Arbre CART, bagging et boosting

Arnaud Callebaut & Lionel Hertzog (illustrations : Pierre Mérian)

30/04/2025

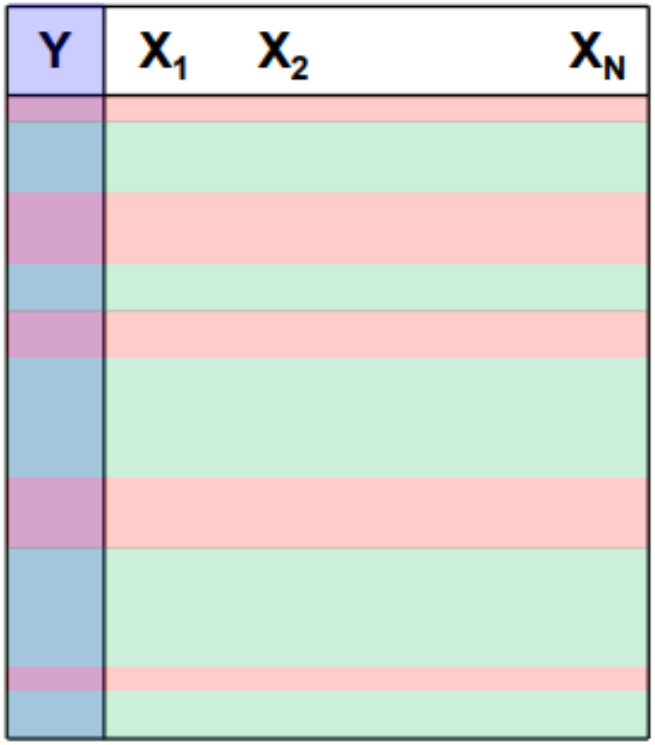
Rappel

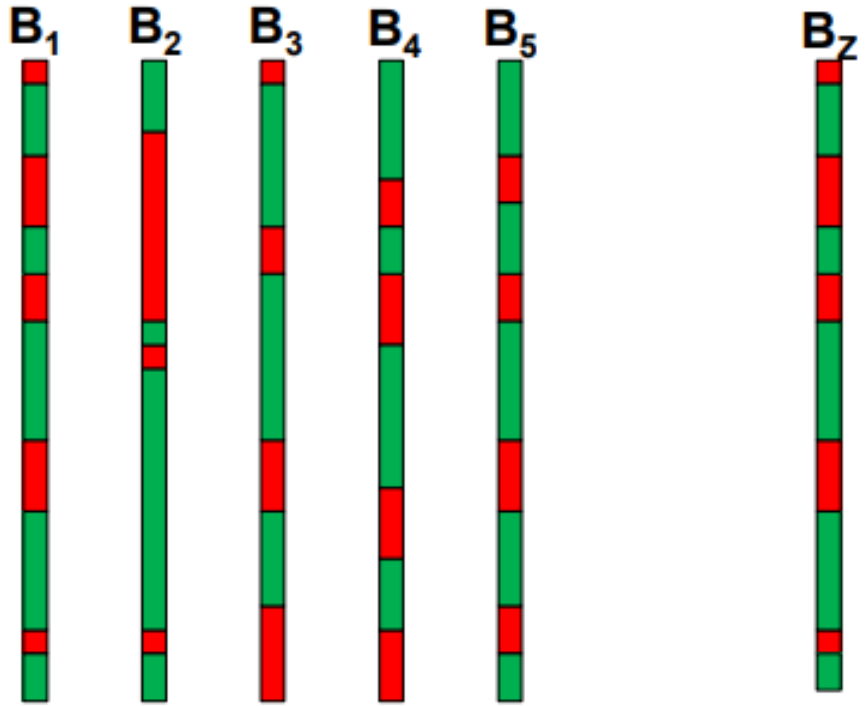
- Les arbres CART partitionnent l'espace des variables pour maximiser l'homogénéité de la variable réponse
- Les méthodes d'ensemble (bagging / boosting) permettent d'obtenir des prédictions plus robustes que des modèles uniques

Bagging sur CART

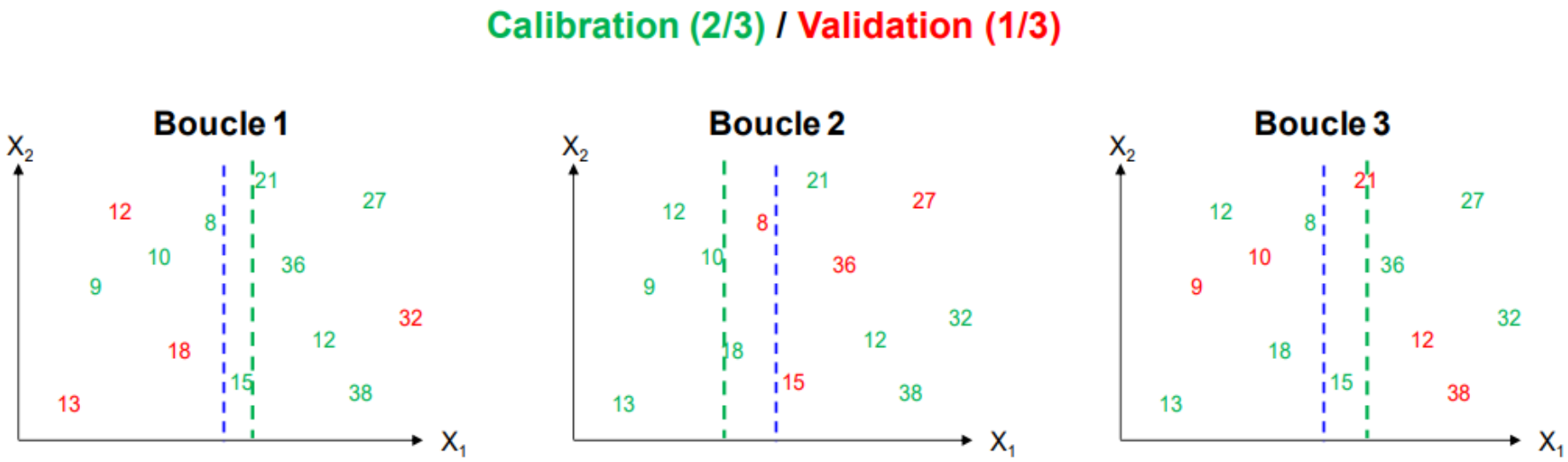
Processus

- On applique une procédure d'échantillonnage avec remise sur les individus.
- Typiquement, on met 66% de calibration, 33% de validation.
- En répétant Z fois, on obtient Z arbres.





Indépendance des arbres



Problème lié au bagging

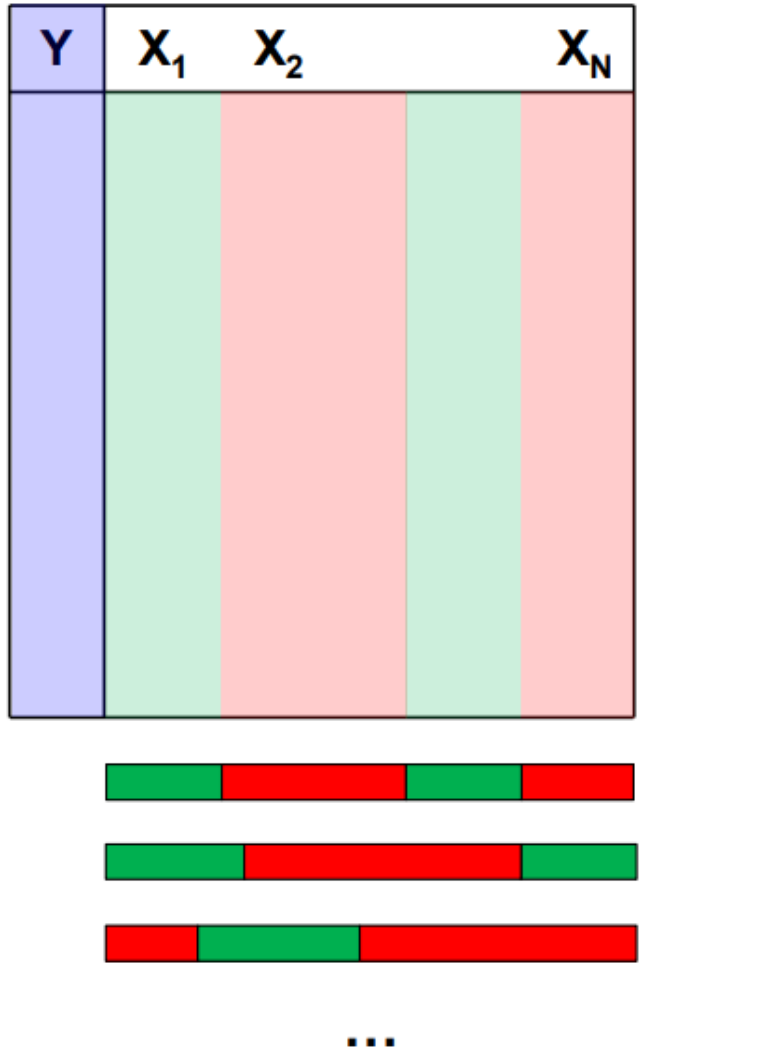
Problème : Quels que soient les jeux de calibration / validation, X1 est toujours sélectionnée comme première variable discriminante. Les arbres ne sont pas indépendants.

Solution : avoir un pool de variables propre à chaque arbre pour que l'ordre de sélection des variables X dépende de l'arbre (et non du jeu de données)

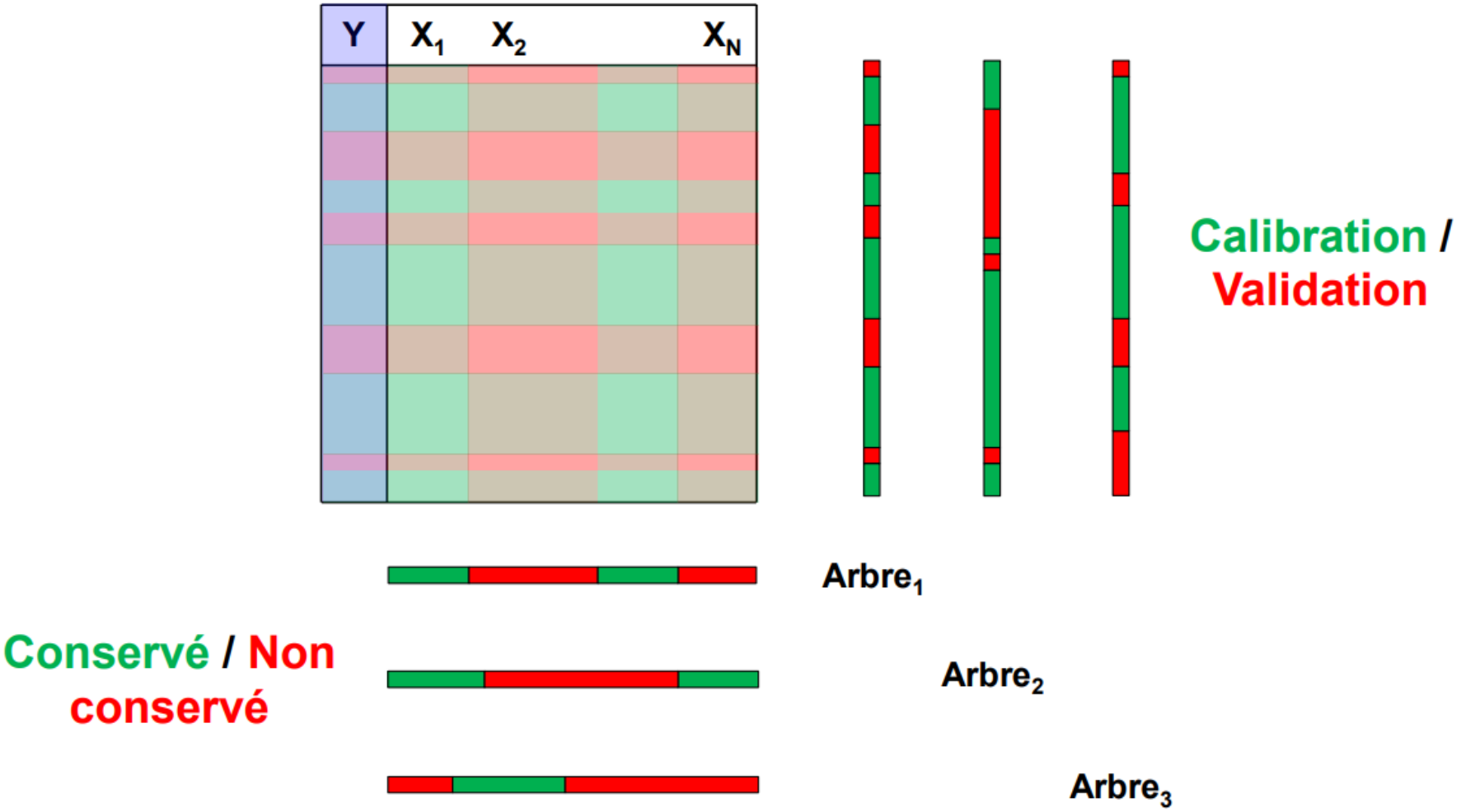
Indépendance des arbres

Choix des variables

- On applique une procédure d'échantillonnage avec remise sur les variables.
- Typiquement, on en choisit $N/3$, $log(N)$, N^{-2}...
- En répétant Z fois, on obtient Z arbres indépendants.



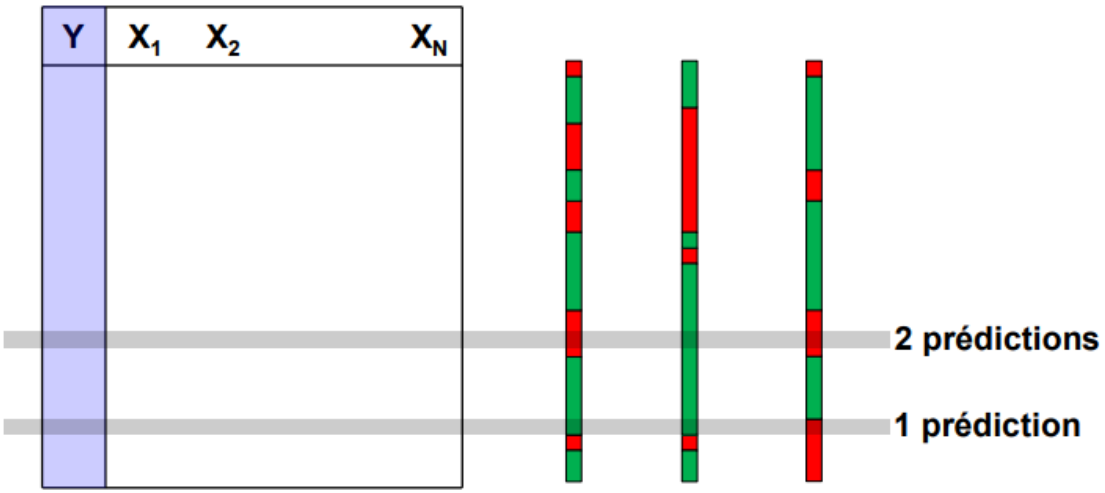
Constitution d'une forêt aléatoire



Evaluer la qualité de notre RF

OOB error (Out-of-bag error)

- On fait une prédiction pour chaque individu en ne prenant que les arbres non construits avec.
- Pour chaque individu, avec Z arbres, on aura donc en moyenne Z/3 prédictions.
- L'OOB error donne une idée de la qualité de prédiction sur le jeu de données.



Calibration = pas de prédiction

/

Validation = prédiction

Evaluer la qualité de notre RF

OOB error (Out-of-bag error)

- On fait une prédiction pour chaque individu en ne prenant que les arbres non construits avec.
- Pour chaque individu, avec Z arbres, on aura donc en moyenne Z/3 prédictions, que l'on agrégera.
- L'OOB error donne une idée de la qualité de prédiction sur le jeu de données.

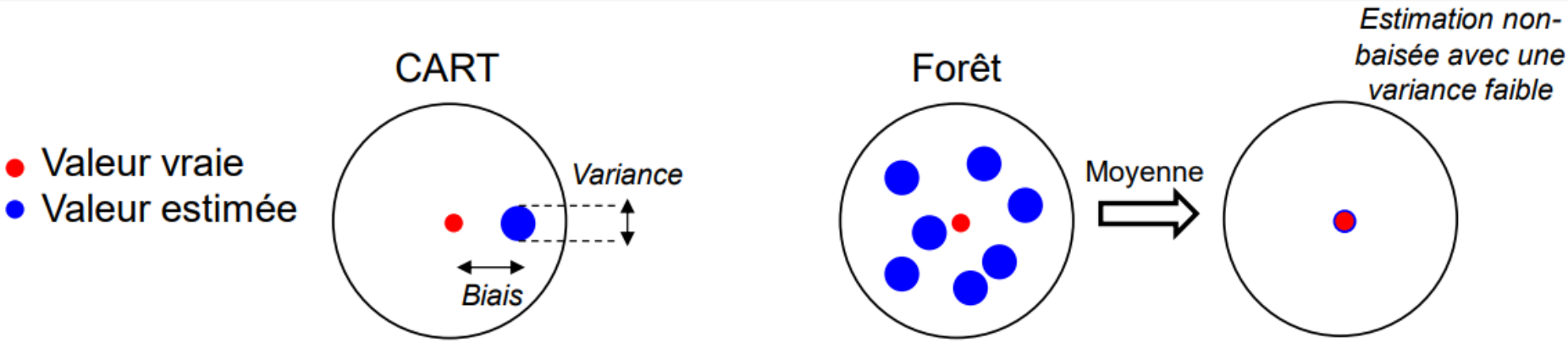
$$\text{OOB Error} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i^{\text{OOB}} - y_i\right)^2 \quad \text{(regression)}$$

$$\text{OOB Error} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(\hat{y}_i^{\text{OOB}} \neq y_i\right) \quad \text{(classification)}$$

Pourquoi les Random Forest ?

Compromis Biais/Variance

- On limite le biais de prédiction grâce au bagging.
- L'indépendance des arbres réduit la variance finale du modèle.



Évaluer l'importance des variables dans une Random Forest

L'évaluation de l'importance de chaque variable est délicat dans une RF. Des méthodes existent, mais sont moins précises et fiables que ce qu'on peut connaître pour les modèles linéaires par exemple. (effet boîte noire)

Importance basée sur la réduction d'impureté

- Chaque fois qu'une variable est utilisée pour un split dans un arbre, elle réduit un peu l'impureté (Gini en classification, variance/MSE en régression).
- On cumule ces réductions sur tous les arbres, pour chaque variable.
- Biaisée vers les variables avec plus de modalités ou à valeurs continues.

Évaluer l'importance des variables dans une Random Forest

L'évaluation de l'importance de chaque variable est délicat dans une RF. Des méthodes existent, mais sont moins précises et fiables que ce qu'on peut connaître pour les modèles linéaires par exemple.(effet boîte noire)

Importance par processus de permutation

- On prend le jeu de données initial, et on permute aléatoirement les valeurs d'une variable.
- La qualité de la RF va à priori diminuer. Plus la variable est importante, plus la performance chute beaucoup.
- Plus fiable et moins biaisé, reflète l'effet réel sur les prédictions.
- Lent, et dur à analyser si les corrélations entre variables sont très fortes.

Avantages, Inconvénients Random Forest

Avantages de la RF

- Robustesse (bagging + randomisation des variables)
- Modèle non-paramétrique
- Liberté dans la nature des variables employées (continues/discrètes)
- Prise en compte d'interactions d'ordres élevés et d'effet non linéaires.

Désavantages de la RF

- Difficile à visualiser.
- Risque de sur-apprentissage.
- Effet des variables dur à analyser.
- Nécéssite une grande puissance de calcul.
- Prédiction hors de la zone de calibration quasi-impossible.

What we saw so far

- CART : Les RF correspondent à un modèle de type bagging dont les arbres ne seraient constitués que d'une partie des variables explicatives.
- Une manière d'évaluer la qualité du modèle est l'OOB error, qui quantifie la moyenne des erreurs pour chaque individu, en ne prenant en compte que les arbres qui n'ont pas été contruits à partir de ces individus.
- L'inférence (= quantification de l'effet des variables) est difficile et mesurée indirectement par deux métriques imparfaites : la réduction d'impureté totale de la forêt et la réduction de l'efficacité par permutation successive des variables.