Evaluer l'importance de différentes sources d'incertitude lors de la validation d'un modèle de simulation

Eric Gozé – UPR Systèmes de culture annuels

1 Introduction – Notations
    Modèle
    MSEP

2 Décomposition de l'erreur sur la fonction f par linéarisation locale.
    Propagation de l'erreur sur les entrées
    Propagation de l'erreur sur les paramètres

3 Décomposition de la variance de l'écart de prédiction

4 Conclusion : acquisition contrôlée et étude de sensibilité
sont les éléments du calcul d'incertitude + Delicas.

1 Introduction - Notations

Estimation expérimentale de l'amplitude de l'erreur de prédiction :
sur échantillon au hasard parmi les situations possibles
n'ayant pas déjà servi pour l'estimation des paramètres.

Répétitions (ou plutôt répliques) locales, permettant une estimation des variances et covariances entre les différentes variables mesurées, les entrées et les sorties du modèle.

Une observation y d'une sortie du modèle s'écrit :

$$y = f(\underline{x},\theta) + \varepsilon$$

avec x le vecteur des entrées,
θ le vecteur des paramètres,
f le modèle
$f(\underline{x},\theta)$ la prédiction,
ε l'écart de prédiction

Critère global pour l'incertitude : *Mean square error of prediction* MSEP = E(ε^2)

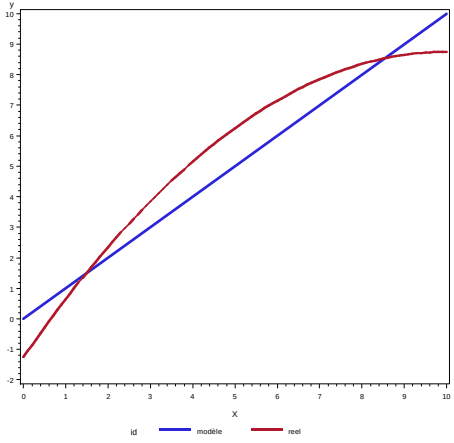
MSEP = **biais** ² + variance ; mais de quel biais veut-on parler ?

Le modèle est une simplification de la réalité :

pour un x donné, le modèle est biaisé, et le biais dépend de x

$$y = f(\underline{x},\theta) + b(\underline{x},\theta) + \varepsilon'$$

$b(\underline{x},\theta)$ est l'erreur de modèle,
ε' l'erreur de mesure de la sortie



sur l'ensemble des situations dont notre échantillon est représentatif, x est aléatoire, l'erreur de modèle aussi

Erreur sur y = erreur sur la fonction f + erreur de modèle + erreurs de mesure et de randomisation

2 **Décomposition de l'erreur sur la fonction f par linéarisation locale**.

Analyse locale d'incertitude, calcul de l'effet de petites variations des entrées et des paramètres, par développement en série de Taylor à l'ordre 1 :

$$\Delta y \approx \Delta x \frac{\partial y}{\partial X} + \Delta\theta \frac{\partial y}{\partial \theta}$$

Cela suppose que :

- les variations des entrées x et des paramètres θ soient suffisamment petites pour que les variations de y soient linéaires en fonction de celles-ci.

- les effets des variations de x et θ soient additifs. Autrement dit, qu'il n'y ait pas d'interaction entre entrées, entre paramètres, et entre entrées et paramètres.

2.1 **Propagation de l'erreur sur les entrées**

Une entrée fixe : $\Delta y \approx \Delta x \frac{\partial y}{\partial x}$

p entrées fixes : $\Delta y \approx \sum_{j=1}^{p} \frac{\partial y}{\partial x_j} \Delta x_j = \left(\frac{\partial y}{\partial x}\right)' (\Delta x)$

produit scalaire entre le vecteur des erreurs et celui des sensibilités à ces erreurs
n.b. on néglige les interactions
n.b. des erreurs peuvent se compenser

Une entrée aléatoire : $\mathrm{Var}(y) \approx \mathrm{Var}(x) \left(\frac{\partial y}{\partial x}\right)^2$

p entrées aléatoires de variance-covariance Σ_x: $\mathrm{Var}(y) \approx \left(\frac{\partial y}{\partial x}\right)' \Sigma_x \left(\frac{\partial y}{\partial x}\right)$

Conclusion : les erreurs sur les entrées se propagent sur les sorties via la sensibilité.
Des erreurs peuvent se compenser ou non : important de connaître leur covariance

2.2 **Propagation de l'erreur sur les paramètres**

de la même façon que l'erreur sur les entrées, via la sensibilité

Biais : $\Delta y \approx \sum_{j=1}^{p} \frac{\partial y}{\partial \theta_j} \Delta\theta_j = \left(\frac{\partial y}{\partial \theta}\right)' (\Delta\theta)$

Variance : $\mathrm{Var}(y) \approx \left(\frac{\partial y}{\partial \theta}\right)' \Sigma_\theta \left(\frac{\partial y}{\partial \theta}\right)$

Les erreurs sur les paramètres sont souvent corrélées quand plusieurs paramètres sont estimés à partir d'une seule et même expérience.

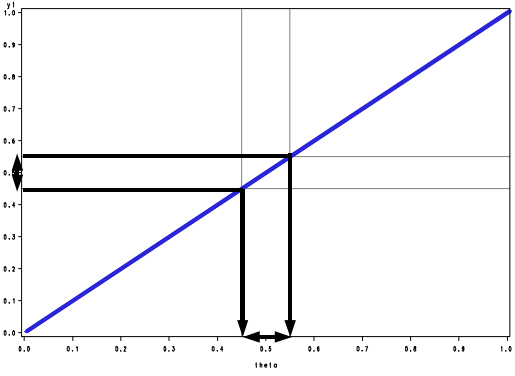
Que vaut l'erreur sur les paramètres ?

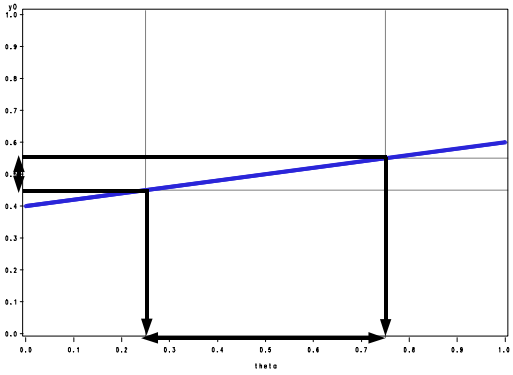
Dépend de la méthode d'estimation :

à dire d'expert : erreur ???? : incertitude à estimer à dire d'expert aussi ???

d'après la littérature : faire suivre la variance

par « optimisation » : on a quelques résultats théoriques





Que vaut l'erreur sur les paramètres ? (suite)

Hypothèses : erreurs indépendantes (=> randomisation)
identiquement distribuées

fonction f deux fois dérivable par rapport aux paramètres
(pas de saut)

Posons $$Z = \left(\frac{\partial \mathrm{f}(x_i, \theta_j)}{\partial \theta_j} \right),$$

matrice des sensibilités locales à θ_j au point x_i .

On peut calculer la variance de l'estimateur $\hat{\theta}$ de θ:

$$\hat{\Sigma}_\theta = \sigma^2 (Z'Z)^{-1}$$

avec σ^2 = variance de l'erreur expérimentale.

Grande sensibilité ou faible erreur expérimentale => faible erreur sur θ, mais aussi…

…**un bon plan d'expérience minimise la variance des erreurs sur θ : plans optimaux**

Pour un modèle linéaire : $y = X\beta + \varepsilon$,

le vecteur des paramètres est β
la matrice des sensibilités est X

la variance de l'estimateur de β est $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$

θ et Z en sont les pendants pour un modèle non-linéaire

Pour un modèle linéaire, on sait faire des expériences dites optimales = qui prennent le moins possible d'observations pour une précision donnée.

Exemple : le plan D-optimal minimise le déterminant de $(X'X)^{-1}$

Ce n'est pas le même plan qui assure la représentativité des situations *x*

L'erreur sur les observations des expériences de calibrage se propage sur les paramètres, puis sur les sorties.

3 **Décomposition de la variance de *y***

$$
\begin{aligned}
y &= \mathrm{f}(x,\vartheta) + b(x,\vartheta) + \varepsilon' \\
&= \mathrm{f}(x_{obs},\hat{\vartheta}) + \left(\mathrm{f}(x,\vartheta) - \mathrm{f}(x_{obs},\hat{\vartheta})\right) + b(x,\vartheta) + \varepsilon' \\
\mathrm{E}(y - \mathrm{f}(x_{obs},\hat{\vartheta})) &= \left(\frac{\partial y}{\partial \theta}\right)' \mathrm{E}(\vartheta - \hat{\vartheta}) + b(x,\vartheta) + 0 \\
\mathrm{var}(y - \mathrm{f}(x_{obs},\hat{\vartheta})) &= \left(\frac{\partial y}{\partial \theta}\right)' \Sigma_\theta \left(\frac{\partial y}{\partial \theta}\right) + \left(\frac{\partial y}{\partial x}\right)' \Sigma_x \left(\frac{\partial y}{\partial x}\right) + \mathrm{var}(b) + \mathrm{var}(\varepsilon')
\end{aligned}
$$

Connaissant les variances des erreurs expérimentales sur les entrées et les sorties par analyse de variance, et leur propagation via les sensibilités, on peut déduire par différence la variance de l'erreur de modèle.

4 **Conclusion** :

La modélisation ne met pas à l'abri des corrélations spatiales, des erreurs expérimentales, des confusions d'effets. Les **bons** (vieux) **dispositifs** que nous utilisons sont là pour en limiter les effets.

Un **plan optimal** pour le **calibrage** est un choix de traitements ou de dates d'observation maximisant l'information sur les paramètres ; la **validation**, au contraire, se fait sur un échantillon représentatif des situations pour estimer la variance de l'erreur de modèle.

La sensibilité locale par rapport aux entrées et par rapport aux paramètres est centrale dans ces calculs de propagation d'erreur.

Pb : y n'est pas forcément localement continue, dérivable et linéaire en fonction de θ au voisinage du vrai θ .

Les **corrélations entre erreurs** sur les paramètres et sur les entrées sont importantes dans cette analyse locale d'incertitude : elles doivent le rester dans une analyse globale.

Projet Anr Delicas : association mapping et phénotypage par modèle pour l'identification de marqueurs moléculaires liés à l'élaboration et à la limitation du rendement chez la canne à sucre

Phénotypage par modèle : expérimentation variétale en vue d'estimer des paramètres de modèle.

Les paramètres sont-ils différents d'une variété à l'autre ?
Quel impact de l'imprécision des paramètres sur les prédictions de l'interaction GxE ?

1ère année : estimer les paramètres par maximum de vraisemblance, avec leur variance-covariance, sur un échantillon de variétés

élaborer un plan optimal pour la deuxième année afin d'alléger les mesures sur les autres variétés