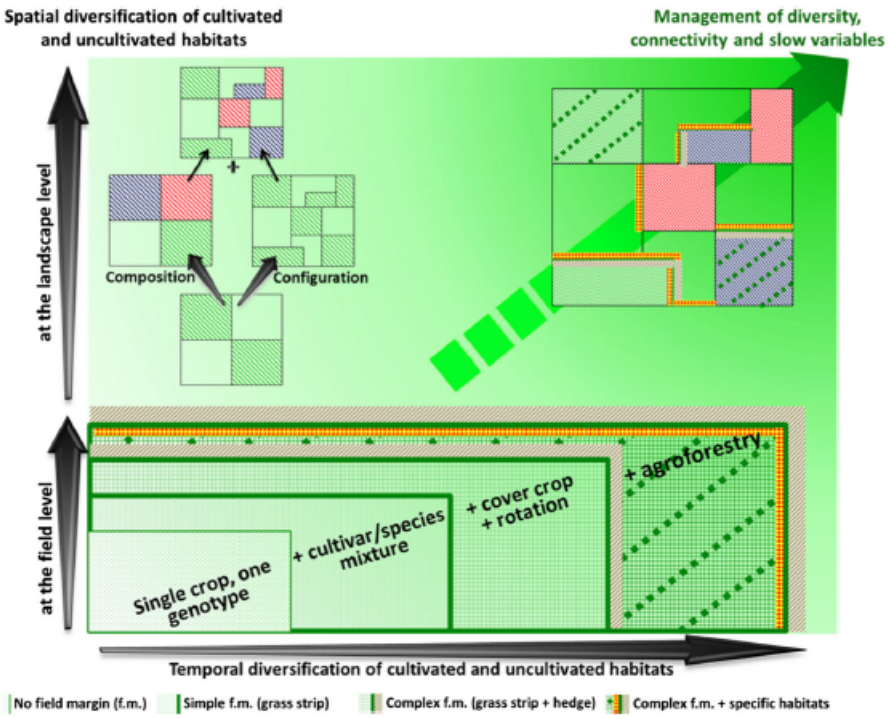


Figure I.2 – Différentes échelles (spatiales et temporelles) de diversification en agriculture (tirée de Duru et al., 2015b). Sur l'axe horizontal, on observe diverses pratiques de diversification (cultures associées, rotations, agroforesterie) se traduisant par une complexification du système cultivé (plus d'espèces, plus de haies/bandes enherbées). Sur l'axe vertical, on voit que la diversification peut également s'appliquer à l'échelle du paysage, se traduisant par des paysages plus complexes et divers.

Au sein des paysages, la diversification prend la forme de la réduction de la taille des parcelles, de l'introduction d'habitats semi-naturels et naturels (haies, prairies permanentes ou temporaires, Veres et al., 2013) pouvant servir d'habitats aux prédateurs de ravageurs (Sirami et al., 2019).

Dans une exploitation agricole, des leviers de diversification temporelle existent, tels que l'inclusion de cultures intermédiaires entre deux cultures de rente (Couëdel et al., 2018), les rotations plus longues composées d'espèces plus variées, ayant notamment pour objectif de casser les cycles des bioagresseurs, comme illustré pour les maladies (Colbach et al., 1997) ou les adventices (Koocheki et al., 2009).

À l'échelle de la parcelle, on peut mobiliser la diversité intra-spécifique, via des mélanges variétaux, ou la diversité inter-spécifique, via la culture de mélanges d'espèces. Cette pratique vise à maintenir une productivité tout en réduisant l'usage d'intrants via une utilisation plus complémentaire des ressources et à améliorer la résistance aux stress biotiques (Kiær et al., 2009, Reiss et Drinkwater, 2018).

Dans ce travail, nous nous focaliserons sur la combinaison de deux leviers de diversification à l'échelle de la parcelle : l'utilisation i) des légumineuses au sein ii) d'associations bispécifiques.

2.3.2 Cultures associées : principes généraux

L'utilisation de cultures associées annuelles est un des leviers de diversification de l'agriculture à l'échelle parcellaire (Martin-Guay et al., 2018). Les cultures associées sont de plus en plus étudiées dans la littérature scientifique, avec un nombre d'articles scientifiques traitant du sujet en constante augmentation depuis 2005 (Lv et al., 2021).

Un des objectifs de cultiver deux espèces sur une même parcelle est de s'appuyer sur les interactions entre plantes plutôt que sur l'apport d'intrants pour assurer le bon fonctionnement des cultures. Il s'agit alors de maximiser les interactions positives entre plantes (complémentarité, facilitation) plutôt que les interactions négatives (compétition) dont nous donnons ici les définitions :

- compétition : les espèces qui coexistent utilisent une même ressource en quantité limitée (Begon et al., 2005),
- complémentarité de niche : les espèces qui coexistent utilisent des ressources différemment réduisant ainsi la compétition inter-spécifique (Hooper et al., 2005). L'utilisation des ressources peut différer selon plusieurs échelles : la forme de la ressource, temporelle et spatiale,
- facilitation : l'une des espèces modifie l'environnement au sein duquel coexistent plusieurs espèces, en améliorant la croissance, la survie ou la reproduction d'une autre espèce (Bronstein, 2009).

Les services écosystémiques (bénéfices) rendus par les cultures associées sont multiples (Maitra et al., 2021). Plus particulièrement, les cultures associées céréale-légumineuse permettent, en contexte bas-intrants (bas niveaux de fertilisation), un rendement plus élevé via une utilisation complémentaire des ressources (Ghaley et al., 2005, Bedoussac et al., 2015, Yu et al., 2015), notamment grâce à la fixation de l'azote atmosphérique par la légumineuse. Cet aspect est particulièrement intéressant pour aller vers une agriculture moins dépendante aux intrants de synthèse. Je décris plus précisément les avantages agronomiques des cultures associées dans le Chapitre III de ce manuscrit.

2.4 État de l'art de la modélisation des cultures associées

La modélisation du fonctionnement des cultures associées est un domaine de recherche en expansion. Ainsi, nombreux sont les projets de recherche (H2020 ReMIX/Intercrop-ValuES, PPR MOBIDIV, etc.) sur les cultures associées ayant un *workpackage* dédié à leur modélisation, ayant pour but l'acquisition de connaissances sur ces systèmes. Leur modélisation est cependant plus complexe que celle des cultures pures conventionnelles plus homogènes (Duru et al., 2015a). En effet, en plus des interactions intraspécifiques, la modélisation devra aussi prendre en compte les interactions interspécifiques (Mao et al., 2015).

Plusieurs approches de modélisation de la performance des cultures associées existent (Figure I.3).

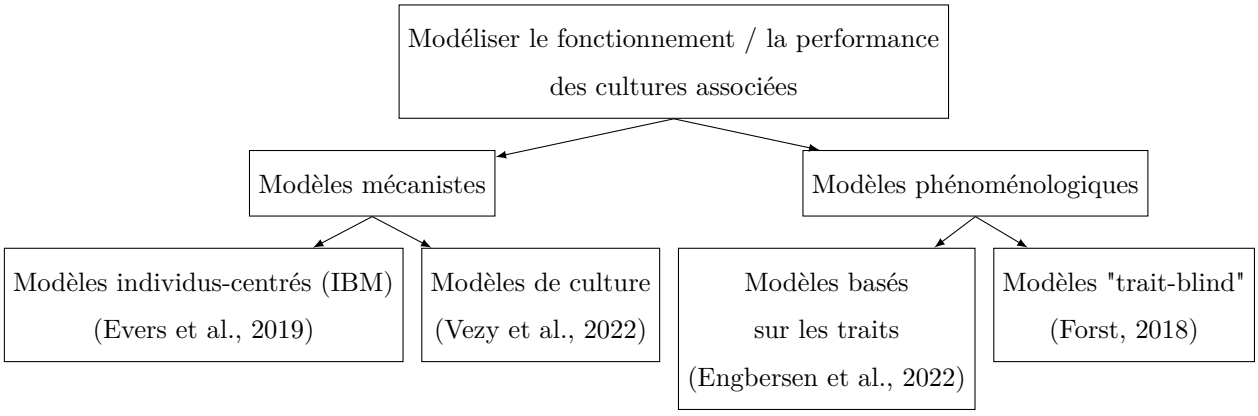


Figure I.3 – Différentes approches de modélisation pour décrire le fonctionnement et la performance des cultures associées.

2.4.1 Modèles mécanistes

La modélisation mécaniste, basée sur les processus écophysiologiques (production de biomasse, photosynthèse, etc.), permet d'explorer les effets de certains facteurs sur le rendement des cultures associées. C'est l'approche de modélisation la plus utilisée actuellement pour prédire le fonctionnement des cultures associées (Gaudio et al., 2019). Parmi ces modèles, deux grands types de modèles, qui diffèrent par l'échelle spatiale considérée, sont développés et constamment améliorés (Evers et al., 2019, Berghuijs et al., 2021) :

- les modèles individus-centrés (*i.e. individual-based models* IBM), dans lesquels chaque plante est représentée avec une architecture plus ou moins explicite, permettent de répondre à des questions de recherche très précises, notamment sur le partage des ressources (ex. : comment évolue le partage de la lumière au sein d'un couvert blé dur - pois, Barillot et al., 2014),
- les modèles de culture, tels que STICS (Paff et al., 2020, Vezy et al., 2022), APSIM (Berghuijs et al., 2021) ou Minimalist Mixture Model (M3, Berghuijs et al., 2020), permettent de simuler des variables à l'échelle du couvert comme le rendement ou la biomasse.

Ces deux types de modèles requièrent la collecte de données très précises à l'échelle de l'individu ou du couvert (Figure I.4). En effet, la description fine des processus écophysiologiques dans les modèles mécanistes nécessite de calibrer de nombreux paramètres (Weih et al., 2022). Les paramètres concernent la phénologie (périodes entre stades développement clés), la physiologie (taille maximale des graines, proportion de la biomasse allouée aux différents compartiments aériens, etc.), la sensibilité à la photopériode et l'efficience d'utilisation de la lumière. Cette complexité fait que peu de caractéristiques des espèces en mélange sont actuellement calibrées dans les différents modèles mécanistes. En mélange, un modèle mécaniste doit être suffisamment générique pour intégrer plusieurs espèces et être capable de simuler les interactions entre plantes. Cette couche de complexité additionnelle rend les modèles mécanistes plus complexes à concevoir en cultures associées qu'en cultures pures.

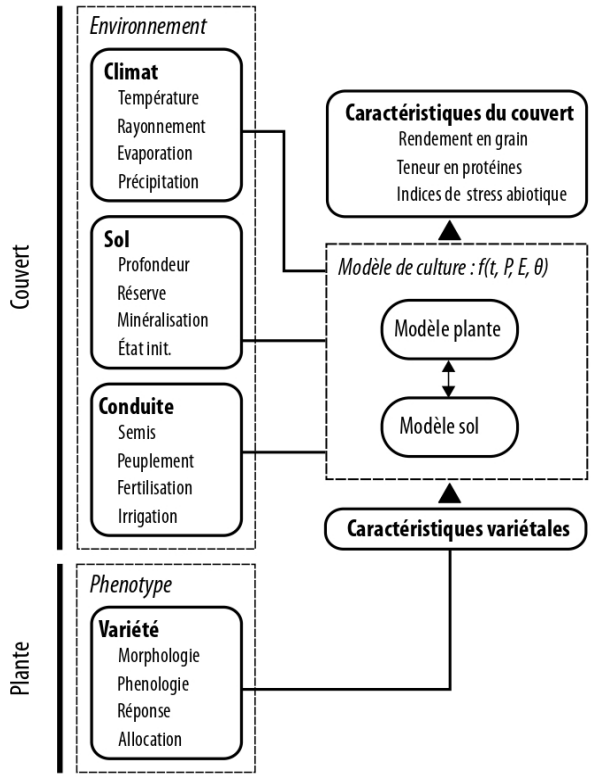


Figure I.4 – Schéma simplifié des entrées et sorties d'un modèle de culture mécaniste

2.4.2 Modèles phénoménologiques

Les modèles phénoménologiques n'ont pas pour objectif la modélisation fine des processus écophysiologiques ayant lieu au sein des cultures associées. Dans le cadre de la compréhension du fonctionnement et de la conception de cultures associées, la littérature scientifique sur la modélisation identifie des approches basées i) sur les valeurs de traits fonctionnels ("*trait-based approaches*") et ii) sur l'estimation de la performance ("*trait-blind approaches*") en association de variétés ou d'espèces sans tenir compte des valeurs de traits fonctionnels (*i.e.* toute caractéristique mesurable à l'échelle individuelle, sans référence à l'environnement ou autre niveau d'organisation, et qui impacte la capacité sélective indirectement via ses effets sur la croissance, la reproduction ou la survie ; Volaire et al., 2020). Les deux approches présentent des avantages et des limites. L'approche "*trait-blind*", initialement développée pour les mélanges variétaux (Knott et Mundt, 1990), est basée sur une estimation de l'aptitude au mélange de chacune des composantes, et la décomposition de cette aptitude entre performance globale (en culture pure), aptitude générale au mélange, et aptitude spécifique au mélange (aptitude de la variété (ou espèce) i à être associée à la variété (ou espèce) j ; Forst, 2018). Cette approche a l'avantage de ne pas nécessiter le phénotypage de traits fonctionnels mais nécessite de tester un grand nombre de combinaisons de variétés/espèces. Les approches dites "*trait-based*" permettent de relier les traits fonctionnels à des services écosystémiques rendus par les plantes (Violle et al., 2007). En décrivant chaque composante du mélange (espèce ou variété) par ses traits fonctionnels, le but est de comprendre le fonctionnement du mélange et son impact sur le rendement pour concevoir des systèmes de culture pertinents pour un ou plusieurs services écosystémiques (Barot et al., 2017). Une des limites de cette approche est qu'elle requiert le phénotypage de nombreuses espèces/variétés en contexte de culture pure et de mélange. Les traits fonctionnels ont des valeurs qui varient selon les environnements, contextes (culture pure / mélange) et dont l'intérêt dépend des pratiques agricoles, ce qui rend le phénotypage encore plus complexe.

2.4.3 Approches complémentaires

Il est également possible de combiner les différentes approches de modélisation via des approches de modélisation hybride (Gaudio et al., 2022). L'idée est d'utiliser plusieurs modèles pour tirer parti des avantages de chacun. Les sorties d'un modèle mécaniste peuvent être les entrées d'un modèle phénoménologique (Ellis et al., 2020). Meunier et al., 2022a ont par exemple proposé de simuler un ensemble de services écosystémiques fournis par des cultures associées céréale-légumineuse via l'utilisation d'une chaîne de modélisation couplant des approches de modélisation différentes

(modèle de culture mécaniste, modèle statistique, modèle qualitatif hiérarchique). La chaîne de modélisation proposée par les auteurs est intégrée dans un jeu sérieux à destination de conseillers/conseillères agricoles, agriculteurs/agricultrices, étudiant.e.s afin de fournir un appui à la mise en place et à la connaissance des cultures associées (Meunier et al., 2022b).

3 Positionnement, questions, et démarche de recherche

Augmenter la diversité végétale dans les systèmes agricoles est un levier de durabilité de l'agriculture. Comprendre des systèmes complexes comme les cultures associées nécessite d'identifier des processus favorables à leur performance (dans notre cas, rendement et dérivés). L'enjeu est de produire des connaissances utiles à la compréhension et à la gestion de ces cultures complexes. Dans ma thèse, je prends le parti que la modélisation phénoménologique est un outil pertinent pour cette production de connaissance.

3.1 Positionnement de la thèse

Je me positionne dans une approche de modélisation phénoménologique basée sur les valeurs de traits. Ce choix est motivé par plusieurs raisons. Premièrement, je pense que la complexité des systèmes considérés (interactions interspécifiques, hétérogénéité des conditions, etc.) en comparaison avec des systèmes plus conventionnels rend la modélisation mécaniste plus complexe. Deuxièmement, les processus à modéliser dans une approche mécaniste sont nombreux et les prendre en compte suppose de disposer des données expérimentales associées. Troisièmement, je souhaite produire des connaissances utiles à la compréhension et à la gestion de ces cultures complexes, et les modèles phénoménologiques sont plus adaptés pour cette tâche. Enfin, les modèles phénoménologiques, incluant des observations phénotypiques (observations concernant des caractères observables des plantes, ex. mesures de hauteur, de surface foliaire, etc.), ont été peu utilisés jusqu'à présent pour la modélisation de la performance des cultures associées.

Pour autant, les modèles que je construis ne sont pas basés sur des valeurs de traits brutes mais sur des prédicteurs pensés pour refléter une certaine connaissance du système (interactions plante-plante, impact du mélange sur chaque espèce). Plus précisément, j'utilise des prédicteurs dont le calcul est basé sur des différences entre valeurs de traits entre espèces au sein du couvert et entre cultures pure et associée. Ce

choix est motivé par l'application de théories issues de l'écologie des communautés, qui suggèrent que deux processus sont particulièrement impliqués dans la performance des couverts mélangés : la complémentarité de niche et la plasticité phénotypique (capacité d'un génotype à exprimer différents phénotypes sous différents environnements ; Litrico et Violle, 2015, Montazeaud et al., 2018). En plus de ces variables, j'utilise des informations sur les pratiques agronomiques. Une hypothèse clé dans mon travail est que la connaissance fine des processus ayant lieu dans une culture associée est difficile à mesurer et n'est pas nécessaire pour prédire le rendement des cultures associées dans les différents pédo-climats considérés (comme ce serait le cas dans des modèles mécanistes). En ce sens, les prédictions réalisées par l'approche proposée sont de types corroboratives. Ainsi, l'hypothèse est que classer ces prédicteurs et hiérarchiser leur importance permettra un gain de connaissances sur ces cultures.

3.2 Questions de recherche

Le contexte décrit jusqu'ici induit plusieurs questions de recherche qui guideront mon travail de thèse. Les prédicteurs étant regroupés en trois grandes catégories (différences entre espèces au sein du mélange, différences entres cultures pures et associées, environnement), une question qui se pose naturellement est l'importance relative de ces prédicteurs dans l'explication de la performance des cultures associées. Les prédicteurs étant construits pour représenter au mieux le fonctionnement de ces agroécosystèmes à partir de théories issues de l'écologie des communautés, la deuxième question qui guide mon travail est de qualifier l'apport de cette discipline à la compréhension du fonctionnement des cultures associées. Enfin, le matériel me permettant de travailler sur ces questions implique également des enjeux méthodologiques qui forment une autre question de recherche.

1. Quels prédicteurs expliquent le mieux la performance des cultures associées ?
2. Quel est l'apport de l'écologie des communautés pour expliquer la performance des cultures associées ?
3. Quels sont les enjeux méthodologiques liés à l'utilisation de jeux de données globaux ?

3.3 Démarche de recherche

3.3.1 Agréger les données pour la construction de modèles

Dans une optique d'approche "*trait-based*", pour remédier au problème de phénotypage coûteux et complexe, une possibilité est de réunir les résultats de nombreuses expérimentations qui ont déjà été menées (Michener et Jones, 2012), ce qui est le cas sur les cultures associées. De nombreuses données sont donc disponibles (Zamir, 2013, Heidorn, 2008). Dans ma thèse, nous disposons d'une base de données de mesures de traits sur 5 espèces de légumineuse et 3 espèces de céréale. Certaines des espèces (pois, féverole, blé dur/tendre, orge, lentille) sont représentées par plusieurs variétés et dans différents environnements (sites, années). Des mesures de traits ont été réalisées sur ces espèces en culture pure et en association. L'agrégation de données issues d'expérimentations hétérogènes, ayant des objectifs et pédo-climats différents, dans un jeu de données global, est un enjeu méthodologique en soi que nous traiterons dans le premier article de cette thèse (Chapitre II, Figure I.5).

3.3.2 Modélisation

L'objectif est de développer des modèles phénoménologiques en partant des données acquises via l'agrégation des expérimentations déjà effectuées. Les prédictions réalisées ont pour objectif de mieux comprendre le fonctionnement des cultures associées. Une des originalités de ce travail est l'utilisation de méthodes statistiques hybrides entre apprentissage statistique et statistique inférentielle (forêt aléatoire avec facteur aléatoire) pour décrire certains des processus ayant lieu au sein des cultures associées céréale-légumineuse. En utilisant des prédicteurs conçus pour quantifier certaines des interactions plante-plante et plante-environnement ayant lieu au sein de la culture associée, un des partis pris est que le classement de l'importance de ces prédicteurs permet d'identifier certains processus d'intérêt au sein de ces cultures.

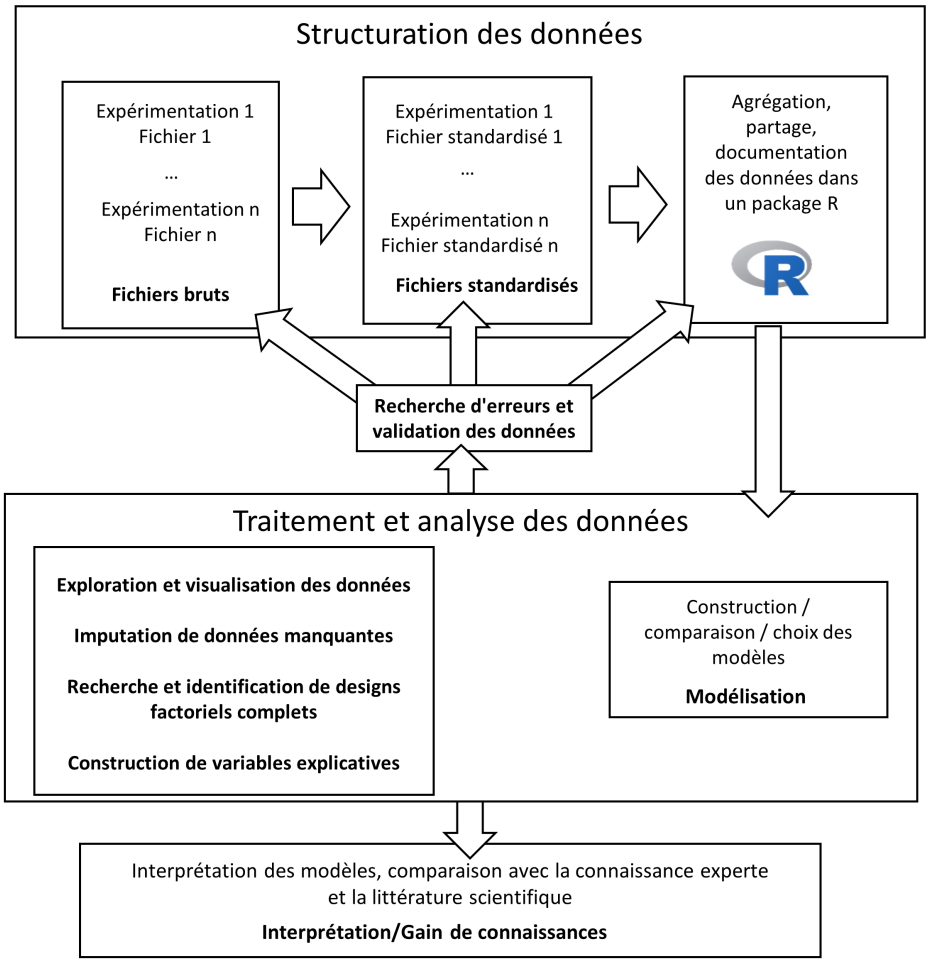


Figure I.5 – Étapes de constitution et d'analyse de jeux de données globaux

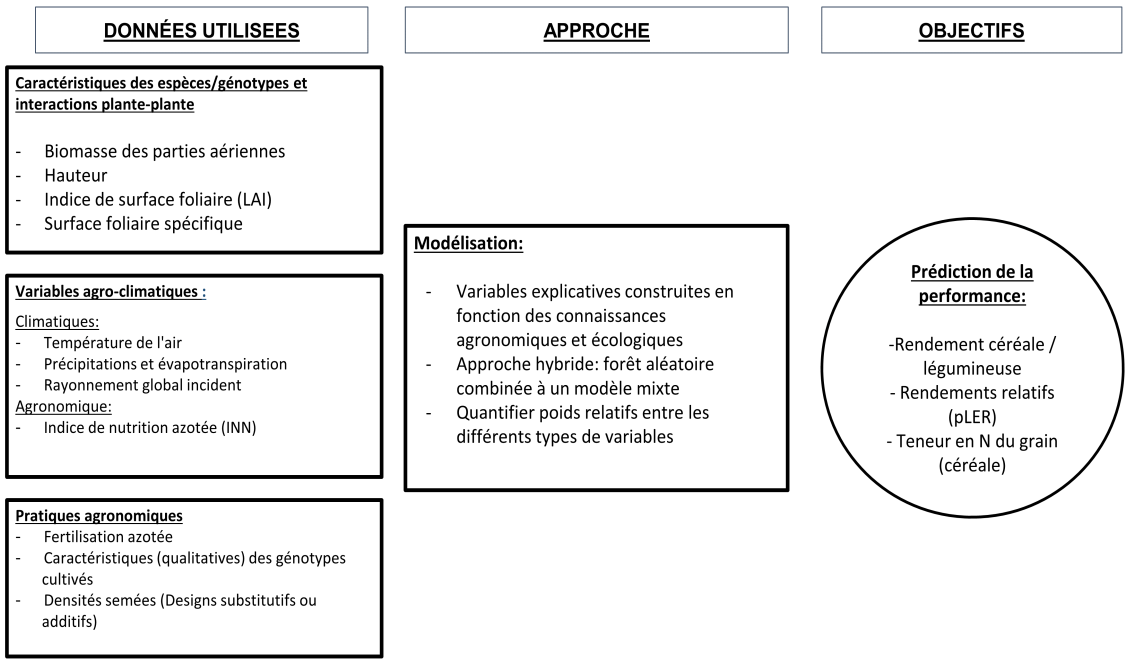


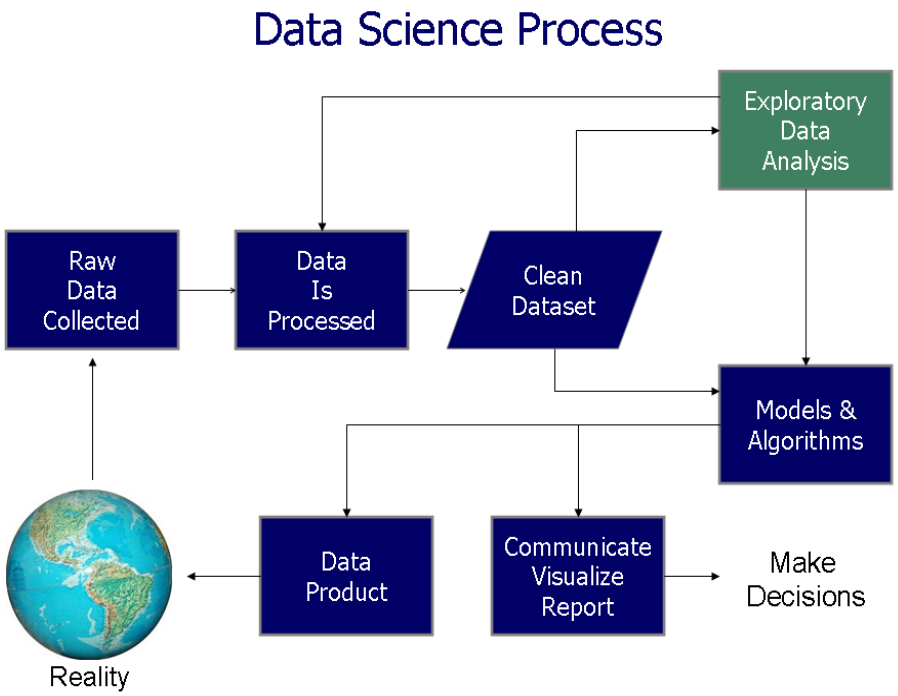
Figure I.6 – Démarche de modélisation de la thèse (de gauche à droite) : 1) Je dispose d'un ensemble de variables de types différents (mesures de traits, variables agro-climatiques, pratiques agronomiques) ; 2) Je modélise mon système en construisant des variables explicatives censées représenter mon système ; 3) Je réalise des prédictions de la performance des cultures associées.

3.4 Plan du manuscrit

Le manuscrit est structuré en 4 chapitres (en plus de l'introduction). Le Chapitre II présente le jeu de données étudié et décrit ensuite, *via* une section sous forme d'article, certains enjeux méthodologiques et épistémologiques qui y sont liés. Dans le Chapitre III, je résume succinctement certains avantages agronomiques des cultures associées dans un premier temps, puis j'étudie le choix de l'espèce et de la fertilisation sur le fonctionnement des cultures associées *via* un article publié dans Agronomy For Sustainable Development. Dans le Chapitre IV, je décris ma démarche de construction de prédicteurs pour la modélisation et l'identification des déterminants de la performance des cultures associées. Enfin, le dernier chapitre de ma thèse est dédié à la discussion générale des résultats présentés dans les différents chapitres, ainsi qu'à une présentation des perspectives liées à mon travail.

Chapitre II

Collecter et structurer des données hétérogènes



1 Enjeux et questionnements autour du jeu de données étudié

Mon travail repose sur la réutilisation de données expérimentales (observations phénotypiques dans différents pédo-climats) collectées dans différents projets de recherche. Bien que certaines de ces données aient déjà été valorisées dans des publications scientifiques, l'agrégation de celles-ci dans un jeu de données global a un fort potentiel de généralisation des connaissances. Dans ce chapitre, je présente d'abord certaines des caractéristiques du jeu de données mobilisé dans ma thèse. Je discute ensuite les enjeux méthodologiques et épistémologiques que présente cette stratégie de recherche dans un article soumis à la revue *Computer and Electronics in Agriculture*.

1.1 Collecte et agrégation de données en agronomie

Certains auteurs utilisent le terme de données obscures ("Dark data" ; Heidorn, 2008) pour désigner les données sans documentation, nettoyées à la main, stockées sur des ordinateurs personnels, rarement réutilisables et réutilisées. En agronomie, la nature des expérimentations, le format de stockage des données et les pratiques de recherche encore peu orientées vers la science ouverte font que de nombreuses données expérimentales sont perdues ou peu valorisées et appartiennent à cette catégorie (Zamir, 2013, Senft et al., 2022). Pour ces raisons, la compilation et l'agrégation de résultats d'expérimentations multiples dans des jeux de données globaux est une pratique rare mais à fort potentiel.

Les jeux de données globaux diffèrent des méta-analyses basées sur les résultats d'expérimentations déjà publiés, où l'information est accessible directement dans les publications constituant la méta-analyse. Ainsi, de nombreuses publications détaillant une méthodologie pour ce type d'analyse existent (Makowski et al., 2019). Ce n'est pas le cas pour les jeux de données globaux. Il est donc nécessaire de traiter certains des enjeux méthodologiques liés à ces jeux de données, notamment i) quelles étapes de conception sont à mettre en oeuvre, ii) comment identifier des sous-jeux de données induisant des plans factoriels complets dans des jeux de données où les plans sont déséquilibrés par construction et iii) comment traiter des données où les observations ne sont pas effectuées au même moment.

Le jeu de données mobilisé dans ma thèse va servir d'étude de cas. Il est le fruit d'un travail collectif. Noémie Gaudio et Catherine Bonnet (INRAE, UMR DYNAFOR) ont

en amont réalisé un travail de collecte et d'homogénéisation des données dans plusieurs fichiers (format csv). Pour ma part, j'ai joué un rôle essentiel dans la constitution et le nettoyage du jeu de données utilisé pour mes analyses. Tout d'abord, j'ai écrit un code pour lire les données. Ce code a été intégré dans un paquet R, ce qui a facilité son utilisation par des collaborateurs et collaboratrices. Ensuite, j'ai consacré du temps à nettoyer le jeu de données en recherchant les données aberrantes, afin d'assurer la qualité et la fiabilité des résultats obtenus dans mes analyses. Notre travail a donc contribué à fournir un jeu de données fiable pour le déroulé et la suite de mon travail.

Il regroupe les résultats de 35 expérimentations impliquant des cultures associées céréale-légumineuse semées et récoltées en même temps (synchrones), ainsi que les cultures pures correspondantes. Il contient 18 (resp. 17) expérimentations mettant en jeu des cultures d'hiver (resp. de printemps). Une partie de ces expérimentations ont déjà été valorisées de manière indépendante (Knudsen et al., 2004, Corre-Hellou et al., 2007, Hauggaard-Nielsen et al., 2008, Hauggaard-Nielsen et al., 2009, Launay et al., 2009, Naudin et al., 2010, Bedoussac et Justes, 2010a, Bedoussac et Justes, 2010b, Pelzer et al., 2012, Naudin et al., 2014, Pelzer et al., 2016, Tang et al., 2016, Kammoun et al., 2021, Moutier et al., 2022). Certaines expérimentations n'avaient par contre pas été valorisées et la constitution de ce jeu de données global a permis de remédier à cela (Meunier et al., 2022a, Louarn et al., 2021, Gaudio et al., 2021a, Mahmoud et al., 2022). Une expérimentation agronomique étant généralement coûteuse humainement et financièrement, la valorisation sous forme de publication scientifique ou autre rendu est une juste reconnaissance du travail effectué.

1.2 Présentation du jeu de données

1.2.1 Définitions clés

Une *expérimentation* est définie comme une combinaison site-parcelle-année. J'utiliserai le terme *unité expérimentale* pour désigner une unique combinaison de pratiques agronomiques (e.g. variété de la céréale/légumineuse, densité de chacune des espèces du mélange ou de la culture pure, niveau de fertilisation azotée, agencement spatial de la culture associée, Figure II.1). Les unités expérimentales de ce jeu de données sont les *individus statistiques*, c'est-à-dire que je les considère comme étant l'élément constitutif de la population dans laquelle j'analyse les observations.

Il existe ainsi une grande diversité de situations rencontrées entre les expérimentations, et donc une forte hétérogénéité dans le jeu de données, que ce soit via les pratiques culturales testées, les espèces associées ou les variables mesurées (Figure II.2).

Figure II.1 – (a) Mélange blé dur / féverole dans le rang et (b) mélange blé dur / pois en rangs alternés (source : Laurent Bedoussac, INRAE UMR AGIR)

1.2.2 Présentation des expérimentations

Les expérimentations qui constituent mon jeu de données sont réparties dans 5 pays européens (France, Danemark, Italie, Allemagne, Angleterre) et couvrent une période allant de 2001 à 2018 (Tableau II.1). Elles regroupent des cultures associées (et leurs cultures pures correspondantes) d'hiver et de printemps, caractérisées par des environnements climatiques contrastrés (Figure II.3).

Figure II.2 – Diversité des traitements dans le jeu de données par facteur (colonnes) et expérimentation (lignes). Au sein de chaque colonne, chaque rectangle est une modalité du facteur considéré. Un rectangle au sein d'une ligne et d'une colonne données indique la présence d'au moins un individu statistique avec la modalité du facteur correspondant. Toutes les expérimentations ont en commun au moins une modalité non fertilisée (rectangles rouges de la colonne "Niveau de fertilisation"), mais seulement 12 ont plusieurs modalités de fertilisation testées. Pour ce qui est des densités relatives, la modalité la plus représentée est la modalité 50% - 50% (rectangles turquoises de la colonne "Densités relatives"). L'arrangement spatial du semis (dans le rang ou en rangs alternés) n'est pas un facteur fréquemment testé dans les expérimentations (hormis dans 3 d'entre elles).

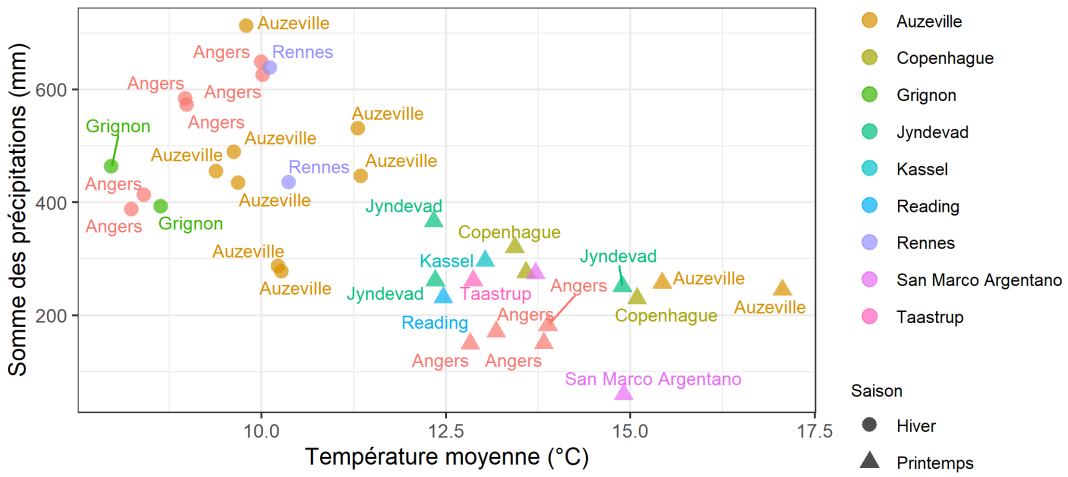


Figure II.3 – Somme des précipitations (mm) et température moyenne (°C) des expérimentations pendant le cycle de culture

Espèces	Pays	Site	Année	Texture du sol (argile - limon - sable, %)	Fertilisation N (kg/ha)	Semis	Densités relatives (céréale - légumineuse)	Arrangement spatial	Nombre de cultivars (céréale / légumineuse)
Orge de printemps / Féverole	Danemark	Copenhague	2001, 2002, 2003	24-29-47	0	substitutif	0.5-0.5	Dans le rang	2 / 1
		Jyndevad		4-9-87	0		0.5-0.5	Dans le rang	2 / 1
Orge de printemps / Lupin		Copenhague		24-29-47	0		0.5-0.5	Dans le rang	2 / 1
		Jyndevad		4-9-87	0		0.5-0.5	Dans le rang	2 / 1
Orge de printemps / Pois		Copenhague		24-29-47	0		0.5-0.5	Dans le rang	2 / 2
		Jyndevad		4-9-87	0		0.5-0.5	Dans le rang	2 / 2
		Taastrup	2003	24-29-47	0	additif-substitutif	0.5-0.5, 0.5-1	Rangs alternés	1 / 1
	France	Angers	2002	6-15-79	0	additif	0.33-1	Rangs alternés	1 / 1
			2003	6-15-79	0-130	additif-substitutif	0.5-0.5, 0.5-1	Rangs alternés	1 / 1
				21-40-39	0		0.5-0.5, 0.5-1	Rangs alternés	1 / 1
			2004	21-40-39	0		0.5-0.5, 0.5-1	Rangs alternés	1 / 1
	Allemagne	Kassel		51-29-20	0		0.5-0.5, 0.5-1	Rangs alternés	1 / 1
	Italie	SMargentano	2003	22-36-42	0	substitutif	0.5-0.5	Rangs alternés	1 / 1
			2004	22-36-42	0		0.5-0.5	Rangs alternés	1 / 1
	Royaume-Uni	Reading	2003	49-32-19	0	additif-substitutif	0.5-0.5, 0.5-1	Rangs alternés	1 / 1
Blé tendre d'hiver / Féverole		Rennes	2018	22-36-42	0	additif	0.7-0.75	Dans le rang	8 / 2

Blé tendre de printemps / Lentille	Auzeville	2015	10-8-82	0	additif-substitutif	0.17-1, 0.33-0.7, 0.33-1, 0.5-0.5, 0.5-1	Rangs alternés-Dans le rang	2 / 4
		2016	18-48-34	0		0.17-1, 0.17-1.3, 0.33-0.7, 0.33-1, 0.33-1.3, 0.5-0.5, 0.5-1	Dans le rang	2 / 4
Blé tendre d'hiver / Pois	Angers	2007	20-38-42	0-45	substitutif	0.5-0.5	Dans le rang	1 / 1
		2008	20-38-42	0-90		0.5-0.5	Dans le rang	1 / 1
		2006	21-40-39	0		0.3-0.7, 0.5-0.5	Dans le rang	1 / 1
		2007	21-40-39	0-30		0.5-0.5, 0.7-0.3	Dans le rang	1 / 1
		2008	21-40-39	0-72		0.5-0.5, 0.7-0.3	Dans le rang	1 / 1
		2009	21-40-39	0-40		0.5-0.5, 0.7-0.3	Dans le rang	1 / 1
	Grignon	2010	11-54-35	0-140	additif-substitutif	0.33-0.66, 0.5-0.5, 0.7-0.5	Dans le rang	1 / 1
		2017	11-54-35	0		0.05-1, 0.15-1, 0.5-0.5, 0.5-1	Dans le rang	1 / 2
	Rennes		19-49-32	0	additif	0.5-0.75, 0.5-1	Dans le rang	8 / 3
		2018	22-36-42	0		0.5-0.75, 0.5-1	Dans le rang	8 / 3
Blé dur d'hiver / Pois chiche		2010	18-48-34	0-140		0.67-1	Rangs alternés	1 / 1
			18-48-34	0-140	additif-substitutif	0.33-0.5, 0.5-0.5, 0.67-0.5, 0.67-1	Rangs alternés-Dans le rang	1 / 1
		2011	18-48-34	0		0.5-0.5	Rangs alternés	1 / 1
			18-48-34	0-140		0.5-0.5	Rangs alternés-Dans le rang	1 / 1

Blé dur d'hiver / Féverole	France	Auzeville	2013	18-48-34	0	substitutif	0.5-0.5	Dans le rang	3 / 4
			2012	10-8-82	0		0.5-0.5	Dans le rang	3 / 4
Blé dur d'hiver / Pois			2015	10-8-82	0	additif-substitutif	0.5-0.5, 0.5-1	Dans le rang	1 / 4
			2007	10-8-82	0-140	substitutif	0.5-0.5	Rangs alternés	4 / 1
			2006	18-48-34	0-180		0.5-0.5	Rangs alternés	1 / 1
			2013	18-48-34	0-140		0.5-0.5	Dans le rang	3 / 5
			2012	10-8-82	0		0.5-0.5	Dans le rang	3 / 4

Table II.1 – Caractéristiques des expérimentations du jeu de données

Au total, le jeu de données réunit 340 unités expérimentales en culture associée et 305 unités expérimentales en culture pure (Figure II.4). Le nombre d'observations phénotypiques (variables x dates de mesure) est logiquement plus élevé puisqu'on compte plus de 15000 observations, dont plus de 9000 en culture associée et plus de 6000 en culture pure.

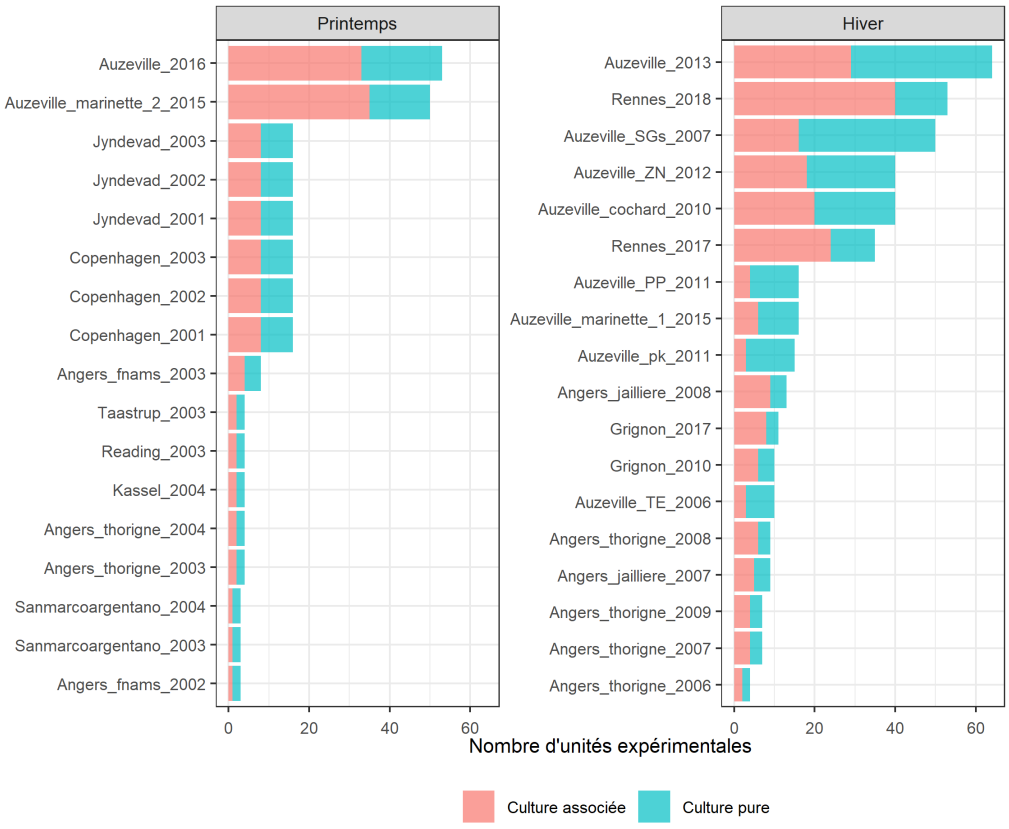


Figure II.4 – Nombre d'unités expérimentales par expérimentation et saison de culture

1.2.3 Mélanges d'espèces cultivés

Les expérimentations regroupent plusieurs espèces de chaque famille végétale :

- trois céréales : le blé dur (*Triticum turgidum* L.), le blé tendre (*Triticum aestivum* L.) et l'orge (*Hordeum vulgare* L.),
- cinq légumineuses : le pois (*Pisum sativum* L.), la féverole (*Vicia faba* L.), la lentille (*Lens culinaris* L.), le pois chiche (*Cicer arietinum* L.) et le lupin

(*Lupinus Angustifolius* L.).

L'orge et le pois sont les espèces présentes dans le plus grand nombre d'expérimentations (Figure II.5). Viennent ensuite le blé (dur et tendre) ainsi que la féverole. Les associations blé tendre / pois et orge / pois sont les mélanges les plus cultivés, respectivement dans 10 et 15 expérimentations. Cependant, les mélanges blé dur / féverole et blé dur / pois sont cultivés dans de nombreuses unités expérimentales, contrairement aux mélanges orge / féverole ou orge / lupin.

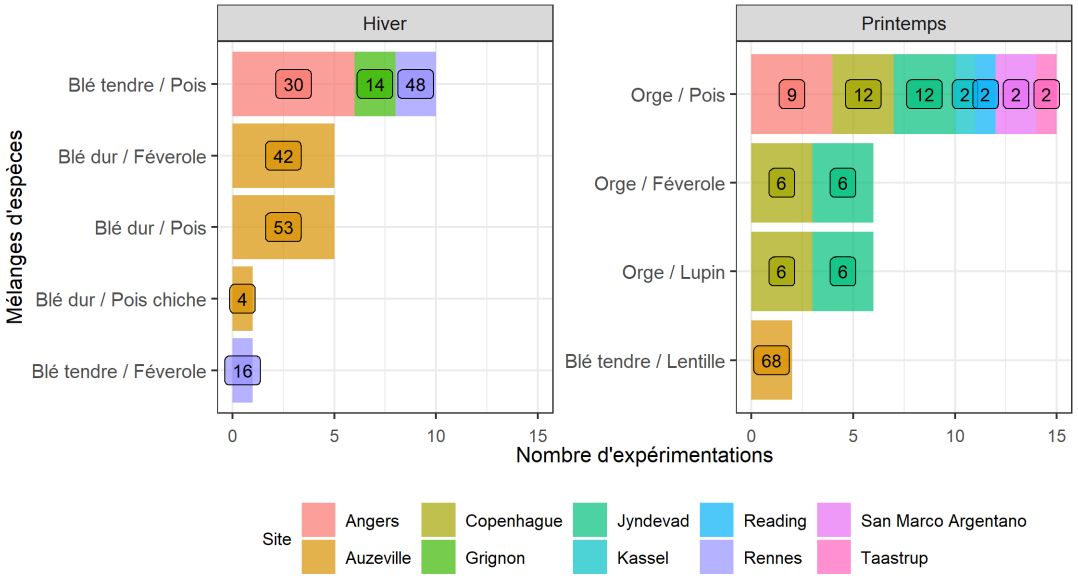


Figure II.5 – Nombre d'expérimentations incluant chaque mélange d'espèces.

1.2.4 Variables mesurées

Dans un objectif de diffusion et d'homogénéisation du jeu de données, les variables le constituant ont été traduites en anglais (Tableau II.2). Plusieurs variables sont mesurées au moins une fois dans différentes expérimentations. J'ai séparé ces variables en 3 grandes catégories, liées à i) des mesures de biomasse, ii) des teneurs (ou quantités) en azote, iii) l'interception de la lumière. Les variables mesurées dans la majorité des expérimentations sont le rendement, la biomasse des parties aériennes de la plante, la teneur en azote (ou azote absorbé) de la plante et la hauteur du couvert (Tableau II.2).

Table II.2 – Principales variables mesurées dans le jeu de données, et utilisées pour l'élaboration des modèles statistiques d'ajustement de la performance des cultures associées

Catégorie	Nom (EN)	Description	Unité	Nombre d'expérimentations
Azote	nitrogen_shoot	Teneur en azote dans les parties aériennes de la plante	$mg.g^{-1}$	31
	nitrogen_abs_shoot	Azote absorbé dans les parties aériennes des plantes	$kg.ha^{-1}$	30
Biomasse	biomass_shoot	Biomasse des parties aériennes	$t.ha^{-1}$	33
	biomass_seed	Biomasse des graines, correspond au rendement	$t.ha^{-1}$	35
Lumière	height	Hauteur totale des plantes	m	29
	lai	Leaf area index (indice de surface foliaire = surfaces photosynthétiques / surface au sol)	-	17
	sla_greenleaftendril	Surface photosynthétique (feuilles, vrilles) rapportée à la masse des parties photosynthétiques de la plante	$cm^2.g^{-1}$	9

La fréquence de mesure de chaque variable peut différer entre les expérimentations (Figure II.6). Dans certaines expérimentations, des variables sont mesurées plus de 10 fois pendant le cycle de culture (ex. hauteur) et parfois une seule fois (en pratique à la fin du cycle de culture). L'extraction d'informations potentiellement utiles de ces variables mesurées de manière dynamique est une étape clé dans le traitement de ces données.

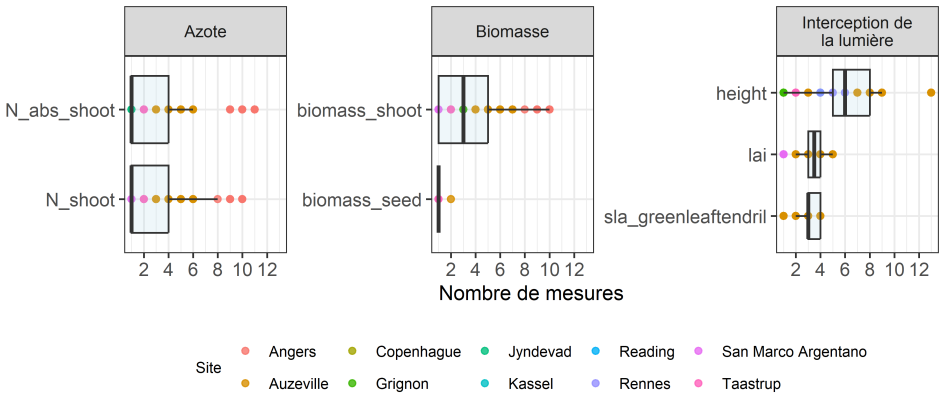


Figure II.6 – Nombre de mesures par variable pendant le cycle de culture pour chaque unité expérimentale

L'ESSENTIEL

Le jeu de données construit et utilisé dans ma thèse est un jeu de données global. Son volume (340 unités expérimentales en culture associée et 305 en culture pure) est à la fois faible par rapport aux standards en statistique mais élevé par rapport aux standards en agronomie. L'utilisation de ce type de structure de données est une pratique rare dans cette discipline.
Trois espèces de céréale et 5 espèces de légumineuse sont représentées dans 35 expérimentations. Plusieurs variables liées à différents processus écophysiologiques sont mesurées dans ces expérimentations, une seule fois ou à plusieurs moments pendant le cycle de culture. Les pratiques agronomiques testées (fertilisation azotée, variétés, densités relatives, etc.) ne sont pas toujours les mêmes entre les expérimentations, ce qui induit une forte hétérogénéité du jeu de données.
La description succincte du jeu de données étudié nous permet de tirer plusieurs conclusions :

1. Le jeu de données est plus fourni que la majorité des jeux de données étudiés en agronomie (grand nombre d'expérimentations pour la majorité des espèces et variétés),
2. Malgré cela, l'absence de plan factoriel complet implique des effets de confusion ainsi que l'impossibilité de certaines analyses,
3. Les questions de recherche pouvant être traitées grâce à un jeu de données de cette sorte sont nombreuses, mais il est nécessaire d'être capable d'identifier et de sélectionner des sous-jeux de données pertinents pour les traiter.

La stratégie de recherche impliquée par ces conclusions requiert une attention particulière sur la méthodologie. Ces enjeux méthodologiques et de positionnement font l'objet d'un article soumis à la revue *Computer and Electronics in Agriculture*.

2 A workflow for processing global datasets : application to intercropping

Article en révision dans la revue Computers and Electronics in Agriculture [1], également présenté à la 31ème International Biometric conference (IBC [2]) et au 17ème congrès de l'European Society of Agronomy (ESA [3]).

Rémi MAHMOUD, Nadine HILGERT, Pierre CASADEBAIG et Noémie GAUDIO

Résumé : Les expérimentations constituent une source conséquente de données et de connaissances en agronomie. Une pratique émergente consiste à compiler les mesures et résultats de ces expérimentations (plutôt que les résultats des publications comme dans les méta-analyses) au sein de jeux de données globaux. L'objectif de notre étude était de fournir plusieurs pistes méthodologiques liées à la conception et à l'analyse des jeux de données globaux. Pour illustrer ces réflexions et apports méthodologiques, nous avons considéré, comme cas d'étude, 35 expérimentations agronomiques pour concevoir un jeu de données global et illustrer comment l'organisation et la structuration des données est une première étape pour aller vers les pratiques de science ouverte dans une optique de diffusion des données. Nous avons développé une méthode pour identifier des sous-jeux de données au design factoriel au sein de ce jeu de données global en mobilisant des outils issus de la théorie des graphes. Une autre contribution méthodologique basée sur des splines de lissage a été appliquée pour exploiter les variables mesurées de manière dynamique. Nous discutons enfin le positionnement des jeux de données globaux dans le continuum entre données et connaissance, comparativement à d'autres approches telles que la méta-analyse. En regard de ce travail, nous pensons que les jeux de données globaux sont un outil approprié pour effectuer des analyses dans le cadre du Big Data. Nous préconisons une utilisation plus large de ce type de jeux de données dans la recherche agronomique.

1. https ://www.sciencedirect.com/journal/computers-and-electronics-in-agriculture
2. https ://www.ibc2022.org/home
3. https ://esa-congress-potsdam2022.de/

2.1 Introduction

Whether on farms or at experimental research stations, field experiments are the traditional and most popular way to gain knowledge in agricultural sciences (Maat, 2011). However, inferring general laws from the results of single and local experiments is not straightforward (Makowski et al., 2014). General laws about agroecosystem functioning are often determined by using meta-analysis (*i.e.* "statistical analysis of a large collection of analysis results from individual studies to integrate the findings" (Glass, 1976) to compare results of multiple scientific studies. In contrast, global datasets are defined as the aggregation of all available observations of multiple experiments.

They differ mainly in that global datasets contain raw experimental results at a fine scale (*e.g.* phenotypic, genomic), while meta-analysis focuses on published results, with few variables available. Potentially available fine-scale measurements are rarely compared between experiments. Except for a few examples (Kattge et al., 2011, Newman et Furbank, 2021, Pappagallo et al., 2021), agricultural datasets contain public statistics of global variables such as nitrogen (N) fertilization and yield (Agreste, 2022, FAOSTAT Statistical Database, 2021). A major strength of agricultural global datasets is that they collect multiple phenotypic observations from a variety of soils and climates, which can be used to generalize locally observed results (Tardieu, 2020), thus avoiding spurious correlations (Tardieu, 2020) and leveraging experimental data that were collected but not yet used in scientific publications (Zamir, 2013).

While global datasets contain more observations than standard experiments, they remain much smaller (*i.e.* hundreds to thousands of observations, megabytes to gigabytes of data) than certain big data applications (terabytes to zetabytes of data) (Kitchin et McArdle, 2016). This reasonable volume of data i) avoids problems of data storage and ii) uses technologies commonly shared by agronomists (*e.g.* R software (R Core Team, 2021)), even those not specialized in statistics or data science.

These advantages are counterbalanced by some weaknesses which are inherent to the building of global datasets. Indeed, gathering data from multiple experiments carried out by different research teams involves a long and tedious step of data collection, standardization and homogenization (Makowski et al., 2014, White et Van Evert, 2008). While no standard experimental files exist among research teams, this step cannot be automatized. Moreover, the different experiments are designed for multiple purposes (*e.g.* testing the impact of different agronomic practices or crop species), and gathering them often leads to incomplete and unbalanced designs. Confounding

factors can also occur, *i.e.* the inadvertent mixing of two or more effects, such that no statistical analysis can separate them (Casler, 2015). Thus, using and analyzing global datasets involve i) good knowledge of the dataset and some caution in interpreting the results, ii) identifying balanced subsets to answer specific questions, and iii) accepting that the effects of some factors cannot be distinguished. For these reasons, the application of big data technics (machine learning) on global datasets is only possible after several data pre-treatments.

Despite these many advantages, global datasets are rarely used to answer specific agricultural questions. To our knowledge, only a few studies have been based on this type of dataset (Licker et al., 2010, Lobell et al., 2020, Newman et Furbank, 2021), whereas many studies have focused on methods used in meta-analysis (Shelby et Vaske, 2008, Philibert et al., 2012, Gurevitch et al., 2018, Makowski et al., 2019), and no publication has focused on methods for global datasets.

Although meta-analysis has provided positive results (Yu et al., 2016, Raseduzzaman et Jensen, 2017, Knapp et van der Heijden, 2018), we argue that crop science would benefit from sharing, combining and jointly studying multiple experiments using global datasets (White et Van Evert, 2008, Zamir, 2013, Cruz et Nascimento, 2019). We thus promote using global datasets as an alternative to meta-analysis. To this end, we recommend using methods for analyzing and describing global datasets that complement the practices described by FAIR principles (*i.e.* Findable, Accessible, Interoperable and Reusable) (Wilkinson et al., 2016).

This is particularly relevant in the current agricultural context, in which crop diversification is a key tool for more sustainable agriculture (Duru et al., 2015b). This need for diversification implies performing more extensive experiments, which in turn requires stronger data-federation efforts. This potential for joint analysis helps clarify the context-dependence of analytical experiments and ultimately increases understanding of the relationship between crop diversity and agroecosystem functioning.

As a use case, we illustrate the design of a global dataset by focusing on intercropping systems, which are defined as the simultaneous growth of two or more species in the same field for a significant period of their growing cycle. Among diversification practices, intercropping has shown many promising results (Martin-Guay et al., 2018). Thus, understanding general laws about the functioning of intercrops is a way to include them more often in agricultural systems. We describe the main steps involved in designing a global dataset of 35 experiments. We also describe and apply an original method for identifying factorial designs, which is a key step for future analysis.

2.2 Development of a global dataset

2.2.1 Main design steps

We show the main steps involved in designing a global dataset, along with the key specifications and methods developed to support future analysis and modeling studies that use this global dataset (Figure II.7).

The first step involves standardizing the raw experiment files (available mainly in spreadsheet software) into a single format and describing the metadata (including all of the variables measured or collected, their definitions and units). The standardized files are then compiled and documented to make the data "analysis-friendly" (Wilson et al., 2017), which enables detection of errors and data exploration, validation and analysis. A good practice is to work with "tidy" data, in which each column is a variable and each row is an observation (Wickham, 2014). Detected errors are corrected in either the raw or standardized files, depending on where they originated.

Once the data are in a tractable format, additional processes are required to render the dataset operational for analytical and modeling studies. In our use case, we illustrated this step with two specific methodological developments. In the first, we identified subsets in the overall dataset to create a complete factorial design that indicates which analytical studies and research issues can be addressed based on the data. In the second, we developed a solution to derive indicators from heterogeneous variables that are measured at different time steps.

2.2.2 Illustration with the intercrop use case

We briefly describe the features of the available field experiments to highlight their richness and heterogeneity. See Gaudio et al., 2021a and Mahmoud et al., 2022 for full details and experimental protocols.

Although combining results from a few experiments (usually two years, often sequential) is common in the intercropping literature, no study included joint analysis of dozens of experiments to infer general laws about intercrop functioning. To this end, we designed, built and analyzed a global dataset that combined the results of 35 field experiments that involved cereal-legume intercrops and the corresponding sole crops. The aim of these field experiments was to compare the growth and grain yield (t/ha) of multiple combinations of species grown in intercrop to their sole-crop reference. The field experiments had been performed in five European countries : France, Denmark, Italy, Germany and England (Figure II.8) from 2001-2018. The

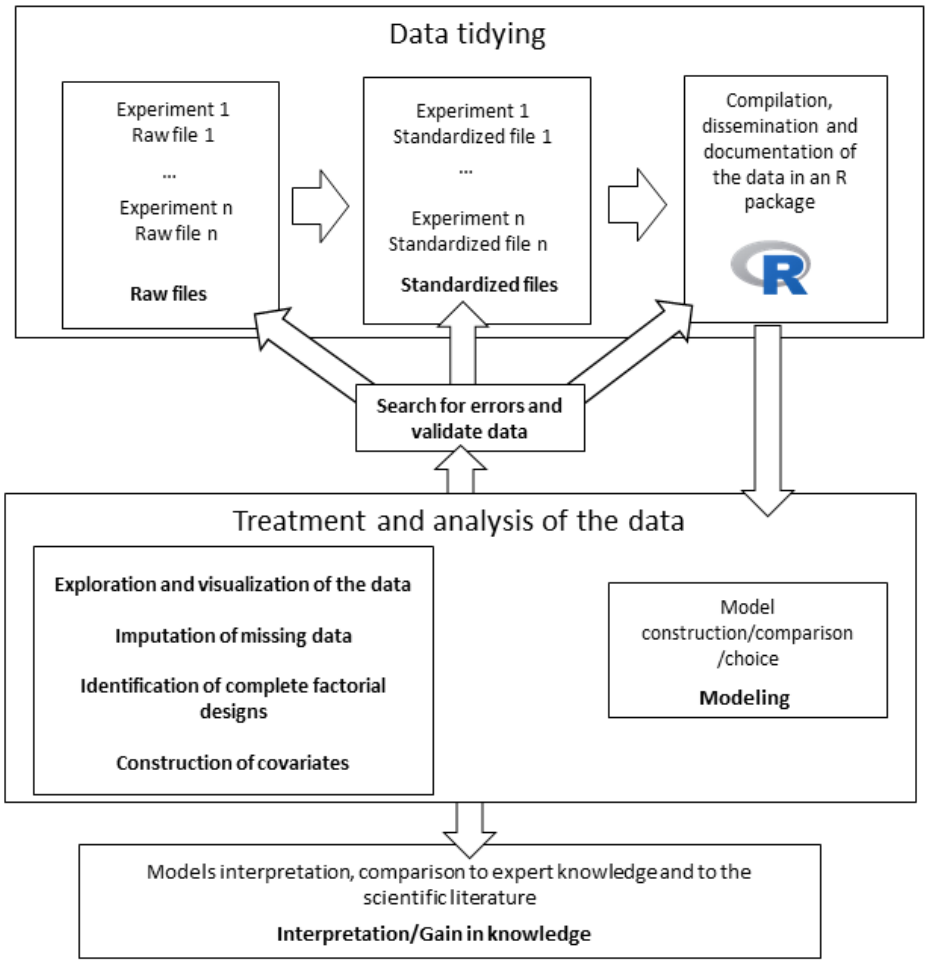


Figure II.7 – Steps of the workflow for processing global datasets

global dataset contained five legume species (*i.e.* chickpea, faba bean, lentil, lupine and pea) and three cereal species (*i.e.* barley, durum wheat and soft wheat).

We transformed raw experimental data tables into standardized rectangular data tables (Figure II.7). The variables and values were placed in tables as a function of the organizational level (*i.e.* individual, population) and the information they provided (*e.g.* plant functioning, climate, soil). One difficulty was that measured variables were related to particular organizational levels ; for instance, height was usually measured at the population level, while inter-node length was measured at the individual level. We defined one table per level. The same distinction was made for variables measured

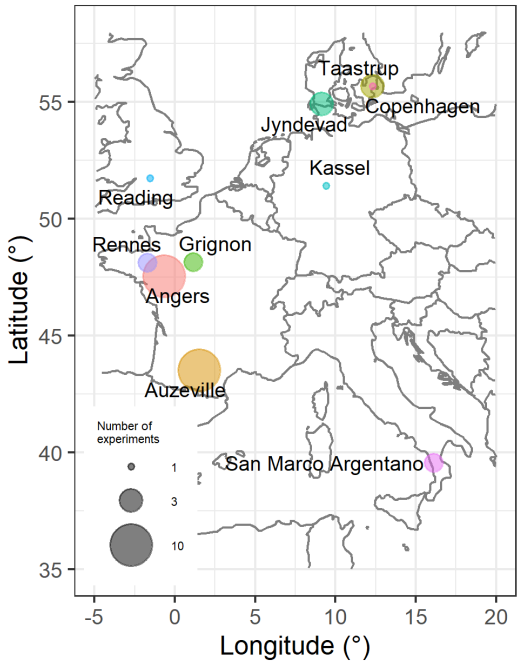


Figure II.8 – Location of the 35 intercrop experiments gathered within the global dataset

for each species vs. those for which the species were not distinguished. Species-specific measurements (*e.g.* height, shoot biomass) were placed in a "plot" table, with one entry per species in the intercrop, while whole-plot measurements (*e.g.* weed biomass, crop cover) that did not distinguish the intercropped species were placed in a "plot_global" table. Plant-specific measurements (*e.g.* inter-node length, number of tillers) were placed in a "plant" table. Information about the experimental site (*e.g.* country, latitude and longitude) and its soil and climate conditions were placed in "site", “soil” and "climate" tables, respectively. Agronomic practices (*e.g.* species sown, cultivar, targeted sowing density, sowing and harvest dates, N fertilization) were placed in a "practices" table. These agronomic practices were assigned a unique identifier for “management” (*e.g.* M1, M2). Depending on the experiment, each management practice was replicated 3-6 times, and each replicate corresponded to one statistical individual. This information was specified in an "index" table that contained the replicate number, x and y positions of the plot at the experimental site and the management identifier. In the dataset considered, one statistical individual was defined as the unique combination of experiment, species, density and spatial arrangement at sowing, and fertilization treatment. Overall, the global dataset contained 340 and 305 statistical individuals in intercrop and sole crop, respectively (Figure II.9).

The number of plant characteristics was much larger (ca. 15,000 observations), since several variables were measured at the plant or crop scale, sometimes several times during the crop cycle.

Figure II.9 – Diversity of the treatments in the global dataset by factor (columns) and experiment (rows). Within each column, each colored rectangle is a level of the factor considered. A rectangle in a given row and column indicates that the corresponding experiment contained at least one statistical individual with the corresponding factor level.

The brief description of the global dataset provides an overview of the diversity of agronomic situations considered (Figure II.9). While the experimental designs had

many similarities (*e.g.* species cultivated, agricultural practices), the latter were not sufficient for immediate analysis of the global dataset. We thus developed methods to identify subsets in the global dataset to form complete factorial designs a posteriori (subsection 2.2.4) and to derive indicators despite heterogeneity in observation times (subsection 2.2.5).

2.2.3 Dissiminating the data

The ability to disseminate data easily is a key feature in designing a dataset, since it determines how other researchers will be able to interact with the data locally or remotely. We specifically designed an R package to read and format the raw experimental files and to document and disseminate the dataset in the computing environment. The package was created to apply FAIR principles as much as possible (Wilkinson et al., 2016), which are defined as follows :

- **F**indable : clear documentation (metadata) makes the data findable,
- **A**ccessible : the procedure for formatting (meta)data is accessible, even if the data are private,
- **I**nteroperable : the (meta)data use a shared vocabulary, and relations among (meta)data are specified.
- **R**eusable : (meta)data are published with a license, and sources of the data are specified

Currently, the global dataset is used privately within a few projects and meets the FAIR principles only partially (Table II.3). We identified several barriers to broader dissemination, including the reticence of some research teams to share the data publicly. Part of the dataset is published (Gaudio et al., 2021b) and may be updated as soon as missing experimental data are released under an open license.

2.2.4 Support for analytic studies

2.2.4.1 Data subsetting to form complete factorial experimental designs a posteriori

Because a global dataset aggregates several datasets, each of which comes from an experiment with its own experimental design (mainly split-plot), it is not supported by an experimental design that can help analyze it.

It is essential to have data derived from factorial experimental designs to be able

Practice	FAIR principle
Clear metadata that describe the variables (*e.g.* unit, description)	Interoperable
R code on GitHub that enables reading and formatting the data	Accessible
Data are documented and disseminated in an R package	Findable
License given in the R package documentation ; data accessible upon reasonable request	Reusable

Table II.3 – FAIR principles applied in the reading and formatting procedure of the global dataset. F : Findable, A : Accessible, I : Interoperable, R : Reusable

to test for statistical effects. Thus, given a set of categorical variables, one objective was to identify one or more data subsets that form a complete factorial design, along with some of the factor levels. This approach can quickly assess whether the dataset is suited to answer a set of scientific questions, as long as the factors of interest are sufficiently represented in the global dataset.

We used tools from graph theory (Phillips et al., 2019) to identify the largest subsets of data associated with complete factorial designs in the global dataset. In graph theory, graph G is a pair $G = (V, E)$ where V is a set of vertices, and E is a set of edges that connect some of the vertices (Table II.4).

Given a set of categorical variables $X_1, ..., X_k$, each having values in a discrete set (*i.e.* $\forall i = 1, ..., k$ $X_i \in \mathcal{A}_i := \{x_{i,1}, ..., x_{i,j_i}\}$, ($j_i \in \mathbb{N}^*$ denoting the number of levels of variable X_i)), a k-partite graph can be derived by setting $V = \bigcup_{i=1}^{k} \mathcal{A}_i$, (*i.e.* each level of each factor is a vertex), and $E = \{(x, y)|$ levels x and y observed together$\}$.

A factorial design is complete if and only if all possible combinations of the factor levels are present. For graph $G = (V, E)$, this is equivalent to identifying a subgraph with an edge between each pair of vertices from independent sets (*i.e.* a k-clique). Thus, the challenge of identifying the largest complete factorial designs within a global dataset can be reduced to counting the number of maximal k-cliques in the graph.

Term	**Definition**
subgraph $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ of a graph $G = (V, E)$	A graph whose vertex set ($\widetilde{V}$) is included in the vertex set of G (i.e $\widetilde{V} \subseteq V$) and whose edge set ($\widetilde{E}$) is included in the edge set of G (i.e $\widetilde{E} \subseteq E$)
complete graph	A graph whose vertices are all connected
clique of a graph G	A complete subgraph of G
maximal clique of a graph G	A clique that cannot be extended by including one more adjacent vertex
k-partite graph	A graph that can be partitioned into k nonempty, vertex-disjoint, edgeless subgraphs
k-partite clique or *k-clique*	A set of vertices that induces a complete k-partite subgraph
maximal k-partite clique	A k-clique that cannot be extended by including one more adjacent vertex

Table II.4 – Some definitions in graph theory (Phillips et al., 2019)

Phillips et al., 2019 developed the Maximum Multipartite Clique Enumeration (MMCE) algorithm to count the number of maximal multipartite cliques in a k-partite graph. In brief, MMCE starts from the observation that if G is k-partite, and if another graph G' is built from G by adding all intrapartite edges, then C is a maximal k-partite clique in G if C is a maximal clique in G' with at least one vertex in each partite set. Thus, the initial question is a matter of a modified problem of maximal clique enumeration, which is a NP-hard problem (Lawler et al., 1980). To address this issue, MMCE uses a graph inflation approach, by adding all possible intrapartite edges to G. It then identifies maximal cliques in the inflated graph using a procedure of Bron et Kerbosch, 1973 and checks whether the cliques identified cover all of the partite sets. We coded MMCE in the R programming language (https://github.com/RemiMahmoud/kclique). Although the problem of identifying maximal k-partite cliques with the maximum number of vertices has also been shown to be NP-hard for any $k \geq 3$ (Phillips et al., 2019), the relatively few vertices ($|V| < 300$) in the global dataset allowed solutions to be found quickly.

2.2.4.2 Application

We used this method to address specific issues in a previous study (Mahmoud et al., 2022). In it, we assessed the intensity of positive plant-plant interactions in unfertilized intercrops and analyzed how N fertilization influenced these interactions. To this end, we used the data subset that contained all of the unfertilized experimental individuals (263 in the 35 experiments). We then looked for experiments that included both N-fertilized and unfertilized treatments by looking for a maximal 2-clique in a graph composed of two sets of vertices : i) field experiments and ii) potential N fertilization (*i.e.* unfertilized and N fertilized levels). The targeted maximal 2-clique needed to contain the two levels of the sets of N-fertilization vertices (Figure II.10). This procedure yielded a data subset of 82 statistical individuals in 11 experiments.

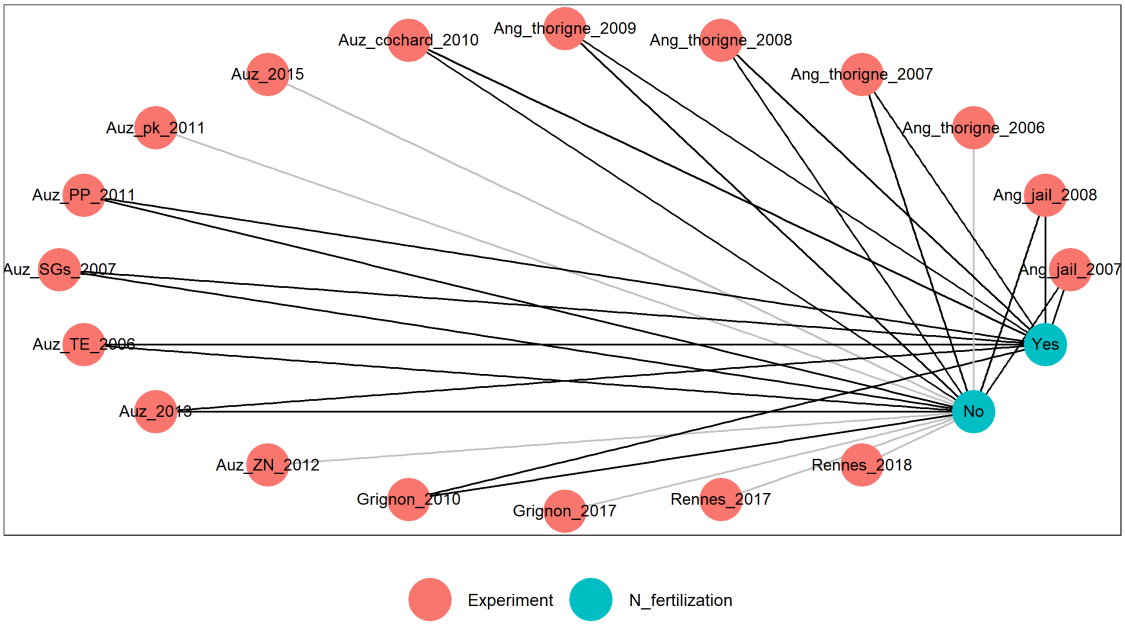


Figure II.10 – Subset of experiments including nitrogen (N)-fertilized and unfertilized treatments studied by Mahmoud et al., 2022, identified using the method developed. Black edges represent the edges belonging to the 2-clique.

Adding additional factors increased the complexity. For instance, focusing on 4 factors (*i.e.* field experiments, N fertilization level, pea cultivars and sowing spatial arrangement), the graph structure became more complex because it contained 115 edges and 33 vertices (Figure II.11). The MMCE algorithm identified 14 maximal 4-cliques, of which we selected two illustrative examples (Figures II.11). The first and

largest 4-clique contained nine experiments, but only one level of each remaining factor (Figure II.11a). Because the only pea cultivar common to several field experiments was *lucy*, it was suitable for studying phenotypic plasticity, since it was grown in nine different environments. The second 4-clique contrasted with the first since it contained only one experiment (Auzeville_2013) but five pea cultivars, two N fertilization levels and one level of the spatial arrangement at sowing (Figure II.11b). Thus, the experiment conducted in Auzeville in 2013 was suitable for studying interspecific diversity of the pea since it contained several cultivars. Identifying complete factorial designs highlights the global dataset's richness and diversity.

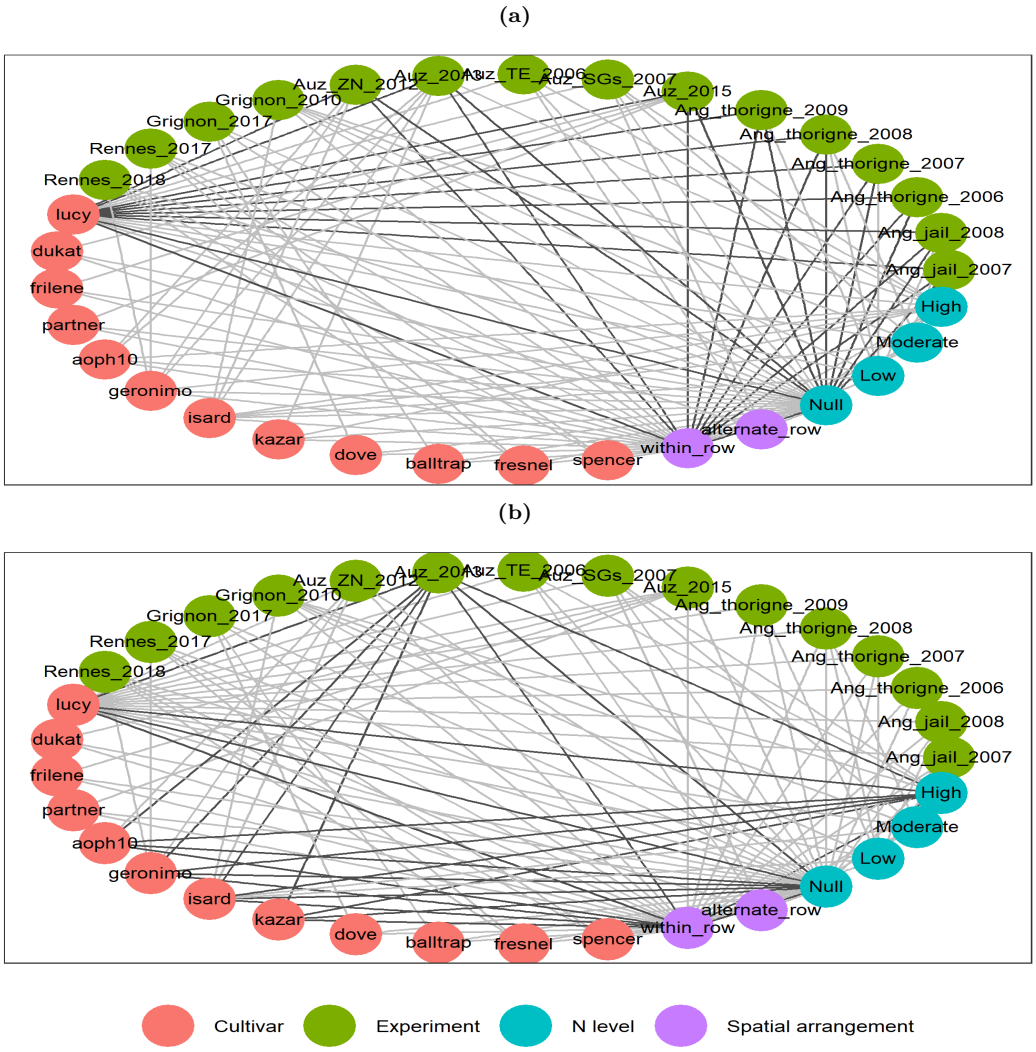


Figure II.11 – Two maximal 4-cliques that represent complete factorial designs with four factors. Black edges represent the edges of the 4-cliques. (a) A use case suited to studying phenotypic plasticity of the pea cultivar *lucy* among nine field experiments, and (b) a use case suited to studying intraspecific diversity of the pea (five cultivars) in one experiment, under unfertilized and highly nitrogen (N)-fertilized conditions.

2.2.5 Support for modeling studies

2.2.5.1 Addressing with different observation times

When experiments with a variety of protocols are combined, temporal data such as plant growth dynamics are observed at different time steps or at different points in time. A traditional way to compensate for differences in the timing of observations is to smooth the curves created by the measurement points of each statistical individual. Most plant growth dynamics follow an "S" shape, which can be fitted by a logistic growth curve or non-linear regression. Many parametric adjustments exist that can describe and fit logistic growth curves (Perperoglou et al., 2019). Variability in sampling times among experiments means that observations sometimes preclude extraction of key features (*i.e.* non-linear regression models do not converge). In this case, non-parametric smoothing (*e.g.* non-linear smoothing splines or kernels) can be applied to reduce the dimensions of the curves to allow key features of the growth cycle to be extracted. For the intercrop use case, we reduced the dimensions of growth curves by applying smoothing splines. We focused on two informative features : i) slope at the inflection point (μ : maximum growth rate) and ii) starting time of the growing phase (λ) (Figure II.12a). Although non-linear regression can be applied when data are sampled at the same sufficiently long time step, a non-parametric method was applied due to the variability in the data and in sampling times.

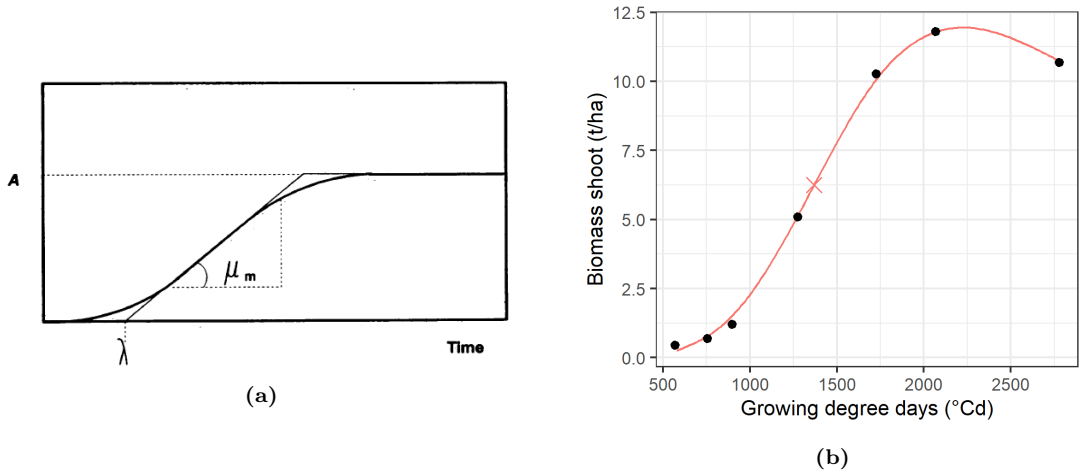


Figure II.12 – II.12a Theoretical growth curve (source : Zwietering et al., 1990), showing the inflection point (μ_m), starting time of growth (λ) and maximal value (A), and (b) observed growth curve of shoot biomass as a function of growing degree days fitted with a smoothing spline. The red cross represents the inflection point.

2.2.5.2 Application

We applied smoothing splines to plant height and shoot biomass, for which the number of observations during the growth cycle ranged from 1-13 per experiment (Figure II.13).

We used smoothing splines to perform the multiple adjustments (*e.g.* 725 adjustments for height and 375 for shoot biomass) since they are often used in biostatistics (Perperoglou et al., 2019), due to their ability to fit data that describe non-linear phenomena. Thus, by combining suitable methods and knowledge of the biological processes, it was possible to reduce the dimension of variables measured several times during a growth cycle. The parameters extracted from these curves can be used to predict yield or in other analyses (Engbersen et al., 2021). Smoothing splines were suitable for these data, but other processes or measurements could involve other methods. Thus, data can be homogenized as long as sufficient knowledge exists about the processes involved and which methods to use.

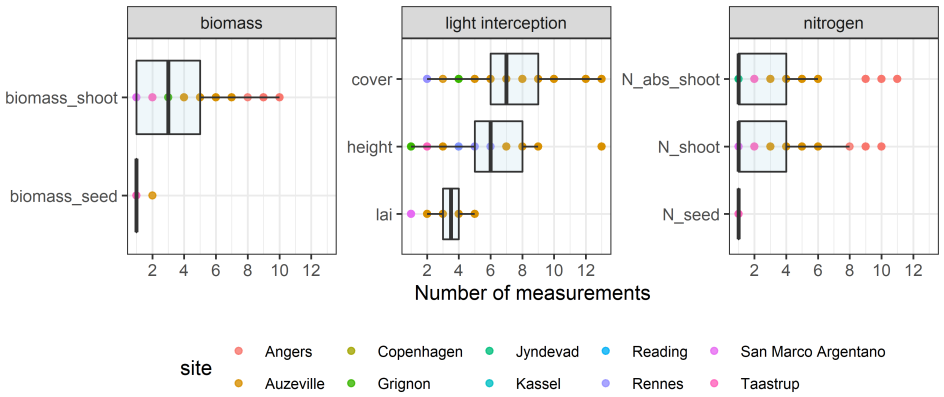


Figure II.13 – Number of measurements for each variable (related to plant biomass, light interception and nitrogen (N)) and each statistical individual. Whiskers represent 1.5 times the interquartile range.

2.3 Discussion

One key reason to use agricultural data is to improve knowledge in crop science, as in other scientific fields, which can be generalized with the Data, Information, Knowledge and Wisdom pyramid (Ackoff, 1989), which describes the continuum between data and the knowledge it provides. Thus, the issue is to use appropriate methods based

on the available data to provide insights and understanding of a studied system's functioning.

Depending on whether data come from experimental data or from scientific publications, methods related to global datasets or meta-analysis, respectively, will be used, and both are useful for studying global issues in agronomy (Table II.5) (Makowski et al., 2014). Two important issues arise from this observation : data availability and the knowledge that one wants to provide.

In meta-analysis, data are available because they are already published, even if it takes a long time to retrieve them. Conducting a meta-analysis is thus time-consuming, especially the pre-analysis search and development of the database, which represent ca. 60% of the working time (Allen et Olkin, 1999). However, combining and formatting data for global datasets likely takes more time. Meta-analysis requires identifying and extracting the values of interest from scientific publications, while being cautious to avoid potential bias. In contrast, building global datasets requires interacting with the research teams that conducted the experiments and adapting their raw experimental files to a standard format (Figure II.7). Moreover, some of the experimental data recovered have not been published. This is an advantage of global datasets, which value the time and energy required to conduct field experiments, but is also a disadvantage, since researchers may be reluctant to share unpublished data. For instance, in our use case, 7 of the 35 experiments (*i.e.* 20%) had not been published before we built the global dataset, while each is now described in 1-3 scientific publications (Gaudio et al., 2021a, Meunier et al., 2022a, Mahmoud et al., 2022). One advantage of global datasets is that they can be expanded by adding new experiments, such as by taking advantage of large research projects with an experimental component (Pappagallo et al., 2021) or using the recent development of high-throughput phenotyping platforms (Yang et al., 2020). In this way, global datasets could fall within the big data framework (Ylijoki et Porras, 2016), thus facilitating the use of big data methods (*e.g.* machine learning) as long as they are sufficiently developed.

The knowledge that is generated depends on the data used to conduct meta-analyses or the approaches used with global datasets. Global datasets are better suited for explaining the influence of individual factors (*e.g.* agronomic practices, climate, plant physiology) on a studied variable, while meta-analyses are better suited to assess the relative importance of these factors as broad trends. The main advantage of global datasets is that they consist of phenotypic observations, which allows for further investigation of potential causalities based on correlations in the data (Gunawardena,

2014). Phenotypic observations help to understand and explain observed correlations (Garside et Bell, 2011). Moreover, since agronomic global datasets contain plant-related variables measured at multiple organizational levels (*e.g.* organ, plant, crop), they may have more potential for reuse than meta-analyses by using data subsets that form factorial designs a posteriori. For instance, researchers who develop functional–structural plant models (Louarn et al., 2020) may be interested in variables measured at the plant scale (*e.g.* number of tillers, inter-node length, plant height), while those who develop crop models to predict yields (H. N. Berghuijs et al., 2021) may be interested in variables measured at the crop scale (*e.g.* crop biomass, crop height). For instance, using the global dataset developed in this study, Gaudio et al., 2021a extracted a subset of 28 experiments (378 statistical individuals) to assess the influence of intercropping on the relation between plant biomass and grain yield, Mahmoud et al., 2022 extracted a subset of 11 experiments (82 statistical individuals) to assess the influence of N fertilization on plant-plant interactions in intercrops, and Meunier et al., 2022a extracted a subset of 31 experiments to calibrate a statistical model used in a modeling chain to predict ecosystem services as a function of the species in cereal-legume intercrops.

The many agronomic experiments that have been conducted may be a huge source of data (Zamir, 2013), and compiling and sharing them is one way to make them usable and broaden the knowledge they can provide. Calculating informative features (subsection 2.2.5) connects the data to knowledge since it shows how to obtain meaningful biological information from raw measurements. However, this must be done carefully, and we argue that global datasets, by contributing to an "educated" big data framework, are suitable for inferring knowledge about systems of interest.

Some authors have argued that the "data deluge makes the scientific method obsolete" (Anderson, 2008), indicating that the large amount of data available in the era of big data decreases recourse to the scientific method (*i.e.* hypothesizing, modeling, testing), since the simple presence of correlations in large datasets can answer scientific questions without needing to develop hypotheses beforehand. Other perspectives assert that the role of science is to explain and interpret the relationships highlighted by studying massive datasets (Pigliucci, 2009) ; or that collecting data is guided by theory, without denying the contribution of data-science approaches (Mazzocchi, 2015) ; or that a correlation can highlight a relationship between certain variables but cannot indicate causality. In this context, modeling can help infer general laws about the functioning of agricultural systems, without asserting that deep understanding is possible simply by examining correlations. Two modeling approaches can be

Criterion	**Meta-analysis**	**Global datasets**
Scope	All practices studied in multiple scientific publications	All practices tested in multiple experiments
Time required to collect and tidy the data	Long to very long (dozen to hundreds of hours)	Very long
Variables used	Often standard variables (*e.g.* yield, nitrogen fertilization)	All available observations (*e.g.* agronomic practices, phenotypic measurements, pests, climate)
Number of observations	Moderate to large (dozens to hundreds)	Large (hundreds to thousands)
Reuse	Possible, but limited to the present variables	Possible once the data are formatted
Data sources	Scientific publications	Experimental files

Table II.5 – Comparison between meta-analysis and global datasets

distinguished (Gunawardena, 2014, Tardieu, 2020) :

- *reverse* modeling (*i.e* "top-down" approaches), which starts from experimental data and identifies potential causalities suggested by correlations observed in the data,
- *forward* modeling, (*i.e* "bottom-up" approaches), which starts from known or suspected causalities to make predictions.

Modeling using global datasets belongs to the reverse-modeling approach since it starts from the experimental data ; however, we argue that the main advantage of global datasets is their many phenotypic measurements, which teams with expertise in given research domains can use to calculate features that indicate broad (agro)system functioning. In the present study, starting from process-based knowledge (*i.e.* the relation between growth rates and intercrop yields), we used data-science methods to identify these indicators in a complex dataset.

The concept of "educated" big data (Tardieu, 2020) is a suitable compromise between reverse and forward modeling. It describes the process of calculating and choosing indicators that have a physiological meaning before performing the regression analysis, instead of performing it directly on tens or hundreds of correlated variables. Global datasets are a suitable tool for analyzing "educated" big data since they help to identify correlations between multiple phenotypic measurements and derived indicators, and can combine them with information about soil and climate conditions. Finally, using data to increase knowledge is not straightforward and requires caution. We argue that analyzing global datasets is an additional tool for doing so. Because it is rarely used in scientific literature, using it more often could help increase scientific knowledge in agriculture.

2.4 Conclusion

We argue that crop science can benefit from global datasets because they decrease the cost of data (reuse) and increase the reproducibility of studies. Ultimately, global datasets contribute to new findings through joint analysis of multiple experiments.

Although science has open-data tools to support the development of global datasets, the initial steps for sharing raw data between experimental and modeling researchers remain lacking.

Global datasets can be a keystone for "educated" big data in agricultural research since they combine many phenotypic observations and multiple variables. Thus, even

though their design and analysis are time-consuming and require caution, the potential insights that they can provide are worthwhile.

Acknowledgement

This study was supported by the French National Research Agency under the Investments for the Future Program (ANR-16-CONV-0004 and ANR-20-PCPA-0006). We thank Michael and Michelle Corson for their helpful comments and English revision.

L'ESSENTIEL

Les jeux de données globaux gagneraient à être plus utilisés en agronomie. Cette pratique est amenée à croître car elle présente des avantages que n'ont pas les méta-analyses (possibilité d'étudier des relations plus finement, réutilisations larges, etc.) malgré des inconvénients dont il faut tenir compte (temps de collecte). Une utilisation plus répandue de ce type de jeu de données implique un besoin en méthodologie. La méthode proposée pour visualiser les combinaisons de modalités disponibles entre plusieurs facteurs et identifier les plans factoriels complets en fait partie, de même que la procédure de réduction de dimension. Par ailleurs, ce type de jeu de données peut aider à la compréhension d'agroécosytèmes complexes, puisque la présence de nombreuses observations phénotypiques permet l'exploration de relations entre leur fonctionnement et leurs phénotypes. Ainsi, nous analysons dans le chapitre suivant certains des avantages agronomiques des cultures associées céréale-légumineuse en fonction de la fertilisation et de l'espèce choisie et dans le Chapitre IV, nous analysons certains déterminants de la performance de ces cultures, en se basant sur des observations phénotypiques.

Chapitre III

Analyse des avantages agronomiques des cultures associées

1 Avantages agronomiques des cultures associées céréale-légumineuse

En France, la part des cultures associées dans la surface agricole utile (SAU) reste faible, bien qu'en légère augmentation (Figure III.1). Parmi les cultures associées, les cultures associées céréale-légumineuse sont un type prometteur et souvent pratiqué (Verret et al., 2020). Dans cette partie, nous revenons sur le fonctionnement des cultures associées et sur certains des avantages agronomiques recherchés par cette pratique. Ces aspects conditionneront en partie nos choix de modélisation.

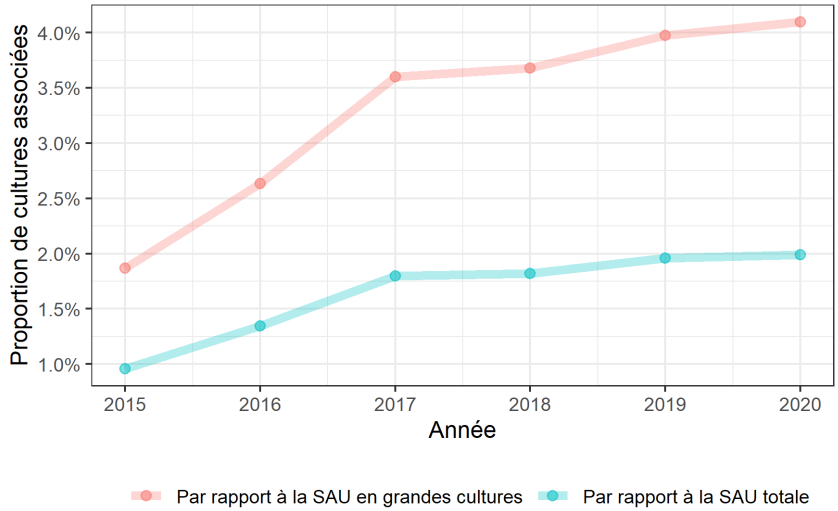


Figure III.1 – Proportion de cultures associées en France depuis 2015, données issues du registre parcellaire graphique (données 2022, extraction réalisée par Élodie Yan (INRAE, UMR SADAPT))

1.1 Fonctionnement des cultures associées céréale-légumineuse

Les deux mécanismes d'interaction plante-plante positifs mis en avant dans les cultures associées céréale-légumineuse sont la complémentarité de niche et la facilitation (Figure III.2). Ces mécanismes peuvent être dus à la nature même des génotypes présents dans le mélange (expression génétique) mais également refléter le résultat de la plasticité phénotypique, i.e. la capacité d'un génotype à s'exprimer différemment selon les conditions environnementales. Je me focaliserai ici sur les principaux processus ayant lieu dans les cultures associées céréale-légumineuse synchrones (espèces semées et

récoltées en même temps), que j'étudie plus particulièrement dans mon travail de thèse.

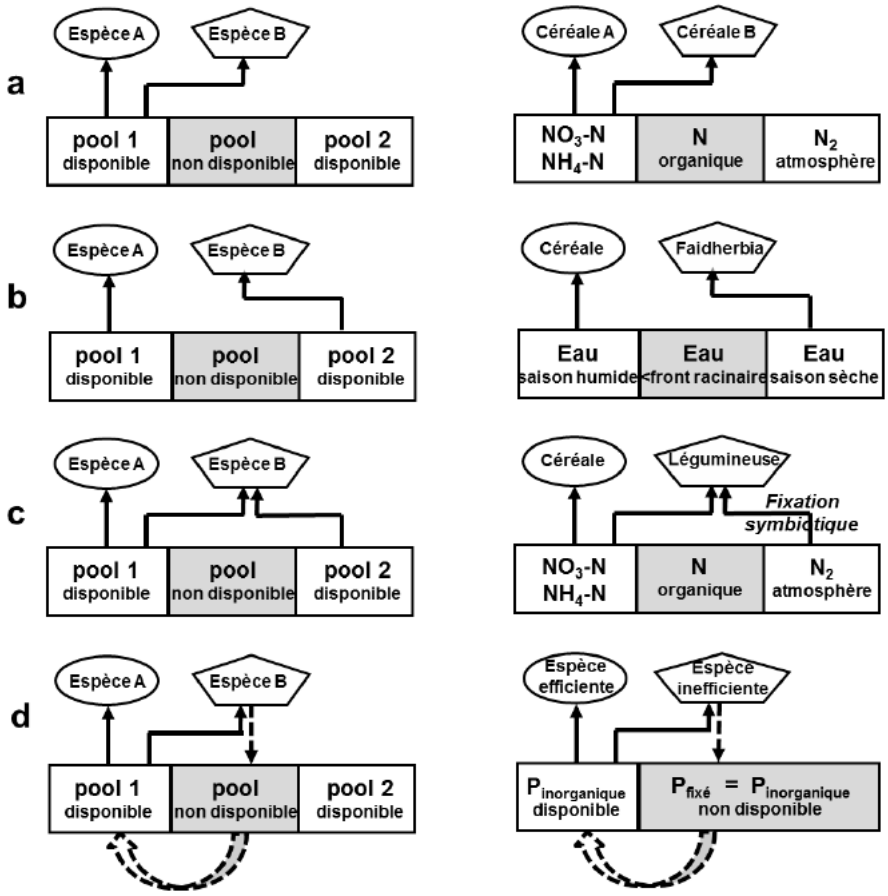


Figure III.2 – Différents phénomènes d'interactions entre plantes au sein de cultures associées (Justes et al., 2014) ; colonne de gauche : principe du phénomène ; colonne de droite : illustration avec un exemple. a) Compétition : les deux espèces utilisent la même ressource, sous la même forme et en même temps ; b) complémentarité de niche : les deux espèces exploitent des ressources différentes ; c) compétition et complémentarité de niche ; d) facilitation : la croissance d'une espèce est favorisée par la présence de l'autre *via* une modification de l'environnement de croissance

Mécanismes de complémentarité

La complémentarité de niche entre deux espèces (Figure III.2b) se définit par une exploitation différente des ressources, se traduisant par une réduction de la compétition entre espèces (compétition inter-spécifique) (Garnier et Navas, 2012). L'utilisation

différentielle d'une ressource donnée peut se faire de différentes manières (temporelle, spatiale, forme de la ressource), les deux principales dans nos agroécosystèmes d'intérêt étant l'utilisation d'une même ressource sous des formes différentes et une différenciation spatiale dans l'acquisition d'une ressource.

Les légumineuses sont la seule famille végétale ayant la capacité de fixer l'azote atmosphérique (N_2) via une symbiose racinaire avec un champignon (Voisin et al., 2003 ; Figure III.3). Cependant, la fixation symbiotique est plus coûteuse en énergie que l'utilisation directe de l'azote minéral du sol (NH_4 et NO_3) (Voisin et Gastal, 2015). Ainsi, en présence d'une céréale très compétitrice pour les ressources azotées souterraines, la légumineuse doit davantage faire appel à la fixation symbiotique. Une complémentarité pour l'utilisation de l'azote a donc lieu entre la céréale et la légumineuse (Figure III.2c), via l'utilisation d'une même ressource sous des formes différentes (N_2 vs NH_4 et NO_3).

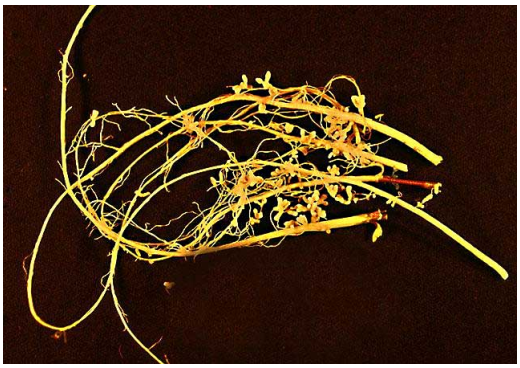


Figure III.3 – Nodosités sur racine de soja (tirée de Wikipédia [1])

Dans les cultures associées céréale-légumineuse synchrones, ce processus est essentiel et largement mis en avant dans la littérature existante comme étant un processus clé de la performance des cultures associées (Hauggaard-Nielsen et al., 2009, Bedoussac et al., 2015).

Des phénomènes de complémentarité spatiale pour l'utilisation des ressources peuvent aussi avoir lieu dans les cultures associées céréale-légumineuse, via l'utilisation d'une même ressource en des endroits différents. Par exemple, une différentiation spatiale de l'utilisation des ressources souterraines (eau et éléments minéraux) peut avoir lieu quand la profondeur et l'agencement des systèmes racinaires permettent aux

1. https ://fr.wikipedia.org/wiki/Nodosité

deux espèces associées d'accéder aux ressources à des profondeurs de sol différentes (Corre-Hellou et al., 2007 pour le pois et l'orge).

Mécanismes de facilitation

La facilitation a lieu quand une espèce est positivement impactée par la présence d'une espèce voisine, via une modification de l'environnement de croissance (Figure III.2d). Par exemple, une facilitation mécanique peut avoir lieu dans les cultures associées via un effet tuteur de la céréale sur la légumineuse. Ceci a été mis en évidence dans des mélanges orge/pois (Podgórska-Lesiak et Sobkowicz, 2013) et blé/lentille (Figure III.4 ; Viguier et al., 2018), permettant une réduction de la verse de la légumineuse, ce qui facilite sa récolte et augmente la quantité de grain récoltée.

Un autre processus de facilitation à l'oeuvre dans les cultures associées céréale-légumineuse est la mobilisation de phosphore insoluble via une acidification du milieu augmentant la disponibilité du phosphore disponible pour l'espèce compagne (Figure III.2d ; Homulle et al., 2021, Tang et al., 2021).

Figure III.4 – Lentille en culture (a) pure et (b) associée (source : Viguier et al., 2018)

1.2 Avantages agronomiques recherchés

Le rendement et ses composantes

Pour les raisons évoquées ci-dessus, l'utilisation des légumineuses en mélange permet de diminuer la dose d'intrants et notamment de fertilisants azotés, tout en maintenant un rendement satisfaisant. Les cultures associées céréale-légumineuse sont donc essentiellement cultivées en conditions bas-intrants (peu ou pas de fertilisation, peu ou pas de contrôle biotique), en particulier pour valoriser des parcelles agricoles assez

pauvres, et en agriculture biologique. Dans ces conditions, les rendements des cultures associées céréale-légumineuse sont généralement plus élevés que ceux des cultures pures correspondantes. Dans une méta-analyse sur les cultures associées céréale-légumineuse basée sur l'analyse de 77 articles scientifiques, les auteurs rapportent qu'il faut 16% (valeur médiane) de surface supplémentaire en culture pure pour obtenir le même rendement qu'en culture associée dans les mêmes conditions de culture (Yu et al., 2016). Cette valeur a été obtenue en contexte de faible fertilisation (cultures majoritairement fertilisées à moins de $< 100\ kgN.ha^{-1}$).

En plus du gain de rendement en contexte bas-intrants, la qualité du grain de la céréale, définie par la teneur en azote (N), est souvent plus élevée en culture associée qu'en culture pure (voir Bedoussac et Justes, 2010a, Timaeus et al., 2022 ; mais teneur en N similaire dans Naudin et al., 2010). Ceci peut s'expliquer par la disponibilité en azote minéral du sol plus élevée pour les céréales dans les mélanges du fait de la fixation symbiotique de la légumineuse (Gooding et al., 2007). Quand il a lieu, le gain constaté est de l'ordre de 2% de teneur en azote dans le grain, gain significatif permettant au blé de prendre de la valeur en passant en qualité supérieure (fourrage $\rightarrow$ blé panifiable $\rightarrow$ blé de qualité supérieure ; Timaeus et al., 2022).

Au-delà du gain de rendement en lui-même, la stabilité inter-annuelle du rendement est un des objectifs recherchés par certains agriculteurs cultivant des cultures associées. L'objectif recherché est un effet tampon/compensation (Justes et al., 2021), à savoir que si une espèce produit moins une année, l'autre espèce compensera. En pratique, les cultures associées céréale-légumineuse semblent effectivement avoir un rendement plus stable que les cultures pures correspondantes (Raseduzzaman et Jensen, 2017), même si cet effet dépend des conditions pédo-climatiques considérées (Weih et al., 2021).

Réduction des stress biotiques

Ces gains ou stabilité de rendement s'expliquent aussi par un effet de l'association sur les stress biotiques (maladies, ravageurs et adventices), qui sont des facteurs majeurs de diminution du rendement des cultures (Savary et al., 2019).

Plusieurs travaux rendent compte d'une réduction des adventices (mauvaises herbes) dans les cultures associées (Stomph et al., 2020). Une méta-analyse a estimé une forte réduction (58%) de la pression adventice dans les cultures associées par rapport à la culture pure la moins résistante et un effet nul par rapport à la culture pure la plus résistante aux adventices (Gu et al., 2021). Les auteurs attribuent cet effet à la densité du couvert de la culture associée supérieure à celle d'une culture pure ainsi qu'à la couverture précoce du sol.

La majorité des travaux montrent également une réduction de l'incidence et de la sévérité des maladies en culture associée (Boudreau, 2013, Stomph et al., 2020). Les causes menant à ces effets sont multiples : effet de dilution (la substitution partielle d'une espèce hôte d'un pathogène par une espèce non-hôte réduit les ressources disponibles pour le pathogène donné), effet barrière (la présence d'une espèce non-hôte réduit la propagation du pathogène), altération du micro-climat.

Dans mon travail de thèse, je ne prendrai pas en compte les stress biotiques car ces derniers ont été contrôlés par pesticides dans la majorité des expérimentations du jeu de données.

1.3 Une pratique plutôt rare

Malgré leurs avantages, les cultures associées restent peu cultivées en France et en Europe (Machado, 2009, Meynard et al., 2018 ; Figure III.1). Plusieurs raisons peuvent expliquer ce décalage apparent entre avantages agronomiques et pratiques agricoles (Brannan, 2021, Mamine et Farès, 2020).

Par exemple, la sélection variétale est historiquement adaptée aux cultures pures. Plusieurs auteurs appellent ainsi à une sélection variétale plus adaptée aux cultures associées (Annicchiarico et al., 2019, Mamine et Farès, 2020, Litrico et Violle, 2015). De plus, le machinisme agricole (semis, récolte) est peu adapté aux cultures associées (Morel et al., 2020). Le tri des grains pour l'alimentation humaine est techniquement difficile/coûteux (Magrini et al., 2013) et les coopératives n'achètent pas forcément de récoltes issues d'associations d'espèces. Ces raisons font que, malgré des avantages agronomiques reconnus dans certains cas, les cultures associées restent une pratique rare.

1.4 Indicateurs de la performance des cultures associées

Pour évaluer la performance des cultures associées, différents indicateurs existent, chacun ayant ses avantages et ses inconvénients (Bedoussac et Justes, 2011). L'un des indicateurs les plus utilisés est le "Land Equivalent Ratio" (LER). Il correspond à la surface relative nécessaire en culture pure pour obtenir le même rendement que la culture associée considérée (Équation (1.1), Willey et Rao, 1980) :

$$\mathrm{LER} = \frac{Y_i}{M_i} + \frac{Y_j}{M_j} \tag{1.1}$$

où Y_i (resp Y_j) est la valeur du rendement pour l'espèce i (resp j) en culture associée et où M_i (resp M_j) est la valeur du rendement pour l'espèce i (resp j) en culture pure.

Le LER a l'avantage d'être simple à calculer et interpréter. En revanche, il présente également certains défauts : un LER peut être élevé car le rendement de la culture pure correspondante est faible (ce qui est fréquemment le cas en contexte bas-intrants), ce qui a pour effet de sur-évaluer la performance du mélange. Cela peut être un problème en contexte de bas-intrants où les rendements des cultures pures sont faibles (3-4 $t.ha^{-1}$) par rapport aux standards en conventionnel (6-7 $t.ha^{-1}$) (Agreste, 2022). En conséquence, certains auteurs recommandent de toujours donner les valeurs de LER en les accompagnant des valeurs de rendement brutes, afin de savoir dans quelles gammes de rendement on se situe (Bedoussac et Justes, 2011).

Dans une optique d'évaluation de la performance des cultures associées, j'ai choisi d'étudier une autre métrique issue de l'écologie des communautés. Cette discipline scientifique se définit comme l'étude de l'organisation et du fonctionnement des assemblages d'espèces dont les populations sont en interaction au sein d'un écosystème donné (Begon et al., 2005). L'étude du lien entre biodiversité et fonctionnement des écosystèmes est une thématique de recherche à part entière dans cette discipline (Biodiversity Ecosystem Functioning ; Loreau et al., 2002). Une hypothèse avancée est que l'augmentation de la biodiversité dans un écosystème peut améliorer son fonctionnement grâce à une meilleure exploitation des ressources (du fait de la présence d'espèces contrastées). Cela peut l'amener à fournir une plus grande gamme de services, et le conduire à être plus stable et productif. Certains travaux proposent de transposer ces concepts de l'écologie des communautés (naturelles) aux agroécosytèmes (Brooker et al., 2021). Une des méthodes les plus utilisées est la décomposition de l'effet de la biodiversité (Loreau et Hector, 2001). Dans une culture associée, l'effet de la biodiversité se traduit par un gain ou une perte de rendement brut entre la culture pure et la culture associée et résulte de deux effets. L'effet de la complémentarité inclut la facilitation et la complémentarité de niche pour l'utilisation des ressources. L'effet de sélection, parfois appelé dominance (Fox, 2005), traduit le fait qu'une espèce fonctionnant déjà bien en culture pure profite du contexte de culture associée, au détriment de l'autre.

Dans le second article de ma thèse, je cherche à quantifier ces processus au sein du jeu de données construit, afin de comprendre le fonctionnement des cultures associées. Je caractérise ensuite l'effet de la fertilisation et du choix de l'espèce sur ces effets. Pour cela, j'ai calculé l'effet de la biodiversité et de ses composantes sur les cultures associées non fertilisées du jeu de données étudié. J'ai ensuite évalué, en me

restreignant à un sous-ensemble de mon jeu de données, l'effet de différents niveaux de fertilisation et du choix de l'espèce sur les différentes composantes de l'effet de la biodiversité.

L'ESSENTIEL

Les cultures associées céréale-légumineuse sont un mélange prometteur en raison de plusieurs interactions interspécifiques à l'oeuvre : complémentarité de niche pour l'utilisation de l'azote, complémentarité spatiale (racines), facilitation. Ces mécanismes assurent un certain nombre de services écosystémiques, dont un rendement généralement plus élevé en contexte bas-intrants, une meilleure qualité de grain pour la céréale ainsi qu'une plus grande stabilité du rendement. Enfin, la régulation des stress biotiques (notamment maladies et adventices) est généralement meilleure dans les cultures associées que dans les cultures pures. Généralement, les gains de rendement des cultures associées vis-à-vis des cultures pures sont mesurés via le Land Equivalent Ratio (LER), indicateur utile mais qui présente certaines limites. Dans le second article de ma thèse, j'étudie l'effet de la biodiversité, un autre indicateur issu de l'écologie des communautés, sur le gain (ou les pertes) de rendement des cultures associées vis-à-vis des cultures pures.

2 Species choice and N fertilization influence yield gains through complementarity and selection effects in cereal-legume intercrops

Rémi Mahmoud, Pierre Casadebaig, Nadine Hilgert, Lionel Alletto, Grégoire T. Freschet, Claire de Mazancourt, Noémie Gaudio

Key words : cereal-legume intercropping, biodiversity effect, complementarity effect, selection effect

DOI : 10.1007/s13593-022-00754-y

Article publié dans la revue Agronomy for Sustainable Development [1], également présenté à la conférence Intercropping for Sustainability [2].

Résumé : Maintenir le rendement tout en réduisant les intrants est l'un des principaux objectifs de l'agriculture durable. Dans ce contexte, les cultures associées céréale-légumineuse sont une pratique qui permet d'obtenir un rendement accru en conditions bas-intrants grâce à l'utilisation complémentaire des ressources abiotiques et des mécanismes de facilitation. Il existe de nombreuses options de gestion pour concevoir des systèmes de culture associées céréale-légumineuse, parmi lesquelles le choix des espèces cultivées et le niveau de fertilisation azotée (N) sont essentiels.

Dans cette étude, nous avons rassemblé les résultats de 35 expérimentations menées en Europe sur des cultures associées céréale-légumineuse combinant différentes espèces associées et différents niveaux de fertilisation azotée. Nous avons d'abord évalué l'intensité de l'effet de la biodiversité et ses composantes dans les cultures associées non fertilisées. Ensuite, nous nous sommes focalisés sur un sous-ensemble de systèmes pour analyser comment la fertilisation azotée influençait les effets de la biodiversité sur trois cultures associées (blé dur / pois, blé tendre / pois et blé dur / féverole). L'effet de biodiversité représente l'écart entre le rendement observé et le rendement attendu d'un mélange. L'effet de complémentarité représente la performance des mélanges par rapport à la performance des cultures pures. L'effet de sélection saisit la mesure dans laquelle une espèce ayant un rendement élevé en culture pure domine un mélange au détriment de l'autre espèce associée.

1. https ://www.springer.com/journal/13593
2. custom.cvent.com/05F40F40EA964BBB968323D70A3E1C38/files/db50436c39344465a6ed36deffe6c5af.pdf

Nos résultats ont confirmé un effet de biodiversité (donc un effet de l'association) globalement positif dans des conditions non fertilisées et dans diverses conditions climatiques (0,86 $\pm$ 0,04 $t.ha^{-1}$). L'effet de complémentarité était le principal moteur de ce gain de rendement puisqu'il représentait 76 % de l'effet de biodiversité, confirmant que la culture associée est une pratique utile dans les systèmes à bas niveaux d'intrants. La fertilisation azotée a réduit l'effet de complémentarité dans les cultures associées blé dur / pois, n'a pas influencé ces effets dans les cultures associées blé tendre / pois et n'a augmenté que l'effet de sélection dans les cultures associées blé dur / féverole. Ces résultats soulignent la nécessité de disposer d'une légumineuse suffisamment compétitive dans les cultures associées lorsque des engrais azotés sont appliqués, afin d'éviter une trop grande perturbation des interactions plante-plante. L'article montre donc que si les performances des cultures associées sont effectivement meilleures que celles des cultures pures en contexte non-fertilisé (conclusion cohérente avec la littérature), cet effet est modulé par l'identité des espèces ainsi que par le niveau de fertilisation.

2.1 Introduction

From 1960-2000, the use of fertilizers, irrigation and pesticides mitigated effects of climatic hazards, soil heterogeneity and pest pressure, and had a large and positive impact on crop yield (Tilman et al., 2002). More recently, especially in Europe, the growing trend of reducing inputs in agricultural systems, due to environmental and social concerns, and the climatic uncertainty caused by climate change have increased the variability in cropping conditions compared to that of the intensive agriculture practiced in the late 20^{th} century. To reduce the negative consequences of climatic uncertainty and continue to produce enough food while reducing the use of inputs (Sadras et Denison, 2016), a promising avenue is to favor functional complementarity of abiotic resource use and biological regulations between plants by designing innovative agricultural practices and systems (Duru et al., 2015a). This can be achieved by selecting relevant plant phenotypes (Lynch 2019) and/or using positive biodiversity effects through plant mixtures, also known as the biodiversity-ecosystem function (BEF) effects (Brooker et al., 2021).

Positive BEF effects on ecosystem services have been widely studied in natural communities (Cardinale et al., 2012), and interest in using them in cropping systems has increased in the past several years (Gurr et al., 2016 ; Martin-Guay et al., 2018 ; Brooker et al., 2021). Analyzing the diversity-productivity relationship enables the effect of biodiversity on primary production of a given system to be estimated and can divide it into complementarity and selection effects (Loreau et Hector, 2001). The former measures the effect due to niche complementarity and/or facilitation, while the latter measures the effect due to the dominance of a given species that fits well with the growth environment. Thus, BEF effects should be viewed as resulting from particularly positive specific interactions rather than explaining underlying processes themselves (Maier, 2012). As Brooker et al., 2021 highlight, a collaboration gap between BEF scientists and crop scientists has led to a poor understanding of "the operation of positive diversity effects in intensive agricultural systems" and thus of how to enhance them.

In agricultural systems, plant diversity can be promoted by a range of intercropping practices (i.e., combining at least two crop species in the same field for most of their growing periods), which may improve crop yield (Li et al., 2020b). Several mechanisms can, for example, improve nitrogen (N) acquisition by the intercrops, including complementary distribution of roots in soil volumes (Postma et Lynch, 2012), use of distinct forms of N in soils (McKane et al., 2002) and fixation of atmospheric N_2 by one species in the intercrop (Jensen et al., 2020). In a context of

input reduction, the use of N_2-fixing legumes is particularly promising. In Europe, this has been widely demonstrated in low-input cereal-legume intercrops, with an increase in total yield and cereal grain quality compared to those of sole crops (Bedoussac et al., 2015). However, supplying too much N fertilizer can cause the cereal to dominate the legume, which decreases positive plant-plant interactions in intercropping systems (Pelzer et al., 2012). Thus, the extent to which N fertilization can be used without compromising BEF effects in such systems remains unclear. More particularly, while recent meta-analyses and reviews generally agree upon positive BEF effects when multiple experiments are assessed, the results of individual experiments have high variability (Bedoussac et al., 2015 ; Gurr et al., 2016 ; Raseduzzaman et Jensen, 2017 ; Martin-Guay et al., 2018). Few recent studies underline a positive effect on intercrops' yield, via temporal niche differentiation (Yu et al., 2016 ; Dong et al., 2018 ; Li et al., 2020c).

In this study, using a database of 35 field experiments (Figure III.5) from five European countries, we first assessed the intensity of the biodiversity effect in winter and spring cereal-grain legume intercrops under unfertilized conditions. Then, focusing on a subset of three winter intercrops – durum wheat (*Triticum turgidum* L.) / pea (*Pisum sativum* L.), soft wheat (*Triticum aestivum* L.) / pea and durum wheat / faba bean (*Vicia faba* L.) – we tested the influence of two levels of N fertilization (moderate and high) on the biodiversity effect depending on the intercropped species considered.

Figure III.5 – Example of a field experiment of winter wheat / pea intercrops (and their corresponding sole crops) conducted at the ARVALIS experimental station, near Angers, France (Photograph courtesy of C. Naudin, ESA, France).

2.2 Materials and methods

2.2.1 Field experiments

To estimate the net biodiversity effect on intercrop productivity in a wide range of environmental conditions, we collected results from 35 factorial experiments conducted in five countries (France, Denmark, Italy, Germany, and the United Kingdom ; Figure III.6A), as detailed hereafter.

We used the following criteria to include set of experiments in our database : (1) grain yield was measured for both species in sole- and intercropping conditions, (2) different species and genotypes were used among cereal and legumes, and (3) a given mixture was observed at least in two locations.

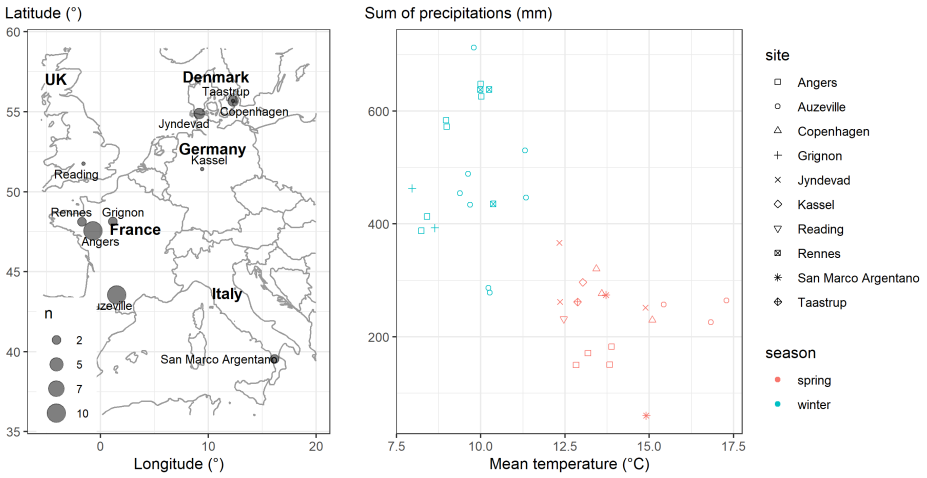


Figure III.6 – Location and main climatic features of the experiments. Panel A displays the number of experiments conducted at each location (different years and cropping systems). Panel B displays the sum of precipitation (mm) as a function of mean temperature (°C) during the crop cycle, with spring and winter crops encoded by colors, and experiment location encoded by symbols.

2.2.1.1 Environmental conditions

Climate conditions of each experiment were characterized using the following variables retrieved from the NASA POWER API : the sum of precipitation (mm) and mean temperature (°C) during the crop cycle (from sowing to harvest dates). The experiments were separated into two groups : winter crops, which had higher precipitation (280-712 mm) and lower mean temperature (6.8-11.3°C) during the crop

cycle, and spring crops, which had lower precipitation (60-366 mm) and higher mean temperature (12.3-17.3°C) (Fig. III.6B).

2.2.1.2 Agricultural management

All experiments included cereal-grain legume intercrops of two annual crop species and their corresponding sole crops for which grain yield ($t.ha^{-1}$) was measured at harvest. Cereals and legumes were each represented by three species : barley (*Hordeum vulgare* L.), durum wheat and soft wheat for the cereals and faba bean, lentil (*Lens culinaris* L.) and pea for the legumes (Table 1). In the database, 39% and 61% of the intercrops were spring or winter crops, respectively. Intercropped species were sown and harvested at the same time. The sowing dates ranged from March 11 to May 3 for spring crops and from October 25 to December 15 for winter crops. The harvest dates for all crops ranged from June 6 to August 23.

Table 1. Description of the 35 cereal-legume experiments analyzed in this study. The *Type* column defines if the experiment is carried on conventional (C) or organic (O) farming.

Intercropped species (cereal / legume)	Country	Year(s)	Soil water capacity (mm)	Soil texture (clay-silt-sand, %)	Type	N treatments ($kg.ha^{-1}$)	Mixture design	Spatial arrangement	No genotypes (cereal / legume)	Relative density in intercrop (cereal / legume)	References
Spring barley / fababean	Denmark	2001, 2002, 2003	173	24-29-47	O	0	substitutive	within row	2-1	0.5-0.5	(Gaudio et al., 2021; Hauggaard-Nielsen et al., 2008; Knudsen et al., 2004)
		2001, 2002, 2003	119	4-9-87	O	0	substitutive	within row	2-1	0.5-0.5	
Spring barley / pea	Denmark	2001, 2002, 2003	173	24-29-47	O	0	substitutive	within row	2-2	0.5-0.5	(Gaudio et al., 2021)
		2001, 2002, 2003	119	4-9-87	O	0	substitutive	within row	2-2	0.5-0.5	
	France	2003	173	24-29-47	O	0	substitutive, additive	alternate row	1-1	0.5-0.5, 0.5-1	
		2002	124	6-15-79	C	0	additive	alternate row	1-1	0.33-1	
		2003	124	6-15-79	C	0-130	substitutive, additive	alternate row	1-1	0.5-0.5, 0.5-1	
		2003, 2004	94	21-40-39	O	0	substitutive, additive	alternate row	1-1	0.5-0.5, 0.5-1	(Gaudio et al., 2021; Hauggaard-Nielsen et al., 2008, 2009; Launay et al., 2009)
	Germany	2004	176	51-29-20	O	0	substitutive, additive	alternate row	1-1	0.5-0.5, 0.5-1	(Gaudio et al., 2021)
	Italy	2003, 2004	169	22-36-42	O	0	substitutive	alternate row	1-1	0.5-0.5	(Gaudio et al., 2021; Hauggaard-Nielsen et al., 2008, 2009; Launay et al., 2009)
	United Kingdom	2003	142	49-32-19	O	0	substitutive, additive	alternate row	1-1	0.5-0.5, 0.5-1	
Spring soft wheat / lentil	France	2015	135	10-8-82	O	0	substitutive, additive	within row	2-4	0.5-1, 0.33-1, 0.3-0.7, 0.17-1	
		2016	187	18-48-34	O	0	substitutive, additive	within row	2-4	0.5-1, 0.33-1.3, 0.33-1, 0.3-0.7, 0.17-1.3, 0.17-1	

Winter durum wheat / fababean	France	2010	187	18-48-34	C	0-60-80-140	substitutive, additive	alternate-, within row	1-1	0.5-0.5, 0.67-0.5, 0.67-1, 0.33-0.5	(Kammoun, 2014)
		2011	187	18-48-34	C	0	substitutive	alternate row	1-1	0.5-0.5	
		2011	187	18-48-34	C	0-140	substitutive	alternate-, within row	1-1	0.5-0.5	
		2012	135	10-8-82	C	0	substitutive	within row	3-4	0.5-0.5	
		2013	187	18-48-34	C	0	substitutive	within row	3-4	0.5-0.5	
Winter durum wheat / pea	France	2006	187	18-48-34	C	0-100-180	substitutive	alternate row	1-1	0.5-0.5	(Bedoussac and Justes, 2010a, 2010b) (Kammoun, 2014)
		2007	135	10-8-82	C	0-60-80-140	substitutive	alternate row	4-1	0.5-0.5	
		2012	135	10-8-82	C	0	substitutive	within row	3-5	0.5-0.5	
		2013	187	18-48-34	C	0-140	substitutive	within row	3-5	0.5-0.5	
		2015	135	10-8-82	C	0	substitutive, additive	within row	1-4	0.5-0.5, 0.5-1	
Winter soft wheat / fababean	France	2018	169	22-36-42	O	0	additive	within row	8-2	0.7-0.75	
Winter soft wheat / pea	France	2010	205	11-54-35	C	0-45-90-140	substitutive, additive	within row	1-1	0.5-0.5, 0.33-0.66, 0.7-0.5	(Pelzer et al., 2016) (Gaudio et al., 2021; Naudin et al., 2010, 2014) (Gaudio et al., 2021)
		2017	205	11-54-35	C	0	substitutive, additive	within row	1-2	0.5-0.5, 0.5-1, 0.15-1, 0.05-1	
		2007	83	20-38-42	C	0-30-45	substitutive	within row	1-1	0.5-0.5	
		2008	83	20-38-42	C	0-30-45-60-90	substitutive	within row	1-1	0.5-0.5	
		2017	197	19-49-32	O	0	additive	within row	8-3	0.5-0.75, 0.5-1	
		2018	169	22-36-42	O	0	additive	within row	8-3	0.5-0.75, 0.5-1	
		2006	94	21-40-39	O	0	substitutive	within row	1-1	0.5-0.5, 0.3-0.7	
		2007	94	21-40-39	O	0-30	substitutive	within row	1-1	0.5-0.5, 0.7-0.3	
		2008	94	21-40-39	O	0-35-72	substitutive	within row	1-1	0.5-0.5, 0.7-0.3	
		2009	94	21-40-39	O	0-40	substitutive	within row	1-1	0.5-0.5, 0.7-0.3	

In the database, 54% of the intercrops were grown in a substitutive design (i.e., the sum of the relative sowing densities of the two species intercropped equals 1), while 46% were grown in an additive design (i.e., the sum of relative sowing densities exceeds 1). A species' relative density is its sowing density in the intercrop relative to that in its reference sole crop. Consequently, the database contained 199 sole crop experimental units and 307 intercrop experimental units (site x year x mix of genotypes x relative densities x N treatment), of which 140 were in an additive design and 167 in a substitutive design. Depending on the experiment, each experimental unit was replicated 2-8 times.

Additional details on experimental designs and management practices are reported in the reference publications of 33 of the 35 experiments (Knudsen et al., 2004 ; Corre-Hellou et al., 2006 ;Hauggaard-Nielsen et al., 2008 ; Hauggaard-Nielsen et al., 2009 ; Launay et al., 2009 ;Bedoussac et Justes, 2010a ; Bedoussac et Justes, 2010b ; Naudin et al., 2010 ; Naudin et al., 2014 ; Pelzer et al., 2016 ; Tang et al., 2016 ; Viguier et al., 2018 ; Gaudio et al., 2021a).

2.2.2 Estimating the biodiversity effect on intercrop performance

For each experimental unit, grain yield ($t.ha^{-1}$) was measured for each species. We calculated the biodiversity effect (BE, Loreau et Hector, 2001) as the observed grain yield minus expected grain yield in intercrops (Eq. (2.1)) :

$$BE = (YO_C + YO_L) - (YE_C + YE_L) \tag{2.1}$$

where YO_C and YO_L are the observed yield of the cereal and legume grown in intercrop, respectively, and YE_C and YE_L are the expected yield of the cereal and legume grown in intercrop, respectively.

Expected yield was estimated from the yield of the species in sole crop weighted by its scaled relative density in intercrop (Eq. (2.2) ; Li et al., 2020c) :

$$YE_C = M_C \frac{RD_C}{RD_C + RD_L} \text{ and } YE_L = M_L \frac{RD_L}{RD_C + RD_L} \tag{2.2}$$

where M_c and M_L are the yield of the cereal and legume in sole crop, respectively, and RD_C and RD_L are the relative density of the cereal and legume in intercrop, respectively. Grain yield in sole crops and intercrops is calculated as the mean from each replicate of every experimental units, within each experiment.

As mentioned, the biodiversity effect can be divided into a selection effect (SE, Eq. (2.3)) and a complementarity effect (CE, Eq. (2.4)) (Loreau et Hector, 2001 ; Li et al., 2020c) :

$$SE = \frac{1}{2} \times \left(\left(\frac{YO_C}{M_C} - \frac{RD_C}{RD_C + RD_L} \right) - \left(\frac{YO_L}{M_L} - \frac{RD_L}{RD_C + RD_L} \right) \right) \times (M_C - M_L) \tag{2.3}$$

$$CE = \frac{M_C + M_L}{2} \times \left(\frac{YO_C}{M_C} - \frac{RD_C}{RD_C + RD_L} + \frac{YO_L}{M_L} - \frac{RD_L}{RD_C + RD_L} \right) = M \times (LER - 1) \tag{2.4}$$

These formulas, used to compute selection and complementarity effects, are only valid in bispecific mixtures.

The first term of Eq. (2.3) calculates the difference in increase or decrease in yield between the two species intercropped, while the second term calculates the difference between their sole crop yields. Thus, a positive selection effect means that the species with the higher yield in sole crop has a higher relative increase in yield in intercrop (i.e., benefits more from intercropping).

Into the equation for the complementarity effect (Eq. (2.4)), we introduced the classic Land Equivalent Ratio, which is used to calculate land-use efficiency ($LER = \frac{Y_C}{M_C} + \frac{Y_L}{M_L}$; Willey et Rao, 1980). Thus, the complementarity effect equals the Land Equivalent Ratio minus 1, multiplied by M, the mean yield in sole crops.

2.2.3 Experimental design, data processing and analysis

The data were curated and formatted in a database. The data were ordered, reshaped and homogenized using the collection of R packages *tidyverse* (Wickham et al., 2019).

The dataset was unbalanced (i.e., groups had different numbers of observations) because the experiments collected were conducted for different purposes and examined many factors (e.g., N fertilization, intercrop design) (Table 1). Thus, the influence of several of the factors on the biodiversity effect and its components could not be analyzed, especially due to the lack of certain treatments in some experiments and to the nesting of factors. For example, only 12 of the 35 experiments tested N fertilization levels, or the species effect also included site and year effects (e.g., spring barley / faba bean intercrops were grown only in Denmark, so they could not be

analyzed properly). The statistical analysis performed was adjusted in response to this unbalanced structure.

We first investigated the overall behavior of mean biodiversity, complementarity and selection effects within the unfertilized cereal-legume intercrops in the 35 experiments, and the correlation between the biodiversity effect and each of its components. Thus, our goal was to assess the influence of N fertilization on the biodiversity effect and its components. N fertilization ranged from 0-180 kg $N.ha^{-1}$, which we split into three levels : null, moderate (30-80 kg $N.ha^{-1}$) and high (> 80 kg $N.ha^{-1}$). A factorial design was then defined between the species intercropped and these levels of N fertilization. The subset of our database with a factorial design of species and N fertilization levels corresponded to three intercrops : durum wheat / pea, soft wheat / pea and durum wheat / faba bean (70 experimental units, among which 62 are in substitutive design, all located in France, Table 1). Durum wheat / pea and durum wheat / faba bean intercrops were grown in experiments with moderate and high levels of N fertilization, while soft wheat / pea intercrops were grown only with a moderate level of N fertilization.

The effect of N fertilization on the biodiversity effect and its components in intercrops was assessed using the Bayesian approach. Bayesian inference is based on reallocating credible values for a parameter (posterior distribution) given prior knowledge (prior distribution) and the adequacy of the data to the model (likelihood). The Bayesian approach provides information about the probability of a hypothesis being true given the data (P(hypothesis|data)). Bayesian estimation for the difference in group means (Kruschke, 2018) is an alternative to the classic Student's t test to compare the means of two groups. This method calculates a posterior distribution for the mean differences between the two groups and derives a 95% highest density interval (HDI), which is defined as the 95% most credible values of the parameter. We performed Bayesian estimation for the difference in mean values of components of the biodiversity effect between N-fertilized (moderate and high) and unfertilized treatments for each of the three intercrops. The null hypothesis (H0) was defined as equal mean biodiversity effect components for N-fertilized and unfertilized intercrops. We applied the following decision rule to the position of the 95% HDI : reject H0 if the 95% HDI excludes 0 but do not reject H0 if it includes 0.

All indicator calculations and statistical analyses were performed with R software, v. 4.0.0 (R Core Team 2020). Bayesian statistical analyses were performed using the R package *BEST* (Kruschke and Meredith 2020).

2.2.4 Definition of references for fertilized legumes

A common assumption when calculating indicators to compare the performance of intercrops to that of sole crops is that N is not a limiting resource for legumes and does not influence their yield (e.g., Pelzer et al., 2012). To test this hypothesis, we performed Bayesian estimation for the difference in group means between N-fertilized and unfertilized legume sole crops. The database contained only three experiments (i.e., 11 experimental units) in which legume sole crops were N-fertilized, because the experiments we collected were designed to conform to agronomic practices of farmers, who rarely fertilize legume sole crops (Magrini et al., 2016). The Bayesian estimation confirmed that N fertilization had no significant influence on the yield of legume sole crops. Given this result and the lack of data on N-fertilized legume sole crops, we used the unfertilized legume sole crops as a reference when calculating the biodiversity effect and its components in all experimental units.

2.3 Results and discussion

2.3.1 Distribution of the biodiversity effect and its components in unfertilized intercrops

On the whole dataset, the mean (± 1 standard error) yield gain in unfertilized intercrops equaled 0.86 ± 0.04 $t.ha^{-1}$(1.04 ± 0.01 $t.ha^{-1}$for additive designs and 0.68 ± 0.00 $t.ha^{-1}$for substitutive designs) for a mean total intercrop yield of 3.54 ± 0.08 $t.ha^{-1}$(Figure III.7A). These results highlight an increase in the yield of cereal-legume intercrops in most experimental units under unfertilized conditions compared to those of the corresponding sole crops, which agrees with results of several studies (Pelzer et al., 2012 ; Pelzer et al., 2014 ; Yu et al., 2016) and confirms the ability of intercropping to increase grain yield in low-input farming systems (Bedoussac et al., 2015).

However, the increase in yield observed was influenced by the cropping conditions used as references to calculate the biodiversity effect. The unfertilized cereal sole crops used as references had lower grain yield (3.2 ± 0.08 $t.ha^{-1}$, all cereals pooled) than cereals grown under conventional farming conditions, which are always N fertilized (i.e., a mean grain yield of 6.1 $t.ha^{-1}$for the cereals of interest in the five European countries considered for the period covered by the experiments (Food and Agriculture Organization of the United Nations ; http://faostat.fao.org/)). Thus, the low yield observed for the unfertilized cereal sole crops contributed greatly to the positive biodiversity effect estimated (Garnier et al., 1997).

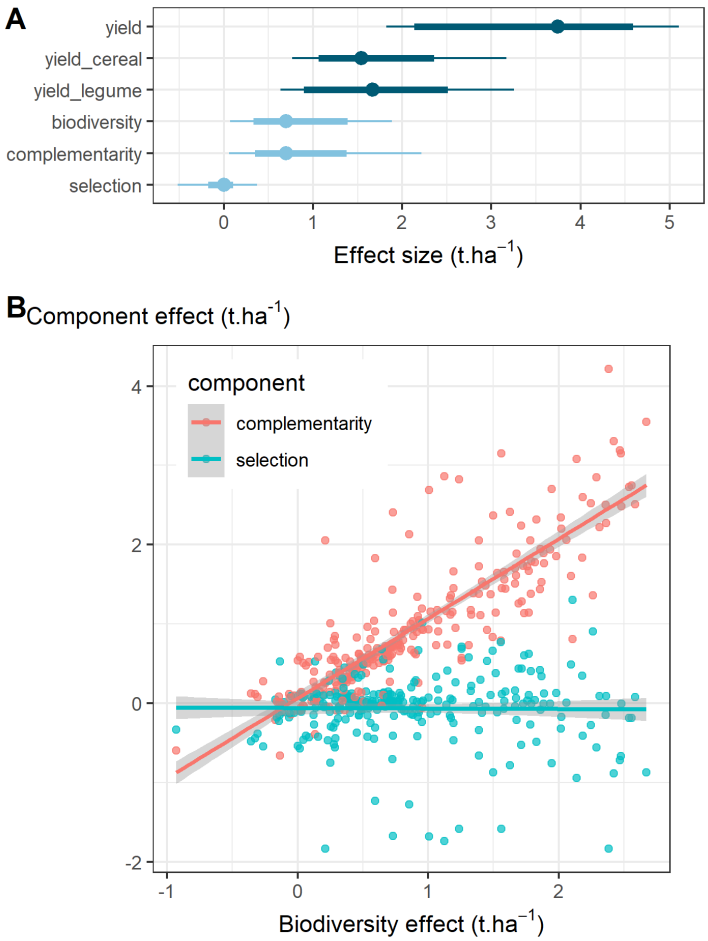


Figure III.7 – **(A)** Distribution of unfertilized cereal-legume intercrop yield and biodiversity effect ($t.ha^{-1}$). Points represent the median, broad lines represent the interquartile range, and thin lines represent the [0.1, 0.9] quantile interval. **(B)** Correlation between biodiversity effect ($t.ha^{-1}$) and complementarity effect ($t.ha^{-1}$) or selection effect ($t.ha^{-1}$) in unfertilized cereal-legume intercrops. Grey zones represent the 95% confidence interval for the linear regressions. Data used : whole dataset (n = 263)

The biodiversity effect was strongly and positively correlated with the complementarity effect ($r = 0.86$, $p < 10^{-15}$), but it was not correlated with the selection effect ($r = -0.01$, $p = 0.87$) (Figure III.7B). Thus, the complementarity effect was the main driver of the yield gain in unfertilized cereal-legume intercrops, meaning that positive plant-plant interactions (i.e., facilitation and / or niche complementarity) rather than the dominance of one of the species increased intercrop yields (Pelzer et al., 2012). However, caution is needed when distinguishing complementarity causes (e.g., niche partitioning, facilitation) of the resulting complementarity effect (Barry et al., 2019). To quantify the relative importance of these processes, specific measurements would be needed, such as symbiotic N_2 fixation to reflect differences in N use between cereals and legumes, or a lodging score to quantify mechanical facilitation (e.g., Podgórska-Lesiak et Sobkowicz, 2013). As Brooker et al., 2021 highlight, explicitly distinguishing facilitation and niche partitioning would help when applying new analytical and conceptual frameworks to design intercrops. Nevertheless, differences in N use in cereal-legume intercrops is a well-known process in which the more competitive cereal usually takes disproportionally more soil mineral N than the legume, which is forced to compensate by increasing symbiotic N_2 fixation (Rodriguez et al., 2020). In a low-input context, this complementarity of N use enables cereals in intercrops to have higher grain yield and quality than cereals in sole crops.

The complementarity effect contributed 76% of the biodiversity effect when the latter was positive (i.e., in 94% of the experimental units), but it contributed only 36% when the latter was negative (i.e., in 6% of the experimental units). In the few cases in which we observed a yield loss in intercrops, the relative contributions of complementarity and selection were reversed : -0.05 ± 0.02 and -0.16 ± 0.02 $t.ha^{-1}$, respectively. In these cases, the total yield of intercrops were lower than those of corresponding sole crops because the competition between cereals and legumes exceeded the complementarity effect (also reported by Pelzer et al., 2016 for soft wheat / pea intercrops and Baxevanos et al., 2017 for oat / pea intercrops).

2.3.2 Influence of N fertilization on the biodiversity effect and its components

The biodiversity effect and its components were altered by N fertilization, which is a key practice in agricultural systems. While the biodiversity effect was positive in 100% of the unfertilized experimental units of the data subset considered (i.e., factorial designs of species and N fertilization levels), the percentage of experimental units with a positive biodiversity effect decreased with N fertilization (i.e., 92% and 67% of the experimental units under moderately and highly N-fertilized conditions, respectively)

(Figure III.8). Overall, the total intercrop yield increased with N fertilization (4.16 ± 0.18, 5.09 ± 0.24 and 4.62 ± 0.21 $t.ha^{-1}$under unfertilized, moderately and highly N-fertilized conditions respectively) ; specifically, mean grain yield decreased for legumes (2.23 ± 0.12, 1.88 ± 0.19 and 1.84 ± 0.16 $t.ha^{-1}$under unfertilized, moderately and highly N-fertilized conditions respectively) but increased for cereals (1.93 ± 0.20, 3.21 ± 0.23 and 2.78 ± 0.15 $t.ha^{-1}$under unfertilized, moderately and highly N-fertilized conditions respectively) with N fertilization (Figure III.8). The same pattern was observed for the complementarity effect, which was positive in 96%, 83% and 56% of the experimental units under unfertilized, moderately and highly N-fertilized conditions, respectively. Conversely, the percentage of experimental units with a positive selection effect increased with N fertilization : 25%, 71% and 61% of the experimental units, under unfertilized, moderately and highly N-fertilized conditions, respectively. Thus, N fertilization tends to decrease positive plant-plant interactions within cereal-legume intercrops by acting on the balance between the two intercropped species to the benefit of the cereal (Pelzer et al., 2012).

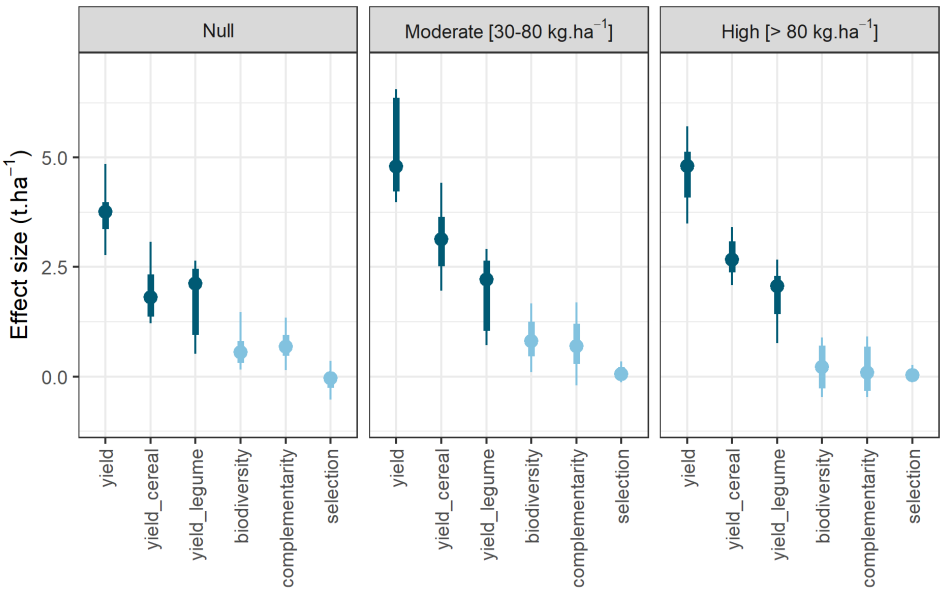


Figure III.8 – Distribution of cereal-legume intercrop yield, cereal and legume yield ($t.ha^{-1}$) and the biodiversity effect ($t.ha^{-1}$) as a function of nitrogen fertilization level. Points represent the median, broad lines represent the interquartile range, and thin lines represent the [0.1, 0.9] quantile interval. Data used : Experiments with a factorial design of species and N fertilization levels (n = 82)

The effect of N fertilization on the biodiversity effect and its components depended on

the species intercropped (Figure III.9). In durum wheat / pea intercrops, even moderate N fertilization decreased the biodiversity effect significantly by 66% compared to that under unfertilized conditions. This moderate N fertilization increased the selection effect significantly by 0.21 $t.ha^{-1}$(99.1% of the posterior values for the difference in group means between N-fertilized and unfertilized conditions were positive), while the complementarity effect decreased by 0.65 $t.ha^{-1}$(99.1% of the posterior values for the difference in means were negative). These effects were emphasized under highly N-fertilized conditions (Figure III.9). When focusing on the yield of both species intercropped, N fertilization disadvantaged the legume, since pea yield decreased by a mean of 37% under N-fertilized conditions compared to that under unfertilized conditions, while the opposite was observed for durum wheat, whose yield increased by a mean of 94%. These results could explain the shift in complementarity and selection effects for durum wheat / pea intercrops between N-fertilized and unfertilized conditions. This behavior is usually highlighted in existing literature related to cereal-legume intercrops (e.g., Naudin et al., 2010). Under N-fertilized conditions, selection effect increases because durum wheat has a competitive advantage over the legume (Mariotti et al., 2009 ; Duchene et al., 2017). Our results showed, however, that choosing a different cereal or legume species can change this effect.

When soft wheat replaced durum wheat in wheat / pea intercrops, N fertilization did not influence the biodiversity effect or its components (Figure III.9). Because the cereal and legume yields tended to increase slightly with N fertilization, the latter did not disrupt the balance between the two species (Table 2). Based on the soil and climate conditions considered, the level of N fertilization (45 kg $N.ha^{-1}$) was probably too low, compared to usual N fertilization rates in conventional agriculture, to increase the yield of one or both species significantly, unlike that of durum wheat / pea intercrops (60-140 kg $N.ha^{-1}$).

Finally, in durum wheat / faba bean intercrops, N fertilization did not influence the biodiversity effect or its complementarity effect, but it did increase the selection effect significantly by 0.3 $t.ha^{-1}$and 0.2 $t.ha^{-1}$under moderately and highly N-fertilized conditions, respectively (95.5% and 95.2% of posterior values for the difference in group means were positive, respectively) (Figure III.9). This increase was due to an increase in faba bean yield, since durum wheat yield changed little in intercrops as N fertilization increased. This behavior contrasts with that of pea yield when intercropped with durum wheat : pea yield decreased as N fertilization increased. Height and biomass differences between two intercropped species have been shown to influence their yields (Gaudio et al., 2021a). Since the faba bean is taller and larger than the pea (Guinet et al., 2018), it showed greater competitive ability (but whether

aboveground for light capture or belowground for nutrient and water acquisition remains to be tested), which explains the lack of shift in the biodiversity effect observed in durum wheat / faba bean intercrops.

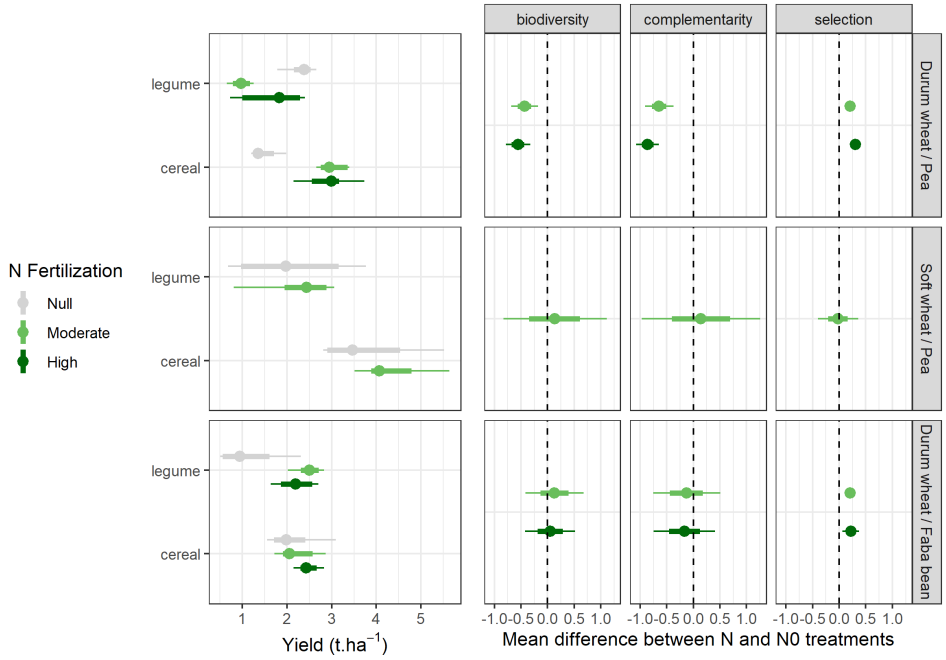


Figure III.9 – Distribution of cereal and legume yields ($t.ha^{-1}$) in three cereal-grain legume intercrops (durum wheat / pea, soft wheat / pea and durum wheat / faba bean) as a function of nitrogen (N) fertilization level : null, moderate (30-80 kg N.ha^{-1}) and high (> 80 kg N.ha^{-1}). For the three intercrops, posterior distributions of the difference in mean of the biodiversity effect between the two N-fertilized (moderate and high) and unfertilized (N0) treatments is illustrated ($t.ha^{-1}$), with dashed lines representing the null value of the posterior difference in means. Points represent the median, broad lines represent the interquartile range, and thin lines represent the [0.1, 0.9] quantile interval. Data used : Experiments with a factorial design of species and N fertilization levels (n = 82).

2.3.3 Pathway to applications

Because cereal-legume intercrops are used mainly to decrease the use of agricultural inputs, most are managed without synthetic inputs. In this way, our study confirmed an increase in productivity under a wide range of unfertilized cropping conditions, with a balance between the two species intercropped (i.e., no species clearly dominated), although the increase depends on the species intercropped (Cheriere et al., 2020). N fertilization can disrupt this balance, shifting positive plant-plant interactions to a

dominance of the cereal at the expense of the legume (e.g., in durum wheat / pea intercrops). This shift appeared at moderate N fertilization levels and even led to lower productivity of intercrops than that of sole crops at the high N fertilization levels applied to wheat sole crops in conventional agriculture (> 100 kg $N.ha^{-1}$).

It would thus be interesting to identify the level of moderate N fertilization that provides benefits from positive effects of intercropping and positive plant-plant interactions, while increasing the total yield by increasing the cereal yield, as farmers often perform in winter intercrops (Verret et al., 2020). Because this N level is likely to differ among species, future research should focus on the interaction between N fertilization and the intercrop species chosen. For instance, recent meta-analysis (Li et al., 2020c) shows high advantages of N fertilization on mixtures including maize (Zea mays L.).

In our study, only one combination of species x N fertilization had a positive interaction on yield (i.e., durum wheat / faba bean intercrops) : cereal yield increased and legume yield remained the same, while in durum wheat / pea intercrops, legume yield decreased. Thus, our results suggest that the legume chosen can be a management mechanism, with the idea that the legume should be sufficiently competitive to counterbalance the increased competition from the N-fertilized cereal (Duchene et al., 2017). Probably, it is the balance of competition between the two components rather than competitiveness of the legume that matters. However, we also observed that the cereal yield stagnated if the N fertilization level was not sufficient (e.g., soft wheat / pea intercrops). Thus, the optimal N fertilization level should depend on the proportion of legume biomass in the intercrop (Naudin et al., 2010). As highlighted by other studies, the species chosen are a relevant mechanism for controlling intercrops' yield (Cheriere et al., 2020) and suitability for the cropping environment in which they grow (Baxevanos et al., 2017). Finally, it is worthwhile to recall that many barriers to adoption of intercrops in Europe exist, beyond the scope of this article, such as. technical and economical ones (Bonke et Musshoff, 2020). Different possibilities (e.g., better communication of scientific results, breeding adapted to intercrops) exist to overcome these barriers (Meynard et al., 2018) and allow intercrops to be more widely cultivated.

2.4 Conclusion

This study highlights that the complementarity between intercropped species is the main driver of the positive biodiversity effect on the performance of cereal-legume intercrops under diverse cropping conditions. If the biodiversity effect depended instead

mainly on the selection effect (i.e., if one intercropped species strongly dominated), growing the dominant species alone would be more practical agronomically, which would shift the balance towards sole crops.

While multiple meta-analyses and reviews highlighted the overall yield gain in intercrops, analysis and tools to derive specific management recommendations for farmers from this general knowledge are still lacking (Brooker et al., 2021). We argue that it may be counterproductive to emphasize that biodiversity has this broad beneficial effect while the specific positive interactions between pairs of species and even more so, cultivars, remain to be identified (Maier, 2012).

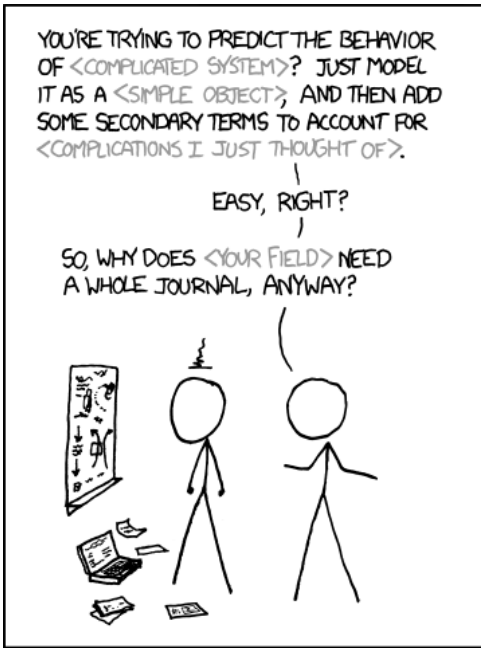
The key question remains how to secure complementarity while intensifying or increasing productivity. When focusing on the response of complementarity processes to N fertilization, we found that behavior differed depending on the species chosen. We highlighted that N fertilization does not always depress complementarity processes as long as the legume species can also benefit from it. Therefore, such shifts in balance need to be understood through the prism of community ecology to develop the use of intercrops in a wider range of agricultural systems besides low-input agriculture.

L'ESSENTIEL

En conditions non fertilisées, la culture associée a un effet global positif sur le rendement, principalement grâce à l'effet de complémentarité entre les deux espèces plutôt qu'à la dominance d'une des deux espèces. Cependant, la fertilisation azotée perturbe les effets de complémentarité et de sélection, et ce selon les espèces associées. La fertilisation a réduit l'effet de la biodiversité au sein des cultures associées blé dur / pois, via une diminution de l'effet de complémentarité et une augmentation de l'effet de sélection. Sur les cultures blé tendre / pois, aucune modification d'effet n'a été constatée. Finalement, un effet positif global de la fertilisation a été constaté sur les cultures associées blé dur / féverole, puisque le rendement global a augmenté sans perturber les effets de complémentarité entre les deux espèces. Les concepts issus de l'écologie des communautés permettent donc de tirer certaines conclusions qui restent assez générales. Il faut cependant analyser les déterminants de manière plus fine, notamment en étudiant la performance de chacune des espèces, pour mieux comprendre les agroécosystèmes considérés. C'est ce que nous proposons de faire dans le prochain chapitre.

Chapitre IV

Modéliser et identifier les déterminants de la performance des cultures associées céréale-légumineuse



1 Description du fonctionnement d'un mélange à travers un ensemble de prédicteurs

Le mélange que l'on observe résulte de l'expression des caractéristiques intrinsèques des espèces (et variétés) et de leurs interactions, sous l'effet de l'environnement de culture. Cet environnement est défini par les conditions pédo-climatiques et les pratiques agronomiques, qui conditionnent le niveau de ressources (lumière, azote minéral, eau, etc.). Chaque composante du mélange impacte et répond à cet environnement (Goldberg, 1990).

L'état actuel des connaissances écophysiologiques, mais aussi le type de données collectées dans les expérimentations agronomiques, ne permettent pas encore une approche de modélisation basée sur les processus ou une approche d'apprentissage utilisant l'ensemble des données phénotypiques brutes sans réflexion.

Nous proposons de résumer et d'organiser les données dont nous disposons en caractéristiques reflétant le fonctionnement des mélanges, en se basant sur des concepts et connaissances de l'écologie et de l'agronomie. Les stratégies des espèces pour accéder aux ressources peuvent être décrites par des valeurs de variables élaborées depuis les données brutes. Les grands types de prédicteurs que nous avons calculés et mobilisés sont donc liés aux interactions entre plantes, à l'environnement pédo-climatique et aux pratiques agronomiques.

L'ensemble des ces prédicteurs sera utilisé pour expliquer la performance des cultures associées. Nous avons sélectionné plusieurs variables de réponse (Tableau IV.1), conditionnellement aux données dont nous disposons. Nous avons ainsi choisi d'analyser les rendements bruts et relatifs (partial Land Equivalent Ratios (pLER), Mead et Willey, 1980) de la céréale et de la légumineuse. Nous nous sommes également intéressés à la qualité du grain de la céréale via la teneur en azote du grain, premier critère de qualité (et donc de prix de vente).

La structure du jeu de données global, la nature de nos prédicteurs et nos choix de modélisation ont conditionné les modalités de mélanges analysés afin de maintenir un effectif suffisant pour analyser l'effet des facteurs considérés (environnement, interactions entre plantes et pratiques). Parmi les mélanges représentés dans le jeu de données, nous nous sommes focalisés sur les cultures associées blé dur / féverole et blé dur / pois (Tableau IV.2).

Variable	Échelle	Formule et notation	Unité	Interprétation
Rendement	Céréale, Légumineuse	Y_C, Y_L	$t.ha^{-1}$	Rendement de la culture
Partial Land Equivalent Ratio	Céréale, Légumineuse	$\text{PLER}_X = \frac{Y_X}{S_X}$	-	Rendement relatif de la culture associée comparé à celui de la culture pure
Teneur en azote du grain	Céréale	N_C	$mg.g^{-1}$	Qualité du grain de la céréale

Table IV.1 – Variables de performance ajustées par nos modèles (X représente la céréale ou la légumineuse)

Mélange d'espèces	Année	Somme des précipitations (mm)	Température moyenne (°C)	Reliquat N (kg/ha)	Fertilisation N (kg/ha)	Date de semis	Date de récolte	Nb. d'observations
Blé dur / Féverole	2010	488.7	9.6	35	0-60-80-140	2009-11-20	2010-07-15	16
	2011	286.4	10.2	53	0-140	2010-12-03	2011-06-30	4
	2012	434.2	9.7	36	0	2011-11-14	2012-07-03	9
	2013	712.5	9.8	41	0	2012-11-20	2013-07-25	10
Blé dur / Pois	2006	454.7	9.4	45	0-100-180	2005-11-08	2006-07-04	3
	2007	530.5	11.3	39	0-60-80-140	2006-11-09	2007-07-10	16
	2012	434.2	9.7	36	0	2011-11-14	2012-07-03	9
	2013	712.5	9.8	41	0-140	2012-11-20	2013-07-25	19

Table IV.2 – Conditions environnementales et caractéristiques des expérimentations (toutes conduites à Auzeville) pour les mélanges considérés dans nos modèles

Au sein de chaque mélange d'espèces, certaines des variables de performance sont

linéairement corrélées entre elles, positivement ou négativement (Figure IV.1). C'est le cas des rendements relatifs et bruts de la féverole et du pois ($\rho = 0.75$ et 0.86 respectivement), ou du rendement relatif du blé et de la féverole ($\rho = -0.63$). On peut également noter l'absence de corrélations entre les rendements bruts et relatifs du blé au sein des associations blé/pois ($\rho = -0.02$).

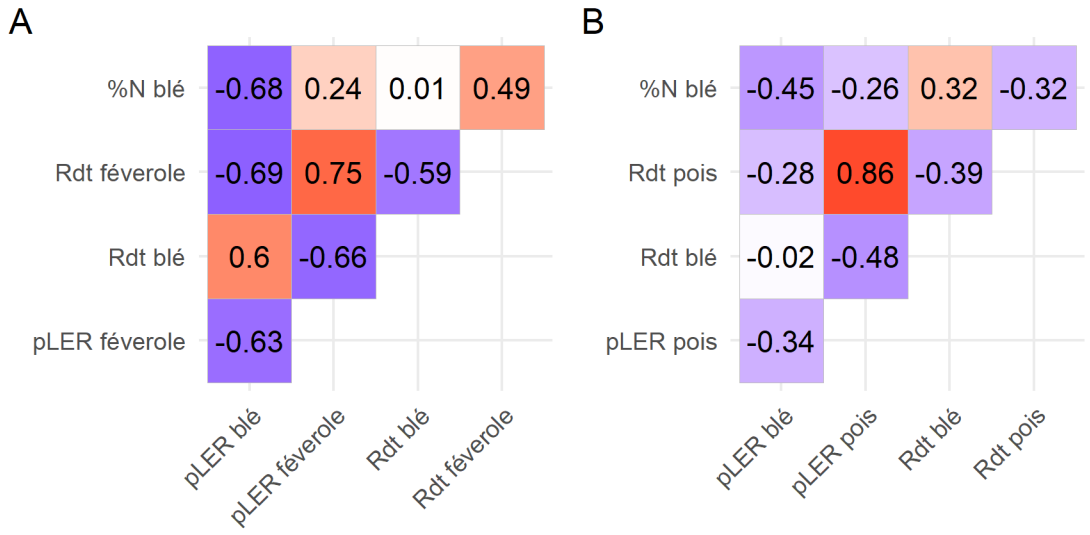


Figure IV.1 – Corrélations entre variables de performance pour A) les mélanges blé dur/féverole et B) blé dur/pois. L'intensité de la couleur des carrés est liée à la corrélation (négative : bleu, positive : rouge). pLER : partial Land Equivalent Ratio ; Rdt : Rendement ; %N : teneur en azote des grains du blé

2 Calcul des variables explicatives de la performance des cultures associées

2.1 Variables mesurées pour caractériser les espèces

Chaque espèce a ses propres caractéristiques, qui sont décrites selon les variables qui ont été mesurées dans les différentes expérimentations. Conditionnellement à notre jeu de données, les traits (caractéristiques des plantes) que nous avons inclus dans nos modèles sont reliés à la biomasse des parties aériennes, la hauteur, l'indice de surface foliaire (LAI) et la surface foliaire spécifique (SLA). Les expérimentations

étant hétérogènes, nous avons souvent dû faire des compromis entre le nombre d'expérimentations à considérer et le nombre de variables prédictives potentielles. Pour illustrer concrètement ceci, je prendrai l'exemple des mélanges blé dur/féverole (Figure IV.2). Parmi les 5 expérimentations ayant des unités expérimentales avec cette association, toutes n'incluent pas les mêmes variables mesurées. J'ai donc identifié l'ensemble des combinaisons expérimentations / variables mesurées, via la méthode des k-cliques proposée dans le Chapitre II. Cette procédure met en évidence le nécessaire compromis évoqué ci-dessus : inclure l'expérimentation nommée "Auzeville_pk_2011" impliquerait de ne pas inclure de variables liées à la surface foliaire comme variables explicatives dans nos modèles. J'ai donc fait le choix d'omettre cette expérimentation afin d'avoir plus de variables explicatives de la performance. Pour les variables mesurées en dynamique, j'ai utilisé les paramètres de courbe de croissance déterminés via la procédure de réduction de dimension décrite dans le Chapitre II.2.

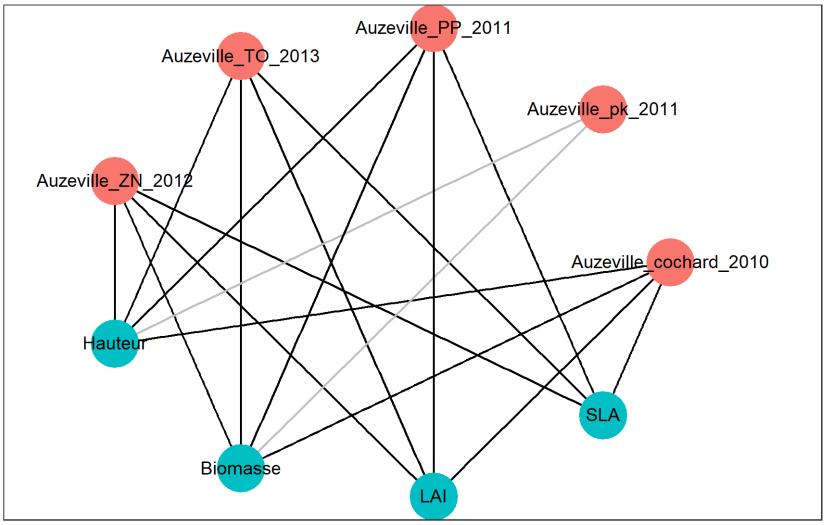


Figure IV.2 – Graphe représentant les combinaisons expérimentations / traits mesurées ensemble. Une expérimentation (noeuds rouges) est reliée à un trait (noeuds bleus) si elle inclut au moins une unité expérimentale où le trait est mesuré. Les arêtes noires correspondent à un ensemble de combinaisons expérimentation / trait formant un jeu de données induisant un plan factoriel complet. Les arêtes grises correspondent aux combinaisons expérimentation / trait non inclues dans la 2-clique choisie.

2.2 Variables représentant les interactions entre plantes

J'ai calculé trois grandes catégories de variables explicatives liées à la manière dont le mélange (et donc les interactions entre plantes) impacte les caractéristiques d'une

espèce (Tableau IV.4). Le choix de ces variables a également été orienté par mes questions de recherche.

Les informations acquises sur les cultures pures sont-elles prédictives de la performance du mélange ?

Certaines caractéristiques d'une espèce donnée peuvent différer selon qu'elle est cultivée en culture pure (interactions intraspécifiques) ou en mélange (interactions interspécifiques). Pour prendre en compte ces changements de comportement, nous avons calculé, pour chaque espèce, les différences de valeurs des paramètres entre les cultures associées et les cultures pures. Si pour un trait donné, la différence entre le mélange et la culture pure est nulle, alors le mélange n'a pas impacté le trait de l'espèce considérée. Dans ce cas, on peut supposer que les informations mesurées en culture pure peuvent être utilisées pour prédire le comportement en culture associée. Plus la différence est élevée, plus le fait d'être en association impacte l'espèce, que ce soit de manière positive (par exemple via une augmentation de la biomasse) ou négative (via une diminution de biomasse).

Les différences de caractéristiques entre espèces sont-elles prédictives de la performance du mélange ?

Dans une culture associée, des interactions positives (complémentarité, facilitation) et négatives (compétition) se produisent entre les deux espèces. En accord avec des travaux antérieurs (Kunstler et al., 2012, Montazeaud et al., 2018), nous posons comme hypothèse que la différence entre les valeurs des traits entre la céréale et la légumineuse au sein de l'association peut être un indicateur de ces interactions. Ces différences reflètent des comportements convergents (différence proche de zéro) ou au contraire divergents (différence importante) entre les deux espèces au sein du mélange. Selon les traits considérés, la convergence ou la divergence peuvent être recherchées (Bernard-Verdier et al., 2012). Ainsi, des divergences de biomasse entre deux espèces pourraient indiquer que l'espèce la plus développée imposerait une forte compétition pour les ressources du milieu sur l'espèce plus petite. A contrario, des divergences de profondeur d'enracinement entre les deux espèces pourraient aboutir à une complémentarité spatiale pour l'utilisation des ressources souterraines.

Le niveau de ressources azotées influence-t-il ces relations entre plantes ?

L'azote (N) est un élément nutritif essentiel qui limite la productivité des cultures (Plett et al., 2020). Le besoin en azote d'une plante est fortement lié à sa biomasse. Ce constat a conduit au concept de "courbe de dilution critique de l'azote" (Justes et al., 1994). Celle-ci donne la concentration minimale d'azote dans les parties aériennes

(Nc, en %) qu'une culture doit maintenir pour atteindre son taux de croissance maximal en fonction de sa biomasse. Nc est définie par deux paramètres a et b qui dépendent de l'espèce, mais sont indépendants des conditions pédo-climatiques. a est la concentration critique de l'azote quand la biomasse est d'une tonne par hectare, b représente la vitesse de décroissance de la courbe de dilution critique de l'azote. L'indice de nutrition azotée (INN) est une mesure du statut azoté d'une culture obtenue en divisant la concentration d'azote de la culture par sa dilution critique d'azote (Nc). Un INN supérieur (respectivement inférieur) à 1 correspond à un excès (respectivement un stress) azoté dans la culture (Lemaire et Meynard, 1997, Figure IV.3). La performance d'un mélange céréale-légumineuse par rapport aux cultures pures est largement déterminé par la disponibilité en azote minéral du sol, avec des cultures associées principalement adaptées à de bas niveaux d'intrants. Au-delà de la dose et date de fertilisation, la caractérisation du statut azoté du mélange est donc essentielle.

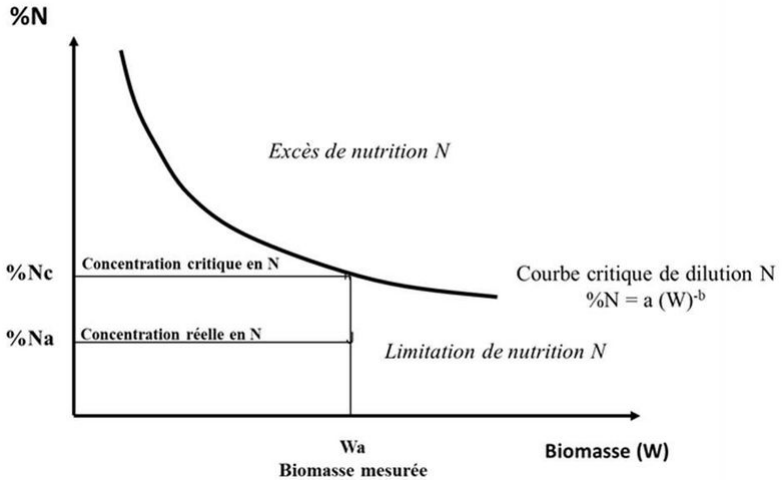


Figure IV.3 – Représentation d'une courbe de dilution critique en azote (N) d'une culture et de la détermination d'un indice de nutrition azotée (INN, figure tirée du site les mots de l'agronomie [1]). Pour une valeur de biomasse donnée (W), la courbe de dilution critique de l'azote est la teneur en azote critique ($\%N_c$) de la plante au-delà de laquelle la plante est en excès de nutrition N. L'INN est calculé comme le ratio de la concentration en N mesurée ($\%N_a$) et $\%N_c$

Cette notion est bien développée pour les cultures pures mais a dû être adaptée pour les cultures associées. Plusieurs méthodes ont été proposées pour évaluer le statut azoté dans les cultures associées (Louarn et al., 2021). Parmi celles-ci, nous en avons

1. https ://mots-agronomie.inra.fr/index.php/Fertilisation_des_cultures_ :_des_bases_scientifiques_renouvelées

choisi une permettant de discriminer le statut azoté des deux espèces en prenant en compte leur proportion respective dans le mélange.

Ainsi, pour évaluer l'INN de l'espèce i associée à l'espèce j :

$$\mathrm{INN}_i = \frac{\%\mathrm{N}_i}{p_i * (a_i * \mathrm{Biomasse}^{-b_i}) + (1 - p_i) * (a_j * \mathrm{Biomasse}^{-b_j})} \tag{2.1}$$

où i) a_i (resp. a_j) représente le paramètre a pour l'espèce i (resp. j), ii) b_i (resp. b_j) représente le paramètre b pour l'espèce i (resp. j), iii) $\%N_i$ est la concentration en azote (en %) de l'espèce i, iv) *Biomasse* représente la biomasse totale de la culture associée et v) p_i représente la proportion de l'espèce i dans la culture associée.

Les valeurs des paramètres a et b sont propres à l'espèce considérée et sont issues de la littérature existante (Tableau IV.3).

Espèces	a (%)	b (sans unité)	Référence
Pois & féverole	5.1	0.32	Louarn et al., 2021
Blé dur d'hiver	5.4	0.44	Justes et al., 1994

Table IV.3 – Paramètres utilisés pour calculer l'INN des cultures associées

Les valeurs de l'INN de chaque espèce de la culture associée ont été calculées en utilisant les valeurs de teneur azotée des parties aériennes à la récolte. L'INN tient compte à la fois des caractéristiques intrinsèques de la plante, des conditions de fertilisation et de l'azote disponible dans le sol.

Variable explicative	Unité	Formule	Interprétation
$\Delta_{\lambda,height}$	$°Cd$	$\lambda_{height,cereal} - \lambda_{height,legume}$	> 0 (resp. < 0) : La céréale (resp. légumineuse) démarre sa croissance (hauteur ou biomasse) en premier. Plus les différences sont élevées, plus les espèces démarrent leur croissance à un moment différent
$\Delta_{\lambda,biomass}$	$°Cd$	$\lambda_{biomass,cereal} - \lambda_{biomass,legume}$	
$\Delta_{\mu,height}$	$m°C^{-1}d^{-1}$	$\mu_{height,cereal} - \mu_{height,legume}$	> 0 (resp. < 0) : La céréale (resp. légumineuse) a une vitesse maximale de croissance (hauteur ou biomasse) plus élevée. Plus les différences sont élevées, plus les espèces ont une vitesse maximale de croissance différente
$\Delta_{\mu,biomass}$	$t.ha^{-1}°C^{-1}d^{-1}$	$\mu_{biomass,cereal} - \mu_{biomass,legume}$	
$\Delta_{\max,height}$	m	$max_{height,cereal} - max_{height,legume}$	> 0 (resp. < 0) : La céréale (resp. légumineuse) a une hauteur maximale plus élevée.
$\Delta_{\max,LAI}$	$m^2.m^{-2}$	$max_{LAI,cereal} - max_{LAI,legume}$	> 0 (resp. < 0) : La céréale (resp. légumineuse) a une surface foliaire plus élevée
$\Delta_{\max,SLA}$	$cm^2.g^{-1}$	$max_{SLA,cereal} - max_{SLA,legume}$	Traduit une différence d'adaptation à la lumière des deux espèces. *Cette variable a été calculée par cohérence avec les autres, mais elle s'est avérée non pertinente à posteriori en terme d'interprétation.*
$\Delta_{IC-SC,\lambda,height,X}$	$°Cd$	$\lambda_{height,IC} - \lambda_{height,SC}$	> 0 (resp. < 0) : L'espèce X a démarre sa croissance (hauteur ou biomasse) plus tard (resp. tôt) en culture associée
$\Delta_{IC-SC,\lambda,biom,X}$	$°Cd$	$\lambda_{biomass,IC} - \lambda_{biomass,SC}$	

$\Delta_{IC-SC,\mu,height,X}$	$m.^{\circ}C^{-1}d^{-1}$	$\mu_{height,IC} - \mu_{height,SC}$	> 0 (resp. < 0) : L'espèce X a une vitesse de croissance (hauteur ou biomasse) maximale plus élevée (resp. faible) en culture associée
$\Delta_{IC-SC,\mu,biom,X}$	$t.ha^{-1}{}^{\circ}C^{-1}d^{-1}$	$\frac{\mu_{biomass,IC}}{\rho} - \mu_{biomass,SC}$	
$\Delta_{IC-SC,max,height,X}$	m	$max_{height,IC} - max_{height,SC}$	> 0 (resp. < 0) : L'espèce X est plus grande (resp. petite) en culture associée
$\Delta_{IC-SC,max,LAI,X}$	$m^2.m^{-2}$	$max_{LAI,IC} - max_{LAI,SC}$	> 0 (resp. < 0) : L'espèce X a une surface foliaire plus élevée (resp. petite) en culture associée
$\Delta_{IC-SC,max,SLA,X}$	$cm^2.g^{-1}$	$max_{SLA,IC} - max_{SLA,SC}$	> 0 (resp. < 0) : L'espèce X réagit au contexte du culture associée en produisant plus (resp. moins) de surface foliaire par unité de masse
$cult_X$	-	-	cultivar de la céréale ou de la légumineuse

Table IV.4 – Variables explicatives calculées (X représente la céréale ou la légumineuse) ; ρ réprésente la densité relative entre culture associée et pure) ; SLA : Surface foliaire spécifique ; LAI : Leaf Area Index

2.3 Variables représentant l'effet de l'environnement sur le mélange

Nous avons tenté de représenter l'effet de l'environnement sur le mélange à travers un effet direct du climat, en mobilisant et développant plusieurs méthodes. J'ai expérimenté deux méthodes reposant sur deux hypothèses différentes : i) classer les expérimentations en groupes climatiques homogènes au sein desquels une culture associée donnée a un comportement relativement similaire, ii) identifier des zones sur lesquelles des variables climatiques ont eu une influence sur le rendement.

Au cours de ma thèse, cette étape s'est chronologiquement déroulée avant la démarche de modélisation de la performance. J'ai donc utilisé toutes les données à ma disposition (i.e. toutes les expérimentations) pour le travail lié à l'environnement.

2.3.1 Classification des expérimentations par groupes d'environnements climatiques

Pour cette approche, j'ai fait l'hypothèse qu'une culture associée se comportera de manière relativement similaire si elle se développe dans des environnements proches en matière de conditions climatiques. J'ai donc cherché à classer les expérimentations dans des groupes de climats similaires, afin de caractériser le climat des expérimentations et monter en généricité dans la procédure de modélisation.

Principe général

Pour classifier les expérimentations, j'ai utilisé des séries temporelles de données climatiques obtenues via l'API NASA POWER. Les variables climatiques considérées sont i) la température ($°C$), ii) le déficit hydrique (Précipitations - Evapotranspiration (mm)) et iii) les radiations ($MJ.m^{-2}$).

Pour un ensemble d'expérimentations partageant une espèce commune, j'ai calculé la distance entre les expérimentations sur la base de ces séries temporelles de données climatiques. Une distance entre deux séries temporelles X et Y peut être définie par la distance euclidienne $d(X, Y)$:

$$d(X, Y) = (\int |X - Y|^2)^{\frac{1}{2}}$$

Numériquement, étant donné N valeurs de X $(X_1, ..., X_N)$ et de Y $(Y_1, ..., Y_N)$, cette intégrale peut être approximée par la formule des trapèzes ($\hat{d}(X, Y) = (\frac{1}{2}(|X_1 - Y_1|^2 + |X_N - Y_N|^2) + \sum_{i=2}^{N-1} |X_i - Y_i|^2)^{\frac{1}{2}}$; Rahman et Schmeisser, 1990). Cette distance

ne peut être calculée que si les séries temporelles sont de longueurs égales. Or, les différentes expérimentations ont des cycles de culture de longueurs différentes (entre 84 et 137 jours pour les expérimentations de printemps et 192 et 262 jours pour les expérimentations d'hiver). Pour homogénéiser le nombre de points entre les expérimentations, j'ai réalisé une interpolation linéaire sur une grille de 500 points pour chacune des séries temporelles.

Ainsi, pour un ensemble de séries temporelles $X_1, ..., X_n$, on peut définir une matrice de distances $D \in \mathbb{M}_n$ par $D(i,j) = d(X_i, X_j)$ (où $\mathbb{M}_n$ est l'ensemble des matrices carrées avec n lignes et n colonnes). Pour chaque variable climatique (respectivement température, radiations et déficit hydrique), on obtient ainsi une matrice de distance inter-expérimentations (respectivement D_{Temp}, D_{Rad} et $D_{\text{P\_ETP}}$).

Afin d'éviter les effets d'échelles (unités différentes entre température, radiations, et déficit hydrique), les valeurs de chaque série temporelle ont au préalable été normalisées entre 0 et 1 : ($z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$).

J'ai agrégé les trois matrices de distance en les moyennant :

$D = \frac{1}{3}(D_{\text{Temp}} + D_{\text{Rad}} + D_{\text{P\_ETP}})$

Les individus (ici les expérimentations) ont ensuite été classés via une classification ascendante hiérarchique réalisée sur les matrices de distance. Le nombre de classes a été déterminé grâce à l'indice silhouette qui est un indice d'homogénéité intra-classes et d'hétérogénéité inter-classes que l'on cherche à maximiser, i.e. plus cet indice est élevé, plus les climats des expérimentations au sein d'un groupe se ressemblent et diffèrent des climats des expérimentations des autres groupes.

Résultats

J'ai réalisé cette classification sur les expérimentations incluant de l'orge et du pois, avec une courbe d'ajustement pour chacune des classes (Figure IV.4). Les résultats montrent que la variabilité des courbes ne permet pas de distinguer des climats particuliers (les courbes d'ajustement sont proches les unes des autres). Une raison probable, au vu des données, est que les climats caractérisant les expérimentations sont finalement relativement homogènes. Les expérimentations incluant l'orge de printemps ont toutes eu lieu en 2002, 2003 et 2004. Ce sont principalement des expérimentations ayant eu lieu au Danemark, qui se retrouvent dans une classe, et des expérimentations ayant eu lieu à Angers, qui se retrouvent dans deux classes distinctes. Les expérimentations incluant le pois d'hiver ont toutes eu lieu en France, à Auzeville, Angers et Rennes. Une classe regroupe les expérimentations ayant eu lieu en 2007, 2008, 2012 et 2018 et l'autre regroupe les expérimentations ayant eu

lieu en 2006, 2009, 2010, 2013 et 2017. Le calcul de distances entre courbes, combiné à la classification hiérarchique, ne permet donc pas de distinguer assez finement les expérimentations entre elles.

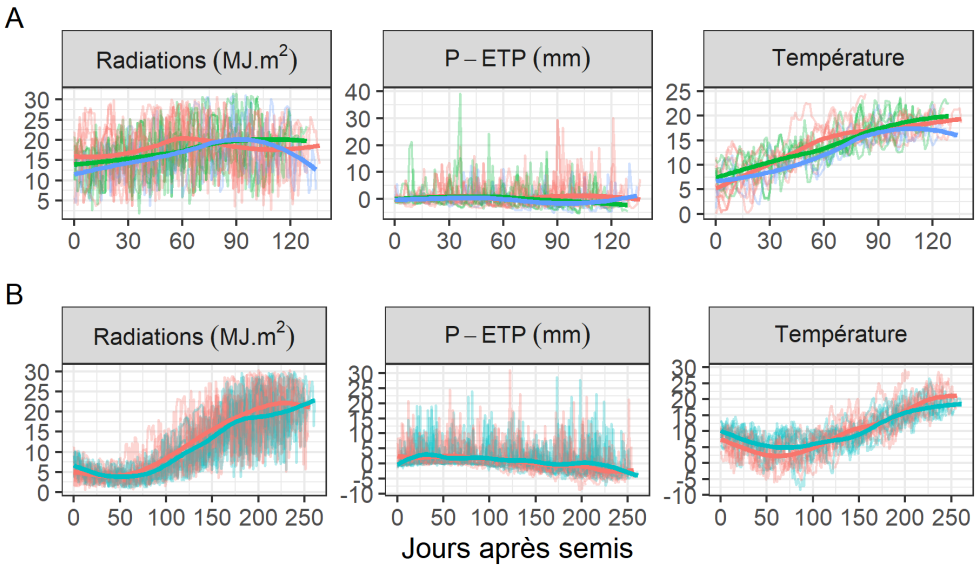


Figure IV.4 – Résultat de la procédure de classification implémentée. Chaque couleur correspond à un groupe d'expérimentations déterminé. Les courbes correspondent à des ajustements par modèle additif généralisé (fonction `geom_smooth` du package ggplot2, Wickham, 2016). A : Expérimentations incluant de l'orge de printemps, B : Expérimentations incluant du pois d'hiver

2.3.2 Sparse Interpretable Slice Inversed Regression (SISIR)

Pour cette seconde approche, mon hypothèse était qu'il est possible d'identifier certains moments climatiques du cycle de croissance qui sont déterminants pour le rendement, alors que d'autres ne le sont pas.

Présentation succincte de la méthode

La méthode SISIR (Picheny et al., 2019) est une méthode de régression de type function-on-scalar, *i.e.* on utilise des variables fonctionnelles (e.g. température, précipitations) pour prédire une variable scalaire (rendement par exemple). Elle est implémentée dans le package `SISIR` (Picheny et al., 2021).

SISIR est originale de par sa capacité à traiter les données en haute dimension (faible nombre d'individus par rapport au nombre de variables) et son objectif d'interprétabilité. Ces caractéristiques en faisaient a priori une méthode adaptée à nos données. Les variables climatiques sont mesurées tous les jours et la taille des séries temporelles correspond donc au nombre de jours du cycle de culture (entre 84 et 137 jours pour les 17 expérimentations de printemps et 192 et 262 jours pour les 18 expérimentations d'hiver).

Formellement, la méthode SISIR considère deux variables aléatoires (X, Y). X correspond à une variable aléatoire fonctionnelle (*i.e.* une série temporelle) observée à des points $t_1, ..., t_p$. On considère un échantillon $(x_i, y_i)_{i=1,...,n}$ de n observations, où $x_i \in \mathbb{R}^p$, $x_i(t_j)$ est une mesure de variable climatique et les y_i sont les rendements.

SISIR suppose que la variable réponse Y et la variable explicative X sont liées par un modèle de la forme : $Y = F(a_1^T X, ..., a_d^T X, \varepsilon)$ où $d < p$, F est une fonction inconnue, les $a_1, ..., a_d$ sont des vecteurs de taille p (longueur de X) et ε est un terme d'erreur (Picheny et al., 2019). Ainsi, l'objectif est de réduire la dimension des prédicteurs, en projetant les observations sur un sous-espace (engendré par les $a_1, ..., a_d$) étant censé contenir toute l'information sur Y contenue dans X. Ce sous-espace est appelé Effective Dimension Reduction (EDR, Li, 1991). L'hypothèse que seuls certains intervalles temporels sont utiles pour la prédiction revient à faire l'hypothèse qu'un grand nombre de valeurs consécutives des vecteurs a_j sont nulles.

Dans le prochain cadre, je décris brièvement les grandes étapes de SISIR. La compréhension fine de la méthode n'est cependant pas nécessaire pour les paragraphes qui le suivent.

Explication du principe de SISIR

1. La plage de variation des valeurs de Y est découpée en H intervalles. *Dans cette étude, cela revient à découper les valeurs de rendement en H intervalles, par exemple rendements faibles / intermédiaires / élevés (si $H = 3$).*
2. On calcule $(\bar{X_1}, ..., \bar{X_H})$, où $\bar{X_h} \in \mathbb{R}^p$ est la moyenne des x_i telles que y_i soit dans le sous-intervalle h (*Dans cette étude, cela revient à moyenner les courbes de température où les rendements sont faibles / intermédiaires / élevés (si $H = 3$).*
3. On calcule la matrice de covariance empirique de X : $\hat{\Sigma} = \frac{1}{n}\sum_1^n (x_i - \bar{X})(x_i - \bar{X})^\top$ ainsi que la matrice de covariance empirique de $\mathbb{E}[X|Y]$ ($\mathbb{E}[X|Y]$ *peut se traduire par "la valeur moyenne du phénomène explicatif étudié (température, etc.) sachant la valeur du rendement Y ")* : $\hat{\Gamma} = \sum_1^H p_h \bar{X_h} \bar{X_h}^\top$, où p_h correspond à la proportion des observations dans l'intervalle h.

En haute dimension ($n \ll p$), la matrice de covariance des observations est dite singulière (c'est-à-dire qu'on ne peut pas l'inverser, Wang et al., 2022). Une pratique courante consiste alors à pénaliser la diagonale de cette matrice en lui ajoutant une petite valeur μ pour la rendre inversible.
Pour rappel, l'espace EDR sur lequel sont projetées les observations est engendré par d vecteurs $a_1, ..., a_d$. On note $A := [a_1|...|a_d]$ la matrice de taille $p \times d$ dont les colonnes sont les vecteurs $a_1, ..., a_d$.

4. L'estimateur de A est constitué des d premiers vecteurs propres de la matrice $(\hat{\Sigma} + \mu\mathbb{I})^{-1}\hat{\Gamma}$
5. Les observations de chaque sous-intervalle h sont projetées sur l'EDR grâce à la matrice A
6. Une étape de "réduction" des valeurs des a_j est effectuée : On multiplie chacune des valeurs de $\hat{a}_j$ par un $\alpha_k (k \in 1, ..., p)$ (*ce vecteur de coefficients est là pour réduire autant que possible la valeur des $\hat{a}_j$*) lui même déterminé par une procédure de type LASSO (*une régression linéaire ou on pénalise la valeur des coefficients de manière à en rendre un maximum nuls*).

À l'issue de ces étapes, on obtient des a_j "réduits" (shrinked) notés $\hat{a}_j^s$. Ces vecteurs sont par construction parcimonieux (*c'est-à-dire contenant beaucoup de 0*).
Une procédure de fusion des intervalles est appliquée pour avoir des intervalles les moins "hachés" (avec beaucoup de valeurs consécutives de a_j nulles) et les plus interprétables possibles (Picheny et al., 2019).
Les paramètres μ, d, H sont des hyperparamètres à ajuster via une procédure de validation croisée implémentée dans le package `SISIR`.

Application à notre cas d'étude

Nous disposons de plusieurs données de rendement par expérimentation. Pour donner le même poids à chaque expérimentation, j'ai programmé une procédure itérative permettant de prendre en compte la variabilité des rendements au sein de chaque expérimentation. À chaque itération $b = 1, ..., B$:

1. Je tire au hasard une valeur de rendement par expérimentation
2. J'ajuste la méthode SISIR sur une grille d'hyperparamètres (H, μ, d)
3. Je choisis la meilleure combinaison d'hyperparamètres
4. Je détermine les zones d'influence des variables climatiques via la méthode décrite dans Picheny et al., 2019

J'ai appliqué cette méthode avec $B = 300$, afin de voir si certains intervalles étaient plus souvent sélectionnés que d'autres, ce qui permettrait de les identifier comme étant particulièrement importants pour expliquer le rendement.

Résultats

Sur la Figure IV.5, on observe le nombre de fois (parmi les B itérations) où la procédure SISIR a sélectionné chaque intervalle. L'échelle des couleurs montre que chaque intervalle de température a été choisi au moins une fois sur deux, sans que des périodes particulières ne ressortent fortement. Ainsi, la procédure appliquée sur mes données n'a pas permis de distinguer des périodes climatiques importantes pour le rendement du pois de printemps et de réduire la dimension de ces courbes. J'ai constaté des résultats similaires sur d'autres courbes (précipitations / radiations) et d'autres espèces.

Plusieurs facteurs peuvent expliquer cette absence de résultats probants :

1. le faible nombre d'individus ($n \approx 15$) en comparaison avec la dimension (nombre de jours dans le cycle de culture) $p \in [84; 262]$,
2. les hypothèses sous-jacentes à la méthode sont trop fortes (homogénéité des zones d'influence de chaque variable climatique entre les expérimentations),
3. la faible variabilité des expérimentations en terme de variables climatiques.

Perspectives : Extension de SISIR au cadre multivarié

La méthode SISIR est conçue pour prendre en entrée une seule variable fonctionnelle. J'ai cependant cherché, en prévision d'une application sur nos données, à étendre la méthode au cadre multivarié pour prendre en entrée plusieurs variables climatiques.

Formellement, au lieu de ne considérer qu'une variable aléatoire X en entrée, on peut considérer un ensemble de variables aléatoires $(X^{(1)}, ..., X^{(s)})$ tel que $X^{(k)} \in \mathbb{R}^{p(k)}, k = 1, ..., s$.

La méthodologie et les équations sont décrites succinctement en Annexes.

Dans le cadre de mon étude, puisque les résultats n'ont pas été probants en univarié, je n'ai pas implémenté la méthode en multivarié, mais la méthodologie et les équations pourront être réutilisées et implémentées dans des projets où les données le permettent.

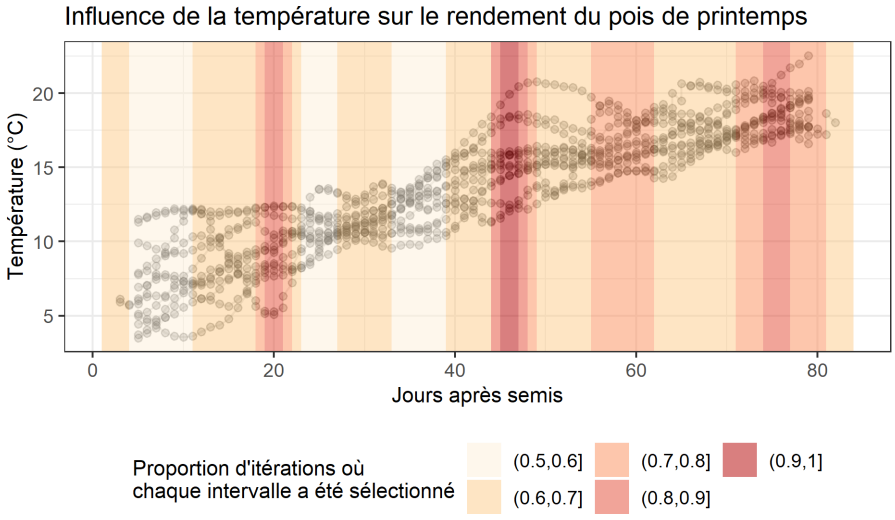


Figure IV.5 – Résultat de la procédure itérative implémentée : la coloration des intervalles représente la proportion des itérations où ils ont été sélectionnés par SISIR comme importants pour le rendement du pois de printemps. Tous les intervalles ont été sélectionnés plus d'une fois sur deux, ne permettant pas de distinguer des zones pour le rendement.

L'ESSENTIEL

Les différentes méthodes testées pour tenir compte explicitement des variables climatiques (température, déficit hydrique et radiations), que ce soit en constituant des groupes d'environnements similaires ou en identifiant des zones influentes dans les variables climatiques, n'ont pas été concluantes. Dans le cas de la classification des expérimentations, les classes identifiées étaient difficilement caractérisables en terme de climat. Dans le cas de l'utilisation de SISIR, toutes les zones des séries temporelles étaient considérées comme influentes plus d'une fois sur deux (au cours de la procédure itérative implémentée). Les hypothèses sous-jacentes aux méthodes utilisées, le faible nombre d'individus statistiques (dans ce cas les expérimentations) ainsi que la relative homogénéité des expérimentations entre elles expliquent probablement l'absence de résultats probants. Cependant, l'environnement (sol, climat) étant une variable clé influençant le rendement, il est nécessaire d'inclure d'une autre manière une variable liée à l'effet expérimentation dans les modèles de prédiction du rendement. Nous avons donc défini la structure des modèles pour tenir compte de ce facteur.

3 Stratégie de modélisation et ajustement des modèles

3.1 Démarche de modélisation

3.1.1 Argumentaire des choix de modélisation

Après avoir construit un ensemble de prédicteurs censés représenter au mieux le fonctionnement des cultures associées céréale-légumineuse, nous avons fait le choix de modéliser le lien entre variables réponses et variables explicatives par un algorithme de forêt aléatoire avec facteur aléatoire. La forêt aléatoire est un algorithme (dont nous détaillons le fonctionnement plus bas) ayant montré de bonnes capacités prédictives dans de nombreux cas (Fernández-Delgado et al., 2014). Ses bonnes performances s'expliquent par sa capacité à prendre en compte les interactions non linéaires entre prédicteurs et les effets de seuil (comportements qui changent au-delà d'une certaine valeur de la variable explicative). De plus, les forêts aléatoires sont robustes à la présence de variables explicatives qui pourraient être peu pertinentes pour expliquer/prédire la variable réponse (Grinsztajn et al., 2022, *preprint*).

Cependant, il existe une structure de dépendance des observations au sein de chaque expérimentation puisque ces observations partagent un même climat. Le climat étant un facteur important dans la conduite de cultures, il est nécessaire de tenir compte de cette structure dans la démarche de modélisation. De plus, d'autres effets environnementaux (pédologie) et techniques (effet expérimentateur) s'ajoutent aux effets climatiques. Une approche possible envisagée était de modéliser finement l'effet du climat et de la pédologie (sol) sur les variables réponses pour tenir compte explicitement de cet effet expérimentation. Cependant, les approches testées n'ont pas été concluantes (Chapitre IV.2.3).

Les analyses statistiques en agronomie se basent de plus en plus sur l'utilisation de modèles mixtes, en mettant l'année en facteur aléatoire. Cette pratique a pour objectif de tenir compte de la dépendance intra-année des observations et du fait que le facteur aléatoire n'est pas intéressant en soi, bien qu'il explique une part de la variabilité de la variable réponse. Pour tenir compte de la dépendance intra-expérimentation des observations, j'ai donc utilisé une méthode combinée de forêt aléatoire et de modèle mixte (conçue par Hajjem et al., 2014 et implémentée par Capitaine et al., 2021) pour comprendre les déterminants de la performance des cultures associées. J'ai utilisé les prédicteurs décrits dans le Tableau IV.4 comme effets fixes (dans la forêt aléatoire) et l'expérimentation comme facteur aléatoire. Cette démarche a pour objectif de

s'appuyer à la fois sur la flexibilité et les bonnes performances prédictives des forêts aléatoires et sur la structure de dépendance induite dans les expérimentations grâce au facteur aléatoire.

3.1.2 Imputation des données

Dans les jeux de données globaux, certaines données peuvent manquer (Figure IV.6). Supprimer les individus/variables avec données manquantes est une procédure courante mais problématique car elle peut biaiser l'échantillon et réduire considérablement sa taille. La littérature scientifique souligne certaines alternatives comme l'imputation multiple, qui permet d'obtenir plusieurs valeurs plausibles pour chaque valeur manquante, tout en gardant une information sur la variabilité due au processus d'imputation (Patrician, 2002). Dans notre cas, certains paramètres des courbes n'étaient pas disponibles pour diverses raisons : i) certaines unités expérimentales n'avaient pas de mesures dynamiques ; ii) certaines des courbes n'étaient pas ajustées correctement.

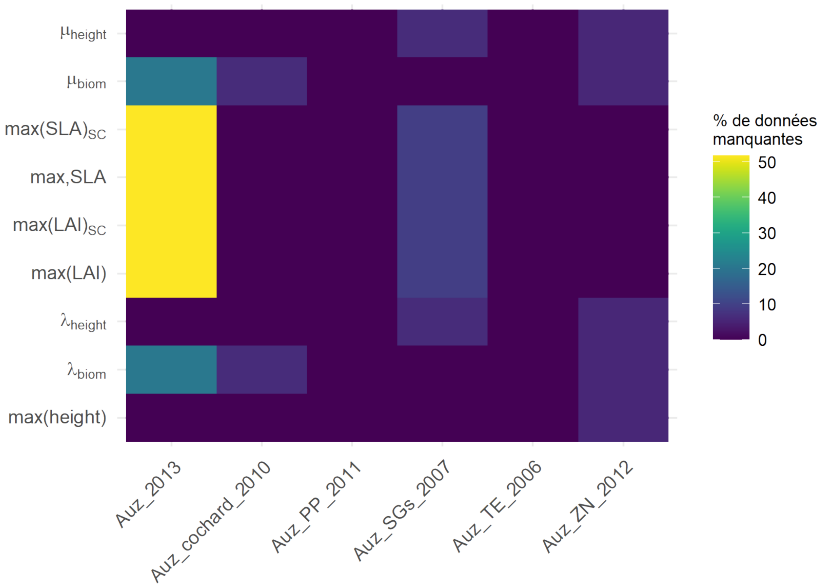


Figure IV.6 – Proportion de données manquantes en fonction des variables et des expérimentations

Nous avons utilisé l'algorithme JointAI (Erler et al., 2021) pour obtenir 10 versions de chaque ensemble de données (chaque mélange d'espèces). Cet algorithme permet d'imputer des données manquantes en se basant sur un modèle linéaire mixte bayésien (ce qui permet de tenir compte de la dépendance des observations au sein de chaque

expérimentation). Nous avons appliqué la procédure de modélisation décrite dans les prochaines sections sur chacun des 10 ensembles de données imputés, ce qui nous a donné 10 valeurs ajustées pour chaque valeur observée. Cela nous a permis d'obtenir une estimation de l'influence du processus d'imputation sur la variabilité globale.

3.1.3 Fonctionnement des forêts aléatoires

Les forêts aléatoires sont un type de modèle de machine learning utilisé pour la classification et la régression (Breiman, 2001). Dans la suite de ce document, je ne considérerai que les forêts aléatoires de régression, même si les principes décrits restent les mêmes en classification. Elles font partie de la catégorie plus large des méthodes dites de bagging (Bootstrap Aggregating), dont le principe est d'agréger (en moyennant par exemple) les prédictions issues d'un ensemble de modèles. Dans le cas des forêts aléatoires, il s'agit d'arbres de régression, c'est-à-dire des modèles permettant de prédire la valeur d'une variable continue à partir d'un ensemble de p variables explicatives. Les arbres de régression sont construits en définissant des partitions binaires successives de l'échantillon sur la base des variables explicatives. À une étape donnée du processus, on choisit une variable explicative et un seuil qui définissent deux groupes (selon que l'individu a une valeur observée au-dessous ou au-dessus du seuil) et qui correspondent au couple de variable et seuil minimisant la variance intra-groupe de la variable réponse.

L'un des avantages des forêts aléatoires est qu'elles sont généralement moins sensibles aux variations des données d'entraînement et plus performantes (qualité d'ajustement/-prédiction) que les arbres de régression individuels. Elles sont également relativement simples à utiliser et ne nécessitent pas beaucoup de paramétrage. De plus, la combinaison des arbres de régression formant la forêt aléatoire permet de prendre en compte des interactions non-linéaires existant entre les variables explicatives.

Voici les principales étapes d'entraînement d'une forêt aléatoire (en régression) sur un jeu de données contenant p variables explicatives :

1. Sélectionner un nombre élevé d'arbres de régression à inclure dans la forêt.
2. Pour chaque arbre de la forêt, sélectionner un sous-ensemble de variables explicatives (généralement $\frac{p}{3}$) et entraîner cet arbre sur un échantillon bootstrapé du jeu de données (sous-ensemble des données d'origine qui est construit par tirage aléatoire avec remise). Chaque arbre entraîné ne contient donc nécessairement pas chaque observation.
3. La valeur *ajustée* de chaque observation est obtenue en moyennant la valeur

prédite par chaque arbre dont l'échantillon d'entraînement ne contenait pas cette observation.

Un des inconvénients des forêts aléatoires est cependant leur manque d'interprétabilité, comparativement à des méthodes plus communes de type régression linéaire ou modèle additif généralisé. Pour contrebalancer cela, j'ai calculé une mesure dite d'importance des variables explicatives. Cette mesure se base sur l'hypothèse suivante : plus une variable est importante, plus sa dégradation (par permutation par exemple) ou sa suppression aura une influence forte sur la qualité de l'ajustement du modèle.

L'importance d'une variable explicative est calculée comme suit :

1. Pour chaque variable explicative, on calcule d'une part l'erreur de prédiction de chaque arbre qui l'utilise à partir des individus qui n'ont pas été utilisés pour entraîner l'arbre et d'autre part l'erreur de prédiction de l'arbre après avoir permuté aléatoirement les valeurs de cette variable.

2. La différence entre les deux erreurs de prédiction donne l'importance de la variable pour un arbre.

3. En moyennant l'importance de la variable pour tous les arbres qui l'utilisent, on obtient l'importance de la variable pour l'ensemble de la forêt aléatoire.

Les modèles de prédiction des rendements (bruts et relatifs) développés ont pour objectif de mieux comprendre le fonctionnement des cultures associées céréale-légumineuse. Les prédictions générées sont donc de type corroboratives (voir Chapitre I). Les variables explicatives calculées sont supposées être de bons prédicteurs de la performance des cultures associées. En ayant des modèles ajustant au mieux cette performance, l'hypothèse que nous faisons est que l'on peut décrire le fonctionnement de ces cultures via une mesure d'importance des variables explicatives.

3.1.4 Sélection de variables

L'algorithme de forêt aléatoire ne prévoit initialement pas de procédure de sélection de variable. Notre objectif n'étant pas d'utiliser aveuglement toute l'information disponible pour décrire notre système, mais plutôt d'identifier quelles variables se révèlent importantes, une étape de sélection de variable nous paraissait essentielle. Sélectionner des variables permet en effet d'être plus parcimonieux en terme d'explication d'une variable réponse par un ensemble de variables explicatives (écarter les variables non-informatives), réduire la complexité du modèle, éliminer des variables redondantes entre elles, faciliter l'interprétation. Plusieurs travaux proposent d'étendre les forêts

aléatoires en ajoutant une couche de sélection de variable (Speiser et al., 2019). Nous avons ajouté à l'algorithme proposé par Hajjem et al., 2014 la méthode de Boruta qui est une procédure de sélection de variables robuste et interprétable (Kursa et Rudnicki, 2010). Cette méthode commence par créer k séries de copies permutées de chaque variable explicative. Elle calcule ensuite l'importance de chaque variable explicative (brute et permutée) ; on dit qu'une variable obtient un "hit" lorsque son importance est supérieure à la plus grande importance des variables dupliquées. Comme cette procédure est appliquée k fois, une variable est sélectionnée si son nombre de hits est supérieur au quantile 95% d'une loi binomiale de paramètres k et 0.5 (i.e on compare le nombre de "hits" au nombre de "hits" sous l'hypothèse que chaque "hit" est le fruit du hasard pur, Figure IV.7).

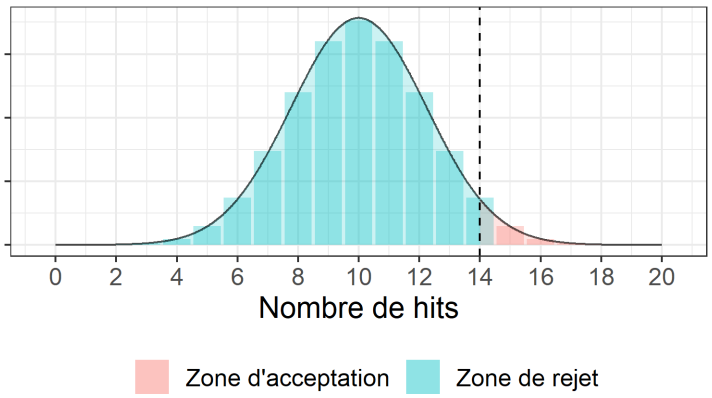


Figure IV.7 – Zones d'acceptation ou de rejet d'une variable explicative dans la procédure Boruta. Si une variable a 14 hits ou plus, on accepte cette variable dans le modèle ($k = 20$ dans la description précédente).

Articulation sélection de variables / imputation multiple

Chaque modèle ajusté sur les 10 versions de chaque jeu de données peut potentiellement sélectionner des variables explicatives différentes pour chacune des versions. J'ai choisi de n'afficher que les importances des variables sélectionnées sur les 10 jeux de données imputés. Ce choix est motivé par ma volonté d'être le plus parcimonieux possible (i.e. limiter le nombre de variables expliquant la performance).

3.1.5 Combinaison des forêts aléatoires et d'un modèle mixte

Définition du modèle

Le modèle de forêt aléatoire avec effet aléatoire se compose i) d'une partie forêt aléatoire modélisant la relation entre la variable réponse et les p facteurs fixes et ii) d'une partie modèle mixte modélisant la relation entre la variable réponse et le facteur expérimentation (constitué de q modalités, une par expérimentation).

La suite de cette section est un peu plus complexe, mais en comprendre précisément le contenu n'est pas nécessaire pour la suite du chapitre.

Formellement, le modèle s'écrit ainsi (Hajjem et al., 2014) :

$$y_i = \underbrace{f(X_i)}_{\text{Forêt aléatoire}} + \underbrace{Z_i b_i}_{\text{Modèle mixte}} + \underbrace{\varepsilon_i}_{\text{Erreur résiduelle}} \tag{3.1}$$

où $y_i = [y_{i1}, \cdots, y_{in_i}]^\top$ est le vecteur de réponses (performance d'une composante du mélange) de taille $n_i \times 1$ pour les n_i individus de l'expérimentation i, X_i est la matrice de taille $n_i \times p$ contenant les valeurs des facteurs fixes (pour l'expérimentation i), Z_i est un vecteur de taille $n_i \times 1$ constitué uniquement de 1, b_i le coefficient du facteur aléatoire lié à l'expérimentation i et ε_i est le vecteur d'erreurs de taille $n_i \times 1$. Le nombre total d'observations est donné par $n = \sum_{i=1}^{n} n_i$.

On suppose les b_i et ε_i indépendants, et $b = (b_i)_{i=1,\ldots,q} \sim \mathcal{N}_q(0, B)$ et $\varepsilon_i \sim \mathcal{N}_{n_i}(0, R_i)$, où B est la matrice de covariance de b avec $B = \gamma^2 I_q$ et R_i est la matrice de covariance des ε_i, supposée diagonale (i.e $R_i = \sigma^2 I_{n_i}$, Hajjem et al., 2014).

Estimation des paramètres

Je décris ici brièvement les différentes étapes d'estimation des paramètres du modèle, on pourra trouver des détails dans la publication de Hajjem et al., 2014

A priori, la fonction f et les coefficients $b_i, (i = 1, \cdots, q)$ ne sont pas connus. Hajjem et al., 2014 utilise donc, pour estimer ces paramètres, un algorithme dit de maximisation de l'espérance ("Expectation Maximization", EM). Les algorithmes EM sont souvent utilisés en statistique quand on dispose d'un jeu de données qu'on suppose généré par un modèle probabiliste dont on cherche à estimer les paramètres.

On note X la matrice de taille $n \times p$ contenant les valeurs des facteurs fixes de toutes les observations. L'algorithme EM estime f et $b_i, (i = 1, \cdots, q)$ via une procédure itérative (Figure IV.8) :

1. Initialisation des paramètres $b_{(0)}$, $B_{(0)}$ et $\sigma^2_{(0)}$
2. Itération r :

(a) Construction du vecteur $y^*_{(r)} := (y^*_{1(r)}, \cdots, y^*_{q(r)})$, avec $y^*_{i(r)} = y_i - Z_i b_{i(r-1)},\ \forall i = 1, \cdots, q$

(b) Ajustement d'une forêt aléatoire sur $y^*_{(r)}$ avec X comme variables explicatives. On obtient un estimé $\hat{f}_{(r)}$ pour f

(c) Ajustement de modèles mixtes sur $y_i - \hat{f}_{(r)}(X_i)$, avec Z_i comme matrice d'incidence des niveaux de l'effet aléatoire. On obtient ainsi des estimés $\hat{b}_{i(r)}$, $\hat{B}_{(r)}$ et $\hat{\sigma^2}_{(r)}$

3. Répétition de l'étape 2 jusqu'à convergence de f et de b (critère de vraisemblance généralisée).

Sélection de variables

J'ai ajouté la procédure de sélection de variables décrite précédemment à l'étape 2(b). Le modèle de régression est mis à jour avec les variables sélectionnées (*via* Boruta) à chaque itération.

Dans la suite du document, $y_i,\ i = 1, \cdots, n$ désignera la i-ème ***observation*** *et $\hat{y}_i,\ i = 1, \cdots, n$ la i-ème* ***valeur ajustée*** *par le modèle.*

Variances et estimateurs des b_i en imputation multiple

Les modèles étant entraînés sur chaque jeu de données imputé, on obtient une valeur $\hat{b_i}^{(m)}$ ainsi qu'une variance associée $\hat{\gamma^2}^{(m)}$. Pour obtenir un estimateur des $\hat{b_i}, (i = 1, \cdots, q)$ sur l'ensemble des imputations ainsi qu'une estimation de leur variance, on applique la règle de Rubin (Marshall et al., 2009) :

$$\bar{b_i} = \frac{1}{M} \sum_{m=1}^{M} \hat{b_i}^{(m)}, \quad \forall i = 1, \cdots, q \tag{3.2}$$

et

$$V = \underbrace{\frac{1}{M} \sum_{m=1}^{M} \hat{\gamma^2}^{(m)}}_{\substack{\text{Moyenne des variances} \\ \text{intra-imputations}}} + (1 + \frac{1}{M}) \underbrace{\frac{1}{M-1} \sum_{m=1}^{M} (\bar{b_i} - \hat{b_i}^{(m)})^2}_{\text{Variance inter-imputations}} \tag{3.3}$$

L'équation (3.2) indique que le meilleur estimé du coefficient correspond à la moyenne des coefficients obtenus au travers des différentes imputations. L'équation (3.3) indique

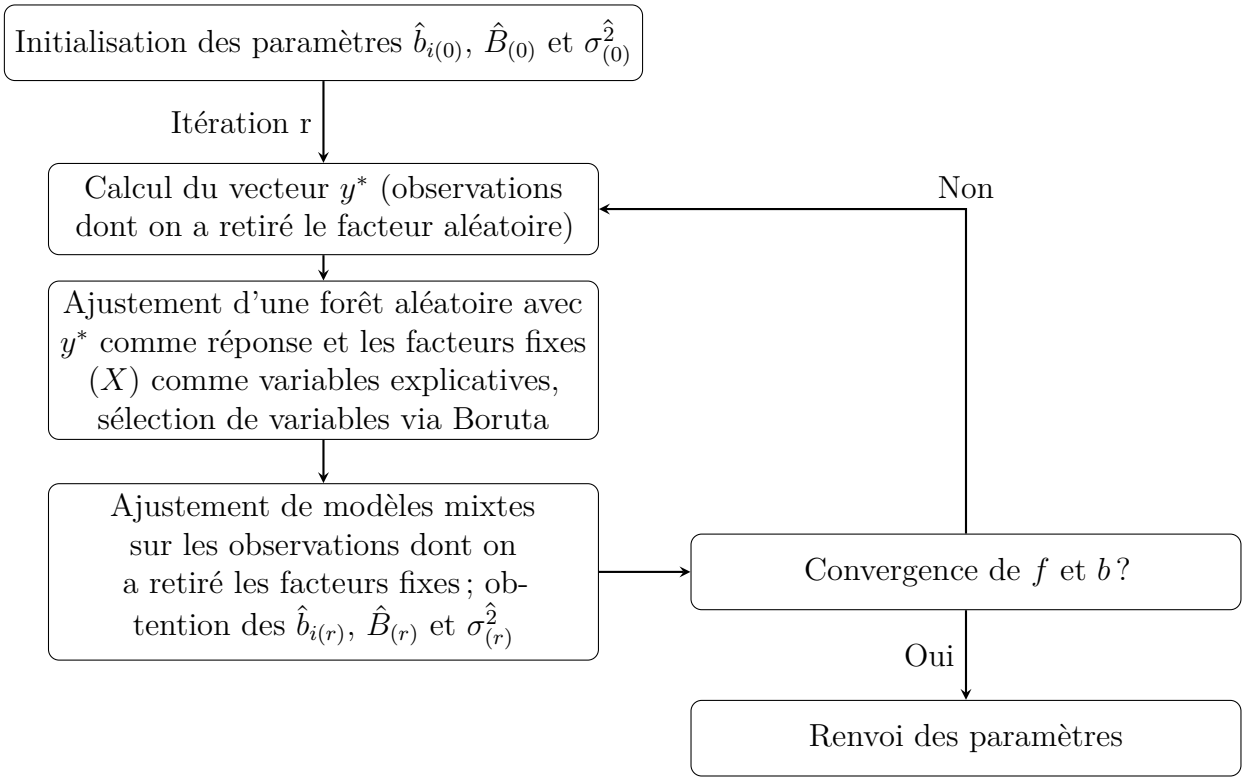


Figure IV.8 – Schéma simplifié l'estimation des paramètres du modèle de forêt aléatoire avec facteur aléatoire ; r correspond au numéro de l'itération

que l'estimateur de la variance des coefficients aléatoires est composé de i) la moyenne des variances intra-imputations (premier terme) et ii) une variance inter-imputations (second terme) dont la contribution est augmentée d'un terme $\frac{1}{M}$ pour tenir compte du fait qu'on a un nombre fini d'imputations. On peut ainsi évaluer la proportion de la variance totale due aux imputations en divisant le second terme de l'équation par V.

3.1.6 Récapitulatif des différentes étapes de modélisation

Nous avons ainsi vu que plusieurs étapes de modélisation se succèdent dans notre démarche (Figure IV.9). En partant des données relatives aux deux mélanges d'espèces considérés, je réduis d'une part la dimension des courbes de hauteur et de biomasse via la procédure décrite dans le Chapitre II. Puis, je réalise une imputation des données manquantes qui me donne 10 versions du jeu de données. Je calcule les variables explicatives décrites dans le Chapitre IV.2. J'ajuste ensuite les modèles décrits ci-dessus i) *via* une procédure de validation croisée pour évaluer la capacité prédictive des modèles et ii) sur l'ensemble du jeu de données.

3.2 Évaluation des modèles

3.2.1 Sens des relations variables réponses / variables explicatives

L'importance en soi n'illustre pas le sens des corrélations existant entre les variables explicatives et les variables réponses (i.e. est-ce que l'augmentation d'une variable explicative entraîne l'augmentation ou la diminution de la variable réponse ?). Les relations modélisées par la forêt aléatoire entre les variables ne sont pas forcément linéaires. Pour qualifier le sens de variation, j'ai donc considéré le signe de la corrélation de Kendall (τ) entre les variables explicatives et les variables réponses. Ce coefficient de corrélation est basé sur les rangs plutôt que les valeurs des deux variables dont on étudie le lien.

J'ai visualisé les relations entre variables réponses et variables sélectionnées par le modèle dans le but de les analyser plus finement. J'ai notamment cherché, via ces visualisations, à distinguer un potentiel effet expérimentation / cultivar / fertilisation azotée sur les facteurs fixes.

3.2.2 Qualité de l'ajustement

Nous avons évalué la qualité de l'ajustement des modèles en utilisant l'erreur quadratique moyenne (RMSE, peut être vue comme une moyenne des erreurs, en donnant

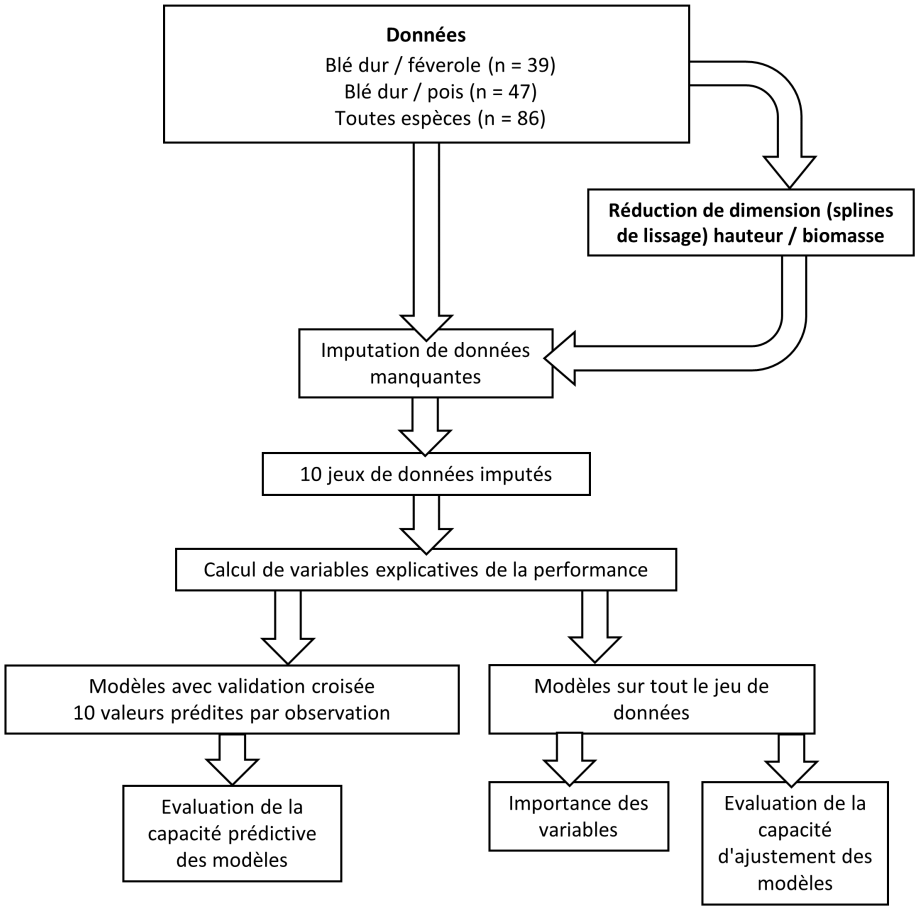


Figure IV.9 – Schéma des différentes étapes de modélisation

un poids plus grand aux grandes erreurs d'ajustement). Pour un modèle donné, la RMSE est donnée par :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Par ailleurs, pour éviter les effets d'échelles (les gammes de variation des variables de sortie diffèrent et donc les RMSE de ces variables), nous avons calculé les RMSE normalisées par la moyenne ($nRMSE$) :

$$\text{nRMSE} = \frac{\text{RMSE}}{\bar{y}}$$

Plus un modèle a des $(n)RMSE$ proches de 0, plus il est performant (en terme d'ajustement ou de prédiction).

Pour évaluer la qualité des modèles en dehors des observations sur lesquelles ils ont appris, nous avons également ajusté des modèles en prenant les $\frac{3}{4}$ des unités expérimentales de chacune des expérimentations considérées (ensemble d'apprentissage) et calculé les métriques précédemment citées sur l'ensemble d'apprentissage et sur le $\frac{1}{4}$ des unités expérimentales restantes (ensemble de validation). En répétant cette opération quatre fois (procédure de validation croisée) et en moyennant l'erreur de prédiction obtenue, on obtient une estimation de la qualité de prédiction du modèle en dehors des données sur lesquelles il a appris.

Décomposition de la variance

L'influence du processus d'imputation sur la variabilité globale a été quantifiée en comparant la variance entre les imputations à la variance globale des résidus. Plus précisément, la proportion de la variance due au processus d'imputation a été calculée comme suit :

$$\frac{\sum_{m=1}^{M}\sum_{i=1}^{n}(\hat{y}_i^{(m)} - \hat{y}_i)^2}{\sum_{m=1}^{M}\sum_{i=1}^{n}(\hat{y}_i^{(m)} - y_i)^2}$$

où M est le nombre de jeux de données imputés, y_i est la i-ième observation, $\hat{y}_i^{(m)}$ sont les valeurs ajustées pour l'imputation m, $\hat{y}_i = \frac{\sum_{m=1}^{M}\hat{y}_i^{(m)}}{M}$ est la moyenne des valeurs ajustées entre les M imputations.

Les détails des calculs sont décrits dans les Annexes.

3.3 Résultats

3.3.1 Performance globale des modèles

La qualité d'ajustement globale des modèles (12 modèles au total : trois types de mélanges et quatre variables de sortie) est satisfaisante. Les erreurs quadratiques normalisées (nRMSE) sont faibles, surtout dans le cas des modèles entraînés sur l'ensemble des données (Figure IV.10). Cette performance globale illustre bien la

bonne capacité d'ajustement des forêts aléatoires, notamment avec le facteur aléatoire additionnel prenant en compte l'effet expérimentation. On observe un bon alignement des valeurs observées et ajustées (Figure IV.11). La différence entre la performance des modèles en validation croisée et en utilisant toutes les données indique un potentiel surajustement du modèle (i.e. quand le modèle serait trop dépendant des conditions d'apprentissage).

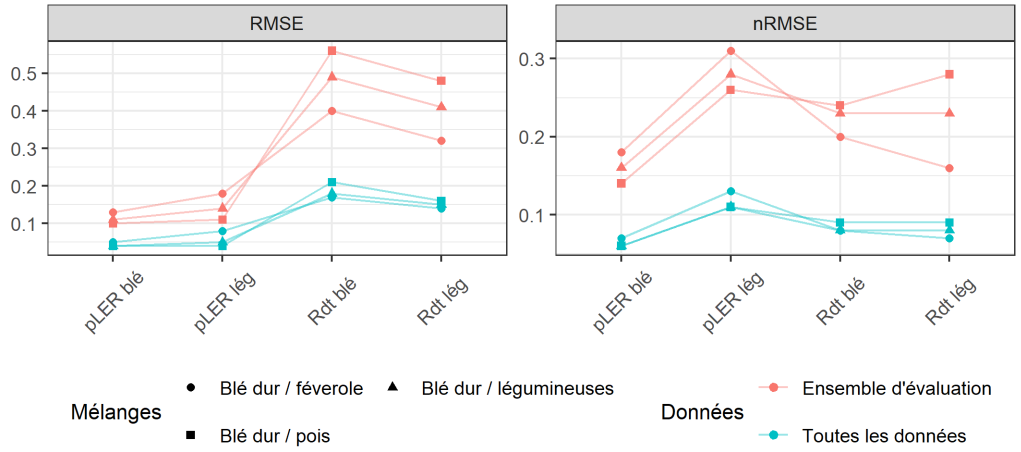


Figure IV.10 – Qualité d'ajustement des différents modèles. Rdt : Rendement ; lég : légumineuse ; pLER : Partial Land Equivalent Ratio. (n)RMSE : (Normalized) Root Mean Squared Error

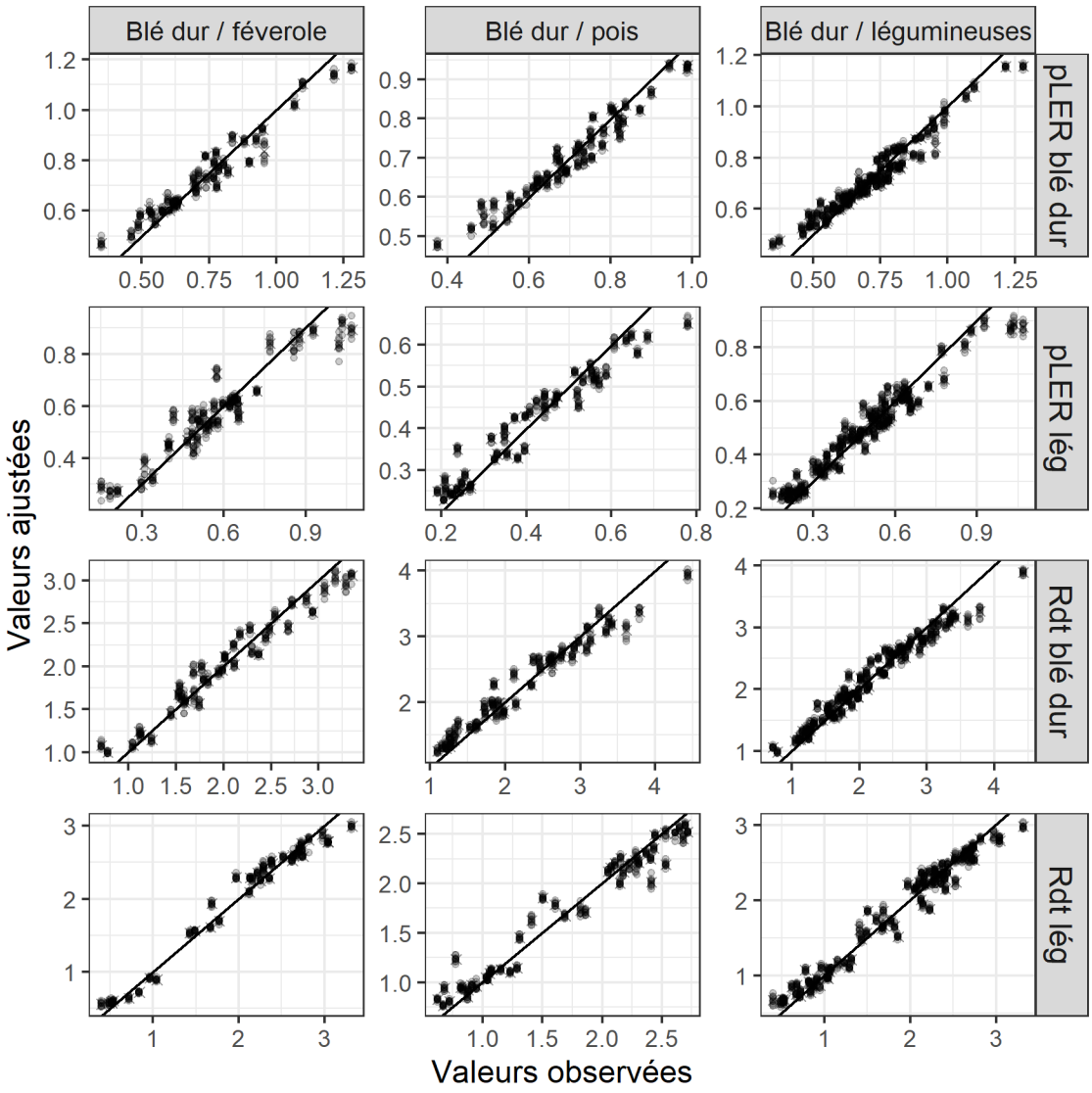


Figure IV.11 – Valeurs ajustées vs observées pour les différents modèles, entraînés sur l'ensemble des données. Les droites d'équation $y = x$ sont affichées. Chaque point représente une valeur ajustée (une pour chaque imputation, 10 pour chaque observation) et les croix correspondent aux valeurs ajustées moyennes (parmi les 10 imputations). Rdt : Rendement ; lég : légumineuse ; pLER : Partial Land Equivalent Ratio

Le poids de l'effet expérimentation dans les différents modèles semble assez faible, puisque les intervalles de confiance autour des coefficients (b_i) du facteur expérimentation contiennent toujours zéro (Figure IV.12). Cela peut sembler surprenant,

puisque l'on attendait un effet fort de l'expérimentation sur les variables de sortie. Une hypothèse plausible est que l'effet expérimentation se reflète en partie dans les facteurs fixes, donnant ainsi peu de poids au facteur aléatoire censé prendre en compte l'effet expérimentation. Les valeurs de ces coefficients peuvent se voir comme une correction de la variable réponse liée à l'expérimentation. Ces valeurs sont donc logiquement liées à la variable réponse en terme d'ordre de grandeur.

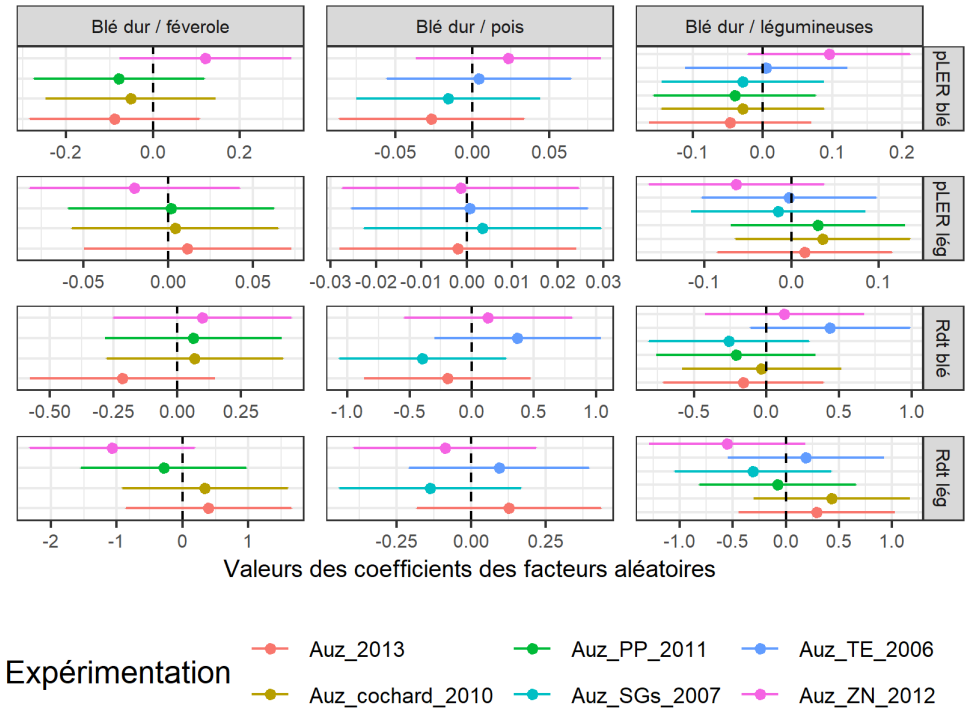


Figure IV.12 – Valeurs moyennes et intervalles de confiance des coefficient du facteur aléatoire (b) de chaque modèle. Les barres d'erreur correspondent aux intervalles de confiance ($\bar{b} \pm 2\sqrt{V}$). Rdt : Rendement ; lég : légumineuse ; pLER : Partial Land Equivalent Ratio.

Quel que soit le modèle, la part de variance du nuage de points des valeurs ajustées vs observées due au processus d'imputation reste faible (<5%, Tableau IV.5). Cet effet est également modeste quand il est comparé à l'importance estimée des variables (Figure IV.13). Enfin, la part de la variance inter-imputations dans la variance des coefficients des facteurs aléatoires est faible (1% en moyenne).

Mélanges	pLER blé	pLER lég	Rdt blé	Rdt lég
Blé dur / légumineuses	3.3%	2.9%	3.9%	3.4%
Blé dur / féverole	4.2%	4.3%	4.4%	3.6%
Blé dur / pois	2.6%	2%	4.2%	2.6%

Table IV.5 – Part de la variance due aux imputations ; Rdt : Rendement ; lég : légumineuse ; pLER : Partial Land Equivalent Ratio

Figure IV.13 – Importances et écart-type (lié à l'imputation) des variables de chaque modèle. Rdt : Rendement ; lég : légumineuse (féverole ou pois) ; pLER : Partial Land Equivalent Ratio ; NEG (respectivement POS) : $\tau < 0$ (respectivement > 0) entre variable réponse et variable explicative ; NA : τ non-calculé car variable qualitative

3.3.2 Interprétation des modèles

Pour identifier et interpréter les variables explicatives des rendements bruts et relatifs (pLER) des mélanges, nous utiliserons une visualisation synthétique de l'importance et du type d'effet (lien positif ou négatif) des variables sélectionnées. Nous effectuerons cette analyse par mélange (blé/féverole, blé/pois, blé/légumineuse). La relation précise entre une variable explicative et une variable de réponse sera détaillée en second temps dans une série de graphiques bivariés, pour analyser la linéarité de la relation et la distribution des effets environnementaux.

3.3.2.1 Mélanges blé dur/féverole (Figure IV.14)

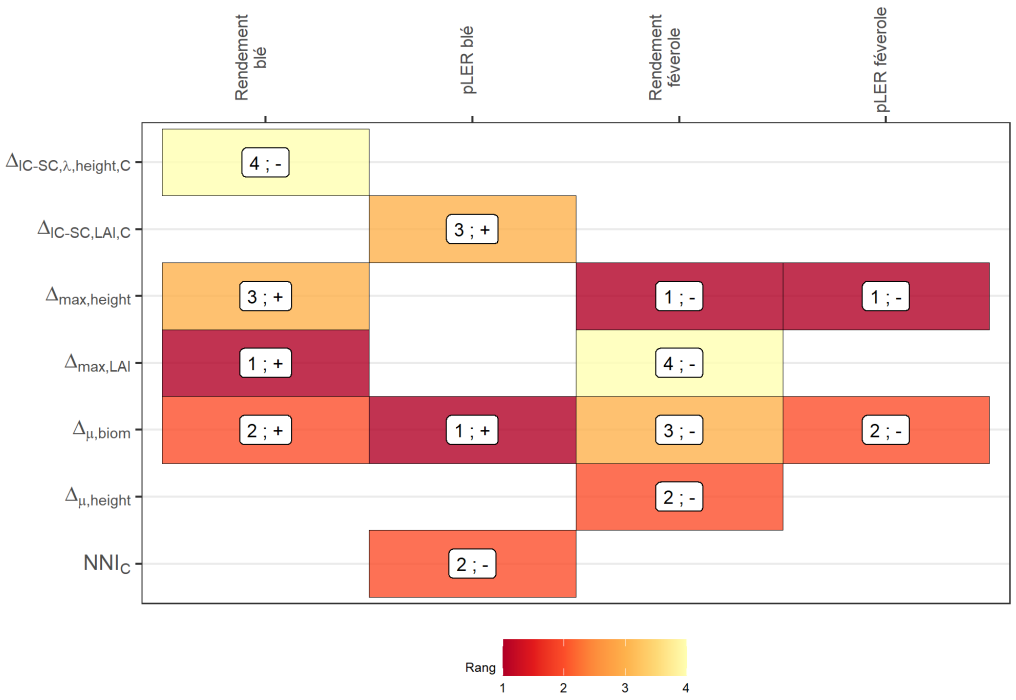


Figure IV.14 – Résumé des variables sélectionnées dans les modèles blé dur/féverole. Chaque colonne concerne un modèle (une variable de performance), chaque ligne correspond à une variable explicative. La couleur des rectangles ainsi que le chiffre inscrit indiquent le rang en terme d'importance de la variable au sein du modèle, le signe correspond au signe du τ de Kendall entre la variable de performance et la variable explicative.

Focus sur les rendements bruts des deux espèces

Les trois premières variables impactant les rendements bruts du blé et de la féverole

sont liées aux interactions entre les deux espèces au sein du mélange. Deux variables caractérisant le niveau de développement relatif d'une espèce (différences de LAI maximal $\Delta_{max,LAI}$, et de vitesse de croissance en biomasse $\Delta_{\mu,biom}$) sont communes aux deux modèles mais agissent dans le sens opposé sur les rendements des deux composantes du mélange. On montre ainsi que les variables induisant la dominance d'une espèce sont liées à une augmentation de son rendement (coefficient de corrélation positif) au détriment de l'autre espèce (coefficient de corrélation négatif). Pour ce mélange, ce comportement correspond à un fort développement du blé par rapport à la féverole (Figure IV.14).

Regardons de manière plus précise les relations entre les rendements du blé (Figure IV.15) et de la féverole (Figure IV.16) et ces deux variables explicatives (Tableau IV.4). Leur effet sur le rendement est relativement linéaire, positif pour le blé ($\rho = 0.83$ et 0.78 pour $\Delta_{max,LAI}$ et $\Delta_{\mu,biom}$ respectivement) et, inversement, négatif pour la féverole ($\rho = -0.61$ et -0.74 pour $\Delta_{max,LAI}$ et $\Delta_{\mu,biom}$ respectivement).

Pour les rendements du blé, l'effet de l'expérimentation n'est pas clairement visible dans ces relations (i.e. les points sont bien répartis, sans structure marquée, le long de la régression). A contrario, pour le rendement de la féverole, une expérimentation ("Auz_ZN_2012") ressort comme ayant des rendements plus faibles que les trois autres expérimentations. Or, la valeur du facteur aléatoire (b_i) est particulièrement négative pour cette expérimentation (Figure IV.12), ce qui indiquerait de manière cohérente un effet prépondérant de l'environnement (climat, sol).

La différence de surface foliaire spécifique (SLA) entre les deux espèces au sein du mélange semble impacter le rendement de la féverole de manière globale ($\rho = 0.64$), avec un fort effet expérimentation. A posteriori, cette variable nous semble peu pertinente en terme de description des interactions entre espèces existant au sein du mélange. En effet, la SLA est un trait qui reflète l'adaptation des plantes à leur environnement lumineux. C'est donc un trait intégratif, résultant des interactions entre plantes. L'interprétation que nous proposons est donc plus incertaine, d'autant plus que la SLA dépend aussi de l'espèce considérée (avec cependant des ordres de grandeurs proches pour le blé et la féverole). Néanmoins, quand la différence de SLA entre le blé et la féverole est négative (i.e. SLA de la féverole supérieure à celle du blé), le rendement de la féverole est plus faible. Or, une SLA élevée correspondrait à une adaptation à une forte compétition pour la lumière (Evans et Poorter, 2001), avec des feuilles ayant une surface plus importante (pour une biomasse plus faible), dans l'optique de capter plus de lumière. Ceci semblerait cohérent avec la dominance du blé sur la féverole.

La hauteur est également sélectionnée comme une variable importante pour expliquer le rendement du blé, que ce soit les différences au sein du mélange entre les deux espèces ($\Delta_{max,height}$) ou entre culture pure et associée pour le blé ($\Delta_{IC-SC,\lambda,height}$). Comme pour le LAI et la biomasse, on observe une relation positive et globalement linéaire entre le rendement du blé et la différence de hauteur entre le blé et la féverole. On observe par contre une corrélation négative entre le rendement du blé et son décalage de démarrage de croissance en hauteur entre la culture pure et le mélange ($\Delta_{IC-SC,\lambda,height}$), ce qui semble indiquer une influence du mélange sur la phénologie et *in fine* le rendement. Par contre, pour cette variable, il semble y avoir une structure plus marquée du nuage de points, avec un regroupement par expérimentation.

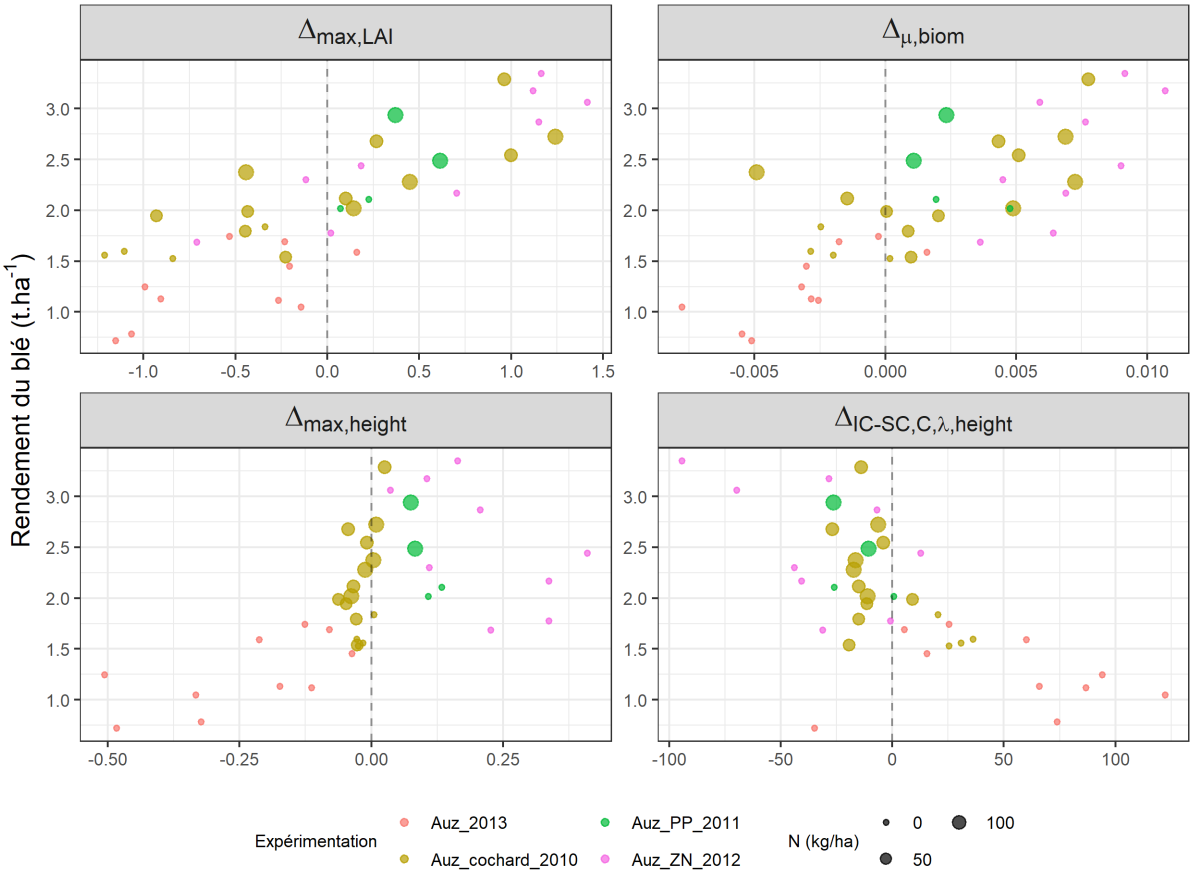


Figure IV.15 – Relations bivariées entre les 4 premières variables explicatives du rendement du blé dans les mélanges blé/féverole. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

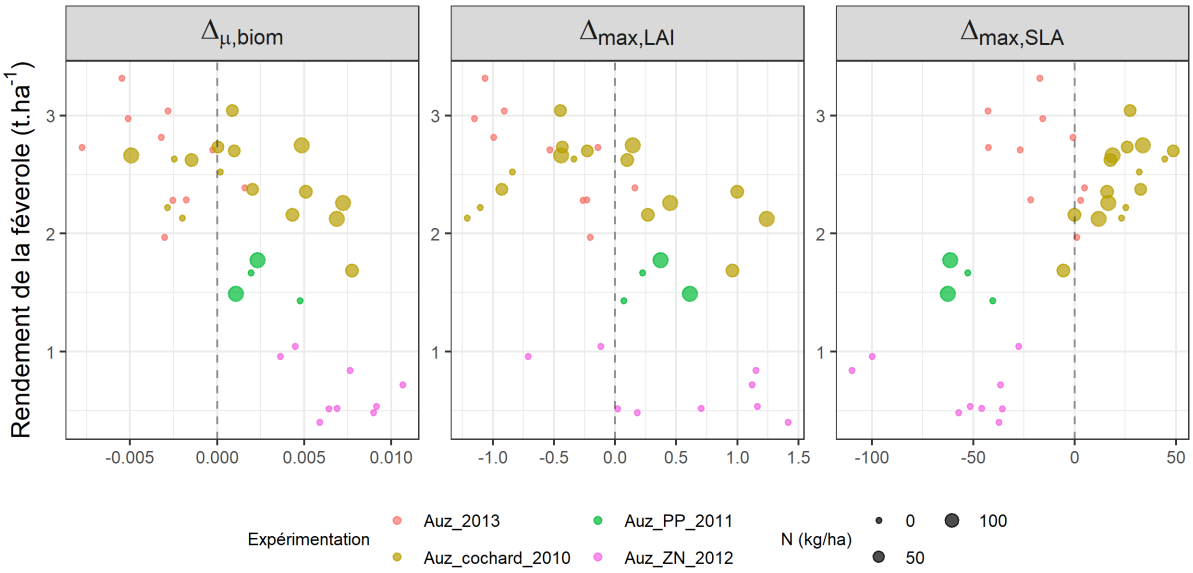


Figure IV.16 – Relations bivariées entre les 3 variables explicatives du rendement de la féverole dans les mélanges blé/féverole. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

Enfin, le statut azoté de la céréale n'est pas sélectionné comme variable explicative des rendements des deux composantes du mélange (Figure IV.14). Nous supposons que pour ce mélange, le fait que cette variable intervienne seulement sur le rendement relatif (pLER) traduit son effet sur la culture pure correspondante (cf. commentaire détaillé ci-après).

Focus sur les rendements relatifs (pLER) des deux espèces

Concernant les rendements relatifs du blé, les trois variables explicatives qui ressortent sont l'indice de nutrition azotée du blé (INN), la différence de croissance en biomasse entre le blé et la féverole ($\Delta_{\mu,biom}$) et la différence de LAI maximal entre cultures associée et pure ($\Delta_{IC-SC,LAI,C}$). La première relation est assez claire et cohérente : le rendement relatif est un ratio entre le rendement en culture associée et le rendement en culture pure ($\frac{Y_C}{S_C}$, Tableau IV.1). En conditions non-fertilisées, le rendement de la céréale en culture pure ($2.3 \pm 0.56\ t.ha^{-1}$) est plus faible qu'en conditions fertilisées ($3.6 \pm 0.51\ t.ha^{-1}$). Ainsi, on peut avoir un rendement relatif élevé associé à un rendement brut faible. Néanmoins, quelles que soient les conditions de fertilisation, le rendement relatif de la céréale est supérieur à 0.5 (36 cas sur 39), indiquant une meilleure performance en culture associée qu'en culture pure. La relation entre la différence de croissance de biomasse et le rendement relatif du blé semble indiquer que le blé arrive à maintenir son rendement en culture associée uniquement s'il est suffisamment compétitif (i.e. croître rapidement) face à la féverole (légumineuse relativement compétitive ; Guinet et al., 2018). Enfin, la dernière relation sépare les céréales ayant une surface foliaire maximale supérieure en culture pure (partie gauche du graphique) de celles pour qui cette surface est supérieure en culture associée (partie droite du graphique). La forme de cette relation est moins robuste que les deux précédentes.

Comme on l'a constaté avec les rendements bruts, l'expérimentation "Auz_ZN_2012" présente des rendements relatifs faibles pour la féverole (0.34 ± 0.15). Pour la relation entre la différence de hauteur maximale entre les deux espèces au sein du mélange ($\Delta_{max,height}$) et le pLER, l'expérimentation d'Auzeville 2013 ressort, où la féverole est plus grande que le blé. On retrouve cette relation pour la différence de croissance en biomasse entre les deux espèces ($\Delta_{\mu,biom}$). Le nombre d'unités expérimentales où la féverole a eu un rendement relatif supérieur en culture associée est plus faible que pour la céréale. Ceci et les relations antagonistes de certains prédicteurs ($\Delta_{\mu,biom}$, $\Delta_{max,height}$) suggèrent que la meilleure performance de la céréale s'est parfois faite au détriment de la féverole.

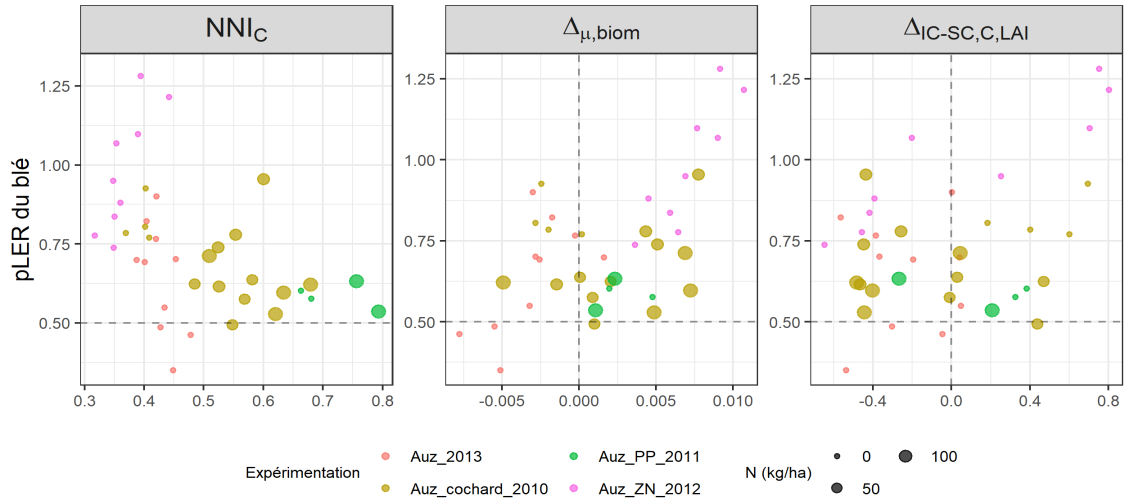


Figure IV.17 – Relations bivariées entre les 3 variables explicatives du pLER du blé dans les mélanges blé/féverole. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

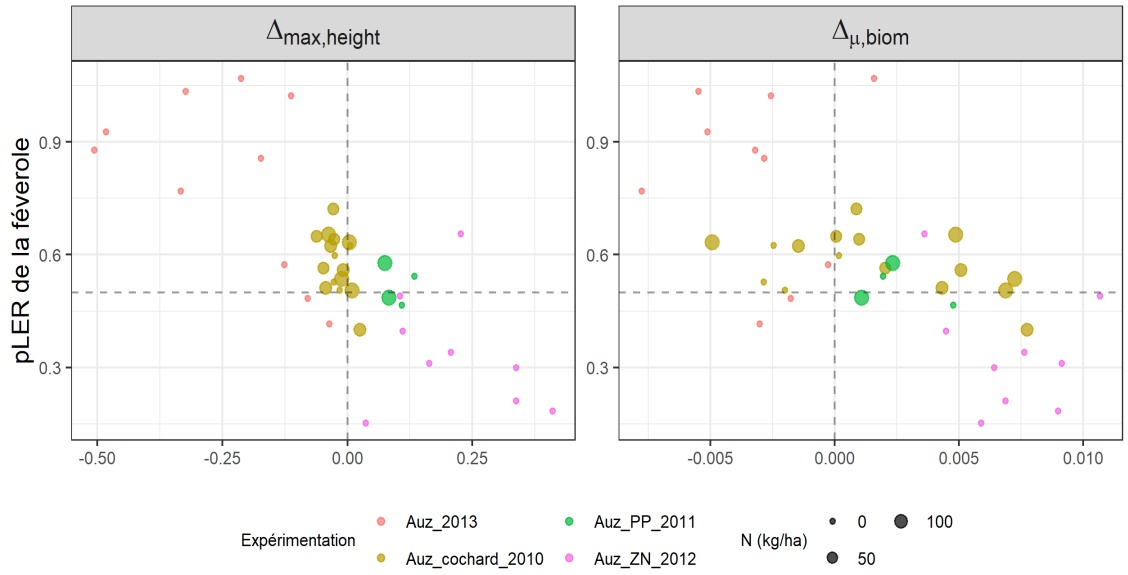


Figure IV.18 – Relations bivariées entre les 2 variables explicatives du rendement relatif (pLER) de la féverole dans les mélanges blé/féverole. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

3.3.2.2 Mélanges blé dur / pois (Figure IV.19)

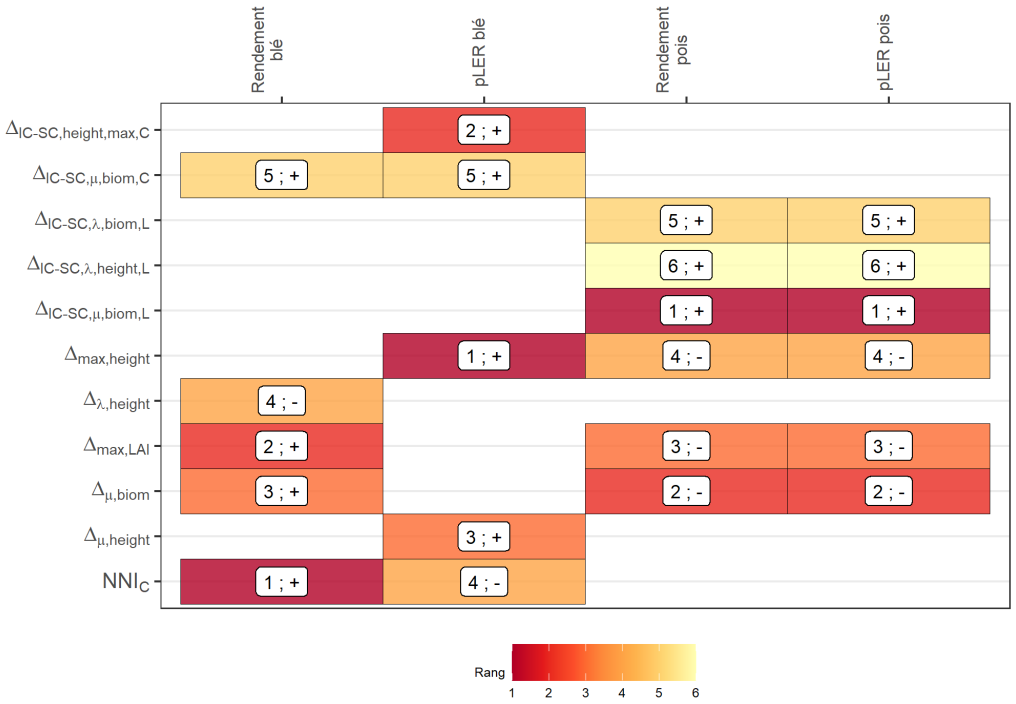


Figure IV.19 – Résumé des variables sélectionnées dans les modèles blé dur/pois. Chaque colonne concerne un modèle (une variable de performance), chaque ligne correspond à une variable explicative. La couleur des rectangles ainsi que le chiffre inscrit indique le rang en terme d'importance de la variable au sein du modèle, le signe correspond au signe du τ de Kendall entre la variable de performance et la variable explicative.

Pour les mélanges blé dur/pois, le statut azoté du blé (INN) est sélectionné comme variable prédominante impactant le rendement brut du blé. Comme pour les mélanges blé dur/féverole, les différences de LAI maximal ($\Delta_{max,LAI}$) et de vitesse de croissance en biomasse ($\Delta_{\mu,biom}$) entre les deux espèces au sein du mélange sont liées aux rendements des deux espèces mais agissent dans un sens opposé, suggérant que si une espèce prend le dessus, c'est au détriment de l'autre espèce. Contrairement au rendement brut, le rendement relatif du blé est principalement impacté par des variables liées à la hauteur alors que les rendements bruts et relatifs du pois sont liés strictement aux mêmes variables, et ce dans le même sens de variation et dans le même ordre.

Focus sur les rendements bruts des deux espèces

Regardons de manière plus précise les relations entre le rendement du blé (Figure IV.20) et du pois (Figure IV.21) et les variables explicatives liées aux différences de LAI maximal et de vitesse de croissance en biomasse. Leur effet sur le rendement est linéaire, positif pour le blé ($\rho = 0.52$ et 0.71 pour $\Delta_{max,LAI}$ et $\Delta_{\mu,biom}$ respectivement) et, inversement, négatif pour le pois ($\rho = -0.68$ et -0.77 pour $\Delta_{max,LAI}$ et $\Delta_{\mu,biom}$ respectivement).

On observe un possible effet expérimentation sur la différence de LAI maximal ($\Delta_{max,LAI}$) où les valeurs sont plus élevées pour l'expérimentation "Auz_SGs_2007". Les rendements du blé sont cependant équivalents entre expérimentations. La fertilisation semble influencer les interactions entre les deux espèces illustrées par ces variables puisque le signe de $\Delta_{max,LAI}$ et $\Delta_{\mu,biom}$ change avec la fertilisation (notamment dans les expérimentations "Auz_2013" et "Auz_SGs_2007").

Une relation moins robuste semble exister entre le rendement du blé et le décalage de démarrage de la croissance entre le blé et le pois ($\rho = -0.23$). L'expérimentation "Auz_ZN_2012" est particulière puisque dans cette expérimentation, le pois s'est beaucoup moins développé que le blé (Figure IV.22). La relation entre le rendement du pois et la différence de hauteur maximale entre les deux espèces au sein du mélange ($\Delta_{max,height}$) suggère un effet seuil (Figure IV.21), puisque pour l'expérimentation "Auz_2013", les rendements du pois sont stables quand le pois est plus grand que le blé, et décroissent au-delà (influence négative du blé). Quand la hauteur maximale du pois est inférieure à celle du blé, on observe une relation négative liant différence de hauteur et rendement du pois, soulignant l'importance de la compétition pour l'accès à la lumière au sein de ce mélange.

Au sein de chaque expérimentation, on note un effet seuil de l'indice de stress azoté (INN) lié à la fertilisation : au-delà de 0.6, les rendements du blé au sein de chaque expérimentation sont plus élevés. L'effet de la fertilisation, et donc de l'INN, sur le rendement du blé est cohérent : plus le statut azoté de la céréale est élevé, plus son rendement l'est (au-delà d'un certain seuil). On observe néanmoins, dans les unités expérimentales fertilisées de l'expérimentation "Auz_SGs_2007", une corrélation négative entre INN et rendement du blé ($\rho = -0.33$), indiquant un possible effet seuil sur le rendement du blé.

On observe un effet expérimentation sur la première variable explicative du rendement du pois. La différence des vitesses de croissance en biomasse entre culture associée et culture pure ($\Delta_{IC-SC,\mu,biom}$) est positive à "Auz_ZN_2013" et négative dans les autres expérimentations (Figure IV.21). Le rendement du pois est positivement corrélé à cette variable ($\rho = 0.78$), que ce soit au sein de la relation globale ou au sein de

chaque expérimentation. Cela suggère que l'adaptation au mélange est un processus clé pour le pois : un pois capable d'augmenter sa vitesse de croissance en biomasse en culture associée par rapport à la culture pure est avantagé, avec un rendement qui augmente.

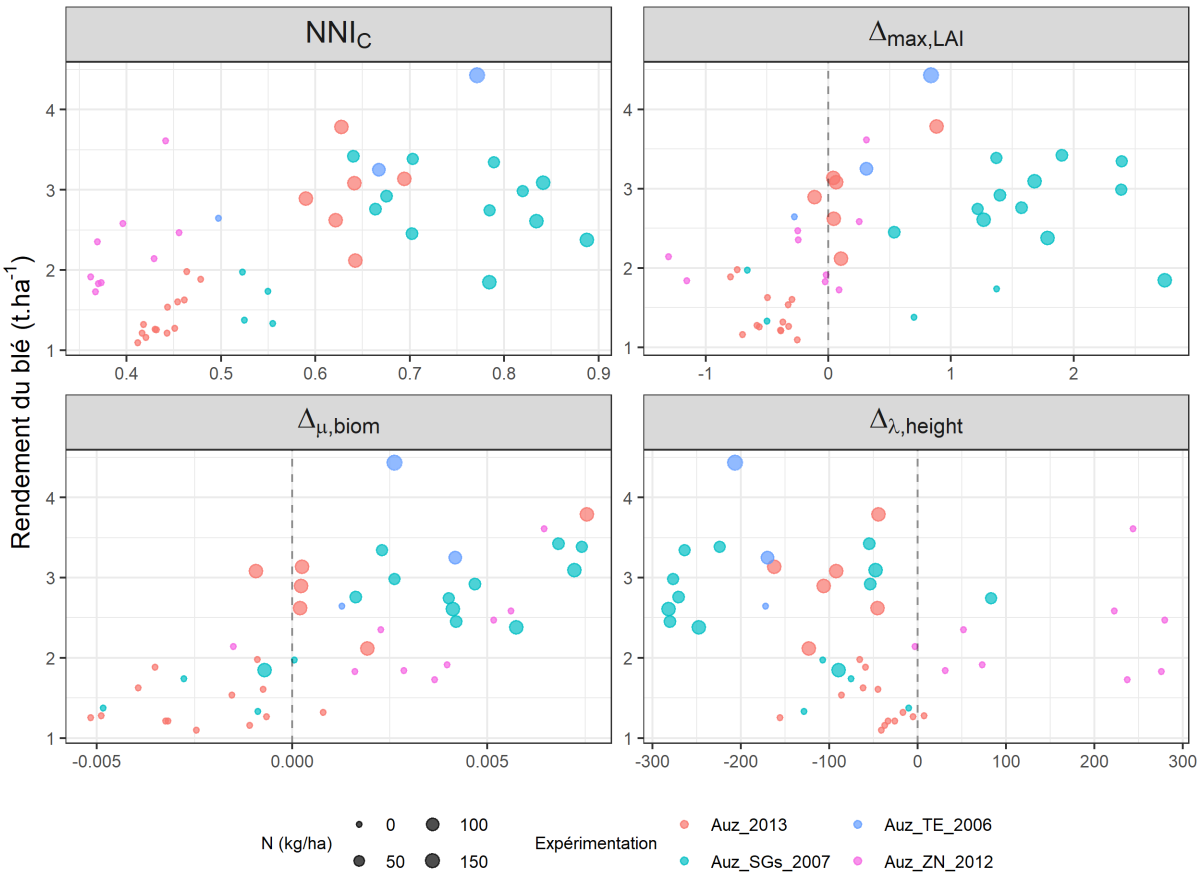


Figure IV.20 – Relations bivariées entre les 4 premières variables explicatives du rendement du blé dans les mélanges blé/pois. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

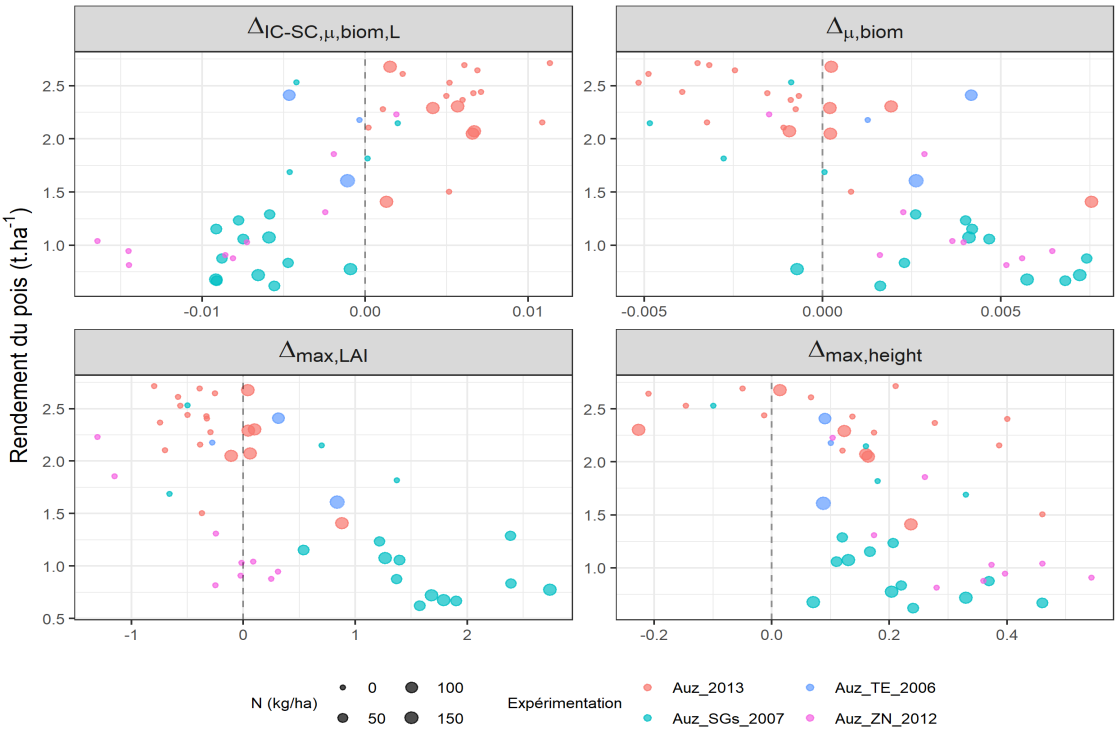


Figure IV.21 – Relations bivariées entre les 4 premières variables explicatives du rendement du pois dans les mélanges blé/pois. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

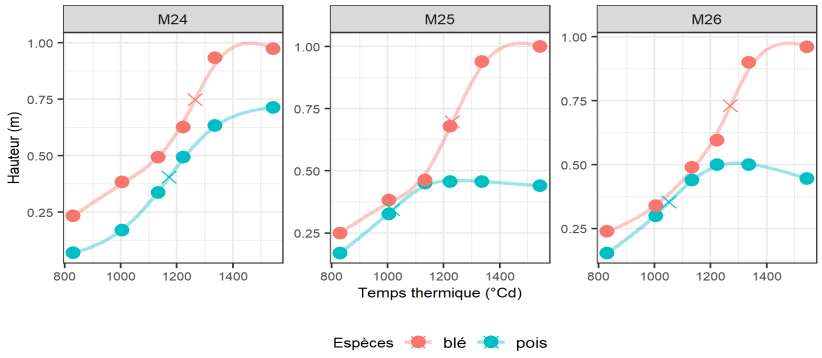


Figure IV.22 – Courbes d'évolution de la hauteur du blé et du pois dans trois unités expérimentales (M24, M25, M26) à Auzeville en 2012.

Focus sur les rendements relatifs (pLER) des deux espèces

Les rendements relatifs du blé sont majoritairement (43 cas sur 47) supérieurs à 0.5, ce qui indique un meilleur rendement du blé en culture associée comparativement aux cultures pures (Figure IV.23). Inversement, les pLER du pois sont majoritairement inférieurs à 0.5, notamment en conditions fertilisées, avec 52% (respectivement 85%) des pLER inférieurs à 0.5 en conditions non-fertilisées (respectivement fertilisées) (Figure IV.24).

Plusieurs variables d'interaction entre espèces au sein du mélange sont liées aux pLER du blé et du pois. La différence de hauteur maximale entre les deux espèces ($\Delta_{max,height}$) est corrélée positivement au pLER du blé ($\rho = 0.63$) et négativement au pLER du pois ($\rho = -0.61$). Ces relations suggèrent, comme pour les mélanges blé/féverole, l'importance de la compétition pour l'accès à la lumière au sein des mélanges blé/pois. La relation entre la différence de vitesse de croissance en hauteur entre les deux espèces au sein du mélange ($\Delta_{\mu,height}$) est principalement tirée par l'expérimentation "Auz_ZN_2012", où le blé a poussé plus vite que le pois, lui assurant un rendement relatif élevé et entraînant une forte compétition sur le pois (Figure IV.22).

Comme pour le rendement brut, des relations négatives sont observées entre le pLER du pois et les différences, entre les deux espèces au sein du mélange, de vitesse de croissance en biomasse ($\Delta_{\mu,biom}$) et de LAI maximal ($\Delta_{max,LAI}$) (ρ = -0.71 et -0.59 respectivement). Ceci souligne encore une fois le rôle de la dominance du blé sur le pois, impactant négativement le rendement de ce dernier. De plus, on observe un effet de la fertilisation sur ces variables (globalement négatives en conditions non-fertilisées et positives en conditions fertilisées), montrant une potentielle perturbation des interactions entre les deux espèces liée à la fertilisation.

Comme pour le rendement brut du pois, son pLER est positivement corrélé à la différence de vitesse de croissance en biomasse entre culture associée et culture pure ($\Delta_{IC-SC,\mu,biom}$). Cela suggère que l'adaptation au mélange est un processus clé pour le pois et que les pois ayant plus subi la compétition du blé (comme à Auzeville en 2012) sont susceptibles d'avoir un rendement relatif plus faible.

La différence de hauteur du blé entre culture pure et associée ($\Delta_{IC-SC,height}$) est clairement impactée par la fertilisation : en culture associée, le blé fertilisé a à peu près la même taille qu'en culture pure (0.5 cm plus grand en moyenne, non-significativement différent de 0 ; Student t-test). En conditions non-fertilisées, le blé en culture associée est plus grand que le blé en culture pure (8 cm en moyenne). Une explication plausible est que le blé en culture associé profite d'une relaxation de la

compétition intraspécifique. En effet, il est associé à du pois, moins compétitif que le blé (i.e. moins de plants de blé par m^2), induisant un rendement en culture associée plus élevé. Cet effet est plus marqué en conditions non-fertilisées, dans lesquels le blé profite de la fixation symbiotique de la légumineuse.

Comme pour dans les mélanges blé dur/féverole, la relation entre le statut azoté (INN) du blé et son pLER est assez claire et cohérente. En conditions non-fertilisées, le rendement du blé en culture pure est plus faible qu'en conditions fertilisées (respectivement 2.37 ± 0.56 $t.ha^{-1}$ et 4.78 ± 0.86 $t.ha^{-1}$). Ainsi, on peut avoir un rendement relatif élevé associé à des rendements bruts faibles.

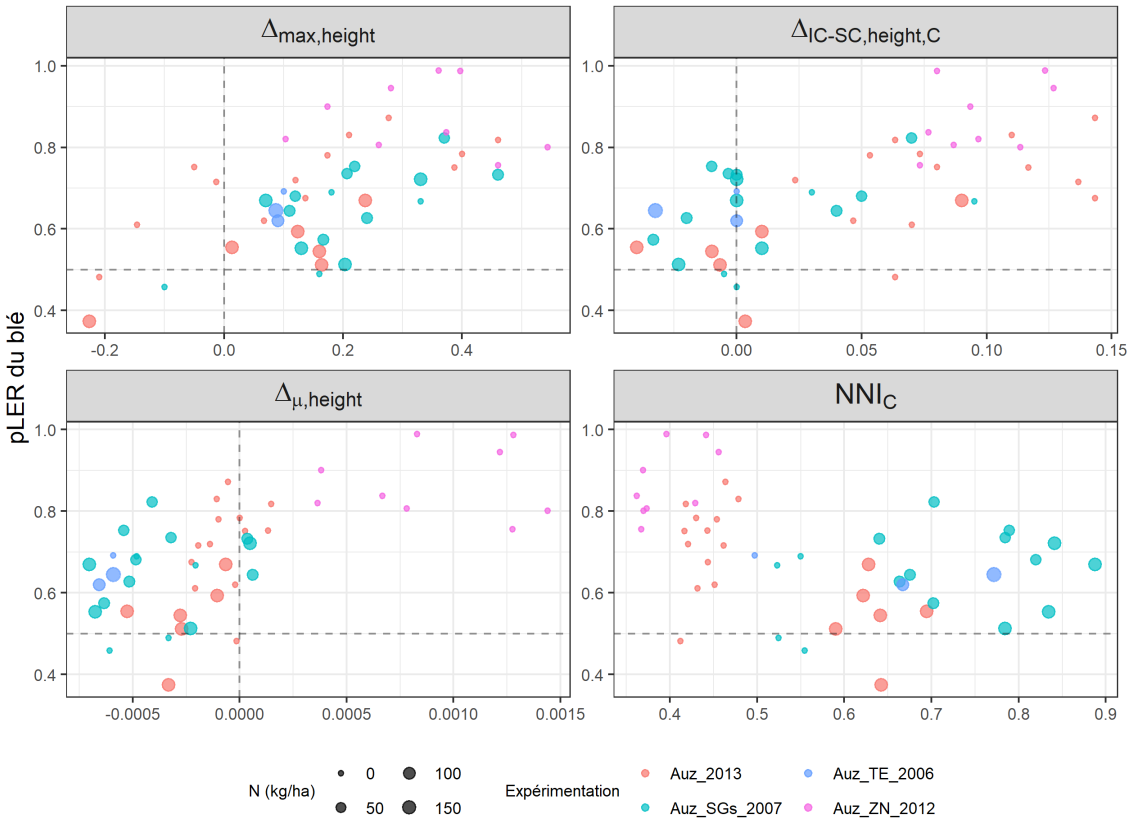


Figure IV.23 – Relations bivariées entre les 4 premières variables explicatives du rendement relatif (pLER) du blé dans les mélanges blé/pois. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

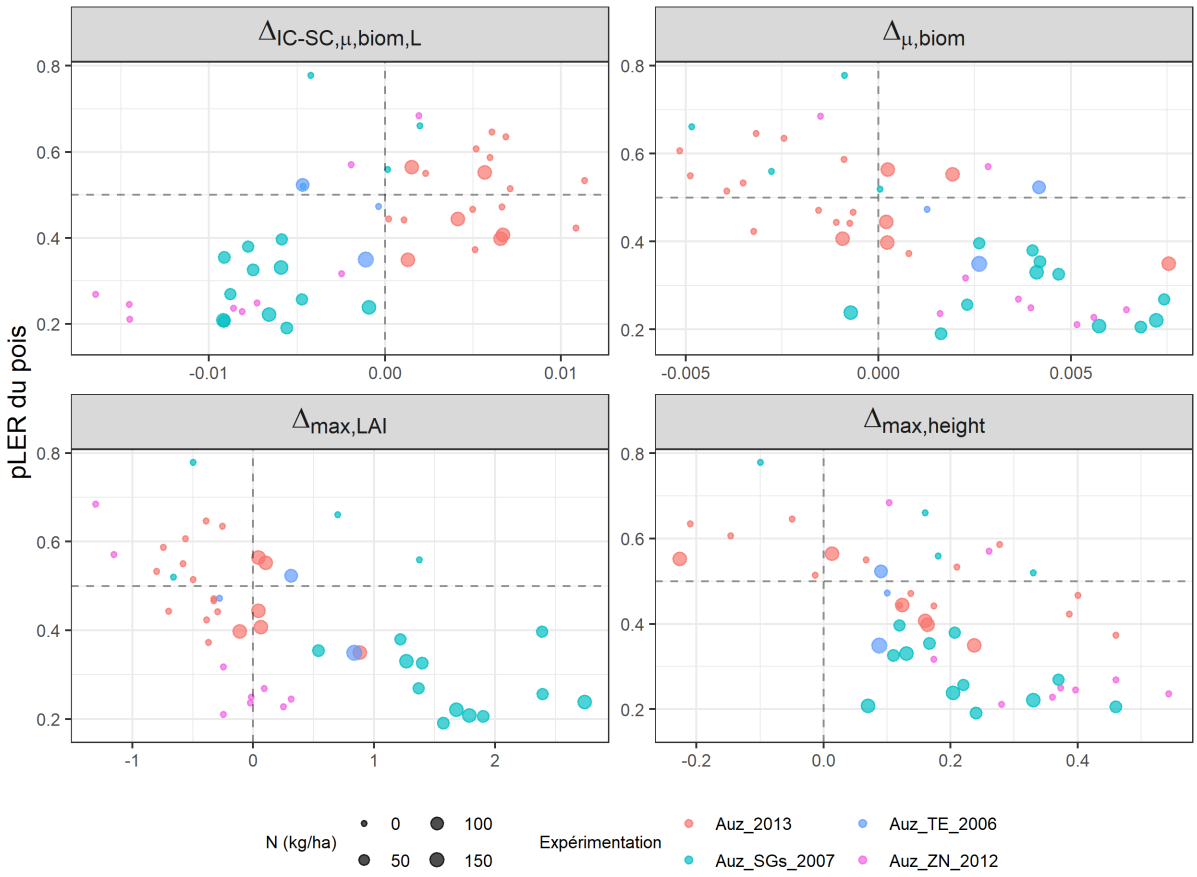


Figure IV.24 – Relations bivariées entre les 4 premières variables explicatives du rendement relatif (pLER) du pois dans les mélanges blé/pois. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

3.3.2.3 Mélanges blé dur / légumineuse (Figure IV.25)

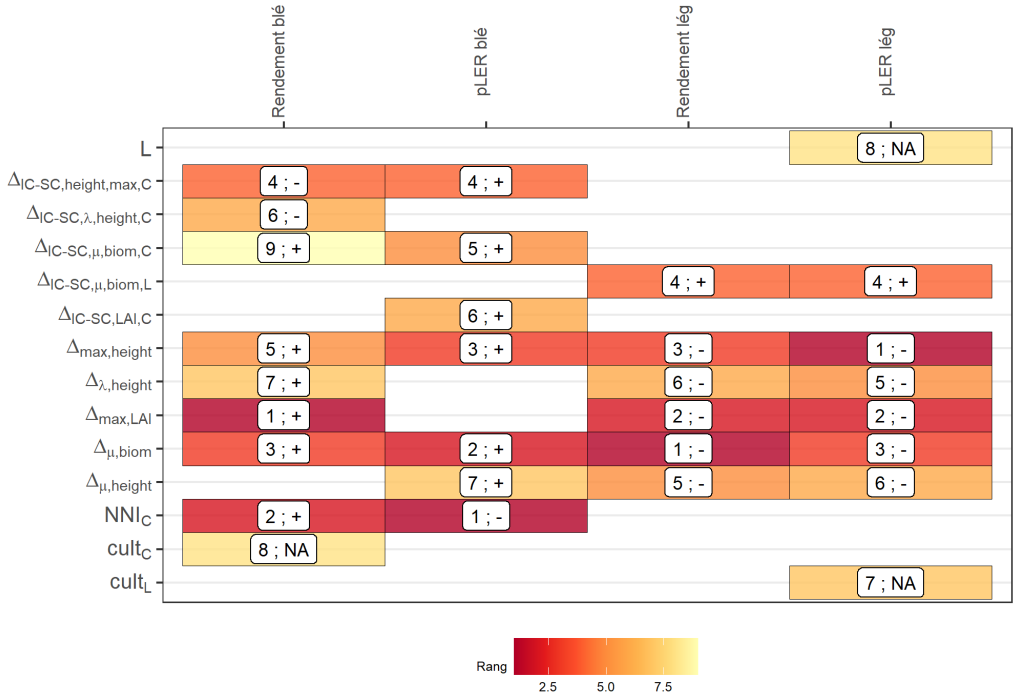


Figure IV.25 – Résumé des variables sélectionnées dans les modèles blé dur/légumineuse (regroupant la féverole et le pois). Chaque colonne concerne un modèle (une variable de performance), chaque ligne correspond à une variable explicative. La couleur des rectangles ainsi que le chiffre inscrit indique le rang en terme d'importance de la variable au sein du modèle, le signe correspond au signe du τ de Kendall entre la variable de performance et la variable explicative. L : légumineuse, $cult_{C/L}$: cultivar de la céréale/légumineuse

Dans les modèles d'ajustement de la performance des mélanges blé dur/légumineuse, on trouve logiquement comme variables explicatives sélectionnées une combinaison de celles sélectionnées dans les modèles d'ajustement de la performance par mélange d'espèces (blé/féverole, blé/pois). Les variables sélectionnées pour les rendements relatifs et bruts de la légumineuse sont presque exclusivement les mêmes. Un point important à souligner est que la variable "légumineuse" ne ressort pas comme étant importante (sauf en 8ème position pour les modèles d'ajustement du rendement relatif de la légumineuse, Figure IV.25). Autrement dit, le fait que ce soit le pois ou la féverole ne ressort pas comme étant important, indiquant une probable dépendance des variables explicatives à l'espèce de la légumineuse. Ainsi, il semblerait que plus que l'espèce (féverole ou pois), c'est la combinaison de valeurs de traits et les différences

de ces valeurs entre le blé et la légumineuse compagne qui sont déterminantes pour expliquer le rendement. Ceci milite pour aborder les espèces et variétés en les regroupant par type fonctionnel (ensemble de variétés partageant des caractéristiques communes).

Focus sur les rendements bruts des deux espèces

Les différences de LAI maximal ($\Delta_{max,LAI}$) et de vitesse de croissance en biomasse ($\Delta_{\mu,biom}$) entre les deux plantes impactent le rendement du blé (ρ = 0.64 et 0.7 respectivement) et de la légumineuse (ρ = -0.64 et -0.72 respectivement) de manière antagoniste. Comme pour le modèle de rendement du pois, la différence de hauteur maximale entre le blé et la légumineuse ($\Delta_{max,height}$) impacte négativement le rendement de la légumineuse ($\rho = -0.68$).

Le statut azoté du blé (INN) est corrélé positivement ($\rho = 0.56$) à son rendement, montrant l'effet positif de la fertilisation sur le rendement du blé.

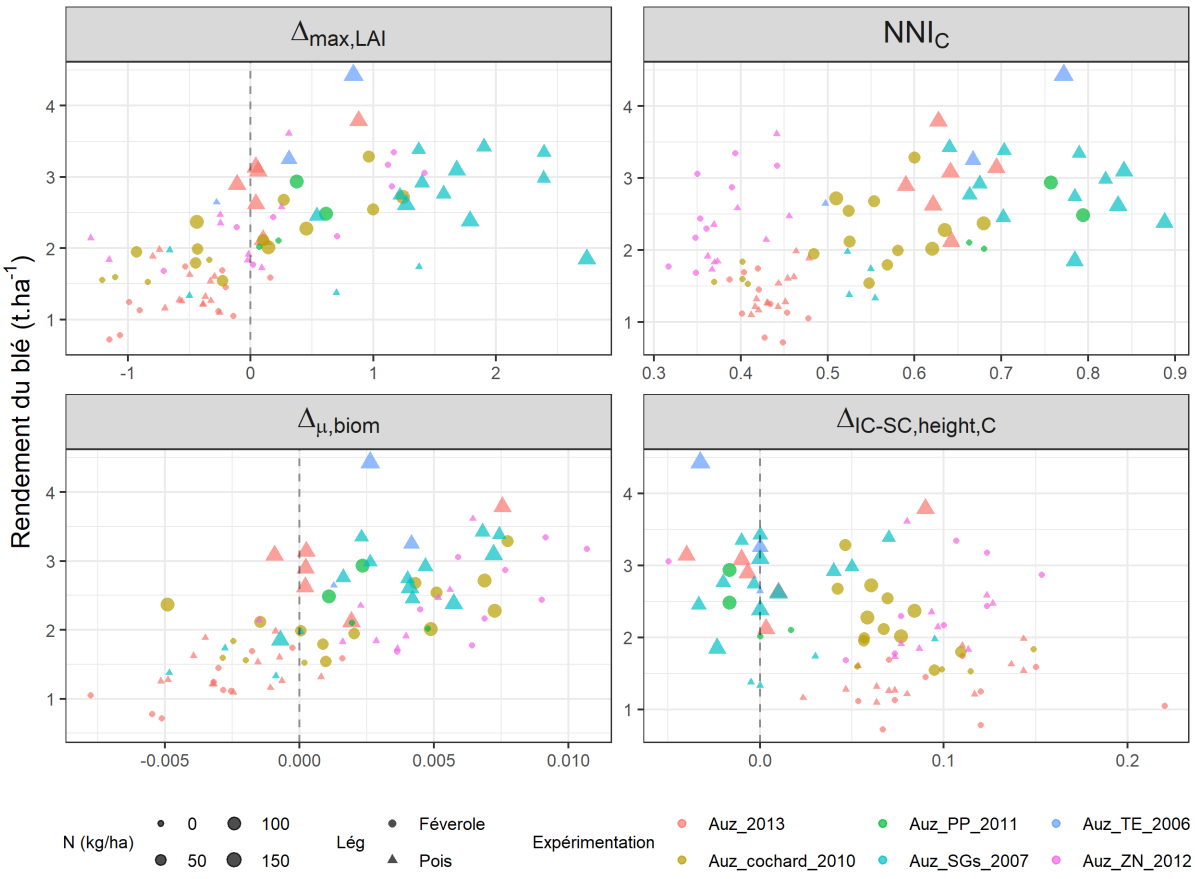


Figure IV.26 – Relations bivariées entre les 4 premières variables explicatives du rendement du blé dans les mélanges blé/légumineuse. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

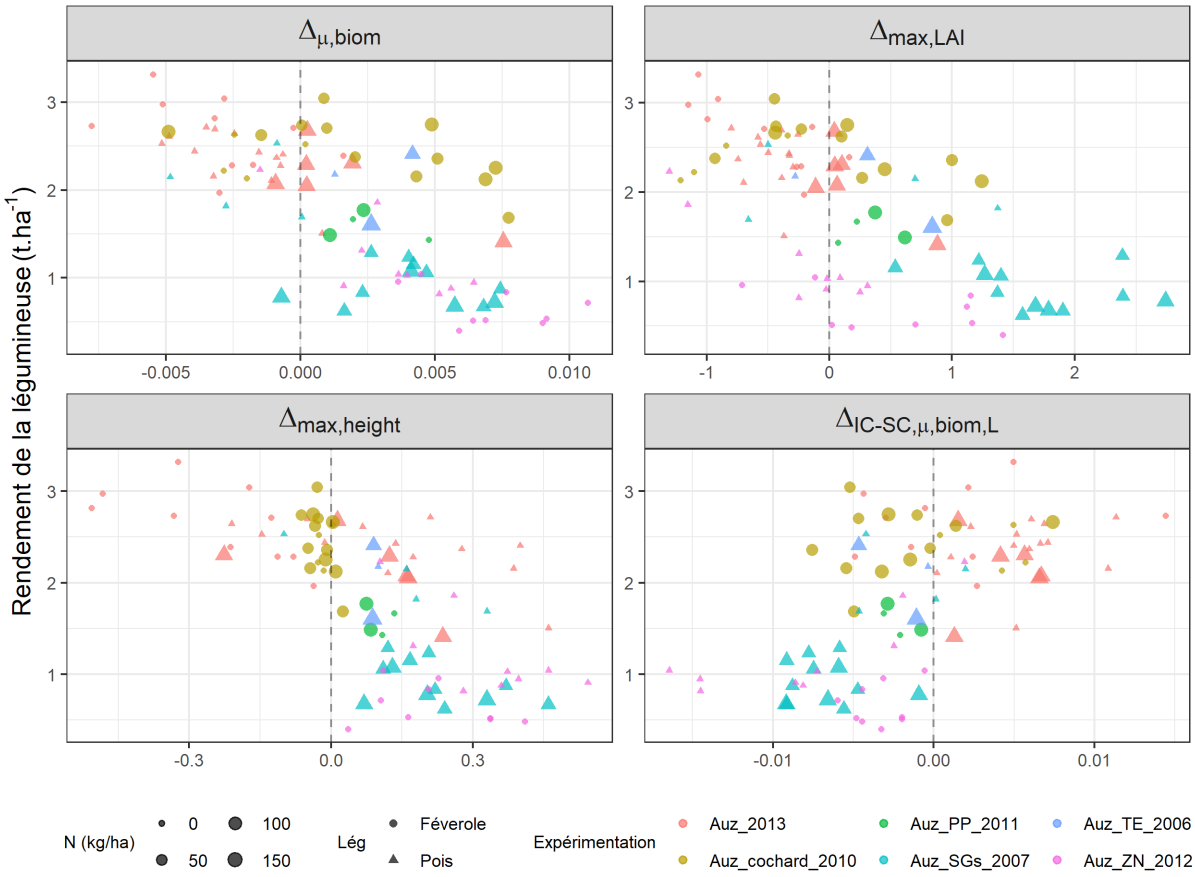


Figure IV.27 – Relations bivariées entre les 4 premières variables explicatives du rendement de la légumineuse dans les mélanges blé/légumineuse. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

Focus sur les rendements relatifs (pLER) des deux espèces

Comme pour les modèles d'ajustement de chaque mélange, l'INN est la première variable explicative du rendement relatif du blé (Figure IV.28), l'explication étant la même : les conditions non-fertilisées entraînent un faible rendement pour les cultures pures de blé, ce qui implique un rendement relatif élevé.

Au sein du mélange, les différences de vitesse de croissance en biomasse ($\Delta_{\mu,biom}$) et de hauteur maximale ($\Delta_{max,height}$) sont, comme pour le rendement brut, liées négativement au rendement relatif de la légumineuse (ρ = - 0.69 et -0.75 respectivement) et positivement à celui du blé (ρ = 0.48 et 0.47 respectivement).

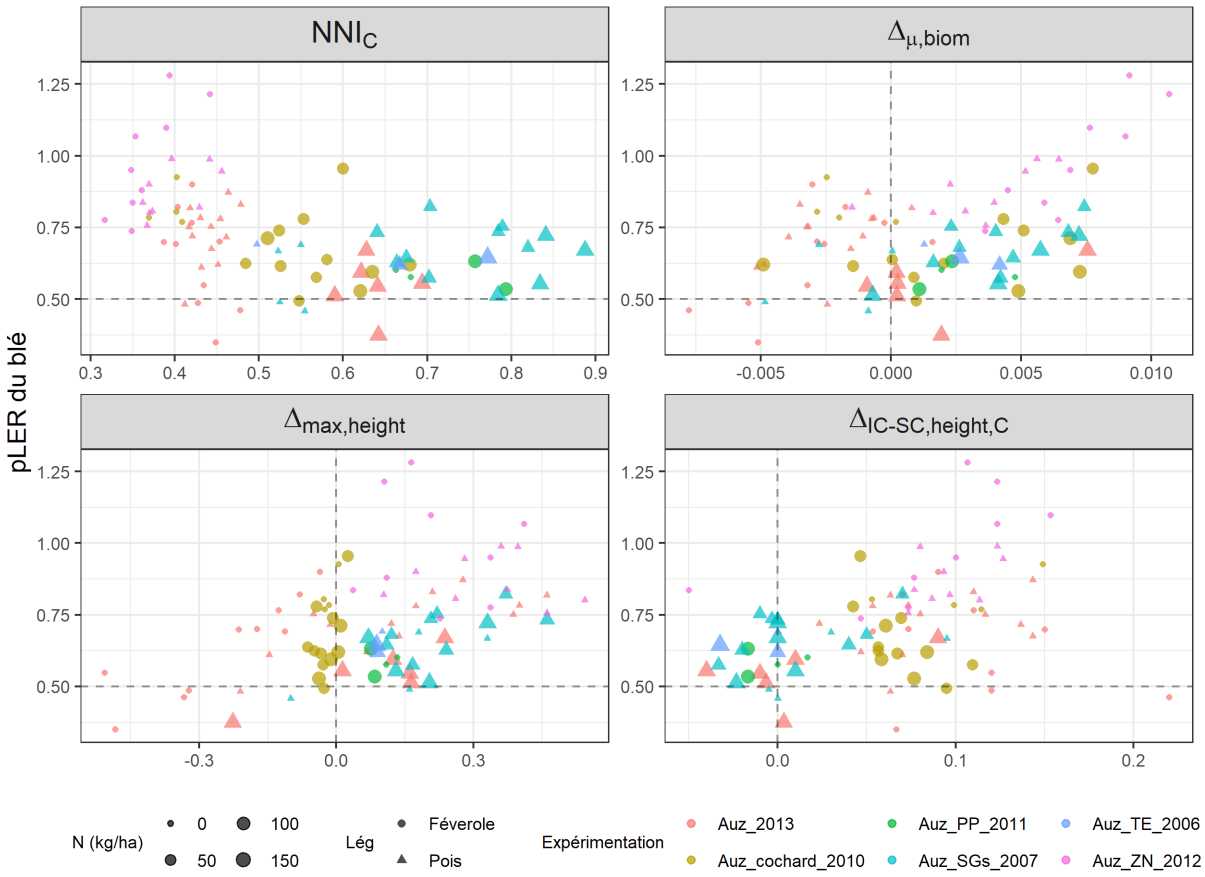


Figure IV.28 – Relations bivariées entre les 4 premières variables explicatives du rendement relatif (pLER) du blé dans les mélanges blé/légumineuse. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

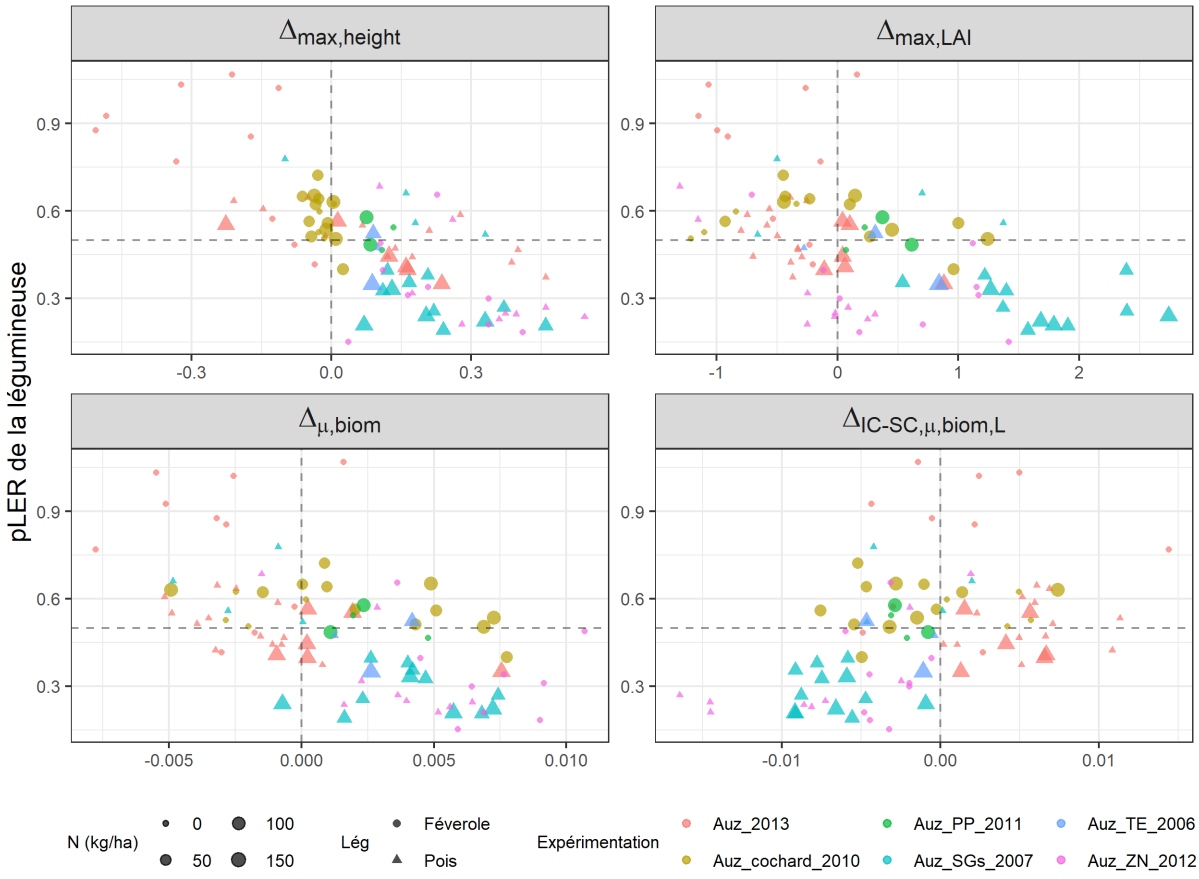


Figure IV.29 – Relations bivariées entre les 4 premières variables explicatives du rendement relatif (pLER) de la légumineuse dans les mélanges blé/légumineuse. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

3.3.3 Teneur en azote (N) du grain du blé dur

Les ajustements des modèles de la teneur en N du grain sont satisfaisants (Figure IV.30).

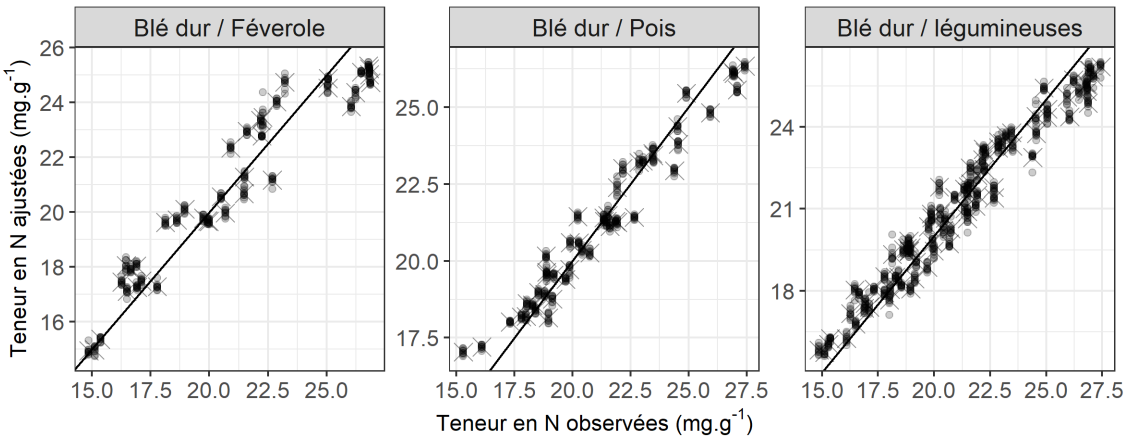


Figure IV.30 – Valeurs ajustées vs observées pour les modèles de teneur en azote du grain, entraînés sur l'ensemble des données, pour les mélanges blé dur/féverole, blé dur/pois et blé dur/légumineuse.

Logiquement, le statut azoté du blé (INN) ressort comme variable explicative la plus importante dans les trois modèles (Figure IV.31), ce qui est un résultat connu dans la littérature et qui s'explique de par la construction même de cette variable (l'INN est fonction de la teneur en azote des parties aériennes des plantes). La variété de blé est également sélectionnée dans tous les modèles, ce qui diffère des modèles d'ajustement des rendements bruts et relatifs. Ceci suggère que l'information liée à la variété de blé est nécessaire à la compréhension de la teneur en N du grain et n'est pas contenue dans les variables explicatives construites à partir des traits (biomasse, hauteur, caractéristiques foliaires).

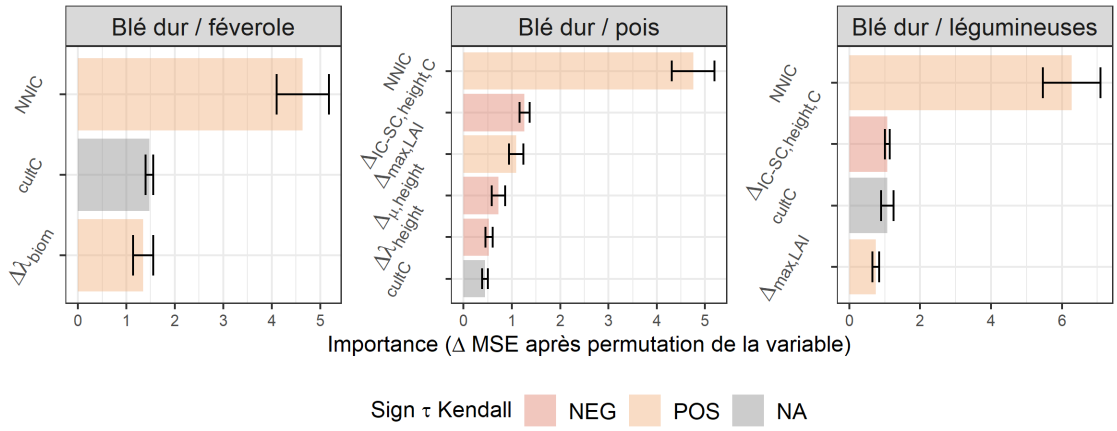


Figure IV.31 – Importances et écart-type (lié à l'imputation) des variables des modèles de teneur en azote du grain pour les trois types de mélange ; NEG (respectivement POS) : $\tau < 0$ (respectivement > 0) entre teneur en azote du grain et variable explicative ; NA : τ non-calculé car variable qualitative

Mélanges blé dur / féverole

Une relation linéaire robuste existe entre la teneur en N du grain et l'INN ($\rho = 0.84$, Figure IV.32). On observe l'effet seuil lié à la fertilisation. Ces observations sont logiques : la fertilisation permet à la céréale d'accumuler plus d'azote dans son grain. L'effet variétal ($cult_C$) détecté par le modèle semble être avant tout un effet fertilisation, puisque les cultivars *sculptur* et *nefer* ne sont cultivés qu'en conditions non-fertilisées alors que le cultivar *L1823* est parfois fertilisé. La relation entre $\Delta_{\lambda,biom}$ et la teneur en N semble peu robuste, et est difficile à interpréter.

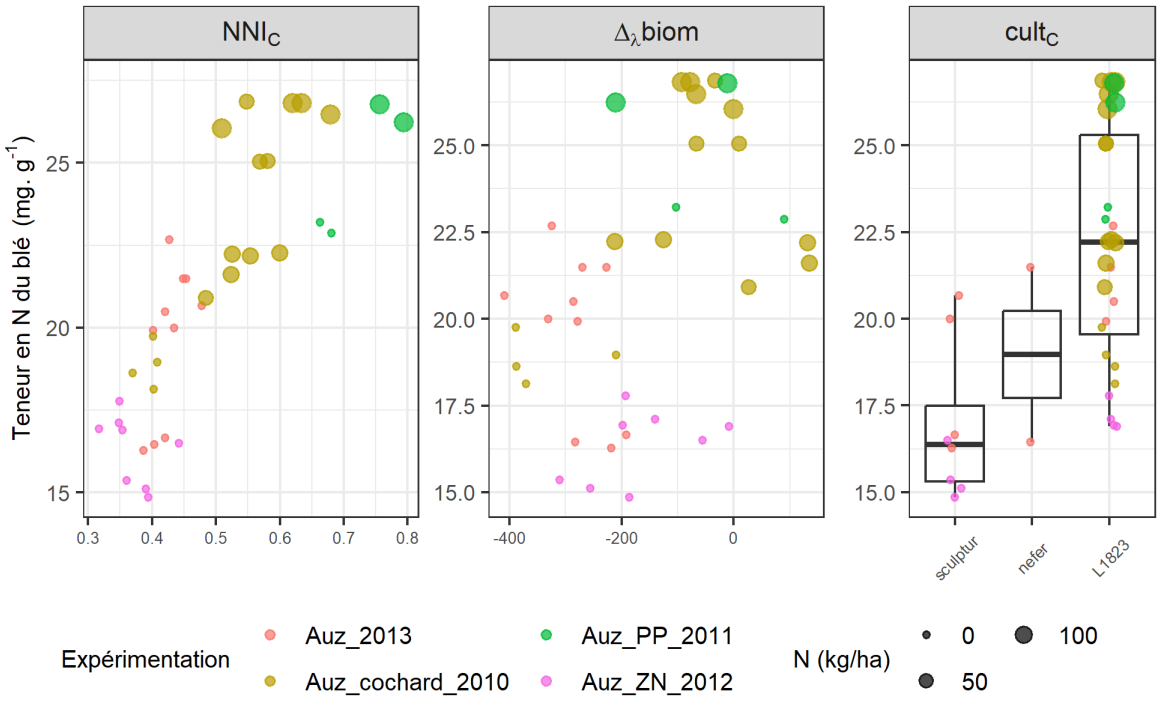


Figure IV.32 – Relations bivariées entre teneur en N du grain et variables explicatives pour les modèles blé dur / féverole. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

Mélanges blé dur / pois

Au-delà de la relation logique entre l'INN et la teneur en N du grain ($\rho = 0.85$), il existe des effets antagonistes sur le pLER du blé et la teneur en N du grain de i) la différence de hauteur du blé entre la culture associée et la culture pure ($\Delta_{IC-SC,height}$), ii) la différence de vitesse de croissance en hauteur au sein du mélange($\Delta_{\mu,height}$) et

iii) la différence de LAI maximal ($\Delta_{max,LAI}$) (Figure IV.33). Ces variables impactent également les rendements brut ($\Delta_{max,LAI}$) et relatif ($\Delta_{IC-SC,height}$, $\Delta_{\mu,height}$) du blé. On peut donc penser que le lien existant entre ces variables et la teneur en N du grain est principalement dû à la fertilisation, qui en améliorant le statut azoté du blé, lui permettent d'atteindre un rendement plus élevé et d'accumuler plus d'azote.

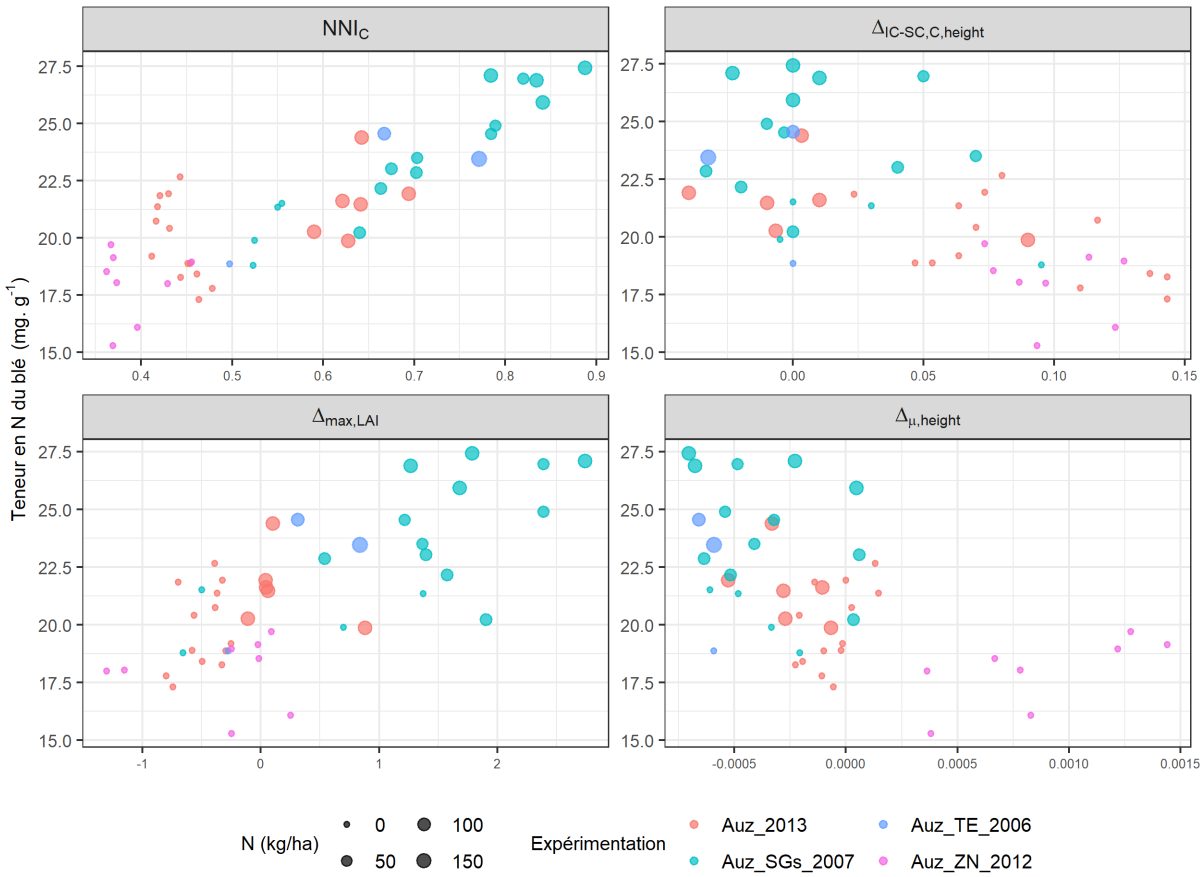


Figure IV.33 – Relations bivariées entre la teneur en azote (N) du grain de blé et les variables explicatives pour les modèles blé dur/pois. La taille et la couleur des points illustrent respectivement le niveau de fertilisation azotée (N) et l'expérimentation.

3.4 Discussion

3.4.1 Forces et limites de l'approche statistique mobilisée

L'approche statistique utilisée (forêt aléatoire combinée à un modèle mixte) avait pour objectif de modéliser le lien entre les différentes variables réponses et les variables explicatives, tout en tenant compte de la dépendance intra-expérimentation des observations. Les bonnes qualités d'ajustement obtenues en font donc une méthode pertinente.

L'objectif de prendre en compte la dépendance intra-expérimentation a cependant été seulement partiellement atteint. En effet, comme le montrent les intervalles de confiance autour des valeurs de b (Figure IV.12), toute la variabilité liée au facteur expérimentation n'a pas été captée par le facteur aléatoire ($Z_i b_i$, Équation (3.1)), suggérant qu'une partie de cette variabilité est contenue dans les facteurs fixes (variables d'interaction entre espèces au sein du mélange et variables d'adaptation au mélange d'une espèce). Cette hypothèse a été corroborée par l'étude des relations bivariées entre variables explicatives et variables réponses. D'autres travaux utilisant le même type de modèle montrent cependant des effets liés aux facteurs aléatoires plus marqués (Pellagatti et al., 2021 ; intervalles de confiance des coefficients ne contenant pas 0). Une perspective possible pour mieux tenir compte de la dépendance intra-expérimentation et estimer plus précisément les coefficients des facteurs aléatoires serait d'ajouter une structure hétéroscédastique au modèle mixte (i.e. supposer une matrice de covariance B dont les valeurs sur la diagonale (γ_i^2) dépendent de l'expérimentation). Néanmoins, il faudrait pour cela avoir suffisamment d'unités expérimentales par expérimentation afin d'estimer la variance correctement (Tableau IV.2). Dans cette perspective, incrémenter notre jeu de données avec des expérimentations contenant plus d'unités expérimentales pourrait être un moyen de prendre en compte l'hétéroscédasticité des expérimentations.

L'imputation des données manquantes a permis d'utiliser une plus grande partie du jeu de données (25% des individus statistiques auraient été perdus si j'avais supprimé chaque ligne pour laquelle au moins une variable était manquante) sans perturber fortement les estimés des paramètres (b) et les valeurs ajustées. La procédure de sélection de variables ajoutée au modèle a permis de limiter le nombre de variables explicatives des variables réponses. La mesure d'importance a permis de hiérarchiser leur influence sur les variables réponses, ce qui améliore notre compréhension des systèmes étudiés. Néanmoins, malgré ma volonté de parcimonie dans le choix des variables (critère restrictif : ne sélectionner que les variables sélectionnées dans les 10 versions du jeu de données), certaines variables ont été sélectionnées par les modèles

sans qu'on puisse interpréter convenablement leurs relations avec les variables réponses (Figure IV.17). Ceci pourrait indiquer i) qu'il existe des interactions entre variables explicatives sélectionnées invisibles sur des relations bivariées ou ii) que le modèle sélectionne des variables non-informatives, via un phénomène de surajustement.

3.4.2 Synthèse des variables explicatives impliquées dans la performance des cultures associées

Variables modélisées			Variables impliquées		
			Interactions entre espèces au sein du mélange	Adaptation au mélange d'une espèce	INN
Blé/féverole	Féverole	Rendement	μ_{biom}, LAI, SLA	-	-
		pLER	max_{height}, μ_{biom}	-	-
	Blé	Rendement	LAI, μ_{biom}, max_{height}	λ_{height}	-
		pLER	μ_{biom}	LAI	×
Blé/pois	Pois	Rendement	μ_{biom}, LAI, max_{height}	λ_{biom}, λ_{height}	-
		pLER	μ_{biom}, LAI, max_{height}	λ_{biom}, λ_{height}	-
	Blé	Rendement	LAI, μ_{biom}, λ_{height}	μ_{biom}	×
		pLER	max_{height}, μ_{height}	max_{height}, μ_{biom}	×

Table IV.6 – Principaux processus impliqués dans la performance des mélanges ; pLER : partial Land Equivalent Ratio

Les modèles d'ajustement de la performance des cultures associées mettent en avant plusieurs variables explicatives des rendements relatifs et bruts du blé et des deux légumineuses (Tableau IV.6).

Dans l'ensemble des modèles, le nombre de variables explicatives rendant compte d'interactions entre les deux espèces au sein du mélange est supérieur ou égal au nombre de variables explicatives rendant compte d'une adaptation au mélange de chacune des composantes (changement des valeurs de traits d'une espèce selon qu'elle pousse en culture pure ou associée). Ceci souligne l'importance des interactions plante-plante dans la performance de chacune des composantes du mélange.

De manière générale, les variables liées à la hauteur et à la biomasse sont souvent sélectionnées pour expliquer les variables de réponse, indiquant que les différences sur

la dynamique et la stature finale des espèces sont cruciales pour la performance de chacune des espèces vis-à-vis de l'autre. Ces variables sont associées à des processus d'acquisition des ressources, mais sont également des résultantes de la compétition (Trinder et al., 2012). Les données à notre disposition ne permettent pas de distinguer l'effet de la compétition de sa résultante. Comme pour la complémentarité, causes et effets de la compétition sont difficiles à distinguer via l'étude des traits (Barry et al., 2019, Streit et Bellwood, 2022). Au sein des mélanges, la différence de vitesse de croissance en biomasse est identifiée comme clé dans la majorité des modèles (7 modèles sur 8). Ce résultat corrobore d'autres études, sur d'autres mélanges (avoine (*Avena sativa*)/lupin (*Lupinus Angustifolius*) et colza (*Brassica napus*)/maïs (*Zea mays*) ou colza/soja (*Glycine max*)), relatant l'importance de cette variable dans la production d'une culture associée (Dong et al., 2018, Engbersen et al., 2021).

Dong et al., 2018 relient le temps d'atteinte du taux de croissance maximal (abscisse du point d'inflexion de la courbe) pour la biomasse (variable similaire à λ_{biom}) à la productivité de la culture associée. L'interprétation proposée est que ce lien entre le démarrage de la croissance et le rendement est la résultante d'une différentiation temporelle de niche ayant lieu au sein du mélange, interprétation corroborée par une méta-analyse basée sur les résultats de 100 études (Yu et al., 2015). Dans mon travail, je montre que les décalages de démarrage de la croissance d'une espèce (leur différence entre culture pure et associée) sont liés aux rendements bruts du blé dans les mélanges blé/féverole et aux rendements bruts (et relatifs) du pois dans les mélanges blé/pois (Tableau IV.6). En revanche, le décalage de démarrage de croissance de hauteur ou de biomasse entre les deux espèces au sein du mélange n'apparaît qu'une seule fois dans les 8 modèles (rendement brut du blé dans les associations blé/pois, Tableau IV.6). Cependant, les études soulignant l'importance de cette variable sont réalisées sur des cultures associées asynchrones, dans lesquelles les deux espèces sont semées à plus d'un mois d'intervalle, induisant de fait une différentiation temporelle de niche plus marquée que dans les cultures associées synchrones, dans lesquelles la céréale et la légumineuse sont semées et récoltées en même temps (ce qui correspond aux mélanges que je modélise). D'autre part, au sein des mélanges, ces différences sur des caractéristiques de démarrage de courbes de croissance sont corrélées aux différences sur la vitesse de croissance, un ensemble de variables qui est souvent sélectionné dans les modèles.

La différence de hauteur entre culture associée et culture pure pour une espèce donnée ($\Delta_{IC-SC,height}$) n'est que peu sélectionnée comme variable explicative (1 modèle sur 8) contrairement à la différence de hauteur entre les deux espèces au sein du mélange (5 modèles sur 8). Pourtant, des relations significatives entre $\Delta_{IC-SC,height}$ et les effets de

biodiversité (effet de complémentarité, de sélection, ou de biodiversité ; Chapitre III.2) ont parfois été mises en avant dans la littérature (Engbersen et al., 2022). Malgré cette divergence (différences entre espèces au sein du mélange dans mes modèles vs entre cultures pure et associée dans les modèles d'Engbersen et al., 2022), ces résultats soulignent l'importance de la stature pour atteindre un rendement élevé. La différence entre les résultats de Engbersen et al., 2022 et les miens est probablement due à la variable réponse considérée. Mon travail cherchait à modéliser le rendement de chaque composante du mélange séparément là où leurs travaux mettaient en lien des variables à l'échelle du mélange. Cette hypothèse pourra être testée dans de futurs travaux.

Dans les modèles blé dur/légumineuse, le point notable qui ressort est que les mêmes variables sont sélectionnées que dans les modèles mélange par mélange. La variété de la céréale ou de la légumineuse sont également absentes des modèles. L'absence de l'effet espèce (pour les légumineuses) ou variété comme variables explicatives dans la majorité des modèles laisse entendre que du point de vue du blé dur (présent dans les deux mélanges), l'important n'est pas tant l'identité de l'espèce compagne que la valeur de certains de ses traits fonctionnels. Cette constatation présage de la possibilité de généraliser l'approche à d'autres espèces et variétés de légumineuses via la mobilisation du concept de groupe fonctionnel (Montazeaud et al., 2020).

L'ESSENTIEL

La démarche de modélisation que j'ai développée avait pour but d'évaluer la valeur prédictive de variables basées sur des différences entre caractéristiques des espèces, en tenant compte de l'hétérogénéité de mes données, ainsi que la dépendance intra-expérimentation des observations. La prise en compte d'un effet environnemental, initialement envisagée par le calcul de variables explicites, puis finalement modélisée via un facteur aléatoire, a été partiellement réussie car la dépendance intra-expérimentation se trouvait également dans les facteurs fixes. Cependant, les différents modèles ont permis d'obtenir des ajustements satisfaisants des valeurs observées, indiquant que les processus représentés par les concepts d'écologie et d'agronomie mobilisés sont bien centraux pour prédire la performance des composantes du mélange. Mon approche indique que les différences dans l'établissement de l'architecture aérienne entre les deux espèces sont cruciales, particulièrement en considérant les dynamique de croissance en biomasse ou hauteur. Même si le rôle de variables d'état du couvert (ex. hauteur et LAI maximal) est ressorti, ces observations indiquent qu'une analyse plus fine des variables qui résument la dynamique (i.e. variables de flux) dans le mélange permettrait d'affiner les déterminants des statures finales des espèces, par exemple en identifiant des stades de développement (phénologiques) clé auxquels la compétition devient déterminante pour la performance de chaque espèce.

Chapitre V

Synthèse et perspectives

1 L'écologie des communautés permet de mieux comprendre les cultures associées

Mon travail de thèse avait pour objectif de qualifier l'apport de l'écologie des communautés dans une démarche de modélisation visant à comprendre le fonctionnement des cultures associées. Plusieurs publications appellent en effet à plus de ponts entre l'écologie et l'agronomie (Barot et al., 2017, Brooker et al., 2015).

1.1 Décomposition de l'effet de la biodiversité

Nous avons pu voir que l'écologie des communautés, via la décomposition de l'effet de la biodiversité (Loreau et Hector, 2001), fournissait un cadre d'étude pertinent pour les cultures associées (Chapitre III). Pour rappel, l'effet de la biodiversité quantifie les gains (ou pertes) de rendement entre la culture associée et les cultures pures correspondantes. Cet effet se décompose en un effet de complémentarité (lié à la facilitation et à la complémentarité de niche) et un effet de sélection (une espèce fonctionnant déjà bien en culture pure profite du contexte de culture associée, au détriment de l'autre). Mes travaux permettent une quantification globale de l'effet de la biodiversité au sein des cultures associées céréale-légumineuse à l'échelle des 35 expérimentations en conditions non-fertilisées avec un gain de rendement moyen de 0.86 $t.ha^{-1}$. Je montre aussi que ce gain de rendement est essentiellement dû aux interactions positives entre la céréale et la légumineuse (effet de complémentarité prépondérant). Néanmoins, ces résultats ont été tempérés par l'effet des pratiques (choix des espèces cultivées et fertilisation azotée). Par exemple, l'effet de complémentarité a diminué en conditions fertilisées dans les associations blé dur/pois, mais pas dans les associations blé tendre/pois ou blé dur/féverole. Ces résultats sont cohérents avec la littérature existante montrant l'influence de la fertilisation sur les effets de complémentarité et de biodiversité (Engbersen et al., 2022, Li et al., 2020a).

Ces résultats soulignent la nécessité de considérer les pratiques au sein des agroécosystèmes comparativement aux écosystèmes naturels. Ainsi, si on peut observer des relations positives et robustes en conditions bas-intrants, la gestion des agroécosystèmes via certains choix de pratiques peut perturber ces relations et est donc primordiale à prendre en compte si l'on veut prédire la performance des cultures associées. En conditions bas-intrants, les rendements sont généralement faibles comparativement aux rendements obtenus en conventionnel. À titre d'exemple, le rendement moyen du blé est de 6-7 $t.ha^{-1}$ en agriculture conventionnelle (Agreste, 2022) et de 3-4 $t.ha^{-1}$ en agriculture biologique (David et al., 2005). L'étude de ces agroécosystèmes

en conditions bas-intrants est cependant pertinente, puisque les cultures associées sont essentiellement pratiquées dans ces conditions (Verret et al., 2020). Nos travaux montrent que la fertilisation pourrait augmenter les rendements respectifs de chacune des composantes du mélange (comme observé dans les mélanges blé dur/féverole). Cependant, l'objectif n'est pas forcément de prôner une fertilisation des cultures associées pour augmenter leur performance mais plutôt de gérer l'azote dans la rotation culturale pour maximiser le rendement de l'association, par exemple en raisonnant la succession des espèces ou en insérant des cultures intermédiaires comme engrais vert (Couëdel et al., 2018). De plus, il est possible que les contraintes réglementaires / géologiques / environnementales qui vont s'imposer dans les prochaines décennies, comme l'épuisement des ressources en phosphore au cours du siècle (Cordell et al., 2009), fassent que les systèmes agricoles se tournent de plus en plus vers des systèmes bas-intrants, rendant les mélanges plurispécifiques plus attractifs (Barot et al., 2017).

1.2 Déterminants de la performance des cultures associées

Dans le Chapitre IV, j'ai utilisé des indicateurs liés aux interactions plante-plante ayant lieu au sein du mélange ainsi que des indicateurs liés à l'adaptation au mélange de chacune des espèces. Les modèles ont montré des ajustements convenables et ont permis de distinguer certaines variables d'intérêt pour la compréhension du fonctionnement des associations. Ainsi, voir les associations sous le prisme d'une communauté, c'est-à-dire un assemblage de deux espèces partageant la même zone géographique en même temps (Begon et al., 2005), permet d'améliorer notre compréhension de leur fonctionnement. Si la performance d'une espèce est conditionnée par la valeur de ses traits, qui lui permettent d'utiliser des ressources pour se développer, elle est également dépendante des valeurs de traits de l'espèce compagne, puisque les deux espèces interagissent entre elles. Dans les cultures associées céréale-légumineuse, nous avons étudié plus précisément des relations impliquant les vitesses (Δ_μ) et le démarrage (Δ_λ) de la croissance, et les hauteurs/biomasses maximales. De plus, nous avons inclus des traits liés à l'interception de la lumière (LAI) et un indicateur du statut azoté des plantes (INN). Les traits liés à l'architecture aérienne et à l'interception de la lumière (hauteur, LAI) ont souvent été sélectionnés comme influençant la performance des composantes du mélange. Les traits reliés à la croissance (dynamique) relative des deux espèces sont ainsi primordiaux, ce qui milite pour réaliser des mesures dynamiques dans les expérimentations agronomiques, si l'on veut élucider les processus physiologiques impliqués dans ces différentiels de croissance. Le travail que j'ai fait sur l'ajustement de ces courbes de croissance illustre que ces mesures doivent être réalisées un nombre suffisant de fois (minimum 4-5 fois pendant le cycle)

à des instants clé du développement de la culture (notamment pendant la période végétative, de croissance et de stagnation), pour permettre d'estimer les paramètres liés au démarrage de la culture (λ) et à la vitesse de croissance (μ).

Les variables que nous avons sélectionnées dans nos modèles sont calculées à partir de traits qui sont traditionnellement mesurés dans les expérimentations agronomiques et qui reflètent principalement des phénomènes de compétition entre les plantes (biomasse, hauteur, LAI). La mesure de traits additionnels reflétant plutôt des interactions positives entre les plantes enrichirait notre compréhension des cultures associées. Parmi les traits auxquels nous n'avions pas accès, les traits racinaires (par ex. longueur, biomasse, profondeur des racines) sont probablement ceux qui manquent le plus. En effet, plusieurs phénomènes d'interaction entre la céréale et la légumineuse sont directement liés à la rhizosphère, notamment via les processus de complémentarité de niche pour l'utilisation de l'azote (Bedoussac et al., 2015), de facilitation pour l'utilisation du phosphore inorganique (Homulle et al., 2021) ou de micro-nutriments, *via* l'acidification de la rhizosphère (Long et al., 2014). Du point de vue expérimental, ce manque n'est pas étonnant puisque ces traits sont les plus coûteux (en terme de travail) à obtenir.

Cependant, nos modèles ajustent de manière satisfaisante la performance des cultures associées, en utilisant uniquement les traits aériens, qui semblent finalement assez intégrateurs de ce qui se passe dans le sol. Ainsi, c'est bien parce qu'on associe une légumineuse capable de fixer l'azote atmosphérique avec du blé que la culture associée fonctionne mieux que les cultures pures en contexte bas-intrants. Ceci est reflété par nos variables explicatives sans pour autant avoir intégré explicitement les interactions souterraines explicitant les processus de complémentarité ou de facilitation (Barry et al., 2019).

1.3 Bilan sur l'apport de l'écologie des communautés

L'écologie des communautés fournit donc des concepts et outils améliorant notre compréhension des cultures associées. La décomposition de l'effet de la biodiversité nous a permis de distinguer et quantifier deux grands processus à l'oeuvre au sein des mélanges. L'utilisation de variables d'interactions entre plantes au sein du mélange et d'adaptation au contexte de culture associée nous a permis de hiérarchiser ces processus et de tirer des connaissances sur ces systèmes complexes.

Néanmoins, les communautés considérées en écologie sont généralement plus diverses et (semi-)naturelles. Le nombre d'espèces interagissant est généralement supérieur à celui des agroécosystèmes. Dans notre cas, les variétés sont des variétés sélectionnées,

dans un environnement partiellement contrôlé (ressources en azote, en eau). Un agroécosystème, même complexe comme une culture associée, reste un environnement modifié artificiellement dans un/des objectif(s) agronomique(s) particulier(s) (rendement d'une espèce, contrôle des adventices, fourniture en N du sol, etc.).

Dès lors, l'écologie des communautés peut à la fois apporter un cadre, des outils et des concepts, mais l'opérationnalisation des connaissances produites reste dépendante de l'expertise agronomique. Par exemple, nous avons montré que certaines différences de traits étaient plus impactantes pour le rendement de certaines des composantes du mélange. L'expertise agronomique peut alors servir à mettre en place les pratiques nécessaires pour atteindre les états du couvert désirables pour les objectifs recherchés.

2 Utiliser la science des données pour comprendre des systèmes complexes

Ma thèse a permis d'explorer certaines des problématiques liées à l'utilisation de jeux de données globaux. J'ai pu explorer les enjeux méthodologiques liés à ce type de jeu de données (Chapitre II.2) et en tirer des connaissances sur les cultures associées en mobilisant et en développant des méthodes statistiques originales et adaptées à mes besoins.

2.1 Fédérer et partager des données expérimentales

Le jeu de données construit et utilisé pendant ma thèse est conséquent et est à ma connaissance le plus étendu sur les cultures associées céréale-légumineuse, une partie étant déjà en ligne[1]. Nous avons cependant vu que cette richesse est aussi synonyme d'hétérogénéité, ce qui complexifie les analyses statistiques sans certaines précautions. Pour aider à l'analyse de ce type de jeux de données, j'ai ainsi développé la méthode des k-cliques (https ://github.com/RemiMahmoud/kclique). Pour rappel, cette méthode utilise la théorie des graphes pour identifier des sous-jeux de données factoriels permettant des analyses statistiques robustes. Cette méthode nous a permis d'identifier certaines questions de recherche intéressantes mais qui ne peuvent pas être traitées "proprement" avec le jeu de données actuel (par manque d'observations et/ou de modalités). La méthode des k-cliques pourra donc être utilisée pour organiser la collecte de données, en donnant la possibilité de qualifier la contribution d'une

1. https ://datadryad.org/stash/dataset/doi :10.5061/dryad.9ghx3ffhv

expérimentation sur l'ensemble des données (en terme de modalités qu'il faudrait ajouter par exemple).

Une autre perspective de mon travail serait d'identifier au sein d'un jeu de données l'ensemble des facteurs de confusion possibles, étant donné plusieurs variables qualitatives. Deux facteurs sont confondus s'ils varient en même temps, empêchant de distinguer les deux effets (Casler, 2015). Par exemple, si une variété A est cultivée dans un environnement a et une variété B est cultivée dans un environnement b, il y a confusion entre les deux facteurs (on ne peut pas distinguer l'effet de l'expérimentation de l'effet de la variété sur une variable de sortie). J'ai brièvement exploré une approche pendant ma thèse dans le but d'identifier les facteurs de confusion possibles au sein de mon jeu de données. Cette approche consistait à comparer plusieurs variables qualitatives 2 à 2 ainsi que les co-occurences de leurs modalités (autrement dit est-ce que les modalités de deux facteurs varient en même temps). Plus précisément, pour chaque couple de variables qualitatives X et Y, j'ai construit une matrice A, dont les éléments A_{ij} valent 1 si la modalité i de la variable X est observée conjointement à la modalité j de la variable Y et 0 sinon (Tableau V.1). On obtient ainsi une matrice contenant uniquement des 0 et des 1, sur laquelle on peut calculer certains indices quantifiant les co-occurences de 0 et de 1 entre lignes et colonnes (Valsecchi et Todeschini, 2020). Par manque de temps, je n'ai pas poussé plus loin cette piste, mais si je devais reprendre ce travail, je chercherais à trouver les couples de variables pour lesquels on observe peu de co-occurences de 1 sur les lignes et colonnes, indiquant un effet de confusion entre les deux variables.

		Expérimentations	
		a	b
Variétés	A	1	0
	B	0	1
	C	0	1

Table V.1 – Tableau de co-occurrence fictif, où il existe un facteur de confusion entre la variété et l'expérimentation. Les cases du tableau valent 1 si la variété de la ligne est cultivée dans l'expérimentation de la colonne et 0 sinon

Le jeu de données global utilisé dans mon travail de thèse va être enrichi de plusieurs autres expérimentations grâce à des projets européens passés ou en cours sur les cultures associées (ReMIX, IntercropValuES), ce qui permettra d'aller plus loin et de consolider les aspects de modélisation évoqués dans la section suivante.

La perspective d'une utilisation plus répandue de ces jeux de données globaux en agronomie est prometteuse. Dans cette optique, j'ai participé avec certains membres de

mon unité d'accueil (UMR AGIR) à constituer un autre jeu de données global sur plus de 40 expérimentations incluant des cultures intermédiaires (cultures implantées entre deux cultures de rente, dont l'objectif est de fournir certains services écosystémiques). Ce jeu de données pourra servir à acquérir des connaissances plus vastes et répondre à certaines questions en suspens sur les cultures intermédiaires (Lamichhane et Alletto, 2022).

Ce travail de fédération et homogénéisation de données expérimentales, dans un but d'acquisition et de généralisation des connaissances, s'inscrit dans le mouvement de la Science Ouverte, qui permet de rendre la recherche scientifique transparente et accessible à tous les niveaux de la société (Vicente-Saez et Martinez-Fuentes, 2018). L'émergence de revues de type "Peer Community In", basées sur une recommandation ouverte par les pairs, fait également partie de ce mouvement (Guillemaud et al., 2019). Je souhaite d'ailleurs soumettre une partie de mon travail décrit dans le Chapitre IV dans une revue de ce type (PCI in Mathematical & Computational Biology [1]).

Si ces initiatives vont dans le bon sens, notamment dans un contexte de défiance vis-à-vis de la science (Nasr, 2021), elles font face à certains freins dont il faut tenir compte. Par exemple, l'évaluation des carrières scientifiques sur le nombre de publications peut freiner le partage des données nécessaire pour les publications, tant que le nombre ou la primauté des publications resteront un critère central d'évaluation (Watson, 2022). Le jeu de données a notamment servi dans deux publications auxquelles je suis associé. L'agrégation et l'homogénéisation de ces données a donc permis d'utiliser des observations phénotypiques déjà acquises pour répondre à d'autres questions scientifiques que celles traitées dans mon travail de thèse, comme par exemple l'impact de l'association et de la fertilisation sur l'allométrie reproductive (Gaudio et al., 2021a) et la constitution d'une chaîne de modélisation pour prédire un ensemble de services écosystémiques dans des cultures associées (Meunier et al., 2022a).

2.2 Développer des approches de modélisation adaptées à des connaissances souvent incomplètes sur les systèmes diversifiés

Ma démarche de modélisation se situe entre les modèles mécanistes, où l'ensemble des processus ayant lieu dans le système sont finement décrits, et qui ont pour but de simuler le fonctionnement du système, et les approches de type méta-analyse, permettant des connaissances relativement génériques en compilant les résultats

1. https ://mcb.peercommunityin.org/

d'expérimentations déjà publiés (voir Chapitre I). Ma démarche a consisté à compiler les résultats d'expérimentations pour mettre en relation la performance des espèces associées (rendements bruts, relatifs et teneur en azote du grain) et les observations phénotypiques, ainsi que des pratiques agronomiques (choix de l'espèce ou de la variété, fertilisation). Mon approche a été conditionnée par un constat sur l'état des connaissances disponibles sur les cultures associées et par mes objectifs de modélisation ('data-driven', Chapitre I). Ce type de démarche est rare en agronomie et a induit des problématiques auxquelles je ne pensais pas être confronté en démarrant ma thèse (données manquantes, combinaisons de variables manquantes, hétérogénéité entre expérimentations, etc.).

En terme de modèle statistique, j'ai exploré une méthode originale, permettant de combiner capacité d'ajustement des forêts aléatoires et flexibilité des modèles mixtes (Hajjem et al., 2014, Capitaine et al., 2021).

En terme de perspectives de modélisation, une piste intéressante à explorer serait d'examiner les relations causales existant entre les variables explicatives et les variables réponses. La littérature sur la détermination de relations causales en statistique est vaste (Pearl, 2009) et nombreuses sont les applications de ces méthodes en écologie et en agronomie (Arif et Macneil, 2022). Une application possible de ces méthodes à mes données consisterait à i) identifier les relations causales potentielles au sein des données (ex. LAI $\Rightarrow$ Biomasse), ii) tester la cohérence de ces relations causales vis-à-vis des données et iii) ajuster un modèle statistique en tenant compte de ces relations (Arif et Macneil, 2022).

Mes recherches réalisées pour prendre explicitement en compte des variables environnementales m'ont permis de découvrir le domaine de l'analyse de données fonctionnelles (Ramsay, 2005) : comment résumer l'information contenue dans des réalisations (courbes) de variables aléatoires appartenant à des espaces de dimensions infinies (espaces de fonctions) et la mettre en relation avec des données d'espaces classiques ($\mathbb{R}$). Notamment, l'étude de la méthode SISIR (Picheny et al., 2019) m'a permis de comprendre certaines méthodes couramment utilisées en analyse de données fonctionnelles (pénalisation de la matrice de covariance, régularisation de coefficients, etc.). J'ai écrit les premiers calculs permettant de généraliser SISIR à un cadre multivarié (voir Annexes). J'ai écrit cette méthode mais ne l'ai pas implémentée ni testée sur mon jeu de données, puisque je n'ai pas pu tirer de résultats concluants en univarié. Je pense cependant que SISIR pourrait donner des résultats probants avec plus d'individus statistiques. Ainsi, cette méthode sera mobilisée dans un projet de recherche Européen (H2020 INVITE) pour analyser la diversité des conditions de culture dans un large réseau d'évaluation de variétés de tournesol (réseau national pluri-annuel

de 650 essais). En multivarié, l'application de cette méthode consisterait à déterminer conjointement les zones d'influence de variables climatiques sur le rendement du tournesol, en préalable à une approche de classification des environnements de culture.

En plus des deux méthodes décrites dans le Chapitre IV.2.3, une méthode que je pourrais explorer est l'utilisation d'analyse en composantes principales fonctionnelle (ACPF) pour décrire mes expérimentations (Ramsay, 2005). Cependant, il me faudrait pour cela plus de variabilité dans les individus statistiques (ici expérimentations). Ce type de méthode permet de réduire la dimension des variables fonctionnelles (courbes). On peut ensuite utiliser l'information contenue dans la dimension réduite pour régresser une variable scalaire (nombre). Plus précisément, l'ACPF fonctionne de manière similaire à l'analyse en composantes principales multivariées classique, la principale différence étant que l'ACPF considère les individus comme des fonctions, réalisations de variables aléatoires appartenant à un espace de dimension infinie. Les composantes principales issues d'une ACPF sont donc des fonctions (i.e. des courbes) et non plus des vecteurs appartenant à un espace de dimension finie. Une description des individus par des composantes principales fonctionnelles permet de caractériser leurs principaux modes de variation. Appliquée à un jeu de données contenant plusieurs variables climatiques, une ACPF permettrait de caractériser ces climats (ex. excès de précipitations en hiver, températures excessives durant le printemps). Ainsi, de récents travaux utilisent les données climatiques liées à 298 exploitations agricoles en France et effectuent une ACP fonctionnelle sur chacune des variables climatiques (Bonneu et al., 2022). Ils caractérisent ainsi chacune des exploitations via leurs coordonnées sur les composantes principales de chacune des variables climatiques. Ils utilisent ensuite ces coordonnées comme variable explicative des rendements du blé dans chacune des exploitations et en déduisent des climats plus ou moins favorables à la production de blé.

3 Contribution de mes travaux pour promouvoir les cultures associées

3.1 Un outil d'aide à la décision, pour quoi faire ?

Lors de l'élaboration de ce projet de thèse, nous avions pensé à l'élaboration d'un outil d'aide à la décision (OAD) à destination des agriculteurs et agricultrices, les aidant dans le choix des espèces à cultiver en fonction des pédo-climats. Nous sommes assez rapidement revenus sur cet objectif pour plusieurs raisons.

La première, très pragmatique, est le peu de données à disposition, relativement à la complexité des systèmes considérés. En effet, mon jeu de données est à la fois riche et pauvre (Chapitre II), dans le sens où le nombre d'unités expérimentales en culture associée est aux alentours de 300, avec plusieurs modalités différentes, notamment en fonction des pédo-climats. Certaines pratiques agronomiques clés dans la conduite des cultures associées ne sont pas testées (par ex. semer/récolter les 2 espèces en même temps ou de manière asynchrone), et les variables explicatives sur lesquelles nous pouvions construire les modèles sont mesurées en cours de cycle, limitant donc les conditions d'application d'un potentiel OAD (qui devrait être capable de préconiser des espèces/variétés à associer avant le semis). La seconde raison est liée à un changement de notre vision du conseil agricole suite aux interactions que nous avons pu avoir dans mon unité d'accueil (UMR AGIR), composée de chercheuses et chercheurs ayant des spécialités très variées (agronomie système, modélisation, phytopathologie, économie de l'innovation, etc.). Pour ma part, issu d'une formation d'ingénieur généraliste (INSA Rennes, Génie Mathématique), je pensais naïvement quand je suis arrivé en stage à INRAE en 2018 qu'un système agricole était (presque) optimisable de la même manière qu'un processus industriel. Or, un système agricole est un lieu d'interactions entre décisions humaines, environnement, évènements extrêmes, stress biotiques et contexte socio-économique. Il est à ce titre loin d'être un système simple et totalement contrôlé. Les décisions prises par les acteurs du monde agricole ne sont pas uniquement guidées par une recherche du rendement, mais par un ensemble de motivations et de possibilités (machinisme agricole, terrain, disponibilité/prix de certaines semences) variant fortement entre les exploitations (Martin-Clouaire, 2017). Ce constat est renforcé dans des systèmes plus complexes comme les cultures associées. Dans ces systèmes pratiqués principalement en conditions bas-intrants, les raisons motivant les agriculteurs à cultiver des mélanges d'espèces sont nombreuses, et les critères de réussite de ces cultures varient beaucoup d'un exploitant à l'autre (Aare et al., 2021, Verret et al., 2020). À titre d'exemple, j'ai pu comprendre, au cours de ces échanges, que le changement vers des pratiques plus vertueuses n'est pas forcément lié à un manque de connaissances des acteurs de ces alternatives mais à un ensemble de verrous socio-techniques pouvant entraver les changements (Meynard et al., 2018). Ainsi, les freins à la pratique des cultures associées sont multiples (tri des grains, machinisme agricole peu développé, sélection variétale non adaptée, etc. ; Mamine et Farès, 2020).

Cette prise de recul vis-à-vis de mes données et des motivations variées poussant un agriculteur ou une agricultrice à cultiver des mélanges m'ont ainsi convaincu de ne pas partir sur le développement d'un outil d'aide à la décision.

3.2 Des pistes pour comprendre et positionner les cultures associées dans des environnements adaptés

Les cultures associées permettent généralement d'obtenir un rendement plus élevé et plus stable qu'en culture pure en conditions bas-intrants (Pelzer et al., 2014). Ainsi, ce type de pratique a toute sa place dans une agriculture économe en intrants, ou dans des environnements peu favorables à une production élevée. Les expérimentations constituant mon jeu de données ont généralement été effectuées en conditions bas-intrants (23 expérimentations / 35 non-fertilisées, fertilisation faible quand appliquée ($< 150\ kg.ha^{-1}$), reliquats modestes). Ainsi, mes travaux fournissent des informations robustes sur le fonctionnement des associations dont le domaine de validité est cohérent avec les conditions de culture des mélanges dans les exploitations agricoles. Parmi les connaissances objectivées et déjà connues, l'influence de la fertilisation, même modérée, sur les interactions positives entre plantes au sein du mélange est corroborée dans mes travaux. Un autre gain de connaissance obtenu et mobilisable (à long terme) est le rôle des différences entre traits au sein du mélange. Nous montrons que dans les mélanges, les différences de traits liés à la stature et l'interception de la lumière sont des variables clés dans la compréhension de la performance. Nous montrons des effets antagonistes de ces différences, illustrant le compromis à faire entre les deux composantes du mélange, selon les objectifs recherchés dans la pratique des cultures associées (i.e. dominance de la céréale, de la légumineuse ou équilibre entre les deux espèces). Enfin, nos travaux contribuent à rendre plus robustes les connaissances déjà acquises sur ces cultures, notamment en terme de relations entre traits et performance.

Dans ce contexte, une perspective pour promouvoir les cultures associées serait de quantifier la fourniture d'autres services écosystémiques, en plus de la performance. Comme je l'ai écrit plus haut, les gammes de rendement dans lesquelles nous nous situons sont plutôt faibles, ce qui rend les cultures associées plus attractives dans les systèmes bas-intrants. Néanmoins, en comparaison avec les rendements atteints dans les systèmes conventionnels, les mélanges restent peu attractifs. Ainsi, prendre en compte les nombreux services fournis par les cultures associées (autres que la performance, par ex. la la stabilité du rendement, la structuration du sol, la réduction de l'utilisation de pesticides etc.), via par exemple une évaluation multi-critères (Naudin et al., 2014, Pelzer et al., 2014) permettrait de promouvoir leur utilisation. Cependant, ça impliquerait un certain changement de paradigme. Par exemple, on pourrait envisager une rémunération des services écosystémiques rendus par les agroécosystèmes entretenus par les agriculteurs et agricultrices (Swinton et al., 2007). Il faudrait également assurer des débouchés en aval aux cultures associées (tri des

grains moins "strict", reprise par l'industrie agro-alimentaire de mélanges pas à 100% pur pour de la transformation (Meynard et al., 2018), etc.). Ces changements sont néanmoins plus complexes à opérer, et vont au-delà du rôle strict de la recherche.

4 Évolution de ma perception de la recherche

4.1 Recherche et changements globaux

En dehors de mon activité de thèse, j'ai depuis quelques années un fort questionnement sur les différentes crises écologiques (changement climatique × extinction de la biodiversité × raréfaction des ressources) que le monde traverse. L'inaction des sociétés humaines (et individus) voire l'attitude de déni vis-à-vis des dangers/défis colossaux que représentent ces différentes crises sont pour moi, comme pour d'autres, un questionnement permanent. Mon point de vue sur ces crises est biaisé : mes interactions quotidiennes avec des spécialistes en agroécologie et mon intérêt pour ce domaine font que j'ai conscience du lien ténu existant entre la vie quotidienne et l'environnement (au sens large). Je cherche néanmoins à toujours prendre du recul sur ces sujets, en essayant de baser mes opinions sur des sources objectives (littérature scientifique/grise).
À une autre échelle, le monde de la recherche se pose de plus en plus la question de son impact environnemental. En France par exemple, le collectif Labos 1.5 [1], créé en 2019 par l'agronome Tamara Ben Ari et l'astrophysicien Olivier Berné a pour but de mener une étude scientifique nationale relative à l'empreinte carbone de la recherche publique française pour nourrir la réflexion sur les leviers d'action permettant de réduire son impact sur le climat et l'environnement. Au sein de ce collectif, des chercheur.e.s ont conçu un outil permettant d'évaluer le bilan Carbone (bilan C) d'une unité de recherche (Mariette et al., 2022). J'ai donc co-initié en 2020 avec Étienne-Pascal Journet le calcul du bilan Carbone de mon unité d'accueil (UMR AGIR) en utilisant cet outil. Nous avons réuni un groupe de travail au sein de l'unité afin de i) mettre à jour annuellement ce bilan C et ii) réduire notre empreinte C via l'établissement d'une charte environnementale pour l'unité. Dans la continuité de ce travail, j'ai établi un bilan Carbone simple des activités liées à ma thèse (méthodologie disponible dans les Annexes). Je profite ensuite de ce bilan pour proposer quelques éléments de réflexion sur le rôle de la recherche dans le contexte des changements globaux que nous vivons.

1. https ://labos1point5.org/

Bilan C de ma thèse

Sur trois ans et demi, mes activités de thèse ont émis 1.4 $teqCO_2$ (Figure V.1). À titre de comparaison, l'atteinte d'une neutralité carbone à l'échelle planétaire nécessiterait que chaque personne émette au maximum 2 $teqCO_2$ /an.

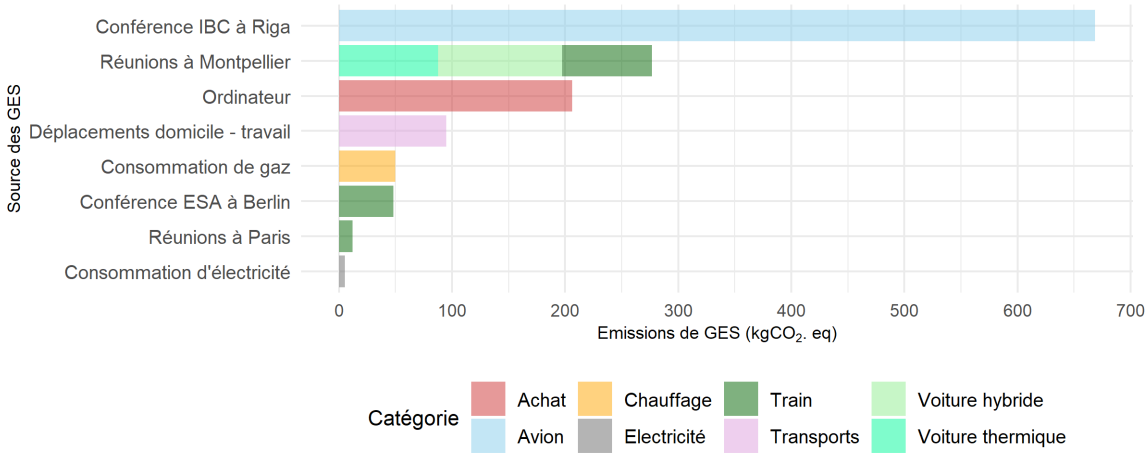


Figure V.1 – Bilan carbone de ma thèse, sur 3 ans et demi

L'impact de mes activités de recherche reste relativement faible en comparaison avec le bilan C de l'UMR AGIR en 2019 (3.6 $teqCO_2$ /agent/an). Ce relativement faible impact de mes activités s'explique par plusieurs aspects, parmi lesquels on peut citer i) l'absence d'expérimentations ou de déplacements sur le terrain, ii) la proximité entre mon domicile et mon lieu de travail, iii) l'absence de contraintes familiales me permettant de prendre plus de temps pour voyager (Berlin, Paris, en train) et iv) deux années de thèse pendant la pandémie de Covid19. Ce bilan C n'a pas pour but de souligner une démarche vertueuse de ma part, mais de servir d'ouverture sur des réflexions plus larges sur la pratique de la recherche. On remarque que le seul trajet en avion représente à lui seul 49% de mon bilan C.

Bénéfices du présentiel et impacts associés

Des travaux récents montrent que le prestige des chercheur.e.s, mesuré par le h-index et/ou le nombre de publications sur 3 ans, est directement lié à leur nombre de vols par an (Berné et al., 2022). Le sens de la causalité, si elle a lieu, entre ces variables n'est pas exploré (les chercheur.e.s souvent cité.e.s sont-ils fréquemment invité.e.s ou sont-ils connu.e.s parce qu'ils sont invité.e.s ?). Néanmoins, ce résultat interroge : une

recherche efficiente, productive et de qualité implique-t-elle le besoin de voyager vite et loin ?

J'ai pu vivre au cours de ma thèse la différence entre réunions/colloques en visio et en présentiel. J'ai par exemple participé à un colloque en distanciel et deux colloques en présentiel (IBC à Riga et ESA à Berlin). La différence était claire : la motivation, les interactions, les échanges étaient beaucoup plus fructueux en présentiel qu'en distanciel. J'ai pu, par exemple, lors de l'ESA en août 2022, être confronté à des visions de l'agriculture radicalement différentes de celles des chercheur.e.s d'AGIR (visions basées sur la robotisation et l'automatisation *vs* visions basées sur l'agrobiodiversité). Ainsi, le présentiel facilite les interactions et les échanges entre chercheur.e.s, éléments clés de la recherche. Une recherche plus durable pourrait donc être celle qui permet des rencontres tout en limitant les distances parcourues, celle favorisant les projets à plus petite échelle.

4.2 Quels rôle et avenir pour la recherche dans un monde fini ?

L'avantage du bilan C est qu'il est à la fois un proxy de l'impact que l'on a sur le réchauffement climatique mais également une mesure de notre dépendance aux énergies fossiles (les émissions de CO_2 étant en majeure partie dues à la combustion d'énergies fossiles). Il est nécessaire de souligner que la recherche scientifique à grande échelle existe grâce aux énergies fossiles. En effet, la croissance du secteur tertiaire (dont fait partie la recherche) observée au cours du XX$^{\text{ème}}$ siècle est probablement une conséquence directe de l'abondance des énergies fossiles, ces dernières ayant entraîné une diminution du nombre d'agriculteurs et d'ouvriers (*via* la mécanisation). Les énergies fossiles sont en quantités finies sur Terre : au-delà d'un certain moment (pic de production ; Hubbert, 1979), la quantité disponible de ces énergies est amenée à se réduire. Pour des raisons que je ne détaillerai pas ici, les énergies de substitution (géothermie, nucléaire, solaire, etc.) ne pourront probablement pas intégralement les remplacer. Penser l'impact de sa recherche est donc de fait une manière de penser une recherche réalisable à moyen/long terme.

La recherche scientifique a un rôle clé dans les questions de transitions. C'est particulièrement le cas en agronomie, puisque les changements globaux, l'extinction de la biodiversité et la raréfaction des énergies fossiles induisent de nouveaux défis (températures extrêmes, nouveaux ravageurs, chutes de populations d'insectes pollinisateurs, etc. ; Harchaoui et Chatzimpiros, 2018).

D'un autre côté, le risque est que la recherche scientifique entretienne l'idée selon laquelle l'ensemble des solutions aux problèmes globaux restent à trouver, où que les problèmes ne sont pas suffisamment compris ("more research is needed"). Pourtant, dans bien des situations, la recherche scientifique fait consensus sur la compréhension globale des problèmes, notamment pour le climat (Pörtner et al., 2022) ou l'extinction de la biodiversité (IPBES, 2022). La non-application de solutions existantes ou le déni des problèmes semblent d'ailleurs relever plus souvent de freins socio-techniques (Meynard et al., 2018) et de barrières psychologiques que d'un réel manque des connaissances des acteurs. L'autre risque qu'entretient la science est celui de la "solution magique" (Lamb et al., 2020). L'innovation et le progrès potentiels sont parfois le moteur de l'inaction. En donnant l'impression que les innovations scientifiques vont permettre de résoudre tous les problèmes, on retarde la mise en route d'un changement systémique. Pourquoi agir contre le changement climatique puisqu'on a découvert un gène du blé permettant la tolérance aux chaleurs et sécheresses extrêmes (Draeger et al., 2020) ou bien puisqu'on pourrait faire pousser du blé hors-sol dans des bâtiments en conditions contrôlées (lumière, concentration en CO_2), permettant de faire 6 récoltes par an et produire 40 fois plus par an (Asseng et al., 2020) ? Pourquoi arrêter/réduire l'usage de l'avion quand (un jour, peut-être), Airbus produira un avion à hydrogène ?

Dans ce contexte, je dois souligner que la recherche scientifique permet parfois de tempérer certaines lubies technosolutionnistes plus souvent issues d'initiatives privées (ex. géoingéniérie ; Bellamy et al., 2012), en rendant compte de la complexité du monde et de ses multiples interactions. Ainsi, contrairement à l'idée reçue selon laquelle les scientifiques seraient technosolutionnistes, j'ai plutôt l'impression que c'est l'opposé qui a lieu. La prise en compte des incertitudes et l'humilité vis-à-vis de ce qui est connu ou pas, propres à la recherche, font des scientifiques des personnes plutôt modérées. Si la recherche aura un rôle clé dans le devenir du XX$^{\text{ème}}$ siècle (du moins je l'espère), que ce soit en terme de réponses à des problèmes ou de description de phénomènes, les scientifiques doivent en connaître les limites, et communiquer dessus.

Références

Aare, A. K., Lund, S. & Hauggaard-Nielsen, H. (2021). Exploring transitions towards sustainable farming practices through participatory research – The case of Danish farmers' use of species mixtures. *Agricultural Systems*, *189*(July 2020). https://doi.org/10.1016/j.agsy.2021.103053

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, *16*(1), 3-9.

Acutis, M., Scaglia, B. & Confalonieri, R. (2012). Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. *European Journal of Agronomy*, *43*, 129-135. https://doi.org/10.1016/J.EJA.2012.06.006

Agreste [Consulté : 03/05/2022]. (2022).

Allen, I. E. & Olkin, I. (1999). Estimating Time to Conduct a Meta-analysis From Number of Citations Retrieved. *JAMA*, *282*(7), 634-635. https://doi.org/10.1001/JAMA.282.7.634

Anderson, C. (2008). The end of theory : The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16-07.

Annicchiarico, P., Collins, R. P., De Ron, A. M., Firmat, C., Litrico, I. & Hauggaard-Nielsen, H. (2019). *Do we need specific breeding for legume-based mixtures ?* (1re éd., T. 157). Elsevier Inc. https://doi.org/10.1016/bs.agron.2019.04.001

Archer, K. J. & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), 2249-2260. https://doi.org/10.1016/J.CSDA.2007.08.015

Arif, S. & Macneil, |. M. A. (2022). Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, e1554. https://doi.org/10.1002/ECM.1554

Asseng, S., Guarin, J. R., Raman, M., Monje, O., Kiss, G., Despommier, D. D., Meggers, F. M. & Gauthier, P. P. (2020). Wheat yield potential in controlled-environment vertical farms. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(32), 19131-19135. https://doi.org/https://doi.org/10.1073/pnas.2002655117

Barillot, R., Escobar-Gutiérrez, A. J., Fournier, C., Huynh, P. & Combes, D. (2014). Assessing the effects of architectural variations on light partitioning within virtual wheat–pea mixtures. *Annals of Botany*, *114*(4), 725-737. https://doi.org/10.1093/aob/mcu099

Barot, S., Allard, V., Cantarel, A., Enjalbert, J., Gauffreteau, A., Goldringer, I., Lata, J. C., Le Roux, X., Niboyet, A. & Porcher, E. (2017). Designing mixtures of varieties for multifunctional agriculture with the help of ecology. A review. *Agronomy for Sustainable Development*, *37*(2). https://doi.org/10.1007/s13593-017-0418-x

Barry, K. E., Mommer, L., van Ruijven, J., Wirth, C., Wright, A. J., Bai, Y., Connolly, J., De Deyn, G. B., de Kroon, H., Isbell, F., Milcu, A., Roscher, C., Scherer-Lorenzen, M., Schmid,

B. & Weigelt, A. (2019). The Future of Complementarity : Disentangling Causes from Consequences. *Trends in Ecology and Evolution*, *34*(2), 167-180. https://doi.org/10.1016/j.tree.2018.10.013

Baxevanos, D., Tsialtas, I. T., Vlachostergios, D., Hadjigeorgiou, I., Dordas, C. & Lithourgidis, A. (2017). Cultivar competitiveness in pea-oat intercrops under Mediterranean conditions. *Field Crops Research*, *214*(April), 94-103. https://doi.org/10.1016/j.fcr.2017.08.024

Bedoussac, L., Journet, E. P., Hauggaard-Nielsen, H., Naudin, C., Corre-Hellou, G., Jensen, E. S., Prieur, L. & Justes, E. (2015). Ecological principles underlying the increase of productivity achieved by cereal-grain legume intercrops in organic farming. A review. *Agronomy for Sustainable Development*, *35*(3), 911-935. https://doi.org/10.1007/s13593-014-0277-7

Bedoussac, L. & Justes, E. (2010a). Dynamic analysis of competition and complementarity for light and N use to understand the yield and the protein content of a durum wheat-winter pea intercrop. *Plant and Soil*, *330*(1), 37-54. https://doi.org/10.1007/S11104-010-0303-8/

Bedoussac, L. & Justes, E. (2010b). The efficiency of a durum wheat-winter pea intercrop to improve yield and wheat grain protein concentration depends on N availability during early growth. *Plant and Soil*, *330*(1), 19-35. https://doi.org/10.1007/S11104-009-0082-2

Bedoussac, L. & Justes, E. (2011). A comparison of commonly used indices for evaluating species interactions and intercrop efficiency : Application to durum wheat-winter pea intercrops. *Field Crops Research*, *124*(1), 25-36. https://doi.org/10.1016/j.fcr.2011.05.025

Begon, M., Townsend, C. R. & Harper, J. L. (2005). Ecology : From Individuals to Ecosystems, 4th Edition (4th, Éd.). *Blackwell Publishing*, 750.

Bellamy, R., Chilvers, J., Vaughan, N. E. & Lenton, T. M. (2012). A review of climate geoengineering appraisals. *Wiley Interdisciplinary Reviews : Climate Change*, *3*(6), 597-615. https://doi.org/https://doi.org/10.1002/wcc.197

Bellon Maurel, V., Brossard, L., Garcia, F., Mitton, N. & Termier, A. (2022). *Agriculture et numérique*. INRIA. https://doi.org/10.17180/wmkb-ty56

Benzécri, J.-P. et al. (1973). *L'analyse des données* (T. 2). Dunod Paris.

Berghuijs, H., Wang, Z., Stomph, T. J., Weih, M., Van der Werf, W. & Vico, G. (2020). Identification of species traits enhancing yield in wheat-faba bean intercropping : development and sensitivity analysis of a minimalist mixture model. *Plant and Soil*, *455*(1), 203-226. https://doi.org/10.1007/s11104-020-04668-0

Berghuijs, H., Weih, M., van der Werf, W., Karley, A. J., Adam, E., Villegas-Fernández, Á. M., Kiær, L. P., Newton, A. C., Scherber, C., Tavoletti, S. & Vico, G. (2021). Calibrating and testing APSIM for wheat-faba bean pure cultures and intercrops across Europe. *Field Crops Research*, *264*, 108088. https://doi.org/https://doi.org/10.1016/j.fcr.2021.108088

Berghuijs, H. N., Weih, M., van der Werf, W., Karley, A. J., Adam, E., Villegas-Fernández, Á. M., Kiær, L. P., Newton, A. C., Scherber, C., Tavoletti, S. & Vico, G. (2021). Calibrating and testing APSIM for wheat-faba bean pure cultures and intercrops across Europe. *Field Crops Research*, *264*, 108088. https://doi.org/10.1016/j.fcr.2021.108088

Bernard-Verdier, M., Navas, M.-L., Vellend, M., Violle, C., Fayolle, A. & Garnier, E. (2012). Community assembly along a soil depth gradient : contrasting patterns of plant trait convergence and divergence in a Mediterranean rangeland. *Journal of Ecology*, *100*(6), 1422-1433. https://doi.org/https://doi.org/10.1111/1365-2745.12003

Berné, O., Agier, L., Hardy, A., Lellouch, E., Aumont, O., Mariette, J. & Ben-Ari, T. (2022). The carbon footprint of scientific visibility. https://doi.org/10.1088/1748-9326/ac9b51

Bonke, V. & Musshoff, O. (2020). Understanding German farmer's intention to adopt mixed cropping using the theory of planned behavior. *Agronomy for Sustainable Development*, *40*(6), 1-14. https://doi.org/10.1007/S13593-020-00653-0/FIGURES/5

Bonneu, F., Makowski, D., Joly, J. & Allard, D. (2022). Machine Learning Based on Functional Principal Component Analysis to Identify Major Influential Factors of Wheat Yield. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4207476

Boudreau, M. A. (2013). Diseases in Intercropping Systems. *Annual review of phytopathology*, *51*, 499-519. https://doi.org/10.1146/ANNUREV-PHYTO-082712-102246

Brannan, T. (2021). *Barriers to Crop Diversification Practices in the European Union : A Narrative Synthesis* (mém. de mast.). Norwegian University of Life Sciences, Ås.

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45 :1*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F. X. & Huard, F. (2010). Why are wheat yields stagnating in Europe ? A comprehensive data analysis for France. *Field Crops Research*, *119*(1), 201-212. https://doi.org/10.1016/J.FCR.2010.07.012

Bron, C. & Kerbosch, J. (1973). Algorithm 457 : Finding All Cliques of an Undirected Graph [H]. *Communications of the ACM*, *16*(9), 575-577. https://doi.org/10.1145/362342.362367

Bronstein, J. L. (2009). The evolution of facilitation and mutualism. *Journal of Ecology*, *97*(6), 1160-1170. https://doi.org/https://doi.org/10.1111/j.1365-2745.2009.01566.x

Brooker, R. W., Bennett, A. E., Cong, W.-f., Daniell, T. J., George, T. S., Hallett, P. D., Hawes, C., Iannetta, P. P. M., Jones, H. G., Karley, A. J., Li, L., Mckenzie, B. M., Pakeman, J., Paterson, E., Sch, C., Shen, J., Squire, G., Watson, C. A., Zhang, C., ... White, P. J. (2015). Improving intercropping : a synthesis of research in agronomy , plant physiology and ecology, 107-117. https://doi.org/10.1111/nph.13132

Brooker, R. W., George, T. S., Zohralyn Homulle, |., Karley, A. J., Newton, A. C., Pakeman, R. J. & Schöb, C. (2021). Facilitation and biodiversity-ecosystem function relationships in crop production systems and their role in sustainable farming. *Journal of Ecology*, *00*, 1-14. https://doi.org/10.1111/1365-2745.13592

Cao, L. (2017). Data science : A comprehensive overview. *ACM Comput. Surv*, *50*(43). https://doi.org/10.1145/3076253

Capitaine, L., Genuer, R. & Thiébaut, R. (2021). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*(1), 166-184. https://doi.org/https://doi.org/10.1177/0962280220946080

Caranta, C., Monod, H., Berry, H., Chelle, M., Génard, M., Jourdan, F., Lannou, C., Maguin, E., Médale, F., Oddou-Muratorio, S., Rogel Gaillard, C. & Traas, J. (2019). *Réflexion prospective interdisciplinaire. Approches prédictives pour la biologie et l'écologie. Rapport de synthèse.* (Other). INRA. https://doi.org/10.15454/1.5783037069682676e12

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., MacE, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S. & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature 2012 486 :7401*, *486*(7401), 59-67. https://doi.org/10.1038/nature11148

Casler, M. D. (2015). Fundamentals of experimental design : Guidelines for designing successful experiments. *Agronomy Journal*, *107*(2), 692-705. https://doi.org/10.2134/AGRONJ2013.0114

Chenu, K., Porter, J. R., Martre, P., Basso, B., Chapman, S. C., Ewert, F., Bindi, M. & Asseng, S. (2017). Contribution of Crop Models to Adaptation in Wheat. *Trends in Plant Science*, *22*(6), 472-490. https://doi.org/10.1016/j.tplants.2017.02.003

Cheriere, T., Lorin, M. & Corre-Hellou, G. (2020). Species choice and spatial arrangement in soybean-based intercropping : Levers that drive yield and weed control. *Field Crops Research*, *256*(March), 107923. https://doi.org/10.1016/j.fcr.2020.107923

Cohen, J. E. (2004). Mathematics Is Biology's Next Microscope, Only Better ; Biology Is Mathematics' Next Physics, Only Better. *PLOS Biology*, *2*(12), e439. https://doi.org/10.1371/JOURNAL.PBIO.0020439

Colbach, N., Duby, C., Cavelier, A. & Meynard, J. M. (1997). Influence of cropping systems on foot and root diseases of winter wheat : Fitting of a statistical model. *European Journal of Agronomy*, *6*(1-2), 61-77. https://doi.org/10.1016/S1161-0301(96)02033-3

Coquillard, P., Hill, D. R. C. & Frontier, S. (1997). *Modélisation et simulation d'écosystèmes des modèles déterministes aux simulations à événements discrets* (B. Masson. Paris Milan, Éd.). Masson.

Cordell, D., Drangert, J. O. & White, S. (2009). The story of phosphorus : Global food security and food for thought. *Global Environmental Change*, *19*(2), 292-305. https://doi.org/10.1016/J.GLOENVCHA.2008.10.009

Corre-Hellou, G., Brisson, N., Launay, M., Fustec, J. & Crozat, Y. (2007). Effect of root depth penetration on soil nitrogen competitive interactions and dry matter production in pea–barley intercrops given different soil nitrogen supplies. *Field Crops Research*, *103*(1), 76-85. https://doi.org/10.1016/J.FCR.2007.04.008

Corre-Hellou, G., Fustec, J. & Crozat, Y. (2006). Interspecific Competition for Soil N and its Interaction with N2 Fixation, Leaf Expansion and Crop Growth in Pea–Barley Intercrops. *Plant and Soil 2006 282 :1*, *282*(1), 195-208. https://doi.org/10.1007/S11104-005-5777-4

Couëdel, A., Alletto, L., Tribouillois, H. & Justes, É. (2018). Cover crop crucifer-legume mixtures provide effective nitrate catch crop and nitrogen green manure ecosystem services. *Agriculture, Ecosystems and Environment*, *254*, 50-59. https://doi.org/10.1016/J.AGEE.2017.11.017

Cruz, S. M. S. d. & Nascimento, J. A. P. d. (2019). Towards integration of data-driven agronomic experiments with data provenance. *Computers and Electronics in Agriculture*, *161*(September 2018), 14-28. https://doi.org/10.1016/j.compag.2019.01.044

David, C., Jeuffroy, M. H., Henning, J. & Meynard, J. M. (2005). Yield variation in organic winter wheat : a diagnostic study in the Southeast of France. *Agronomy for Sustainable Development*, *25*(2), 213-223. https://doi.org/10.1051/AGRO:2005016

De Notaris, C., Rasmussen, J., Sørensen, P. & Olesen, J. E. (2018). Nitrogen leaching : A crop rotation perspective on the effect of N surplus, field management and use of catch crops. *Agriculture, Ecosystems and Environment*, *255*, 1-11. https://doi.org/10.1016/j.agee.2017.12.009

Dong, N., Tang, M. M., Zhang, W. P., Bao, X. G., Wang, Y., Christie, P. & Li, L. (2018). Temporal Differentiation of Crop Growth as One of the Drivers of Intercropping Yield Advantage. *Scientific Reports*, *8*(1), 1-11. https://doi.org/10.1038/s41598-018-21414-w

Draeger, T., C. Martin, A., Alabdullah, A. K., Pendle, A., Rey, M. D., Shaw, P. & Moore, G. (2020). Dmc1 is a candidate for temperature tolerance during wheat meiosis. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *133*(3), 809-828. https://doi.org/10.1007/S00122-019-03508-9

Duchene, O., Vian, J. F. & Celette, F. (2017). Intercropping with legume for agroecological cropping systems : Complementarity and facilitation processes and the importance of soil microorganisms. A review. https://doi.org/10.1016/j.agee.2017.02.019

Duru, M., Therond, O. & Fares, M. (2015a). Designing agroecological transitions ; A review. https://doi.org/10.1007/s13593-015-0318-x

Duru, M., Therond, O., Martin, G., Martin-Clouaire, R., Magne, M. A., Justes, E., Journet, E. P., Aubertot, J. N., Savary, S., Bergez, J. E. & Sarthou, J. P. (2015b). How to implement biodiversity-based agriculture to enhance ecosystem services : a review. https://doi.org/10.1007/s13593-015-0306-1

Ellis, J. L., Jacobs, M., Dijkstra, J., van Laar, H., Cant, J. P., Tulpan, D. & Ferguson, N. (2020). Review : Synergy between mechanistic modelling and data-driven models for modern animal production systems in the era of big data. *Animal*, *14*, s223-s237. https://doi.org/10.1017/S1751731120000312

Engbersen, N., Brooker, R. W., Stefan, L., Studer, B. & Schöb, C. (2021). Temporal Differentiation of Resource Capture and Biomass Accumulation as a Driver of Yield Increase in Intercropping. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.668803

Engbersen, N., Stefan, L., Brooker, R. W. & Schöb, C. (2022). Using plant traits to understand the contribution of biodiversity effects to annual crop community productivity. *Ecological Applications*, *32*(1), e02479. https://doi.org/10.1002/eap.2479

Erler, N. S., Rizopoulos, D. & Lesaffre, E. M. (2021). JointAI : Joint Analysis and Imputation of Incomplete Data in R. *Journal of Statistical Software*, *100*(20), 1-56. https://doi.org/10.18637/JSS.V100.I20

Evans, J. R. & Poorter, H. (2001). Photosynthetic acclimation of plants to growth irradiance : The relative importance of specific leaf area and nitrogen partitioning in maximizing carbon gain. *Plant, Cell and Environment*, *24*(8), 755-767. https://doi.org/10.1046/j.1365-3040.2001.00724.x

Evers, J. B., Van Der Werf, W., Stomph, T. J., Bastiaans, L. & Anten, N. P. (2019). Understanding and optimizing species mixtures using functional-structural plant modelling. *Journal of Experimental Botany*, *70*(9), 2381-2388. https://doi.org/10.1093/jxb/ery288

FAOSTAT Statistical Database [Consulté : 03/05/2022]. (2021).

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ? *Journal of Machine Learning Research*, *15*(90), 3133-3181. https://doi.org/https://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf

Figueroa, L. L., Grab, H., Ng, W. H., Myers, C. R., Graystock, P., McFrederick, Q. S. & McArt, S. H. (2020). Landscape simplification shapes pathogen prevalence in plant-pollinator networks (U. Brose, Éd.). *Ecology Letters*, *23*(8), 1212-1222. https://doi.org/10.1111/ele.13521

Fisher, R. (1935). Design of experiments by fisher.pdf. *Oliver and Boyd*.

Forst, E. (2018). Développement de méthodes d'estimation de l'aptitude au mélange pour la prédiction des performances et la sélection de mélanges variétaux chez le blé tendre et co-conception d'idéotypes de mélanges adaptés à l'agriculture biologique.

Fox, J. W. (2005). Interpreting the 'selection effect' of biodiversity on ecosystem function. *Ecology Letters*, *8*(8), 846-856. https://doi.org/10.1111/j.1461-0248.2005.00795.x

Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of statistics*, 1189-1232. https://doi.org/10.1214/aos/1013203451

Garnier, E., Navas, M. L., Austin, M. P., Lilley, J. M. & Gifford, R. M. (1997). A problem for biodiversity-productivity studies : How to compare the productivity of multispecific plant mixtures to that of monocultures ? *Acta Oecologica*, *18*(6), 657-670. https://doi.org/10.1016/S1146-609X(97)80049-5

Garnier, E. & Navas, M. L. (2012). A trait-based approach to comparative functional plant ecology : Concepts, methods and applications for agroecology. A review. *Agronomy for Sustainable Development*, *32*(2), 365-399. https://doi.org/10.1007/S13593-011-0036-Y/FIGURES/7

Garside, A. L. & Bell, M. J. (2011). Growth and yield responses to amendments to the sugarcane monoculture : Towards identifying the reasons behind the response to breaks. *Crop and Pasture Science*, *62*(9), 776-789. https://doi.org/10.1071/CP11055

Gaudio, N., Louarn, G., Barillot, R., Meunier, C., Vezy, R. & Launay, M. (2022). Exploring complementarities between modelling approaches that enable upscaling from plant community functioning to ecosystem services as a way to support agroecological transition. *in silico Plants*, *4*(1). https://doi.org/10.1093/INSILICOPLANTS/DIAB037

Gaudio, N., Escobar-Gutiérrez, A. J., Casadebaig, P., Evers, J. B., Gérard, F., Louarn, G., Colbach, N., Munz, S., Launay, M., Marrou, H., Barillot, R., Hinsinger, P., Bergez, J.-E., Combes, D., Durand, J.-L., Frak, E., Pagès, L., Pradal, C., Saint-Jean, S., ... Justes, E. (2019). Current knowledge and future research opportunities for modeling annual crop mixtures. A review. *Agronomy for Sustainable Development*, *39*(2), 20. https://doi.org/10.1007/s13593-019-0562-6

Gaudio, N., Violle, C., Gendre, X., Fort, F., Mahmoud, R., Pelzer, E., Médiène, S., Hauggaard-Nielsen, H., Bedoussac, L., Bonnet, C., Corre-Hellou, G., Couëdel, A., Hinsinger, P., Steen Jensen, E., Journet, E. P., Justes, E., Kammoun, B., Litrico, I., Moutier, N., ... Casadebaig, P. (2021a). Interspecific interactions regulate plant reproductive allometry in cereal–legume intercropping systems. *Journal of Applied Ecology*, *58*(11), 2579-2589. https://doi.org/10.1111/1365-2664.13979

Gaudio, N., Violle, C., Gendre, X., Fort, F., Mahmoud, R., Pelzer, E., Médiène, S., Hauggaard-Nielsen, H., Bedoussac, L., Bonnet, C., Corre-Hellou, G., Couëdel, A., Hinsinger, P., Steen Jensen, E., Journet, E.-P., Justes, E., Kammoun, B., Litrico, I., Moutier, N., ... Casadebaig, P. (2021b). Interspecific interactions regulate plant reproductive allometry in cereal-legume intercropping systems, Dryad, Dataset [available at : https://datadryad.org/stash/dataset/doi:10.5061/dryad.9ghx3ffhv (May. 2022)].

Ghaley, B. B., Hauggaard-Nielsen, H., Høgh-Jensen, H. & Jensen, E. S. (2005). Intercropping of wheat and pea as influenced by nitrogen fertilization. *Nutrient Cycling in Agroecosystems*, *73*(2-3), 201-212. https://doi.org/10.1007/s10705-005-2475-9

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, *5*(10), 3-8. https://doi.org/https://doi.org/10.2307/1174772

Goldberg, D. E. (1990). Components of resource competition in plant communities. *Perspectives on plant competition*, 27-49.

Gonthier, D. J., Ennis, K. K., Farinas, S., Hsieh, H.-Y., Iverson, A. L., Batáry, P., Rudolphi, J., Tscharntke, T., Cardinale, B. J. & Perfecto, I. (2014). Biodiversity conservation in agriculture requires a multi-scale approach. *Proceedings of the Royal Society B : Biological Sciences*, *281*(1791), 20141358. https://doi.org/10.1098/rspb.2014.1358

Gooding, M. J., Kasyanova, E., Ruske, R., Hauggaard-Nielsen, H., Jensen, E. S., Dahlmann, C., Von Fragstein, P., Dibet, A., Corre-Hellou, G., Crozat, Y., Pristeri, A., Romeo, M., Monti, M. & Launay, M. (2007). Intercropping with pulses to concentrate nitrogen and sulphur in

wheat. *Journal of Agricultural Science*, *145*(5), 469-479. https://doi.org/10.1017/S0021859607007241

Gosseau, F., Blanchet, N., Varès, D., Burger, P., Campergue, D., Colombet, C., Gody, L., Liévin, J. F., Mangin, B., Tison, G., Vincourt, P., Casadebaig, P. & Langlade, N. (2019). Heliaphen, an outdoor high-throughput phenotyping platform for genetic studies and crop modeling. *Frontiers in Plant Science*, *9*, 1908. https://doi.org/https://doi.org/10.3389/fpls.2018.01908

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data ? https://doi.org/10.48550/ARXIV.2207.08815

Gu, C., Bastiaans, L., Anten, N. P., Makowski, D. & van der Werf, W. (2021). Annual intercropping suppresses weeds : A meta-analysis. *Agriculture, Ecosystems & Environment*, *322*, 107658. https://doi.org/10.1016/J.AGEE.2021.107658

Guillemaud, T., Facon, B. & Bourguet, D. (2019). Peer Community In : A free process for the recommendation of unpublished scientific papers based on peer review. *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées*. https://doi.org/10.4000/PROCEEDINGS.ELPUB.2019.23

Guinet, M., Nicolardot, B., Revellin, C., Durey, V., Carlsson, G. & Voisin, A. S. (2018). Comparative effect of inorganic N on plant growth and N2 fixation of ten legume crops : towards a better understanding of the differential response among species. *Plant and Soil*, *432*(1-2), 207-227. https://doi.org/10.1007/s11104-018-3788-1

Gunawardena, J. (2014). Models in biology : 'Accurate descriptions of our pathetic thinking'. *BMC Biology*, *12*(1), 1-11. https://doi.org/https://doi.org/10.1186/1741-7007-12-29

Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature 2018 555 :7695*, *555*(7695), 175-182. https://doi.org/10.1038/nature25753

Gurr, G. M., Lu, Z., Zheng, X., Xu, H., Zhu, P., Chen, G., Yao, X., Cheng, J., Zhu, Z., Catindig, J. L., Villareal, S., Van Chien, H., Cuong, L. Q., Channoo, C., Chengwattana, N., Lan, L. P., Hai, L. H., Chaiwong, J., Nicol, H. I., ... Heong, K. L. (2016). Multi-country evidence that crop diversification promotes ecological intensification of agriculture. *Nature Plants 2016 2 :3*, *2*(3), 1-4. https://doi.org/10.1038/nplants.2016.14

Haccoun, R. R. & Cousineau, D. (2009). Statistiques : Concepts et applications (2e édition)., 462.

Hajjem, A., Bellavance, F. & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313-1328. https://doi.org/10.1080/00949655.2012.741599

Han, Z. Q., Liu, T., Liu, H. F., Hao, X. R., Chen, W. & Li, B. L. (2019). Derivation of species interactions strength in a plant community with game theory. *Ecological Modelling*, *394*, 27-33. https://doi.org/10.1016/J.ECOLMODEL.2018.12.018

Harchaoui, S. & Chatzimpiros, P. (2018). Can agriculture balance its energy consumption and continue to produce food ? A framework for assessing energy neutrality applied to French agriculture. *Sustainability (Switzerland)*, *10*(12). https://doi.org/10.3390/su10124624

Hauggaard-Nielsen, H., Gooding, M., Ambus, P., Corre-Hellou, G., Crozat, Y., Dahlmann, C., Dibet, A., von Fragstein, P., Pristeri, A., Monti, M. & Jensen, E. S. (2009). Pea-barley intercropping for efficient symbiotic N2-fixation, soil N acquisition and use of other nutrients in European organic cropping systems. *Field Crops Research*, *113*(1), 64-71. https://doi.org/10.1016/j.fcr.2009.04.009

Hauggaard-Nielsen, H., Jørnsgaard, B., Kinane, J. & Jensen, E. S. (2008). Grain legume - Cereal intercropping : The practical application of diversity, competition and facilitation in arable

and organic cropping systems. *Renewable Agriculture and Food Systems*, *23*(1), 3-12. https://doi.org/10.1017/S1742170507002025

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280-299. https://doi.org/10.1353/lib.0.0036

Homulle, Z., George, T. S. & Karley, A. J. (2021). Root traits with team benefits : understanding belowground interactions in intercropping systems. *Plant and Soil*. https://doi.org/10.1007/s11104-021-05165-8

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J. H., Lodge, D. M., Loreau, M., Naeem, S., Schmid, B., Setälä, H., Symstad, A. J., Vandermeer, J. & Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning : a consensus of current knowledge. *Ecological Monographs*, *75*(1), 3-35. https://doi.org/10.1890/04-0922

Houlahan, J. E., McKinney, S. T., Anderson, T. M. & McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1-7. https://doi.org/10.1111/OIK.03726

Hubbert, M. K. (1979). Hubbert estimates from 1956 to 1974 of us oil and gas. In M. GRENON (Éd.), *Methods and Models for Assessing Energy Resources* (p. 370-383). Pergamon. https://doi.org/https://doi.org/10.1016/B978-0-08-024443-3.50038-8

Huck, S. W., Cormier, W. H. & Bounds, W. G. (1974). *Reading statistics and research*. Harper & Row New York.

Huth, N., Thorburn, P., Radford, B. & Thornton, C. (2010). Impacts of fertilisers and legumes on N2O and CO2 emissions from soils in subtropical agricultural systems : A simulation study [Estimation of nitrous oxide emission from ecosystems and its mitigation technologies]. *Agriculture, Ecosystems & Environment*, *136*(3), 351-357. https://doi.org/https://doi.org/10.1016/j.agee.2009.12.016

IPBES. (2022). *Thematic Assessment Report on the Sustainable Use of Wild Species of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (J. M. Fromentin, M. R. Emery, J. Donaldson, M. Danner, A. Hallosserie & D. Kieling, Éd.). https://doi.org/10.5281/zenodo.6448567

IPCC. (2013). *Climate Change 2013 : The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

Jain, Y., Ding, S. & Qiu, J. (2019). Sliced inverse regression for integrative multi-omics data analysis. *Statistical applications in genetics and molecular biology*, *18*(1). https://doi.org/https://doi.org/10.1515/sagmb-2018-0028

Jensen, E. S., Carlsson, G. & Hauggaard-Nielsen, H. (2020). Intercropping of grain legumes and cereals improves the use of soil N resources and reduces the requirement for synthetic fertilizer N : A global-scale analysis. *Agronomy for Sustainable Development*, *40*(1), 1-9. https://doi.org/10.1007/s13593-020-0607-x

Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S., Keating, B. A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C. & Wheeler, T. R. (2017). Brief history of agricultural systems modeling. *Agricultural Systems*, *155*, 240-254. https://doi.org/https://doi.org/10.1016/j.agsy.2016.05.014

Justes, E., Mary, B., Meynard, J. M., Machet, J. M. & Thelier-Huche, L. (1994). Determination of a Critical Nitrogen Dilution Curve for Winter Wheat Crops. https://doi.org/10.1006/anbo.1994.1133

Justes, E., Bedoussac, L., Corre-Hellou, G., Fustec, J., Hinsinger, P., Jeuffroy, M.-H., Journet, E.-P., Louarn, G., Naudin, C. & Pelzer, E. (2014). *Les processus de complémentarité de niche et de facilitation déterminent le fonctionnement des associations végétales et leur efficacité pour l'acquisition des ressources abiotiques* (rapp. tech.).

Justes, E., Bedoussac, L., Dordas, C., Frak, E., Louarn, G., Boudsocq, S., Journet, E. P., Lithourgidis, A., Pankou, C., Zhang, C., Carlsson, G., Jensen, E. S., Watson, C. & Li, L. (2021). The 4C approach as a way to understand species interactions determining intercropping productivity. *Frontiers of Agricultural Science and Engineering*, *8*(3), 387-399. https://doi.org/10.15302/J-FASE-2021414

Kamilaris, A., Kartakoullis, A. & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*(October), 23-37. https://doi.org/10.1016/j.compag.2017.09.037

Kammoun, B., Journet, E.-P., Justes, E. & Bedoussac, L. (2021). Cultivar Grain Yield in Durum Wheat-Grain Legume Intercrops Could Be Estimated From Sole Crop Yields and Interspecific Interaction Index. *Frontiers in Plant Science*, *12*(October), 1-14. https://doi.org/10.3389/fpls.2021.733705

Kattge, J., Diaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J. et al. (2011). TRY–a global database of plant traits. *Global change biology*, *17*(9), 2905-2935.

Keating, B. A. & Thorburn, P. J. (2018). Modelling crops and cropping systems—Evolving purpose, practice and prospects. *European Journal of Agronomy*, *100*, 163-176. https://doi.org/10.1016/J.EJA.2018.04.007

Kiær, L. P., Skovgaard, I. M. & Østergård, H. (2009). Grain yield increase in cereal variety mixtures : A meta-analysis of field trials. *Field Crops Research*, *114*(3), 361-373. https://doi.org/https://doi.org/10.1016/j.fcr.2009.09.006

Kitchin, R. & McArdle, G. (2016). What makes Big Data, Big Data ? Exploring the ontological characteristics of 26 datasets : *http ://dx.doi.org/10.1177/2053951716631130*, *3*(1). https://doi.org/10.1177/2053951716631130

Knapp, S. & van der Heijden, M. G. (2018). A global meta-analysis of yield stability in organic and conservation agriculture. *Nature Communications 2018 9 :1*, *9*(1), 1-9. https://doi.org/10.1038/s41467-018-05956-1

Knott, E. A. & Mundt, C. C. (1990). Mixing ability analysis of wheat cultivar mixtures under diseased and nondiseased conditions. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *80*(3), 313-320. https://doi.org/10.1007/BF00210065

Knudsen, M. T., Hauggaard-Nielsen, H., Jørnsgård, B. & Steen Jensen, E. (2004). Comparison of interspecific competition and N use in pea-barley, faba bean-barley and lupin-barley intercrops grown at two temperate locations. *Journal of Agricultural Science*, *142*(6), 617-627. https://doi.org/10.1017/S0021859604004745

Koocheki, A., Nassiri, M., Alimoradi, L. & Ghorbani, R. (2009). Effect of cropping systems and crop rotations on weeds. *Agronomy for Sustainable Development*, *29*(2), 401-408. https://doi.org/10.1051/agro/2008061

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270-280. https://doi.org/10.1177/2515245918771304

Kunstler, G., Lavergne, S., Courbaud, B., Thuiller, W., Vieilledent, G., Zimmermann, N. E., Kattge, J. & Coomes, D. A. (2012). Competitive interactions between forest trees are driven by species'

trait hierarchy, not phylogenetic or functional similarity : Implications for forest community assembly. *Ecology Letters*, *15*(8), 831-840. https://doi.org/10.1111/j.1461-0248.2012.01803.x

Kursa, M. B. & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1-13. https://doi.org/10.18637/JSS.V036.I11

Lamb, W. F., Mattioli, G., Levi, S., Roberts, J. T., Capstick, S., Creutzig, F., Minx, J. C., Müller-Hansen, F., Culhane, T. & Steinberger, J. K. (2020). Global Sustainability cambridge.org/sus Intelligence Briefing Discourses of climate delay. https://doi.org/10.1017/sus.2020.13

Lamichhane, J. R. & Alletto, L. (2022). Ecosystem services of cover crops : a research roadmap. *Trends in Plant Science*, *27*(8), 758-768. https://doi.org/10.1016/J.TPLANTS.2022.03.014

Lamichhane, J. R., Constantin, J., Schoving, C., Maury, P., Debaeke, P., Aubertot, J.-N. & Dürr, C. (2020). Analysis of soybean germination, emergence, and prediction of a possible northward establishment of the crop under climate change. *European Journal of Agronomy*, *113*, 125972. https://doi.org/https://doi.org/10.1016/j.eja.2019.125972

Launay, M., Brisson, N., Satger, S., Hauggaard-Nielsen, H., Corre-Hellou, G., Kasynova, E., Ruske, R., Jensen, E. S. & Gooding, M. J. (2009). Exploring options for managing strategies for pea–barley intercropping using a modeling approach. *European Journal of Agronomy*, *31*(2), 85-98. https://doi.org/10.1016/J.EJA.2009.04.002

Lawler, E. L., Lenstra, J. K. & Rinnooy Kan, A. H. G. (1980). Generating All Maximal Independent Sets : NP-Hardness and Polynomial-Time Algorithms. *SIAM Journal on Computing*, *9*(3), 558-565. https://doi.org/10.1137/0209042

Lee, D. S., Fahey, D. W., Skowron, A., Allen, M. R., Burkhardt, U., Chen, Q., Doherty, S. J., Freeman, S., Forster, P. M., Fuglestvedt, J., Gettelman, A., De León, R. R., Lim, L. L., Lund, M. T., Millar, R. J., Owen, B., Penner, J. E., Pitari, G., Prather, M. J., ... Wilcox, L. J. (2021). The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmospheric Environment*, *244*, 117834. https://doi.org/10.1016/J.ATMOSENV.2020.117834

Lemaire, G. & Meynard, J. M. (1997). Use of the Nitrogen Nutrition Index for the Analysis of Agronomical Data. *Diagnosis of the Nitrogen Status in Crops*, 45-55. https://doi.org/10.1007/978-3-642-60684-7{_}2

Li, C., Hoffland, E., Kuyper, T. W., Yu, Y., Li, H., Zhang, C., Zhang, F. & van der Werf, W. (2020a). Yield gain, complementarity and competitive dominance in intercropping in China : A meta-analysis of drivers of yield gain using additive partitioning. *European Journal of Agronomy*, *113*(November 2019), 125987. https://doi.org/10.1016/j.eja.2019.125987

Li, C., Hoffland, E., Kuyper, T. W., Yu, Y., Zhang, C., Li, H., Zhang, F. & van der Werf, W. (2020b). Syndromes of production in intercropping impact yield gains. *Nature Plants*, *6*(6), 653-660. https://doi.org/10.1038/s41477-020-0680-9

Li, C., Hoffland, E., Kuyper, T. W., Yu, Y., Zhang, C., Li, H., Zhang, F. & van der Werf, W. (2020c). Syndromes of production in intercropping impact yield gains. *Nature Plants*, *6*(6), 653-660. https://doi.org/10.1038/s41477-020-0680-9

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316-327. https://doi.org/10.1080/01621459.1991.10475035

Licker, R., Johnston, M., Foley, J. A., Barford, C., Kucharik, C. J., Monfreda, C. & Ramankutty, N. (2010). Mind the gap : How do climate and agricultural management explain the 'yield gap' of croplands around the world ? *Global Ecology and Biogeography*, *19*(6), 769-782. https://doi.org/10.1111/j.1466-8238.2010.00563.x

Litrico, I. & Violle, C. (2015). Diversity in Plant Breeding : A New Conceptual Framework. *Trends in Plant Science*, *20*(10), 604-613. https://doi.org/10.1016/j.tplants.2015.07.007

Lobell, D. B., Deines, J. M. & Tommaso, S. D. (2020). Changes in the drought sensitivity of US maize yields. *Nature Food*, *1*(11), 729-735. https://doi.org/10.1038/s43016-020-00165-w

Long, L., Tilman, D., Lambers, H. & Zhang, F.-S. (2014). Plant diversity and overyielding : insights from belowground facilitation of intercropping in agriculture. *New Phytologist*, *203*(1), 63-69. https://doi.org/https://doi.org/10.1111/nph.12778

Loreau, M. & Hector, A. (2001). Partitioning selection and complementarity in biodiversity experiments. *412*(July), 5.

Loreau, M., Naeem, S. & Inchausti, P. (2002). Biodiversity and ecosystem functioning : synthesis and perspectives, 294.

Louarn, G., Barillot, R., Combes, D. & Escobar-Gutiérrez, A. (2020). Towards intercrop ideotypes : Non-random trait assembly can promote overyielding and stability of species proportion in simulated legume-based mixtures. *Annals of Botany*, *126*(4), 671-685. https://doi.org/10.1093/aob/mcaa014

Louarn, G., Bedoussac, L., Gaudio, N., Journet, E. P., Moreau, D., Steen Jensen, E. & Justes, E. (2021). Plant nitrogen nutrition status in intercrops– a review of concepts and methods. *European Journal of Agronomy*, *124*. https://doi.org/10.1016/j.eja.2021.126229

Lv, W., Zhao, X., Wu, P., Lv, J. & He, H. (2021). A Scientometric Analysis of Worldwide Intercropping Research Based on Web of Science Database between 1992 and 2020. https://doi.org/https://doi.org/10.3390/su13052430

Maat, H. (2011). The history and future of agricultural experiments. *NJAS - Wageningen Journal of Life Sciences*, *57*(3-4), 187-195. https://doi.org/10.1016/j.njas.2010.11.001

Machado, S. (2009). Does intercropping have a role in modern agriculture ? https://doi.org/10.2489/jswc.64.2.55A

Magrini, M., Anton, M., Cholez, C., Corre-Hellou, G., Duc, G., Jeuffroy, M. H., Meynard, J. M., Pelzer, E., Voisin, A. S. & Walrand, S. (2016). Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits ? Analyzing lock-in in the French agrifood system. *Ecological Economics*, *126*(June 2016), 152-162. https://doi.org/10.1016/j.ecolecon.2016.03.024

Magrini, M., Triboulet, P. & Bedoussac, L. (2013). Pratiques agricoles innovantes et logistique des coopératives agricoles. Une étude ex-ante sur l'acceptabilité de cultures associées blé dur-légumineuses. *http ://journals.openedition.org/economierurale*, (338), 25-45. https://doi.org/10.4000/ECONOMIERURALE.4145

Mahmoud, R., Casadebaig, P., Hilgert, N., Alletto, L., Freschet, G. T., de Mazancourt, C. & Gaudio, N. (2022). Species choice and N fertilization influence yield gains through complementarity and selection effects in cereal-legume intercrops. *Agronomy for Sustainable Development*, *42*(2). https://doi.org/10.1007/s13593-022-00754-y

Maier, D. S. (2012). Theories of Biodiversity Value. *International Library of Environmental, Agricultural and Food Ethics*, *19*, 159-307. https://doi.org/10.1007/978-94-007-3991-8{_}6

Maitra, S., Hossain, A., Brestic, M., Skalicky, M., Ondrisik, P. & Gitari, H. (2021). Intercropping — A Low Input Agricultural Strategy for Food and Environmental Security, 1-29. https://doi.org/https://doi.org/10.3390/agronomy11020343

Makowski, D., Nesme, T., Papy, F. & Doré, T. (2014). Global agronomy, a new field of research. A review. *Agronomy for Sustainable Development*, *34*(2), 293-307. https://doi.org/10.1007/s13593-013-0179-0

Makowski, D., Piraux, F. & Brun, F. (2019). From Experimental Network to Meta-analysis Methods and Applications with R for Agronomic and Environmental Sciences.

Mamine, F. & Farès, M. (2020). Barriers and Levers to Developing Wheat–Pea Intercropping in Europe : A Review. *Sustainability*, *12*(17). https://doi.org/10.3390/su12176962

Mao, L.-L., Zhang, L.-Z., Zhang, S.-P., Evers, J. B., van der Werf, W., Wang, J.-J., Sun, H.-Q., Su, Z.-C. & Spiertz, H. (2015). Resource use efficiency, ecological intensification and sustainability of intercropping systems. *Journal of Integrative Agriculture*, *14*(8), 1542-1550. https://doi.org/https://doi.org/10.1016/S2095-3119(15)61039-5

Mariette, J., Blanchard, O., Berné, O., Aumont, O., Carrey, J., Ligozat, A., Lellouch, E., Roche, P.-E., Guennebaud, G., Thanwerdas, J. et al. (2022). An open-source tool to assess the carbon footprint of research. *Environmental Research : Infrastructure and Sustainability*, *2*(3), 035008. https://doi.org/10.1088/2634-4505/ac84a4

Mariotti, M., Masoni, A., Ercoli, L. & Arduini, I. (2009). Above- and below-ground competition between barley, wheat, lupin and vetch in a cereal and legume intercropping system. *Grass and Forage Science*, *64*(4), 401-412. https://doi.org/10.1111/J.1365-2494.2009.00705.X

Maris, V., Huneman, P., Coreau, A., Kéfi, S., Pradel, R. & Devictor, V. (2018). Prediction in ecology : promises, obstacles and clarifications. *Oikos*, *127*(2), 171-183. https://doi.org/https://doi.org/10.1111/oik.04655

Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation : current practice and guidelines. *BMC Medical Research Methodology*, *9*(1), 57. https://doi.org/10.1186/1471-2288-9-57

Martin-Clouaire, R. (2017). Modelling Operational Decision-Making in Agriculture. *Agricultural Sciences*, *8*(7), 527-544. https://doi.org/10.4236/AS.2017.87040

Martin-Guay, M. O., Paquette, A., Dupras, J. & Rivest, D. (2018). The new Green Revolution : Sustainable intensification of agriculture by intercropping. *Science of the Total Environment*, *615*, 767-772. https://doi.org/10.1016/j.scitotenv.2017.10.024

Mazzocchi, F. (2015). Could Big Data be the end of theory in science ? *EMBO reports*, *16*(10), 1250-1255. https://doi.org/https://doi.org/10.15252/embr.201541001

McKane, R. B., Johnson, L. C., Shaver, G. R., Nadelhoffer, K. J., Rastetter, E. B., Fry, B., Giblin, A. E., Kielland, K., Kwlatkowski, B. L., Laundre, J. A. & Murray, G. (2002). Resource-based niches provide a basis for plant species diversity and dominance in arctic tundra. *Nature 2002 415 :6867*, *415*(6867), 68-71. https://doi.org/10.1038/415068a

Mead, R. & Willey, R. W. (1980). The concept of a 'land equivalent ratio' and advantages in yields from intercropping. *16*, 217-228. https://doi.org/10.1017/S0014479700010978

Meunier, C., Alletto, L., Bedoussac, L., Bergez, J. E., Casadebaig, P., Constantin, J., Gaudio, N., Mahmoud, R., Aubertot, J. N., Celette, F., Guinet, M., Jeuffroy, M. H., Robin, M. H., Médiène, S., Fontaine, L., Nicolardot, B., Pelzer, E., Souchère, V., Voisin, A. S., ... Martin, G. (2022a). A modelling chain combining soft and hard models to assess a bundle of ecosystem services provided by a diversity of cereal-legume intercrops. *European Journal of Agronomy*, *132*(October 2021). https://doi.org/10.1016/j.eja.2021.126412

Meunier, C., Casagrande, M., Rosiès, B., Bedoussac, L., Topp, C. F., Walker, R. L., Watson, C. A. & Martin, G. (2022b). Interplay : A game for the participatory design of locally adapted cereal–legume intercrops. *Agricultural Systems*, *201*, 103438. https://doi.org/10.1016/J.AGSY.2022.103438

Meynard, J.-M., Charrier, F., Fares, M., Le Bail, M., Magrini, M.-B., Charlier, A. & Messéan, A. (2018). Socio-technical lock-in hinders crop diversification in France. *Agronomy for Sustainable Development*, *38*(5), 1-13. https://doi.org/10.1007/s13593-018-0535-1

Michener, W. K. & Jones, M. B. (2012). Ecoinformatics : supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85-93. https://doi.org/10.1016/J.TREE.2011.11.016

Montazeaud, G., Violle, C., Fréville, H., Luquet, D., Ahmadi, N., Courtois, B., Bouhaba, I. & Fort, F. (2018). Crop mixtures : does niche complementarity hold for belowground resources ? An experimental test using rice genotypic pairs. *Plant and Soil*, *424*(1-2), 187-202. https://doi.org/10.1007/s11104-017-3496-2

Montazeaud, G., Violle, C., Roumet, P., Rocher, A., Ecarnot, M., Compan, F., Maillet, G., Fort, F. & Fréville, H. (2020). Multifaceted functional diversity for multifaceted crop yield : Towards ecological assembly rules for varietal mixtures. *Journal of Applied Ecology*, *57*(11), 2285-2295. https://doi.org/10.1111/1365-2664.13735

Morel, K., Revoyron, E., Cristobal, M. S. & Baret, P. V. (2020). Innovating within or outside dominant food systems ? Different challenges for contrasting crop diversification strategies in Europe. *PLoS ONE*, *15*(3), 1-24. https://doi.org/10.1371/journal.pone.0229910

Moutier, N., Baranger, A., Fall, S., Hanocq, E., Marget, P., Floriot, M. & Gauffreteau, A. (2022). Mixing Ability of Intercropped Wheat Varieties : Stability Across Environments and Tester Legume Species. *Frontiers in Plant Science*, *13*, 1495. https://doi.org/10.3389/FPLS.2022.877791/BIBTEX

Nasr, N. (2021). Overcoming the discourse of science mistrust : how science education can be used to develop competent consumers and communicators of science information. *Cultural Studies of Science Education*, *16*(2), 345-356. https://doi.org/10.1007/S11422-021-10064-6/FIGURES/2

Naudin, C., Corre-Hellou, G., Pineau, S., Crozat, Y. & Jeuffroy, M.-H. (2010). The effect of various dynamics of N availability on winter pea–wheat intercrops : Crop growth, N partitioning and symbiotic N2 fixation. *Field Crops Research*, *119*(1), 2-11. https://doi.org/https://doi.org/10.1016/j.fcr.2010.06.002

Naudin, C., Van Der Werf, H. M., Jeuffroy, M. H. & Corre-Hellou, G. (2014). Life cycle assessment applied to pea-wheat intercrops : A new method for handling the impacts of co-products. *Journal of Cleaner Production*, *73*, 80-87. https://doi.org/10.1016/j.jclepro.2013.12.029

Newman, S. J. & Furbank, R. T. (2021). A multiple species, continent-wide, million-phenotype agronomic plant dataset. *Scientific Data*, *8*(1), 1-8. https://doi.org/10.1038/s41597-021-00898-8

Neyton, S., Abbady, D. & Jean-Pierre, S. (2018). Agroécosystème : Définition [Consulté : 05/10/2022].

Paff, K., Munz, S., Vezy, R., Gaudio, N., Bedoussac, L. & Justes, E. (2020). Calibration and Evaluation of the STICS Intercrop Model for Two Cereal-Legume Mixtures. *ICROPM 2020 - Crop Modelling for the future*.

Pappagallo, S., Brandmeier, J., Banfield-zanin, J. A., Kiær, L., Shaw, P., Raubach, S., Karley, A. J. & Scherber, C. (2021). Coordinating data collection in intercropping : A feasible example. *Aspects of Applied Biology*, (146), 249-256.

Parris, K. (2011). Impact of agriculture on water pollution in OECD countries : Recent trends and future prospects. *International Journal of Water Resources Development*, *27*(1), 33-52. https://doi.org/10.1080/07900627.2010.531898

Patrician, P. A. (2002). Multiple imputation for missing data. *Research in Nursing and Health*, *25*(1), 76-84. https://doi.org/10.1002/NUR.10015

Pearl, J. (2009). Causal inference in statistics : An overview. *https ://doi.org/10.1214/09-SS057*, *3*(none), 96-146. https://doi.org/10.1214/09-SS057

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572. https://doi.org/10.1080/14786440109462720

Pearson, R. & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species : are bioclimate envelope models useful ? *Global Ecology and Biogeography*, *12*(5), 361-371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Pellagatti, M., Masci, C., Ieva, F. & Paganoni, A. M. (2021). Generalized mixed-effects random forest : A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining*, *14*(3), 241-257. https://doi.org/10.1002/sam.11505

Pelzer, E., Bazot, M., Guichard, L. & Jeuffroy, M. H. (2016). Crop management affects the performance of a winter pea–wheat intercrop. *Agronomy Journal*, *108*(3), 1089-1100. https://doi.org/10.2134/agronj2015.0440

Pelzer, E., Bazot, M., Makowski, D., Corre-hellou, G., Naudin, C., Al, M., Baranger, E., Bedoussac, L., Biarnès, V., Boucheny, P., Carrouée, B., Dorvillez, D., Foissy, D., Gaillard, B., Guichard, L., Mansard, M.-c., Omon, B., Prieur, L., Yvergniaux, M., ... Jeuffroy, M.-h. (2012). Pea – wheat intercrops in low-input conditions combine high economic performances and low environmental impacts. *European Journal of Agronomy*, *40*, 39-53. https://doi.org/10.1016/j.eja.2012.01.010

Pelzer, E., Hombert, N., Jeuffroy, M. H. & Makowski, D. (2014). Meta-analysis of the effect of nitrogen fertilization on annual cereal–legume intercrop production. *Agronomy Journal*, *106*(5), 1775-1786. https://doi.org/10.2134/agronj13.0590

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, *19*(1), 46. https://doi.org/10.1186/s12874-019-0666-3

Philibert, A., Loyce, C. & Makowski, D. (2012). Assessment of the quality of meta-analysis in agronomy. *Agriculture, Ecosystems and Environment*, *148*, 72-82. https://doi.org/10.1016/j.agee.2011.12.003

Phillips, C. A., Wang, K., Baker, E. J., Bubier, J. A., Chesler, E. J. & Langston, M. A. (2019). On Finding and enumerating maximal and maximum k-partite cliques in k-partite graphs. *Algorithms*, *12*(1). https://doi.org/10.3390/a12010023

Picheny, V., Servien, R. & Vialaneix, N. (2021). *SISIR : Sparse Interval Sliced Inverse Regression* [R package version 0.1-2].

Picheny, V., Servien, R. & Villa-Vialaneix, N. (2019). Interpretable sparse SIR for functional data. *Statistics and Computing*, *29*(2), 255-267. https://doi.org/10.1007/s11222-018-9806-6

Pigliucci, M. (2009). The end of theory in science ? *EMBO reports*, *10*(6), 534-534. https://doi.org/10.1038/embor.2009.111

Pimentel, D. (2006). Soil Erosion : A Food and Environmental Threat. *Environment, Development and Sustainability*, *8*(1), 119-137. https://doi.org/10.1007/s10668-005-1262-8

Pingali, P. L. (2012). Green Revolution : Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences*, *109*(31), 12302-12308. https://doi.org/10.1073/PNAS.0912953109

Plett, D. C., Ranathunge, K., Melino, V. J., Kuya, N., Uga, Y. & Kronzucker, H. J. (2020). The intersection of nitrogen nutrition and water use in plants : new paths toward improved crop productivity. *Journal of Experimental Botany*, *71*(15), 4452-4468. https://doi.org/10.1093/JXB/ERAA049

Podgórska-Lesiak, M. & Sobkowicz, P. (2013). Prevention of pea lodging by intercropping barley with peas at different nitrogen fertilization levels. *Field Crops Research*, *149*, 95-104. https://doi.org/10.1016/j.fcr.2013.04.023

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Pörtner, H., Roberts, D., Adams, H., Adler, C., Aldunce, P., Ali, E., Ara Begum, R., Betts, R., Bezner Kerr, R., Biesbroek, R., Birkmann, J., Bowen, K., Castellanos, E., Cissé, G., Constable, A., Cramer, W., Dodman, D., Eriksen, S., Fischlin, A., ... Zaiton Ibrahim, Z. (2022). *Climate change 2022 : impacts, adaptation and vulnerability*. IPCC.

Postma, J. A. & Lynch, J. P. (2012). Complementarity in root architecture for nutrient uptake in ancient maize/bean and maize/bean/squash polycultures. *Annals of Botany*, *110*(2), 521-534. https://doi.org/10.1093/AOB/MCS082

Praveen, B. & Sharma, P. (2019). A review of literature on climate change and its impacts on agriculture productivity. *Journal of Public Affairs*, *19*(4), e1960. https://doi.org/10.1002/PA.1960

Preisler, H. K., Ager, A. A., Johnson, B. K. & Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics*, *15*(7), 643-657. https://doi.org/10.1002/ENV.636

R Core Team. (2021). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rahman, Q. I. & Schmeisser, G. (1990). Numerische MathematJk Characterization of the Speed of Convergence of the Trapezoidal Rule. *Numer. Math*, *57*, 123-138. https://doi.org/10.1007/BF01386402

Ramsay, B., J.O. ; Silverman. (2005). *Funtional Data Analysis*.

Raseduzzaman, M. & Jensen, E. S. (2017). Does intercropping enhance yield stability in arable crop production ? A meta-analysis. *European Journal of Agronomy*, *91*, 25-33. https://doi.org/10.1016/j.eja.2017.09.009

Reiss, E. R. & Drinkwater, L. E. (2018). Cultivar mixtures : a meta-analysis of the effect of intraspecific diversity on crop yield. *Ecological Applications*, *28*(1), 62-77. https://doi.org/https://doi.org/10.1002/eap.1629

Rodriguez, C., Carlsson, G., Englund, J. E., Flöhr, A., Pelzer, E., Jeuffroy, M. H., Makowski, D. & Jensen, E. S. (2020). Grain legume-cereal intercropping enhances the use of soil-derived and biologically fixed nitrogen in temperate agroecosystems. A meta-analysis. *European Journal of Agronomy*, *118*(July 2019). https://doi.org/10.1016/j.eja.2020.126077

Sadras, V. O. & Denison, R. F. (2016). Neither crop genetics nor crop management can be optimised. *Field Crops Research*, *189*, 75-83. https://doi.org/10.1016/J.FCR.2016.01.015

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N. & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution 2019 3 :3*, *3*(3), 430-439. https://doi.org/10.1038/s41559-018-0793-y

Senft, M., Stahl, U. & Svoboda, N. (2022). Research data management in agricultural sciences in Germany : We are not yet where we want to be. *PLOS ONE*, *17*(9), e0274677. https://doi.org/10.1371/JOURNAL.PONE.0274677

Shelby, L. B. & Vaske, J. (2008). Understanding meta-analysis : A review of the methodological literature. *Leisure Sciences*, *30*(2), 96-110. https://doi.org/10.1080/01490400701881366

Sirami, C., Gross, N., Baillod, A. B., Bertrand, C., Carrié, R., Hass, A., Henckel, L., Miguet, P., Vuillot, C., Alignier, A., Girard, J., Batáry, P., Clough, Y., Violle, C., Giralt, D., Bota, G., Badenhausser, I., Lefebvre, G., Gauffre, B., ... Fahrig, L. (2019). Increasing crop

heterogeneity enhances multitrophic diversity across agricultural regions. *Proceedings of the National Academy of Sciences*, *116*(33), 16442-16447. https://doi.org/10.1073/pnas.1906419116

Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93-101. https://doi.org/10.1016/J.ESWA.2019.05.028

Stomph, T. J., Dordas, C., Baranger, A., de Rijk, J., Dong, B., Evers, J., Gu, C., Li, L., Simon, J., Jensen, E. S., Wang, Q., Wang, Y., Wang, Z., Xu, H., Zhang, C., Zhang, L., Zhang, W. P., Bedoussac, L. & van der Werf, W. (2020). Designing intercrops for high yield, yield stability and efficient use of resources : Are there principles ? *Advances in Agronomy*, *160*(1), 1-50. https://doi.org/10.1016/BS.AGRON.2019.10.002

Street, D. J. (1990). Fisher's Contributions to Agricultural Statistics. *Biometrics*, *46*(4), 937. https://doi.org/10.2307/2532439

Streit, R. P. & Bellwood, D. R. (2022). To harness traits for ecology, let's abandon 'functionality'. *Trends in Ecology & Evolution*. https://doi.org/https://doi.org/10.1016/j.tree.2022.11.009

Swinton, S. M., Lupi, F., Robertson, G. P. & Hamilton, S. K. (2007). Ecosystem services and agriculture : Cultivating agricultural ecosystems for diverse benefits. *Ecological Economics*, *64*(2), 245-252. https://doi.org/10.1016/J.ECOLECON.2007.09.020

Tang, X., Placella, S. A., Daydé, F., Bernard, L., Robin, A., Journet, E. P., Justes, E. & Hinsinger, P. (2016). Phosphorus availability and microbial community in the rhizosphere of intercropped cereal and legume along a P-fertilizer gradient. *Plant and Soil*, *407*(1-2), 119-134. https://doi.org/10.1007/S11104-016-2949-3/FIGURES/6

Tang, X., Zhang, C., Yu, Y., Shen, J., van der Werf, W. & Zhang, F. (2021). Intercropping legumes and cereals increases phosphorus use efficiency ; a meta-analysis. *Plant and Soil*, *460*(1), 89-104. https://doi.org/10.1007/s11104-020-04768-x

Tappert, C. C. (2019). Who is the father of deep learning ? *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, 343-348. https://doi.org/10.1109/CSCI49370.2019.00067

Tardieu, F. (2020). Educated big data to study sensitivity to drought. *Nature Food*, *1*(11), 669-670. https://doi.org/10.1038/s43016-020-00187-4

Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. (2002). Agricultural sustainability and intensive production practices. *Nature 2002 418 :6898*, *418*(6898), 671-677. https://doi.org/10.1038/nature01014

Timaeus, J., Weedon, O. D. & Finckh, M. R. (2022). Harnessing the Potential of Wheat-Pea Species Mixtures : Evaluation of Multifunctional Performance and Wheat Diversity. *Frontiers in Plant Science*, *13*, 683. https://doi.org/10.3389/FPLS.2022.846237/BIBTEX

Trinder, C., Brooker, R., Davidson, H. & Robinson, D. (2012). Dynamic trajectories of growth and nitrogen capture by competing plants. *New Phytologist*, *193*(4), 948-958. https://doi.org/https://doi.org/10.1111/j.1469-8137.2011.04020.x

Valsecchi, C. & Todeschini, R. (2020). Similarity/diversity indices on incidence matrices containing missing values. *Match*, *83*(2), 239-260.

Veres, A., Petit, S., Conord, C. & Lavigne, C. (2013). Does landscape composition affect pest abundance and their control by natural enemies ? A review. *Agriculture, Ecosystems & Environment*, *166*, 110-117. https://doi.org/https://doi.org/10.1016/j.agee.2011.05.027

Vergé, X., De Kimpe, C. & Desjardins, R. (2007). Agricultural production, greenhouse gas emissions and mitigation potential. *Agricultural and Forest Meteorology*, *142*(2), 255-269. https://doi.org/https://doi.org/10.1016/j.agrformet.2006.06.011

Verret, V., Pelzer, E., Bedoussac, L. & Jeuffroy, M. H. (2020). Tracking on-farm innovative practices to support crop mixture design : The case of annual mixtures including a legume crop. *European Journal of Agronomy*, *115*(July 2019). https://doi.org/10.1016/j.eja.2020.126018

Vezy, R., Munz, S., Launay, M., Lecharpentier, P. & Justes, E. (2022). Modelling intercrops functioning to advance the design of innovative agroecological systems. https://doi.org/10.21203/RS.3.RS-1930394/V1

Vicente-Saez, R. & Martinez-Fuentes, C. (2018). Open Science now : A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428-436. https://doi.org/10.1016/J.JBUSRES.2017.12.043

Viguier, L., Bedoussac, L., Journet, E.-P. & Justes, E. (2018). Yield gap analysis extended to marketable grain reveals the profitability of organic lentil-spring wheat intercrops. *Agronomy for Sustainable Development*, *38*(4), 39. https://doi.org/10.1007/s13593-018-0515-5

Violle, C., Navas, M. L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I. & Garnier, E. (2007). Let the concept of trait be functional ! *Oikos*, *116*(5), 882-892. https://doi.org/10.1111/J.0030-1299.2007.15559.X

Voisin, A. & Gastal, F. (2015). Nutrition azotée et fonctionnement agrophysiologique spécifique des légumineuses [Avec la contribution de : Guénaëlle Corre-Hellou, Jean-Jacques Drevon, Gérard Duc, Pierre Jouffret, Eric Justes, Bernadette Julier, Christophe Naudin, Anne Schneider, Pascal Thiébeau, Françoise Vertès.]. *Les légumineuses pour des systèmes agricoles et alimentaires durables* (512 p.). Editions Quae.

Voisin, A., Salon, C., Jeudy, C. & Warembourg, F. R. (2003). Symbiotic N2 fixation activity in relation to C economy of Pisum sativum L. as a function of plant phenology. *Journal of Experimental Botany*, *54*(393), 2733-2744. https://doi.org/10.1093/jxb/erg290

Volaire, F., Gleason, S. M. & Delzon, S. (2020). What do you mean "functional" in ecology ? Patterns versus processes. *Ecology and Evolution*, *10*(21), 11875-11885. https://doi.org/10.1002/ece3.6781

Wang, C., Du, J. & Fan, X. (2022). High-dimensional correlation matrix estimation for general continuous data with Bagging technique. *Machine Learning*, *111*(8), 2905-2927. https://doi.org/10.1007/S10994-022-06138-3/FIGURES/5

Watson, C. (2022). Many researchers say they'll share data - but don't. *Nature*, *606*(7916), 853. https://doi.org/10.1038/D41586-022-01692-1

Weih, M., Adam, E., Vico, G. & Rubiales, D. (2022). Application of Crop Growth Models to Assist Breeding for Intercropping : Opportunities and Challenges. *Frontiers in Plant Science*, *13*, 182. https://doi.org/10.3389/FPLS.2022.720486/BIBTEX

Weih, M., Karley, A. J., Newton, A. C., Kiær, L. P., Scherber, C., Rubiales, D., Adam, E., Ajal, J., Brandmeier, J., Pappagallo, S., Villegas-Fernández, A., Reckling, M. & Tavoletti, S. (2021). Grain yield stability of cereal-legume intercrops is greater than sole crops in more productive conditions. *Agriculture (Switzerland)*, *11*(3), 1-18. https://doi.org/10.3390/agriculture11030255

White, J. W. & Van Evert, F. K. (2008). Publishing agronomic data. *Agronomy Journal*, *100*(5), 1396-1400. https://doi.org/10.2134/agronj2008.0080F

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23. https://doi.org/10.18637/jss.v059.i10

Wickham, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., D', L., Mcgowan, A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/JOSS.01686

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). Comment : The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1-9. https://doi.org/10.1038/sdata.2016.18

Willey, R. W. & Rao, M. R. (1980). A competitive ratio for quantifying competition between intercrops. *Experimental Agriculture*, *16*(2), 117-125. https://doi.org/10.1017/S0014479700010802

Willey, R. W. (1979). Intercropping Its Importance And Research Needs Part 1. Competition And Yield Advantages Vol-32.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, *13*(6), e1005510. https://doi.org/10.1371/JOURNAL.PCBI.1005510

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., Xiong, L. & Yan, J. (2020). Crop Phenomics and High-Throughput Phenotyping : Past Decades, Current Challenges, and Future Perspectives. *Molecular Plant*, *13*(2), 187-214. https://doi.org/10.1016/J.MOLP.2020.01.008

Ylijoki, O. & Porras, J. (2016). Perspectives to Definition of Big Data : A Mapping Study and Discussion. *Journal of Innovation Management*, *4*(1), 69-91. https://doi.org/10.24840/2183-0606{_}004.001{_}0006

Yu, Y., Stomph, T. J., Makowski, D. & van der Werf, W. (2015). Temporal niche differentiation increases the land equivalent ratio of annual intercrops : A meta-analysis. *Field Crops Research*, *184*, 133-144. https://doi.org/10.1016/j.fcr.2015.09.010

Yu, Y., Stomph, T. J., Makowski, D., Zhang, L. & van der Werf, W. (2016). A meta-analysis of relative crop yields in cereal/legume mixtures suggests options for management. *Field Crops Research*, *198*, 269-279. https://doi.org/10.1016/j.fcr.2016.08.001

Zamir, D. (2013). Where Have All the Crop Phenotypes Gone ? *PLoS Biology*, *11*(6), 1-4. https://doi.org/10.1371/journal.pbio.1001595

Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & van 't Riet, K. (1990). Modeling of the bacterial growth curve [PMC184525[pmcid]]. *Applied and environmental microbiology*, *56*(6), 1875-1881. https://doi.org/10.1128/aem.56.6.1875-1881.1990

Annexes

1 SISIR multivarié

Introduction

On généralise la méthode SISIR (Picheny et al., 2019) au cas multivarié. Certaines notations sont reprises de Jain et al., 2019.

L'essentiel des équations est une généralisation des équations de Picheny et al., 2019. Un indice (k) apparaît donc sur de nombreux termes.

Ecriture du modèle dans le cadre multivarié

Adapté de la section 3 de Picheny et al., 2019

Soit un ensemble de variables aléatoires $(X^{(1)}, ..., X^{(s)}, Y)$ tel que $X^{(k)} \in \mathbb{R}^{p(k)}, k = 1, ..., s$. Dans la suite de cette partie, pour alléger les écritures, on suppose $p(k) = p,\ \forall k = 1, ..., s$.

Soient $(x_i^{(1)}, ..., x_i^{(s)}, y_i)$ n réalisations i.i.d de $(X^{(1)}, ..., X^{(s)}, Y)$.

On note $\mathbf{x_i^{(k)}} = (x_i^{(k)}(t_j))_{j=1,...,p} \in \mathbb{R}^p$ la i-ème observation de la variable k. Soit $\mathbf{X^{(k)}} \in M_{n,p}(\mathbb{R}), k = 1, ..., s$ la collection de matrices de taille $n \times p,\ (\mathbf{x_1^{(k)}}, ..., \mathbf{x_n^{(k)}})^\top$

Etape Ridge

Adapté de la section 3.1 de Picheny et al., 2019

On réécrit l'équation (4) de Picheny et al., 2019 dans le cas multivarié :

$$\mathcal{E}(A, C) = \sum_{k=1}^{s} \sum_{h=1}^{H} \hat{p}_h C_h^{(k)\top} A^{(k)\top} (\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p) A^{(k)\top} C_h^{(k)\top} - 2 \sum_{k=1}^{s} \sum_{h=1}^{H} \hat{p}_h (\bar{X}_h^{(k)} - \bar{X}^{(k)}) A^{(k)} C_h^{(k)\top}.$$

Les estimateurs de $A^{(k)}$ sont composés des d premiers vecteurs propres $(\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p)$-orthonormaux de $(\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p)^{-1} \hat{\Gamma}^{(k)}$ associés aux d plus grandes valeurs propres, où

$$\hat{\Gamma}^{(k)} = \sum_{h=1}^{H} \hat{p}_h (\bar{X}_h^{(k)} \bar{X}^{(k)}).$$

Pour calculer ces vecteurs propres, pour $k = 1, ..., s$, les vecteurs orthonormaux $(\hat{b}_j^{(k)})_{j=1,...d}$ des matrices $(\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p)^{-1/2} \hat{\Gamma}^{(k)} (\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p)^{-1/2}$ sont calculés.

Les matrices $A^{(k)}$ sont obtenues via :

$$A^{(k)} = (\widehat{\Sigma}^{(k)} + \mu_2 \mathbb{I}_p)^{-1/2} B^{(k)},$$

où B est la matrice de taille $p \times d$ dont les colonnes sont les $(\hat{b_j^{(k)}})_{j=1,...d}$.

Enfin, on a $\hat{C_h}^{(k)} = A^{(k)\top}(\bar{X}_h^{(k)} - \bar{X}^{(k)})$.

Etape Sparse

Adaoé de la section 3.2 de Picheny et al., 2019

Une fois qu'on a estimé les collections de matrices $(A^{(k)})_{k=1,...,s}$ et $(C^{(k)})_{k=1,...,s}$, on peut estimer les projections des espérances conditionnelles $(\hat{\mathbb{E}}(X^{(k)}|Y = y_i))_{i=1,...n}$ sur l'EDR via :

$\mathcal{P}_{\hat{A}^{(k)}}(\hat{\mathbb{E}}(X^{(k)}|Y = y_i)) = (\bar{X}_h^{(k)} - \bar{X}^{(k)})^\top \hat{A}^{(k)}$

En gardant les notations de Picheny et al., 2019, on note pour tout $k \in 1, ..., s$, $(\mathcal{P}_i^{(k),1}, ..., \mathcal{P}_i^{(k),d})$ le vecteur $\mathcal{P}_{\hat{A}^{(k)}}(\hat{\mathbb{E}}(X^{(k)}|Y = y_i))$ et on note $\mathbf{P}^{(k),j}$ le vecteur $(\mathcal{P}_j^{(k),1}, ..., \mathcal{P}_j^{(k),d})$.

D coefficients sont estimés pour chaque covariable : $\alpha^{(\mathbf{k})} = (\alpha_1^{(k)}, ..., \alpha_D^{(k)})_{k=1,...,s} \in \mathbb{R}^D$, (1 pour chaque intervalle $(\tau_k)_{k=1,...,D}$).

En généralisant l'équation (7) de Picheny et al., 2019, on déduit qu'on estime les $(\alpha^{(k)})_{k=1,..,s}$ comme solution du problème suivant :

$$\hat{\alpha}^{(\mathbf{k})} = \arg\min_{\alpha^{(k)} \in \mathbb{R}^D} \|\mathbf{P}^{(k)} - \Delta(\mathbf{X}^{(k)} \hat{A}^{(k)}) \alpha^{(\mathbf{k})}\|^2 + \mu_1 \|\alpha^{(\mathbf{k})}\|_{\ell_1}.$$

Si on suppose $\alpha^{(k)} = \alpha,\ \forall k = 1, ..., s$, la version multivariée donne :

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^D} \sum_{k=1}^{s} \|\mathbf{P}^{(k)} - \Delta(\mathbf{X}^{(k)} \hat{A}^{(k)}) \alpha\|^2 + \mu_1 \|\alpha\|_{\ell_1} \tag{1.1}$$

où $\mathbf{P}^{(k)} = \begin{pmatrix} \mathbf{P}^{(k),1} \\ ... \\ \mathbf{P}^{(k),d} \end{pmatrix}$, vecteur de taille dn et $\Delta(\mathbf{X}^{(k)} \hat{A}^{(k)}) \in M_{dn,D}(\mathbb{R})$, telle que définie

dans Picheny et al., 2019 : $\Delta(\mathbf{X}^{(k)}\hat{A}^{(k)}) = \begin{pmatrix} \mathbf{X}^{(k)}\Delta(\hat{a}_1^{(k)}) \\ ... \\ \mathbf{X}^{(k)}\Delta(\hat{a}_p^{(k)}) \end{pmatrix}$ Les $\Delta(\hat{a}_j^{(k)})$, $k = 1, ..., s$ sont les matrices de taille $(p \times D)$ telles que

$$\Delta_{dc}(\hat{a}_j^{(k)}) = \begin{cases} \text{entrée d du vecteur } \hat{a}_j^{(k)} & \text{si } t_d \in \tau_c \\ 0 & \text{sinon.} \end{cases}$$

Les vecteurs sparse sont définis via les coefficients $\hat{\alpha}^{(k)}$, $k = 1, ..., s$:

$$\forall d = 1, ..., p,\ \hat{a}_{jl}^{(k),s} = \hat{\alpha}_c^{(k)}\hat{\alpha}_{jl}^{(k)} \text{ pour } c \text{ tel que } t_l \in \tau_c.$$

Choix des intervalles

Problématique : jusqu'à présent, les intervalles $(\tau_k)_{k=1,...,D}$ étaient donnés à priori. Picheny et al., 2019 propose une procédure afin de déterminer ces intervalles optimalement.

Rappel : Pour un ensemble d'intervalles $(\tau_k)_{k=1,...,D}$ et pour $k = 1, ..., s$, on a des coefficients $\hat{\alpha}^{(k)}$ si on utilise l'équation (7) de Picheny et al., 2019.
Si on utilise la généralisation au cadre multivarié (équation (1.1)), on obtient un vecteur $\hat{\alpha}$.

Possibilités :

1. On peut appliquer la même méthode que la section 4 de Picheny et al., 2019, pour chaque $\hat{\alpha}^{(k)}$, $k = 1, ..., s$ i.e : $\forall k = 1, ..., s$ définir $(\hat{\alpha}^{(k)+})_{i=1,...,D}$ et $\hat{\alpha}^{(k)-})_{i=1,...,D}$ comme étant les premières solutions parmi le chemin de solutions (qui varient selon la valeur de $\mu_1^{(k)}$) telles que au maximum (resp. au minimum) une proportion P des coefficients sont différents de 0 (resp. égaux à 0). On peut ainsi déduire des intervalles d'influence via la procédure décrite en section 4 et en figure 2 de Picheny et al., 2019. On obtient des intervalles d'influence différents entre les covariables.
2. On peut déduire les intervalles d'influence via la procédure décrite en section 4 et en figure 2 de Picheny et al., 2019, via le vecteur α obtenu grâce à l'équation (1.1). On obtient des intervalles d'influence communs entre les covariables.

La solution 1. revient finalement à appliquer SISIR sur chaque série temporelle séparément.

La solution 2. risque de se heurter à des problèmes d'échelles entre les différentes covariables (P-ETP ; T° ; Radiations). On pourrait remédier à cela via des procédures de normalisation.

2 Décomposition de la somme des résidus totaux (imputations comprises)

On dispose d'un ensemble d'observations $(y_i)_{i=1,\dots,n}$ et d'un ensemble de valeurs ajustées liées $(\hat{y}_i^{(m)})_{i=1,\dots,n,m=1,\dots,M}$, avec n le nombre d'observations et M le nombre de valeurs imputées.

La somme des résidus totaux au carré s'écrit :

$$\text{SSRT} = \sum_{m=1}^{M} \sum_{i=1} (\hat{y}_i^{(m)} - y_i)^2. \tag{2.1}$$

On pose $\hat{y}_i = \frac{\sum_m \hat{y}_i^{(m)}}{M}$ (i.e $\hat{y}_i$ est la moyenne des valeurs ajustées (entre toutes les imputations) pour l'observation i).

Décomposons cette somme :

$$\begin{aligned}
\sum_{m=1}^{M} \sum_{i=1}^{n} (\hat{y}_i^{(m)} - y_i)^2 &= \sum_m \sum_i (\hat{y}_i^{(m)} - \hat{y}_i + \hat{y}_i - y_i)^2 \\
&= \underbrace{\sum_m \sum_i (\hat{y}_i^{(m)} - \hat{y}_i)^2}_{\text{somme carrés inter-imputations}} + \underbrace{\sum_m \sum_i (\hat{y}_i - y_i)^2}_{\text{somme carrés résiduels}} + 2 \underbrace{\sum_m \sum_i (\hat{y}_i^{(m)} - \hat{y}_i)(\hat{y}_i - y_i)}_{\alpha}
\end{aligned}$$

Or,

$$\begin{aligned}
\alpha &= \sum_m \sum_i \hat{y}_i^{(m)} \hat{y}_i - \sum_m \sum_i \hat{y}_i^{(m)} y_i - \sum_m \sum_i \hat{y}_i^2 + \sum_m \sum_i \hat{y}_i y_i \\
&= \sum_i M \hat{y}_i^2 - \sum_i M \hat{y}_i y_i - \sum_i M \hat{y}_i^2 + M \sum_i y_i \hat{y}_i = 0
\end{aligned}$$

D'où

$$\text{SSRT} = \sum_m \sum_i (\hat{y}_i^{(m)} - \hat{y}_i)^2 + M \sum_i (\hat{y}_i - y_i)^2.$$

On peut donc quantifier la part de variance résiduelle totale due aux imputations :

$$\%\text{Variance imputation} = \frac{\sum_m \sum_i (\hat{y}_i^{(m)} - \hat{y}_i)^2}{\text{SSRT}}$$

et le pourcentage de variance inexpliquée :

$$\%\text{Variance inexpliquée} = \frac{M \sum_i (\hat{y}_i - y_i)^2}{\text{SSRT}}$$

On souhaite que le part de variance résiduelle totale due aux imputations soit *faible*, *i.e* le processus d'imputation multiple n'a pas eu beaucoup d'impact sur la variabilité des résidus.

3 Méthodologie pour l'estimation du bilan C de ma thèse

Périmètre

Le périmètre, c'est à dire l'ensemble des émissions à comptabiliser ou non dans un bilan C est un élément crucial. Dans mon cas, j'ai choisi de me focaliser *sur mes activités de recherche seules*, sans prendre en compte l'impact des expérimentations qui constituent ma base de données ou l'alimentation. Ce choix est motivé par i) la difficulté d'estimation du bilan Carbone de toutes les expérimentations liées à ma thèse, ii) le fait que ces expérimentations ont eu lieu indépendamment de ma thèse. Néanmoins, un bilan Carbone réalisé à AGIR en 2019 a montré que les expérimentations représentaient une part non négligeable des émissions par personne.

Les facteurs d'émissions (i.e les émissions par unité de consommation (kilomètres, kWh etc.) sont principalement issus de la base carbone de l'Ademe [1].

Aviation

J'ai fait le choix de représenter l'impact de l'aviation sans traînées de condensation. Ce choix est motivé par le fait que l'impact de ces traînées, formant des cirrus dans le ciel est encore sujet à débat dans la littérature scientifique (Lee et al., 2021). Les cirrus sont les nuages qui sont hauts dans le ciel, et qui ont un pouvoir réchauffant très élevé (ils réfléchissent beaucoup plus de rayons infrarouges vers le sol qu'ils ne réfléchissent de rayons ultraviolets vers l'espace) . L'impact CO_2 de l'aviation est élevé en soi, et a lieu sur le long terme (un molécule de CO_2 mettant un temps très long à se dégrader dans l'atmosphère). L'impact des traînées est direct, mais court-terme. Autrement dit, si tous les avions arrêtaient de voler du jour au lendemain, le réchauffement dû aux traînées s'arrêteraient également alors que l'impact lié au CO_2 continuerait des milliers d'années. Comptabiliser ou non les impacts des traînées de condensation est donc un choix de prise en compte des impacts sur le court ou long terme, et est encore un sujet de débat.

Chauffage

J'ai estimé l'impact de l'utilisation du chauffage (gaz) en divisant l'impact carbone calculé pour les bâtiments lors de la réalisation d'un bilan C de l'unité AGIR par le nombre d'agent.e.s.

Déplacements (congrès et réunions)

J'ai compté les déplacements de mes encadrant.e.s de thèse quand ils m'accompagnaient à Montpellier, ainsi que les déplacements de ma directrice de thèse quand elle venait à Toulouse, toujours dans une optique de compter toutes les activités nécessaires à mon activité de recherche. Je n'ai pas comptabilisé les déplacements des membres du jury, n'ayant pas l'information de leur moyen de venue à ma soutenance.