21/03/2018 – Paris

Arnaud Gauffreteau
Celia Pontet
Margaux d'Orchymont
Josiane Lorgeou

# De la modélisation à la prédiction des variations de rendement variétaux dans un réseau d'essais virtuels en tournesol

## Intérêts et limites des approches statistiques classiques

GEVES
Expertise & Performance

ARVALIS
Institut du végétal

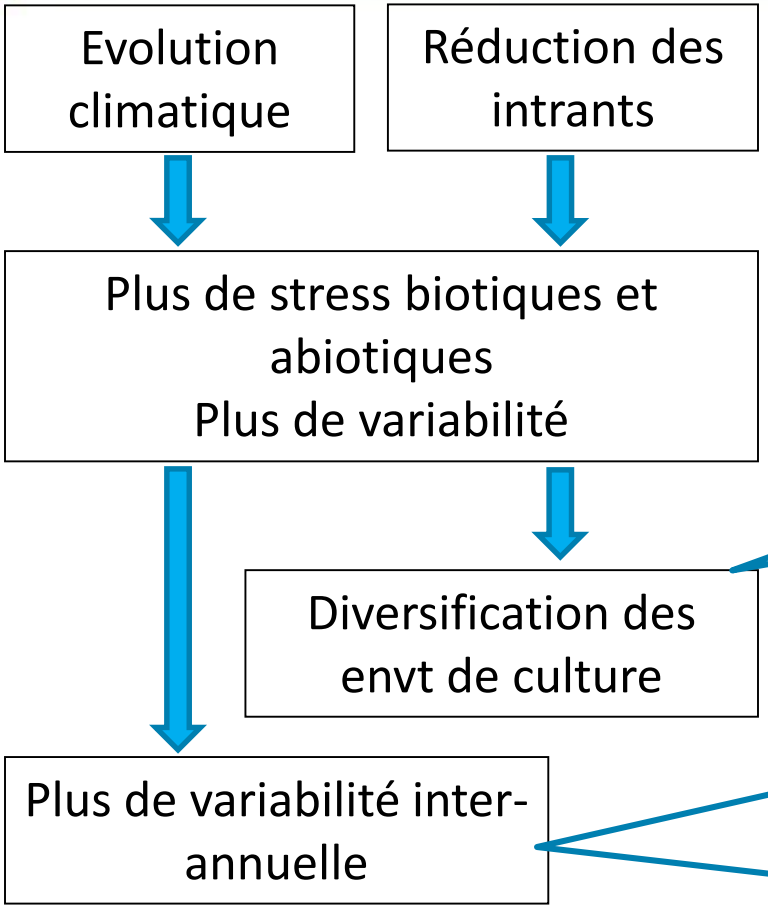
INRA
SCIENCE & IMPACT

Institut Technique de la Betterave ITB

Terres Inovia
l'agronomie en mouvement

Grande Culture
GCHP2E
Hautes Performances Economiques et Environnementales

# Contexte




- Adaptation des variétés aux environnements de culture
- Besoin de variétés :
  **Diversifiées** : et adaptées à une large gamme d'environnements
  **Caractérisées** : Performances moyennes et réponse aux différentes conditions de cultures au champ

→Modéliser les IGE dans les réseaux d'essais variétaux (MET)

- Importance de la stabilité des variétés
- Réflexion à l'échelle de bouquets de variétés

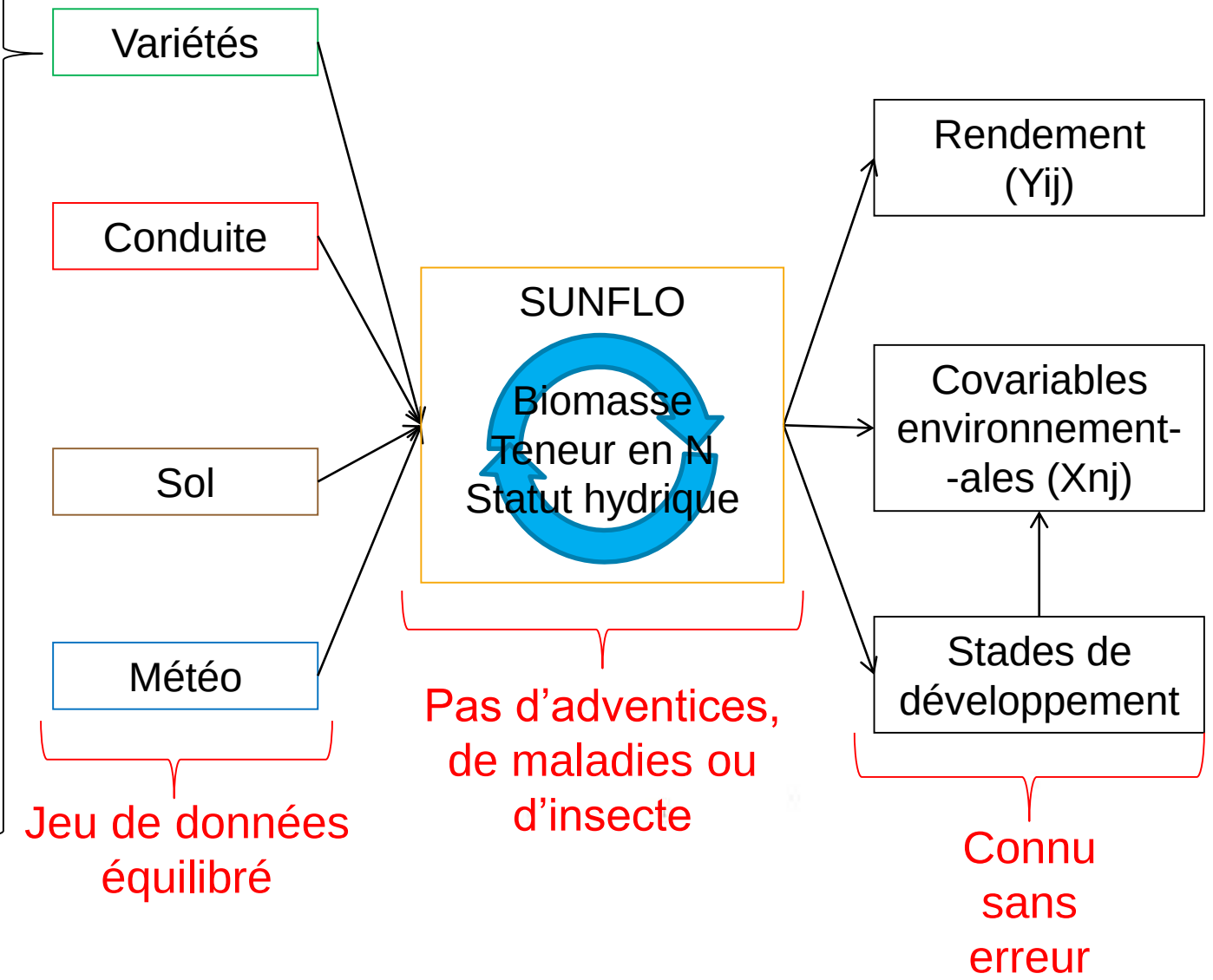
→Estimation et prédiction de la stabilité de variétés et bouquets variétaux

Qualité prédictive des modèles d'IGE modeste sur réseaux d'essais variétaux

- Qualité des données (erreurs sur mesures et variables environnementales)?
- Pertinence des variables environnementales?
- Pertinences des méthodes statistiques?

# Données générées par modélisation

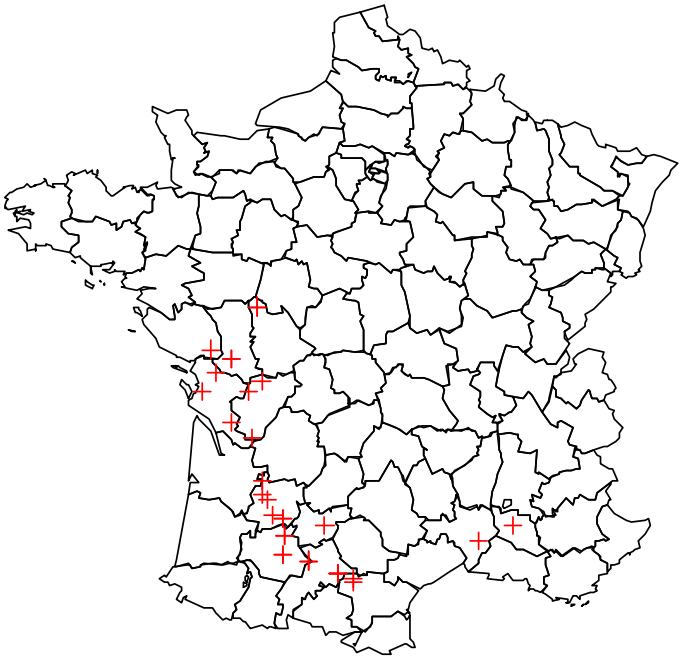
- Précocité floraison
- **Précocité Maturité**
- Nb feuilles pot à flo
- **Rang de la feuille la plus large à flo**
- **Surface pot de la feuille la plus large à flo**
- Coef d'extinction lumineuse pendant la croissance végétative
- **Seuil pour impact de stress H sur expansion foliaire**
- **Seuil pour impact de stress H sur conductance stomatique**



# Réseau d'essais et covariables environnementales

- 7 années : 2008-2014
- 30 sites (23 stations météo)
- RU représentatives des parcelles en tournesol
- 32 variétés virtuelles
- 12 covariables mesurant pour chaque environnement
  - Stress hydrique
  - Stress azoté
  - Offre en rayonnement
  - Offre en température
  - Hautes Température
  - Basses températures

  pendant
  - La période végétative
  - La floraison
  - Le remplissage



| Variable | Écart-type (q/ha) |
|---|---|
| Environnement | 6.2 |
| Génotypes | 1.7 |
| GxE | 1.3 |

# Modèles statistiques

1. Estimation des effets environnementaux et IGE

$$Y_{ij} = \mu + G_i + E_j + GxE_{ij} + \varepsilon_{ij}$$

2. Expression de $E_j$ et $GxE_{ij}$ en fonction des variables environnementales par différentes méthodes statistiques

$$E_j = f(X_{1j}, \ldots, X_{nj}) + \varepsilon'_j$$

$$GxE_{ij} = f(G_i, X_{1j}, \ldots, X_{nj}) + \varepsilon'_{ij}$$

Méthodes mobilisées :

- Régression linéaire simple
- Régression pénalisée lasso
- Régression PLS1 et PLS2
- Random forest

# Performance en estimation

| Méthode | $R^2$ sur Effet E | $R^2$ sur Effet GxE |
|---|---|---|
| reg linéaire | 93.1% | 66.4% |
| reg lasso | 92.9% | 65.8% |
| Random Forest | 98.3% | 89.4% |
| PLS1 | 91.8% (2 axes) | 64.8% (5 axes) |
| PLS2 | | 60.7% (3 axes) |

- Avantage à la RF
- Sinon une part d'IGE expliquée plus limitée

➔ Mauvaise prise en compte des interactions entre covariables?

| Variables envt | $R^2$ $(\widehat{E})$ | $R^2$ $(\widehat{IGE})$ | $R^2$ $(\widehat{IGE})$ / $R^2$ $(\widehat{E})$ |
|---|---|---|---|
| DH_VEG | 53.7% | 40.1% | 0.75 |
| DH_FLO | 38.4% | 46.8% | 1.22 |
| DH_REM | 1.0% | 2.5% | 2.42 |
| Autres_VEG | 3.1% | 5.2% | 1.70 |
| Autres_FLO | 3.7% | 3.3% | 0.88 |
| Autres_REM | 0.1% | 2.1% | 21.98 |

- Les stress précoces génèrent des interactions difficiles à modéliser

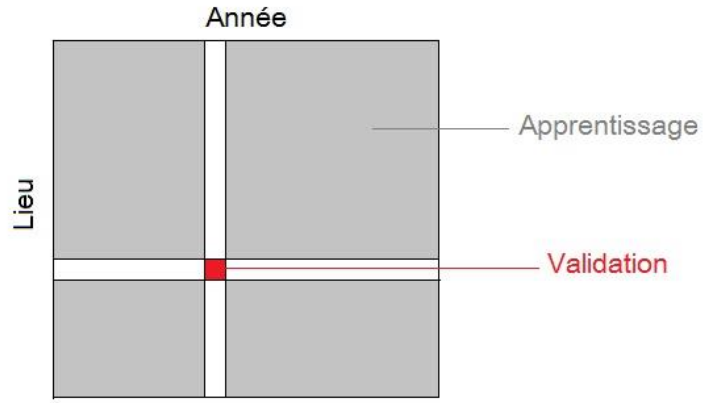
# Performance en prédiction

1. Estimation des effets environnementaux et IGE sur toute la BdD

$$Y_{ij} = \mu + G_i + E_j + GxE_{ij} + \varepsilon_{ij}$$

2. Validation croisée sur les $E_j$ et $GxE_{ij}$ : On suppose $G_i$ connus parfaitement

   Validation croisée sur années et sites

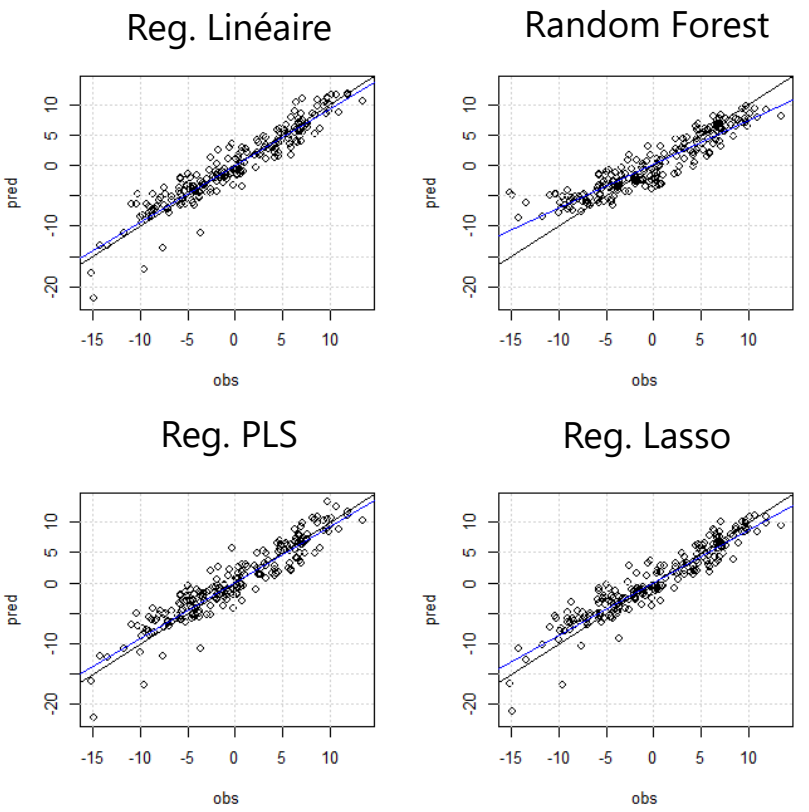


# Prédiction de l'effet E

| Méthode | $R^2$ (estimation) | $R^2$ (prédiction) |
|---|---|---|
| reg linéaire | 93.1% | 89.4% |
| random forest | 98.3% | 83.4% |
| reg PLS (2 axes) | 91.8% | 87.6% |
| reg lasso | 92.9% | 88.5% |

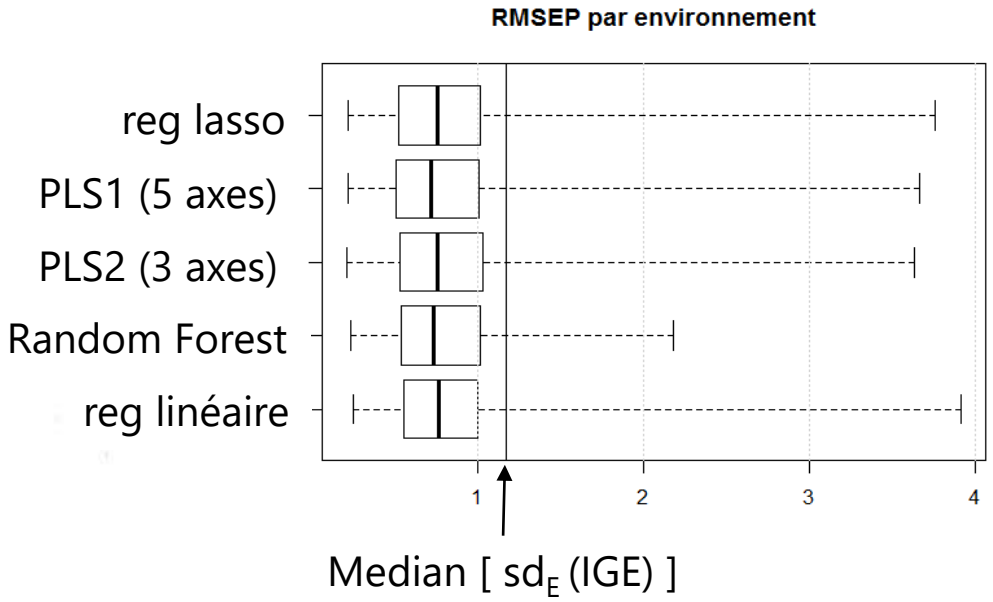
- Bonne prédiction des variations moyenne de rendement
- Random Forest décroche un peu plus que les autres méthodes

➔ Stress agissant de façon plutôt additive et linéaire



# Prédiction des IGE

| Méthode | $R^2$ (estimation) | R2 (prédiction) |
|---|---|---|
| reg linéaire | 66.4% | 42.9% |
| Random Forest | 89.4% | 55.5% |
| reg PLS2 (3 axes) | 60.7% | 47.9% |
| reg PLS1 (5 axes) | 64.8% | 48.3% |
| reg lasso | 65.8% | 47.4% |




- Perte significative entre estimation et prédiction (de 21 à 38% de la variabilité expliquée n'est pas prédite par les modèles)

➔Des interactions entre variables Moins de robustesse dans la prédiction des IGE

- Random Forest reste la meilleure méthode même si baisse forte entre qualité explicative et prédictive

➔ Des interactions entre variables environnementales à considérer

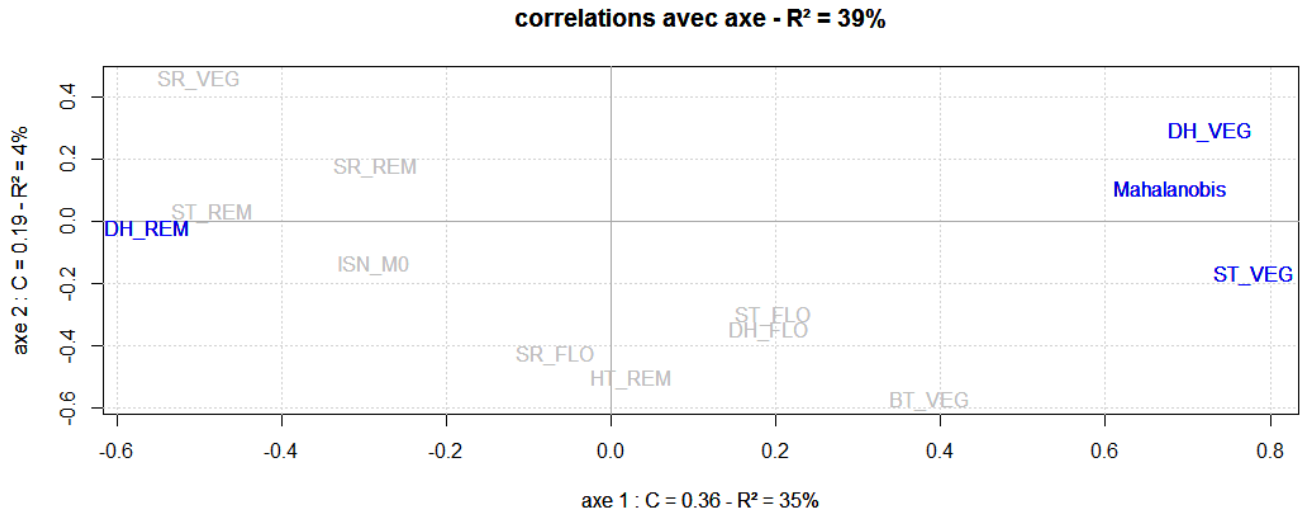
- Une prédiction améliorée par les modèles dans une majorité d'environnements mais des erreurs de prédiction variables

➔Pour quels environnements les IGE sont-elles mal prédites

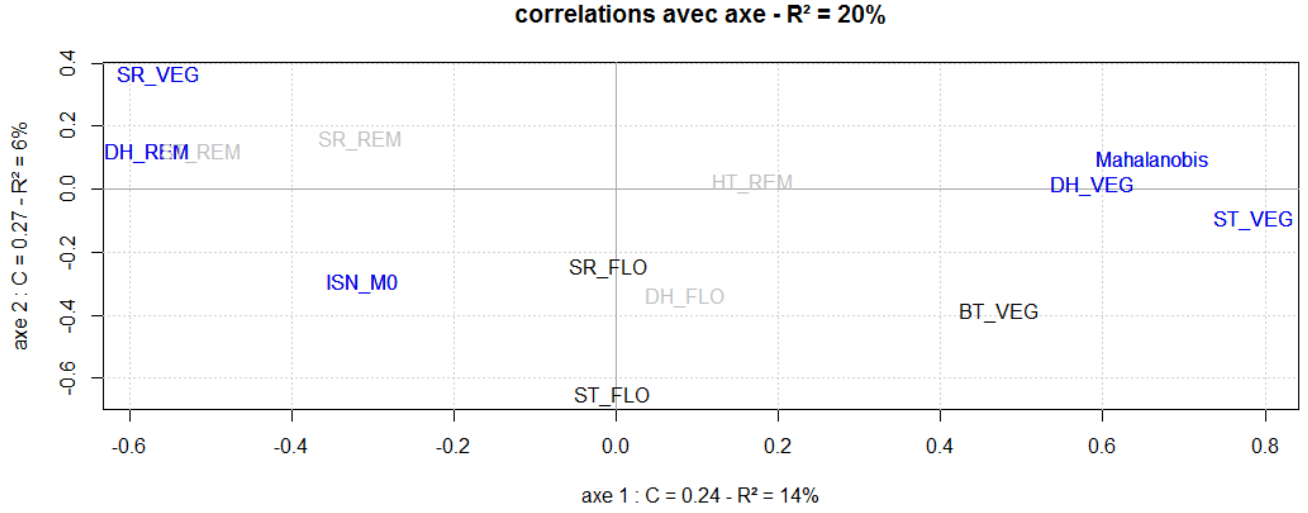
# Quels sont les environnements dont les IGE sont mal prédites?

**Régression PLS** : $RMSEP_j = f(X1_j, ..., Xn_j, Mahalanobis_j) + \varepsilon_j$

Régression Linéaire




Random Forest



# En conclusion

- Quelques covariables environnementales permettent de correctement prédire les variations moyennes de rendement entre environnements (utiles pour classer les environnements)
- Même dans un contexte très optimiste avec des données générées parfaitement connues sans erreurs sur les X et les Y, les modèles de GxE n'améliorent pas autant qu'on l'attendrait la prédiction des IGE
- Une partie de cette erreur de prédiction est directement due à un problème d'estimation
- Effet de la singularité des milieux d'essais
- Les interactions précoces sont difficilement modélisables et prédictibles : Elles induisent des statuts différents entre plantes précocement dans le cycle et qui entrainent par la suite des comportements différents face aux stress et aux ressources

➔ Reprendre le jeu de données pour différencier l'effet de la précocité des stress et l'effet de la singularité des milieux sur la qualité prédictive des modèles

➔ D'autres approches à explorer : composantes de rendement, approches bayésiennes, modèles d'apparentement…