Analyse par simulation de l'interaction climat / rendement

Soutenance de stage

Master Statistique-Econométrie

Présenté par :

Rolande C. B. KPEKOU TOSSOU

Stage encadré par :

Victor Picheny & Nathalie Villa-Vialaneix

3 Septembre 2015

TSE
Toulouse School of Economics
École d'économie de Toulouse

INRA
SCIENCE & IMPACT

Organisme d'accueil

✿ INRA

Organisme public de la recherche scientifique qui se charge des questions liées à l'agriculture, l'alimentation, l'environnement, ...

✿ MIAT

Développer et mettre à jour des méthodes et des compétences en mathématiques et ou en informatique.

✿ MAD

Modélisation des Agro-écosystèmes et Décision

Organisme d'accueil

✿ INRA

Organisme public de la recherche scientifique qui se charge des questions liées à l'agriculture, l'alimentation, l'environnement, ...

✿ MIAT

Développer et mettre à jour des méthodes et des compétences en mathématiques et ou en informatique.

✿ MAD

Modélisation des Agro-écosystèmes et Décision

Organisme d'accueil

✿ INRA

Organisme public de la recherche scientifique qui se charge des questions liées à l'agriculture, l'alimentation, l'environnement, ...

✿ MIAT

Développer et mettre à jour des méthodes et des compétences en mathématiques et ou en informatique.

✿ MAD

Modélisation des Agro-écosystèmes et Décision

Plan

1. Problématique et Objectifs
2. Données de l'étude
3. Méthodologie de l'étude
4. Présentation des résultats
5. Discussion
6. Conclusion

Problématique

- Influence forte des conditions climatiques sur les systèmes de culture

⇓

- Intégrer l'aléa climatique aux études d'optimisation de rendement

⇓

- La complexité des interactions climat/plante

⇓

- Recours aux approches dites par simulation

Problématique

- Influence forte des conditions climatiques sur les systèmes de culture

⇓

- Intégrer l'aléa climatique aux études d'optimisation de rendement

⇓

- La complexité des interactions climat/plante

⇓

- Recours aux approches dites par simulation

Problématique

- Influence forte des conditions climatiques sur les systèmes de culture

⇓

- Intégrer l'aléa climatique aux études d'optimisation de rendement

⇓

- La complexité des interactions climat/plante

⇓

- Recours aux approches dites par simulation

Problématique

- Influence forte des conditions climatiques sur les systèmes de culture

⇓

- Intégrer l'aléa climatique aux études d'optimisation de rendement

⇓

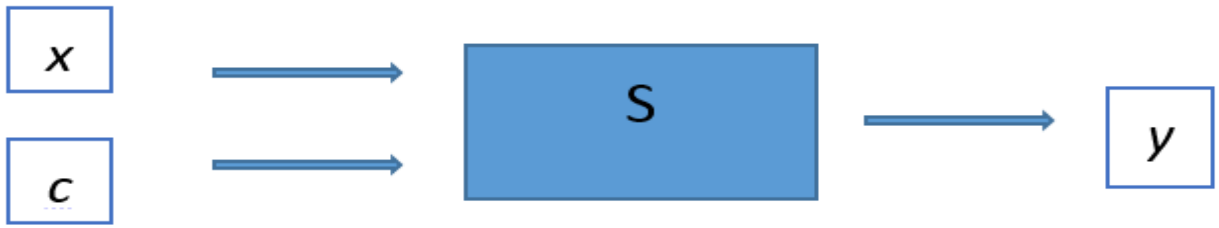
- La complexité des interactions climat/plante

⇓

- Recours aux approches dites par simulation

Le modèle SUNFLO

- SUNFLO : simulateur numérique qui reconstitue de façon dynamique l'évolution jour après jour d'une variété de tournesol en fonction de son milieu et de la conduite culturale.



$$\begin{array}{rccc} y & : & \mathbb{R}^8 \times \mathbb{R}^{5\times 183} & \to & \mathbb{R} \\ & & x, c & \mapsto & S(x,c) \end{array}$$

Objectifs

Objectif général

Comprendre la manière dont le climat influence le rendement du tournesol

De manière spécifique

- Mettre en valeur les motifs climatiques les plus influents pour le rendement du tournesol
- Identifier les intervalles de temps les plus influents
- Approche boîte noire

Objectifs

Objectif général

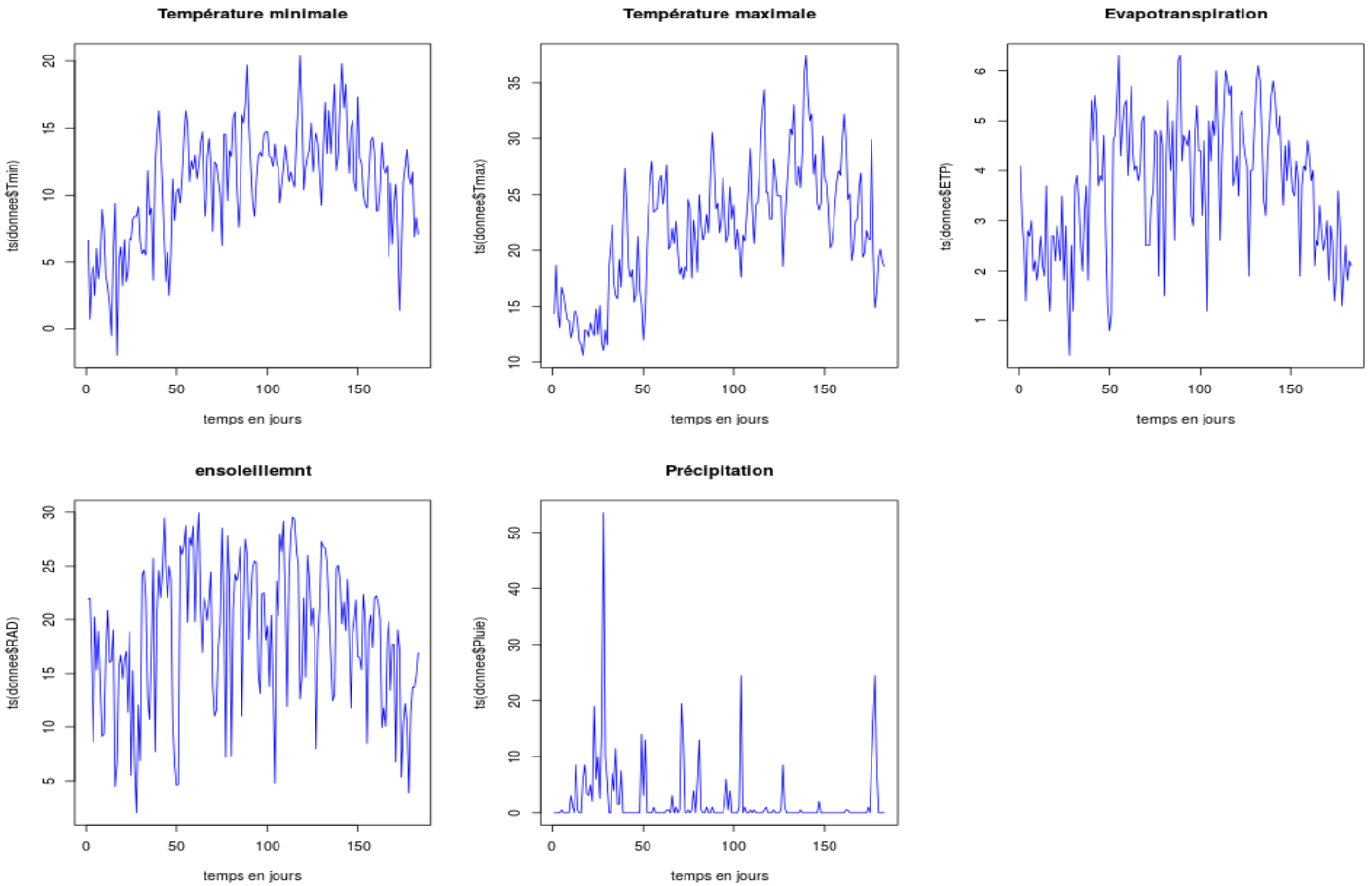
Comprendre la manière dont le climat influence le rendement du tournesol

De manière spécifique

- Mettre en valeur les motifs climatiques les plus influents pour le rendement du tournesol
- Identifier les intervalles de temps les plus influents
- Approche boîte noire

Données climatiques

- 190 relevés climatiques journaliers sur la période 1975-2012 (Dijon, Lusignan, Reims, Blagnac et Avignon)



Traits phénotypiques et rendement

1000 variétés de tournesol caractérisées par 8 traits phénotypiques

- Durée de la phase levée-floraison (TDF1)
- Durée de la phase levée-maturité (TDM3)
- Seuil de réponse de la conductance stomatique à une contrainte hydrique (TR)
- Coefficient d'extinction du rayonnement lors de la phase végétative (K)
- etc

Données sur le rendement

- Plan d'expérience LHS : croisement des 1000 variétés par les 190 relevés climatiques (soit 190000 observations)
- Calcul du rendement à partir du modèle SUNFLO

Traits phénotypiques et rendement

1000 variétés de tournesol caractérisées par 8 traits phénotypiques

- Durée de la phase levée-floraison (TDF1)
- Durée de la phase levée-maturité (TDM3)
- Seuil de réponse de la conductance stomatique à une contrainte hydrique (TR)
- Coefficient d'extinction du rayonnement lors de la phase végétative (K)
- etc

Données sur le rendement

- Plan d'expérience LHS : croisement des 1000 variétés par les 190 relevés climatiques (soit 190000 observations)
- Calcul du rendement à partir du modèle SUNFLO

Méta-modélisation

Méta-modélisation

Construction d'un modèle simplifié possédant de bonnes performances prédictives pour résumer au mieux le modèle initial

Méta-modèles utilisés

1. Le modèle linéaire
2. Les forêts aléatoires

Méta-modélisation

Méta-modélisation

Construction d'un modèle simplifié possédant de bonnes performances prédictives pour résumer au mieux le modèle initial

Méta-modèles utilisés

1. Le modèle linéaire
2. Les forêts aléatoires

Méta-modélisation

Méta-modélisation

Construction d'un modèle simplifié possédant de bonnes performances prédictives pour résumer au mieux le modèle initial

Méta-modèles utilisés

❶ Le modèle linéaire
❷ Les forêts aléatoires

Méta-modélisation

Démarche

- La période de culture (avril-septembre) a été subdivisée en intervalles de temps réguliers
- Différents niveaux de décomposition ont été testés (mois, deux semaines, une semaine)
- Des méta-variables (moyenne, écart-type, maximum) ont été construites sur chaque intervalle de temps (pour 26 semaines, 186 variables au total)
- Analyse de sensibilité sur le modèle linéaire et calcul d'importance sur les forêts aléatoires
- **Notations** : le jème facteur sera noté X^j et la sortie Y

Analyse de sensibilité sur modèle linéaire

Définition (approche globale)

L'objectif de l'analyse de sensibilité est de quantifier l'impact de chaque entrée d'un modèle sur sa sortie.

$$S_i = \frac{V(E[Y/X^j])}{V(Y)}$$

Les indices PCC (Partial correlation coefficient) cf B. Iooss et P. Lemaître

Ils évaluent la sensibilité de Y à X_j en éliminant l'effet des autres variables, sous l'hypothèse de linéarité

$$PCC_j = \frac{cov(Y, X_j|X_{-j})}{\sqrt{V(Y/X_{-j})V(X_j|X_{-j})}}$$

Analyse de sensibilité sur modèle linéaire

Définition (approche globale)

L'objectif de l'analyse de sensibilité est de quantifier l'impact de chaque entrée d'un modèle sur sa sortie.

$$S_i = \frac{V(E[Y/X^j])}{V(Y)}$$

Les indices PCC (Partial correlation coefficient) cf B. Iooss et P. Lemaître

Ils évaluent la sensibilité de Y à X_j en éliminant l'effet des autres variables, sous l'hypothèse de linéarité

$$PCC_j = \frac{cov(Y, X_j|X_{-j})}{\sqrt{V(Y/X_{-j})V(X_j|X_{-j})}}$$

Les forêts aléatoires

Définition (cf L. Breiman, 2001)

- Technique d'apprentissage très performante, à la fois pour des problèmes de classification et de régression.
- Construction et agrégation d'un grand nombre d'arbres de régression obtenus par la méthode CART.
- **Erreur OOB (Out-of-Bag)** : erreur basée sur du bootstrap.

Objectif : Calcul d'importance de variables

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.

Importance groupée (cf B. Gregorutti *et al.*, 2015)

Perte de qualité de prédiction induite par une permutation des valeurs observées des variables d'un groupe de variables donné.

Les forêts aléatoires

Définition (cf L. Breiman, 2001)

- Technique d'apprentissage très performante, à la fois pour des problèmes de classification et de régression.
- Construction et agrégation d'un grand nombre d'arbres de régression obtenus par la méthode CART.
- **Erreur OOB (Out-of-Bag)** : erreur basée sur du bootstrap.

Objectif : Calcul d'importance de variables

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.

Importance groupée (cf B. Gregorutti *et al.*, 2015)

Perte de qualité de prédiction induite par une permutation des valeurs observées des variables d'un groupe de variables donné.

Les forêts aléatoires

Définition (cf L. Breiman, 2001)

- Technique d'apprentissage très performante, à la fois pour des problèmes de classification et de régression.
- Construction et agrégation d'un grand nombre d'arbres de régression obtenus par la méthode CART.
- **Erreur OOB (Out-of-Bag)** : erreur basée sur du bootstrap.

Objectif : Calcul d'importance de variables

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.

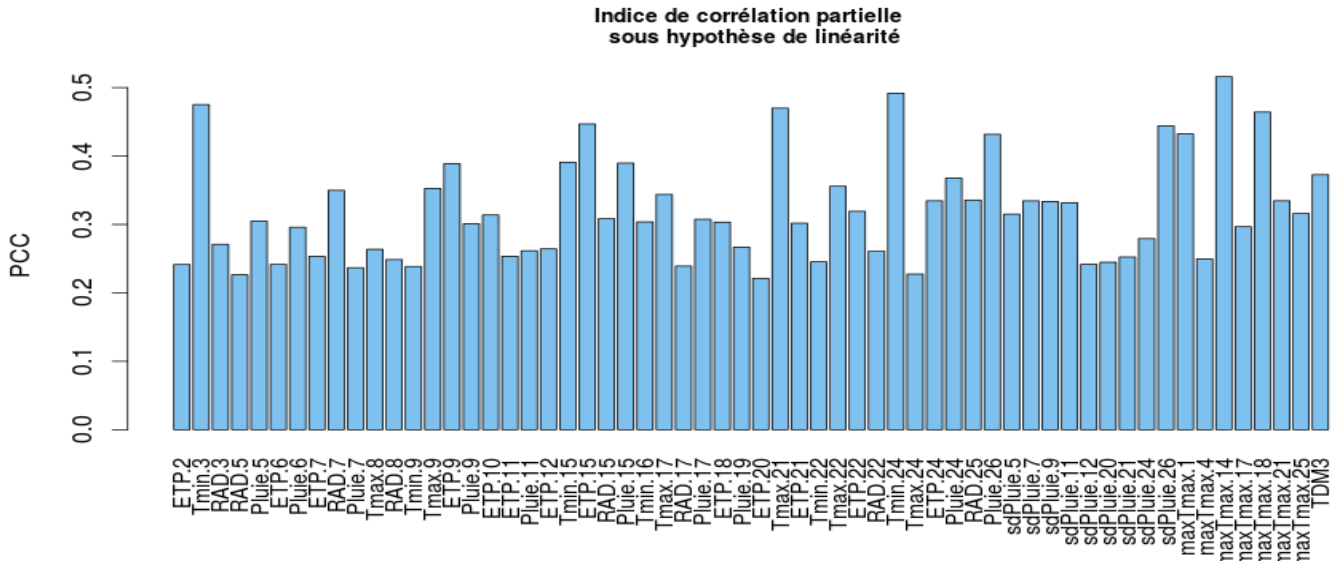
Importance groupée (cf B. Gregorutti *et al.*, 2015)

Perte de qualité de prédiction induite par une permutation des valeurs observées des variables d'un groupe de variables donné.

Modèle linéaire et analyse de sensibilité

- Variable à expliquer : le rendement du tournesol
- Variables explicatives : climat et phénotypes

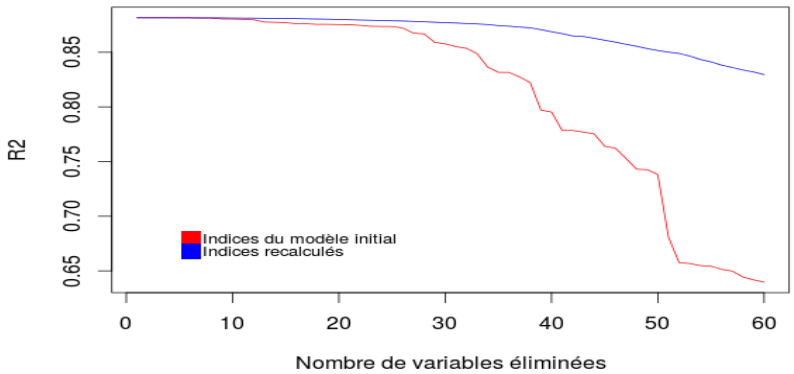
Modèles	**Nombre de variables**	R^2 ajusté en %
Analyse mensuelle	50	17.76
Analyse bihebdomadaire	99	42.35
Analyse hebdomadaire	186	**88.15**



Modèle linéaire et analyse de sensibilité

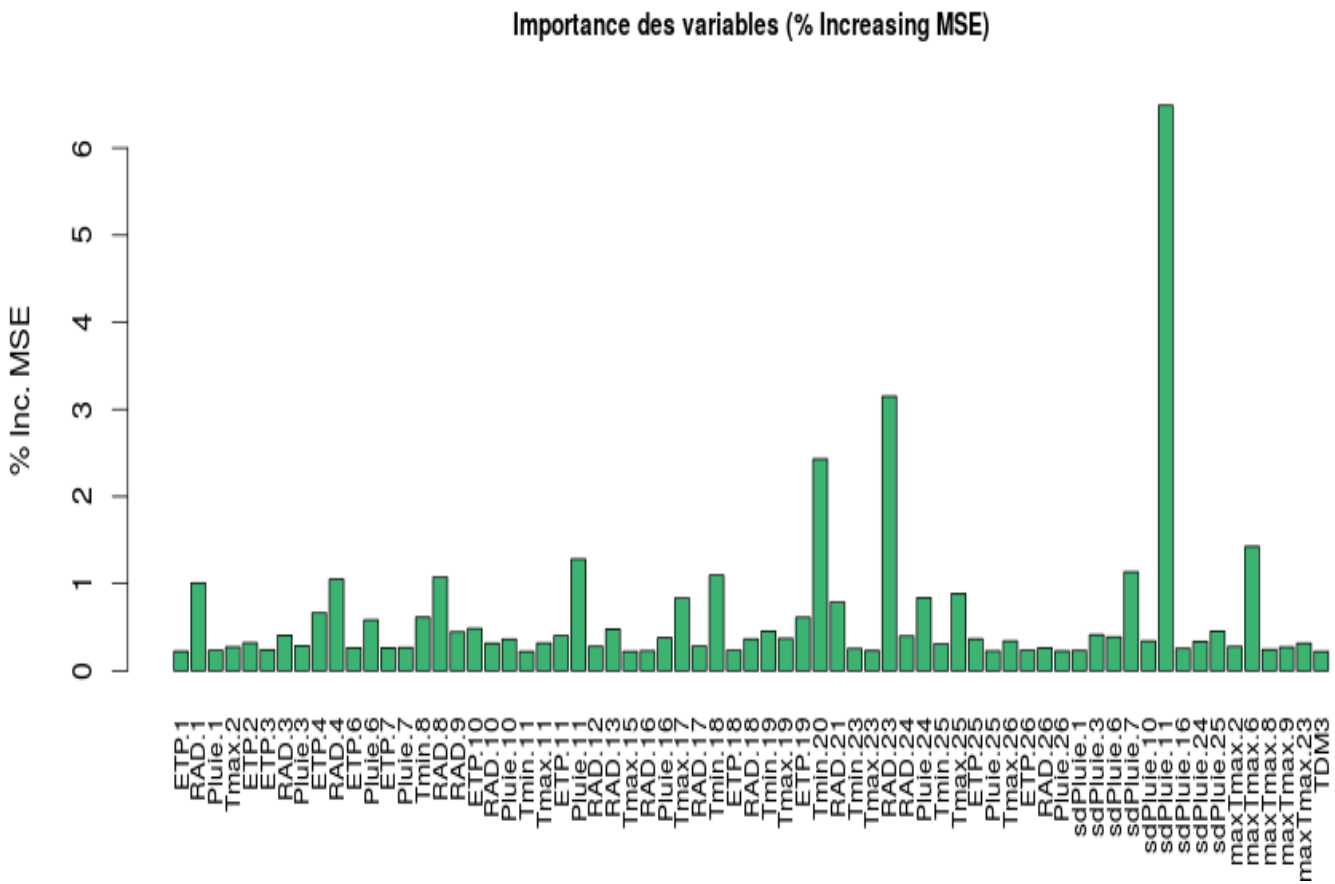
☛ Elimination de variables à partir des indices de sensibilité

- Eliminer à chaque itération la variable la moins influente à partir des indices du modèle initial.
- Eliminer la variable la moins influente en recalculant les indices après chaque élimination.



- Il est possible de réduire le nombre de variables en conservant le pouvoir explicatif du modèle.

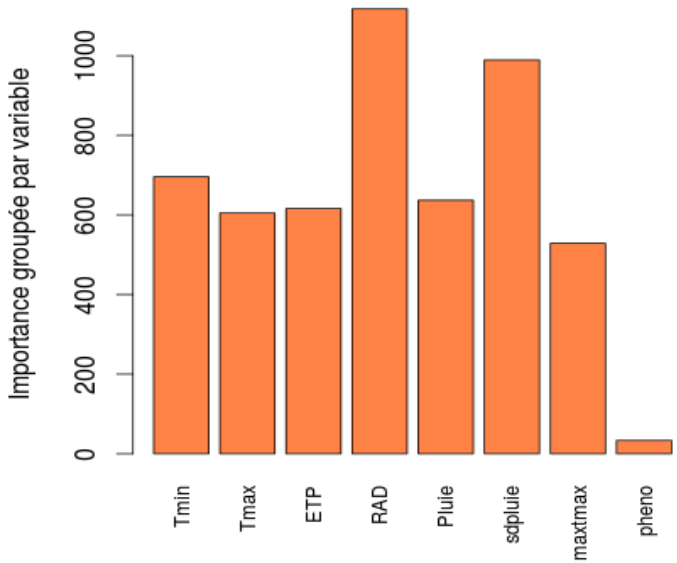
Forêts aléatoires

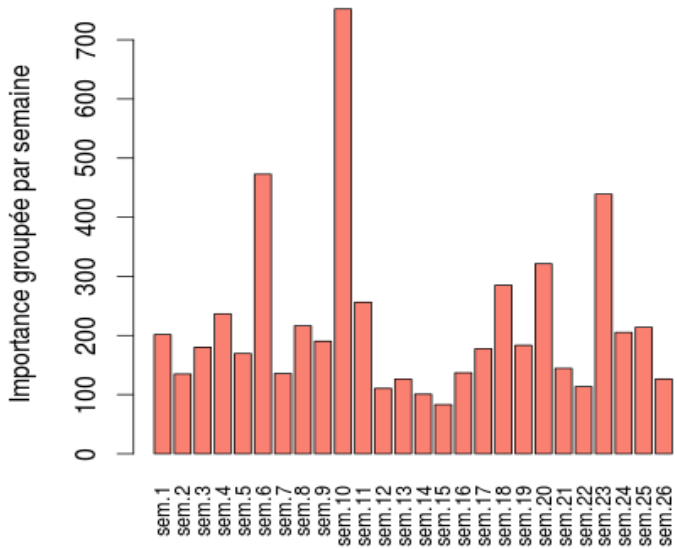


Forêts aléatoires

☛ Calcul d'importances groupées par variable et par semaine (package RFgroove)

- Groupes selon le type de variable : l'influence de chaque variable sur toutes les semaines.
- Groupes selon la semaine : l'influence de toutes les variables sur chaque semaine.

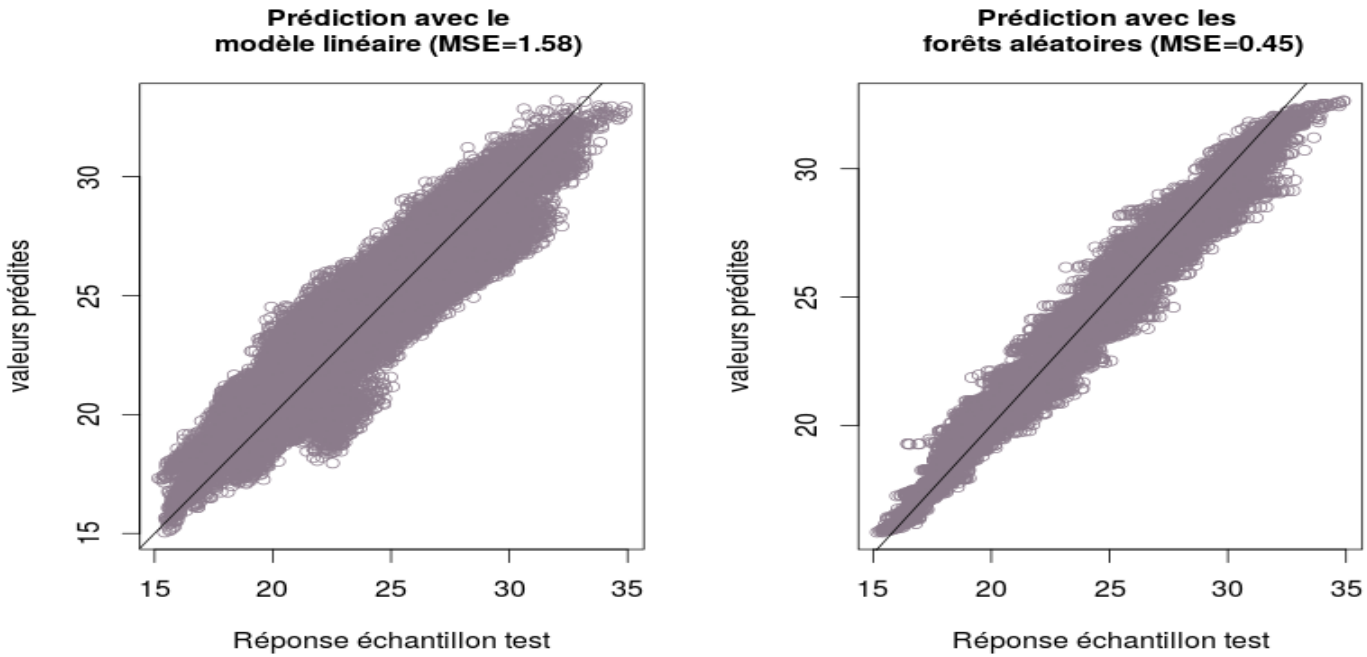




Comparaison de modèles

☛ Modèle linéaire et analyse de sensibilité : influence de toutes variables sur l'ensemble des semaines.

☛ Forêts aléatoires : semaines 6, 10 et 23 (mi-mai, mi-juin, fin septembre)

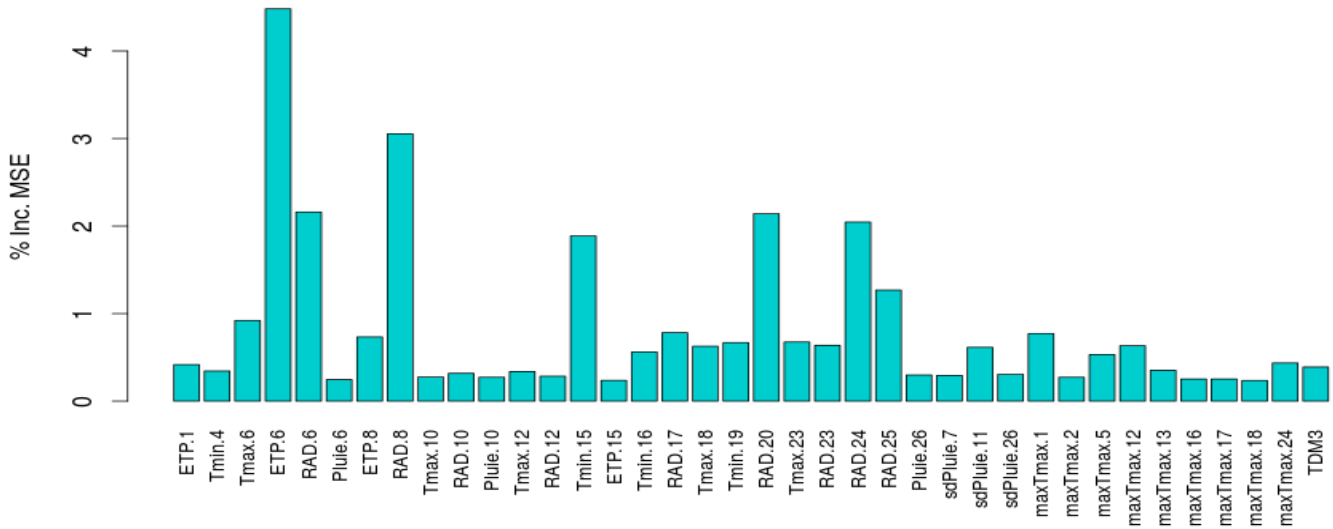


MSE (Mean Square Error) : $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

Forêts aléatoires

☛ Nouveau jeu de données

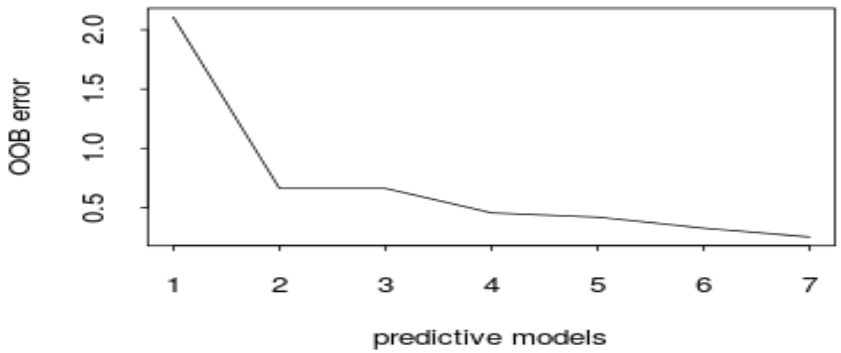
- 50 climats correspondant aux années 2003-2012 sur les 5 stations
- 50 phénotypes choisis aléatoirement parmi les 1000



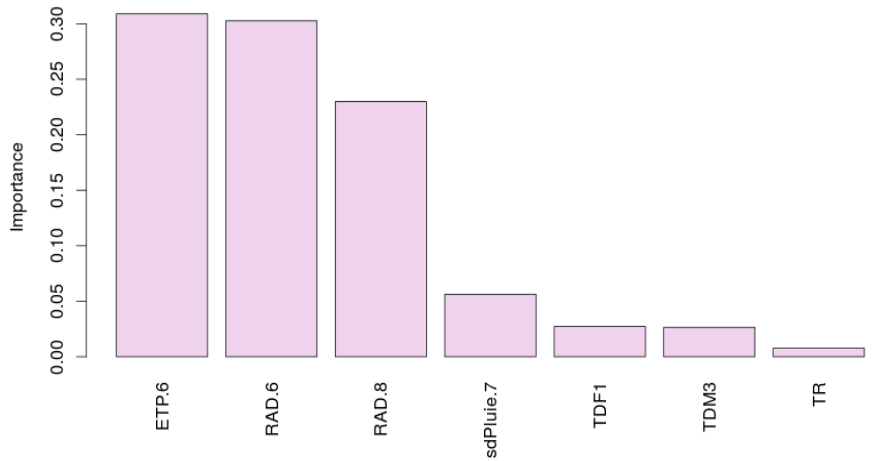
Forêts aléatoires et Sélection de variables (package VSURF de **R**)

(Genuer, 2010)

- Phase d'élimination : élimination préliminaire à partir des importances de variables.



- Introduction ascendante pas à pas des variables $\Rightarrow$ fournit 2 sous-ensembles de variables :
  - Phase d'interprétation : 183 variables
  - Phase de prédiction : 7 variables



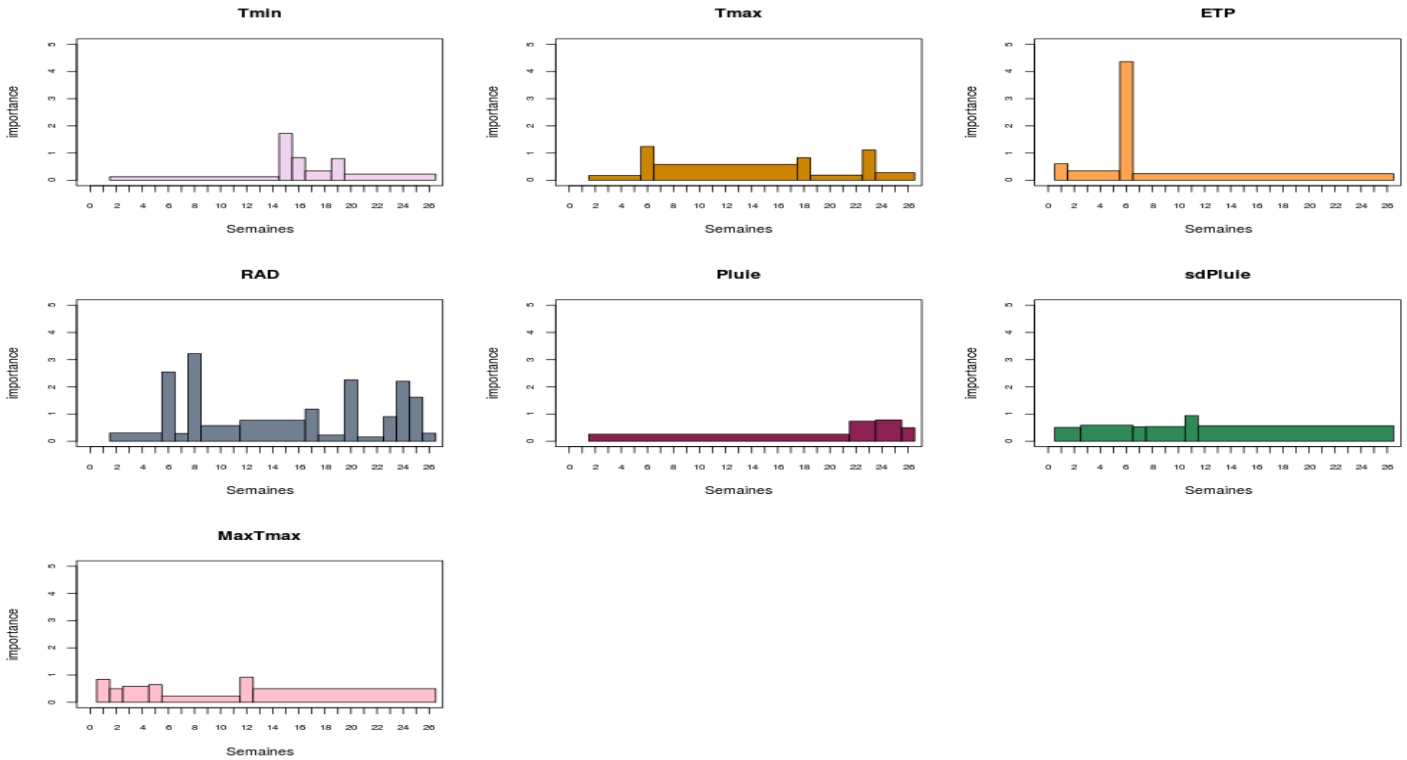
Proposition d'une nouvelle approche : Fusion de variables

☛ Démarche

- Former des paires de variables de la même série, du même résumé et sur des intervalles de temps consécutifs,
- Calculer des importances groupées pour l'ensemble des paires de variables formées,
- Fusionner les variables formant le groupe le moins influent,
- Estimer à nouveau le modèle de forêts aléatoires et calculer l'erreur OOB,
- Répéter le processus jusqu'à dégradation considérable de l'erreur OOB,
- On autorise des fusions sur plus de 2 semaines.

Forêts aléatoires et Fusion de variables

- Importance par variable sur 26 semaines



- Fusion et sélection aboutissent à la même conclusion.

Prédiction

☛ Comparaison de 3 modèles :

- le modèle initial avec 186 variables
- le modèle de sélection avec 7 variables
- le modèle de fusion avec 56 variables

☛ Sur 3 types d'échantillon test :

- échantillonnage global
  - Apprentissage : 80% de l'échantillon global
  - Test : 20% de l'échantillon global
- échantillonnage des climats
  - Apprentissage : 80% des climats et tous les phénotypes
  - Test : 20% des climats et tous les phénotypes
- échantillonnage des phénotypes
  - Apprentissage : 80% des phénotypes et tous les climats
  - Test : 20% des phénotypes et tous les climats

Prédiction (erreur MSE en test)

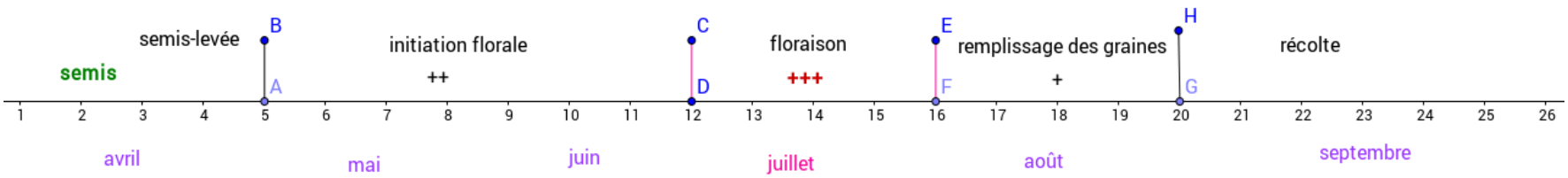
Erreur MSE (Mean Square Error)			
Modèles	**Global**	**Climats**	**Phénotypes**
Modèle à 186 variables	0.18	0.21	0.17
Sélection (7 variables)	0.24	0.26	0.19
Fusion (56 variables)	0.18	0.22	0.17

Discussion

✎ Les variables influentes se retrouvent au mois de mai

- Tmin : forte influence sur la période mi-juillet à août (semaines 15 à 19)
- Tmax : forte influence sur la période mi-mai et fin septembre (semaines 6 et 23),
- ETP : forte influence durant la semaine 6 (mi-mai),
- RAD : forte influence sur les semaines 6, 8, 20, 23 et 24 (mai,fin août et fin septembre),
- Pluie : faible influence

✎ Le tournesol est sensible au mois de juillet



- Forte influence de : Tmax, ETP, Pluie courant le mois de juillet

Comment expliquer cette différence

☛ 2 hypothèses

- SUNFLO est un modèle, donc une représentation de la réalité
- Le type de méta-modèle
- Forte corrélation entre les variables

Conclusion

☛ Objectifs du stage atteints

- Identifier les variables et les intervalles de temps les plus influents du tournesol à partir de SUNFLO

☛ Bilan technique

- Nouveaux outils : analyse de sensibilité, forêts aléatoires,
- Renforcement de mes connaissances du logiciel **R**,
- Nouveau logiciel : GIT,
- Apprentissage : Latex

☛ Bilan personnel

- Découverte du milieu de la recherche,
- Application de la statistique à un nouveau domaine : agriculture

MERCI DE VOTRE ATTENTION

Annexes

Les indices SRC(Standardized Regression coefficient)

Ils expriment la part de la variance de la réponse de Y due à la variance de la variable X_j lorsque les entrées sont indépendantes.

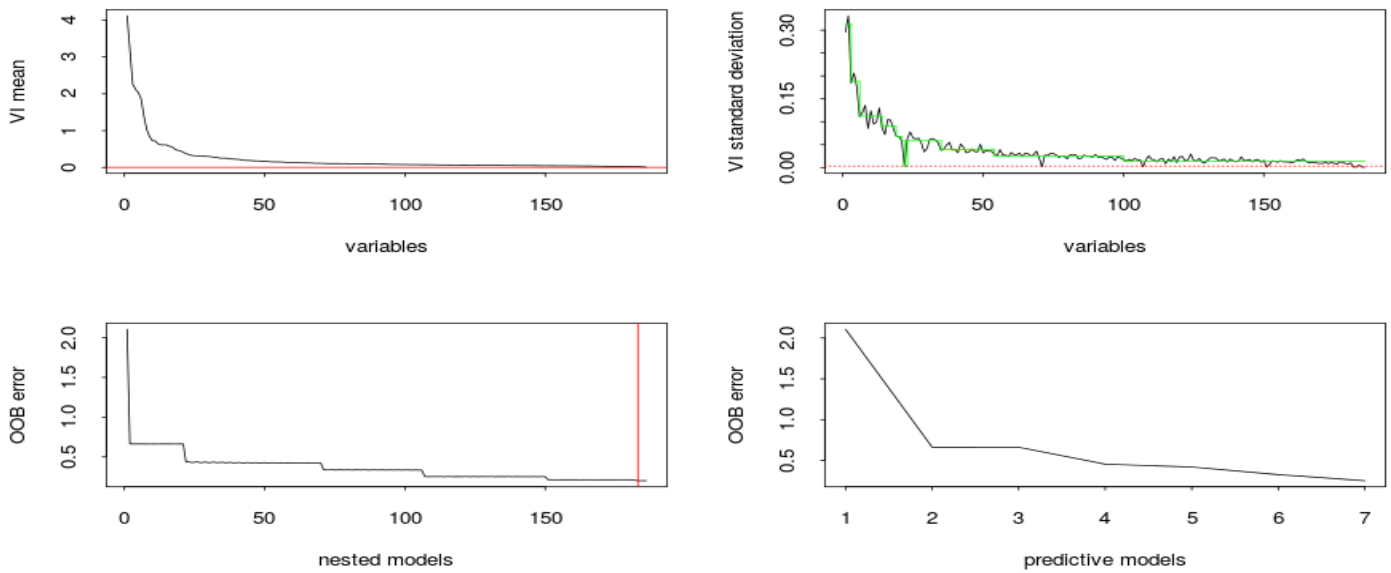
$$SRC_j = \frac{\beta_j^2 \, Var(X_j)}{Var(Y)}$$

Erreur OOB

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

avec $\hat{y}_i^{OOB}$ la moyenne des prédictions faites par les arbres pour lesquels l'observation i ne faisait pas partie de l'échantillon bootstrap.

Sélection de variables



Fusion de variables