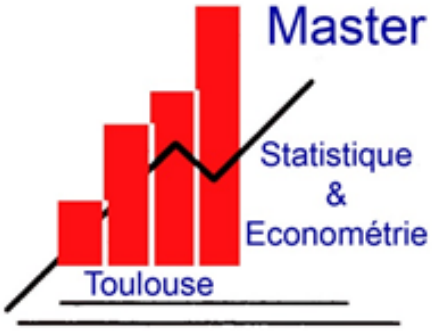
Toulouse School of Economics
École d'économie de Toulouse



Analyse par simulation de l'interaction Climat/Rendement

Rolande C. B. KPEKOU TOSSOU

Encadrement: **Victor Picheny** et **Nathalie Villa-Vialaneix**

Période de stage: Du 7 Avril 2015 au 15 Septembre

Remerciements

Ce travail ne se serait jamais concrétisé sans le concours de nombre de personnes qu'il est impératif de remercier.
Qu'il me soit permis d'exprimer ma profonde gratitude à mes encadrants **Victor Picheny** et **Nathalie Villa-Vialaneix** de m'avoir accompagnée durant tout mon stage, partageant mes difficultés et mes inquiétudes. Leur collaboration et leur orientation m'ont été d'une grande utilité. Je voudrais aussi exprimer toute ma reconnaissance envers **Robert Faivre** pour sa disponibilité et ses orientations toutes les fois que je l'ai sollicité.

Mes remerciements vont également à l'endroit de tout les membres de l'unité MIAT pour leur accueil et le très bon environnement dans lequel ils m'ont permis d'effectuer mon stage. Enfin, je remercie Franck Boizard, en stage à l'unité MIAT en même temps que moi, pour sa collaboration.

Sommaire

Liste des abréviations et acronymes

CART	Classification And Regression Trees
LHS	Latin Hypercube Sampling
MCO	Moindres Carrés Ordinaires
OOB	Out-Of-Bag
PCC	Partial Correlation Coefficient
SRC	Standardized Regression Coefficient
VSURF	Variables Selection Using Random Forets

Table des figures

Liste des tableaux

Analyse par Simulation de l'Interaction Climat/Rendement

Résumé

Cette étude intitulée “Analyse par simulation de l'interaction climat rendement”, a pour objectif de comprendre la manière dont le climat, ou plutôt les motifs climatiques, influencent le rendement du tournesol. Pour atteindre cet objectif, j'ai travaillé à partir d'un jeu de données composé de diverses séries climatiques annuelles (qui correspondent à 190 relevés climatiques sur cinq stations françaises), d'une part, et de divers ensembles de traits phénotypiques du tournesol, d'autre part. Ces deux catégories de variables sont utilisées comme des paramètres d'entrée d'un simulateur numérique dynamique (SUNFLO) pour le calcul du rendement du tournesol pour tous les croisements de ces deux données.

Le modèle SUNFLO permet de calculer les sorties (la valeur du rendement) pour une entrée donnée mais est coûteux en temps de calcul : il est donc difficile voire impossible de l'utiliser pour une analyse de sensibilité ou une étude exhaustive de l'influence des valeurs des variables d'entrée sur la sortie. Nous avons donc eu recours à une approche dite par méta-modélisation afin d'identifier les variables les plus influentes et les périodes les plus sensibles : ce type d'approche consiste à construire un modèle statistique permettant d'approcher le modèle SUNFLO avec des temps de calcul moindres tout en conservant de bonnes performances prédictives puis à effectuer une analyse de sensibilité ou 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 dans un premier temps et dans un second temps, une approche plus flexible, les forêts aléatoires. Dans le cas du modèle linéaire, des indices de sensibilité ont été calculés et dans le cas des forêts aléatoires, l'importance des variables a été utilisée. Il a été nécessaire de faire une décomposition en semaines (26 semaines) de la période de culture (avril-septembre) et de construire des méta-variables sur ces semaines pour avoir un méta-modèle explicatif du rendement du tournesol. L'estimation du modèle linéaire indique la significativité de toutes les variables sur presque toutes les semaines. Un tel résultat ne permettant pas la recherche de motifs climatiques, d'autres approches telles que la sélection de variables, la fusion de variables, etc, toutes basées sur les forêts aléatoires, ont été utilisées pour isoler les variables et les intervalles de temps les plus influents. Les résultats obtenus sur le modèle SUNFLO sont partiellement en adéquation avec le cycle de production du tournesol.

Mots clés : méta-modélisation, analyse de sensibilité, forêts aléatoires, importance des variables.

Abstract

This study entitled "Analysis by simulation of the climate/yield interaction" aims at understanding how the climate, or rather the climate patterns, affect the yield of sunflower. To achieve this goal, I worked from a data set composed of various annual climate series (corresponding to 190 climate records in five French stations), on the one hand, and various sets of phenotypic traits of sunflower, on the other hand. These two categories of variables are used as input parameters of a dynamic numerical simulator (SUNFLO) for calculating the yield of sunflower for all crosses of these two data.

SUNFLO model allows us to compute the yield (the value of yield) for a given input but is costly in terms of computation time, so it is difficult or impossible to use it for a sensitivity analysis or exhaustive study of the influence of values of the input variables to the output. We therefore used an approach called meta-modeling to identify the most influential variables and the most sensitive periods that consist of building a statistical model to approach the SUNFLO model with less computation time while maintaining good predictive performances then to perform a sensitivity analysis or research variables that contribute most to the quality of the prediction. In this study, I used as meta-models, the linear model at first and in a second time, a more flexible approach, random forests. In the case of the linear model, sensitivity indices were calculated and in the case of random forest variable importance was used. It was necessary to decompose the sunflower culture period into weeks (26 weeks) and build meta-variables on these weeks to have an explanatory meta-model. The estimation of linear model shows the significance of all variables on almost every week. Such a result does not allow the search for climate patterns, other approaches such as variables selection, variables merging, etc, all based on random forests, have been used to isolate the most influential variables and time intervals . The results obtained on SUNFLO model are partially consistent with sunflower production cycle.

Keywords : meta-modeling, sensitivity analysis, random forests, variable importance.

Introduction

Le stage de cinq mois effectué à l'Institut National de la Recherche Agronomique sous la responsabilité de Victor Picheny et de Nathalie Villa-Vialaneix conclut une formation de Master 2 en Statistique-Économétrie à l'École d'Économie de Toulouse. Ce stage porte sur l'analyse par simulation de l'interaction climat/rendement du tournesol. En effet, le climat a une influence très forte sur la production agricole, qui peut être considérée comme l'une des activités humaines les plus dépendantes des conditions météorologiques. Cette influence peut être entre autres le décalage du calendrier agricole, la modification de la qualité des produits et surtout la baisse ou l'augmentation du rendement.

Le but de ce travail est donc de tenir compte de l'aléa climatique dans les études de rendement. Nous nous intéressons dans cette étude aux variétés du tournesol à l'aide d'un simulateur dynamique : le modèle SUNFLO. SUNFLO est un modèle de simulation qui permet de prédire le rendement et la teneur en huile du tournesol à l'échelle d'une parcelle à l'issu d'une période culturale d'un an. Les données utilisées par ce modèle concernent le milieu (sol, climat), la conduite de culture (date de semis, irrigation, etc) et la variété (les traits phénotypiques). L'originalité de ce modèle tient au fait qu'il permet de tenir compte des différences entre les variétés sur différents critères (leur phénologie, leur comportement face au stress, etc). Dans le cadre de ce stage, nous avons analysé l'influence d'entrées (le climat, les traits phénotypiques) sur une sortie (le rendement du tournesol). L'objectif final est d'identifier les motifs climatiques les plus influents sur le rendement du tournesol. Différentes approches ont été explorées telles que : l'analyse de sensibilité, le calcul d'importance de variables, la sélection de variables, etc. Ce stage relève bien donc d'un travail de statisticien et est donc en parfaite adéquation avec ma formation.

La suite du document sera organisée de la manière suivante : nous allons présenter la problématique du stage et les données utilisées, la méthodologie adoptée et les outils statistiques utilisés et nous terminerons par les résultats obtenus. Mais avant tout cela, nous allons faire une brève présentation de la structure d'accueil.

Structure d'accueil

1.1 Institut National de la Recherche Agronomique

L'Institut National de la recherche Agronomique (INRA) est un organisme de la recherche scientifique public, 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. Il a pour missions de :

❶ produire et diffuser des connaissances scientifiques ;

❷ concevoir des innovations et des savoir-faire pour la société ;

❸ éclairer par son expertise, les décisions des acteurs publics et privés ;

❹ développer la culture scientifique et technique et participer au débat science/société ;

❺ former à la recherche et par la recherche.

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)

Elle 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 avoir les autres départements de l'INRA. L'unité est composée de deux grandes équipes de recherche :

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

Ces deux équipes sont rattachées à des plateformes pour mener à bien leur travaux de

recherche (GIS GENOTOUL, RECORD et SIGENAE). 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.

Problématique et données

Depuis les années 80, le tournesol demeure une composante majeure des assolements du grand Sud-Ouest de la France. Sa culture est appréciée par les agriculteurs pour sa place dans la rotation, sa rusticité, sa conduite économe en intrants. À côté de ces atouts agronomiques et environnementaux, le tournesol bénéficie de débouchés assurés tant sur le plan alimentaire (alimentation humaine et animale) qu'industriel. Pour conforter sa place dans les systèmes de grande culture et renforcer sa compétitivité, la productivité du tournesol doit s'améliorer et gagner en régularité. plusieurs voies sont à explorer, dont la maîtrise de l'interaction climat/rendement du tournesol.

2.1 Problématique du stage

De nombreux travaux récents en agronomie portent sur la modélisation de systèmes de culture (à différentes échelles : paysage, parcelle, plante) tenant compte de conditions climatiques (ensoleillement, pluviométrie, etc.) représentées typiquement sous forme de séries temporelles. Ces conditions ayant une influence très forte sur le comportement de la majorité de ces systèmes, intégrer l'aléa climatique aux études d'analyse de risque ou d'optimisation de rendement est une question primordiale. Cependant, la complexité des interactions plante/climat à l'œuvre ne permet pas d'étudier les 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 le cadre de ce stage, on s'intéresse à l'étude de variétés de tournesol à l'aide d'un modèle de simulation dynamique (SUNFLO). Dans ce cadre-ci, il est important de comprendre la manière dont le climat influence le rendement. L'objectif final de cette étude sera de mettre en valeur les motifs climatiques les plus influents pour le rendement, indépendamment de la plante ou en fonction de celle-ci, et d'estimer à partir de ceux-ci la distribution du rendement conditionnellement aux caractéristiques de la plante. D'un point de vue théorique, les entrées climatiques du

modèle SUNFLO peuvent être comprises soit comme des séries temporelles multivariées, soit comme des variables aléatoires fonctionnelles.

2.2 Présentation des données

Les données utilisées dans cette étude sont relatives au climat et aux traits phénotypiques du tournesol. Ces deux catégories de variables seront utilisées comme des paramètres d'entrée du simulateur dynamique (SUNFLO) pour le calcul du rendement. Cette section présente successivement le simulateur SUNFLO, les variables climatiques et les traits phénotypiques.

2.2.1 Le modèle SUNFLO

SUNFLO est un modèle de simulation du fonctionnement de la culture du tournesol construit par l'INRA en collaboration avec le CETIOM (Centre Technique Interprofessionnel des Oléagineux Métropolitains). Il a été mis en place pour représenter de manière dynamique l'interaction entre une variété, son milieu (le climat, le sol) et la conduite culturale. Il prend en compte trois familles de paramètres en entrées et renvoie deux sorties :

En entrée :

❶ les traits phénotypiques du tournesol ;

❷ les données climatiques ;

❸ le contexte de culture.

À partir de ces trois types de variables, le modèle SUNFLO calcule :

— le rendement sur la parcelle en quintal par hectare ;

— la teneur en huile en pourcentage.

Dans cette étude, seuls les deux premiers paramètres d'entrées (climat, phénotypes) seront utilisés et seul le rendement sera calculé. Pour plus de détails sur ce modèle, voir [1, 2]

Le modèle SUNFLO peut être schématisé comme sur la figure 2.1.

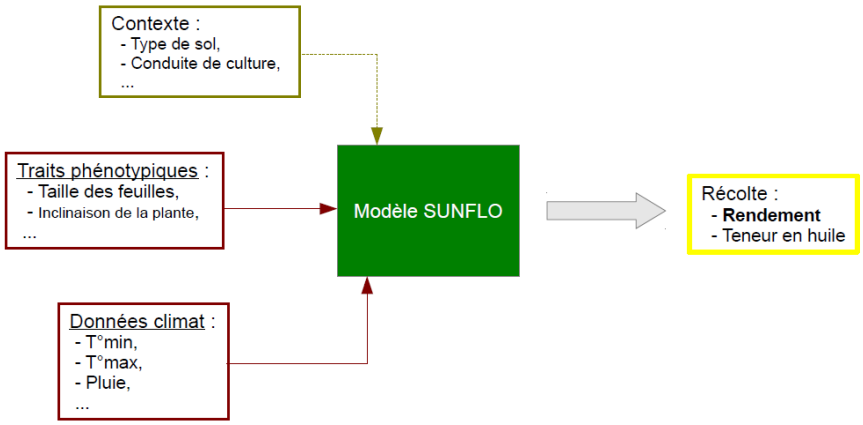


FIGURE 2.1 – Représentation schématique du modèle SUNFLO

2.2.2 Les données climatiques

Les données climatiques sont des relevés journaliers de certaines caractéristiques décrites sous forme de séries temporelles. Elles concernent 5 variables : la température minimale (Tmin), la température maximale (Tmax), l'évapotranspiration (ETP), l'ensoleillement (RAD) et les précipitations (Pluie). Cet ensemble de données contient 190 relevés climatiques sur cinq stations (Avignon, Blagnac, Dijon, Poitiers et Reims). La période de culture du tournesol étant fixée entre avril et septembre, les analyses seront faites uniquement sur cette période. Les données climatiques sont donc composées de 5 séries de 183 mesures journalières pour 190 années de simulation, comprises entre 1975 et 2012. Un exemple de données climatiques est illustré sur la figure 2.2.

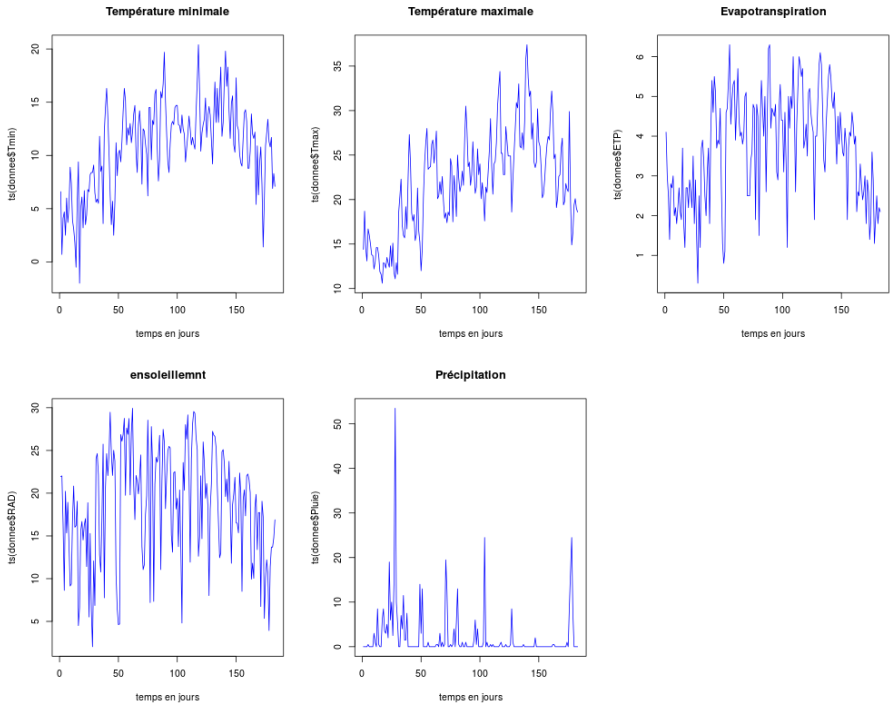


FIGURE 2.2 – Un exemple de données climatiques

2.2.3 Les traits phénotypiques

Huit traits phénotypes caractérisent les variétés de tournesol :

❶ durée de la phase levée-floraison (TDF1) ;

❷ durée de la phase levée-maturité (TDM3) ;

❸ nombre de feuilles potentiel (TLN) ;

❹ coefficient d'extinction du rayonnement lors de la phase végétative (K) ;

❺ rang (depuis le sol) de la plus grande feuille du profil foliaire à la floraison (LLH) en feuilles ;

❻ surface de la plus grande feuille de profil foliaire à la floraison (LLS) en cm^2 ;

❼ seuil de réponse de l'expansion foliaire à une contrainte hydrique (LE) ;

❽ seuil de réponse de la conductance stomatique à une contrainte hydrique (TR).

Un plan d'expérience LHS a été généré en croisant 1000 variétés de tournesol aux 190 climats et le rendement a été calculé pour ces 190000 croisements.

Méthodologie et outils

L'objectif de cette étude est d'identifier les variables climatiques et les intervalles intertemporels les plus influents du rendement du tournesol. Le problème peut se formaliser sous la forme d'une question dans laquelle on cherche à comprendre l'influence de variables d'entrées $X \in \mathbb{R}^d$ (le climat et les traits phénotypiques) sur une variable de sortie $Y \in \mathbb{R}$ (le rendement). Le modèle SUNFLO 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 permettant d'approcher le modèle SUNFLO avec des temps de calcul moindres et qui permet d'interpréter facilement la relation entrée-sortie. Le méta-modèle est construit à partir d'observations qui ont été préalablement générées par le modèle SUNFLO : $(x_1, y_1), \ldots, (x_n, y_n)$. On peut ensuite faire une analyse de sensibilité sur le méta-modèle obtenu. Cette section présente les méta-modèles utilisés et l'analyse de sensibilité. Dans la suite, on notera en particulier :

- $\mathbf{X}$ la matrice de dimension $n \times d$ des observations d'entrée, $\mathbf{X} = (x_1^T, \ldots, x_n^T)^T$, et $\mathbf{Y}$ le vecteur de taille n des observations de sortie, $\mathbf{Y} = (y_1, \ldots, y_n)^T$;
- lorsqu'on travaille sur une variable d'entrée particulière, $j \in \{1, \ldots, d\}$, 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

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 SUNFLO. 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 $Y \approx \text{metamodele}(X) + \epsilon$. ϵ est la différence entre les valeurs du modèle initial et les valeurs fournies par le méta-modèle.

Plan d'expérience

Un plan d'expérience est une sélection de combinaisons de valeurs des facteurs qui fournira, à moindre coût, des informations sur la relation entrées-sortie. Initialement définis pour l'expérimentation réelle, les plans d'expérience peuvent être étendus au contexte des expériences issues de simulateur, que l'on nommera expériences numériques ou simulées. Un plan d'expérience peut être vu comme un ensemble de d vecteurs d'entrée tels que les n observations soient bien réparties sur l'ensemble de définition des variables explicatives. Il existe plusieurs types de plan d'expérience, le plus simple est le plan de Monte Carlo avec une répartition des vecteurs d'entrée suivant une loi prédéfinie, ceux-ci sont réajustés selon les bornes inférieure et supérieure des entrées du modèle. Nous avons également les plans Hypercubes latin ou LHS (Latin Hypercube Sampling), qui sont créés en découpant chaque dimension de l'espace en n intervalles de même longueur. Un point est ensuite choisi par intervalle pour chaque dimension, ce qui donne un maillage de dimension d composé de n^d cellules de même taille.

3.1 Modèle linéaire et analyse de sensibilité

Il existe plusieurs méthodes de méta-modélisation telles que les modèles linéaires, les réseaux de neurones, les forêts aléatoires. Nous allons commencer par le modèle linéaire, le plus simple et le plus connu.

3.1.1 Modèle linéaire

Le principe du modèle linéaire est d'exprimer la sortie sous forme de combinaison linéaire des facteurs (entrées). On cherche une fonction f, linéaire en $X = (X^1, \ldots, X^d)$, telle que $Y = f(X)$. Le modèle peut s'écrire :

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

ou sous forme matricielle : $\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\beta + \epsilon$ avec $\beta = (\beta_1, \ldots, \beta_d)^T$, $\mathbf{1}_n$ le vecteur composé de n fois la valeur 1 et $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$.

Les paramètres du modèle peuvent être déterminés par les moindres carrés ordinaires (MCO), ce qui revient à résoudre ce programme de minimisation sur un échantillon de taille n :

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta_i - \beta_0)^2 \tag{3.2}$$

La résolution de ce programme donne l'estimateur MCO de la forme :$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Les paramètres ainsi déterminés peuvent alors permettre de prédire la sortie Y à partir de nouvelles données d'entrée.

3.1.2 Analyse de sensibilité

L'objectif de l'analyse de sensibilité est de quantifier l'impact de chaque entrée d'un modèle sur sa sortie. Elle étudie comment les perturbations sur les variables d'entrée du modèle engendrent des perturbations sur la variable de sortie. L'analyse de sensibilité permet de :

- hiérarchiser les facteurs (entrées) : identifier les variables qui contribuent le plus à la variabilité de la réponse du modèle, et donc nécessitent plus de précision.
- Mieux appréhender et comprendre le phénomène modélisé, en éclairant les relations entre les variables.

On peut distinguer trois groupes de méthodes d'analyse de sensibilité :

1. Les méthodes d'analyse locale : étudient quantitativement l'impact de la modifica-

tion d'un facteur à un niveau local, c'est-à-dire dans le voisinage d'une valeur x_o donnée des facteurs.

2. Les méthodes de criblage : elles consistent en une analyse qualitative de la sensibilité de la variable de sortie aux variables d'entrées du modèle.
3. Les méthodes d'analyse globale : elles s'intéressent à la variabilité de la sortie du modèle dans l'intégralité de son domaine de variation. On regarde de manière globale la part de la variance de sortie qui est due à telle entrée ou tel ensemble d'entrées.

Cette étude se focalise sur les méthodes d'analyse de sensibilité globale, car elles permettent non seulement de hiérarchiser les facteurs, mais aussi de prendre en compte tout le domaine de variation de la réponse. Pour faire une analyse de sensibilité globale, on se sert des indices de sensibilité.

Indices de sensibilité sur modèle linéaire

Nous nous intéressons aux méthodes d'analyse de sensibilité basées sur l'étude de la variance. Elles consistent à déterminer quelle part de variance de la réponse est due à la variance de chaque variable d'entrée. Nous définissons alors sous l'hypothèse de linéarité les indices de sensibilité suivants :

☛ **Indice SRC (Standardized Regression Coefficient)**

Il exprime la part de la variance de la réponse de Y due à la variance de la variable X^j. Cet indice compris entre 0 et 1, est le carré du coefficient de corrélation linéaire entre la réponse du modèle et ses variables d'entrée. Soit le modèle 3.1, les X^j étant indépendants, la variance de Y peut s'écrire :

$$V(Y) = \sum_{j=1}^{d} (\beta_j)^2 V(X^j)$$

où $(\beta_j)^2 V(X^j)$ est la part de la variance de la réponse due à la variable X^j. On définit alors l'indice de sensibilité de Y à X^j par le rapport de la part de variance de Y due à la

variance de X^j sur la variance totale.

$$SRC_j = \frac{(\beta_j)^2 Var(X^j)}{Var(Y)} = \frac{(\beta_j)^2 Var(X^j)}{\sum_{j=1}^{d} (\beta_j)^2 V(X^j)}$$

☛ **Indice PCC (Partial Correlation Coefficient)**

Dans la pratique, il arrive que la corrélation entre Y et X^j soit due à une autre variable, surtout lorsqu'il existe une forte corrélation entre les variables. L'indice PCC a été proposé pour pallier cet effet. Cet indice exprime la sensibilité de Y à X^j en éliminant l'effet des autres variables, sous l'hypothèse de linéarité. Il est donné par :

$$PCC_j = \frac{cov(Y, X^j | \mathbf{X}^{-j})}{\sqrt{V(Y/\mathbf{X}^{-j}) V(X^j | \mathbf{X}^{-j})}} \tag{3.3}$$

avec $\mathbf{X}^{-j}$ qui représente $\mathbf{X}$ privé de sa composante j. Pour plus de détails sur l'estimation des indices PCC, voir [3].

3.2 Forêts aléatoires et calcul d'importance de variables

Introduites par Brieman [4] en 2001, les forêts aléatoires font partie des techniques d'apprentissage automatique. Elles se basent sur la construction d'un grand nombre d'arbres de régression obtenus par la méthode CART (Classification And Regression Trees). Elles peuvent être utilisées sur des variables de sortie de type qualitative ou quantitative. Ce sont des méthodes non paramétriques, c'est-à-dire qui ne nécessitent pas d'hypothèse à priori sur la forme du lien entre la sortie Y et les entrées X. Les forêts aléatoires présentent aussi l'avantage d'être applicables lorsqu'on a un nombre très élevé de facteurs ou un échantillon de taille faible. La construction des forêts aléatoires étant basée sur les arbres de régression, il est nécessaire de commencer par une brève présentation des arbres de régression.

CART correspond à deux situations distinctes selon que la variable à expliquer est qualitative (arbre de classification) ou quantitative (arbre de régression). Nous allons nous

limiter aux arbres de régression dans la suite. L'idée derrière les arbres de régression est de faire un découpage binaire et itératif de l'espace des variables explicatives. On commence par découper la racine de l'arbre qui contient toutes les observations en deux nœuds fils. Un nœud est défini par un élément de la forme $\{X^j \leq c\} \cup \{X^j > c\}$, où $j \in \{1, \ldots, d\}$ et $c \in \mathbb{R}$.

Étant donné un nœud $\tau \subset \{1, \ldots, n\}$, ce nœud est divisé en deux nœuds fils en sélectionnant la variable X^j et le seuil c qui minimisent critère d'hétérogénéité suivant :

$$\sum_{i:\, i\in\tau \text{ et } x_i^j<c} (y_i - \bar{y}_1)^2 + \sum_{i:\, i\in\tau \text{ et } x_i^j\geq c} (y_i - \bar{y}_2)^2 \tag{3.4}$$

où $\bar{y}_k$ $(k = 1, 2)$ est la moyenne des y_i telles que $i \in \tau$ et x_i^j est (respectivement) inférieur ou supérieur à c. On itère ensuite le processus jusqu'à atteindre une règle d'arrêt. Une règle d'arrêt simple est, par exemple, de ne pas découper des nœuds qui contiennent moins qu'un certain nombre d'observations.

La règle de prédiction associée à une valeur $x \in \mathbb{R}^d$ donnée est réalisée de la manière suivante : en parcourant l'arbre, x est associé à un nœud terminal τ (appelé aussi feuille) pour lequel la valeur prédite $\hat{y}$ est la valeur moyenne des observations qu'il contient $\frac{1}{\tau}\sum_{i\in\tau} y_i$. Les étapes de la construction des forêts aléatoires sont les suivantes :

- On génère T échantillons bootstrap : un échantillon bootstrap étant le tirage aléatoire de n observations avec remise dans l'échantillon d'apprentissage.
- On construit sur chaque échantillon un arbre de régression maximum par la procédure CART : pour ce faire, on tire aléatoirement m variables parmi les d et on cherche la meilleure coupure suivant ces m variables. Le tirage à chaque nœud des m variables se fait sans remise et uniformément parmi toutes les variables. Chaque variable a la probabilité $\frac{m}{d}$ d'être sélectionnée.
- La collection d'arbres est enfin agrégée pour donner le prédicteur : la prédiction associée à une valeur $x \in \mathbb{R}^d$ est obtenue en faisant la moyenne empirique des prédictions de l'ensemble des arbres pour x : soit $\hat{y} = \frac{1}{T}\sum_{l=1}^{T} \hat{y}_l(x)$, où $\hat{y}_l(x)$ est la prédiction obtenue pour x avec l'arbre l.

La valeur de m est fixée par l'utilisateur, Breiman suggère $m = \sqrt{d}$ en classification et $m = \frac{d}{3}$ en régression.

Outre la prédiction, les forêts aléatoires peuvent fournir des informations pertinentes par le calcul et la représentation graphique d'indices permettant de quantifier le pouvoir prédicteur d'une variable donnée que l'on appelle "importance" d'une variable. Avant de présenter le mode de calcul de l'importance des variables, nous allons définir la notion d'erreur "out of bag", qui sera utile pour la suite.

Erreur Out-Of-Bag (OOB) Pour une observation (x_i, y_i), on note $\mathcal{T}_i$ l'ensemble des arbres dans $\{1, \ldots, T\}$ pour lesquels l'observation i faisait partie de l'échantillon bootstrap d'apprentissage. On peut alors définir la prédiction OOB de i comme la moyenne des prédictions faites par les arbres n'étant pas dans $\mathcal{T}_i$: $\hat{y}_i^{\text{OOB}} = \frac{1}{T-|\mathcal{T}_i|} \sum_{l \notin \mathcal{T}_i} \hat{y}_l(x_i)$. L'erreur OOB est l'erreur moyenne (au sens de l'erreur quadratique) de l'erreur commise sur la prédiction OOB de l'ensemble des observations : $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{\text{OOB}} - y_i)^2$.

Importance d'une variable L'importance de la variable X^j correspond à la perte de qualité de prédiction induite par une permutation des valeurs observées de cette variable. De manière plus précise, pour un arbre donné $l \in \{1, \ldots, T\}$, l'ensemble des indices des observations de l'échantillon bootstrap associé à cet arbre est noté B_l. On note alors errOOB_l, l'erreur quadratique moyenne commise par l'arbre l sur les observations $\{i \notin B_l\}$ et $\widetilde{\text{errOOB}}_l^{\,j}$ l'erreur quadratique moyenne commise par l'arbre l sur les observations $\{i \notin B_l\}$ lorsque l'on permute aléatoirement les valeurs de la j-ème variable de ces observations. L'importance de la variable X^j est alors définie par

$$\text{Imp}(X^j) = \frac{1}{T} \sum_{l=1}^{T} (\widetilde{\text{errOOB}}_l^{\,j} - \text{errOOB}_l). \tag{3.5}$$

Dans la sortie de la fonction `randomForest` (package randomForest de **R**), cet indicateur est noté `%IncMSE`.

On peut aussi mesurer l'importance d'une variable à partir de la décroissance de l'hétérogénéité définie à partir du critère de Gini (`%IncNodePurity`). L'importance d'une variable est alors une somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un nœud.

Importance groupée Définie par Gregorutti [5], l'importance d'un groupe de variables est basée sur le même principe que l'importance par permutation définie précédemment [4]. Ici, les variables explicatives sont structurées en groupes de variables et la même permutation est utilisée pour l'ensemble des variables du groupe.
Supposons que $J = \{j_1, \ldots, j_q\}$ est un ensemble de q indices dans $\{1, \ldots, d\}$ et $\mathbf{X}^J = (\mathbf{X}^{j_1}, \ldots, \mathbf{X}^{j_q})$ le sous-vecteur de $\mathbf{X}$ associé. L'importance du groupe $\mathbf{X}^J$ est donnée par :

$$\mathrm{Imp}(X^J) = \frac{1}{T}\sum_{l=1}^{T}(\widetilde{\mathrm{errOOB}}_l^{\,J} - \mathrm{errOOB}_l). \tag{3.6}$$

A la différence de la formule précédente (3.5), $\widetilde{\mathrm{errOOB}}_l^{\,J}$ est l'erreur OOB commise par l'arbre l sur les observations $\{i \notin B_l\}$ lorsque l'on permute aléatoirement les valeurs des variables formant le groupe J. L'importance groupée peut être calculée à partir de la fonction `varImpGroup` (package RFgroove de **R**). Notons que la constitution des groupes est guidée par la connaissance que l'on a des variables. Par exemple, il peut être pertinent de regrouper des variables corrélées ou des variables ayant des caractéristiques communes.

Présentation des résultats

Ce chapitre décrit les travaux réalisés et les différents résultats obtenus.
L'objectif de l'étude était d'identifier les variables climatiques et phénotypiques qui rendent plus sensible le rendement du tournesol et les intervalles de temps les plus influents. Pour ce faire, des méta-variables ont été créées. Celles-ci correspondent à des résumés (moyenne, écart-type, maximum) des relevés climatiques journaliers pour une période de temps d'une ou plusieurs semaines. La période de culture du tournesol (avril-septembre, soit 6 mois) a été subdivisée en intervalles de temps réguliers (un mois, deux semaines, une semaine). La moyenne de chaque variable climatique a été calculée sur chacun des intervalles de temps. A cela, s'ajoutent l'écart-type de la variable "Pluie" (pour tenir compte de la variabilité des précipitations), la valeur maximale de la variable "Température maximale" (pour un éventuel effet de seuil) et les 8 traits phénotypiques. Des méta-modèles ont été ensuite construits sur ces méta-variables.
Les résultats sont structurés en deux grandes parties. La première présente les résultats du modèle linéaire, de l'analyse de sensibilité et des forêts aléatoires sur l'ensemble des données (190000 observations). La deuxième partie est une analyse plus approfondie sur un sous-ensemble de données (2500 observations). Dans cette partie, j'ai repris le modèle de forêts aléatoires (qui s'est avéré le plus performant lors de la première étude) sur un jeu de données réduit dans le but de mettre en place une méthodologie permettant d'utiliser un nombre réduit de variables d'entrée pour construire un méta-modèle pertinent.

4.1 Les résultats obtenus sur l'ensemble des données

Les données utilisées dans cette section sont constituées des 190 relevés climatiques croisés par les 1000 variétés de tournesol avec une décomposition en semaines, soit une jeu de données de 190000 observations et 186 variables.

4.1.1 Régression linéaire et analyse de sensibilité

Une régression linéaire a été effectuée sur chaque jeu de données (mensuel, hebdomadaire, bihebdomadaire), suivie d'une analyse de sensibilité. Nous allons présenter dans cette section les différents résultats obtenus.

✺ **Résultats du modèle linéaire**

La variable que nous cherchons à expliquer est le rendement du tournesol. Pour cela, nous utilisons les 8 traits phénotypiques et les relevés climatiques résumés sous forme de méta-variables sur des intervalles de temps réguliers. Par exemple, pour l'approche par semaine, nous avons un total de 186 variables explicatives. Le tableau 4.1 présente les R^2 obtenus avec les différents modèles.

Modèles	**Nombre de variables**	R^2 **en %**
Analyse mensuelle	50	17.76
Analyse bihebdomadaire	99	42.35
Analyse hebdomadaire	186	**88.15**

Tableau 4.1 – Evolution du R^2 en fonction du nombre de subdivisions

On observe qu'il faut aller jusqu'à une décomposition hebdomadaire de la période de culture du tournesol pour avoir un modèle dont la qualité de prédiction est suffisamment bonne. Ce modèle explique plus de 88% de la variabilité du rendement du tournesol. Les coefficients associés à toutes les variables sont significatifs au seuil de 5%, sauf pour 2 variables : la température minimale de la semaine 5 (Tmin.5) et celle de la semaine 19 (Tmin.19). L'objectif étant d'identifier les variables et les semaines rendant sensibles la variabilité du rendement du tournesol, des indices de sensibilité ont été calculés sur ce modèle linéaire.

✺ **Résultats de l'analyse de sensibilité**

Les indices PCC décrits précédemment (équation 3.3) ont été calculés grâce au package "sensitivity" de **R**. Les valeurs absolues de ces indices ont été ensuite représentées

graphiquement pour en faciliter l'interprétation. La figure 4.1 donne les variables qui ont un indice supérieur ou égal à 0.22.

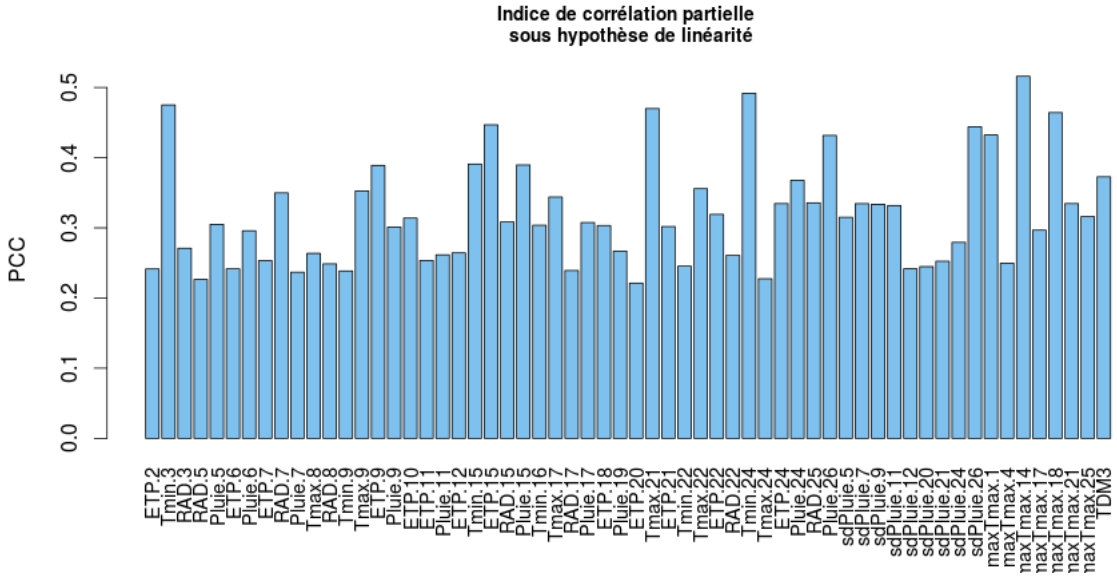


FIGURE 4.1 – Indices PCC

Données simulées à partir de SUNFLO

La distribution des indices semble être uniforme, nous ne pouvons pas encore dégager des motifs climatiques qui influencent le rendement du tournesol. Nous avons ensuite fait un regroupement par variable et par semaine en sommant les indices, ce qui n'est pas justifié mathématiquement mais permettait de faciliter la lecture dans la figure 4.2.

Il ressort de ces deux graphiques que l'évapotranspiration (ETP), la température maximale (Tmax), la valeur maximale de la température maximale (maxTmax) et les précipitations (Pluie) semblent être les variables les plus pertinentes. En ce qui concerne les semaines, les semaines 9, 15 et 24 (i.e. début juin, 3ième semaine de juillet et fin septembre) semblent avoir une forte influence.

Le nombre de variables dans le modèle linéaire étant élevé, j'ai procédé à une élimination itérative de variables tout en observant la conséquence sur le comportement du R^2 ajusté. Pour ce faire, 2 approches ont été comparées. La première part des indices calculés avec le modèle initial et élimine à chaque étape la variable ayant l'indice PCC le plus faible. Après chaque élimination, le modèle linéaire est estimé sur le nouveau jeu de

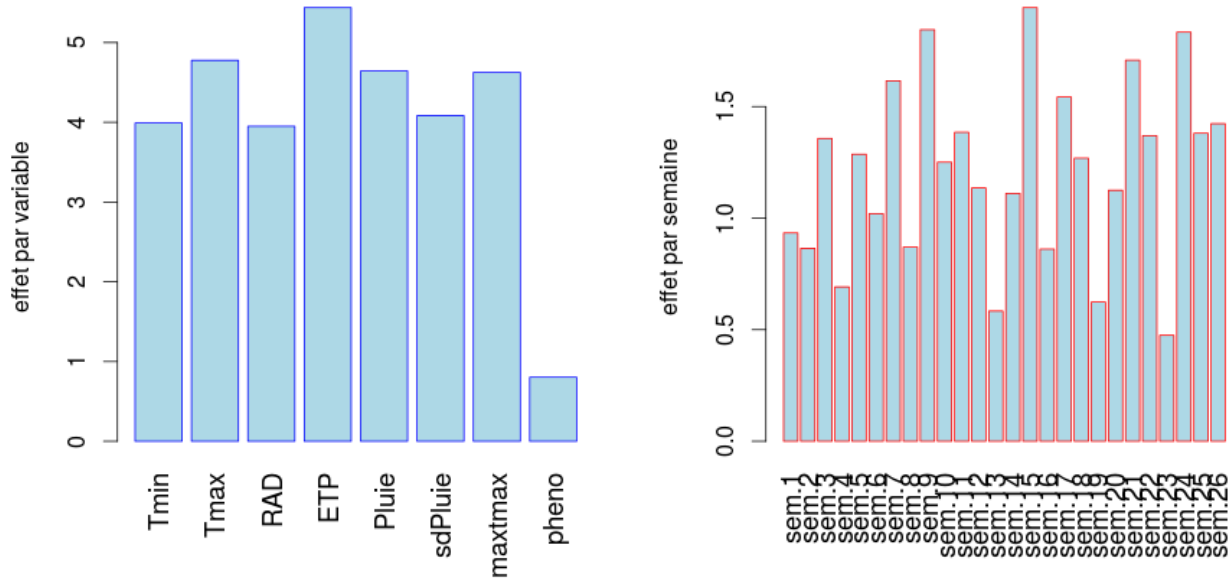


FIGURE 4.2 – Indices PCC cumulés par semaine et par variable

Données simulées à partir de SUNFLO

données et le R^2 est recalculé. La seconde approche utilise un principe similaire mais dans celle-ci les indices sont recalculés à chaque étape et l'élimination à l'étape $t+1$ est faite à partir des indices PCC calculés à l'étape t. L'évolution du R^2, lorsqu'on élimine jusqu'à 60 variables sur 186 est illustrée, sur la figure 4.3. On remarque que le R^2 se dégrade moins vite dans le cas où l'indice est recalculé à chaque étape que dans le cas où les indices du modèle initial sont utilisés. Dans le premier cas, le R^2 se dégrade de 0.05 point lorsqu'on passe de 186 à 126 variables, alors que dans le second il se dégrade de 0.25 point. Bien que toutes les variables soient significatives, il semble donc être possible de réduire leur nombre en conservant la capacité de prédiction du modèle.

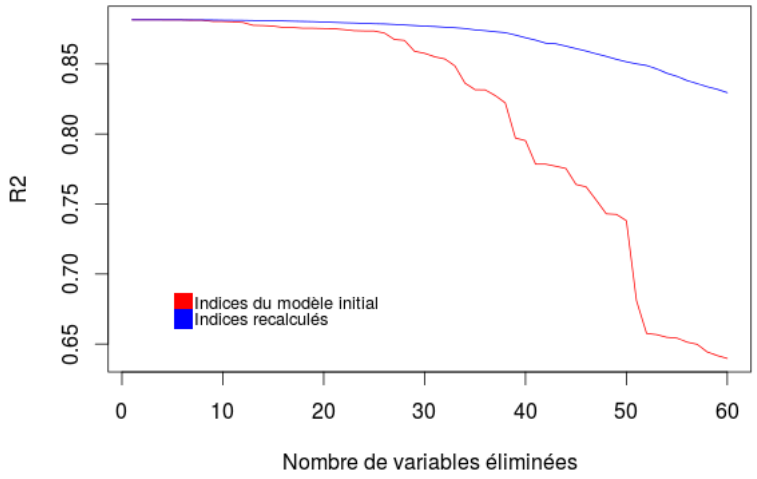


FIGURE 4.3 – Evolution du R^2 en fonction du nombre de variables éliminées

Données simulées à partir de SUNFLO

✺ **Calcul d'importance de variables sur modèle linéaire**

Afin de rendre comparables les résultats obtenus avec le modèle linéaire et ceux obtenus avec les forêts aléatoires, j'ai calculé des importances de variables en m'inspirant du principe de permutation utilisé dans les forêts aléatoires. La différence dans le cas du modèle linéaire est qu'on n'utilise pas de bootstrap mais l'estimation de l'accroissement de l'erreur quadratique moyenne sur les données d'apprentissage directement. Pour calculer l'importance d'une variable X^j dans le cas du modèle linéaire :

- On estime le modèle linéaire et on calcule l'erreur MSE.
- On permute ensuite aléatoirement les valeurs de cette variable et on réestime le modèle linéaire.
- L'importance de la variable est alors mesurée par la croissance de l'erreur MSE entre le modèle perturbé et le modèle initial.

La figure 4.4 présente les variables qui ont un indice supérieur à 0.15. Les résultats sont cohérents avec les indices PCC calculés plus haut (figure 4.1) : on a les mêmes pics sauf qu'ici il y a moins de variables influentes. Les variables les plus influentes selon ces deux approches sont : la température minimale des semaines 3 et 24, la température maximale de la semaine 21, la valeur maximale de la température maximale de la semaine 14.

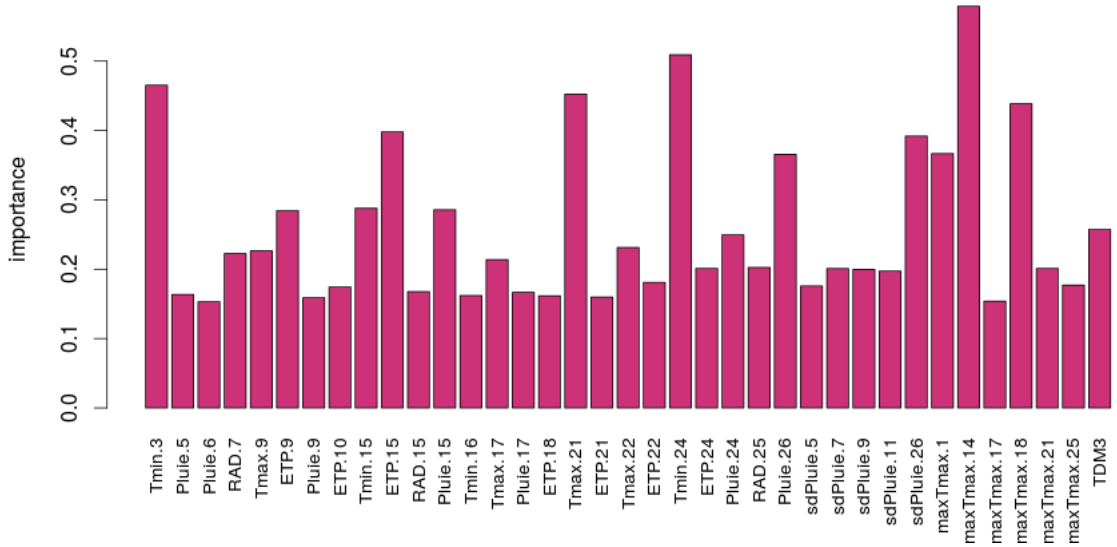


FIGURE 4.4 – Importance de variables sur modèle linéaire

Données simulées à partir de SUNFLO

Conclusion partielle : on retient des résultats du modèle linéaire et de l'analyse de sensibilité qu'une décomposition temporelle fine (semaine) est nécessaire pour expliquer le rendement du tournesol à partir des variables climatiques. Les résultats indiquent que toutes les variables sont significatives et importantes, il ne serait donc pas facile d'utiliser ces résultats pour chercher les motifs climatiques influents. Cela pourrait être expliqué par la forte corrélation entre les variables et l'hypothèse de linéarité qui peut ne pas être vérifiée.

4.1.2 Résultats des forêts aléatoires

✺ Calcul d'importances individuelles

Les forêts aléatoires constituent le second méta-modèle utilisé. Ici, aucune hypothèse n'est faite a priori sur la relation entre les entrées et la sortie. Pour la mise en œuvre sur **R**, nous avons utilisé le package "randomForest". L'objectif est d'identifier les entrées qui contribuent le plus à la qualité de la prédiction. Pour cela, nous avons calculé l'importance

de chaque entrée (voir la formule (3.5)). La méthode étant aléatoire, nous avons entraîné 10 forêts différentes pour étudier la stabilité du modèle et des variables qui ressortaient comme importantes. A la suite de ces 10 constructions du modèle, les mêmes variables se sont révélées importantes, on conclut alors que le modèle est stable. Les importances obtenues sur un des dix modèles sont ensuite représentées graphiquement. La figure 4.5 indique que la pluviométrie de la semaine 11 (fin juin) et son écart-type, l'ensoleillement de la semaine 23 (fin septembre) et la température minimale de la semaine 20 (fin août) semblent être les variables les plus influentes. Ces mêmes variables sont obtenues quelque soit la forêt. On remarque que les importances de variables calculées sur les forêts aléatoires et celles calculées sur le modèle linéaire (figure 4.4) n'aboutissent pas à la même conclusion.

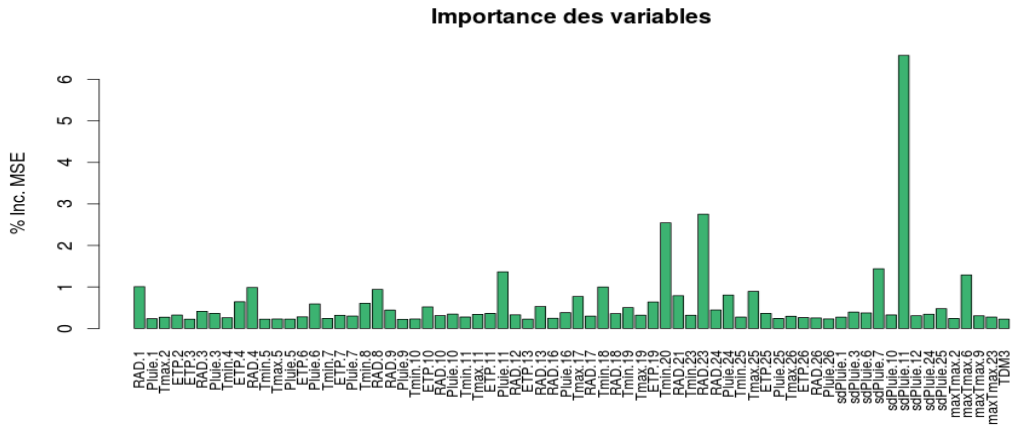


FIGURE 4.5 – Importances des variables calculées à partir des forêts aléatoires (valeur > 0.22)

Données simulées à partir de SUNFLO

✺ Calcul d'importances groupées

Des importances groupées ont été calculées sur deux types de groupes. Les premiers groupes ont été formés en faisant un regroupement par type de variable et le second en faisant un regroupement par semaine. Huit groupes ont été donc formés pour la première approche dont 7 constitués des différents résumés sur les 26 semaines et le dernier constitué des traits phénotypiques. Pour la seconde approche, 26 groupes correspondant aux 26 semaines ont été formés. Cette dernière approche permettra d'identifier la semaine la plus influente. Les importances obtenues pour ces différents groupes sont représentées sur la figure 4.6. On remarque que lorsqu'on considère les 26 semaines, les variables ensoleillement

(RAD) et écart-type de la pluie (sdpluie) ont les importances les plus élevées. On note également que tous les traits phénotypiques regroupés n'ont qu'une faible influence sur le rendement. Les regroupements en semaines indiquent que les semaines les plus influentes sont les semaines 10 (mi-juin), 6 (mi-mai) et 23 (3ième semaine de septembre). On observe aussi une forte stabilité des ces résultats sur plusieurs constructions du modèle. Ces regroupements ne peuvent pas être comparés à ceux présentés sur la figure 4.2, les calculs n'étant pas fondés sur le même principe.

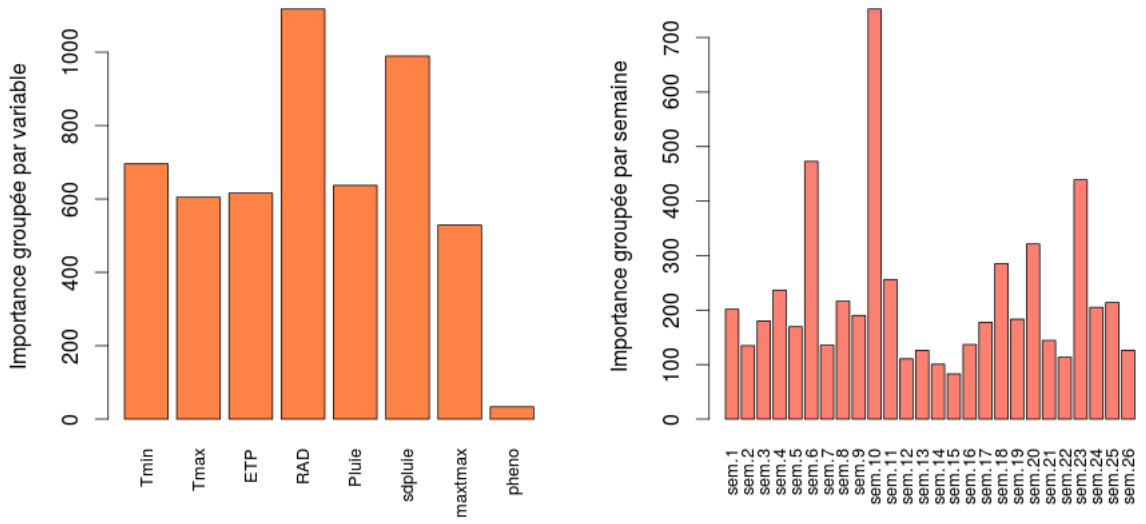


FIGURE 4.6 – Importances groupées par variable et par semaine

Données simulées à partir de SUNFLO

Conclusion partielle : les 2 méta-modèles utilisés dans cette première partie sur l'ensemble des données n'aboutissent pas à la même conclusion. Le modèle linéaire montre une influence plus ou moins forte des variables sur presque toutes les semaines, alors que le modèle de forêts aléatoires distingue les semaines 6 (mi-mai), 10 (mi-juin) et 23 (3ième semaine de septembre). Cette différence entre les résultats n'est pas un résultat surprenant, puisque le premier méta-modèle est linéaire et le second non linéaire. Par ailleurs, le modèle de forêts aléatoires a une qualité de prédiction meilleure à celle du modèle linéaire, comme le montre la figure 4.7. Lorsqu'on apprend ces 2 méta-modèles sur 80%

des données et qu'on les teste sur les 20% restantes, on obtient une MSE de 1.58 sur le modèle linéaire et de 0.45 sur les forêts aléatoires. Pour la suite, nous allons nous focaliser sur les forêts aléatoires en utilisant un jeu de données réduit.

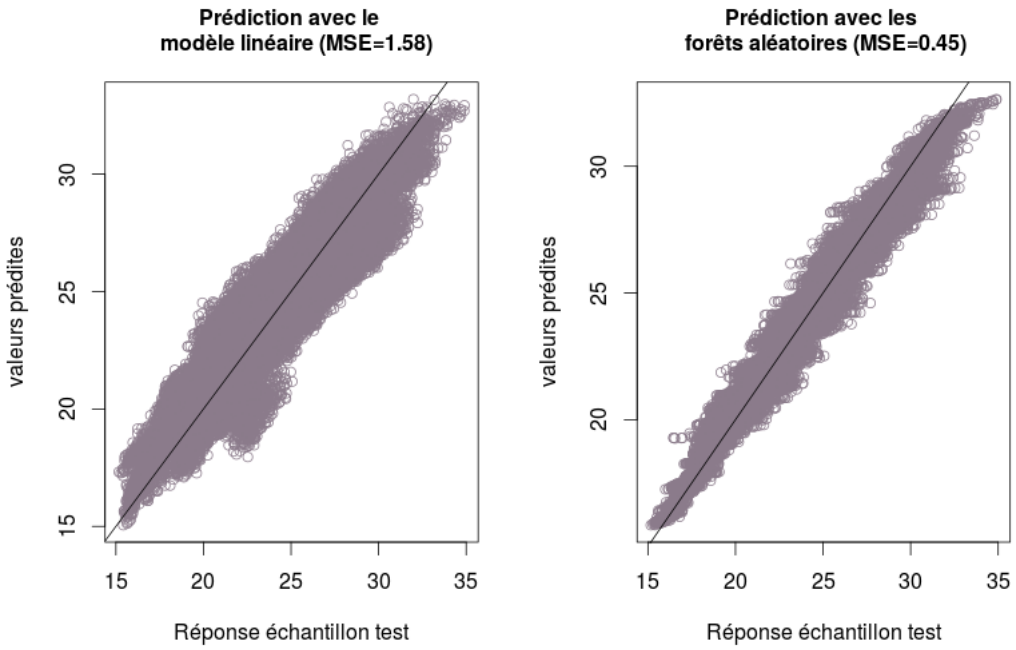


FIGURE 4.7 – Comparaison de modèles

Données simulées à partir de SUNFLO

4.2 Analyse approfondie sur un sous-ensemble de données

Dans cette section, il est question de faire une analyse plus approfondie en utilisant un jeu de données réduit, du fait du temps de calcul des algorithmes utilisés. Ce nouveau jeu de données est obtenu à partir de 50 relevés climatiques. Ces relevés correspondent à ceux des 10 dernières années (2003-2012) de la période d'étude (1975-2012) sur chacune des 5 stations. 50 phénotypes ont été tirés de manière aléatoire parmi les 1000 phénotypes. Le croisement des 50 climats par les 50 phénotypes donne un jeu de données de 2500

observations avec 186 méta-variables hebdomadaires.

4.2.1 Importance des variables sur les forêts aléatoires

Le calcul des importances des variables sur le jeu de données réduit offre des contrastes significatifs. On voit (figure 4.8) qu'au début de la période de culture (les 5 premières semaines), l'influence des variables climatiques n'est pas significative. On observe ensuite une forte influence sur les semaines 6 et 8 (les 2 dernières semaines du mois de mai) de l'ensoleillement et de l'évapotranspiration. On observe un creux sur la période juin à mi-août (semaines 9 à 19, excepté l'influence de Tmin.15). On peut émettre l'hypothèse qu'il n'y a pas une grande différence entre les différentes séries climatiques sur cette période. Les variables climatiques deviennent influentes vers la fin de la période de culture (semaine 20, 23, 24 et 25). Les variables les plus influentes, lorsqu'on considère les relevés climatiques des dernières années sont : l'évapotranspiration de la semaine 6 (ETP.6), l'ensoleillement (RAD) des semaines 6, 8, 20, 24 et 25 et la température minimale de la semaine 15 (Tmin.15). Ce modèle, tout comme celui obtenu sur l'ensemble des données est stable sur plusieurs itérations.

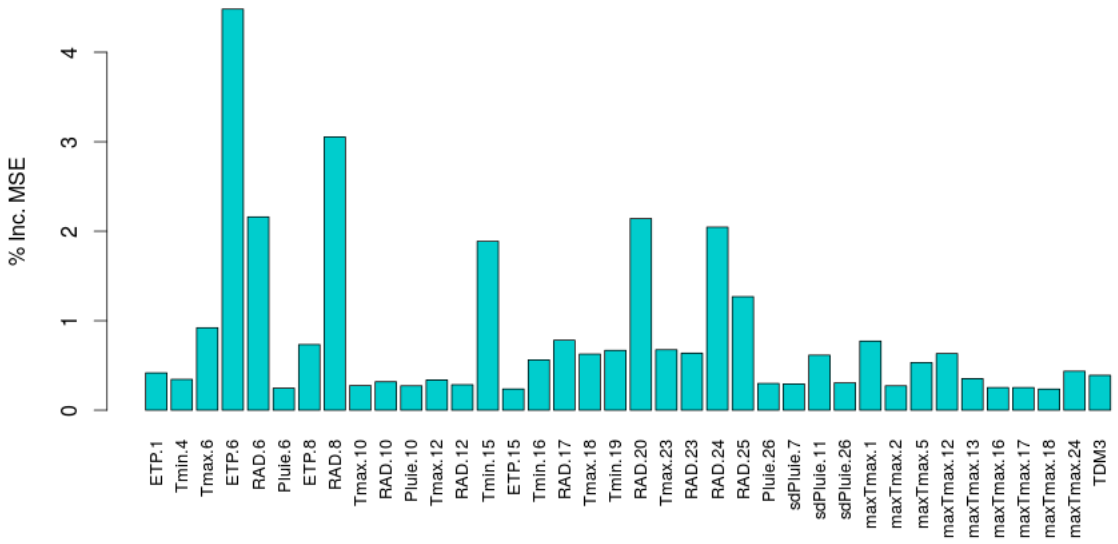


FIGURE 4.8 – Importances des variables sur un sous-ensemble de données (valeur > 0.22)

Données simulées à partir de SUNFLO

4.2.2 Forêts aléatoires et sélection de variables

On retient de tout ce qui précède qu'avec la décomposition en semaines, on aboutit à 186 variables explicatives parmi lesquelles, seulement quelques unes sont influentes. Il serait donc intéressant d'explorer d'autres approches pour isoler ces variables. J'ai utilisé à cet effet une approche de sélection de variables basée sur les forêts aléatoires pour identifier les variables les plus influentes. Pour cela, j'ai utilisé la fonction `VSURF` (package VSURF de **R**) [6]. Ce package fait une sélection de variables en se basant sur les importances de variables. Le principe consiste à faire un classement préliminaire des variables à partir des indices d'importance des forêts aléatoires. Il utilise ensuite un algorithme d'introduction ascendante pas à pas et fournit deux sous-ensembles de variables. Le premier donne les variables importantes pour l'interprétation et le second est un sous-ensemble parcimonieux à l'aide duquel on peut faire de bonnes prédictions. Le principe se résume en deux étapes :

☞ Étape 1 : élimination préliminaire
- l'algorithme calcule les importances des variables et les ordonne par ordre décroissant d'importance,
- il supprime toutes les variables ayant une faible importance, soit p ($p < d$) le nombre de variables restant,

☞ Étape 2 : sélection de variables
- Pour l'interprétation : il construit ensuite une collection de forêts imbriquées sur les k premières variables, $k = 1, \ldots, d$ et sélectionne les variables qui donnent les plus petites erreurs OOB.
- Pour la prédiction : à partir des variables ordonnées pour l'interprétation, il construit une séquence de forêts aléatoires par une introduction pas à pas (ascendante) des variables. Les variables du dernier modèle sont alors retenues.

Les résultats obtenus à partir du package VSURF se résument dans le tableau 4.2.

Ce tableau fournit les variables retenues avec la procédure de sélection. A la première étape, celle de l'élimination préliminaire, toutes les 186 variables ont été retenues. Seulement 3 variables ont été éliminées parmi les 186 à la phase d'interprétation, ce qui confirme

Elimination préliminaire	**Interprétation**	**Prédiction**
186 variables retenues	183 variables retenues	7 variables
	variables éliminées ETP.14, LLH, TLN	ETP.6, RAD.6, RAD.8, sd-pluie.7, TDM3, TDF1, TR

Tableau 4.2 – Forêts aléatoires et sélection de variables

l'importance de toutes las variables obtenue avec le modèle linéaire. Par contre lorsqu'on passe à l'étape de prédiction, on ne retient que 7 variables, cela signifie que les variables ont une influence sur le rendement du tournesol mais contribuent peu à la qualité de sa prédiction.

Les différentes étapes de la procédure de sélection sont résumées sur la figure 4.9. La phase d'élimination est représentée par les deux graphiques du haut. Le premier graphique (à gauche) donne l'importance moyenne des variables par ordre décroissant ; le seuil d'importance à partir duquel une variable est éliminée est fixée à 0 (ligne rouge du graphique). On remarque qu'aucune variable n'a une importance en dessous de 0, ce qui justifie le choix de toutes les variables à la première étape. Les écarts-types des importances sont représentés sur le second graphique de haut (droite) dans l'ordre décroissant des importances (courbe en noir). La courbe en vert est une estimation de ces écarts-types obtenue à partir d'un arbre CART.

Le graphique en bas à gauche correspond à l'étape d'interprétation. Il donne l'erreur OOB moyenne des forêts aléatoires imbriquées (de celle avec une seule variable comme prédicteur, à celle avec toutes les variables conservées après la première étape). La ligne rouge verticale indique le modèle retenu. Enfin, le dernier graphique, représentant la phase de prédiction, montre la décroissance de l'erreur OOB des différentes forêts imbriquées à la différence que les variables sont introduites de manière pas-à-pas (ascendante) et 7 variables sont retenues à cette dernière étape. On obtient la même erreur OOB sur ces 7 variables que sur les 183 variables. Les importances de ces 7 variables qui contribuent le plus à la qualité de la prédiction sont représentées sur la figure 4.10.

On remarque que les variables climatiques qui contribuent le plus à la qualité de la prédiction sont l'ensoleillement et l'évapotranspiration et la période la plus sensible est le mois de mai (semaines 6 à 8). En ce qui concerne les traits phénotypiques on a : la durée de la phase levée-floraison (TDF1), la durée de la phase levée-maturité (TDM3) et le seuil

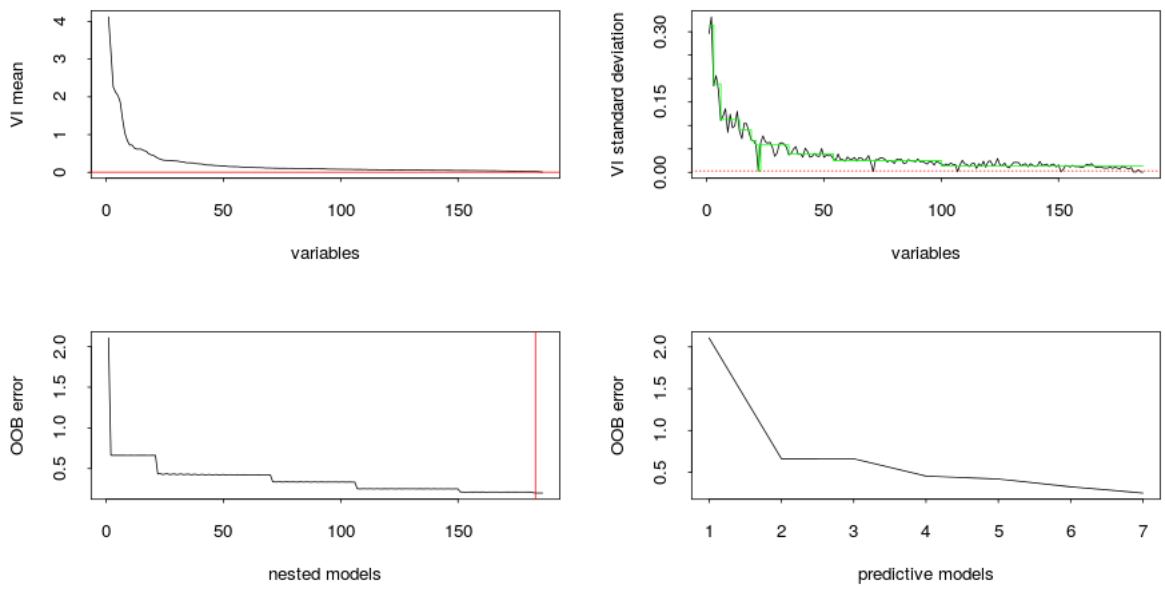


FIGURE 4.9 – Forêts aléatoires et sélection de variables

Données simulées à partir de SUNFLO

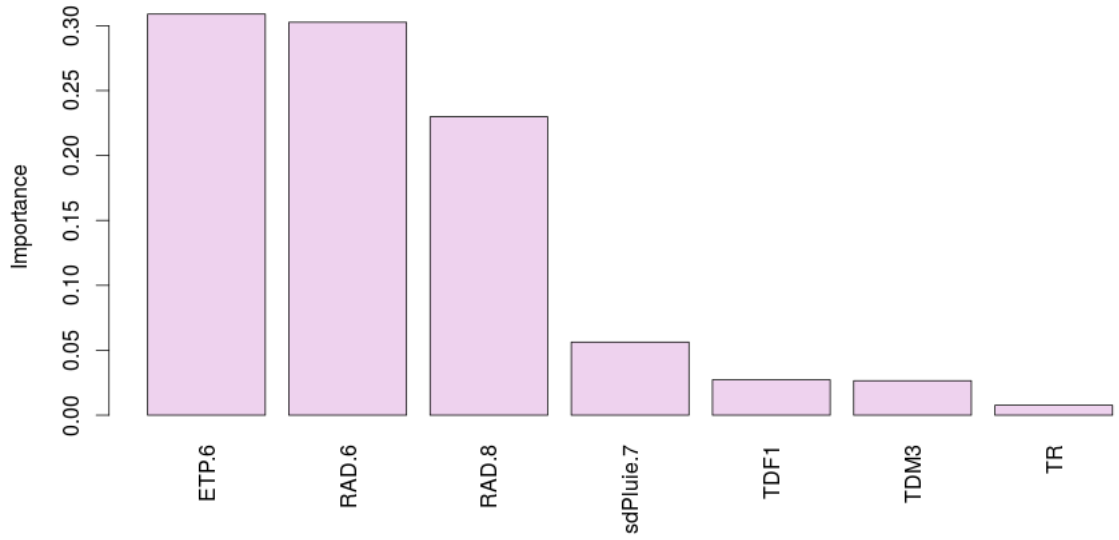


FIGURE 4.10 – Variables retenues pour la prédiction

Données simulées à partir de SUNFLO

de réponse de la conductance stomatique à une contrainte hydrique (TR) qui contribuent aussi à la qualité de la prédiction, mais gardent une importance faible sur le rendement. Nous avons étudié la stabilité de ce modèle en entraînant plusieurs itération, on remarque

les 3 variables climatiques les plus influentes et les 3 traits phénotypiques ne changent pas quelque soit l'itération. On conclut que le modèle est stable.

4.2.3 Fusion de variables

Dans cette partie, j'ai utilisé une nouvelle approche basée toujours sur les forêts aléatoire pour réduire le nombre de variables. Elle consiste à fusionner de manière itérative les variables les moins influentes sur des semaines consécutives. L'algorithme se présente comme suit :

❶ On forme des groupes contigus de variables, chaque groupe contenant 2 variables de la même série, du même résumé et sur des intervalles de temps consécutifs (par exemple, regrouper Tmin.1 et Tmin.2).

❷ On calcule ensuite des importances groupées pour l'ensemble des paires de variables formées.

❸ On identifie le groupe de variables le moins influent (importance groupée la plus faible) et on fusionne les deux variables formant ce groupe. On fait la moyenne s'il s'agit des moyennes, on prend le maximum si c'est un maximum et on recalcule l'écart type si c'est un écart-type.

❹ On estime de nouveau un modèle de forêt aléatoire sur ce nouveau jeu de données et on calcule l'erreur OOB commise.

❺ On répète le processus jusqu'à dégradation considérable de l'erreur OOB. Il est à noter que l'algorithme autorise des fusions sur plus de deux semaines.

La figure 4.11 montre comment se comporte l'erreur OOB lorsqu'on fusionne jusqu'à 130 variables, c'est-à-dire lorsqu'on passe de 186 variables à 56 variables. On observe une très faible augmentation de l'erreur lorsqu'on passe de 186 à 56 variables. Cela signifie qu'on peut réduire considérablement le nombre de variables en regroupant les variables les moins influentes sans dégrader la qualité de la prédiction.

Après la fusion, l'importance de chaque série climatique a été représentée sur chaque semaine, comme le montrent les graphiques de la figure 4.12. L'influence de la température minimale s'observe à partir de la semaine 15 jusqu'à la semaine 19, ce qui correspond à la période mi-juillet à août. La température maximale quant à elle, a une forte influence

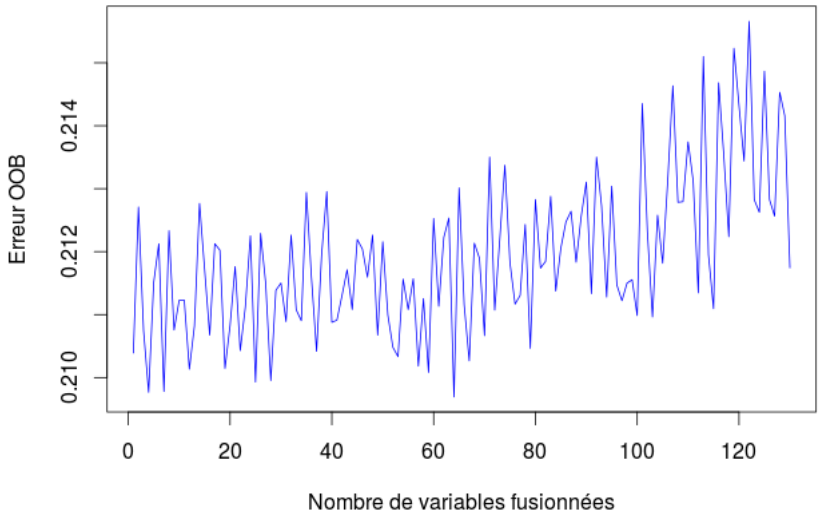


FIGURE 4.11 – Evolution de l'erreur OOB en fonction du nombre de variables fusionnées

Données simulées à partir de SUNFLO

durant les semaines 6 (mi-mai) et 23 (fin-septembre) et une influence moyenne sur la période juin-août. L'influence de l'évapotranspiration sur le rendement ne s'observe que sur la semaine 6. En ce qui concerne l'effet de l'ensoleillement, 3 phases sont à distinguer. La première fin mai (semaine 6 et 8) est marquée par une forte influence, la deuxième allant de fin juin à fin juillet (semaines 12-16), est marquée par une influence moyenne et la dernière (fin août et fin septembre) est aussi marquée par une forte influence. Un fait surprenant est la faible influence de la pluviométrie sur toutes les semaines.

Conclusion partielle : les résultats de la procédure de fusion de variables confirment ceux de la procédure de sélection de variables. A la phase d'interprétation de la procédure de sélection, 183 variables ont été sélectionnées sur 186, ce qui confirme également la significativité de toutes les variables obtenue avec le modèle linéaire. Les variables climatiques RAD.6, RAD.8, RAD.20 et ETP.6 ont été retenues comme les plus influentes que ce soit par la procédure de sélection que par la procédure de fusion.

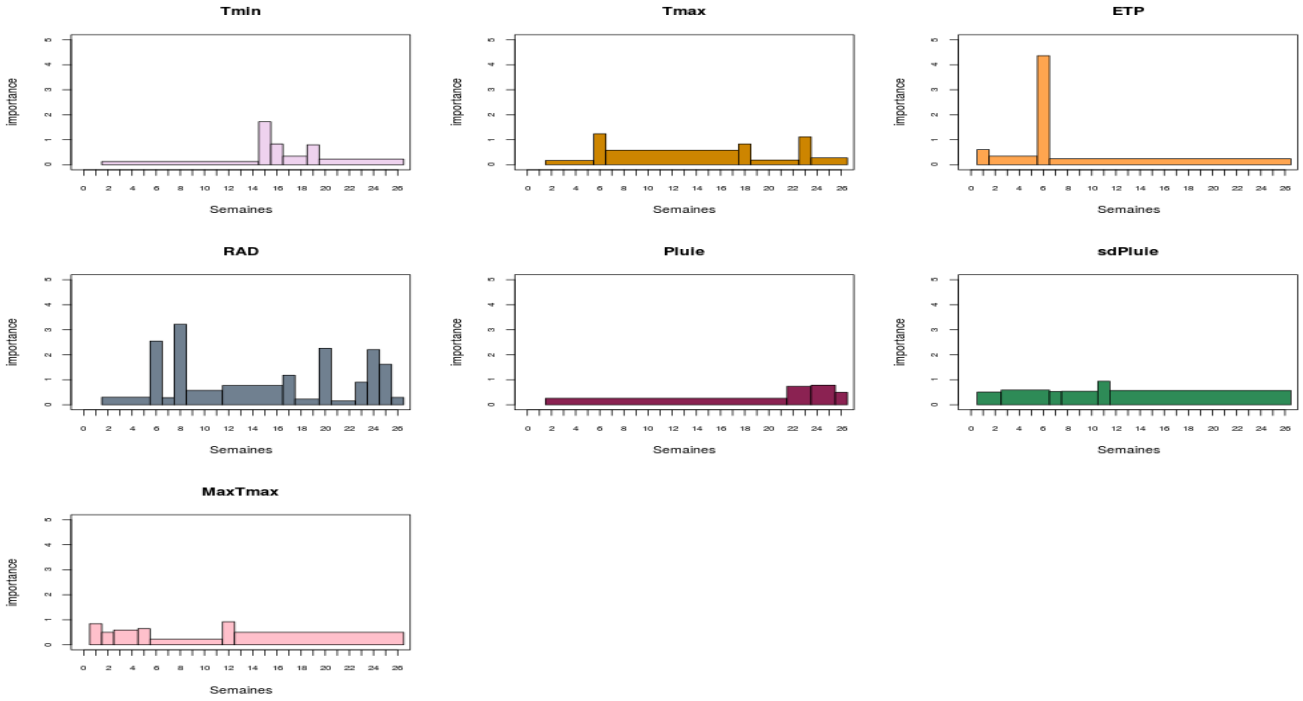


FIGURE 4.12 – Importance par variable sur les 26 semaines

Données simulées à partir de SUNFLO

4.2.4 Prédiction

7 variables ont été retenues pour la prédiction par la procédure de sélection de variables et la procédure de fusion aboutit à 56 variables. Nous allons dans cette partie comparer le pouvoir prédictif de ces deux modèles (7 variables et 56 variables) à celui de toutes les variables sur trois types d'échantillonnage (tableau 4.3). Sur le premier type d'échantillonnage, l'apprentissage est fait sur 80% de l'ensemble des données et le reste a servi d'échantillon test. Dans le deuxième type d'échantillonnage, celui des climats, l'échantillon d'apprentissage a été constitué en choisissant de manière aléatoire 80% des climats et tous les phénotypes, puis le reste des climats avec tous phénotypes pour l'échantillon test. Le même principe a été utilisé pour l'échantillonnage des phénotypes. Les prédictions ont été faites avec le modèle forêt aléatoire. Les valeurs prédites ont été représentées en fonction des vraies valeurs et les erreurs quadratiques moyennes (MSE) ont été calculées (tableau 4.4). La figure 4.13 présente les résultats.

Les 3 premiers graphiques en ligne correspondent à la prédiction obtenue lorsque toutes

Echantillonnge global	**Echantillonnage des climats**	**Echantillonnage des phénotypes**
Apprentissage : 80% de l'ensemble	Apprentissage : 80% des climats et tous les phénotypes	Apprentissage : 80% des phénotypes et tous les climats
Test : 20%	Test : 20% des climats et tous les phénotypes	Test : 20% des phénotypes et tous les climats

Tableau 4.3 – Description des échantillons

les variables sont utilisées comme prédicteurs sur les trois types d'échantillonnage. Les 3 graphiques du milieu donnent la prédiction lorsqu'on utilise les 7 variables retenues dans la procédure de sélection et les 3 derniers graphiques correspondent à la procédure de fusion. Lorsqu'on compare les types d'échantillonnage, on remarque que dans les 3 cas (7, 56 ou 186 variables), la meilleure prédiction est obtenue lorsqu'on effectue la prédiction pour de nouveaux phénotypes et la plus mauvaise est obtenue lorsqu'on effectue la prédiction pour de nouveaux climats. On peut donc dire qu'il n'y a pas une grande différence entre les phénotypes contrairement aux climats, ce qui pourrait justifier la faible influence des phénotypes sur le rendement. En ce qui concerne la comparaison des modèles ayant servi à la prédiction, on note que les erreurs les plus faibles sont commises sur le modèle à 186 variables et que les erreurs obtenues sur ce modèle sont sensiblement égales à celles obtenues sur le modèle de fusion. Ce résultat est cohérent puisque dans la procédure de fusion toutes les variables ont été utilisées, sauf qu'elles ont été "compressées". on observe une augmentation de l'erreur lorsqu'on passe à 7 variables, mais cette augmentation est raisonnable (une augmentation de 0.06, 0.05 et 0.02 points respectivement sur les 3 échantillons) par rapport au gain de parcimonie du modèle. On conclut alors que les 7 variables peuvent prédire autant que les 186.

Modèles	**Echantillonnage global**	**Echantillonnage des climats**	**Echantillonnage des phénotypes**
Modèle initial (186 variables)	$MSE = 0.18, OOB = 0.20$	$MSE = 0.21, OOB = 0.20$	$MSE = 0.17, OOB = 0.21$
Sélection (7 variables)	$MSE = 0.24, OOB = 0.27$	$MSE = 0.26, OOB = 0.26$	$MSE = 0.19, OOB = 0.28$
Fusion (56 variables)	$MSE = 0.18, OOB = 0.21$	$MSE = 0.22, OOB = 0.20$	$MSE = 0.17, OOB = 0.21$

Tableau 4.4 – Comparaison de modèles

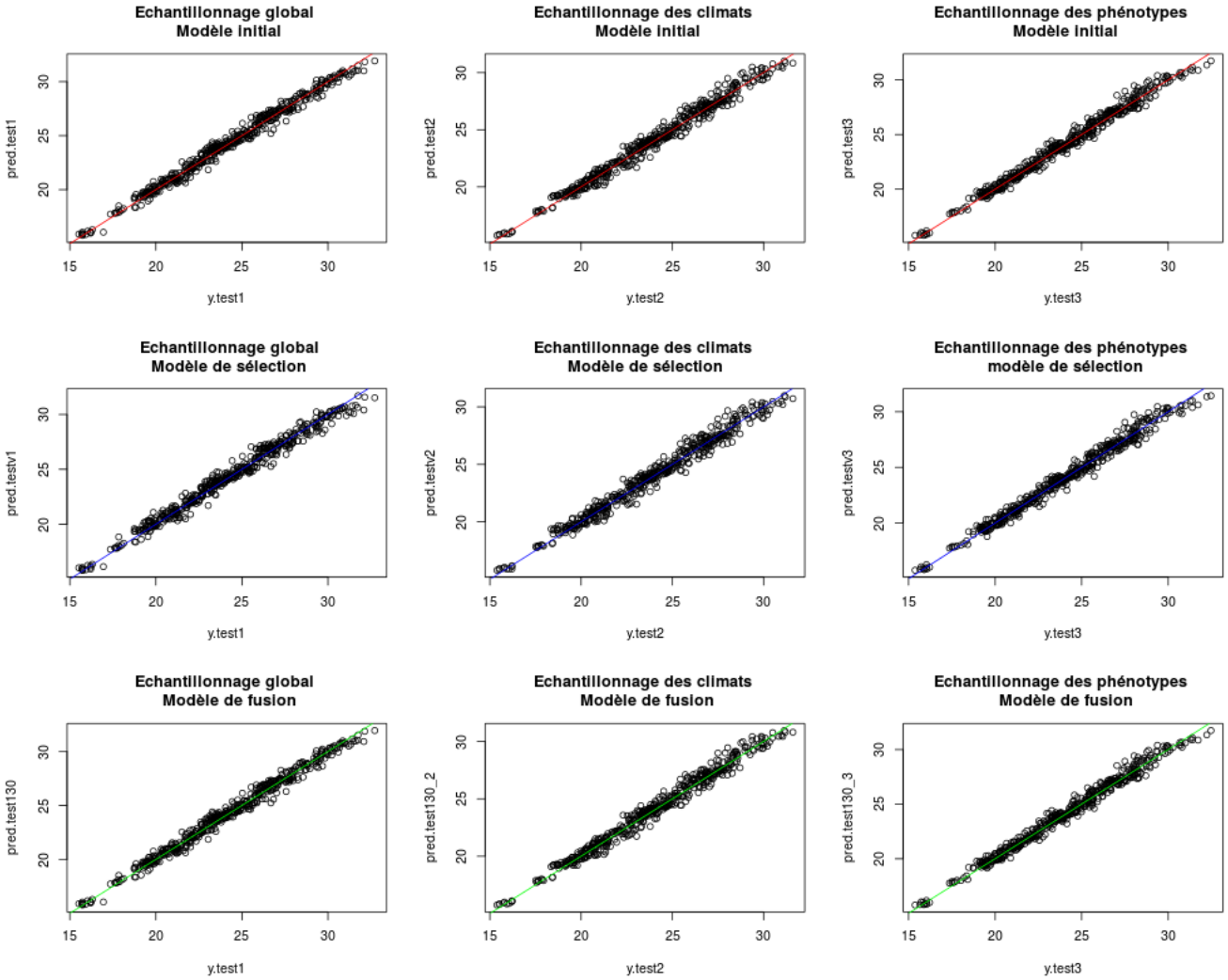


FIGURE 4.13 – Rendements prédits vs rendements réels

Données simulées à partir de SUNFLO

Note : Rendements prédits en fonction des rendements réels pour trois échantillons tests

- première colonne : échantillonnage global
- deuxième colonne : échantillonnage des climats
- troisième colonne : échantillonnage des phénotypes
- première ligne : modèle avec 186 variables
- deuxième ligne : modèle de sélection avec 7 variables
- troisième ligne : modèle de fusion avec 56 variables

Discussion

✿ Interprétation agronomique des résultats

Le tournesol est généralement cultivé sur la période avril-septembre. Les dates de semis et de récolte varient selon le producteur et la région. Dans le modèle SUNFLO, la date de semis est fixée au 14 avril (semaine 2). Les différentes phases de la culture du tournesol sont représentées sur la figure 5.1.

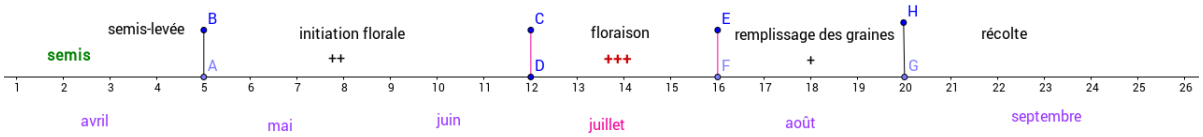


Figure 5.1 – Cycle de production du tournesol

Selon les dires d'experts, la phase de floraison est la phase la plus importante pour le rendement du tournesol. Elle devrait être marquée par une forte influence des variables climatiques, notamment la pluie, la température maximale et l'évapotranspiration. La phase d'initiation florale devrait être aussi marquée par une influence de l'évapotranspiration. Après la floraison, on devrait avoir une influence de certaines variables, en particulier l'ensoleillement. En confrontant ces faits aux résultats obtenus, on remarque que l'effet de l'évapotranspiration (ETP), de la température maximale (Tmax) et de l'ensoleillement (RAD) est proche des résultats attendus. Par contre le résultat le moins attendu est la faible influence de la pluie sur presque toutes les semaines. Le tournesol dans la réalité est sensible au mois de juillet alors que le modèle SUNFLO est sensible au mois de mai. Plusieurs hypothèses sont possibles pour expliquer ces différences. La première est que SUNFLO étant un modèle, donc une représentation d'une culture de tournesol, il n'est pas forcément fidèle à la réalité. Une autre hypothèse qui pourrait expliquer ce phénomène est la forte corrélation entre les variables. Cette corrélation existe non seulement entre les séries climatiques elles-mêmes (l'évapotranspiration et l'ensoleillement par exemple), mais aussi entre les résumés d'une même série. En effet, le fait d'utiliser des résumés hebdomadaires crée une certaine redondance de l'information. Sur 2 semaines consécutives

par exemple, l'information peut être la même pour une série donnée. Ainsi, une variable peut avoir été retenue non pas pour son influence sur le modèle, mais pour sa capacité à représenter un sous-ensemble de la série climatique.

✿ Difficultés

La principale difficulté rencontrée est la forte corrélation entre mes variables d'études, et la nature multi-dimensionnelle des données, c'est-à-dire le fait que plusieurs variables fonctionnelles doivent être utilisées simultanément. Cette forte corrélation ne permet pas l'utilisation des méthodes d'analyse classiques (les indices SRC par exemple) et rend peu interprétable les résultats. Au cours des différentes études, j'ai pu constater que toutes les variables sont influentes et seulement quelques unes ont un pouvoir prédictif élevé. Par ailleurs, la grande dimension des données a rendu très long le temps d'exécution des algorithmes utilisés, ce qui a retardé le travail.

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 tournesol, 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

Les données climatiques utilisées concernent 5 stations (Avignon, Blagnac, Dijon, Poitiers et Reims) et couvrent la période 1975-2012. Dans cette étude, toutes les stations et toutes les années ont été traitées de la même manière, alors que le climat change d'une année à une autre et d'une station à une autre. Ce changement peut aussi décaler la période de culture. Il serait donc intéressant d'inclure d'autres paramètres tenant compte de la nature de l'année (humide, sèche, chaude, froide, par exemple) et aussi de la station. Par ailleurs, le plan d'expérience utilisé dans cette étude est fait de telle manière que le même phénotype est croisé avec les 190 climats, ce qui entraîne une certaine redondance dans les données. C'est ce qui explique la grande dimension des données et pourrait également expliquer le fait que l'influence des phénotypes n'a pas pu réellement être captée.
Pour la suite du travail, on pourrait envisager de faire une ACP fonctionnelle et utiliser

les composantes principales pour l'analyse afin de contourner la corrélation. Une autre piste de réflexion serait de partir sur des intervalles de temps plus larges (un mois, deux mois) et de subdiviser les intervalles de temps les plus influents, plutôt que de partir sur des semaines et de regrouper les moins influentes.

✿ Conclusion personnelle

Mon passage à l'Unité Mathématique et Informatique Appliquées (MIAT) de l'INRA m'a permis de mettre en application les connaissances acquises au cours de ma formation. 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 (analyse de sensibilité, forêts aléatoires, 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. A côté de toutes ces compétences techniques, j'ai aussi eu l'occasion de découvrir le monde de l'agriculture et de me familiariser au domaine de la recherche, ce qui me motive à continuer dans ce sens.

Bibliographie

[1] Pierre Casadebaig, Lydie Guilioni, Jérémie Lecoeur, Angélique Christophe, Luc Champolivier, and Philippe Debaeke. Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and Forest Meteorology*, 151(2) :163 – 178, 2011.

[2] J.-E. Bergez, P. Chabrier, C. Gary, M.H. Jeuffroy, D. Makowski, G. Quesnel, E. Ramat, H. Raynal, N. Rousse, D. Wallach, P. Debaeke, P. Durand, M. Duru, J. Dury, P. Faverdin, C. Gascuel-Odoux, and F. Garcia. An open platform to build, evaluate and simulate integrated models of farming and agro-ecosystems. *Environmental Modelling and Software*, 39(1) :39–49, 2013.

[3] A. Saltelli, K. Chan, and E.M. Scott. *Sensitivity analysis*. Wiley, 2000.

[4] L. Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001.

[5] B. Gregorutti, B. Michel, and P. Saint Pierre. Grouped variable importance with random forest and application to multilple functional data analysis. hal-01084301v2, 2005.

[6] R. Genuer. *Forêts aléatoires, aspects théoriques, sélection de variables et application*. PhD thesis, Université Paris-Sud 11, 2010.

[7] H.Cukier, R.I. Levine, and K.Shuler. Non linear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26 :1–42, 1978.

[8] R.I.Cukier, J.H. Schaibly, and K.E. Shuler. Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients. *Journal of Chemical Physics*, 63 :1140–1149, 1975.

[9] R. Faivre, B. Iooss, D. Makowski, and H. Monod. *Analyse de sensibilité et exploration de modèles Application aux sciences de la nature et de l'environnement*. Quæ, 2013.

[10] J. Ali, R. Ahmad, and I. Maqsood. Random Forests and Decision Trees. *International Journal of Computer Science Issues*, 9, Issue 5, No 3, 2012.

[11] P. Lemaître B. Iooss. A review on global sensitivity analysis methods. arXiv :1404.2405v1[math.ST], 2014.

[12] K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measure. *Computational Statistics and Data Analysis*, 52 :2249–2260, 2008.

[13] J. Jacques. *Contribution à l'analyse de sensibilité et à l'analyse discriminante généralisée*. PhD thesis, Université Joseph-Fourier-Grenonle I, 2005.

[14] D. M. Hamby. A review of techniques for parameter sensitiviy analysis of environmental models. *Environmental Monitoring and Assessment*, 32 :135–154, 1994.

[15] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. *VSURF : Variable Selection Using Random Forests*, 2015. R package version 1.0.0.

[16] Gilles Pujol, Bertrand Iooss, and Alexandre Janon. *sensitivity : Sensitivity Analysis*, 2014. R package version 1.10.1.

Annexes

Analyse des résidus du modèle linéaire

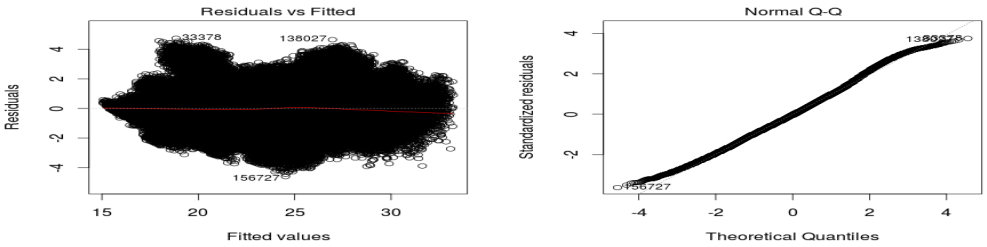


Figure 2 – Analyse des résidus

Données simulées à partir de SUNFLO

Résultats intermédiaires de la procédure de fusion de variables

— Fusions de 100 variables

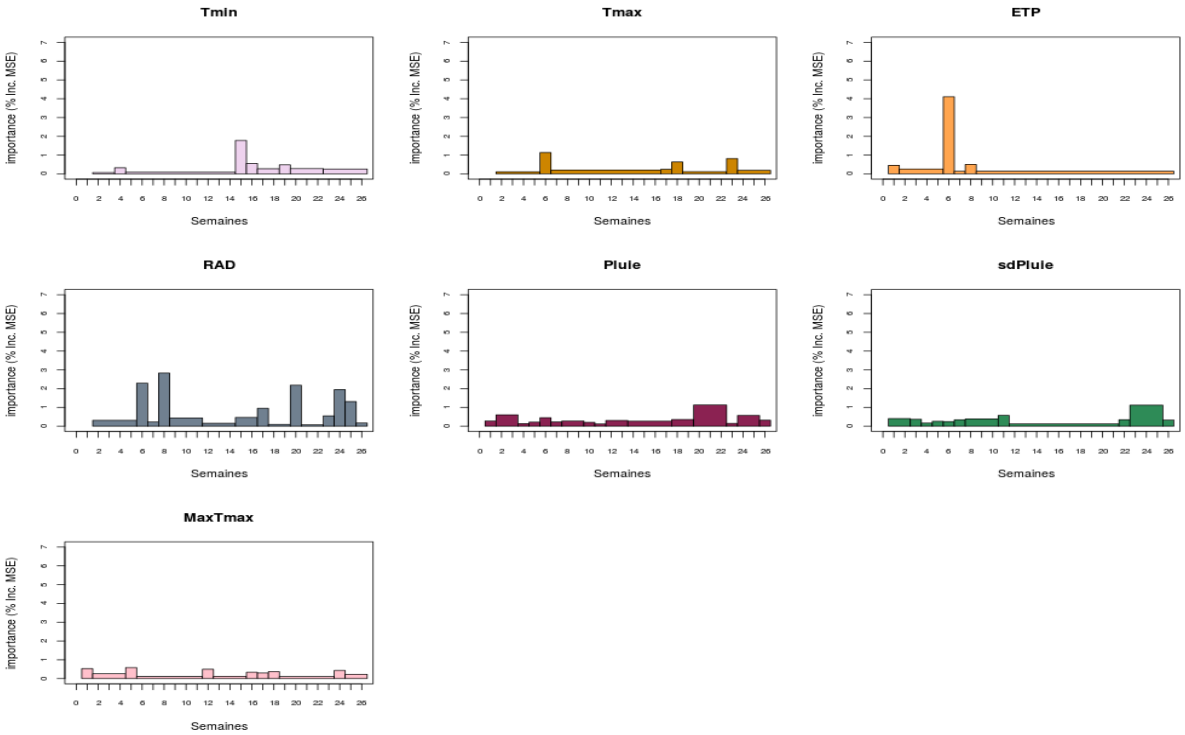


Figure 3 – Importances des variables après 100 fusions

Données simulées à partir de SUNFLO

— **Fusions de 160 variables**

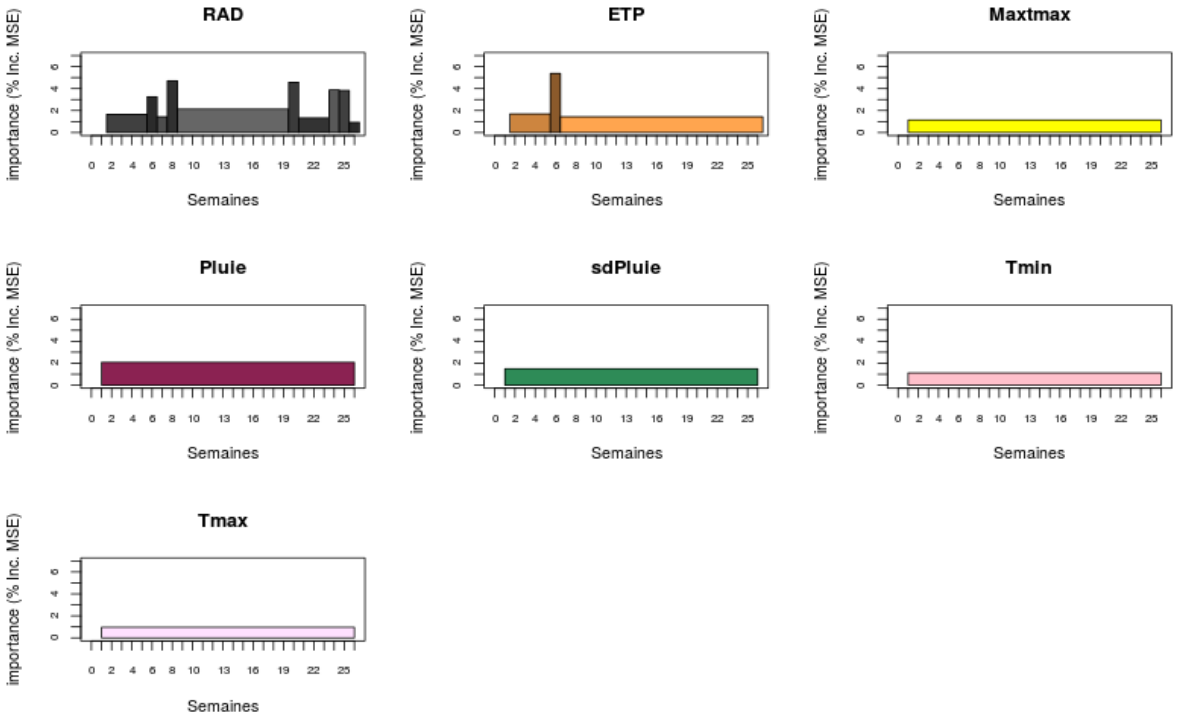


FIGURE 4 – Importances des variables après 160 fusions

Données simulées à partir de SUNFLO

Table des matières