Evaluation des modèles.

François Brun
(contributions de D.Wallach)

Formation Introduction à la modélisation pour la protection intégrée des cultures.
30 novembre au 4 décembre 2015

Objectifs

- Qu’est-ce qu’évaluer un modèle ?
- Mieux comprendre la différence entre différents usages des modèles (hypothèses scientifiques vs outils d'ingénierie)
- Quantifier l'ajustement du modèle aux données
- Faire la différence entre l’ajustement aux données et la qualité de prédiction
- Estimer la qualité de prédiction

Évaluation des modèles et outils

- **Validation analytique**
  - On va se concentrer là-dessus

- **Validation par l'usage**
  - Adéquation de l'outil à l'usage, pour répondre à la question
  - Facilité l'utilisation (l'ergonomie, échanges de données,…)
  - Pas de méthodes standardisées
  - Recherche d'indicateurs pertinents
  - Démarche de co-conception

Validation analytique : définitions

- **Vérification**
  - le formalisme du modèle est correct
    - Logique (Analyse de dimension, conservation matière,…)
    - Programmation : correspondance avec ce que l'on souhaite
- **Calibration**
  - estimation des paramètres du modèle
  - limites
- **Validation**
  - Précision suffisante par rapport aux objectifs du modèle ?
  - Définition d'un domaine de validité ?

(Rykiel, 1996; Sinclair et Séligman, 2000)

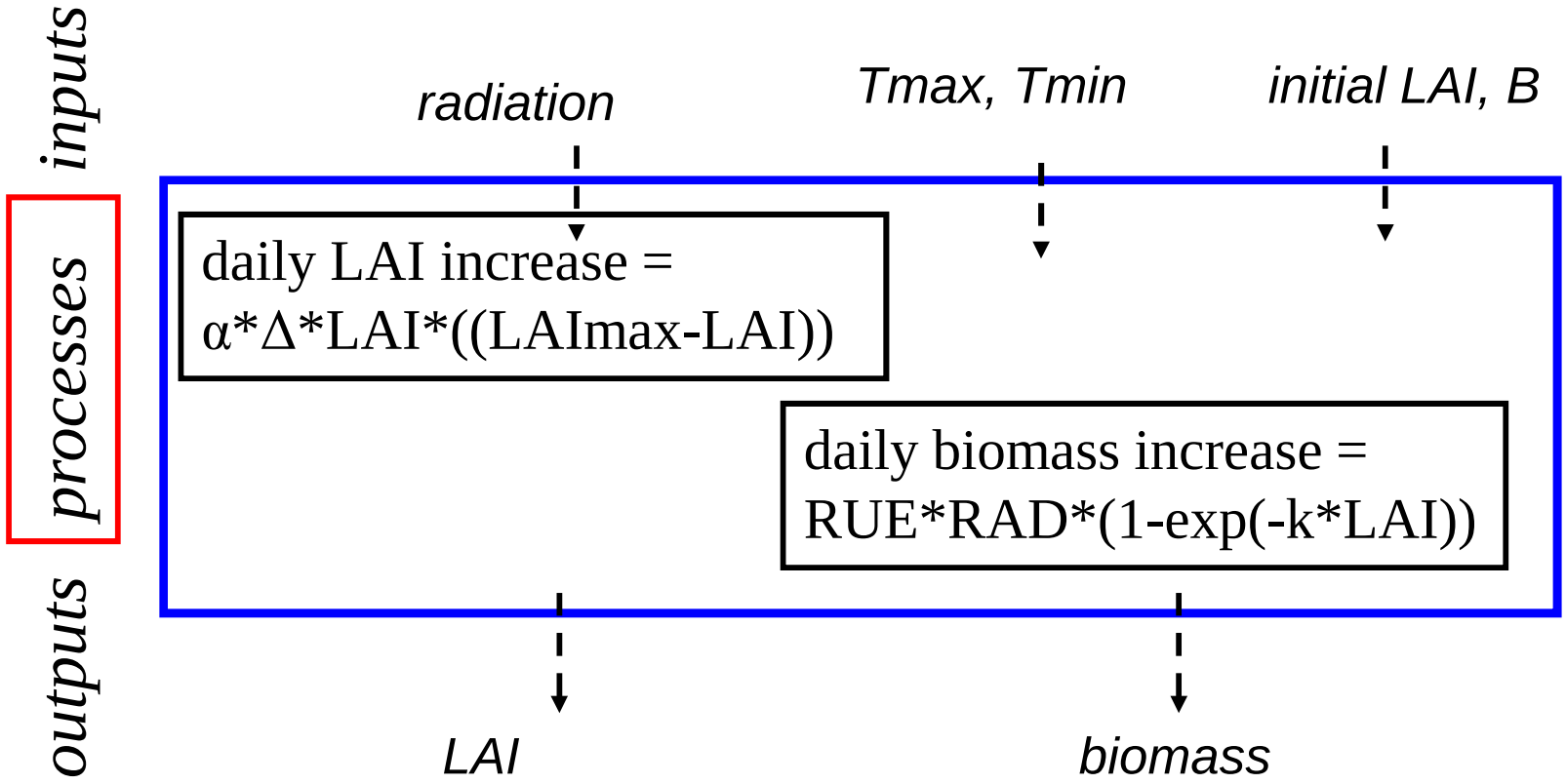
Validation analytique

- Un modèle…
  - « représentation simplifié du monde réelle »
  - manque de connaissance sur le déterminisme de certains processus
  - Disponibilité et qualité des données pour paramétrage
- Sources d'erreurs
  - Le **système considéré** dans le modèle (processus pris en compte et limite du système)
  - Les **formalismes** décrivant les processus (équations)
  - Les **paramètres** des équations
  - L'erreur sur les **variables d'entrée** (météo, sol, conditions initiales,…)

En fonction du type de modèle

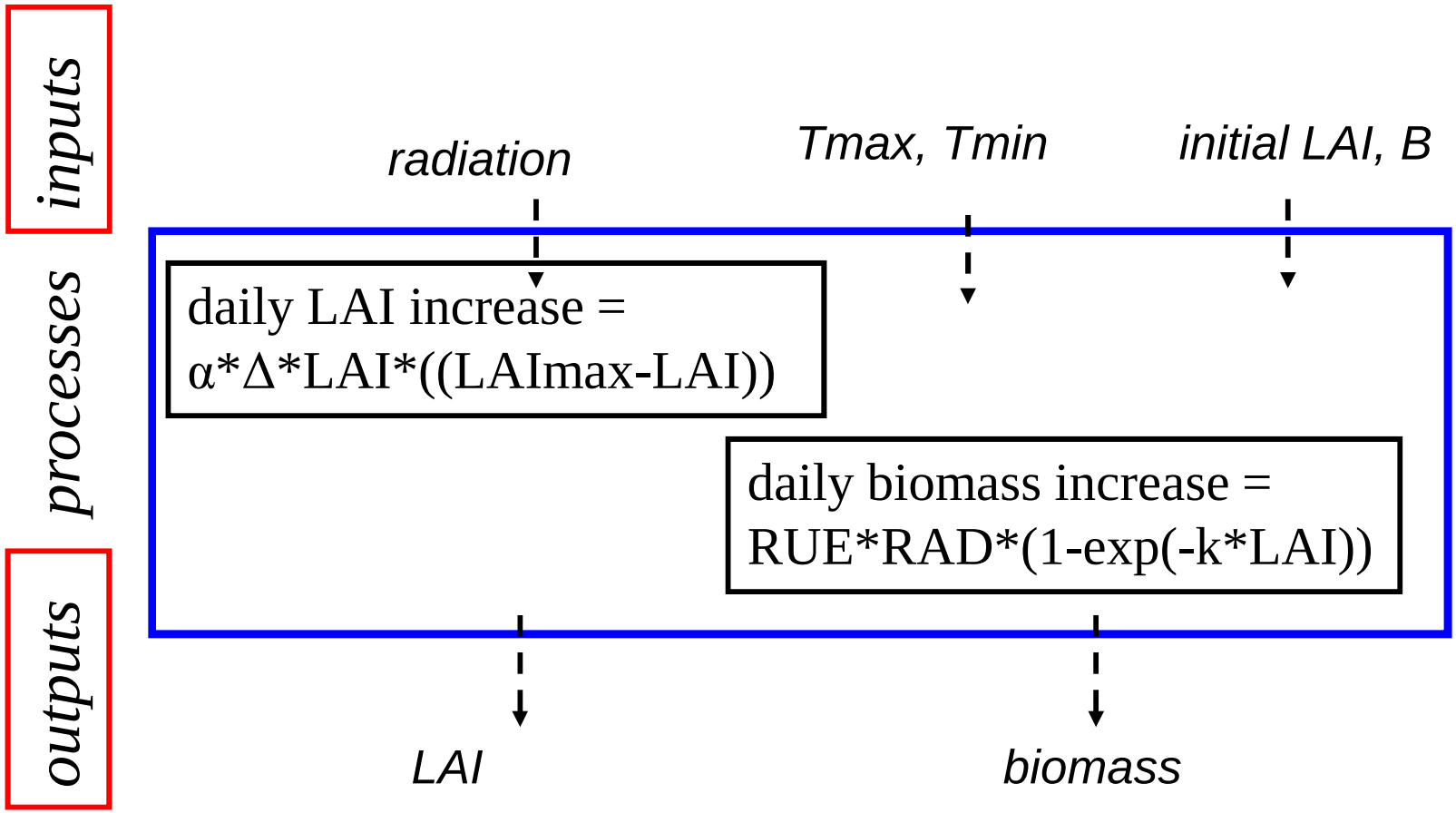
Un modèle : une hypothèse scientifique

- L'hypothèse : la description des processus du système, et leurs interactions, est correcte.
  - Intérêt pour ce qui se passe dans le modèle
  - Description valide ou pas : OUI ou NON



Un modèle : un outil d'ingénieur

- A quel point il se comporte comme on le souhaiterait ?
  - Exemple : est-ce que les prédictions sont bonnes ?
  - On s'intéresse à la qualité des relations entre entrée et sorties
  - Le résultat est, en général, un nombre (l'erreur de prédiction)



Un modèle : une hypothèse scientifique

- Comparer les résultats de simulation à des données.
- Sur cette comparaison, on veut tirer une conclusion sur le fait que les processus inclus dans le modèle ressemble aux véritables processus.

Problème 1. Impossible de prouver que le modèle est exact

- SI les résultats du modèle diffèrent franchement aux résultats d'observations, ALORS l'hypothèse est fausse.
- SI les résultats du modèle ressemblent aux résultats d'observations, ALORS cela ne prouve pas que le modèle est valide.
  - peut ressembler sans que la représentation soit juste
  - peut ressembler dans ces conditions, mais pas d'autres

⇒Comme une théorie, un modèle ne peut que être invalidé, et jamais être validé définitivement.

⇔ Karl Popper

Problème 2. le modèle est faux

- Car un modèle reste toujours une simplification de la réalité.

- On l’a construit ainsi.

Est-ce inutile de considérer le modèle comme une hypothèse scientifiques ?

- Pas vraiment inutile, mais besoin d'être un peu moins ambitieux…
- Tester si le modèle donne des résultats similaires aux observations
- C'est un résultat utile, si :
  - L'hypothèse est innovante
  - Comparer les hypothèses concurrentes

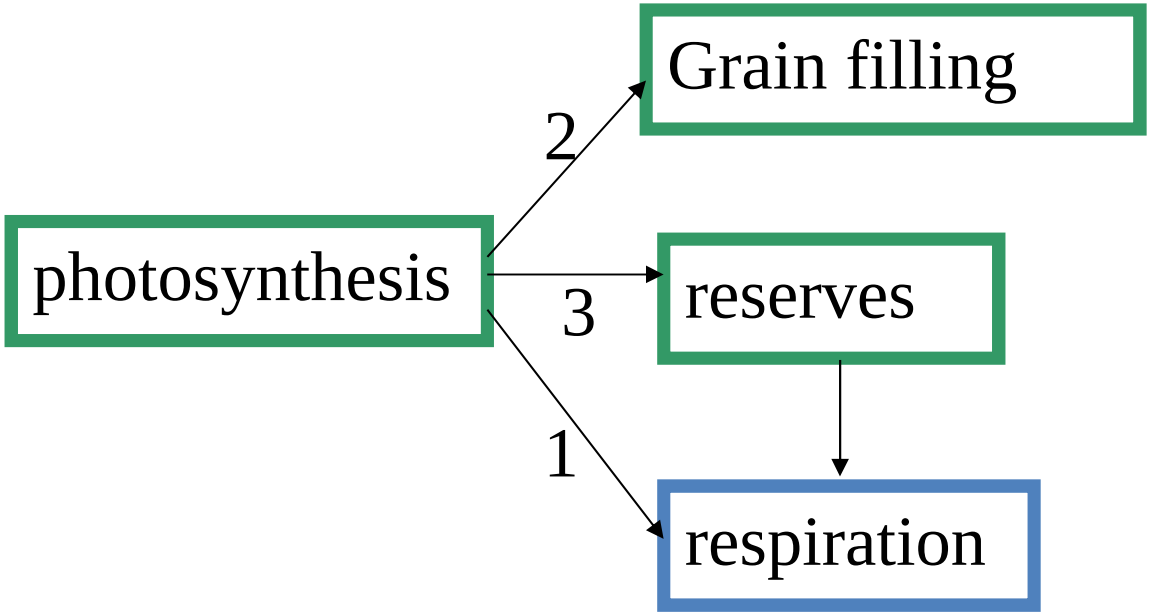
Exemple d’une comparaison de modèles

- Quelle est la source de carbone utilisée pour la respiration pendant le remplissage du grain de blé ?

$\Rightarrow$Tester deux modèles différents, qui correspondent à deux hypothèses différentes.

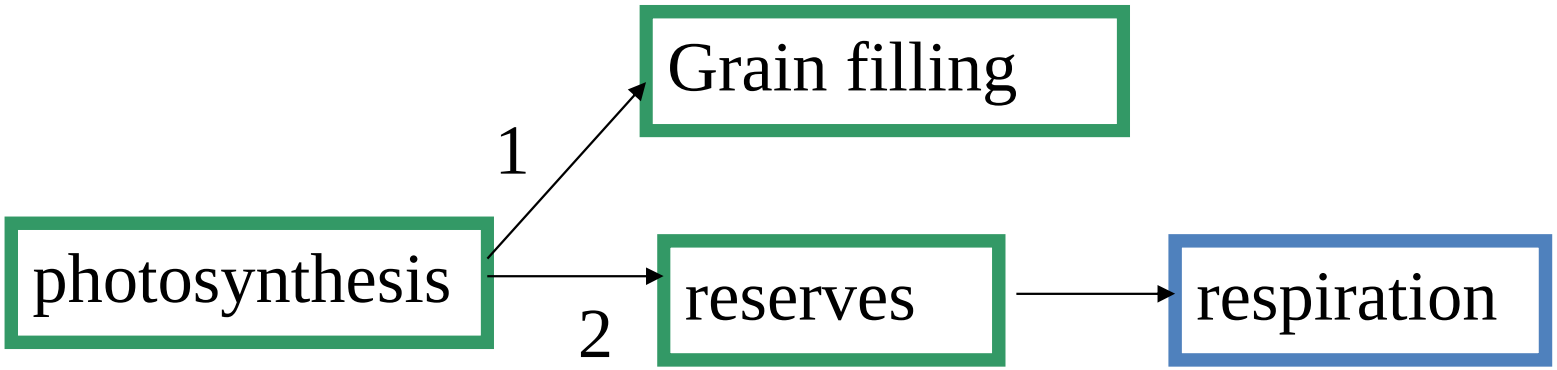
Hypothèse 1 : le Carbone pour la respiration vient directement de l'assimilation.

- Entrée : période d'application de $^{14}C0_2$ (marqué)
- Sortie : concentration du ^{14}C dans le $C0_2$ respiré et dans le grain
- Développer un modèle dynamique 1 basé sur l'hypothèse 1



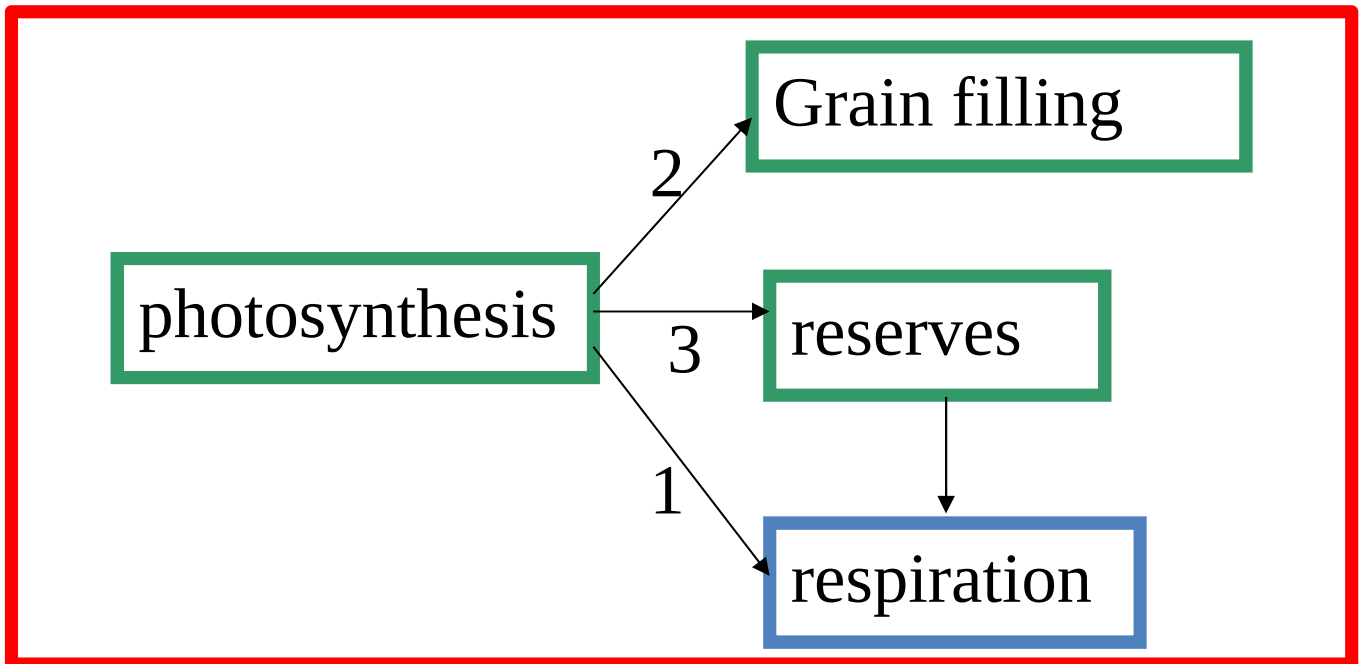
Hypothèse 2 : le Carbone vient des réserves

- Entrée : période d'application de $^{14}C0_2$ (marqué)
- Sortie : concentration du ^{14}C dans le $C0_2$ respiré et dans le grain
- Développer un modèle dynamique 2 basé sur l'hypothèse 2



Experimentation

- Entrée : application de $^{14}C0_2$ (marqué) pour une durée limitée
- Sortie : mesure de la concentration du ^{14}C dans le $C0_2$ respiré et dans le grain (même variable que dans le modèle)
- Comparaison des sorties du modèle et des expérimentations
- Résultats : le modèle basé sur l'hypothèse 1 présente des résultats plus proche des observations



- Est-ce que l'on peut conclure que l'hypothèse 1 est correcte ?
  - Non, car on ne peut pas valider un modèle/une théorie
  - De tout façon, les résultats ne sont pas strictement semblables
- Que conclure ?
  - L'hypothèse 1 est une meilleure description de la respiration que l'hypothèse 2.
  - Un modèle a amélioré notre compréhension des processus sous-jacent

Un cas complexe

- On compare deux modèles complexes
- Si le modèle 1 donne des résultats plus proche des observation pour un ensemble d'expérimentations
- Peut-on conclure que les processus représentés dans le modèle 1 constituent une meilleure représentation de la réalité ?
  - Non, certains processus peuvent être mieux représentés, d'autres moins bien.
  - On ne peut pas tirer des informations processus par processus.

- Utiliser un modèle comme une hypothèse scientifique est en général utile pour des modèles relativement simple
  - Tester une seule hypothèse
- En général, pas pertinent pour les modèles agronomiques

Un modèle : un outil d’ingénieur.

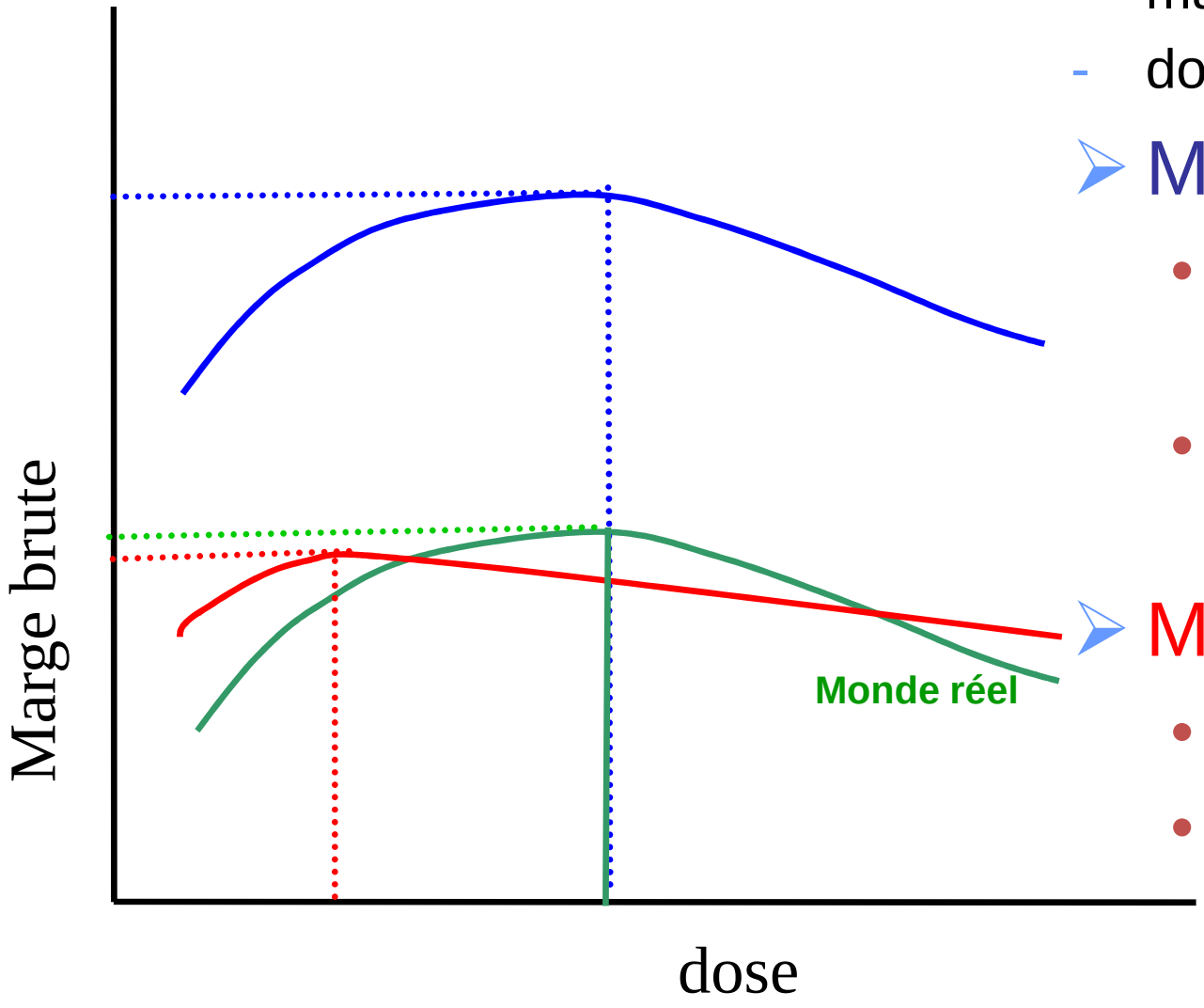
Evaluer les performances du modèle.

L’évaluation est essentielle

- Pour le modélisateur
  - Si l’on ne peut pas évaluer un modèle, on ne peut pas progresser
- Pour l’utilisateur
  - Comment prendre des décisions, si l’on n’a pas d’information sur la fiabilité des informations ?

En fonction de l'objectif : variables d'intérêt à bien préciser

- Prédiction vs Décision



- 2 variables
  - marge
  - dose optimale
- Modèle 1 (bleu)
  - Très mauvais en prédiction
  - Mais OK pour décision
- Modèle 2 (rouge)
  - bon en prédiction
  - Mais mauvais pour décision

L’évaluation prend différentes formes

- À quel point le modèle prédit bien ?
  - Surface foliaire, Sévérité de la maladie, …
  - Est-ce que les décisions basées sur le modèle sont bonnes ?
  - Comparaison de différents scénarios, de pratiques par exemple, pour choisir la meilleur combinaison de pratique
- A quel point le modèle classe bien ?
  - Est-ce que le niveau d’une maladie est au dessus ou au dessous d’un seuil ?

Qualité d'ajustement.
A quel point le modèle représente bien les données disponibles?

Le point de départ de l'évaluation…

Types de comparaisons

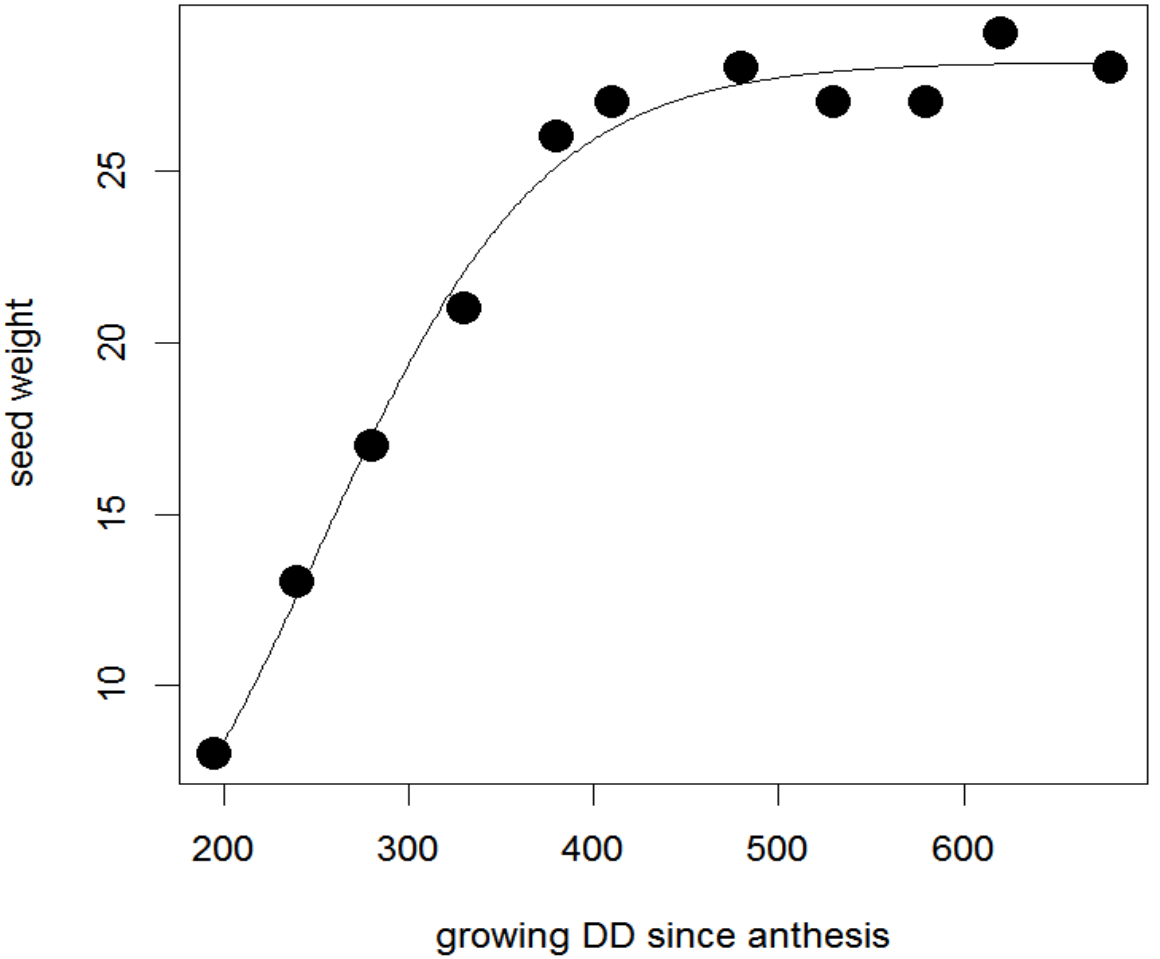
- Graphiques
  - Pour représenter la concordance entre modèle et observation
- Mesures quantitatives
  - Résumer cette concordance avec un indicateur numérique
- Mesures de performance
  - Comparer le modèle avec un estimateur naïf
- Décomposition en différents termes
  - Mieux comprendre l'origine des discordances

Comparaisons graphiques

Variables dynamiques

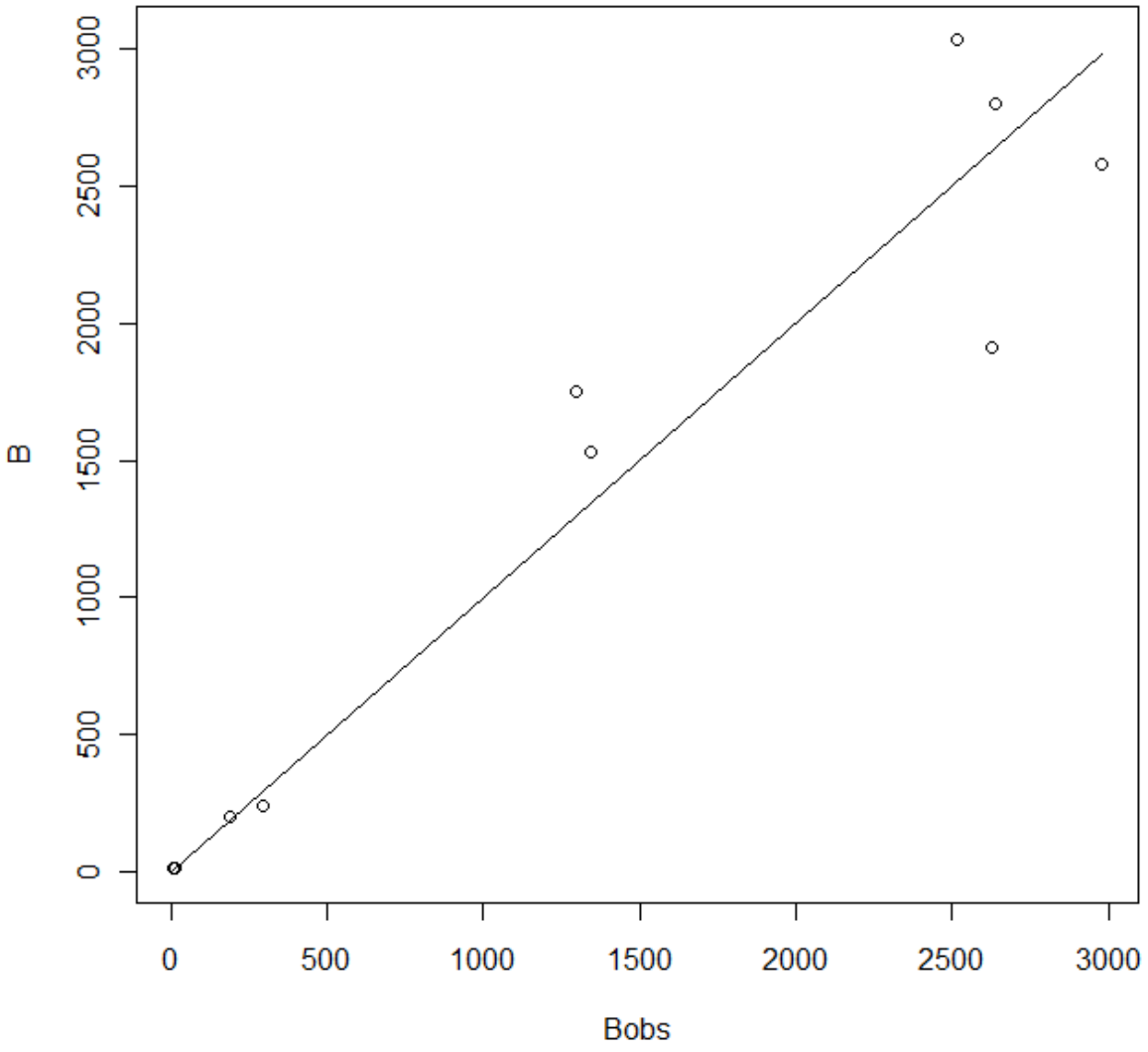
- représenter les valeurs simulées et observées sur le même graphique

- Convention
  - ligne : modèle
  - points : observation



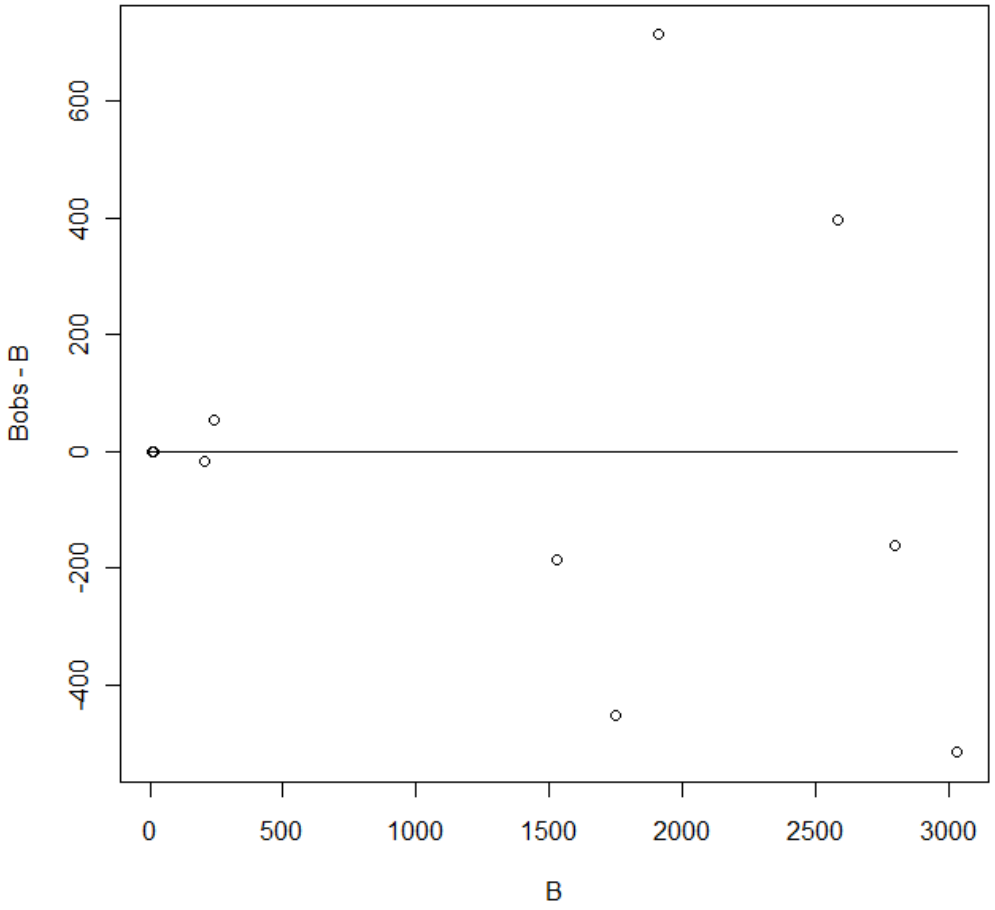
Observations en fonction des prédictions

- La ligne diagonale représente l'égalité entre les valeurs observées et valeurs prédites



Graphique des résidus (observation-prediction)

- Résidus en fonction des valeurs prédites
- On peut aussi faire des graphiques en fonction d'autres variables (par exemple, des indicateurs météorologiques)
- C'est important aussi pour vérifier les hypothèses statistiques (Variance constante et indépendance)



Mesures quantitatives de l’adéquation entre observations et simulations

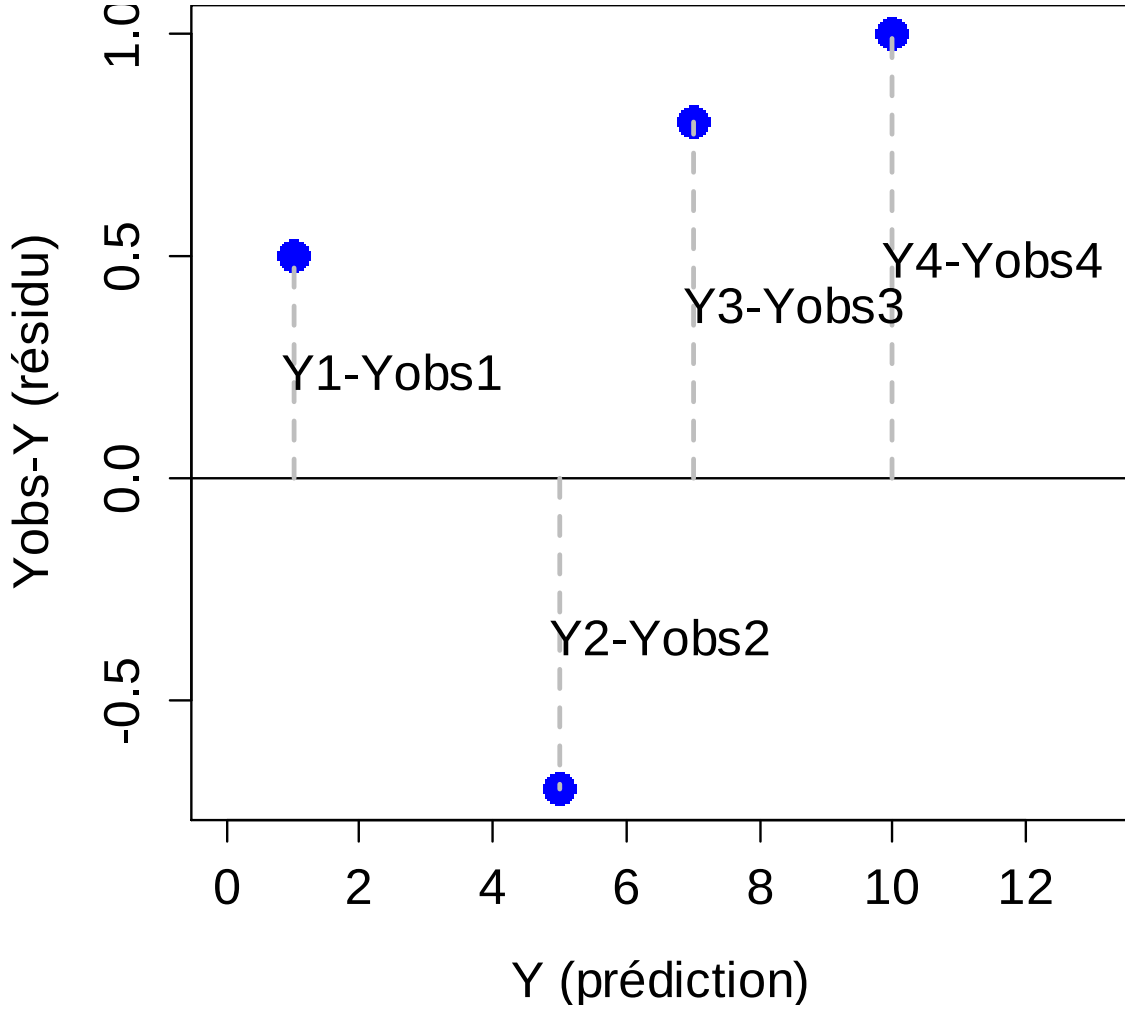
Mean squared error (MSE)

- moyenne des carrés des erreurs
- Probablement le plus courant et populaire

$$MSE = (1/N) \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

Yobs	Y	résidus	résidus2
1.5	1	0.5	0.25
4.3	5	-0.7	0.49
7.8	7	0.8	0.64
11.0	10	1.0	1.00

MSE = 0.595



Root mean squared error (RMSE)
Relative RMSE (RRMSE)

$$RMSE = \sqrt{MSE}$$

$$RRMSE = RMSE / \bar{y}$$

- RMSE : même unité que la variable Y, plus facile à interpréter.
- RRMSE : sans unité. Utile pour évaluer la multi-performance d'un modèle (différentes variables de sortie)

Mean absolute error (MAE)

$$MAE = (1/N) \sum_{i=1}^{N} \left| Y_i - \hat{Y}_i \right|$$

- Avantage par rapport à RMSE?
- Pourquoi RMSE reste plus largement utilisé ?

Critères de performance

Efficience

$$EF = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \overline{Y})^2}$$

- Si le modèle est parfait, alors :
  - EF=1
- EF < 1
- On considère le modèle naïf : toute les prédictions égales à la moyenne ($\overline{Y}$)
  - EF=0
  - Si EF est égale à 0, alors il n'est pas plus performant que de prendre la moyenne.
  - On peut aussi avoir EF < 0…

Efficience

$$EF = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

- Si EF est proche de 1, alors le modèle est performant (explique bien la variabilité des observations disponibles)
- Si EF est proche de 0, ou négatif, alors le modèle n'est pas mieux qu'une prédiction naïve

Efficience

$$EF = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{MSE}{\mathrm{var}(Y)(N-1)/N}$$

- dépend de l'erreur (variabilité de la réponse non expliquée par le modèle) : MSE
- relativement à la variabilité totale de la réponse dans le jeu de données considérés.
- ⇔ définition similaire au R^2
- EF dépend de manière monotone au MSE, mais inclus une comparaison à un modèle naïf (la moyenne)

Décomposition de l’erreur

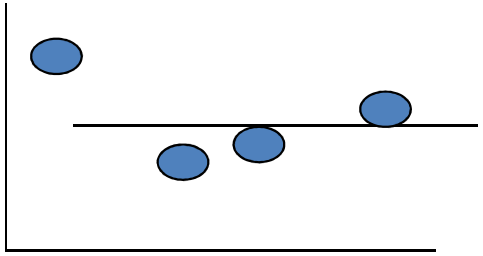
- Mieux décerner les contributions aux erreurs
- Suggestions de pistes d’amélioration

MSE=biais² + (différence entre les écart-types)² + résiduel

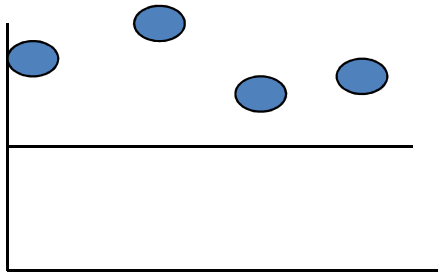
Biais au carré

$$\text{biais}^2 = \left[(1/N) \sum y_i - (1/N) \sum \hat{y}_i \right]^2$$

Peu de biais



Biais important

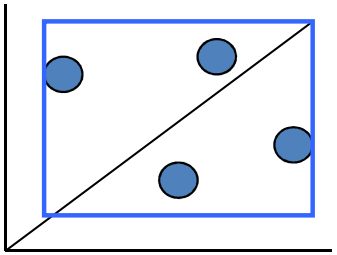


- Si ce terme est important :
  - Peut-être qu'un effet n'est pas pris en compte (ou mal pris en compte)
  - Exemple : pas de prise en compte de nuisibilité, pas de prise en compte d'un stress,…

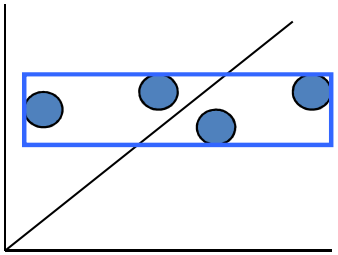
(Différence entre écart-type)²

$$SDSD = (\sigma_Y - \sigma_{\hat{Y}})^2 (N-1)/N$$

SDSD petit



SDSD important



- Si ce terme est important :
  - Peut-être qu'un effet n'est pas pris en compte (ou mal pris en compte), mais cet effet peut être positif ou négatif
  - Exemple : pas de prise en compte de l'effet variété, mais en moyenne, c'est bon.

Terme résiduel

$$LCS = 2\sigma_Y \sigma_{\hat{Y}} (1 - \text{correlation coefficient})*(\text{N-1})/\text{N}$$

- Lié à la construction de la décomposition pour faire apparaitre les deux termes précédents…
- Difficile à interpréter…

Exemple : SeptoLIS

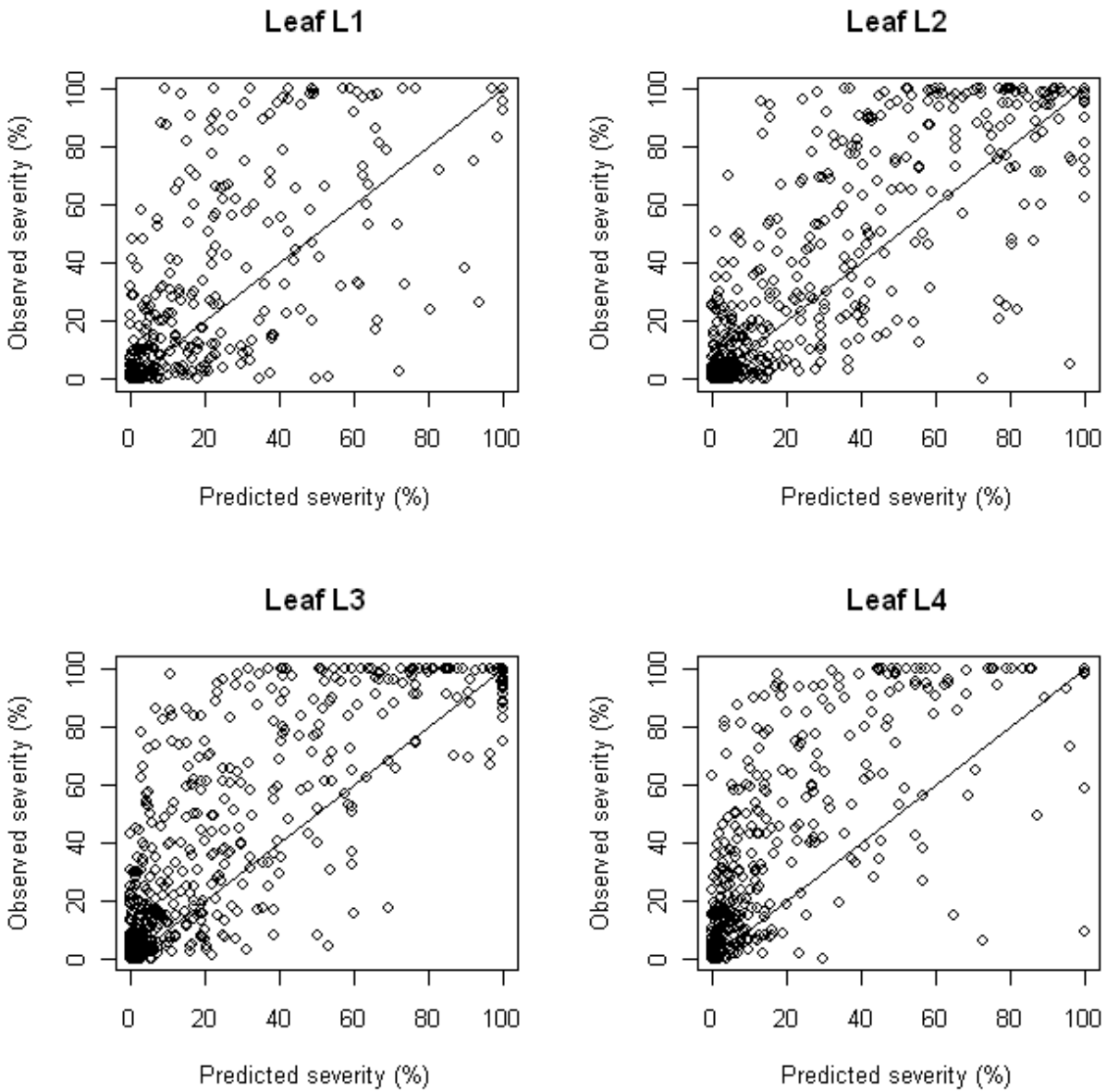


Fig. 2. Observed versus predicted severity, all observations of L1-L4.

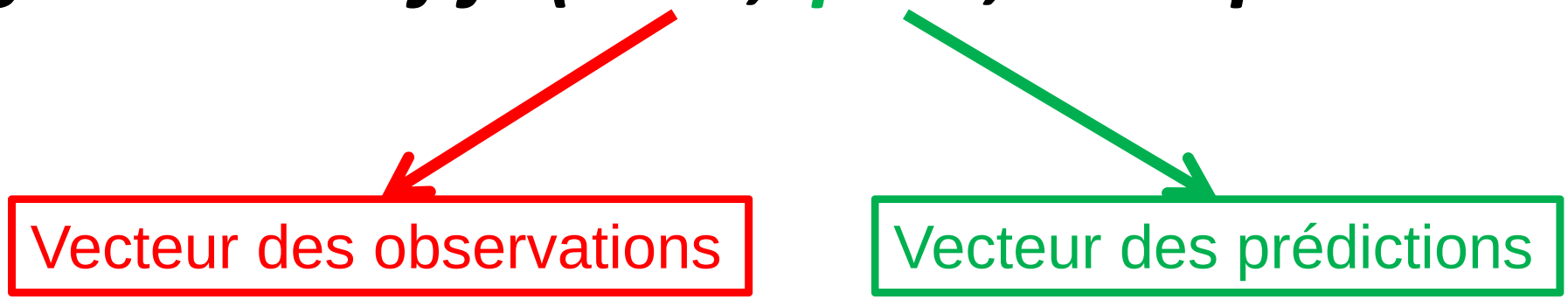
Exemple: SeptoLIS

	MSE	Bias²	SDSD	Remainder	Eff
avant calibration	1518	594	371	553	0.51
après calibration	478	26	3	449	0.62

Aspect pratique

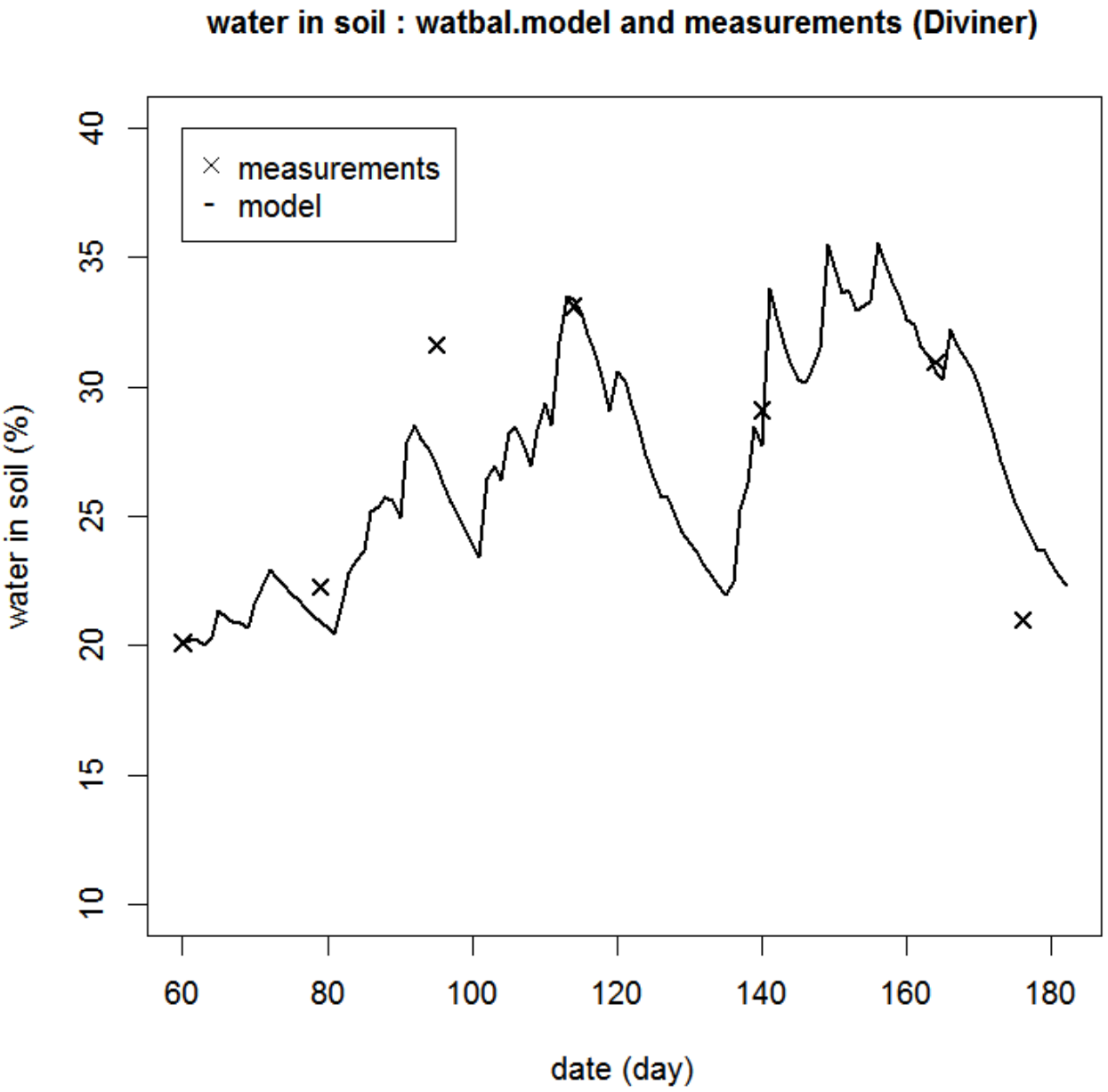
- *Sous R, le package ZeBook*
- *critères d'évaluation en utilisant la fonction qui fait les calculs de critères classiques*

goodness.of.fit(Yobs,Ypred,draw.plot=FALSE)



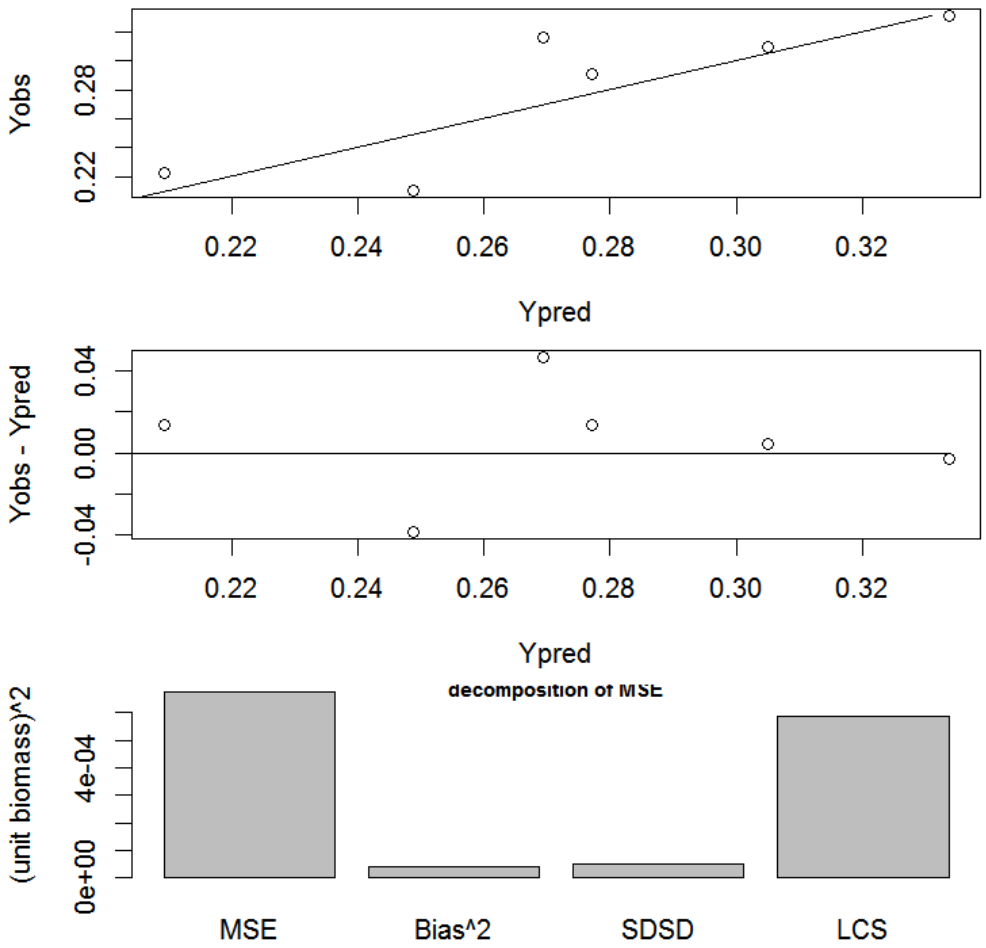
Evaluation of the water balance model (1)

```
library(ZeBook)

head(watbal.simobsdata)

plot(watbal.simobsdata$day,watbal.sim
      bsdata$WATp*100, xlab="date
      (day)", ylab="water in soil (%)",
      ylim=c(10,40),,type="l",  lwd=2)
title("water in soil : watbal.model and
      measurements (Diviner)",cex.main =
      1)
legend(60, 40,
      c("measurements","model"), pch =
      c(4,45), cex =1)
points(watbal.simobsdata$day,watbal.si
      mobsdata$WATp_SF.mean*100,
      cex=1.15, pch=4,lwd=2)
```



Evaluation of the water balance model (2)

```
# first value has been used to
   initiate the model
simobs=watbal.simobsdata[-1,]
Ypred<- simobs$WATp
Yobs<- simobs$WATp_SF.mean

evaluation.criteria(Ypred,Yobs,dra
   w.plot=TRUE)
```




```
 Nobs mean.Yobs mean.Ypred   std.Yobs  std.Ypred         SSE          MSE       RMSE         r
1    6 0.2800066  0.2738266 0.04667444 0.03961418 0.004061729 0.0006769548 0.02601836 0.840745
  bias.Squared         SDSD          LCS       EF
1 3.819224e-05 4.984732e-05 0.0005889153 0.689257
```

Comparer des modèles

- Peut on utiliser la valeur de MSE (ou de l'efficience) pour comparer les modèles
  - En prenant le modèle avec le MSE le plus petit
  - (ou Efficience la plus grande)
- Quels implications ?
  - Comment MSE varie avec la complexité du modèle ?
    - Ex : ajout de variables, d'équations,…

$\Rightarrow$MSE décroit (ou reste identique)

$\Rightarrow$le modèle est plus flexible et s'ajuste mieux au modèle

Problème lié à l'utilisation de MSE

- Cela implique ce choisir toujours le modèle le plus complexe… (MSE diminue)
  - Un modèle plus complexe est il toujours meilleur ?
  - Doit on inclure toute nos connaissances dans un modèle ?
  - Même des variable ayant peu d'effet ?
  - Même si les effets ne sont pas bien connus ?
- Non, ce n'est pas un bon critère
  - MSE mesure l'AJUSTEMENT
  - On est plus intéressé par la prédiction

Qualité de prédiction

Erreur de prédiction

- Souvent, l'objectif principal est la prédiction
- Pas vraiment intéressé par le passé…
  - La comparaison aux situations passées nous intéresse uniquement si cela fournit des informations utiles pour les prédictions

$\Rightarrow$Evaluer l'erreur de prédiction

- Besoin d'un critère
- Besoin d'une méthode pour l'estimer

Un critère pour mesurer l'erreur de prédiction : MSEP

- **MSEP : Mean squared error of prediction**
- Mesure populaire pour l'erreur des prédictions
- MSE représente l'erreur au carré, moyennée sur les données
- Définition de MSEP

$$MSEP = E\left[\left(Y - \hat{Y}\right)\right]^2$$

$$MSE = (1/N)\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

- **L'erreur au carré, moyennée sur toutes les situations pour lesquelles on veut faire une prédiction**

MSEP est liée à une population cible particulière

- Ce sont les situations pour lesquelles on veut faire des prédictions
- Besoin de définir ces situations
  - Exemple : sols, pratiques, climats (donc régions)

- Aussi pour des variables d'intérêts particulières
  - Sévérité de la maladie, LAI, etc.
  - Bien les définir, en lien aussi avec les variables observables

Estimation de MSEP

- En général, MSEP ne peut pas être mesuré
  - Lié à la population entière et donc inaccessible (par exemple climats possibles)
- On peut seulement obtenir une estimation de MSEP, pas la valeur exacte

MSE, une estimation de MSEP ?

- MSE ressemble beaucoup à MSEP
  - Les deux : moyenne de carré des erreurs (des résidus)
  - MSE : pour un échantillon

≠

  - MSEP : pour l'ensemble de la population cible
- MSE = estimateur raisonnable (sans biais) de MSEP
  - SI : l'échantillon est représentatif de la population représentatif de la population cible & si le modèle n'a pas été calibré avec ces données
  - SINON : MSE n'est pas un estimateur raisonnable (peu sous-estimer largement l'erreur de prédiction)

Condition 1. MSE doit être basé sur la population cible.

- Si l'on veut prédire le niveau d'une maladie du blé
  - population = parcelles de blé en France
  - Échantillon = parcelles de blé en agriculture biologique

- Est-ce que MSE est un bon estimateur pour la population ?
- NON, car ce n'est pas du tout les mêmes variétés, pas les mêmes pratiques,…

Condition 2. MSE ne doit pas être calculé sur les données utilisées pour calibrer le modèle

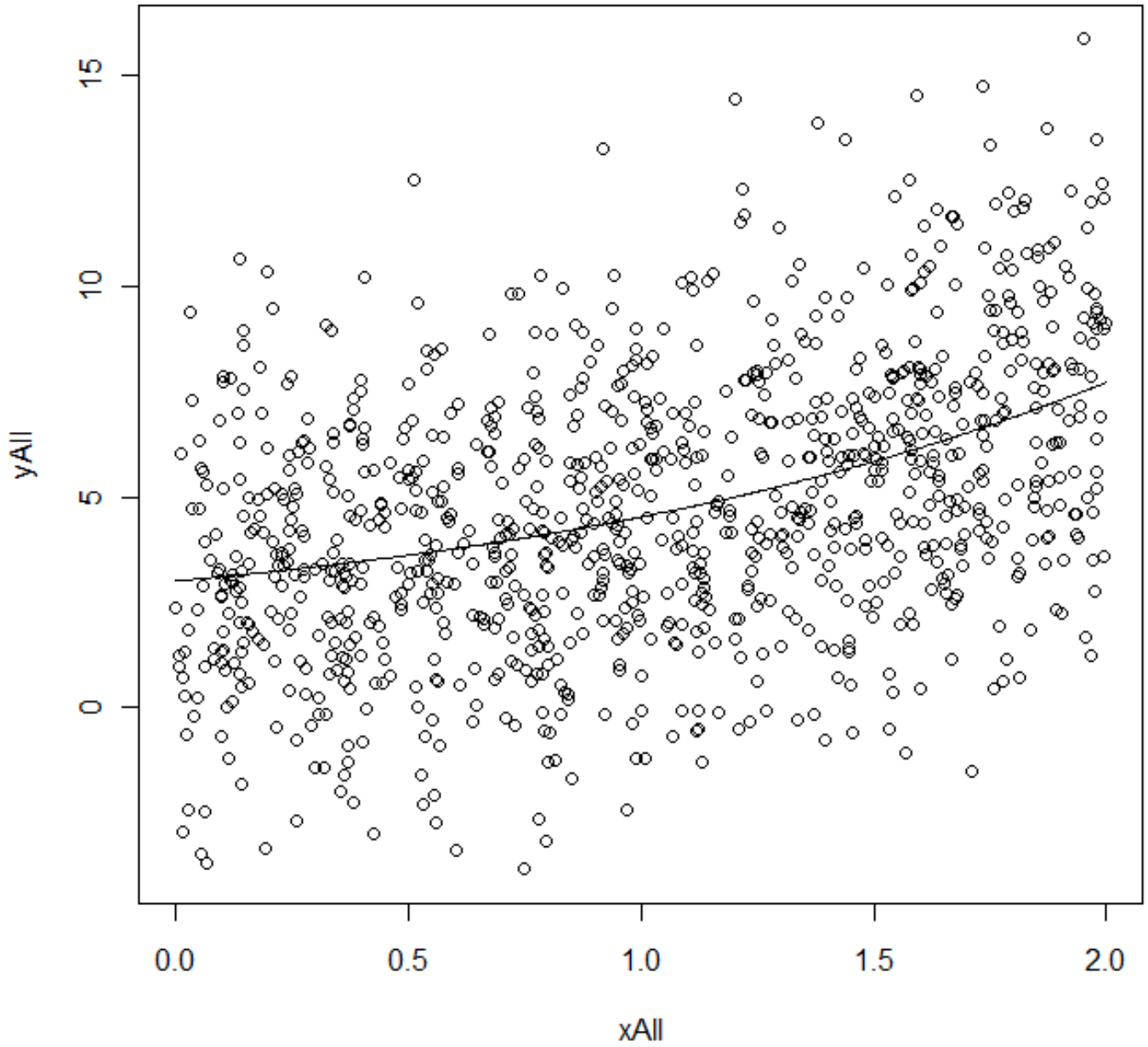
- Une fois que le modèle est ajusté aux données de l'échantillon, ces données deviennent spéciales.
  - Les paramètres du modèle sont spécialement adaptés à ces données

Un exemple pour comprendre la différence entre MSE et MSEP

- Illustration avec un modèle statistique simple

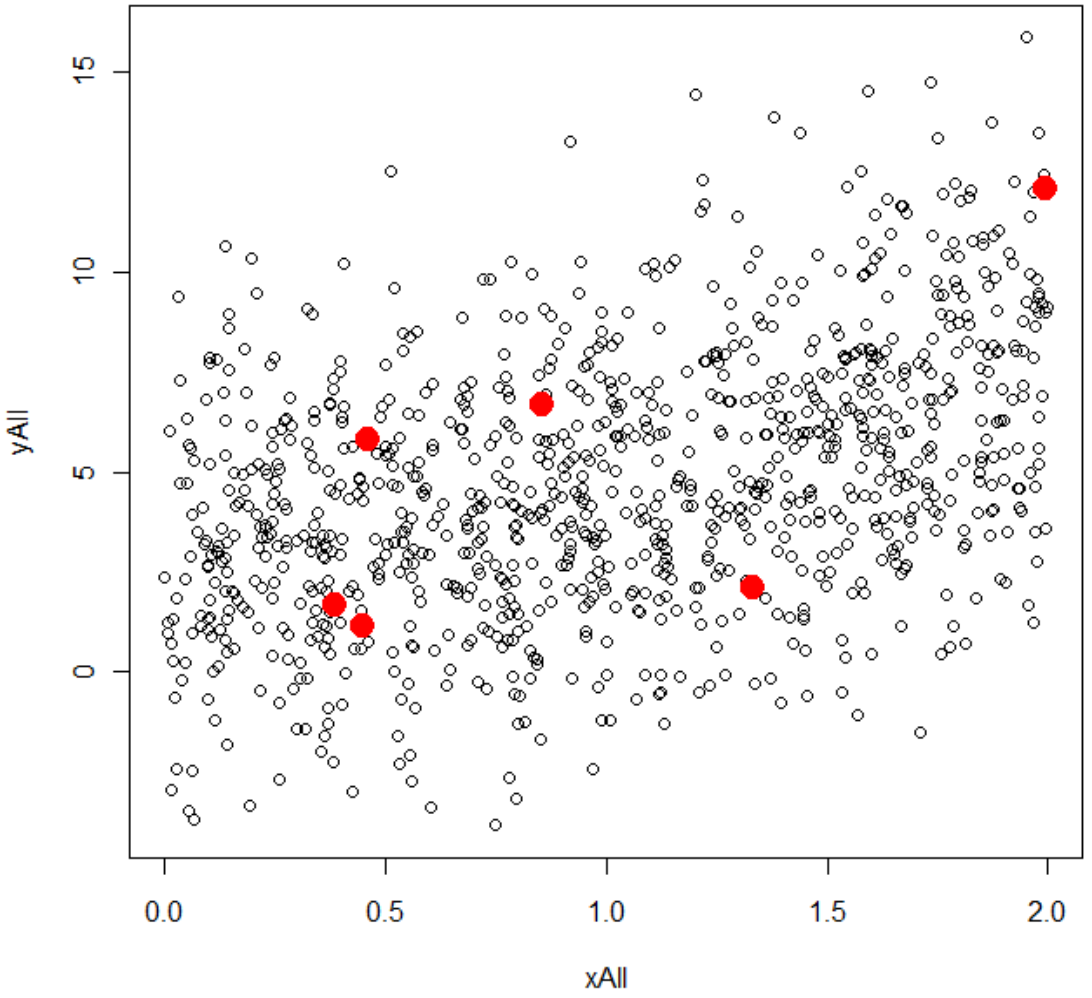
- $y=3+x+0.4x^2+0.1x3+0.02x4+\varepsilon \quad \varepsilon \sim N(0,3^2)$
  - y est la variable à expliquer
  - x est la variable d'entrée
    - nombreuses valeurs de x tirées au hasard dans [0,2]
    - notre population cible
  - ε est une variable aléatoire (x n'explique pas tout)

La population entière



On tire un échantillon de taille 6

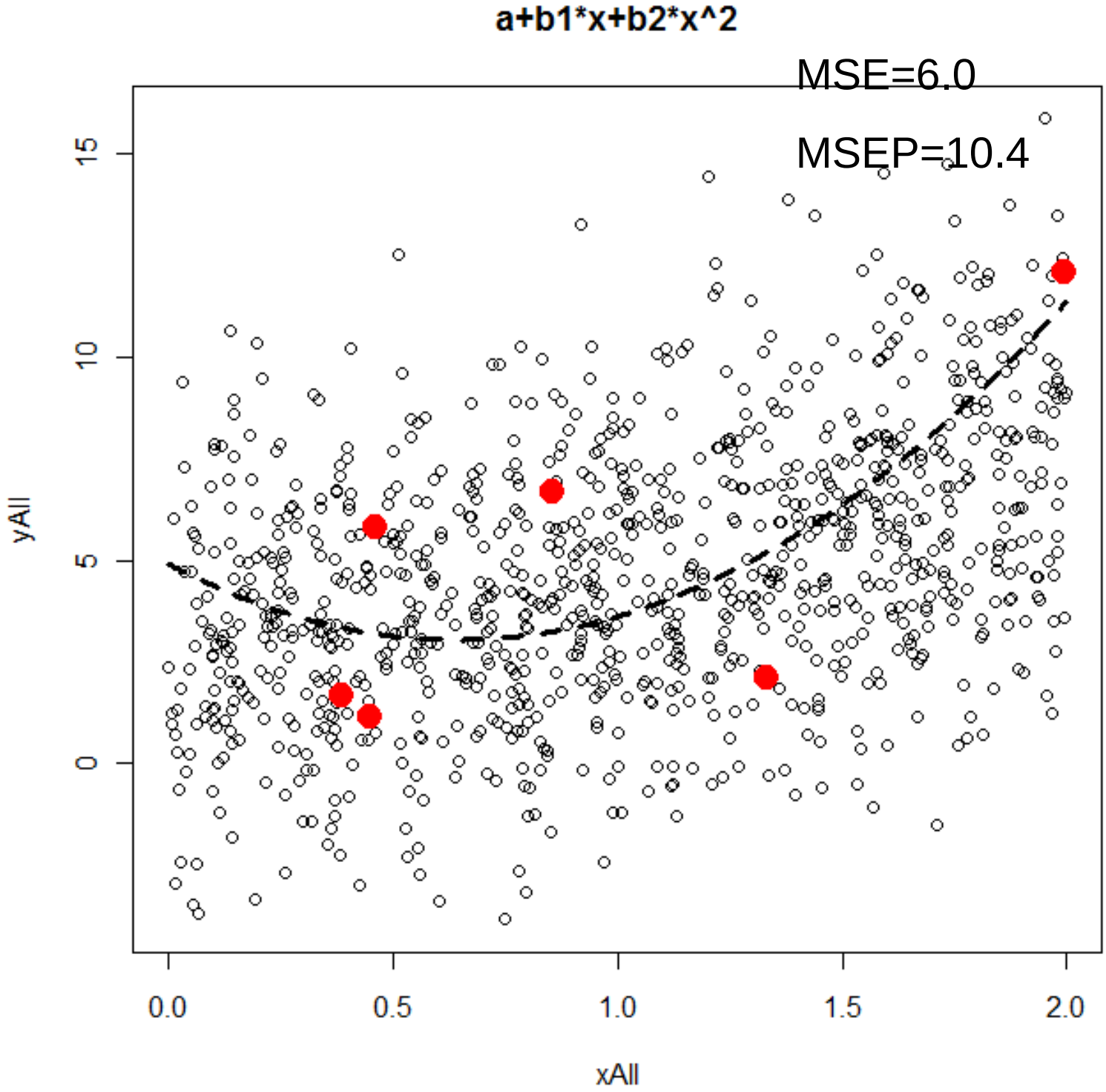
- MSE est l'erreur du modèle pour l'échantillon considéré
- MSEP est l'erreur pour l'ensemble de la population

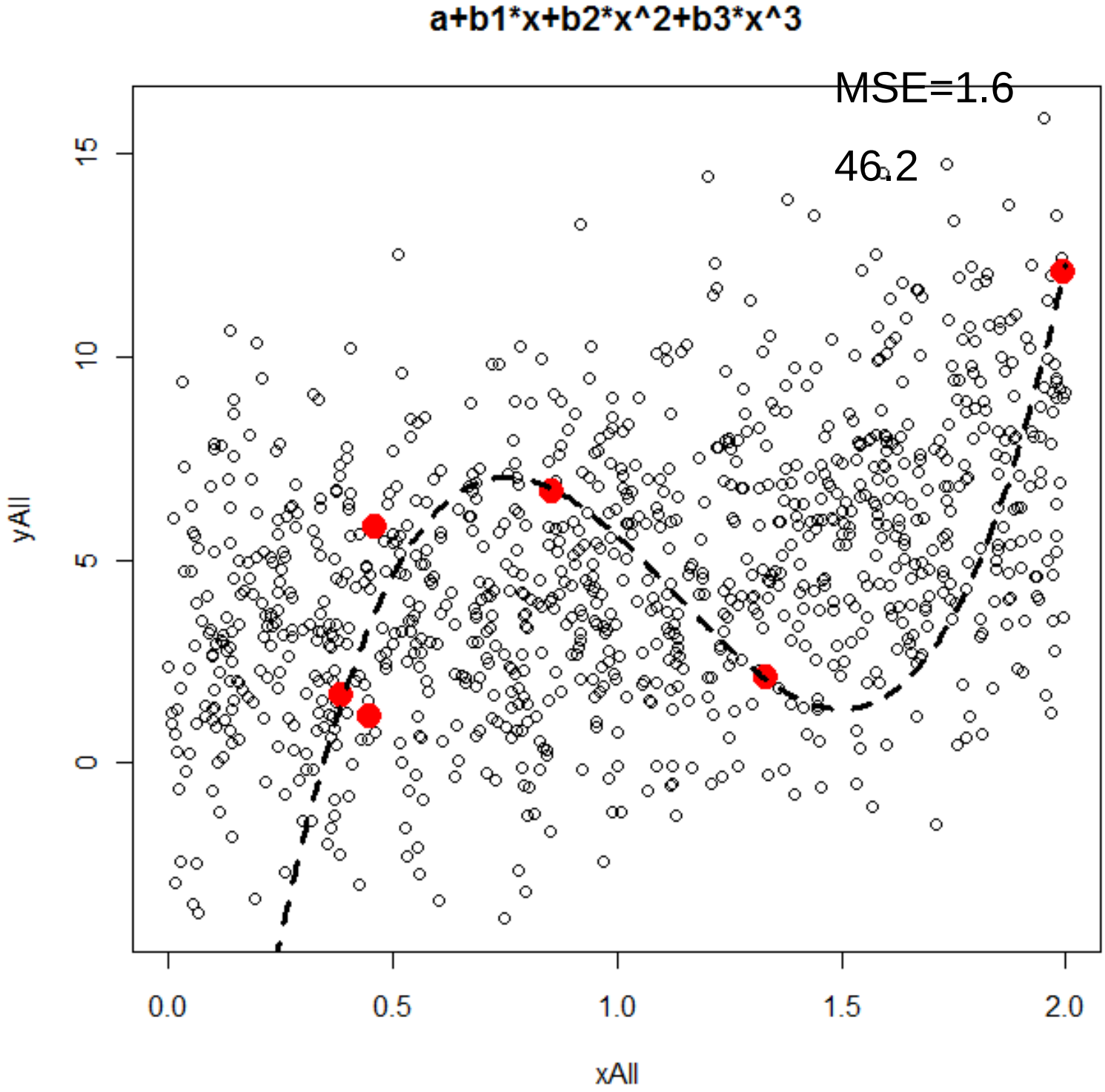


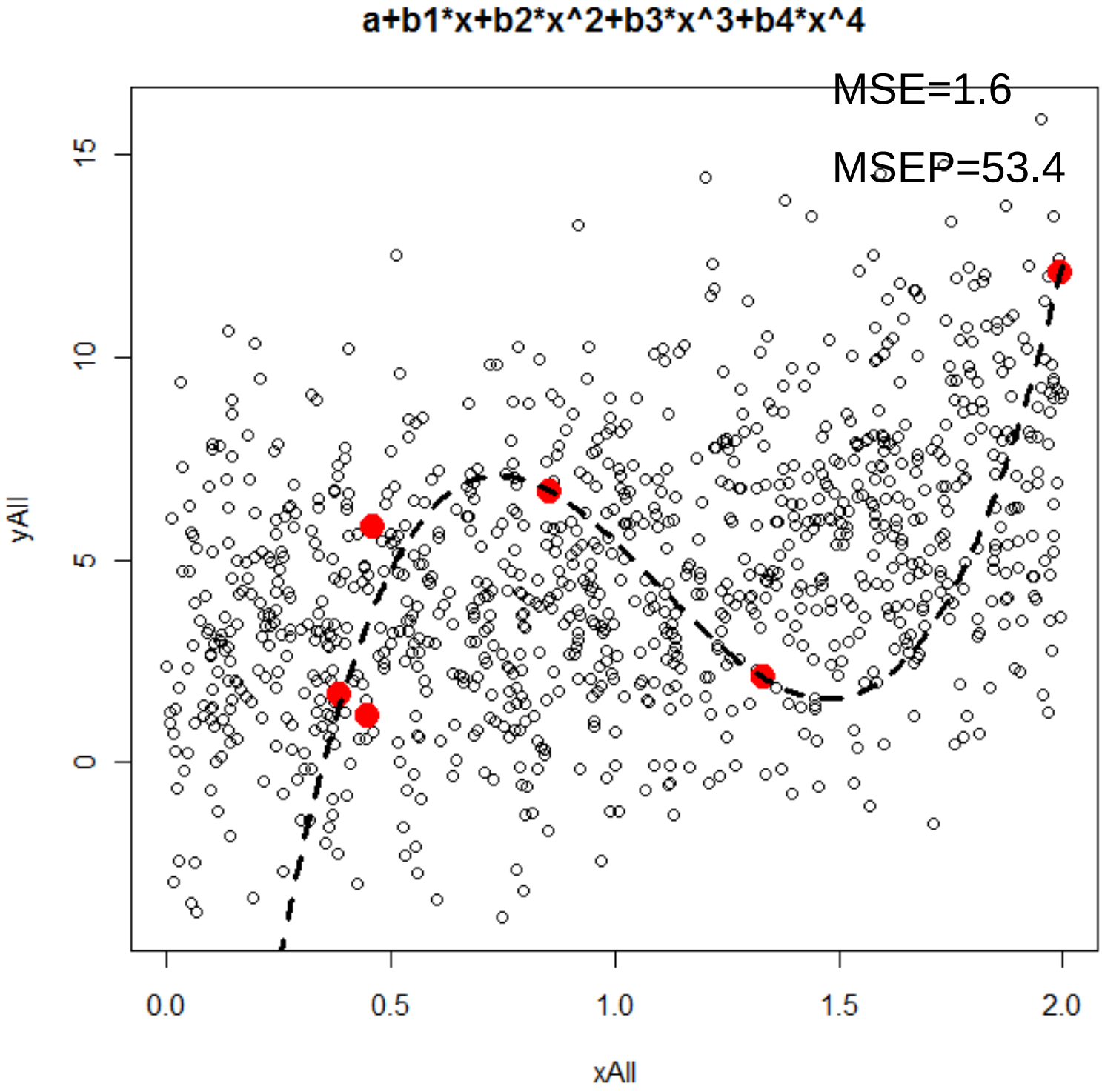
Ajuster les modèles aux données

- Tester différents modèles
  - $y=a+b1*x$
  - $y=a+b1*x+b2*x^2$
  - $y=a+b1*x+b2*x^2+b3*x3$
  - $y=a+b1*x+b2*x^2+b3*x3+b4*x4$

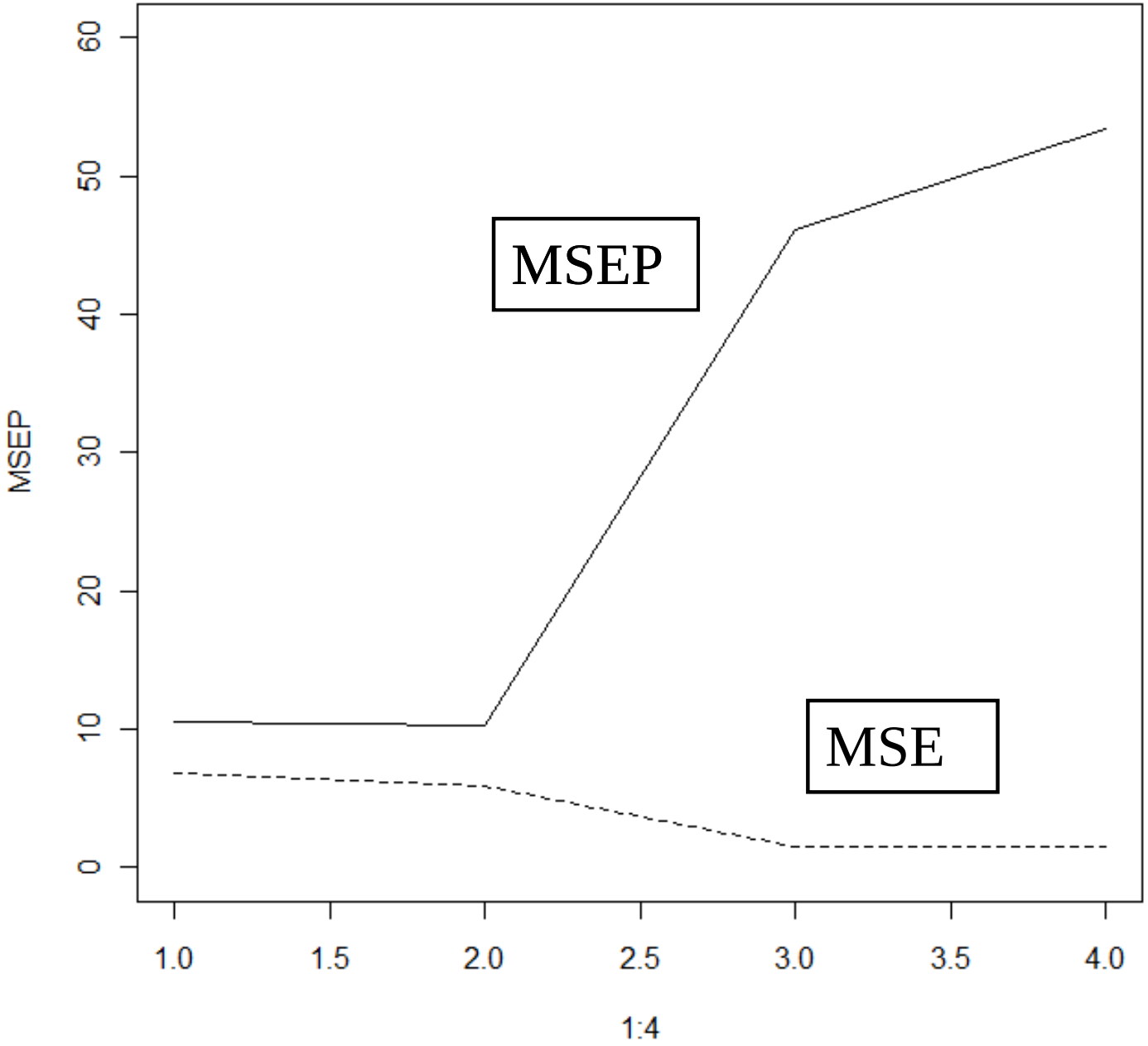
- Estimer les paramètres pour chaque modèle (avec les moindres carrés)
- Calculer MSE et MSEP pour chaque modèle
- Quel modèle présentera le plus petit MSE?


a+b1*x+b2*x^2
MSE=6.0
MSEP=10.4
yAll
xAll
0
5
10
15
0.0
0.5
1.0
1.5
2.0


a+b1*x+b2*x^2+b3*x^3
MSE=1.6
46.2
yAll
xAll
0
5
10
15
0.0
0.5
1.0
1.5
2.0


a+b1*x+b2*x^2+b3*x^3+b4*x^4
MSE=1.6
MSEP=53.4
yAll
xAll
0
5
10
15
0.0
0.5
1.0
1.5
2.0

MSE, MSEP en fonction du nombre de paramètres



Pour un modèle ajusté aux données

- MSE diminue en ajoutant des détails/des processus (plus de paramètres, plus “flexible”)
- MSEP atteint une valeur minimum pour un niveau intermédiaire de complexité
- Plus on rajoute de la complexité, plus MSE est trompeur…
- Sur-paramétrisation: estimer trop de paramètres. MSE semble bien, mais MSEP est beaucoup plus important
  - AIC prend en compte une pénalisation liée au nombre de paramètres

Comment estimer MSEP ?

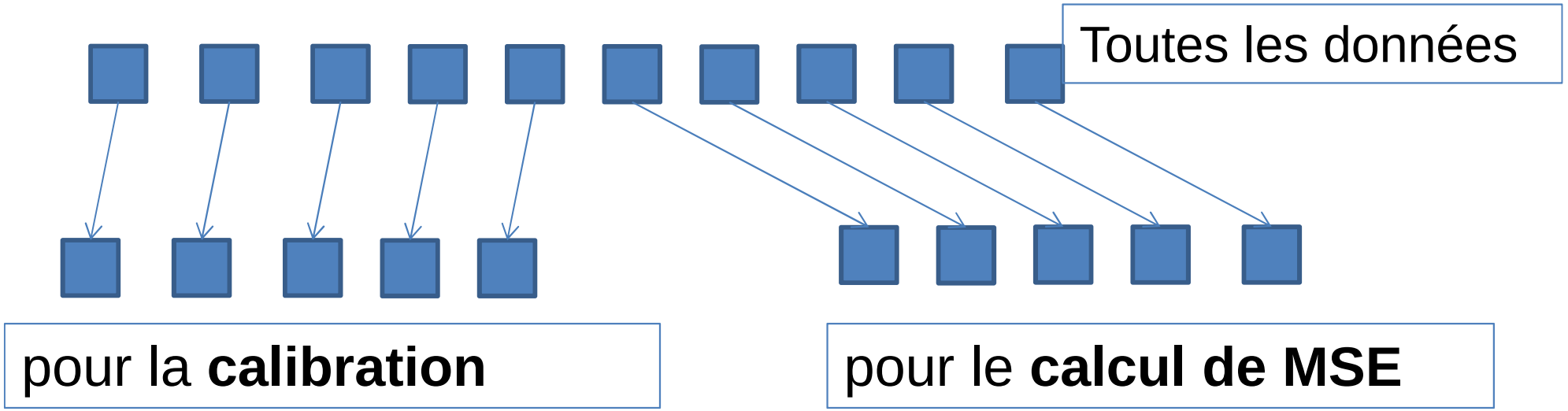
La situation

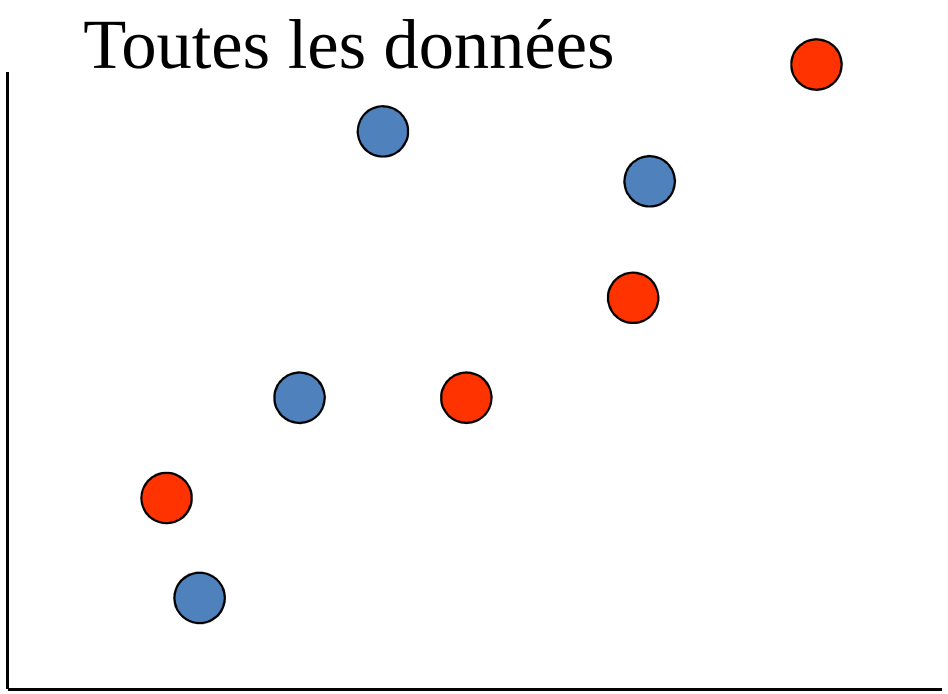
- On a des données représentatives de la population cible.
- On veut utiliser ces données:
  - Pour estimer les paramètres
  - Mais, souhait d'estimer MSEP également

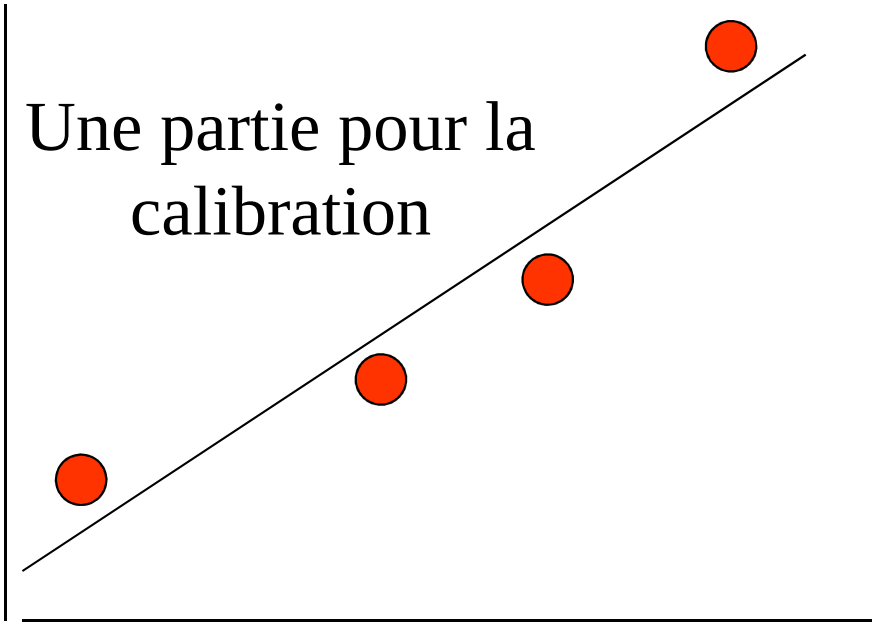
$\Rightarrow$on utilise ces données pour la calibration, MSE n'est pas un bon estimateur de MSEP

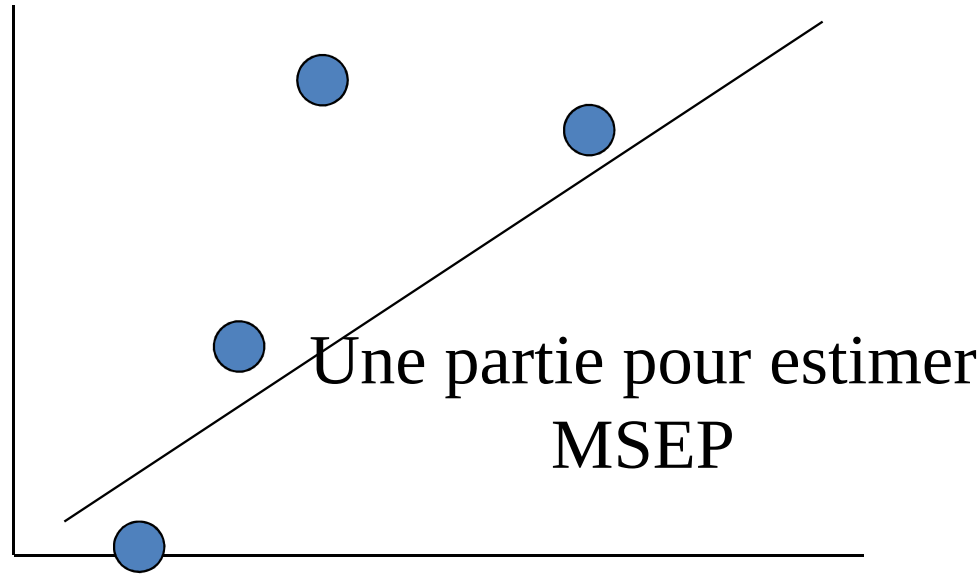
Comment faire ?

- Diviser le jeu de données en deux parties : une pour la calibration, l'autre pour l'estimation de MSEP.
  - La deuxième partie n'étant pas utilisée pour la calibration, MSE est une estimation de MSEP.
  - (avec l'hypothèse que c'est un échantillon représentatif de la population cible)




Toutes les données


Une partie pour la calibration

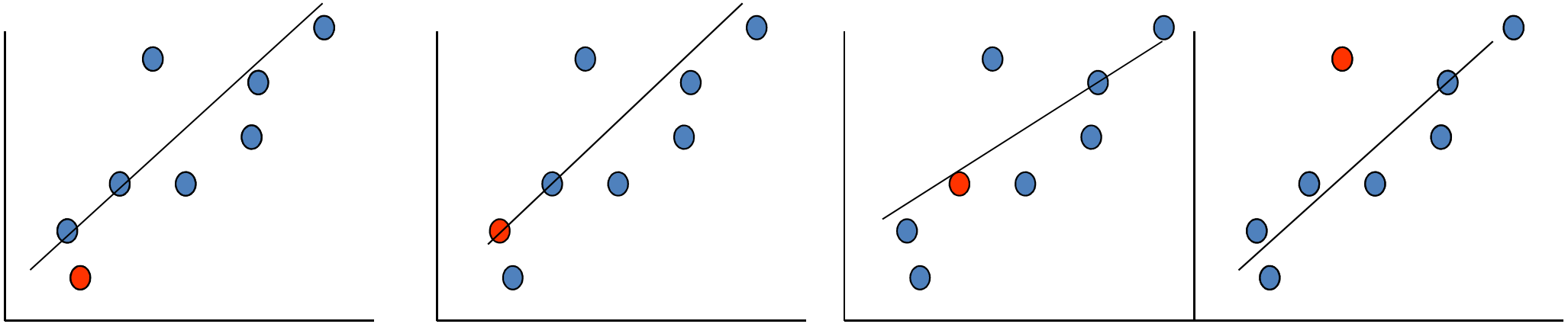

Une partie pour estimer MSEP

Est-ce que cela pose problème ?

- oui
- La partition/division du jeu de donnée est arbitraire
- L'estimation des paramètre ne profite pas de l'ensemble des données
- L'estimation du MSEP ne profite pas de l'ensemble des données

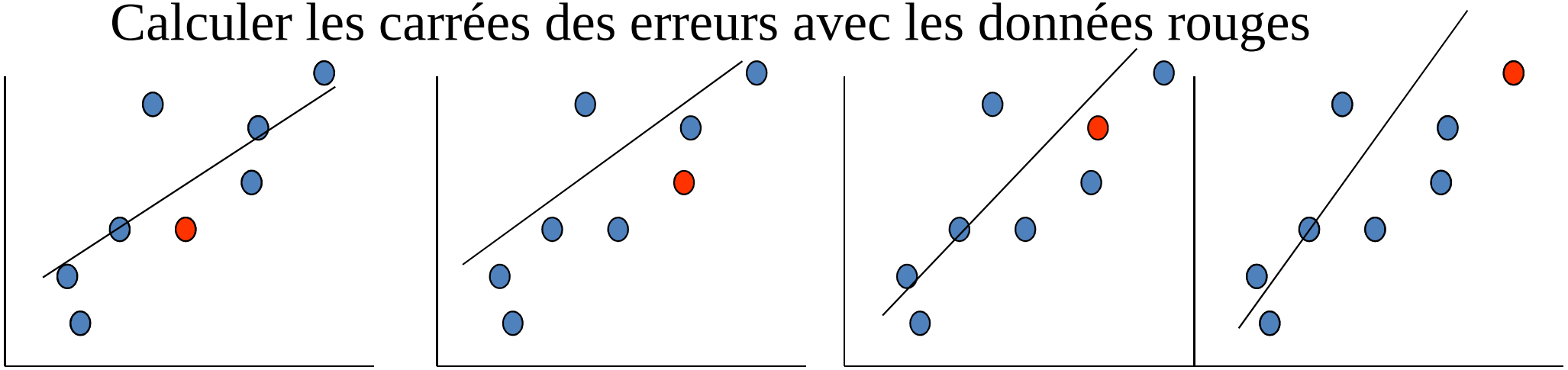
Validation croisée

- Un moyen plus efficace d'utiliser les données.
- Pour N données, faire N partitions différentes du jeux de données.
- Pour chaque partition
  1. calibrer avec les N-1 données
  2. estimer MSEP avec une seule donnée.



Estimer les paramètres avec les données bleus

Calculer les carrées des erreurs avec les données rouges



Estimation finale du MSEP : moyenne des carrées des erreurs.

Calcul

Y1 Y2 Y3 ……. YN $(Y_1-\hat{Y}_1(\theta_{-1}))^2$

Y1 Y2 Y3 ……. YN $(Y_2-\hat{Y}_2(\theta_{-2}))^2$

Y1 Y2 Y3 ……. YN $(Y_N-\hat{Y}_N(\theta_{-N}))^2$

$$\hat{MSEP} = 1/N \sum \left[Y_i - \hat{Y}_i(\theta_{-i}) \right]^2$$

Est-ce que cela pose problème ?

- On obtient N modèles différents…
- Quel modèle fournir aux utilisateurs ?
- La valeur de MSEP calculée correspond à quel modèle ?

Peut-on faire mieux ?

- Les limites : on ne peut pas utiliser toutes les données à la fois pour la calibration et l'estimation de MSEP.
- Rappel de nos priorités :
  1. avoir le meilleur modèle possible
  2. avoir un bon estimateur de MSEP pour ce modèle.

La validation croisée est une assez bonne solution

1. Utiliser **TOUTES LES DONNEES pour la calibration**. Le modèle obtenu est celui à utiliser pour la prédiction.
2. Utiliser la **validation croisée pour estimer MSEP**. MSEP est en fait moyenné sur un ensemble de modèles très similaires au modèle de prédiction
   - Mêmes équations.
   - Même jeu de données, sauf une partie.

Quels effets liés à l'ajout d'un détail au modèle ?

- Diminution de la variance de Y pour les mêmes situations
  - Le modèle explique a priori plus de la variabilité de Y
- Mais l'erreur du modèle augmente
  - Car besoin de plus d'équations et de paramètres
- Si l'ajout d'un détail permet de beaucoup mieux expliquer la variabilité, MSEP diminue
- Si c'est l'ajout d'un détail peu important (peu explicatif), alors MSEP augmente
- Il y a un optimum au niveau de la complexité pour obtenir le MSEP minimum

Conclusions pratiques

- Spécifier la population cible et les variables d'intérêt
- Partir d'un "bon modèle" : éviter la sur-paramétrisation pendant la calibration
  - Estimer uniquement les paramètres importants (effet important sur le MSE)
  - Une méthode pour pénaliser l'utilisation de trop de paramètre (AIC,…)
- Explorer les données et l'ajustement (résidus)
- Utiliser un estimateur raisonnable de MSEP (sans tricher !)
  - S'assurer que l'échantillon représente bien la cible
  - S'assurer de l'indépendance des parties du jeu de données lors de la validation croisée