Rapport de stage

Pour obtenir le

Diplôme de master en Recherche Opérationnelle

Présenté et soutenu publiquement le 04 Septembre 2018

Par

Yousra Khattali

Sélection de variables dans les données climatiques pour les modèles agro-écologiques

---

Encadré par :

Mme. Aude Rondepierre

Mme. Nathalie Vialaneix

Mr. Rémi Servien

Mr. Victor Picheny

Période de stage: 06 Mars 2018 - 31 Août 2018

Remerciements

Tout d'abord, j'adresse mes remerciements respectueux à **Sylvain Jasson** pour m'avoir accueillie à l'unité MIAT de l'INRA dont il est le directeur.
Je souhaite également remercier **Marcel Mongeau** pour la formation qu'il nous a offerte et dont il est le responsable.˙

Je tiens également à remercier mes responsables de stage **Victor Picheny**, **Nathalie Vialaneix** et **Rémi Servien** pour m'avoir intégrée rapidement au sein de l'INRA et m'avoir accordée toute leur confiance ; pour le temps qu'ils m'ont consacrée tout au long de cette période, sachant répondre à toutes mes interrogations.

Aussi, je remercie **Aude Rondepierre**, encadrante académique du stage, pour sa gentillesse et sa modestie, ainsi que toute l'équipe pédagogique du master « Recherche Opérationnelle» de l'université Paul Sabatier pour leurs enseignements m'ayant permis d'effectuer ce stage.

Je voudrais aussi exprimer toute ma reconnaissance envers **Alain Perault** et **Fabienne Ayrignac** pour leur disponibilité toutes les fois que je les ai sollicités.

Aussi, un grand merci à tous mes collègues stagiaires et doctorants pour avoir contribué à créer une ambiance de travail agréable. Je souhaite également n'oublier personne, que les personnes ayant participées de près ou de loin à ce projet soient remerciées.

Table des matières

Table des figures

Liste des tableaux

Résumé

Notre étude a comme objectif de comprendre la manière dont le climat, ou plutôt les motifs climatiques, influencent le rendement du blé. Pour atteindre cet objectif, j'ai travaillé à partir d'un jeu de données composé de 1000 séries climatiques annuelles. Ces variables sont utilisées comme des paramètres d'entrée d'un simulateur numérique de culture (STICS) pour le calcul du rendement du blé.

Le modèle STICS permet de simuler les sorties (la valeur du rendement) pour une entrée donnée mais vu que nos entrées sont des séries complexes fonctionnelles, de plus, la complexité des interactions climat/plantes ne permet pas d'étudier les systèmes de manière explicite : il est donc difficile voire impossible de l'utiliser pour une étude exhaustive de l'influence des valeurs des variables d'entrée sur la sortie.

Il est alors utile d'avoir recours à une approche dite par méta-modélisation afin de faciliter l'optimisation et d'identifier les variables climatiques les plus influentes et les intervalles temporels les plus sensibles : ce type d'approche consiste à construire un méta-modèle statistique permettant d'approcher le modèle original STICS avec un temps de calcul optimisé tout en conservant de bonnes performances prédictives puis à effectuer une recherche des variables contribuant le plus à la qualité de la prédiction. Dans cette étude, j'ai utilisé comme méta-modèles, le modèle linéaire basé dans la construction des méta-variables sur la moyenne dans un premier temps et basé sur la régression par PLS dans un second temps. Enfin, j'ai construit des méta-variables pour le modèle linéaire en prenant des degrés jours au lieu du jour de culture.

Dans les trois cas du modèle linéaire, des erreurs quadratiques moyennes ont été calculées et une analyse de significativité des variables a été utilisée. Vu que les prédictions du modèle initial sont faiblement corrélées par rapport aux valeurs initiales et que le modèle n'a pas un bon pouvoir prédictif, il a été donc nécessaire de faire un découpage de la période de culture traitée (223 jours) en plusieurs subdivisions. Pour ce faire, on a construit des méta-variables sur ces intervalles pour avoir un méta-modèle explicatif du rendement du blé. L'estimation du modèle linéaire par agrégation par moyenne par intervalle nous a donné des bons résultats interprétatifs alors que celle par PLS par intervalle est considérée comme approche détériorée. Les résultats obtenus sur le modèle STICS sont partiellement en adéquation avec le cycle de production du blé.

Mots clés : méta-modélisation, P-valeur, régression PLS,degré jour.

Abstract

Our study aims to understand how climate, or rather climatic patterns, influence wheat yield. To achieve this goal, I worked based on dataset composed of 1000 annual climate series. These variables are used as input parameters of a digital crop simulator (STICS) for calculating wheat yield.

The STICS model allows to simulate the outputs (the value of the yield) for a given input but our inputs are complex functional series, moreover, the complexity of the climate / plant interactions does not allow to study the systems in an explicit way :

it is therefore difficult and even impossible to use it for an exhaustive study of the influence of the values of the input variables on the output.

So, it is useful to use a so-called meta-modeling approach to facilitate optimization and to identify the most influential climate variables and the most sensitive time intervals :

this type of approach consists of constructing a meta-statistical model allowing to approach the original STICS model with optimized computation time while maintaining good performance prediction and then search for the variables that contribute the most to the quality of the prediction. In this study, I used as meta-models, the linear model based on the average at first and based on PLS regression in a second time. Finally, I built meta-variables from degrees days instead of the day of culture as input of linear model.

Since the initial model is weakly correlated and does not have a good predictive power, it was therefore necessary to split the treated crop period (223 days) into several subdivisions. To do this, we built meta-variables on these intervals to have an explanatory meta-model of wheat yield.

By constructing interval meta-variables to provide an explanatory meta-model of wheat yield. The estimation of the linear model by mean interval aggregation gave us good interpretative results that PLS per interval is considered to be a deteriorated approach. The results obtained on the STICS model are partially in line with the wheat production cycle.

Keywords : meta-modeling, P-value, PLS regression, Degree day.

Introduction

Le changement climatique a créé des défis pour le secteur agricole et continuera à le faire en particulier pour la productivité agricole. En effet, il est le facteur dominant qui a une grande influence sur le rendement de divers types de cultures. Dans ce contexte, mon stage porte sur l'étude de l'interaction culture/climat. Je m'intéresse principalement à la culture du blé dont le rendement est simulé à l'aide du modèle de culture appelé STICS (Simulateur mulTIdisciplinaire pour les Cultures Standard). C'est un modèle développé à l'INRA depuis 1996 qui permet d'évaluer le rendement de certains types de culture. Ce modèle prend en entrée des données climatiques (température, pluviométrie ...) et, à partir d'un code de calcul complexe, calcule un rendement associé. Les données utilisées par ce modèle dépendent du milieu (type de sol), de la conduite de culture (date de semis, type d'irrigation) et de la variété (deux types de blé).

L'objectif de ce travail est de comprendre l'interaction culture/climat et de déterminer les intervalles au sein des données climatiques (donc en entrée du modèle) permettant d'expliquer et de prédire les variations du rendement, la sortie des modèles. Pour ce faire, il existe plusieurs approches qui ont été explorées telles que : la construction de "méta-variables" qui résument chaque intervalle, la régression linéaire voir [1]

, la régression des moindres carrés partiels PLS et la sélection de variables significatives qui ont beaucoup d'influence sur le rendement.

La suite du document sera organisée de la manière suivante : nous allons tout d'abord faire une brève présentation de l'organisme d'accueil. Puis nous présenterons la problématique du stage et les données utilisées. Ensuite, nous détaillerons la méthodologie adoptée et les outils statistiques utilisés. Puis nous présenterons les résultats obtenus et enfin nous terminerons en évoquant les perspectives possibles.

Section I : Organisme d'accueil

1.1 Institut National de la Recherche Agronomique

✺ **Historique :**

L'INRA est un organisme de recherche scientifique publique, placé sous la double tutelle du Ministère de l'Enseignement Supérieur et de la Recherche et du Ministère de l'Alimentation, de l'Agriculture et de la Pêche. Il a été créé en 1946 et est constitué aujourd'hui de 14 départements scientifiques, répartis sur 19 centres régionaux. Ses recherches se concentrent sur les questions liées à l'agriculture, à l'alimentation et à la sécurité des aliments, à l'environnement et à la gestion des territoires, avec une perspective de développement durable.

✺ **Missions :**

Il a pour objectif de :

- Produire et diffuser des connaissances scientifiques ;
- Contribuer à l'innovation par le partenariat et le transfert ;
- Former à la recherche et par la recherche ;
- Élaborer la stratégie de recherche européenne et nationale ;
- Éclairer les décisions publiques ;
- Contribuer au dialogue entre sciences et société.

Pour cela, l'INRA est présent au niveau mondial et est en permanence au contact des acteurs académiques, économiques, associatifs et territoriaux. Tous ces différents acteurs agissent au travers de branches scientifiques très diversifiées : les sciences de la vie en majorité (68% des compétences scientifiques de l'INRA), les sciences des milieux et des procédés (12%), l'ingénierie écologique, les écotechnologies et les biotechnologies (8%), de même que les sciences économiques et sociales (8%) et les sciences du numérique (4%).

1.2 Département de Mathématiques et Informatique Appliquées (MIA)

Le stage a été effectué dans le département de recherche en Mathématiques et Informatique Appliquées (MIA). Les recherches de ce département sont axées sur :

— la bioinformatique, au sens de l'ensemble des méthodes relevant des mathématiques et l'informatique appliquées à l'exploitation des données de génomique et de post génomique ;

— la modélisation des systèmes complexes dans les champs de l'agriculture, de l'environnement et de l'analyse des risques et des procédés industriels.

Ses chercheurs participent au développement de méthodes et logiciels et à leurs mises en œuvre dans des projets en partenariat avec les thématiciens de l'INRA. Le département est composé de 8 unités de recherche : 2 unités de recherche propres, 2 unités mixtes de recherche associées avec d'autres organismes de recherche ou d'enseignement, 2 unités de recherche pluri-départementales, une unité mixte de recherche sous contrat et une unité mixte de service.

Le stage s'est déroulé dans une unité de recherche propre : l'unité Mathématiques et Informatique appliquées de Toulouse (MIAT).

1.3 Unité Mathématiques et Informatique Appliquées de Toulouse (MIAT)

L'unité MIAT a pour mission de développer et de mettre à jour des méthodes et des compétences en mathématiques et/ou en informatique pour la résolution des problèmes que peuvent rencontrer les autres départements de l'INRA. L'unité est composée de deux équipes de recherche :

- SaAB : Statistique et Algorithmique pour Biologie
- MAD : Modélisation des Agro-écosystèmes et Décision.

Ce stage est rattaché l'équipe MAD et à la plateforme RECORD. Les travaux de cette équipe s'articulent autour de la modélisation, la simulation, l'exploration et l'optimisation

des systèmes dans les champs de l'agriculture, de l'environnement et de l'analyse des risques alimentaires. La plateforme RECORD offre un cadre et des outils informatiques adéquats pour la mise en oeuvre de ces travaux.

Section II : Cadre du projet

Les céréales sont considérées comme la base des grandes civilisations et ont constitué l'une des premières activités agricoles, fournissant un moyen d'alimentation régulier, autour duquel l'activité humaine pouvait s'organiser. C'est ainsi que les civilisations en Europe se sont construites autour du blé. En Méditerranée, le blé occupe une place essentielle dans les sociétés où il est considéré comme une matière première vitale pour l'homme [2]. Afin de conforter sa place dans les systèmes de grande culture et de renforcer sa compétitivité, la productivité du blé doit s'améliorer et gagner en régularité. Pour ce faire, plusieurs voies sont à explorer, dont la maîtrise de l'impact du climat sur le rendement du blé.

2.1 Problématique du stage

Une interaction permanente entre climat et culture trouve un écho majeur dans le bassin méditerranéen. En effet, de nombreux travaux récents en agronomie portent sur la modélisation de systèmes de culture à différentes échelles : paysage, parcelle, plante, etc., et sur l'utilisation de ces modèles pour aider à l'amélioration de ces systèmes vis-à-vis d'aspects variés tels que la productivité, l'impact écologique ou la résilience au changement climatique. Dans le cadre de mon stage, je m'intéresse en particulier au modèle de culture STICS. Les entrées de notre modèle sont des données climatiques sous forme des courbes fonctionnelles et complexes, ce qui rend difficile l'étude des systèmes de manière explicite. On a alors recours aux approches dites par simulation, où les modèles sont traités comme des boîtes noires et les relations entrées/sorties inférées à l'aide d'outils statistiques à partir d'un échantillon simulé. Dans ce cadre, il est important de comprendre la manière dont le climat influence le rendement. C'est pourquoi notre objectif final sera de mettre en valeur les motifs climatiques les plus importants pour le rendement et ainsi d'identifier les intervalles les plus influents, indépendamment de la plante ou en fonction de celle-ci. D'un point de vue théorique, les entrées climatiques du modèle STICS peuvent être comprises soit comme des séries temporelles multivariées, soit comme des variables aléatoires

fonctionnelles et les sorties sont le rendement du blé.

2.2 Description des données

2.2.1 Le modèle STICS

STICS est un modèle de simulation du fonctionnement de divers types de culture, en particulier le blé, développé par l'INRA depuis 1996. Il est caractérisé par son dynamisme et sa robustesse, il permet de simuler le système sol-atmosphère-culture. De nos jours, le modèle est adapté à près d'une vingtaine d'espèces cultivées, annuelles, pérennes, herbacées ou ligneuses, ce qui le rend très reconnu sur la scène internationale. Il prend en compte trois familles de paramètres en entrées et renvoie deux sorties.
Le modèle STICS peut être schématisé comme sur la figure 2.1.

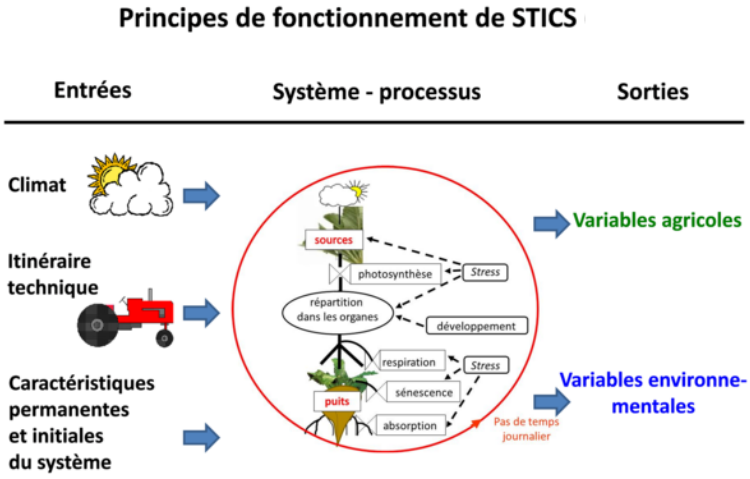

FIGURE 2.1 – Représentation schématique du modèle STICS

Le modèle a besoin de variables climatiques caractérisant, pour chaque jour de la simulation, l'état de l'atmosphère au voisinage du système : rayonnement, température, pluie, etc. Il simule également l'effet des pratiques agricoles sur le système, ce qui nécessite de lui fournir des informations telles que la date de semis qui va déterminer en partie la levée ou encore le calendrier des irrigations. De plus, l'état initial du système doit être décrit, comme la teneur en eau et en nutriments dans le sol. Les entrées sont donc

— Les données climatiques ;

— L'itinéraire technique ;

— Les caractéristiques initiales du système.

Les sorties de STICS sont de deux types : soit d'intérêt agronomique telle que le rendement et les consommations d'intrants, soit d'intérêt environnemental (pertes d'eau et de nitrates, émission de N2O, ...).

Dans cette étude, seul le premier paramètre d'entrée (données climatiques) sera utilisé et seul le rendement sera calculé. Pour plus de détails sur ce modèle, nous renvoyons le lecteur à [3, 4]

Nous avons traité des données issues du simulateur de données météorologiques WACS-Gen. Celui-ci a été calibré pour reproduire les caractéristiques du site de Lleida (Catalogne, Espagne) pour les années 1981 et 1982 (la période culturale étant à cheval sur deux années calendaires).

Les données générées sont utilisées par le modèle de culture STICS pour simuler le rendement d'une culture de blé sous différentes configurations. Chaque configuration diffère d'une autre par plusieurs modalités culturales qui sont :

— la profondeur du sol : deep (profond) ou shallow (peu profond)

— la date de semis (en jour julien) : 287, 203 ou 332

— le mode d'irrigation : irrigation automatique ou pas d'irrigation

— la variété du blé : tendre ou fin

On a testé deux configurations en changeant à chaque fois le mode d'irrigation. Le tableau 2.1 présente les deux traitées durant cette étude.

Tableau 2.1 – Modalité culturale pour le blé

Configuration	**Mode d'irrigation**	**Date se semis**	**Profondeur du sol**	**variété du blé**
1	**automatique**	287	peu profond	tendre
2	**sans**	287	peu profond	tendre

2.2.2 Les données d'entrée : données climatiques

Les données climatiques sont des relevés journaliers de certaines caractéristiques décrites sous forme de séries temporelles. Elles sont présentées par 5 courbes discrétisées en 223 points qui sont :

— T_{min} : la température minimale (C)

— T_{max} : la température maximale (C)

— R : le rayonnement globale (MJ/m)

— P : la pluviométrie (mm)

— ETM : l'évapotranspiration potentielle (mm/jour).

Cet ensemble de données contient 1000 séries climatiques virtuelles. La période de culture du blé étant fixée entre octobre et juillet, les analyses seront faites uniquement sur cette période qui dure 223 jours. Les données climatiques sont donc composées de 5 séries de 223 mesures journalières, comprises entre les deux années 1981 et 1982. Après une analyse qualitative des données, nous avons présenté les composantes de l'entrée du modèle en fonction du temps.

Un exemple d'entrée du modèle est donné sur la figure 2.4.

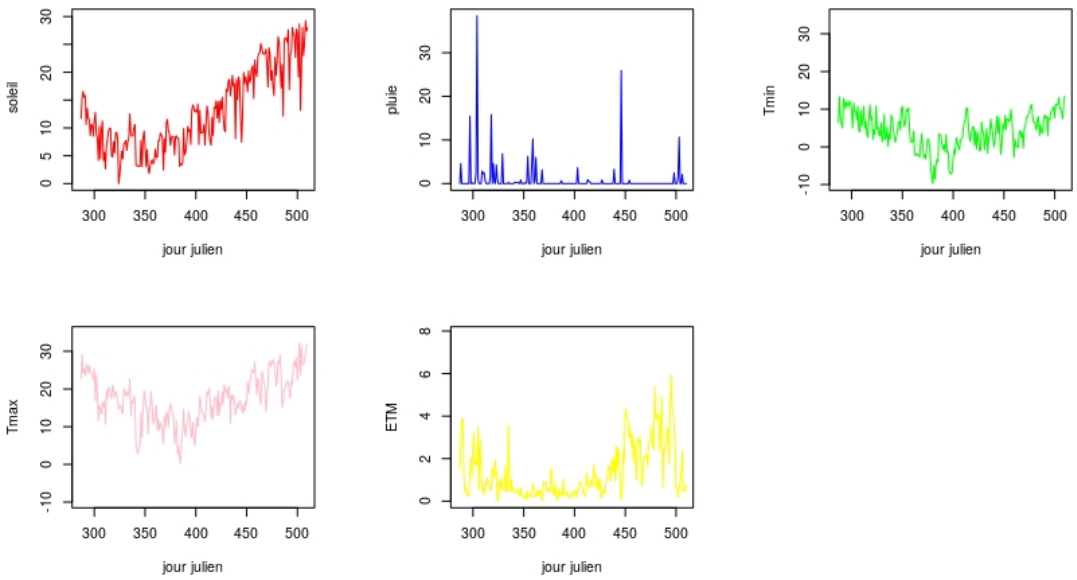

FIGURE 2.2 – Un exemple de données climatiques.

Nos données, comme les montre la figure 2.2, sont des séries fonctionnelles et complexes ce qui nous empêche d'étudier les systèmes de façon explicite. Nous avons donc recours aux approches dites par simulation qui sont détaillées dans le chapitre suivant.

2.2.3 Les données de sortie : rendement du blé

La sortie de notre modèle est le rendement du blé, on a commencé notre traitement par une étude graphique sur la distribution du rendement dans le cas où le type d'irrigation est automatique. Ensuite on a traité le cas d'absence d'irrigation.

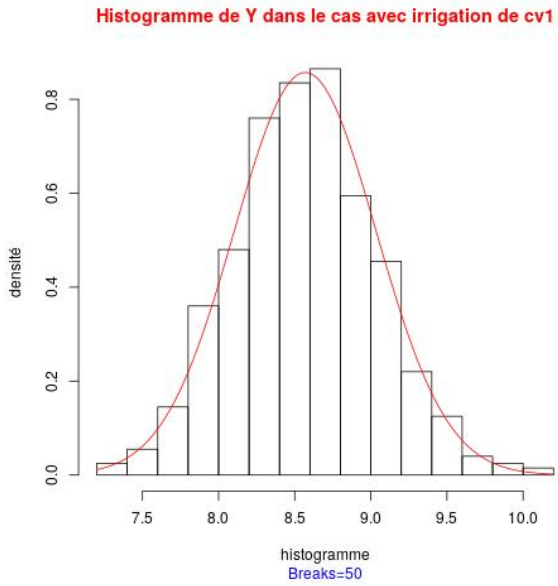

FIGURE 2.3 – Histogramme du rendement dans le cas avec irrigation.

La variable Y portée en ordonnée donne le rendement du blé sur l'histogramme (fig 2.3). On remarque que le rendement, dans ce cas, est une variable gaussienne dont les valeurs observées sont encadrées entre 7 et 10 tonnes/hectare.

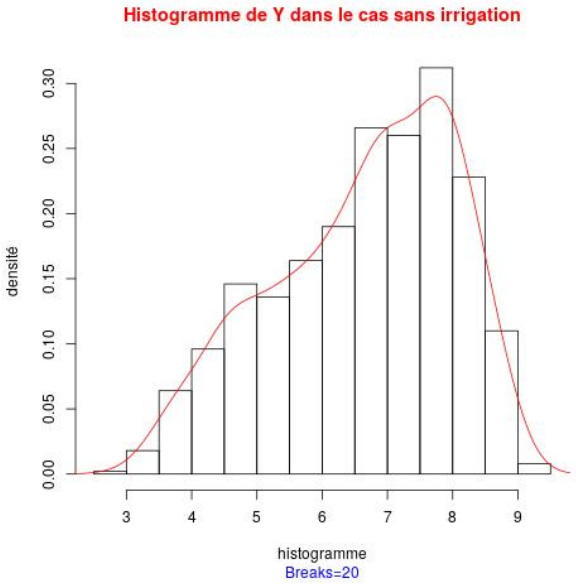

FIGURE 2.4 – Histogramme du rendement dans le cas sans irrigation.

On remarque dans le cas où il n' y a pas d'irrigation la distribution du rendement du blé n'est pas gaussienne comme la première modalité. D plus les valeurs sont faibles et encadrées entre 2 et 9 tonnes/hectare. C'est pour cela, on va commencer l'analyse de la configuration dont le mode d'irrigation est automatique. Plusieurs voies sont à explorer, dont la maîtrise de l'interaction climat/rendement du blé. Dans ce contexte, la troisième section sera consacrée à décrire les outils et les approches utilisées.

Section III : Cadre méthodologique

L'objectif principal de ce stage est d'identifier les motifs climatiques et les intervalles temporels les plus influents à la variation du rendement du blé. Le problème peut se formaliser alors sous la forme d'une question dans laquelle on cherche à comprendre l'influence de variables d'entrées $X \in \mathbb{R}^{5\times 223}$ (les données climatiques) sur une variable de sortie $Y \in \mathbb{R}$ (le rendement du blé). Le modèle STICS permet de calculer explicitement les sorties pour une entrée donnée mais les résultats ne sont pas facilement interprétables. On a alors recours à un méta-modèle, c'est-à-dire un modèle statistique qui nous permet d'approcher le modèle STICS avec des temps de calcul moindres et qui permet d'interpréter plus facilement la relation entrée-sortie. Le méta-modèle est construit à partir d'observations qui ont été préalablement générées par le modèle STICS : $(x_1, y_1), \ldots, (x_n, y_n)$. On peut ensuite faire une analyse de sensibilité sur le méta-modèle obtenu. Cette section présente les méthodes utilisées pour construire les méta-modèles ainsi que l'analyse de sensibilité.

Dans la suite, on notera en particulier :

- $\mathbf{X}$ la matrice de dimension $5\times223\times1000$ des observations d'entrée, $\mathbf{X} = (x_1^T, \ldots, x_{1000}^T)^T$, et $\mathbf{Y}$ le vecteur de taille 1000 des observations de sortie, $\mathbf{Y} = (y_1, \ldots, y_{1000})^T$;
- lorsqu'on travaille sur une variable d'entrée particulière, $j \in \{1, \ldots, 1000\}$, celle-ci est indiquée en exposant : x_i^j désigne la j-ème variable de l'observation x_i, X^j la j-ème variable en général et $\mathbf{X}^j$ la j-ème colonne de la matrice des observations d'entrée.

Méta-modélisation

Pour plusieurs raisons, les recherches en sciences agronomiques nécessitent des traitements de bases de données de plus en plus importantes. Il est donc de plus en plus nécessaire de pouvoir extraire des données quantitatives à partir des publications de la littérature.

En conséquence, les méthodes de méta-modélisation statistiques des bases de données

deviennent essentielles et il importe de les mettre en œuvre d'une façon adéquate, (voir [5]).

Modéliser c'est représenter un phénomène réel par un modèle mathématique dans le but de le simplifier. Cependant, il arrive que le modèle obtenu soit encore trop complexe et coûteux en temps de calcul, comme c'est le cas pour le modèle STICS. On a alors recours à la méta-modélisation, qui consiste à créer un méta-modèle afin de simplifier et de pouvoir explorer facilement le modèle initial. La méta-modélisation consiste à construire un modèle simplifié qui devra posséder de bonnes performances prédictives pour résumer au mieux le modèle initial. La construction d'un méta-modèle nécessite un jeu de données composé d'observations des variables explicatives et d'observations des variables à expliquer correspondantes. Le modèle initial est tel que $Y = \text{modele}(X)$ et on cherche un méta-modèle tel que $Y = metamodele(X) + \epsilon$ ou $Y \approx \text{metamodele}(X) + \epsilon$ où ϵ est la différence entre les valeurs du modèle initial et les valeurs fournies par le méta-modèle.

Plan d'expériences

Un plan d'expériences est défini comme une suite ordonnée d'essais d'une expérimentation dont chacun permet d'expliquer la relation entré/sortie avec une bonne économie (un maximum d'informations en un minimum d'expériences). Il peut être vu comme un ensemble de n vecteurs d'entrée tels que m observations soient bien réparties sur l'ensemble de définition des variables explicatives. Celui-ci se décrit par trois éléments, des données X (les données climatiques), une cible Y (le rendement) et une fonction d'erreur qui permet d'évaluer la distance entre la prédiction et la cible.

Il existe plusieurs méthodes de méta-modélisation telles que les modèles linéaires, les réseaux de neurones, les arbres de régression, les forêts aléatoires etc.

On a entamé notre étude par l'application d'une régression linéaire basée sur des résumés sur chaque intervalle de temps. Pour obtenir ces résumés des méta-variables, trois approches statistiques ont été étudiées et mises en œuvre :

- moyenne par intervalle ;
- PLS par intervalle ;
- degré jour au lieu du jour de culture.

3.1 Régression linéaire

Le modèle linéaire est un modèle statistique où on cherche à exprimer une variable Y (sortie) linéairement en fonction d'une variable explicative X (entrées). Le modèle peut s'écrire sous la forme suivante :

$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_i^j + \epsilon_i \tag{3.1}$$

On définit β_0 la constante inconnue, β le vecteur des paramètres du modèle (β_0,...,β_d), x_i le vecteur des variables explicatives pour l'individu i, d le nombre d'observation et ϵ_i l'erreur du modèle de moyenne nulle et de variance σ^2.
Ce modèle permet l'étude de la relation entre entrées et sortie mais aussi une éventuelle prédiction de la sortie à partir de nouvelles données d'entrée. voir [6]

3.1.1 Indicateurs statistiques

Tout au long de cette étude, je me suis basée sur deux indicateurs statistiques : le premier pour mesurer la qualité du modèle qui est l'erreur quadratique moyenne (MSE). Puis, pour étudier la significativité des variables et des intervalles, nous avons regardé le deuxième indicateur qui l'analyse de variance en regardant les indices p-valeur de chacune d'entre elles.

❶ Erreur Quadratique Moyenne (MSE) :
Nous souhaitons avoir des modèles avec de bonnes qualités de prévisions. Une validation du modèle se fait en terme de qualité de prévision et non en terme de qualité d'ajustement donc un meilleur modèle n'est pas celui qui s'ajuste le mieux aux données d'apprentissage mais plutôt celui qui a une faible erreur de prévision sur un échantillon de test.
Pour chaque modèle, on calcule donc un indicateur qui résume la série d'erreurs apparues au cours de la période étudiée.
Il s'agit généralement de l'erreur absolue moyenne (MAE) ou de l'erreur quadratique moyenne (MSE). Par rapport au MAE, la MSE pondère plus fortement les erreurs impor-

tantes, c'est pour cela on a choisi la MSE comme indice de validation du modèle.
On définit l'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ d'un paramètre θ par la moyenne arithmétique des carrés des écarts entre les prévisions et les observations.
Elle est alors définie par la formule suivante :

$$\mathrm{MSE}(\hat{\theta}) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Dans ce qui suit, on a évalué le critère de comparaison, l'erreur de prédiction MSE, sur l'échantillon test.
❷ Analyse de la variance (ANOVA, p-valeur) :
Dans notre étude, la deuxième partie est d'étudier le meilleur modèle qu'on a trouvé et de savoir la significativité de chaque variable par intervalle dans ce modèle. Pour ce faire, on s'est basé sur une analyse de la variance en regardant p-valeur.
Les p-valeurs sont des indices compris entre 0 et 1. En effet, une variable significative est une variable qui doit être conservée dans le modèle. La p-valeur est utilisée pour quantifier la significativité statistique des variables.
Le statisticien Ronald Fisher a introduit les termes de significativité et l'utilisation de la p-valeur, il a précisé des seuils qui sont généralement pris pour référence comme indique le tableau 3.1 .

p-valeur	**siginificativité**
p-valeur ≤ 0.01	très forte
$0.01 <$ p-valeur ≤ 0.05	forte
$0.05 <$ p-valeur ≤ 0.1	faible
p-valeur ≥ 0.1	Pas de siginificativité

Tableau 3.1 – Seuil de p-valeur et son degré de significativité

La variable que nous cherchons à expliquer est le rendement du blé. Pour cela, nous utilisons les relevés climatiques résumés sous forme de méta-variables sur des intervalles de temps réguliers.
Maintenant, dans ce qui suit, je vais détailler comment on a construit ces méta-variables en exploitant des différentes agrégations.

3.1.2 Régression linéaire : Agrégation par moyenne

Dans cette section, on note $(X_t)_{t=1,\dots,T}$ une série climatique d'entrée, mesurée aux pas de temps $t = 1, \dots, T$ et Y la sortie du modèle. On dispose de n observations du couple $((X_t)_t, Y)$ que l'on note $((x_{it})_t, y_i)_{i=1,\dots,n}$.

On a entamé notre recherche par une étude préliminaire en prenant les moyennes des variables par intervalle sur les données X .

Ces nouvelles variables construites appelées méta-variables V forment des matrices moyennes de X et sont définies par la formule suivante pour tout $t = 1, \dots, T$ et I_u le u$^{\text{e}}$ intervalle de temps :

$$V_{iu} = \frac{1}{n} \sum_{t \in I_u} x_{it} \tag{3.2}$$

D'où notre modèle devient de la forme :

$$y_i = \alpha_0 + \sum_{u=1}^{5 \times d} \alpha_u v_{iu} + \epsilon_i \tag{3.3}$$

Tout d'abord, nous avons séparé les données en deux échantillons : un échantillon d'apprentissage sur lequel nous estimons les paramètres du modèle et un échantillon de test sur lequel nous évaluons sa qualité à l'aide du critère de comparaison MSE.

Pour le premier échantillon, on a estimé notre modèle par un modèle de régression linéaire en appliquant la moyenne pour différents découpages des variables temporelles en intervalles.

Pour chacun de ces modèles, on a étudié la significativité de l'intervalle dans le modèle en se reposant à une analyse de variance et en évaluant les p-valeur de chaque intervalle. Durant cette méthode,nous avons réalisé trois analyses.

Première analyse : Choix du nombre d'intervalles

Comme on a dit précédemment, pour évaluer la qualité du modèle, on s'est basé sur l'indicateur de qualité MSE. Après la création du modèle à partir du jeu de données apprentissage, on a appliqué une prévision sur l'échantillon test. Puis, on a étudié la relation entre Y_{pred_i} (la valeur prédite) en fonction du Y_{test_i} (la valeur initiale) pour le modèle initial qui est basé sur moyenne annuelle (donc sur un unique intervalle). La MSE dans notre cas est définit par :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_{pred_i} - Y_{test_i})^2 \tag{3.4}$$

avec : n est le nombre d'observations de l'échantillon test traité.

Après le traçage du modèle initial (créé à partir de l'échantillon apprentissage sur un intervalle durant 223 jours), on a constaté que ses prédictions sont faiblement corrélées par rapport aux valeurs initiales et que le modèle n'a pas un bon pouvoir prédictif. Pour l'améliorer, on a pensé à optimiser le nombre de variables en créant des métas-variables. Pour ce faire, on a appliqué plusieurs découpages de la période de culture en intervalles temporels.

On a divisé la période de culture étudiée en plusieurs intervalles égaux, puis, on a mesuré la qualité de chaque modèle par intervalle en calculant la MSE sur chaque morceau.

Notre but, comme on a dit, est de savoir quelles sont les variables et les intervalles qui influencent le rendement du blé. Après l'estimation du modèle de la régression linéaire sur nos données initiales, et suite à une prévision faite sur le jeu de données test, on a remarqué qu'il n'a pas un bon pouvoir prédictif.

Pour améliorer les résultats obtenus, on a pensé alors à appliquer d fois la division sur la période de la culture étudiée. Cette division permet de mieux étudier l'influence des variables en augmentant leur nombre. On a varié le nombre de divisions de 1 à 50 en obtenant ainsi un nombre de modèles variant de 1 à 50. La régression linéaire est ensuite appliquée pour chaque modèle.

Comme on a décrit précédemment, les données d'entrée de notre modèle initial sont une matrice de dimension (223,5). Après chaque division, les données d'entrée du modèle de la $d$ième découpage seront une matrice de dimension ($223{,}5 \times i$).

On a appliqué n découpages des variables initiales avec n varie entre 1 et 50.

L'étape suivante est de varier nos échantillons d'apprentissage et de test pour voir la variation de la MSE en fonction du nombre de découpages. Pour ce faire, on repose sur les boîtes à moustaches dites boxplot.

La boîte à moustaches est une traduction de Box & Whiskers Plot, est une invention de TUKEY (1977) pour représenter schématiquement une distribution. Cette représentation graphique peut être un moyen pour approcher les concepts abstraits de la statistique.[7]

Pour évaluer la qualité de chaque modèle, on a comparé la MSE de chacun selon les 50 groupes de découpage, on juxtapose sur le même graphique les 50 boîtes à moustaches définies respectivement par le groupe de découpage d avec d varie entre 1 et 50. Puis, on identifie les intervalles temporels les plus influents pour le rendement qui sont définis par les intervalles possédant la valeur de MSE la plus faible.

Deuxième analyse : Étude du meilleur modèle

Après avoir identifier le découpage adopté, on passe à la deuxième étape qui est l'étude du meilleur modèle dont on va examiner l'influence des intervalles sur la sortie. On regarde ici la significativité des variables du meilleur modèle en se basant sur l'analyse de variance ANOVA à l'aide des p-valeurs pour retrouver les motifs climatiques les plus influents

Troisième analyse : Étude du modèle réduit

Cette étude consiste à simplifier le modèle en ne gardant que les variables importantes, pour cela, on supprime celles de faible importance c'est-à-dire les variables qui ont une grande p-valeur (p-valeur > 0.1) puis on refait l'estimation du modèle sur le nouveau jeu de données réduit.

3.1.3 Régression linéaire : Agrégation par PLS

Dans cette partie, dans l'espoir améliorer l'approche précédente. Donc, pour la constructions des méta-variables on a remplacé l'agrégation par moyenne par intervalle par la projection des individus sur le premier axe de la régression PLS par intervalle.

Tout au long de cette approche, on exploite les mêmes indicateurs statistiques utilisés antérieurement.

1. Histoire

   L'algorithme de la régression PLS (régression des moindres carrés partiels ou en anglais Partial Least Squares) est une méthode itérative née des recherches de Svante Wold au début des années 1966 largement utilisée, notamment en chimiométrie dans l'agroalimentaire, permet la construction de modèles prédictifs quand les variables sont nombreuses et fortement corrélées entre elles. Lors de l'analyse de données spectrales (Near Infra-Red ou HPLC) discrétisées et donc toujours de grandes dimensions. Aujourd'hui, il connaît de nombreuses applications dans des domaines extrêmement variés.[8]

2. Présentation

   Pour régresser une variable Y (centrée) sur k variables explicatives $X^{(1)}$, $X^{(2)}$, ...,$X^{(k)}$ (centrées), la méthode PLS propose de trouver de nouveaux facteurs qui joueront le même rôle que les variables explicatives initiales. Ces nouveaux facteurs sont appelés variables latentes ou composantes. Chaque composante est une combinaison linéaire des variables $X^{(1)}$, $X^{(2)}$, ...,$X^{(k)}$. Les nouvelles composantes sont présentées par $t^{(1)}$, $t^{(2)}$, ...,$t^{(k)}$.

3. Objectif

   Son but principal est de construire de nouvelles variables qui soient combinaison linéaire des variables initiales sur lesquelles la variable réponse est régressée ainsi que créer des modèles prédictifs quand les variables sont nombreuses et fortement corrélées entre elles sur la base de maximisation de la covariance.

   Elle est appliquée lorsqu'il existe une forte colinéarité entre les variables explicatives, aussi lorsque le nombre de colonnes est plus grand que le nombre de lignes ou encore lorsqu'il y a des données manquantes. Pour plus de détails sur la PLS voir [9]

4. Pourquoi la régression PLS ?

   On a choisi cette méthode pour améliorer les résultats trouvés par la régression linéaire par moyenne vu qu'une régression PLS est une méthode d'analyse des données extrêmement puissante permettant d'améliorer la construction des métavariables en remplaçant le calcul de la moyenne par des composantes orthogonales les unes aux autres issues des variables explicatives. de départ. voir [10]

5. Principe :

   Cet algorithme est basé sur différentes étapes, pour plus de détail sur le principe de cette méthode voir [11].

   Le principe de la régression PLS est de construire des nouvelles variables explicatives $t^{(1)}$, $t^{(2)}$, ...,$t^{(k)}$, combinaisons linéaires des variables de départ $X^{(1)}$, $X^{(2)}$ telles que $t^{(j)}$=X c_j, qui soient orthogonales entre elles et classées par ordre d'importance.

On est dans le cas où X est une matrice de taille (223,5) et Y est univarieé, la PLS ici est appelée PLS1 et se définit itérativement.

✺ **Première étape :** X est noté $X^{(1)}$ et Y noté $Y^{(1)}$. La première composante PLS $t^{(1)} \in \mathbb{R}^n$ est choisie telle que

$t^{(1)} = \underset{t=X^{(2)}w, w\in\mathbb{R}^{\text{l}}, ||w||^2=1}{\text{argmax}} < t, Y^{(1)} >$

Ensuite, nous effectuons la régression univariée de $Y^{(1)}$ sur $t^{(1)}$ et donc

$$Y^{(1)} = \beta_1 t^{(1)} + \epsilon_1 \tag{3.5}$$

où β_1 est le coefficient de la régression estimé et ϵ_1 est le résidu de la régression simple sans constante.

✺ **Deuxième étape :** soit

$$Y^{(2)} = P_{t^{(1)}} Y^{(1)} = \hat{\epsilon_1} \tag{3.6}$$

la partie non encore expliquée de Y. soit

$$X^{(2)} = P_{t^{(1)}} x^{(1)} = \hat{\epsilon_1} \tag{3.7}$$

la partie de $X^{(1)}$ n'ayant pas encore servi à expliquer. La seconde composante PLS est choisie telle que : $t^{(2)} = \text{argmax}_{t=X^{(2)}w, w\in\mathbb{R}^{\text{l}}, ||w||^2=1} < t, Y^{(2)} >$ Ensuite, nous effectuons la régression univriée de $Y^{(2)}$ sur $t^{(2)}$.

$$Y^{(2)} = \beta_2 t^{(2)} + \epsilon_2 \tag{3.8}$$

où β_2 est le coefficient de la régression estimé et $\epsilon_2 == P_{t^{(1)}} Y^{(1)}$ est le résidu de la régression simple estimé par MC ✺ k^e **étape** : soit $Y^{(k)}$ la partie non encore expliquée de Y Soit

$$X^{(k)} = P_{t^{(k-1)}} X^{(k-1)} = \hat{\epsilon_1} \tag{3.9}$$

la partie de $X^{(k-1)}$ n'ayant pas encore servi à expliquer.La kème composante PLS est choisie telle que : $t^{(k)} = \text{argmax}_{t=X^{(k)}w, w\in\mathbb{R}^{\text{l}}, ||w||^k=1} < t, Y^{(k)} >$. Ensuite, nous

effectuons la régression univriée de $Y^{(k)}$ sur $t^{(k)}$.

$$Y^{(k)} = \beta_k t^{(k)} + \epsilon_k \tag{3.10}$$

où β_k est le coefficient de la régression estimé et ϵ_k== $P_{t^{(k)}} Y^{(k)}$ est le résidu de la régression simple.

La régression PLS cherche donc une suite de composantes PLS qui soient orthogonales entre elles et cela par construction. Puisque $t^{(j)}$ est une combinaison linéaire des colonnes de $X^{(j)}$, qui est par construction dans l'orthogonale de ($t^{(1)}$, $t^{(2)}$, ...,$t^{(j-1)}$). Ces composantes sont choisies comme maximisant la covariance (empirique) entre Y et une composante t quand X et Y sont centrées au préalable.

3.1.4 Régression linéaire : Agrégation par degré jour

Dans cette section, pour construire les nouvelles méta-variables, on a exploité une nouvelle approche en remplaçant les jours de culture par le degré jour.
Selon les agriculteurs, le développement de la culture est piloté par un indice thermique qui est le degré jour.

1. Définition

   Le degré jour de croissance est une mesure empirique utilisée pour calculer l'accumulation de chaleur qui sert à estimer la durée d'un développement agronomique tel que la croissance d'une plante en tenant compte de la température. Cette notion est particulièrement utilisée dans le domaine de l'agronomie.

2. Formule de culture

   Un degré-jour est calculé comme l'intégrale d'une fonction du temps qui varie généralement avec la température.

   Aussi pour le calculer, il faut établir une température de base qui est généralement la température en dessous de laquelle la croissance des plantes est de zéro (0°C pour le blé par exemple).

   Le degré jour se calcule ainsi :

   DJ= [(T_{max}+T_{min}) / 2] - T_{base}

3. Principe

   Après avoir calculé les degrés jour, on a réconstruit les nouvelles méta-variables en prenant des moyennes sur intervalle.Puis, on a créé le modèle par régression linéaire à partir de l'échantillon apprentissage afin de le prédire à partir de l'échantillon test.

Section IV : Résultats

4.1 Les résultats obtenus sur l'ensemble des données

L'objectif de l'étude était d'identifier les motifs climatiques qui rendent plus sensibles le rendement du blé ainsi que les intervalles temporels les plus influents. Pour ce faire, on a créé des méta-variables qui correspondent dans la première étude à des résumés (par moyenne) des relevés climatiques journaliers pour une période de temps puis dans la deuxième étude, ils correspondent à des nouvelles variables créées par la régression PLS. Enfin la troisième étude, on a remplacé les jours de culture par des degrés jour.
La période de culture du blé durant 223 jours a été subdivisée en intervalles de temps ce qui diminue le nombre de variable pour bien étudier leur influence.
On a calculé la moyenne de chaque variable climatique sur chacun des intervalles de temps puis on a appliqué la régression linéaire à ces méta-variables construites afin d'obtenir des méta-modèles. Ce chapitre est réparti en trois grandes parties. La première présente les résultats du modèle linéaire par agrégation par moyenne sur intervalle, la deuxième partie est dédiée à présenter les résultats de la régression linéaire par PLS. Enfin, la dernière partie est faite pour les résultats issus de la méthode utilisant le degré jour.

4.2 Résultats de la régression linéaire par moyenne

✿ Première analyse : Choix du nombre d'intervalles

On a commencé par tracer la régression linéaire du modèle initial basé sur moyenne par variable climatique en étudiant la relation entre la valeur prédite et la valeur initiale (traitée à partir du jeu de données test) et on a obtenu la figure 4.1.

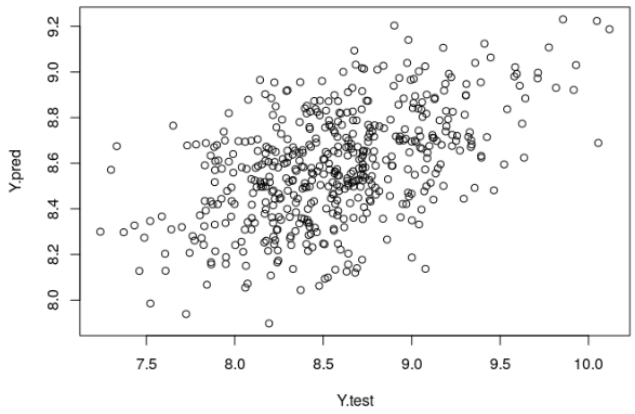

FIGURE 4.1 – Modèle basé sur moyenne annuelle sur un intervalle.

On sait qu'un modèle parfait est un modèle dont tous les points sont situés sur la diagonale. Ici on remarque que les prédictions sont faiblement corrélées aux valeurs initiales du jeu de donnée test et que le modèle n'a pas un bon pouvoir prédictif. Pour l'améliorer, on a pensé à diminuer le nombre de variables en créant des méta-variables. Pour ce faire, on a appliqué plusieurs découpages de la période de culture en intervalles temporels.
Nous présentons ici une simulation permettant d'illustrer l'évolution de MSE pour la sélection des intervalles recherchés.
Le graphique de la figure 4.2 représente, sous la forme de boxplot, les erreurs quadratiques moyennes de prédictions obtenues, évaluées sur un échantillon test pour 10 simulations différentes pour chaque découpage.

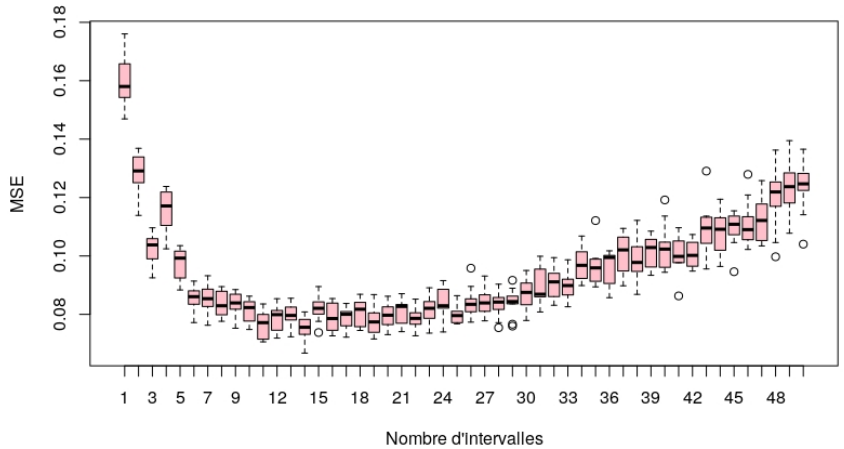

FIGURE 4.2 – Évolution du MSE en fonction du nombre de découpages

Selon cette figure, on remarque que la MSE était élevée au début et avec l'augmentation du nombre de variables, elle commence à se baisser. On remarque également que les performance sont très proches entre 10 et 22 découpages.
Parmi ces intervalles influents, on trouve que le modèle en 20 intervalles a une valeur moyenne de MSE petite. On peut dire alors que le modèle du découpage en 20 intervalles est le plus performant au sens du MSE qui vaut 0.08.

✿ Deuxième analyse : Étude du meilleur modèle

-Dans la première étude ; en effectuant plusieurs essais sur la base de données et en appliquant la technique de segmentation et découpage des intervalles, on constate que le modèle du 20ème découpage est le plus performant vu qu'il a la valeur de la mse la plus petite=(0.101).

On a regardé la significativité des variables du meilleur modèle (donc avec 20 intervalles) en se basant sur l'analyse de variance ANOVA à l'aide des p-valeurs et en tenant compte des deux autres modèles de divers nombre de découpage (10 et 35) et on a résumé cela dans la figure 4.3 .

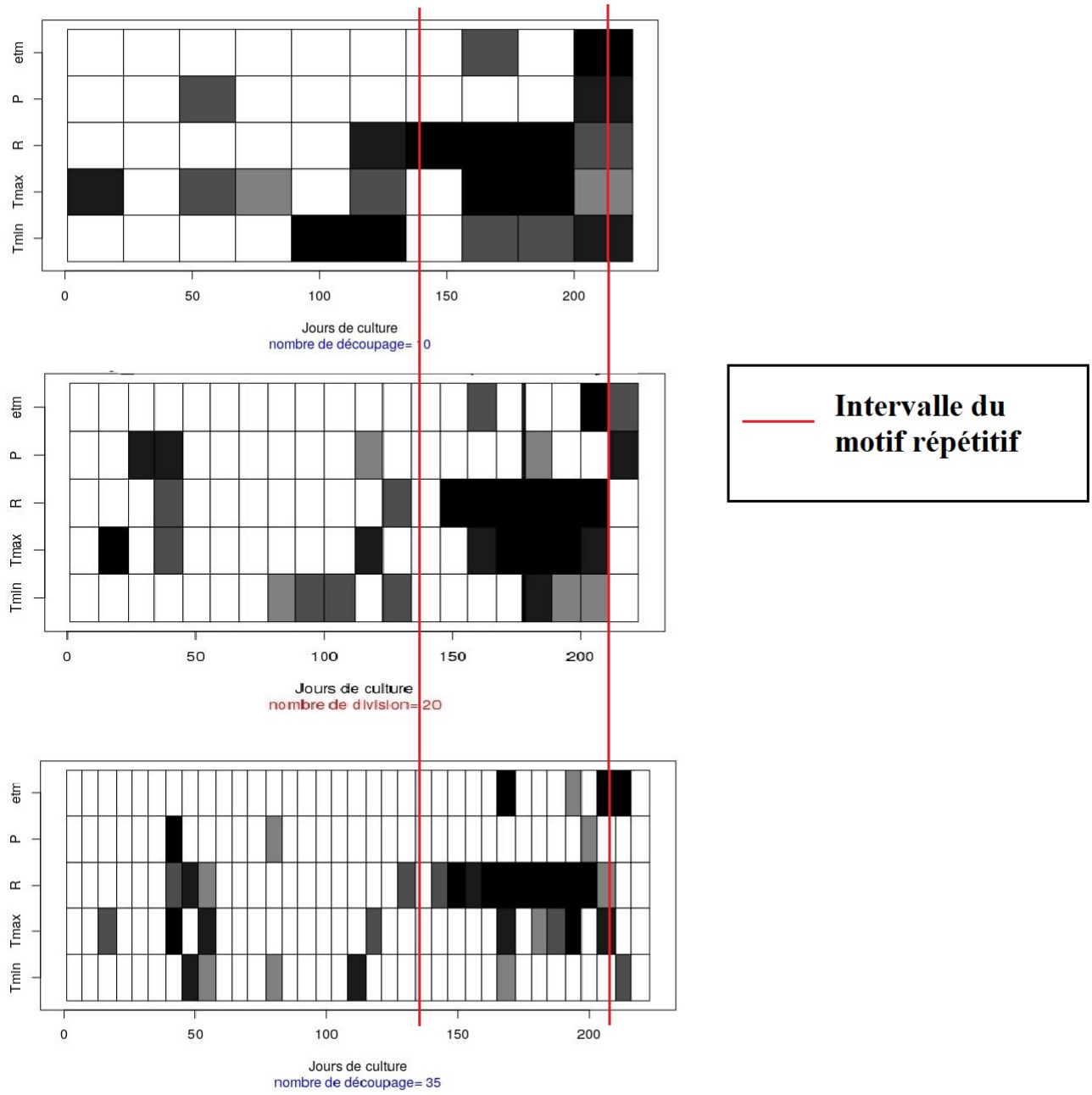

FIGURE 4.3 – Significativité Statistique des variables du modèle optimale.

Chaque case correspond à un couple donné (série climatique/intervalle temporel) . La couleur des cases se dégrade selon leur significativité : par exemple, les cases blanches correspondent aux couples les moins significatives c'est à dire les couples qui ont la p-

valeur ≥ 0.1 , alors que les cases noires sont associées aux plus significatifs c'est à dire qui ont la p-valeur ≤ 0.01.

On remarque qu'il y a une stabilité du modèle en passant d'un découpage à un autre et donc il y a un motif répétitif qui est présenté dans la figure entre les deux traits rouges.

L'étude des résultats donnés dans cette figure nous permet d'affirmer que les variables Tmax (température maximale) et R (ensoleillement) sont les variables les plus importantes et que l'intervalle le plus influent est [150,200] en jours de culture. Après la conversion de cet intervalle en jours juliens, on trouve que la période la plus importante est dans l'intervalle [71,121] c'est-à-dire la période entre la deuxième semaine de mars et la troisième semaine de mai.

Aussi, à l'aide de cette figure, on peut voir qu'il existe beaucoup de variables et intervalles qui ne sont pas significatifs, donc nous allons tenter de simplifier le modèle en réduisant le domaine traité en supprimant les variables et les intervalles qui n'ont pas une grande influence sur le rendement.

✿ Troisième analyse : Étude du modèle réduit

Le nombre de variables dans le modèle linéaire par moyenne étant élevé, j'ai procédé à une élimination de variables tout en observant la conséquence sur le comportement du MSE ajusté. Pour ce faire, une approche a été appliquée. Il s'agit de voir les variables qui ont une faible significativité dans le modèle initial et les éliminer. Après chaque élimination, le modèle linéaire est estimé sur le nouveau jeu de données et la MSE est recalculée toujours sur l'échantillon test.

Puis nous avons tracé la figure 4.4 qui décrit l'évolution de la MSE après la suppression de ces variables.

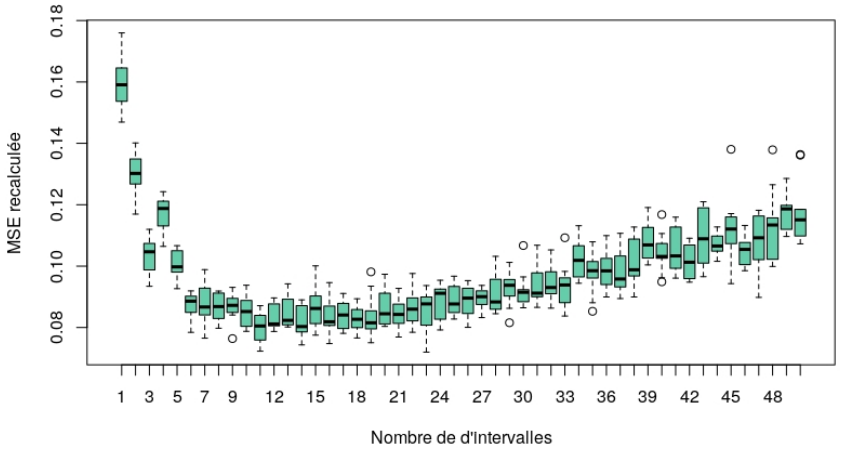

FIGURE 4.4 – Évolution de la MSE après la réduction.

Tout d'abord, on note que le graphique de la figure 4.4 est très proche de celui de la figure 4.2, aussi on constate que la MSE baisse puis réaugmente. De plus, on note que le modèle en 21 découpages est toujours le plus performant au sens de la MSE. On mesure maintenant la qualité de ce modèle optimal et on trace la figure 4.5 qui décrit la relation entre les nouvelles valeurs prédictives et les valeurs initiales de l'échantillon test.

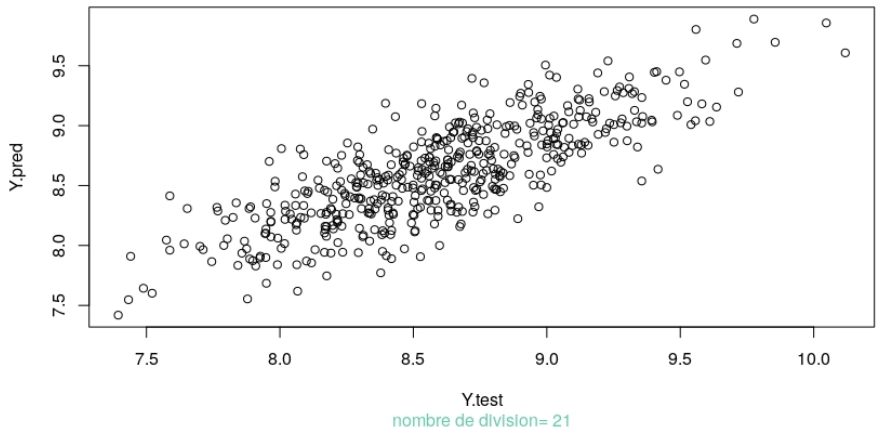

FIGURE 4.5 – Modèle optimal.

Par rapport au modèle initial, on observe selon ce graphique une amélioration au niveau de la corrélation entre la valeur prédite et la valeur observée du meilleur modèle qui a maintenant un bon pouvoir prédictif.

Conclusion

On retient des résultats de la méthode de régression linéaire par moyenne qu'une décomposition temporelle est très importante pour expliquer la sortie (le rendement du blé) à partir des données climatiques. Mais les résultats qu'on a trouvés avec cette méthode ne sont qu'une solution approximative et dans l'espoir d'obtenir une meilleure solution, on a remplacé cette approche par une autre basée sur la régression PLS par intervalle qui sera décrite dans le paragraphe suivant.

4.3 Résultats de la régression linéaire par PLS

À ce stade, on a changé l'agrégation 'moyenne par intervalle' par 'PLS par intervalle' et on a refait la même stratégie c'est-à-dire on se base toujours sur les deux indicateurs statistiques précédents.

✿ **Première analyse : Choix du nombre d'intervalles**

Comme précédemment, on a entamé l'étude par tracer la régression linéaire du modèle initiale basé sur la PLS et on a obtenu la figure ci-dessous.

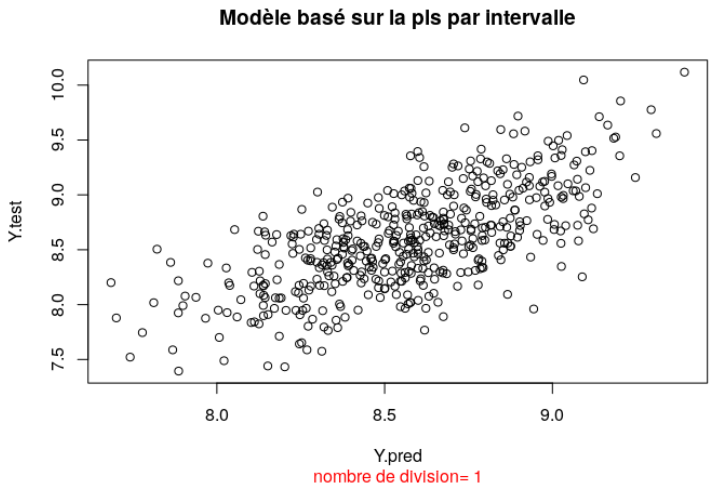

FIGURE 4.6 – Modèle sans division basé sur la PLS.

Selon cette figure, on peut voir que les valeurs prédites sont fortement corrélées et le modèle a un bon pouvoir prédictif ce qui améliore les résultats trouvées par l'agrégation

par moyenne.

Maintenant, on refait la même technique du découpage des intervalles toujours dans le but d'avoir un bon méta-modèle explicatif du rendement du blé et de sélectionner les variables climatiques les plus influentes.

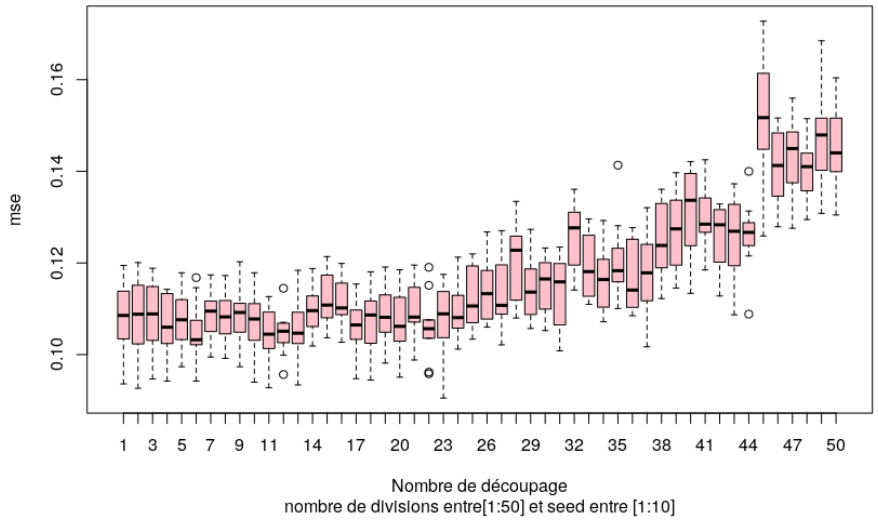

FIGURE 4.7 – Évolution de la MSE en fonction du nombre de découpages.

Selon cette figure, la meilleure valeur de la MSE correspond au découpage en 21 intervalles (comme dans la partie précédente) et qui vaut MSE=0,10.

La figure 4.8 présente l'évolution de p-valeur en fonction du nombre d'intervalles.

✿ Deuxième analyse : Étude du meilleur modèle

Comme on fait dans l'approche précédente, on a regardé la significativité des variables du meilleur modèle (donc avec 21 intervalles) en examinant toujours les P-valeurs et en tenant compte des deux autres modèles (en 10 découpages et en 35 découpages), puis on a résumé cela dans la figure 4.8

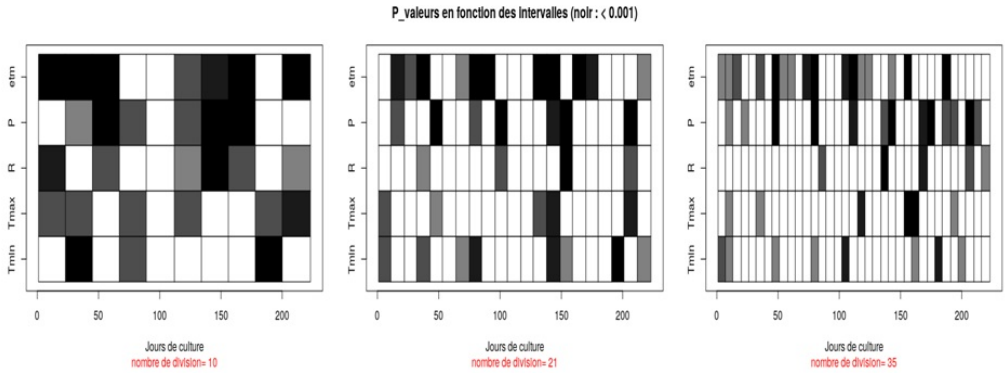

FIGURE 4.8 – Significativité statistique des variables du modèle optimal par PLS.

Nous avons étudié la stabilité de ce modèle en entraînant plusieurs découpages d'intervalle, à l'aide de cette figure, on a pu identifier les motifs climatiques les plus influents mais, contrairement à la partie précédente, on n'est pas apte à faire ressortir facilement d'intervalles influents. De plus, les variables climatiques les plus influentes changent d'un découpage à un autre. On conclut alors le modèle n'est pas stable.

✿ **Troisième analyse : Étude du modèle réduit**

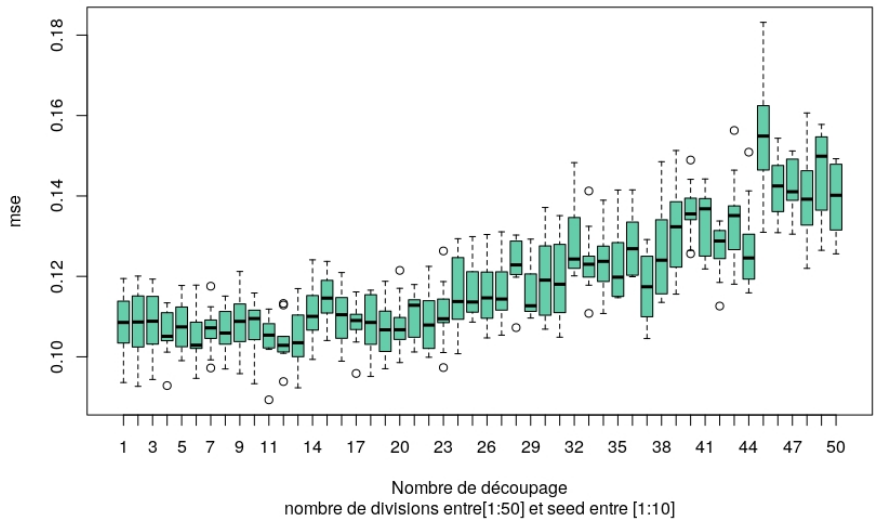

FIGURE 4.9 – MSE de la PLS après la réduction des variables.

En enlevant les variables les moins influentes, on trouve que les valeurs de MSE varient légèrement par rapport à celles d'avant. De plus, ils sont très proches d'un découpage

à un autre (elles sont encadrées entre 0.10 et 0.18), ce qui nous permet de garder les informations initiales du modèle.

Conclusion

À l'aide de la méthode de régression linéaire par PLS par intervalle, on a réussi à identifier les périodes les plus influentes mais on n'a pas réussi à les interpréter.
Aussi, en terme de la MSE, on note que cette approche nous donne le modèle considéré le plus performant (le modèle en 21 intervalles) avec une valeur de MSE qui est égale à 0.10 qui est donc plus grande que celle de la première approche (MSE=0.08).
On conclut, d'après nos résultats, que la régression par PLS est une méthode détériorée. Afin d'améliorer ces résultats, nous avons cherché à savoir si une procédure de méta-modélisation basée sur le degré jour était plus adaptée.

4.4 Résultats de la régression linéaire par degré jour

Dans cette section, il est question de faire une analyse plus approfondie en utilisant le degré jour au lieu du jour de culture. Tout d'abord, on a commencé par tracer un exemple des données climatiques sur le nouveau jeu de données. La figure 4.10 décrit les séries climatiques

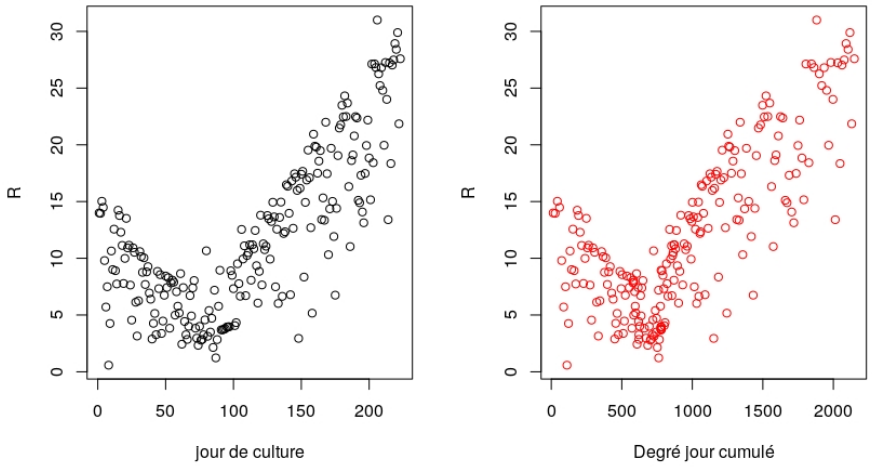

FIGURE 4.10 – L'ensoleillement en fonction de jour de culture et de degré jour cumulé.

Le graphique 4.10 nous présente un exemple de données climatiques qui est l'ensoleillement R. La figure à gauche correspond à l'évolution de l'ensoleillement en fonction de jours de culture (jeu de données initial) alors que celle à droite présente cette évolution en fonction du degré jour (jeu de données transformé).
Selon ces deux figures, on voit que les séries sont des séries fonctionnelles et complexes.
Pour appliquer cette approche, on a refait la même stratégie des deux approches précédentes en se basant toujours sur les deux indicateurs statistiques (MSE et l'analyse de variance ANOVA en regardant les p-valeurs).
La figure 4.10 donne la relation entre les valeurs prédites par cette approche et les valeurs initiales du jeu de données test.

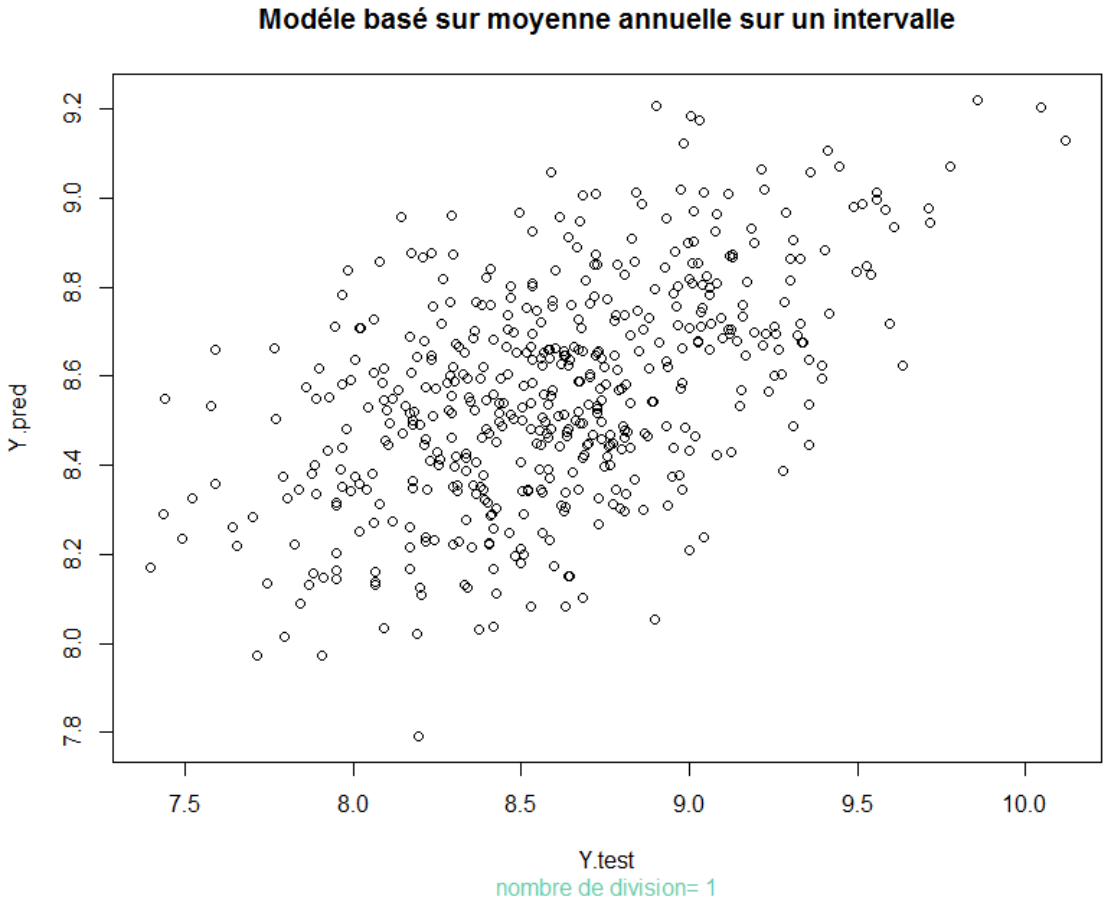

FIGURE 4.11 – Modèle linéaire basé sur le degré jour sur un intervalle.

Selon cette figure, on remarque que les prédictions du modèle ont une faible corrélation aux variables initiales de l'échantillon test ce qui ne permet pas au modèle d'avoir un bon pouvoir prédictif.

On a alors recours aux plusieurs découpages de la période de culture. Après avoir vu la MSE et la significativité de chaque intervalle, on a pu distinguer le meilleur modèle qui est le modèle en 11 intervalles avec la meilleure valeur de MSE égale à 0.07. La figure 4.11 résume l'étude de significativité des variables du modèle optimal trouvé.

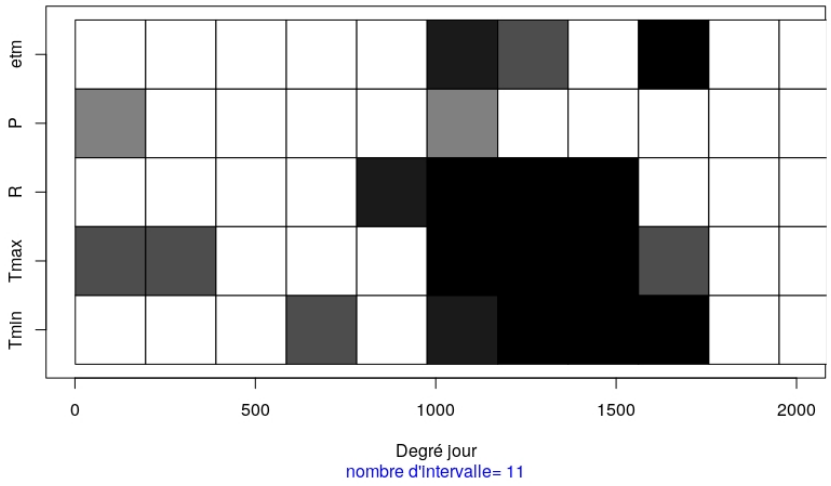

FIGURE 4.12 – Meilleur modèle en degré jour.

À l'aide de cette figure, on peut identifier le motif climatique le plus influent qui s'étale sur l'intervalle en degré jour [1000 : 1750] ainsi que sélectionner les variables climatiques qui contribuent le plus à la qualité de la prédiction sont la température minimale T_{min}, la température maximale T_{max} et l'ensoleillement R.

Conclusion

On retient des résultats de notre étude sur les trois approches le tableau 4.1 qui résume les résultats trouvées.

Tableau 4.1 – Résumé des résultats

Approche	**Modèle optimale**	**MSE**	**Intervalle important**	**Variables climatiques**
Moyenne	**20**	0.08	[71 :121]	T_{max} et R
PLS	**21**	0.10	Difficile à interpréter	Difficile à interpréter
Degré jour	**11**	0.07		T_{min}, T_{max} et R

Selon ce tableau, on peut retenir que la méthode de régression linéaire par degré jour donne le meilleur modèle au sens de la MSE.

Section V : Discussion

✿ Difficultés

La principale difficulté rencontrée est la nature multi-dimensionnelle des données, c'est-à-dire le fait que plusieurs variables fonctionnelles doivent être utilisées simultanément. De plus, j'ai rencontré des difficultés au niveau de la méthode PLS pour le choix des projections des individus au début ce qui a retardé le travail. Ainsi que la forte corrélation entre les nouvelles valeurs prédites sur ce jeu de données et les valeurs initiales sur l'échantillon de test : Cette forte corrélation rend peu interprétables les résultats.

Malgré toutes ces difficultés, les objectifs du stage à savoir : identifier les motifs climatiques et les intervalles de temps les plus influents du rendement du blé, ont pu être atteints. Cependant ce sujet reste largement ouvert, et on pourrait aller plus loin dans l'analyse en explorant d'autres pistes.

✿ Perspectives

Pour la suite du travail, on pourrait envisager de faire une régression non linéaire tel que forêt aléatoire, arbre de décision , SIR (Sliced Inverse Regression) dans l'espoir d'identifier les motifs climatiques qui rendent plus sensibles le rendement du blé ainsi que les intervalles temporels les plus influents. Une autre piste de réflexion serait de partir sur des intervalles de temps plus larges et de subdiviser les intervalles de temps les plus influents c'est-à-dire qui ont des faibles p-valeur, plutôt que de partir sur des intervalles et de regrouper les moins influentes.

✿ Conclusion

Mon passage à l'Unité Mathématique et Informatique Appliquées (MIAT) de l'INRA Toulouse m'a permis de mettre en application les connaissances acquises en statistiques et machine learning. C'était pour moi une expérience très enrichissante tant sur le plan professionnel que personnel. Ce stage m'a permis d'acquérir de nouvelles compétences puisque la plupart des outils utilisés (régression PLS, Git, etc) étaient nouveaux pour moi. J'ai également pu améliorer mes connaissances du logiciel statistique **R** surtout en programmation. J'ai eu la chance d'avoir des encadrants patients et disponibles pour répondre à toutes mes difficultés. À côté de toutes ces compétences techniques, j'ai aussi

eu l'occasion de découvrir le monde des statistiques dans le domaine d'agriculture et fort de cette expérience et en réponse à ses enjeux, j'aimerai beaucoup par la suite essayer de m'orienter via les statistiques et la science des données.

Bibliographie

[1] R Kpekou-Tossou. Analyse par simulation de l'interaction climat-rendement. *Master 2 intership report*, 40.

[2] S. Abis. Le blé en méditéerrannée : société, commerce et stratégie. 241, 2012.

[3] Nadine Brisson, Christian Gary, Eric Justes, R Rocheand Bruno Maryand Dominique Ripoche, Daniel Zimmer, Jorge Sierra, Patrick Bertuzzi, and P. Burger. An overview of the crop model stics. european journal of agronomy. 18(3) :309 – 332, 2003.

[4] Equipe Projet Stics. Les principes de fonctionnement stics. 2013.

[5] D. Sauvant, P. Schmidely, and J.J. Daudin. Les méta-analyses des données expérimentales : applications en nutrition animale. *INRA Prod Anim*, 73 :63–73, 2005.

[6] G. Bizouard. Méta-modélisation : état de l'art et application. 54 :25–35, 2012.

[7] M. Le Guen. La boîte à moustaches de tukey, un outil pour initier à la statistique. *Statistiquement Votre - SFDS,*, pages 1–3, 2001.

[8] E. Jakobowicz. Les modèles d'équations structurelles à variables latentes. 65 :34–43.

[9] C. Binard. Introduction à la régression pls. 44 :8–13, 2012.

[10] S.Bougard, A. Eslamiand E. Qannari, and S. Legley. Analyse en composante principal et régression pls multigroupes, application à l'usage du cannabis dans 13 pays européens. 6.

[11] P. André Cornillon and E. Matner Lober. Régression avec r. 242 :198–200, 2011.

[12] J. Jacques. Contribution à l'apprentissage statistique à base de modèles génératifs pour données complexes. 94 :55–70, 2012.

Annexes

Résultats intermédiaires de l'approche régression linéaire par moyenne par intervalle

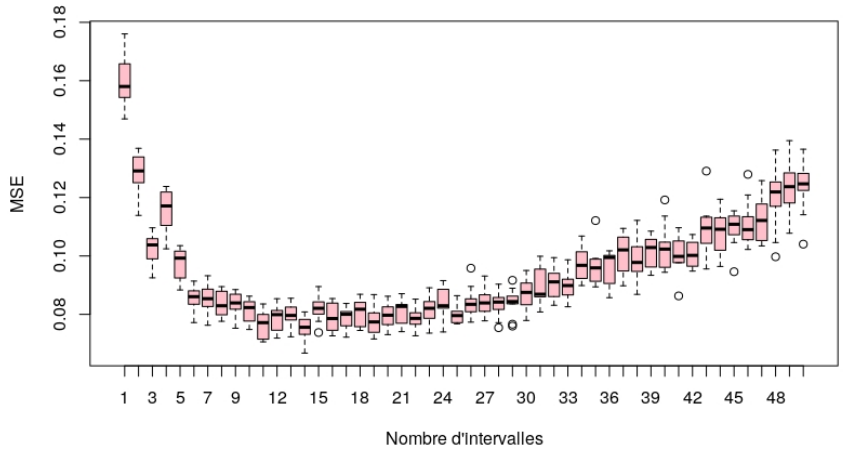

Figure A.1 – Mse après la réduction des variables (approche par moyenne)

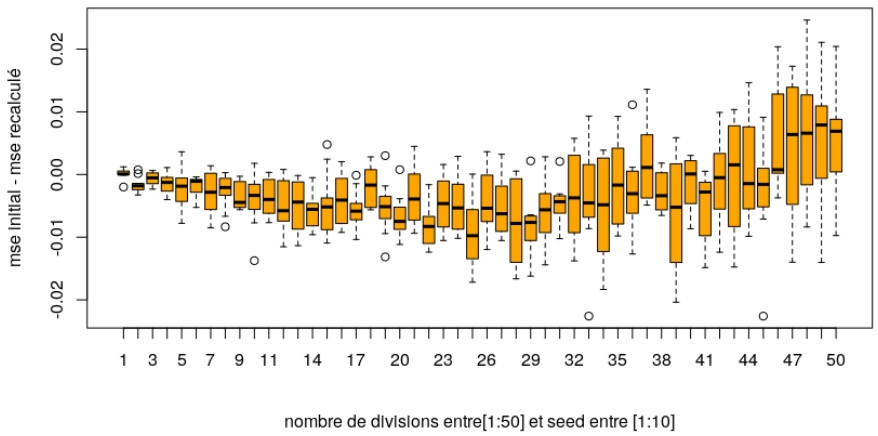

Figure A.2 – Boxplot de la différence du mse

Résultats intermédiaires de l'approche régression linéaire par PLS par intervalle

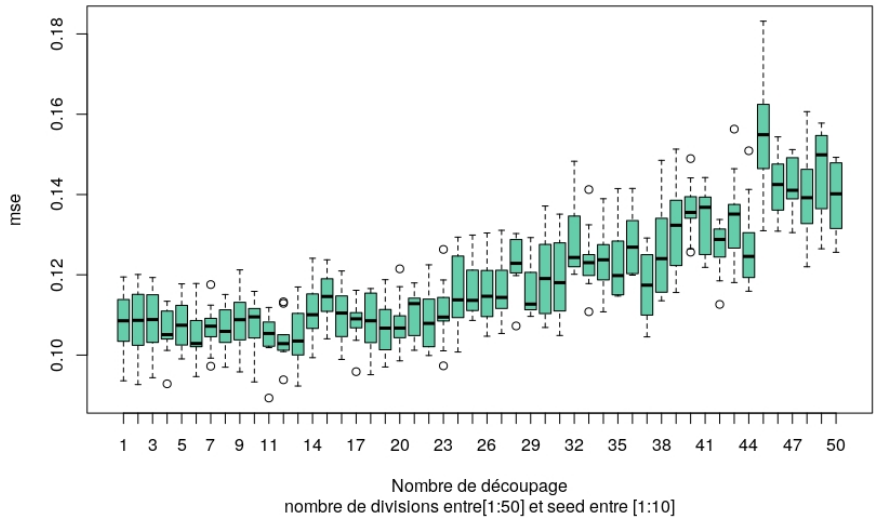

FIGURE A.3 – Mse après la réduction des variables (approche par PLS)

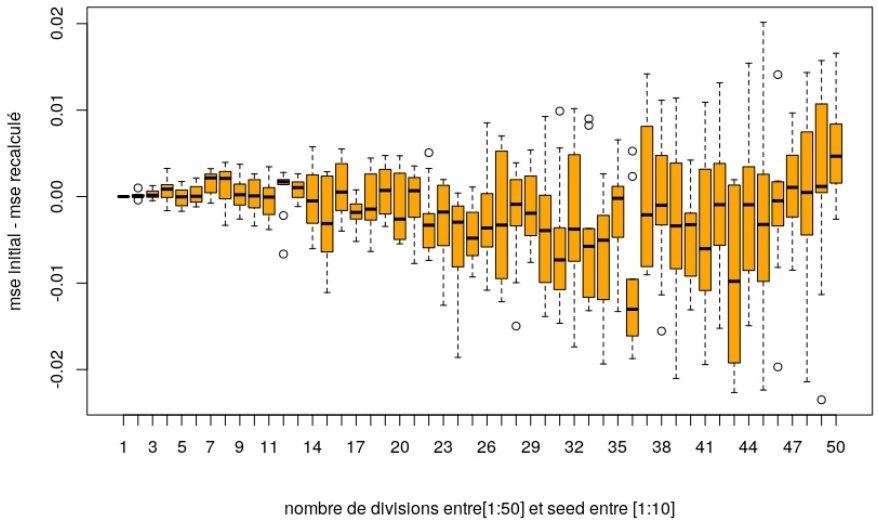

FIGURE A.4 – Différence entre mse avant et après la réduction,pls

Résultats intermédiaires de l'approche régression linéaire par degré jour par intervalle

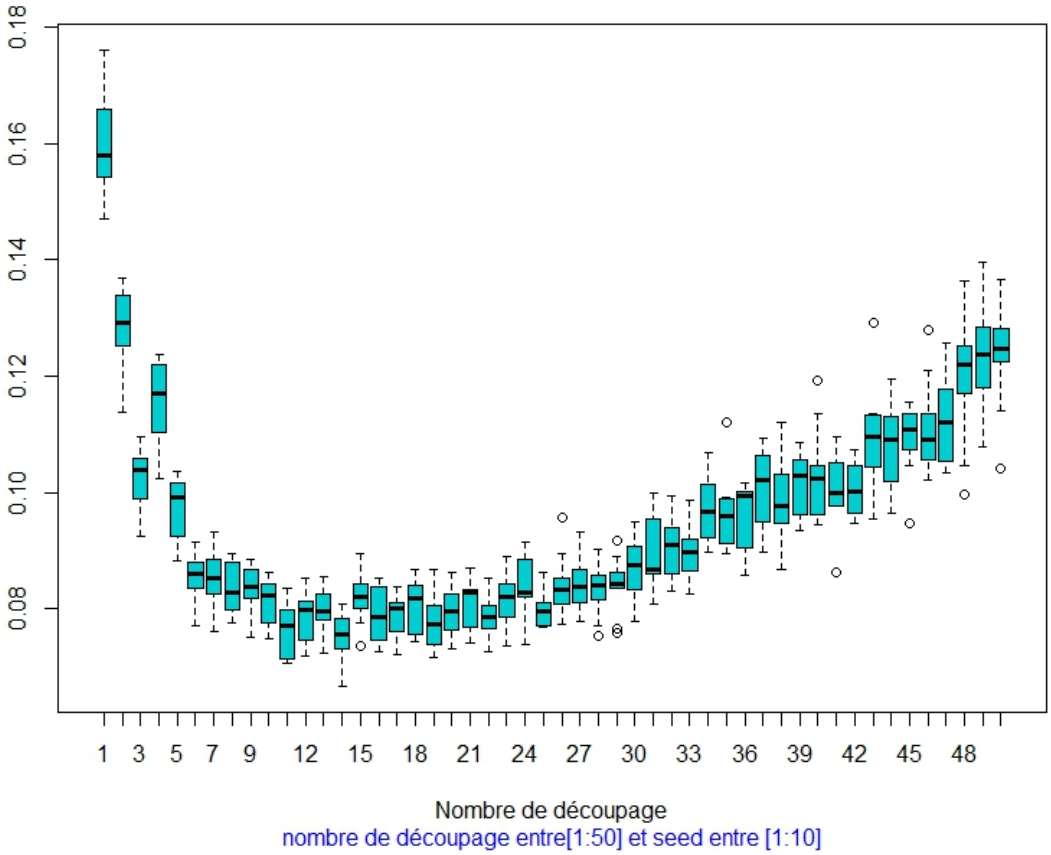

FIGURE A.5 – Mse avant la réduction des variables (approche par degré jour)

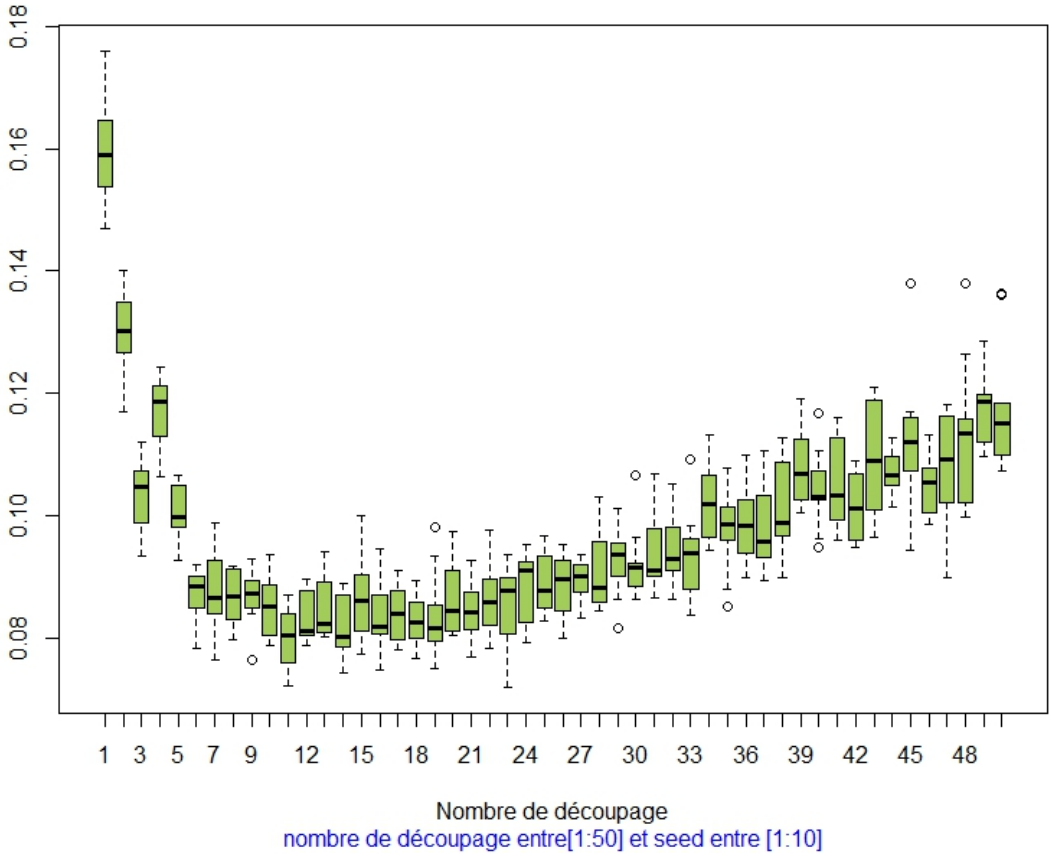

FIGURE A.6 – Mse après la réduction des variables (approche par degré jour)

Modèle	**Nombre de variables initiales**	**Nombre de variables restantes**	**MSE**
10	50	10	0.12
20	100	31	0.08
35	175	37	0.09

Tableau A.1 – Tableau récapitulatif pour l'approche par moyenne par intervalle

Modèle	**Nombre de variables initiales**	**Nombre de variables restantes**	**MSE**
10	50	10	0.108
21	100	12	0.101
35	175	13	0.115

Tableau A.2 – Tableau récapitulatif pour l'approche par PLS par intervalle

Dans cette partie, j'ai mis trois tableaux récapitulatifs des résultats de trois approches.

Modèle	**Nombre de variables initiales**	**Nombre de variables restantes**	**MSE**
6	30	8	0.08
11	55	8	0.07
31	155	11	0.09

Tableau A.3 – Tableau récapitulatif pour l'approche par degré jour