

Évaluation de modèles déterministes complexes

François Husson

husson@agrocampus-ouest.fr

UP de mathématiques appliquées
Institut Agro - Rennes

1 / 39

Plan

- Introduction sur les modèles
- Présentation d'un modèle de croissance de colza et de la problématique associée
- Évaluation du modèle avec des données indépendantes des données utilisées pour la construction du modèle
 - Choix des données permettant l'évaluation
 - Présentation des méthodes permettant l'évaluation
- Évaluation du modèle avec les mêmes données que celles utilisées pour la construction du modèle
- Discussion

2 / 39

Introduction

Modèles de simulation :

- représentation simplifiée de phénomènes réels
- outils de réflexion et de synthèse pluridisciplinaire
- critères de prévision, et plus récemment, comme outils de test

Quelle que soit l'utilisation du modèle, nécessité de vérifier sa qualité : **étape d'évaluation**

Décisions prises à partir d'un modèle non validé \iff décisions prises sans modèle

All models are false, some are useful (George E. P. Box)

3 / 39

Introduction



Production de colza (1996) = 2 800 000 tonnes sur 850 000 ha
(2009) = 5 300 000 tonnes sur 1 500 000 ha

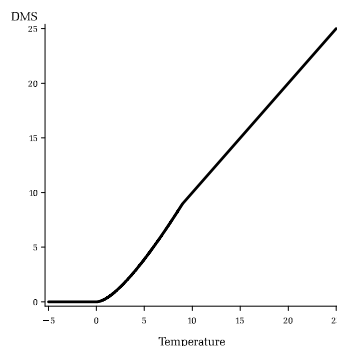
Culture gourmande en azote \Rightarrow risque de pollution par les nitrates
 \Rightarrow Doses d'azote doivent optimiser le rendement et minimiser les lessivages des nitrates

4 / 39

Description du modèle CECOL

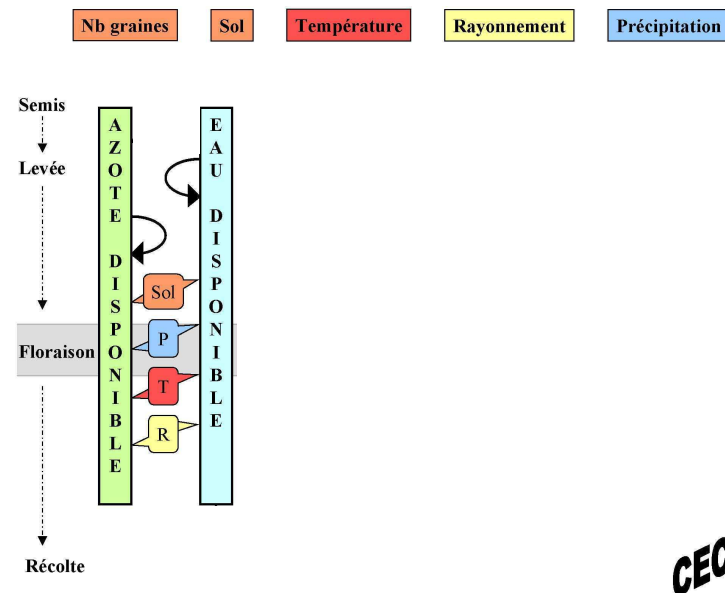
CECOL : modèle de croissance du colza d'hiver

- Modèle complexe
- Construit comme beaucoup de modèles agronomiques
- Modèle sol (CERES-N Maize) + un modèle plante (Colibri)
- 2000 lignes de Fortran, 69 paramètres « plante »
- Équations non linéaires



5 / 39

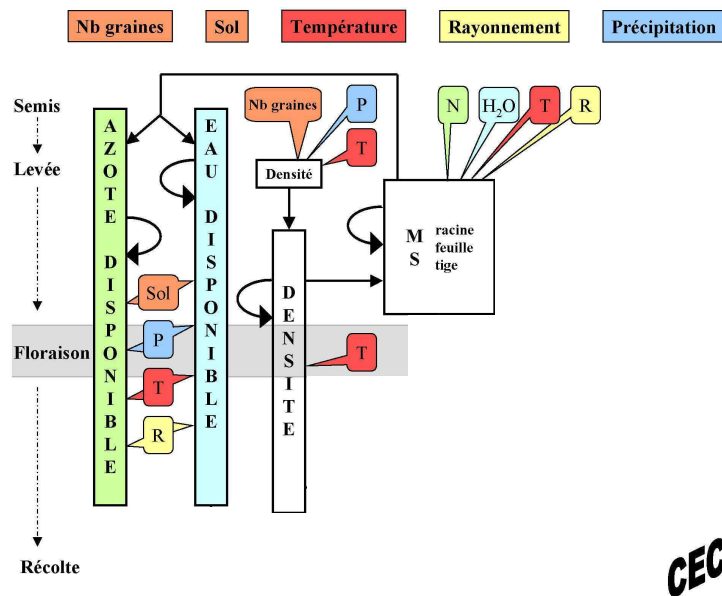
Description du modèle CECOL



CECOL

6 / 39

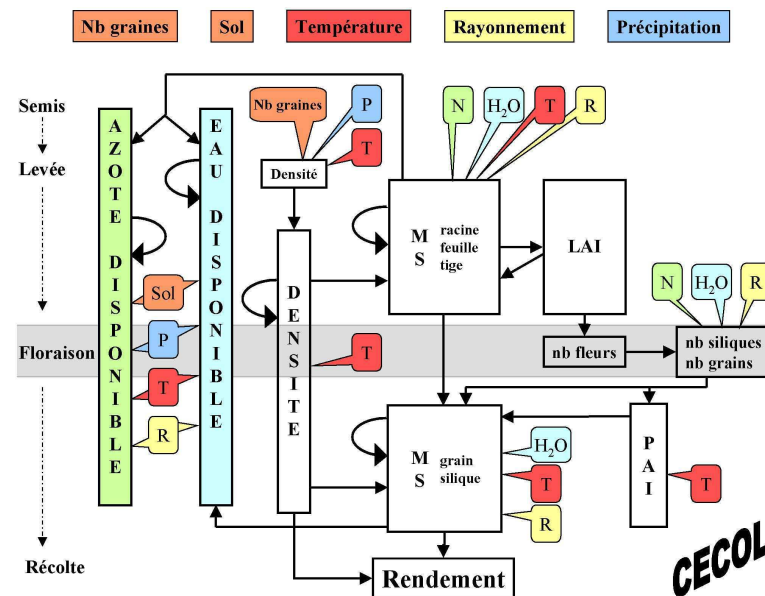
Description du modèle CECOL



CECOL

6 / 39

Description du modèle CECOL

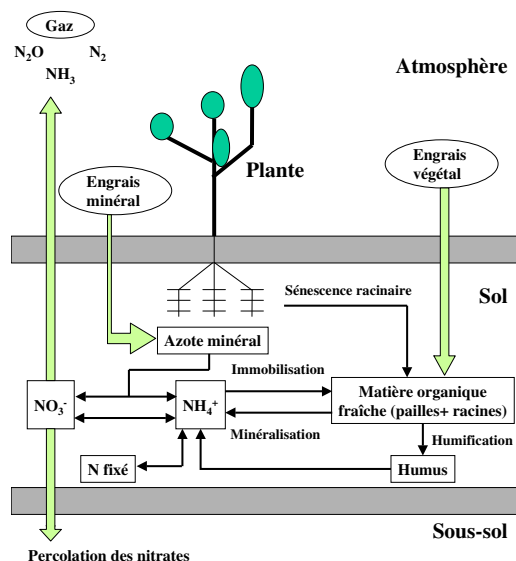


CECOL

6 / 39

Description du modèle CECOL

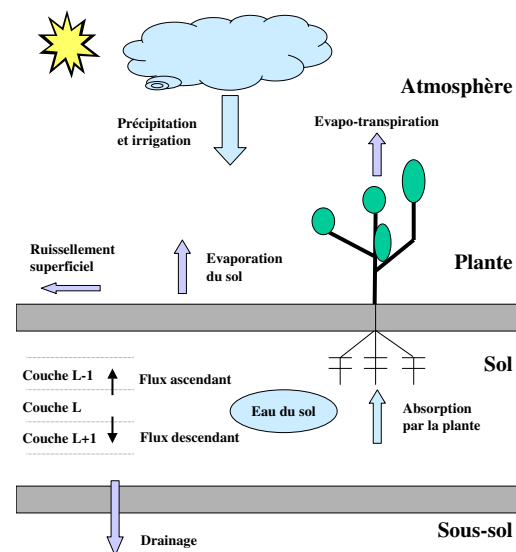
Cycle de l'azote dans le système sol-plante



7 / 39

Description du modèle CECOL

Cycle de l'eau dans le système sol-plante



8 / 39

Évaluation

L'évaluation d'un modèle est :

- toujours **indispensable** avant d'utiliser un modèle
- parfois laissée de côté, souvent très succincte
- dépend des applications et de l'utilisation du modèle
- pourtant : bonne évaluation donne des pistes pour améliorer le modèle

⇒ **But** : proposer une démarche (*i.e.* plusieurs techniques complémentaires) permettant de valider des modèles complexes linéaires, non linéaires, dynamiques, statiques, mécanistes

But n'est pas l'acceptation ou le rejet du modèle, mais l'évaluation des qualités du modèle

9 / 39

Définition du modèle

y : résultat d'un phénomène complexe, dépend de paramètres (notés β) et de variables explicatives (noté x)

Exemple : y rendement de maïs, un des β est le coefficient de conversion de la lumière et des x sont les données météo

y est prédite par une fonction de β et de x notée $f(\beta, x)$

f : une seule équation ou résultat de nombreuses équations

Hypothèses :

- tous les paramètres ont été estimés par des expériences antérieures ⇒ $\hat{\beta}$ est un vecteur aléatoire
- les variables explicatives sont connues

$$y = f(\hat{\beta}, x) + \hat{e}$$

10 / 39

Rappels sur la régression multiple

Définition du modèle : $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$

avec $\mathcal{L}(\varepsilon_i) = \mathcal{N}(0, \sigma)$ et $\text{cov}(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \forall i' \neq i$

Estimation des paramètres : $\hat{\beta} = (X'X)^{-1}X'Y$

$$\text{Critère : } R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Pb : $R^2 \nearrow$ quand nb de variables explicatives \nearrow
 \Rightarrow utilisation du test de significativité du R^2 :

$$\frac{R^2}{1 - R^2} \frac{n - p - 1}{p} \sim F_{n-p-1}^p \text{ sous l'hypothèse } R^2 = 0$$

Quelles données choisir pour l'évaluation ?

- Déterminer le **domaine d'évaluation** du modèle
Sur ce domaine, tests et évaluation ; en dehors, extrapolations
- Données représentatives de tout le domaine, sinon extrapolations
- Données dépendent de l'utilisation du modèle : pour gestion de l'azote, données précisément enregistrées et suffisamment différentes pour cette variable
- Données de validation indépendantes des données ayant servi à estimer les paramètres, sinon :
 - si données nombreuses, les séparer en 2 groupes indépendants : 1 pour l'estimation, l'autre pour l'évaluation
 - si données peu nombreuses : validation croisée ou Jackknife
- Données sur des variables intermédiaires permettent de tester certaines équations ou parties du modèle
- Données nombreuses \Rightarrow évaluation + précise

Description des données

- 321 traitements dont 9 ayant des répétitions
- 2 variétés : Bienvenu, Darmor
- Essais sains : ni maladie, ni ravageurs
- 12 sites-années (10 à 72 traitements par site-année)

Lieu	année	nb d'essais
Cher	84-85	57
	85-86	10
Haute Garonne	85-86	23
	86-87	19
	87-88	19
Indre	84-85	72
	86-87	15
Meurthe et Moselle	84-85	2
	85-86	55
	86-87	40
	87-88	54

Description des données

Variable modèle	Valeur mini	Valeur maxi
Date semis	20 août	15 octobre
Dose semis	47	80
Dose azote automne	0	115
Dose azote printemps	0	220
Dose azote tardif	0	80
Dose azote total	50	395
Profondeur de sol	30	150
Jour floraison	5 avril	14 mai
Poids 1000 grains	3.20	6.00
Rendement observé	1.999	5.132

Gammes de variations importantes sauf pour les doses de semis

Analyse de la variance

Utiliser les répétitions pour estimer la qualité des données
Les traitements donnent-ils des rendements significativement différents ?

$$y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$$

μ effet moyen, α_i effet du traitement i , ε_{ik} la résiduelle

On teste l'hypothèse nulle : $H_0 : \forall i \alpha_i = 0$ contre $H_1 : \exists i \alpha_i \neq 0$

$$\text{Statistique de test : } F_{obs} = \frac{SCM/(I-1)}{SCR/(N-I)} = 2.19$$

Si H_0 vraie, $\mathcal{L}(F_{obs}) = F_{N-I}^{I-1}$

Accepter H_0 signifie que les traitements donnent des rendements similaires (\Rightarrow évaluation sur un domaine très restreint)

$F_{obs} > F_{45}^8(0.95)$ donc rejet de H_0 . Les traitements ont un effet significatif sur les rendements observés

15 / 39

Description du modèle

Modèle statistique utilisé est hiérarchique :

$$y_{ijkl} = f(\hat{\beta}, x_{ij}) + \hat{\varepsilon}_{ijkl}$$

y_{ijkl} : observation

$f(\hat{\beta}, x_{ij})$: prédiction par CECOL

$\hat{\varepsilon}_{ijkl}$: erreur de prédiction du modèle

- i : numéro de la variable explicative prépondérante (le site-année)
- j : numéro de la combinaison des niveaux des variables explicatives utilisées par le modèle (variété, date de semis, 3 apports d'azote, irrigation)
- k : numéro de la combinaison des niveaux des variables explicatives non utilisées par le modèle (fongicide, régulateur de croissance, insecticide, soufre)
- l : numéro de la répétition

16 / 39

Utilisation de l'analyse de sensibilité

AS = étude de sensibilité du modèle aux valeurs de ses paramètres

Buts :

- Comprendre le comportement du modèle
- Déterminer les paramètres influents du modèle
- N'évalue pas directement la qualité du modèle
- Utile pour extension ou mise à jour du modèle

$$\text{On calcule } \kappa = f(\hat{\beta}, x) - f(\hat{\beta} + \eta, x) \quad \text{avec } \eta \ll \hat{\beta}$$

κ grand \Rightarrow paramètre(s) sensible(s) à estimer précisément

η a p composantes non nulles si on évalue la sensibilité à p paramètres simultanément

Résultats de l'AS dépendent des $x \Rightarrow$ explorer tout le domaine d'évaluation. Ex : paramètre de stress hydrique sans effet dans climat pluvieux et déterminant dans climat sec

17 / 39

Simulations en conditions tranchées

Buts :

- Étudier le comportement du modèle dans conditions particulières
- Évaluation qualitative : pas de comparaison aux observations mais aux connaissances générales sur le phénomène étudié
- Simulations en conditions tranchées et AS se ressemblent : ici, variables explicatives varient et non paramètres

Comparaison des prédictions en conditions « normales » et extrêmes : souvent difficile mais la tendance suffit (à dire d'experts)

18 / 39

Simulations en conditions tranchées

Climat	Prof sol (cm)	Date se-mis	Dens se-mis	Qté d'N $u.ha^{-1}$	Date floraison	MS flo $t.ha^{-1}$	Rdt $t.ha^{-1}$
Sans stress	70	15 sep	60	200	07 avr	11.10	4.20
Sans stress	18	15 sep	60	200	07 avr	7.62	3.21
Sans stress	70	31 oct	60	200	12 mai	8.62	3.34
Sans stress	70	2 août	60	200	24 mar	10.39	3.59
Sans stress	70	15 sep	5	200	07 avr	10.91	4.15
Sans stress	70	15 sep	60	0	07 avr	5.88	3.67
Sans stress	18	15 sep	60	0	07 avr	3.22	2.72
Sec	18	15 sep	60	200	13 avr	6.62	2.12
Froid	70	15 sep	60	200	13 mai	3.77	2.92

19 / 39

Construction de courbes prédit/observé

Buts :

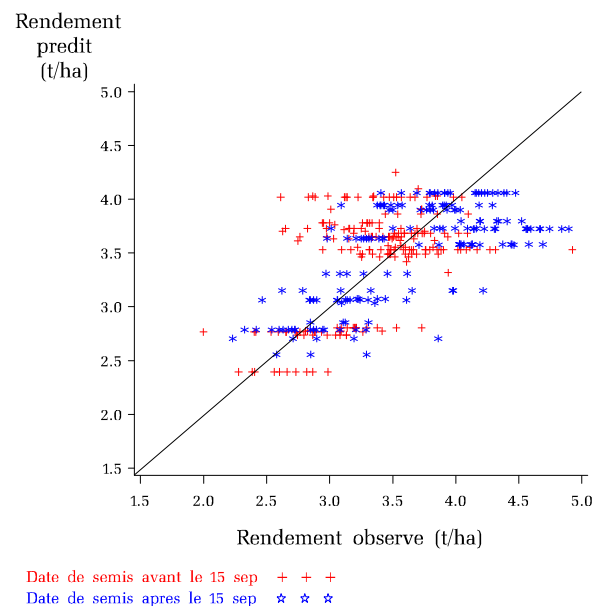
- repérer données aberrantes
- avoir une idée de la réponse du modèle
- Déterminer si l'effet d'une variable particulière est bien pris en compte par le modèle

Avantages :

- Simples à établir
- Possibilité de construction en fonction de variables influentes (date de semis ou doses d'apport d'azote)

20 / 39

Construction de courbes prédit/observé



21 / 39

Construction de courbes prédit/observé

Hypothèse H_0 : « dans la régression des valeurs observées par les valeurs prédites, la pente est de 1 et l'ordonnée à l'origine de 0 »

$$y_{ik} = \beta_0 \times f(\hat{\beta}, x_i) + \beta_1 + \varepsilon_{ik}$$

$$SCR = \sum_{ik} \left(y_{ik} - \hat{\beta}_0 \times f(\hat{\beta}, x_i) + \hat{\beta}_1 \right)^2 = \sum_{ik} \hat{\varepsilon}_{ik}^2$$

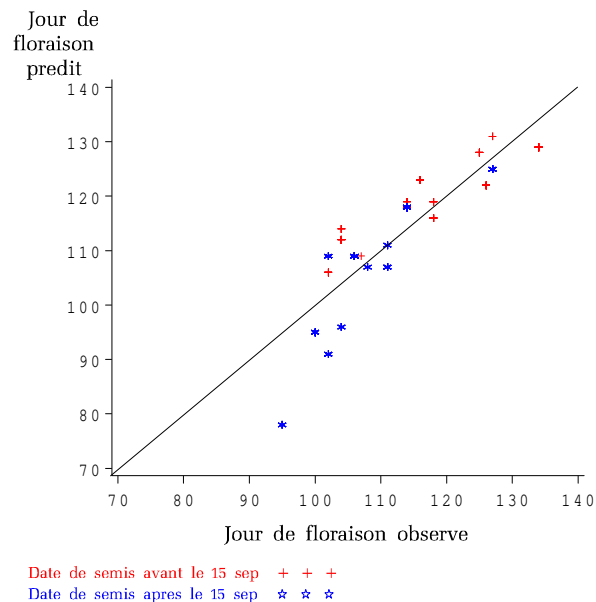
$$\text{Sous } H_0, \beta_0 = 1 \text{ et } \beta_1 = 0 \text{ donc } SCM = \sum_{ik} \left(y_{ik} - f(\hat{\beta}, x_i) \right)^2$$

$$\text{Statistique de test : } U = \frac{(SCM - SCR)/2}{SCR/(n-2)}$$

Sous H_0 , U suit une loi de Fisher $F_{2,n-2}$. On accepte l'hypothèse la pente est de 1 et l'ordonnée à l'origine de 0 si $U < F_{n-2}^2(0.95)$

22 / 39

Construction de courbes prédit/observé



23 / 39

Construction de courbes

But : déterminer si l'effet d'une variable particulière est bien pris en compte par le modèle :

- on ne s'intéresse pas à la justesse des prévisions
- étape importante si utilisation du modèle pour rechercher une conduite optimale

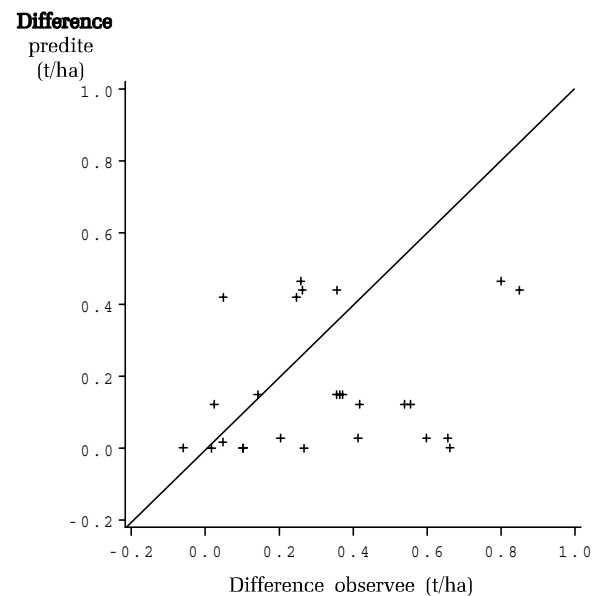
On trace le graphe :

$y_{ijt} - y_{ijs}$ en fonction de $f(\hat{\beta}, x_{ijt}) - f(\hat{\beta}, x_{ijs})$

avec s et t les deux modalités de la variable particulière

24 / 39

Construction de courbes prédit/observé



25 / 39

Critères d'évaluation

Définition	Notation	Equation
Coefficient de détermination	R^2	
Modelling Efficiency	EF	$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$
Moyenne de la valeur absolue de l'erreur	MAE	$\frac{\sum y_i - \hat{y}_i }{n}$
% moyen de la valeur absolue de l'erreur	MA%E	$\frac{100}{n} \sum \left(\frac{ y_i - \hat{y}_i }{ y_i } \right)$
Racine de la moyenne du carré des erreurs	RMSE	$\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$
Erreur quadratique moyen relatif	RRMSE	$\frac{100}{\bar{y}} \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$
Erreur quadratique moyenne de prédiction	MSEP	$\frac{\sum (y_i - \hat{y}_i)^2}{n}$

R^2 et EF proche de 1 \iff modèle très bon

MAE plus robuste que $RMSE$ aux valeurs extrêmes des résidus

Évaluation des qualités prédictives du modèle \implies MSEP

26 / 39

MSEP

Avantages de la MSEP :

- répond à l'objectif d'évaluation du modèle en tant que prédicteur (MSEP mesure la qualité de prédiction et non la qualité d'ajustement aux données du passé)
- signification précise :
 - $MSEP = 0 \iff$ prédiction parfaite
 - $MSEP \gg 0 \implies$ modèle très mauvais prédicteur
- peut servir de critère de comparaison entre modèles
- se décompose en différentes contributions spécifiques de l'erreur
- se calcule simplement si les données d'évaluation sont indépendantes des données d'estimation

$$MSEP = \frac{\sum (y_i - \hat{y}_i)^2}{n} = 0.2241 (t \text{ ha}^{-1})^2$$

Hypothèse : $\hat{\beta}$ estimé par données indépendantes

27 / 39

MSEP du modèle « Moyen »

But : comparer qualités prédictives CECOL et modèle simple

Idée : variance des rendements faible \implies rendement moyen est une bonne estimation du rendement (pb d'extrapolation)

$$y_{ijkl} = \mu + \delta_{ijkl}$$

$$MSEP_M = E((y^* - \hat{\mu})^2) = E((\mu + \delta_{i'j'k'l'} - \hat{\mu})^2)$$

$$\hat{\mu} = \frac{1}{n} \sum_{i,j,k,l} y_{ijkl} \Rightarrow E(\hat{\mu}) = \mu, \quad Var(\hat{\mu}) = \frac{1}{n^2} n \sigma_\delta^2 = \frac{1}{n} \sigma_\delta^2$$

$$MSEP_M = Var(\hat{\mu}) + \sigma_\delta^2 = \frac{n+1}{n} \sigma_\delta^2$$

$$\widehat{MSEP}_M = 0.3325 (t \text{ ha}^{-1})^2$$

$$\widehat{MSEP}_{Cecol} = 67\% \widehat{MSEP}_M$$

28 / 39

Décomposition de la MSEP

But : Déterminer d'où provient l'essentiel de l'erreur

$$MSEP(\hat{\beta}) = \Lambda + \Delta(\hat{\beta}) + \Gamma(\hat{\beta})$$

Λ : MSEP minimum qui peut être atteinte avec les variables explicatives utilisées

$\Delta(\hat{\beta})$: contribution à la MSEP du biais du modèle

$\Gamma(\hat{\beta})$: contribution à la MSEP des erreurs de mesure dans données d'entrée

29 / 39

Estimation de Λ

$$\hat{\Lambda} = \frac{\sum_{i,j} \sum_{k,l} (y_{ijkl} - \bar{y}_{ij..})^2}{\sum_{i,j} (n_{ij} - 1)} = 9.84 \cdot 10^{-2} (t \cdot \text{ha}^{-1})^2$$

avec n_{ij} le nombre de traitements du groupe i ayant les variables explicatives j

Λ = variabilité intra-individu + variabilité due aux variables non utilisées par le modèle

2^e terme est important \implies certaines variables non utilisées par le modèle ont des effets significatifs sur les observations et devraient être incorporées dans le modèle

30 / 39

Estimation de Γ

Données initiales (% MO, % NH₄, T°, ...) + ou - précises :

$$\hat{x} = x + \varepsilon_x$$

x : vraie valeur, ε_x : erreur de mesure i.i.d., $\varepsilon_x \sim \mathcal{N}(0, \sigma_x)$

Γ estime, par techniques de Monte Carlo, l'effet de ces approximations sur les prédictions :

- Définir les distributions des mesures initiales (σ_x)
- Par traitement, calculer la prédiction du modèle pour V ensembles de données initiales \hat{x} provenant de simulations
- Par traitement, estimer la variance des V prédictions
- Calculer la moyenne des variances

31 / 39

Variabilité des variables d'entrée

Table – Coefficient de variation des erreurs de mesure des variables explicatives

Variables d'entrés	Coefficient de variation ($\frac{\sigma}{\mu}$)
Profondeur du sol	0.3
Concentration de matière organique dans le sol	0.2
Capacité au champ	0.2
Point de flétrissement	0.2
Point de saturation	0.2
Quantité de nitrate initiale	0.4
Quantité d'ammonium initiale	0.4
Quantité d'eau initiale dans le sol	0.4
Quantité d'azote apportée	0.02
Température	0.00
Rayonnement	0.00
Pluviométrie	0.00

32 / 39

Estimation de Γ

$$\hat{\Gamma}(\hat{\beta}) = \frac{1}{J} \sum_{i,j} \frac{1}{V-1} \sum_{v=1}^V \left(f(\hat{\beta}, x_{ijv}) - \overline{f(\hat{\beta}, x_{ijv})} \right)^2 = 2.23 \cdot 10^{-2} (t.ha^{-1})^2$$

J nb de traitements, $f(\hat{\beta}, x_{ijv})$ prédiction pour le traitement ij et le v^e ensemble de valeurs de \hat{x} et $\overline{f(\hat{\beta}, x_{ijv})}$ moyenne des V prédictions $f(\hat{\beta}, x_{ijv})$

Rq : $\Gamma(\hat{\beta})$ peut être estimé sans observation et est estimable pour toutes les variables de sortie du modèle

33 / 39

Estimation de Δ

La contribution du biais sur la MSEP est estimée par différence :

$$\hat{\Delta}(\hat{\beta}) = \widehat{MSEP}(\hat{\beta}) - \hat{\Lambda} - \hat{\Gamma}(\hat{\beta}) = 10.34 \cdot 10^{-2} (t.ha^{-1})^2$$

34 / 39

Décomposition de la MSEP

L'essentiel de l'erreur provient :

- si Λ important \Rightarrow du choix des variables explicatives
- si $\Gamma(\hat{\beta})$ important \Rightarrow d'erreurs de mesure dans variables d'entrée
- si $\Delta(\hat{\beta})$ important \Rightarrow provient de la forme du modèle

La MSEP se décompose en 4 termes :

- variabilité dans les mesures des variables observées : indépendant du modèle
- contribution des variables explicatives du modèle : identique pour tout modèle utilisant les mêmes variables d'entrée
- contribution des erreurs de mesure des variables d'entrée : peut diminuer si les variables d'entrée sont plus précises
- biais du modèle

35 / 39

Récapitulatif évaluation de CECOL

Source de variation	Espérance ($t \text{ ha}^{-1}$) ²	Variance ($t \text{ ha}^{-1}$) ⁴
Variance totale	0.2886	5.21 10^{-4}
Biais moyen de CECOL (en $t \text{ ha}^{-1}$)	$-7.94 \cdot 10^{-3}$	
Effet erreurs de mesure sur variables d'entrée	$2.23 \cdot 10^{-2}$	$2.13 \cdot 10^{-8}$
Biais du modèle	0.1034	
Effet des variables explicatives	$9.84 \cdot 10^{-2}$	$4.31 \cdot 10^{-4}$
MSEP de CECOL	0.2241	$9.25 \cdot 10^{-4}$
MSEP du modèle «Moyen»	0.3325	

36 / 39

Démarche si les données de validation ont servi à la construction du modèle

- Analyse de sensibilité
- Simulations en conditions tranchées
- Construction de graphes
Attention : donnent une idée optimiste de la qualité du modèle surtout si le nombre de paramètres estimés est important
- Calcul de la MSEP par validation croisée :

$$\widehat{MSEP} = \frac{1}{n} \sum_{i=1}^n \left(y_i - f(\hat{\beta}_{-i}, x) \right)^2$$

Inconvénient : temps de calcul très longs

- Comparaison de modèles par MSEP
- Décomposition de $MSEP(\hat{\beta})$

$$MSEP(\hat{\beta}) = \Lambda + \Delta(\hat{\beta}) + \Gamma(\hat{\beta})$$

Λ , $\Gamma(\hat{\beta})$ et $\Delta(\hat{\beta})$ sont estimés comme précédemment

37 / 39

Discussion

Démarche proposée permet d'évaluer :

- modèle simple ou complexe
- modèle linéaire, non linéaire, statique ou dynamique

Hypothèse : les données doivent être indépendantes

Démarche inadaptée aux prédictions de valeurs successives dans le temps (par exemple mesure de MS) : dans ce cas, observations non indépendantes

Méthodes d'évaluation présentées sont complémentaires \Rightarrow **les utiliser toutes**

38 / 39

Discussion

CECOL : meilleur prédicteur que la simple moyenne des observations du passé pour prédire les rendements futurs, mais avantage seulement modéré

Valeur de Λ faible : le jeu de variables explicatives utilisé par CECOL semble bon

Valeur de Δ important : biais du modèle important \implies modèle peut être amélioré en modifiant ses équations ou en réestimant certains paramètres

Améliorations du modèle possibles :

- date de floraison mal prédite : trop précoce pour semis précoces et trop tardive pour semis tardifs \implies identifier la cause exacte de ce comportement
- Effets de l'irrigation mal pris en compte