Forêt aléatoire distributionnelle (DRF)

Arnaud Callebaut & Lionel Hertzog

30/04/2025

Rappel

- Les arbres CART partitionne 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
- Les forêts aléatoires combinent les arbres CART, le bagging et une sélection aléatoire des variables à évaluer à chaque découpage (mtry)
- (BRT)

Allez plus haut

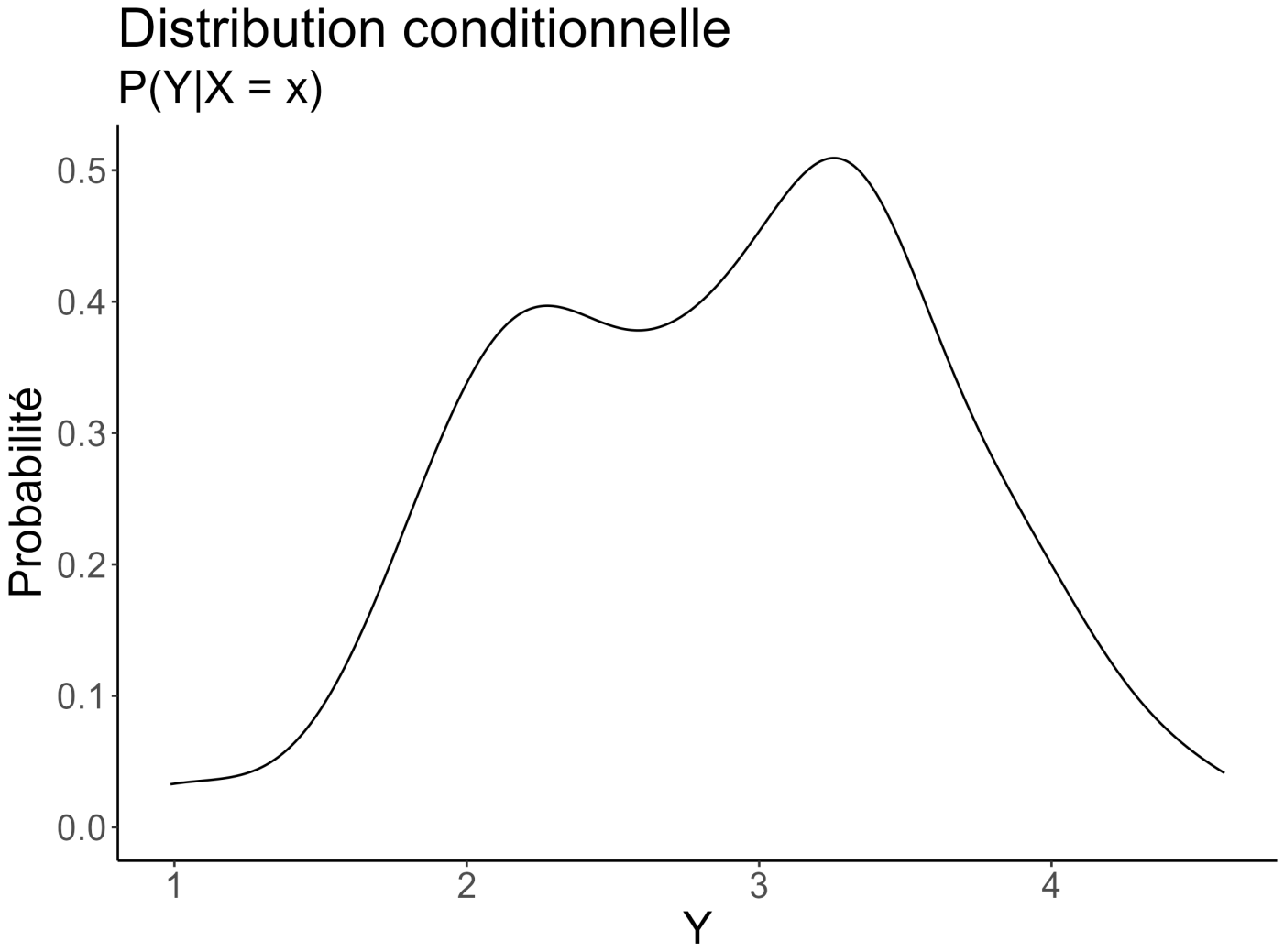
Limites des forêts aléatoires

1. Pas d'estimation d'incertitude des prédictions
2. Une seule variable de réponse

[add fig from RF TP]

Une distribution cible

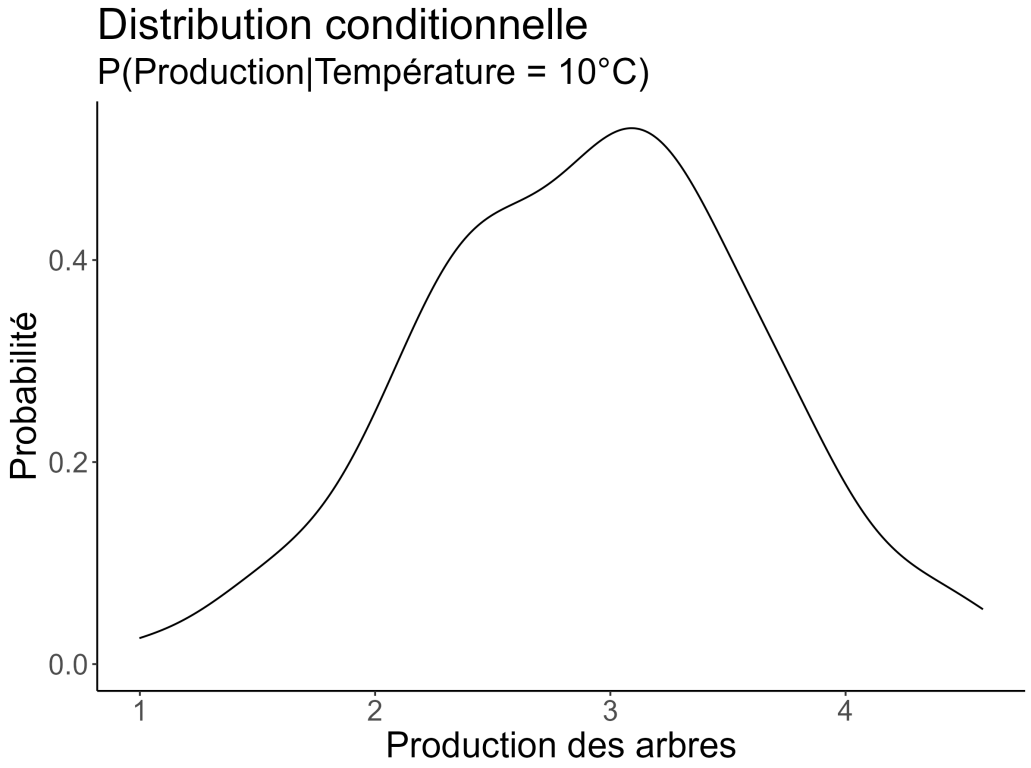
La distribution conditionnelle



Une distribution cible

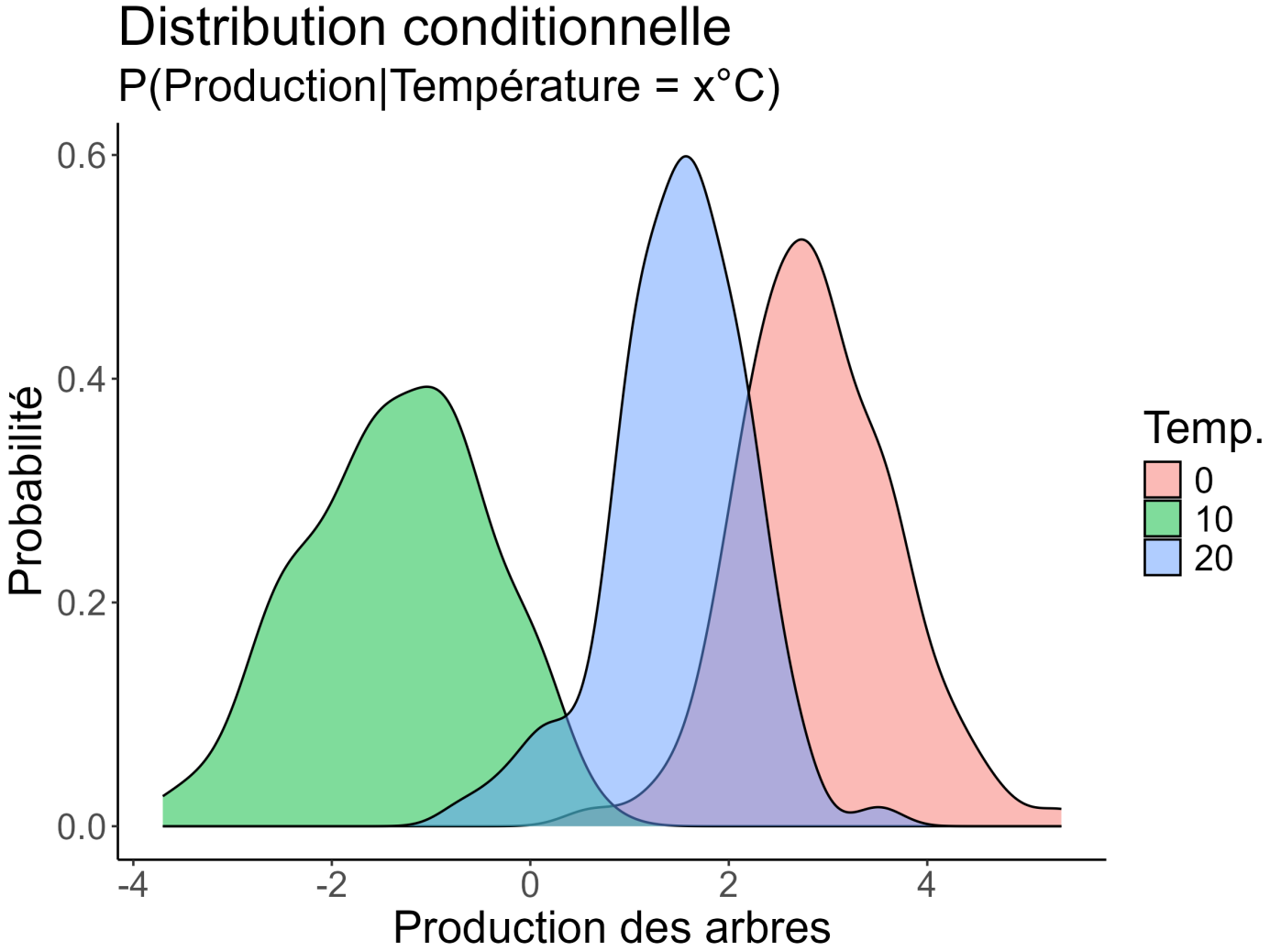
A partir de cette distribution de nombreuses statistiques peuvent être dérivé

- Moyenne, Médiane, Mode
- Intervalle de confiance
- Probabilité, p.ex. $P(Prod.|Temp. = x) > 2$



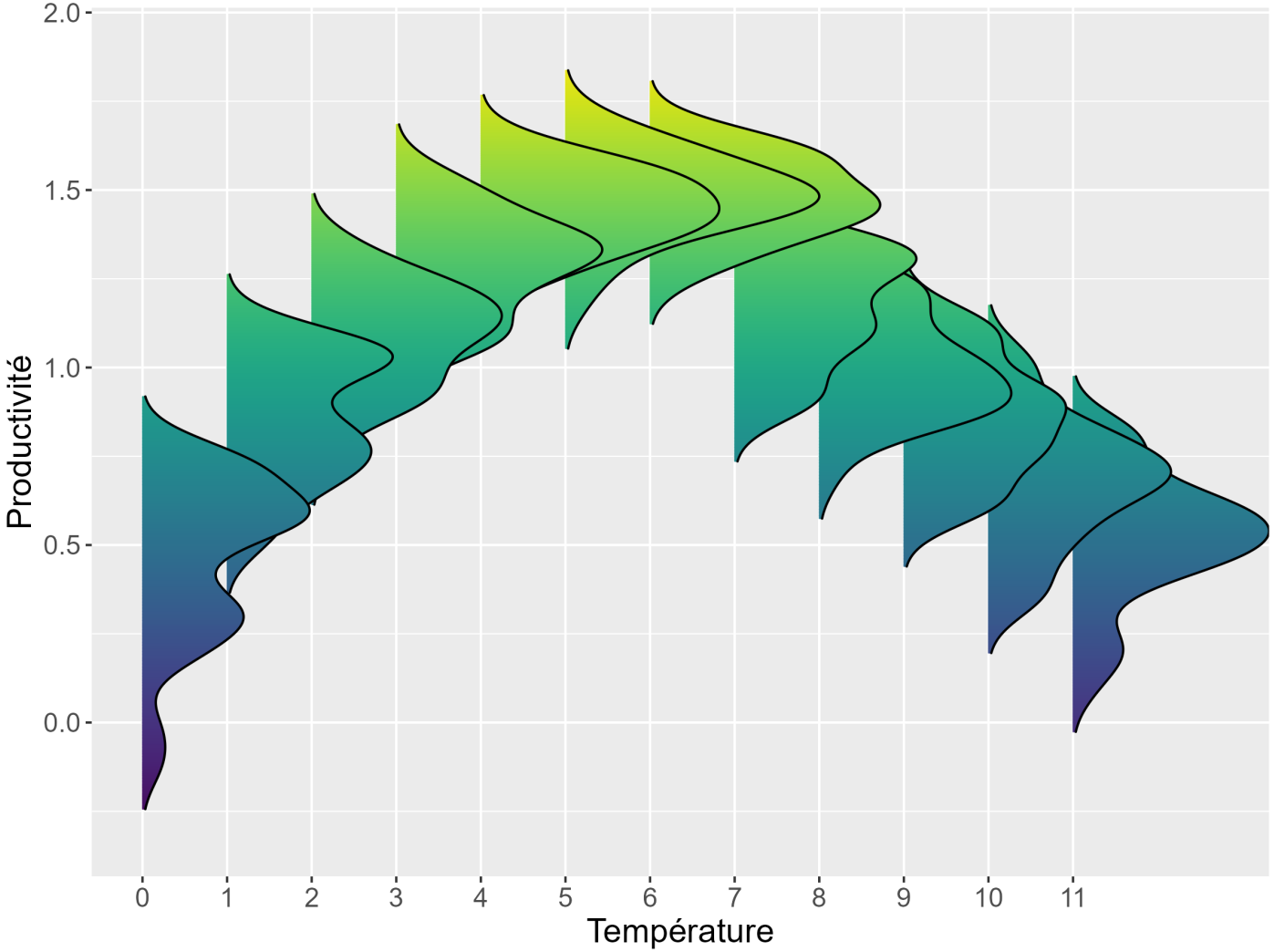
Une distribution cible

La distribution conditionnelle est dérivée à un point donné de l'espace des variables



Une distribution cible

La distribution conditionnelle est dérivée à un point donné de l'espace des variables



Comment estimer la distribution conditionnelle ?

Fonction de poids

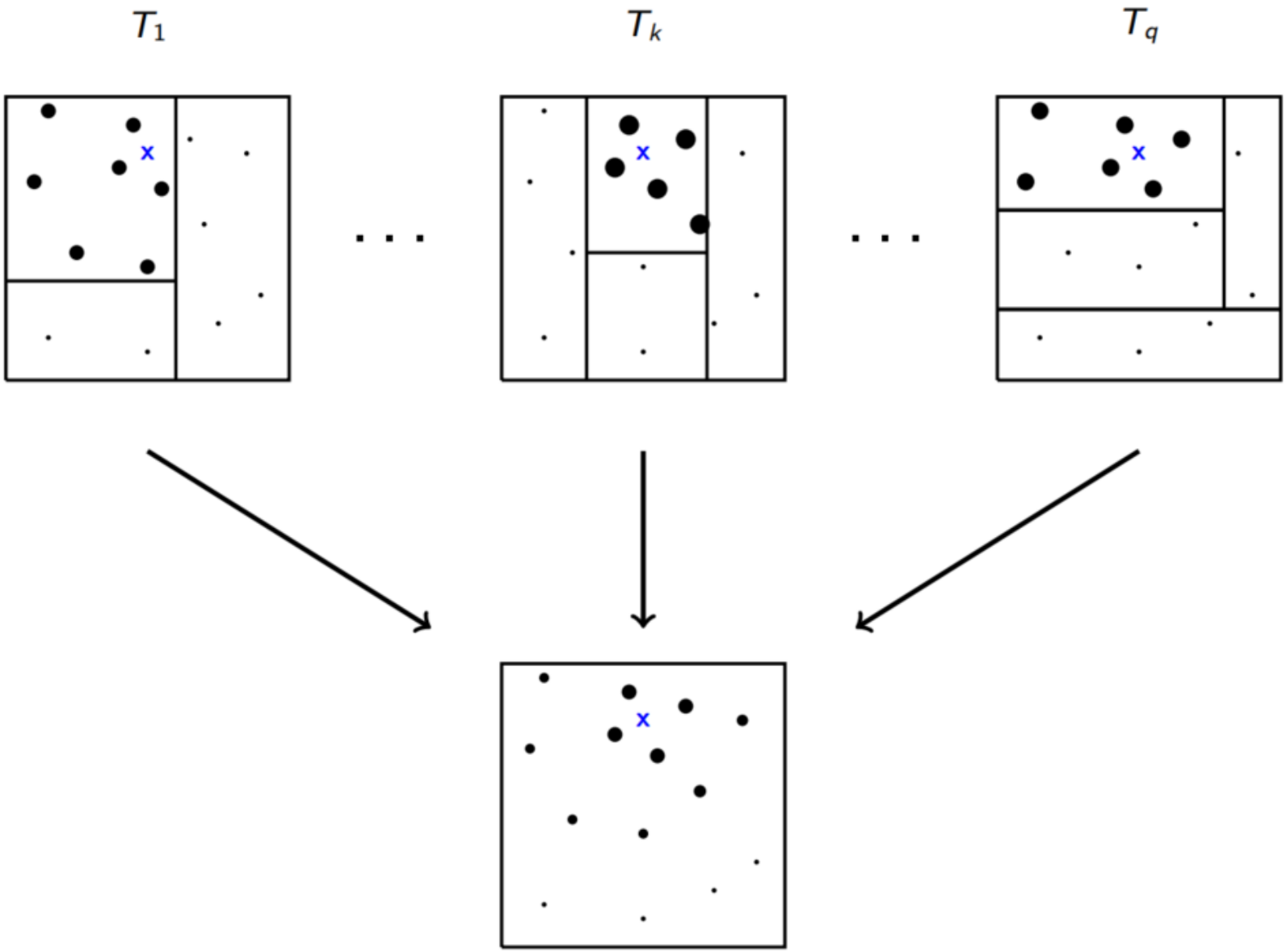
Une fonction de poids, $w_i(x)$, est définis pour chaque observation (i) pour un point de l'espace des variables (x) donnée sur les q arbres :

$$w_i(x) = \frac{1}{q} \sum_{k=1}^{q} \frac{\mathbf{1}(i \in leaf_k(x))}{|leaf_k(x)|} \tag{1}$$

En français

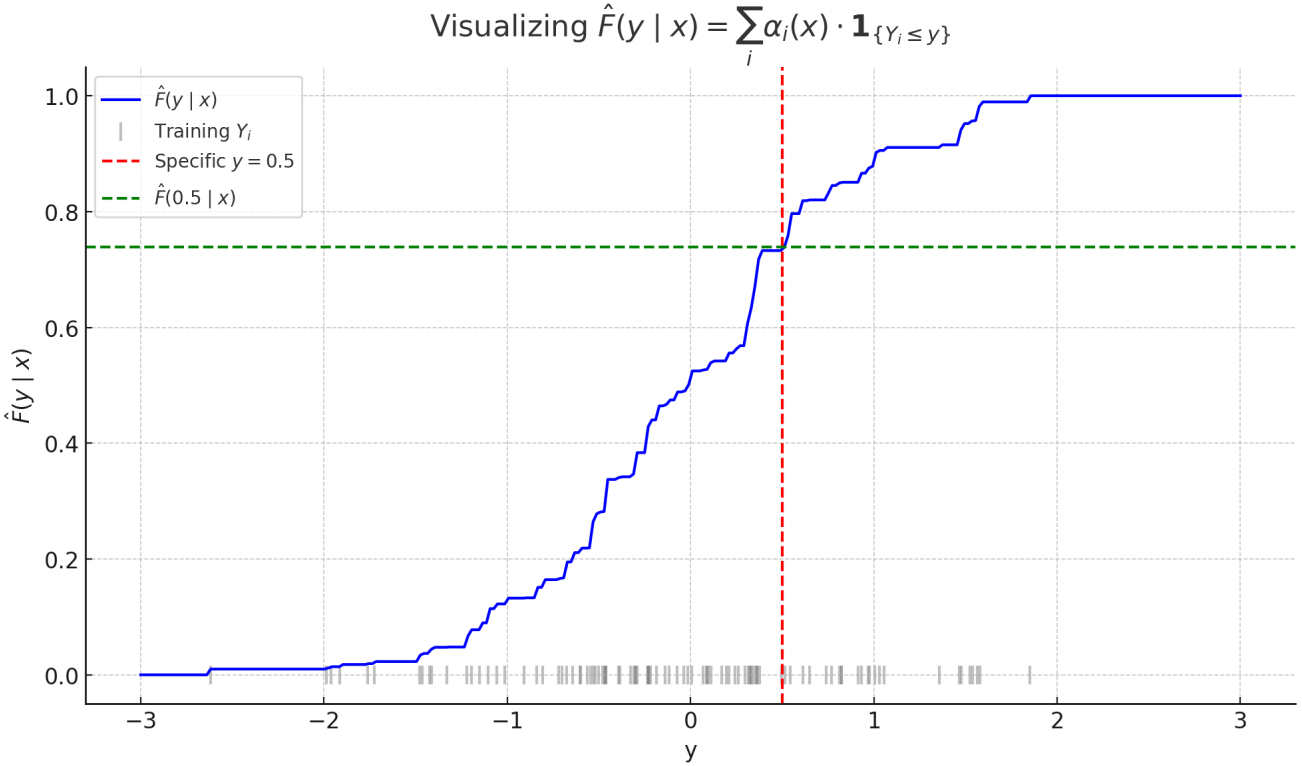
Chaque observation dans la feuille correspondant au point x contribue un poids non-nul pondérée par le nombre d'observation dans la feuille et moyenné sur l'ensemble des arbres.

La fonction de poids - en image



Du poids vers la distribution

La fonction de poids est ensuite utilisé pour estimer la distribution conditionnelle $P(Y|X = x)$ via une fonction de densité cumulative (CDF) :



$$F(y|X = x) = \sum_{i=1}^{n} w_i(x) * \mathbf{1}_{Y_i \le y} \tag{2}$$

Du poids vers la distribution - en pratique

En pratique cette distribution (et toutes les statistiques associées) peut être estimé par approche de Monte Carlo par ré-échantillonnage pondéré des y.

Intuition

Le vecteur de y ré-échantillonné avec pondération est un échantillon aléatoire de la distribution $P(Y|X = x)$

Limites :

1. Approximation correcte pour des tailles d'échantillons élevées
2. Si de nombreux w sont 0, le ré-échantillonage manque en diversité

En pratique - avec R

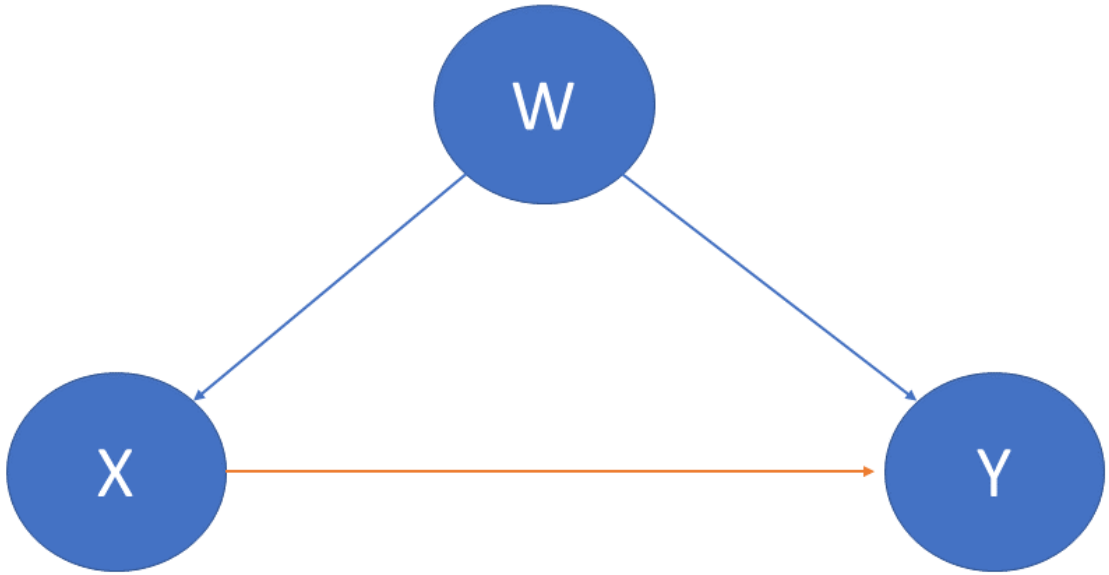
```r
# weighted resampling of y
y_samp <- sample(Y, probs = w, replace = TRUE)
# get estimated mean
y_mean <- mean(y_samp)
# get 75% quantile
y_q75 <- quantile(y_samp, probs = 0.75)
# get probability that y > 2
prob_2 <- sum(y_samp > 2)
```

Le package sous R contient également une méthode predict permettant d'obtenir de nombreuses statistiques (moyenne, quantile ...).

DRF et causalité

Problème

Lors de l'estimation d'un effet de causalité entre une variable X (un traitement) et une variable Y, des effets de confusions W peuvent ne pas avoir été contrôlé par un protocole expérimental.



DRF et causalité

2 étapes

1. Utiliser un DRF avec comme variables d'entrées les W et en sortie les X **et** les Y (et oui DRF marche aussi en multivarié)
2. Extraire les poids du DRF et les utilisé pour pondéré le modèle estimant la relation X - Y

Avantages

- Le DRF est une approche flexible non-paramétrique et non-linéaire
- Le second modèle peut être n'importe quel type de modèle permettant une pondération : (G)LM(M), GAM(M) ...
- L'estimation des poids étant dépendante de la position dans l'espace des W, cette approche permet d'étudier des interactions entre W et X sur Y

To take à la maison

- Les forêts aléatoires distributionnelles sont une extension des forêts aléatoires
- En se basant sur l'estimation d'une fonction de poids, une distribution conditionnelle $P(Y|X = x)$ peut être estimé et utilisé pour dériver un ensemble de statistique au-delà de la seule moyenne.
- Les DRF permettent également d'estimer plusieurs variables réponses (plusieurs Y)
- Les effets confondants peuvent également être contrôler en utilisant la pondération issue de DRF.