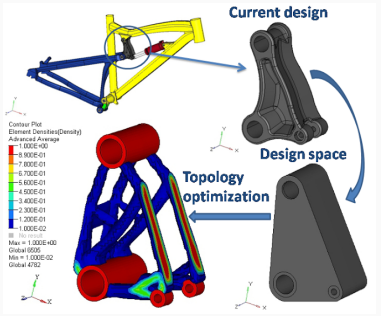
Revue des méthodes de régression quantile pour les boîtes noires aléatoires

Léonard Torossian, Victor Picheny, Robert Faivre, Aurélien Garivier

Problèmes associant variables de contrôle/variables subies



Réponse à un médicament

Variables de contrôle : Choix de la molécule, posologie...

Variable subies : Age, sexe, prédispositions génétiques...

Optimisation d'un portefeuille financier

Variable de contrôle : Stratégie d'investissement.

Variable subie : Comportement d'une grande quantité de personne.

Design industriel

Variables de contrôle : Matériaux, procédure de fabrication, géométrie...

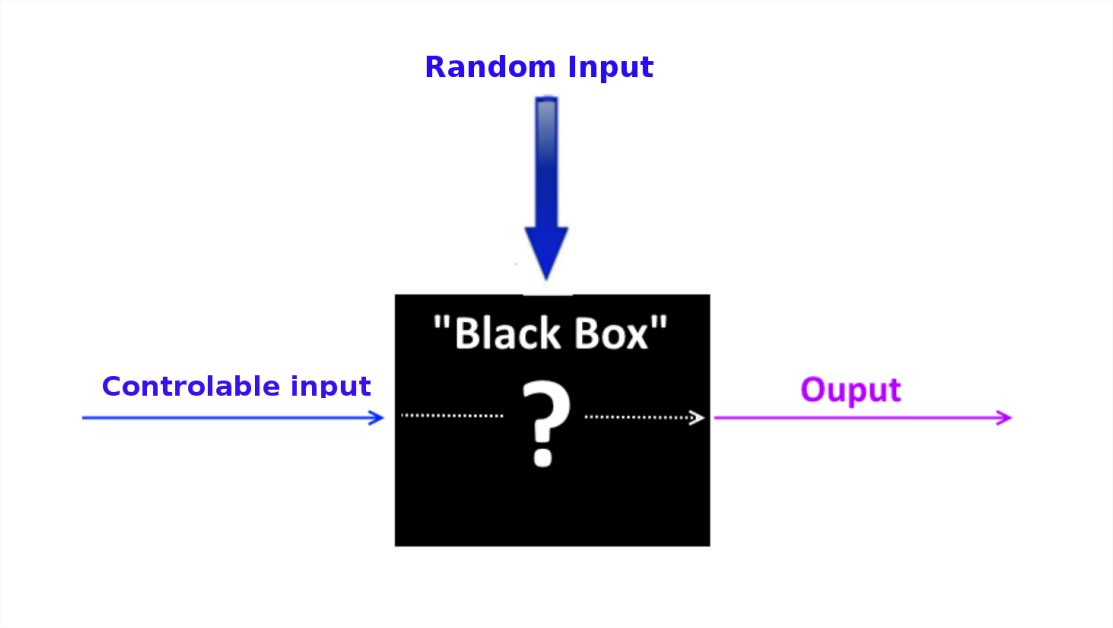
Variable subies : Défauts de fabrication, conditions d'utilisation.

Production agricole

Variables de contrôle : Choix de la variété, stratégie d'exploitation...

Variable subie : Le climat.

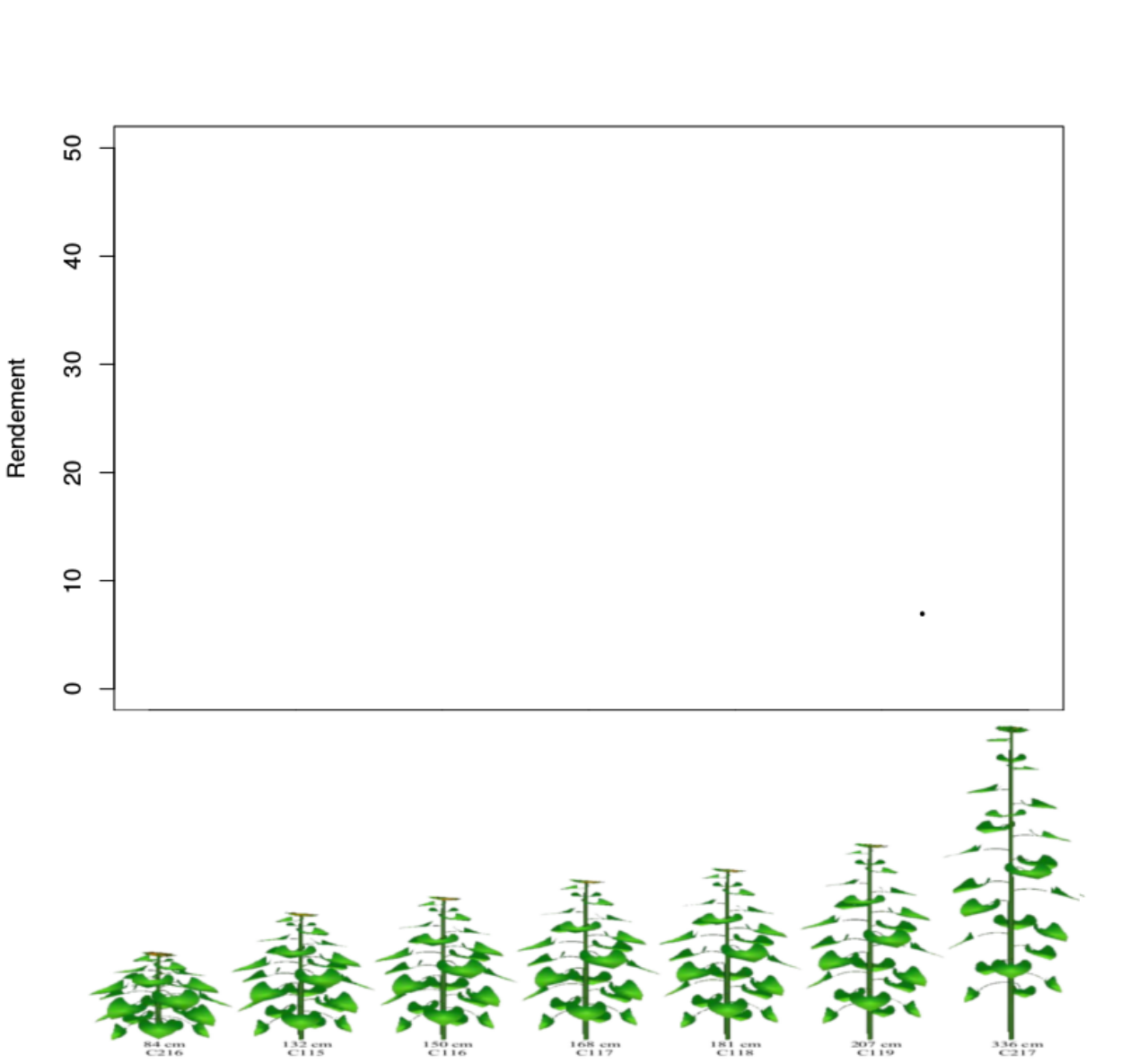
Boîte noire aléatoire



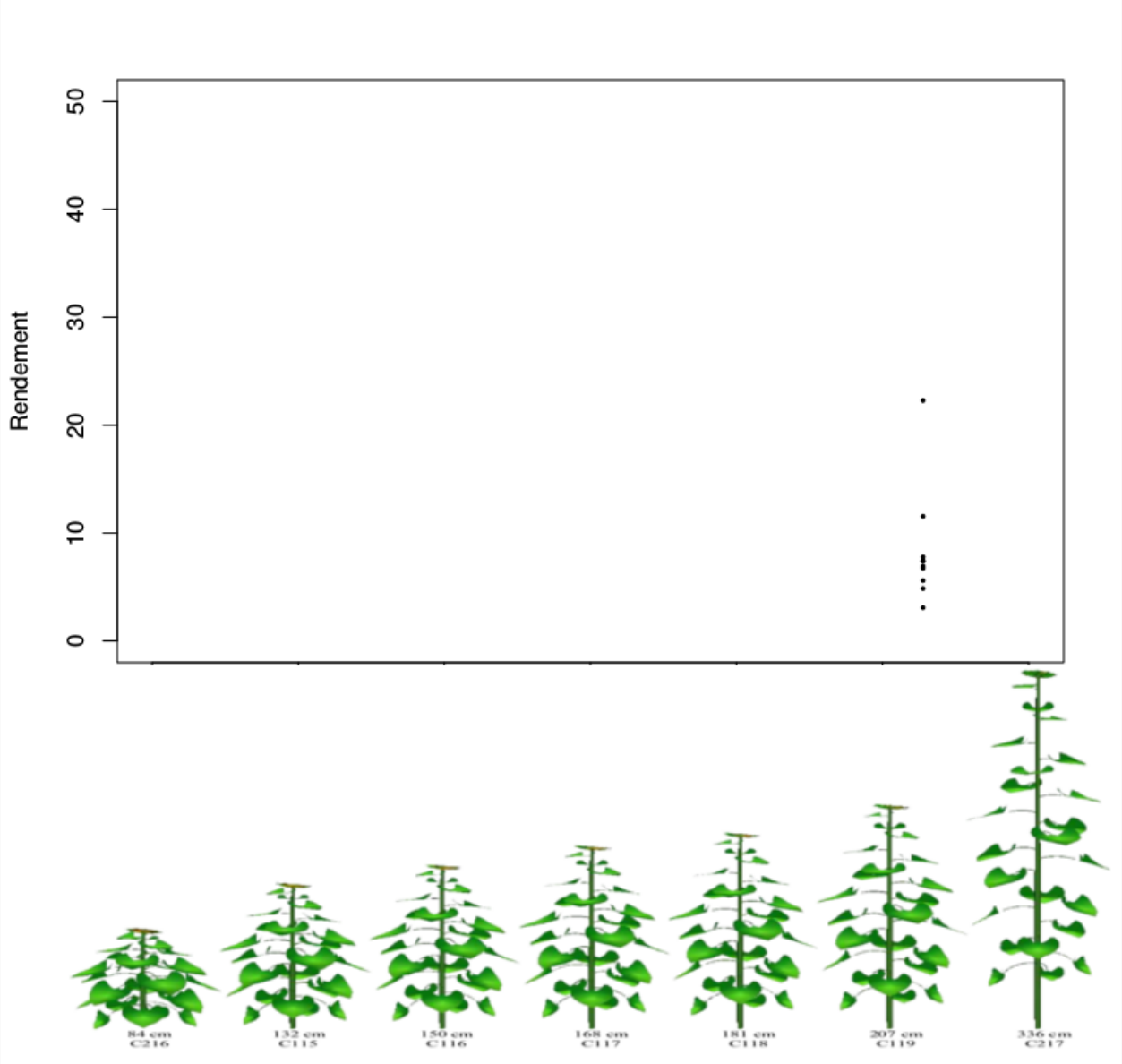
- Une boîte noire aléatoire est une fonction $f : \mathbb{X} \subset \mathbb{R}^d \times \Omega \to \mathbb{R}$
- L'entrée $x \in \mathbb{X}$ représente un paramètre fixé, il peut être choisi de manière déterministe ou aléatoire.
- L'entrée ω représente le caractère stochastique du code.
- **L'objectif est d'obtenir des informations sur la loi de**

$$Y(x) = f(x, \omega) \quad \text{pour tout} \quad x \in \mathbb{X}$$

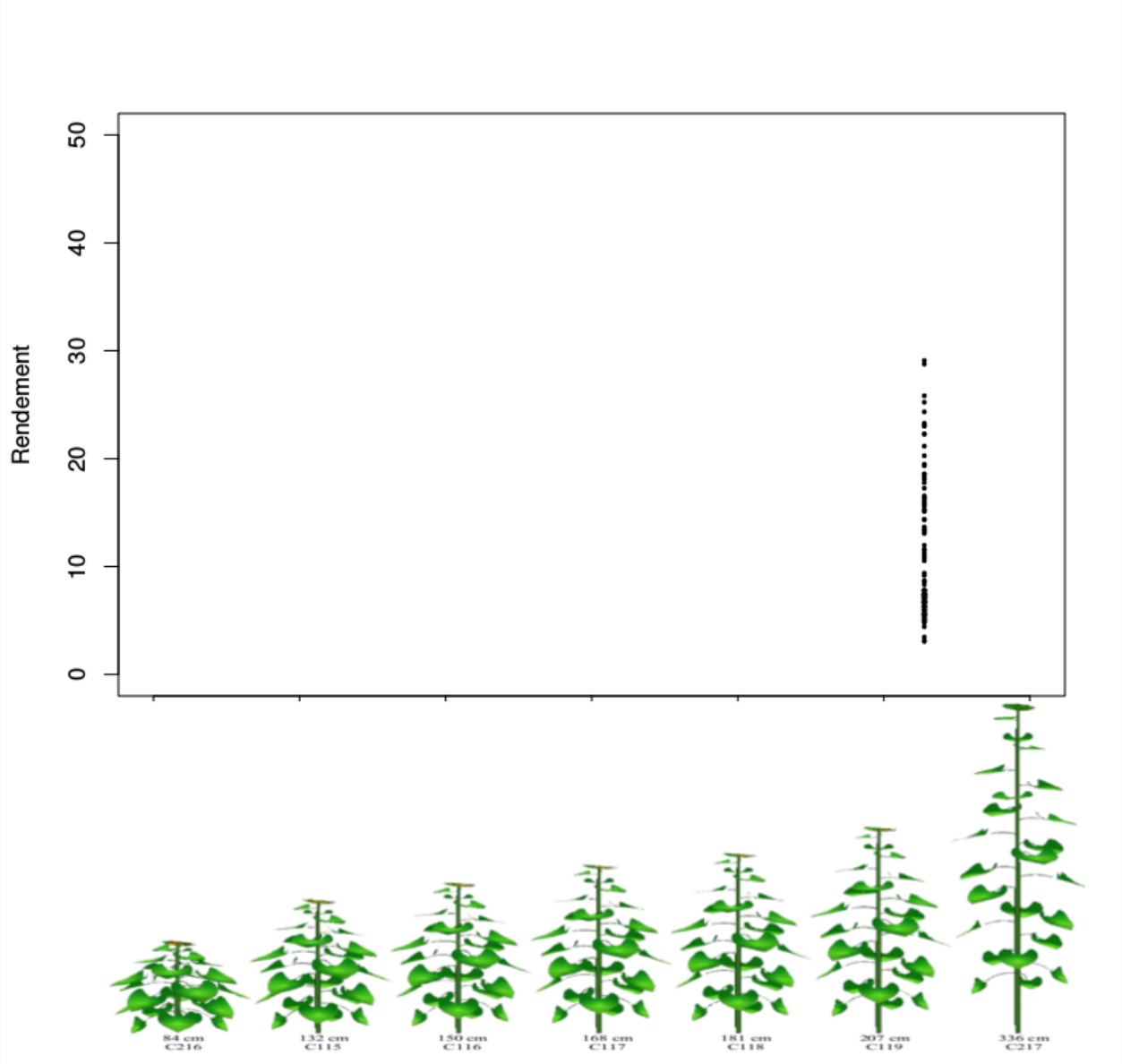
Boîte noire aléatoire : Illustration



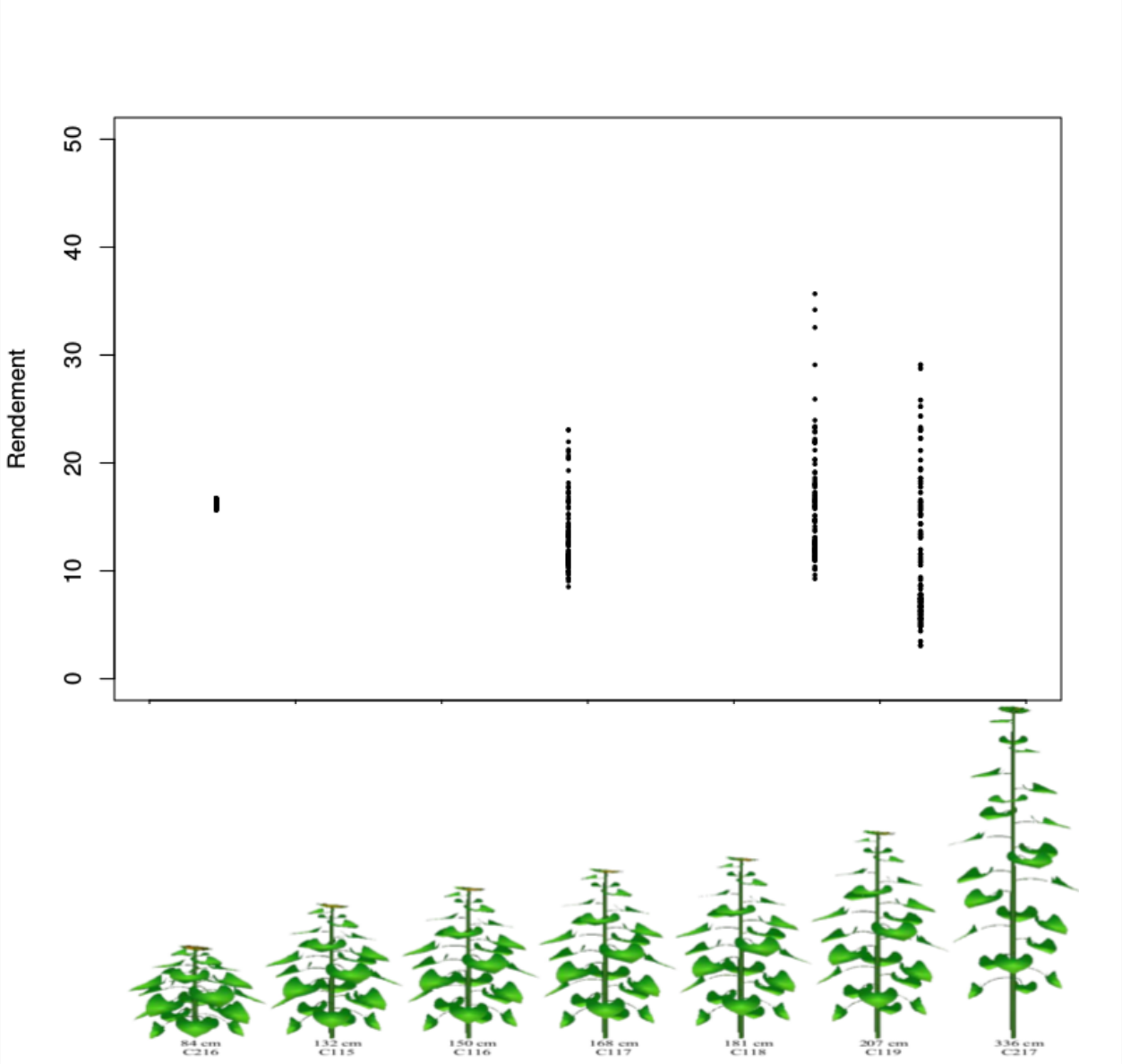
Boîte noire aléatoire : Illustration



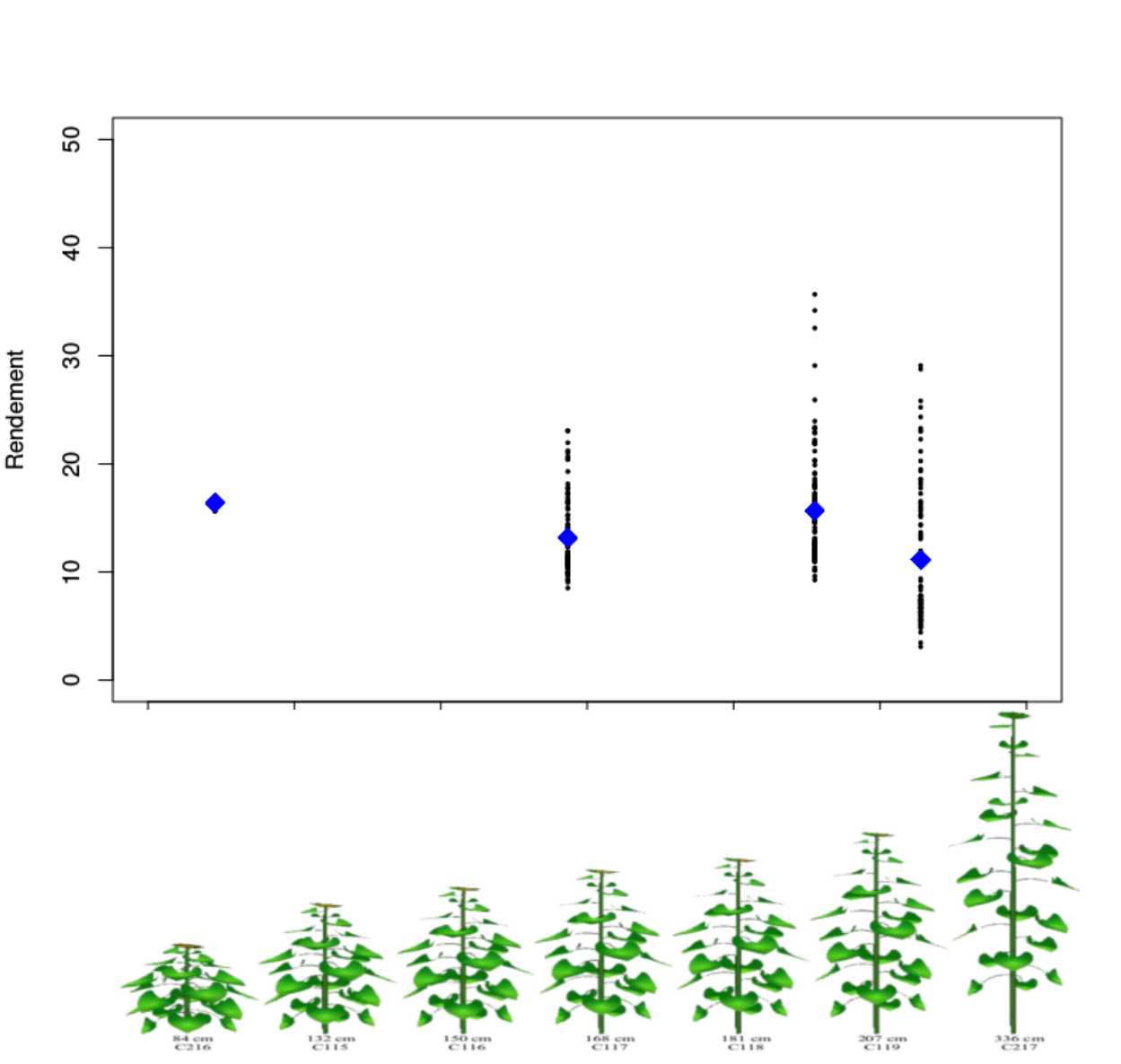
Boîte noire aléatoire : Illustration



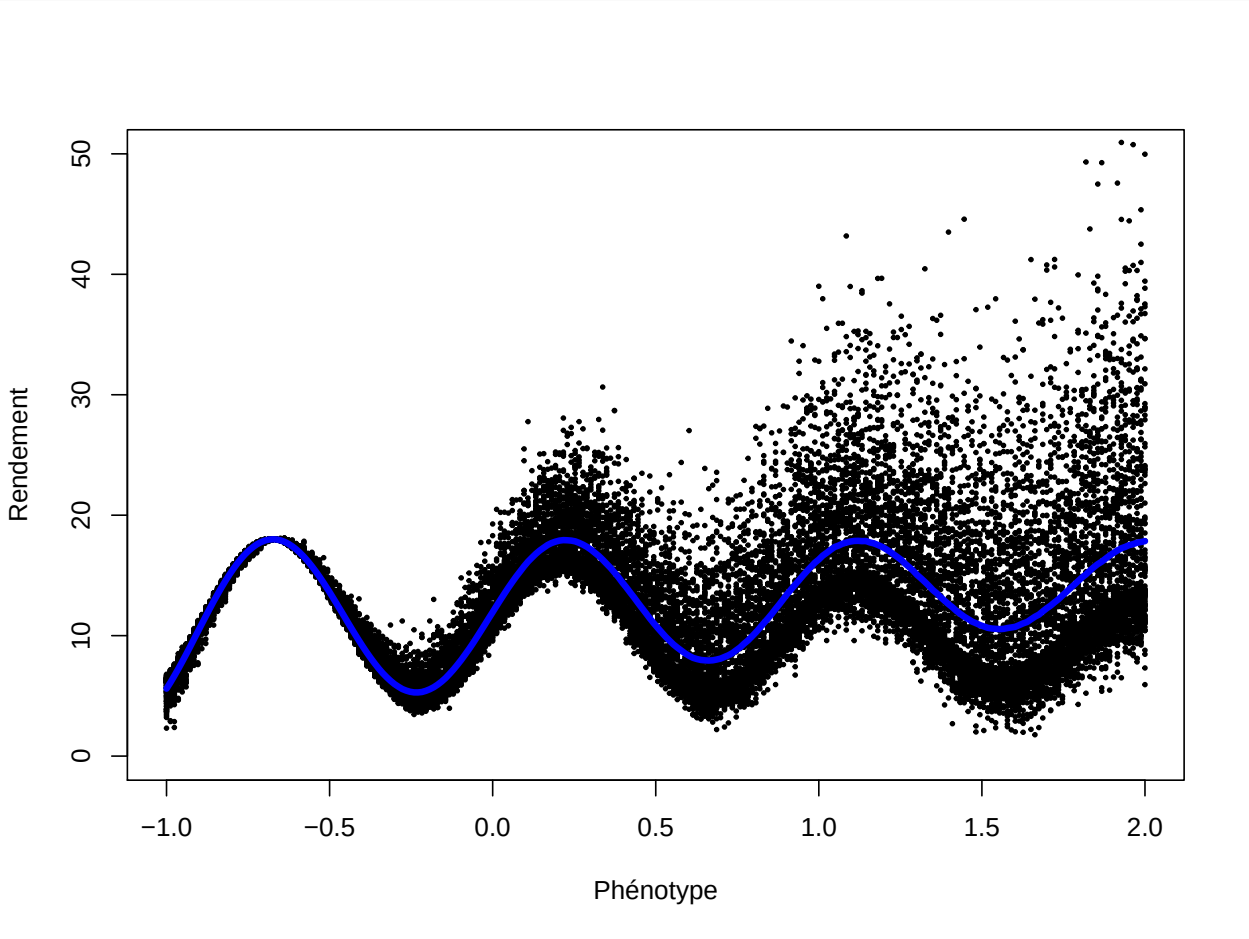
Boîte noire aléatoire : Illustration



Boîte noire aléatoire : Illustration

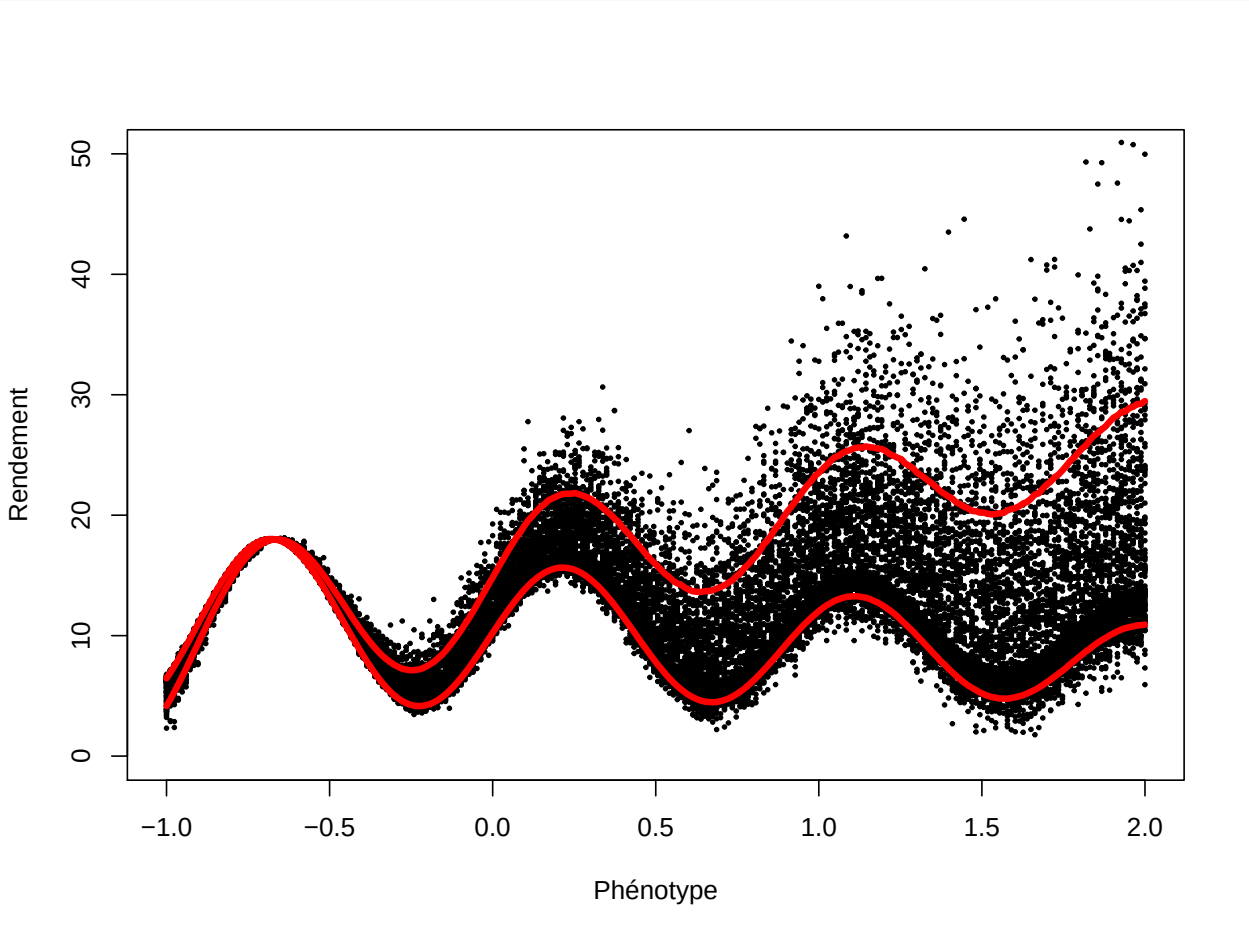


Les limites de la moyenne



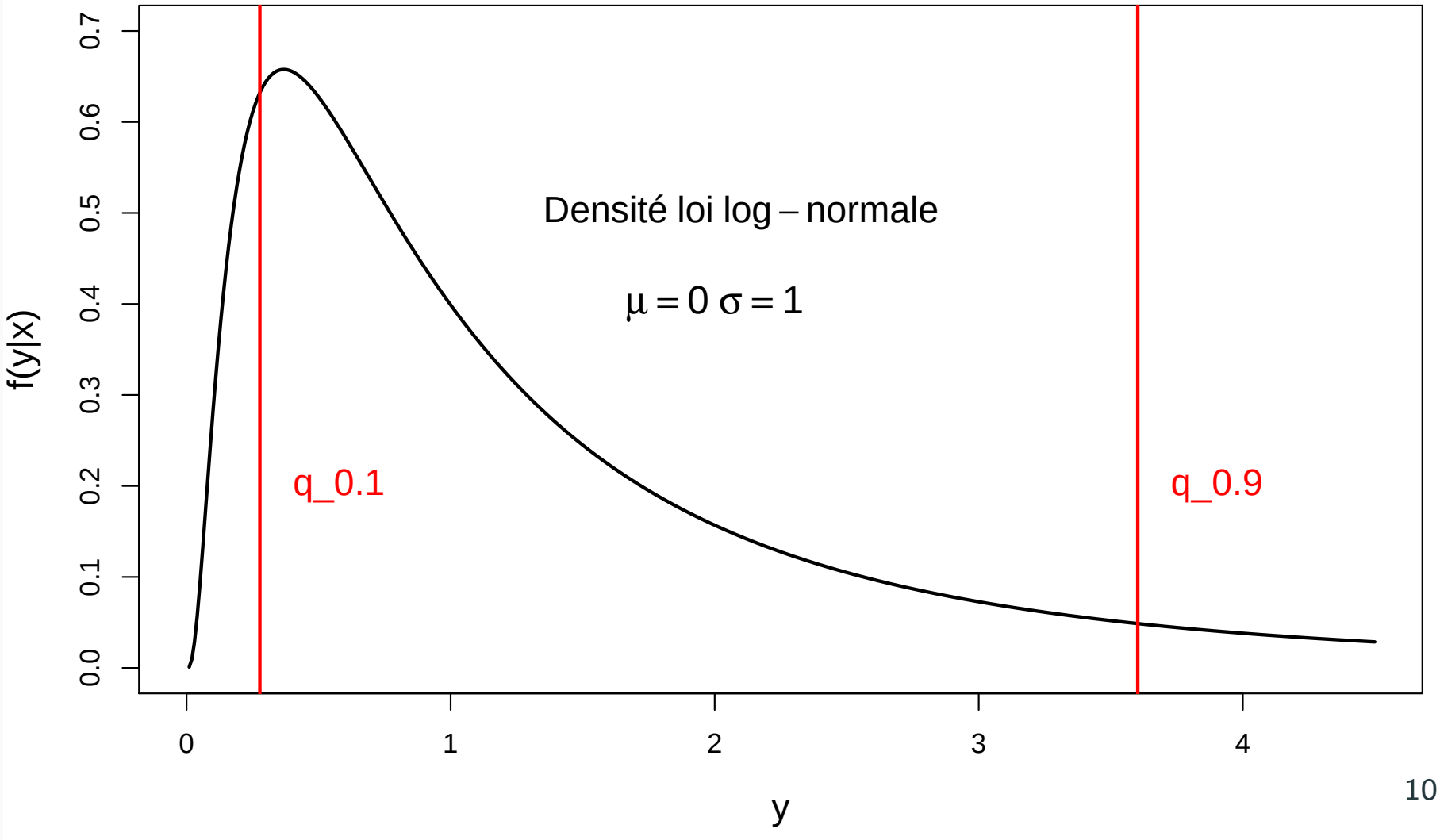
La moyenne donne peu d'information sur la loi $\mathbb{P}_x(Y)$.

Les quantiles



Les quantiles apportent plus d'information sur la loi $\mathbb{P}_x(Y)$.

Les quantiles



Cadre

- f est accessible uniquement par des évaluations ponctuelles (x, y)
- f peut être non linéaire en $\mathbb{X}$.
- La variance de Y et la forme de $\mathbb{P}_x(Y)$ peuvent dépendre de $\mathbb{X}$.
- Un appel à f peut être très coûteux.
- La taille de l'échantillon $\mathcal{D}_n = ((x_1, y_1), .., (x_n, y_n))$ est considérée petite.

$\rightarrow$ **Utilisation de métamodèles** *aka* **émulateurs statistiques.**

Questions motivants l'article

- Quelles sont les grandes classes de métamodèle de quantile ?
- Y a-t-il un métamodèle meilleur que tous les autres ?
- Comment se comportent les métamodèles en fonction du nombre de points, de la dimension et de la valeur de la densité au voisinage du quantile ?
- Les métamodèles sont-ils capables d'estimer les quantiles d'une distribution dont la forme ou la variance varie fortement en espace ?

Les métamodèles

Métamodèles basés sur les statistiques d'ordre

- K-plus proches voisins (KN)
  [Bhattacharya and Gangopadhyay, 1990]
- Forêts aléatoires (RF) [Meinshausen, 2006]

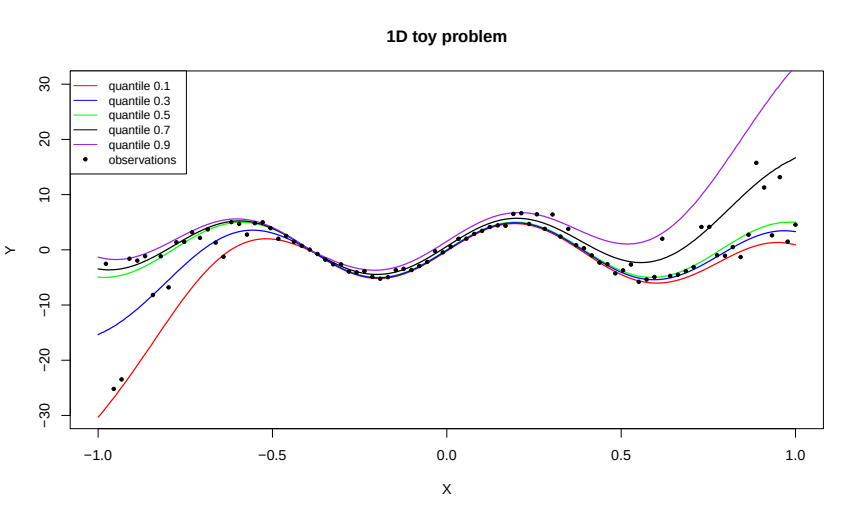
Métamodèles basés l'analyse fonctionnelle

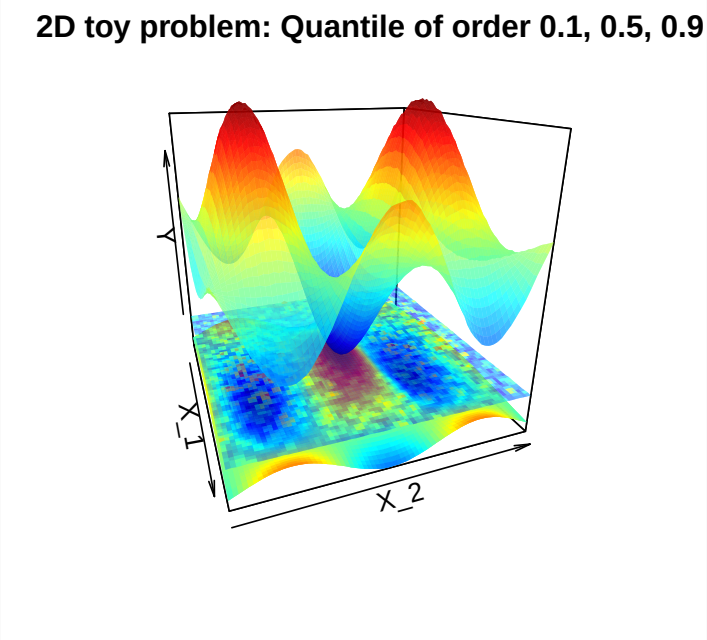
- Réseaux de neurones (NN) [Cannon, 2011]
- Régression RKHS (RK)[Takeuchi et al., 2006]

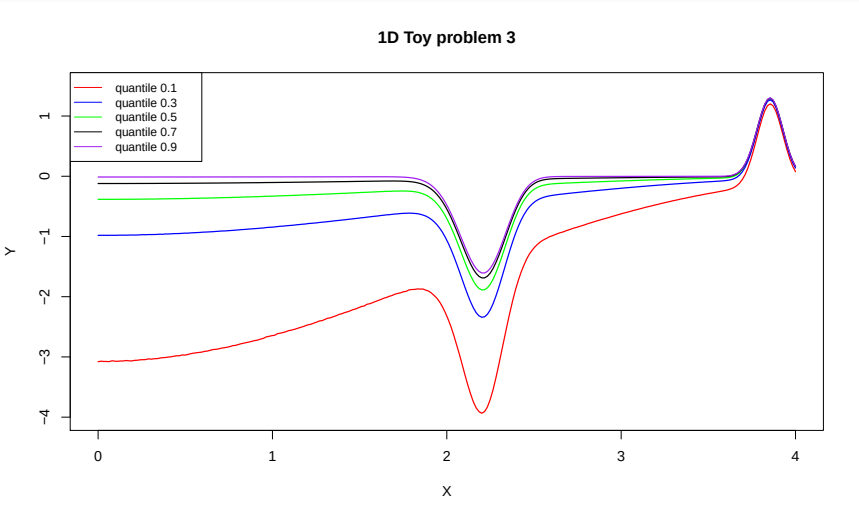
Métamodèles basés sur les processus aléatoires/approche bayésienne

- Quantile kriging (QK) [Plumlee and Tuo, 2014]
- Regréssion bayésienne variationnelle (BV)
  [Abeywardana and Ramos, 2015]

Cas tests [Casadebaig et al., 2011]







Plan d'expérience

Dim	Data size (no repetitions)				Data size (with repetitions)			
1	40	80	160	320	5 (8)	10 (8)	10 (16)	16 (20)
2	100	200	400	800	10 (10)	20 (10)	25 (16)	40 (20)
9	250	500	1000	2000	25 (10)	50 (10)	100 (10)	100 (20)

- Pour chaque taille on génère 10 plans d'expériences sous le critère *maxminLHS*.

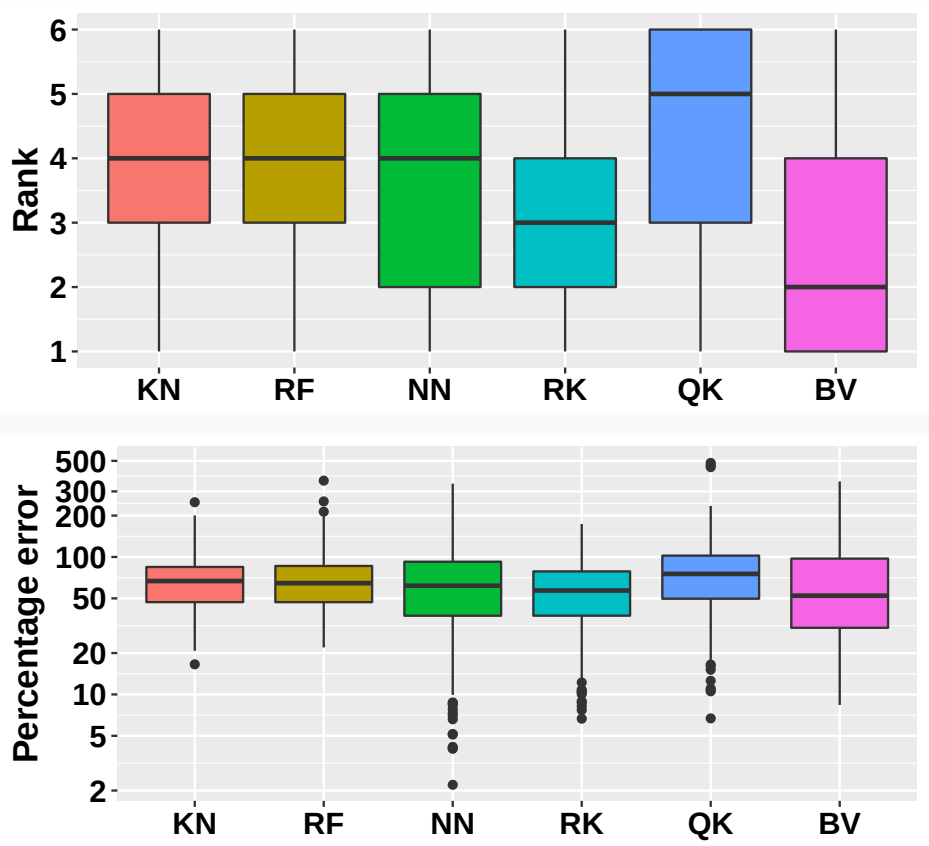
Critères de comparaison

Les **problèmes** et les **résultats** obtenus sont très **hétérogènes**.

Pour **agréger** les résultats deux indicateurs ont été sélectionnés :

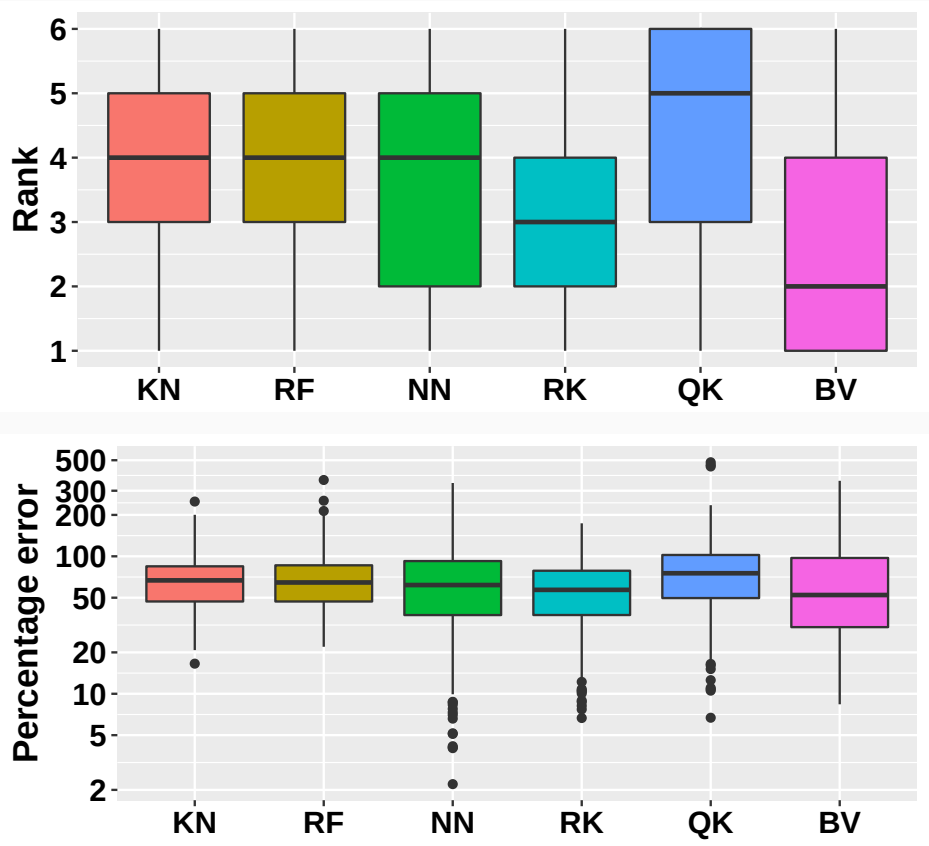
- $$E_{cq}(method_i) = \frac{E_{L2}(method_i)}{E_{L2}(QC)} \times 100$$
- Le rang par rapport à l'erreur L^2 sur chaque problème

Y a-t-il un métamodèle meilleur que tous les autres ?



Y a-t-il un métamodèle meilleur que tous les autres ?

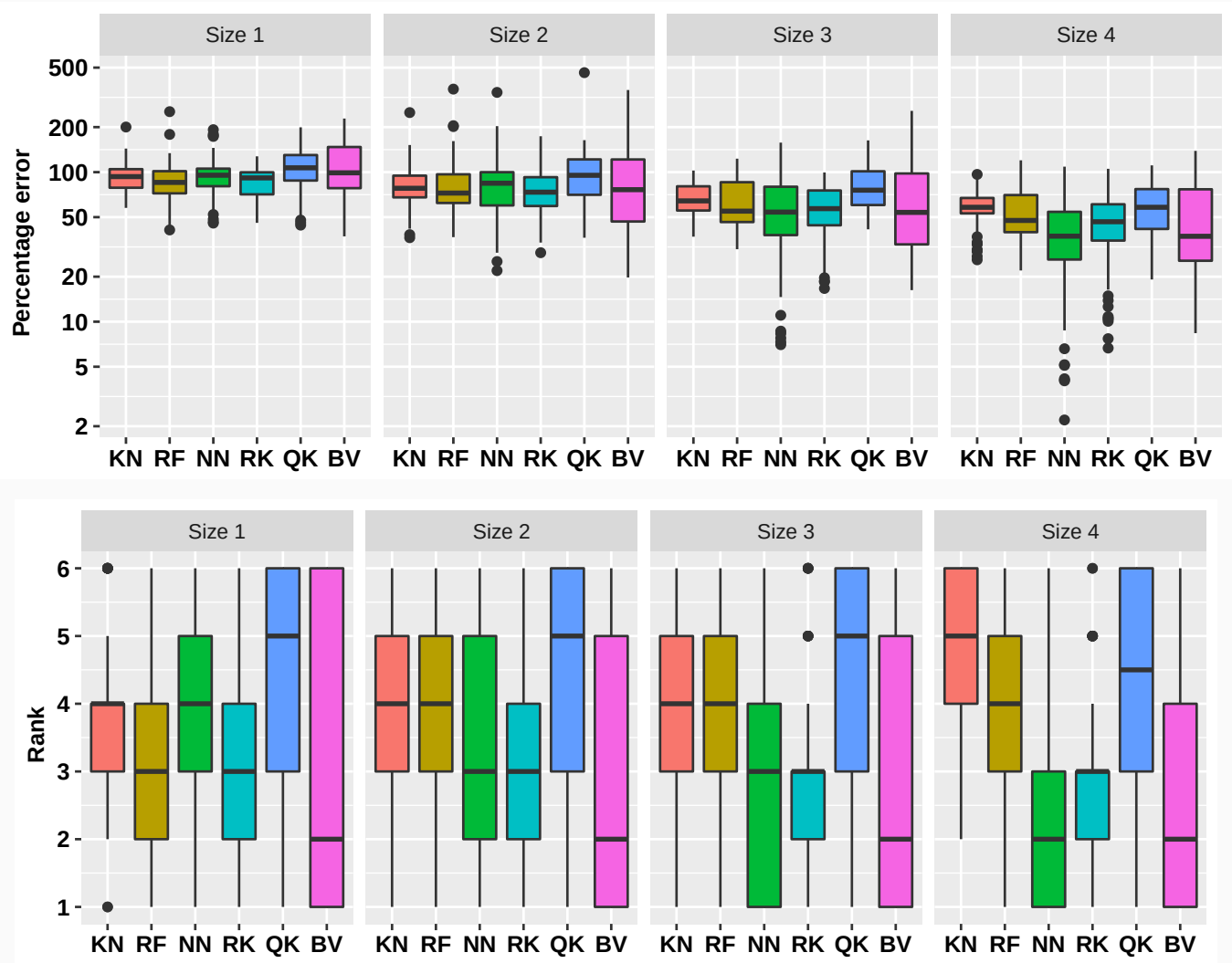
Y a-t-il un métamodèle meilleur que tous les autres ?



Y a-t-il un métamodèle meilleur que tous les autres ?

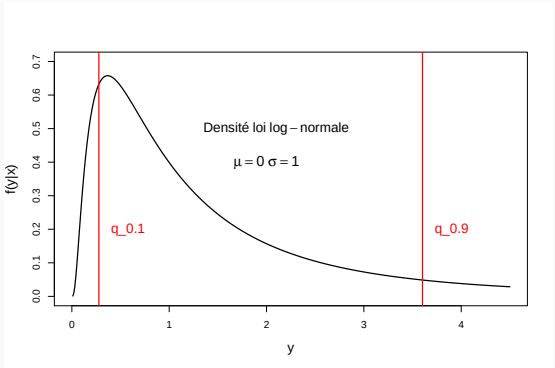
$\rightarrow$ **BV est en tête mais le contraste n'est pas net.**

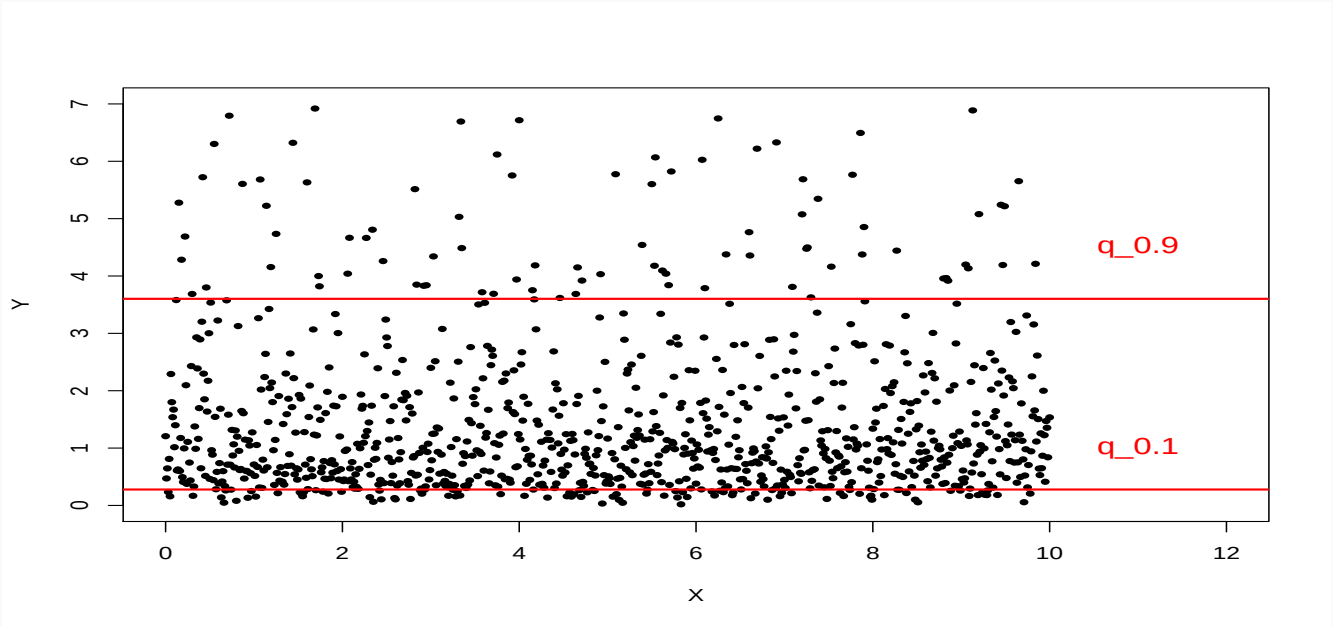
Peut-on extraire des comportements propres à chaque méta-modèle ? Focus sur l'influence du nombre de points



Peut-on extraire des comportements propres à chaque méta-modèle ? Focus sur l'influence de la valeur de la densité

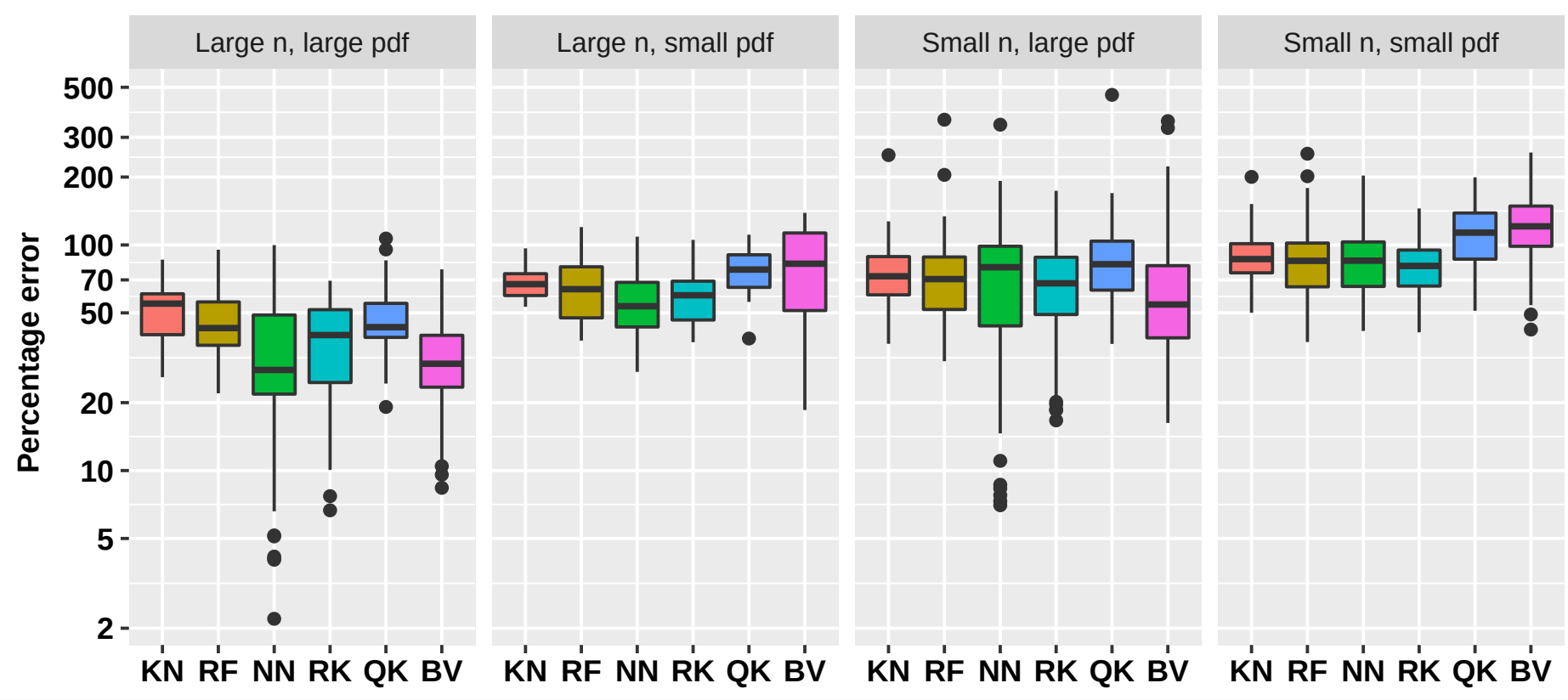
Supposons que pour tout $x \in \mathcal{X}, \mathbb{P}_X(Y)$ est log-normale avec $\mu = 0,\ \sigma = 1$.





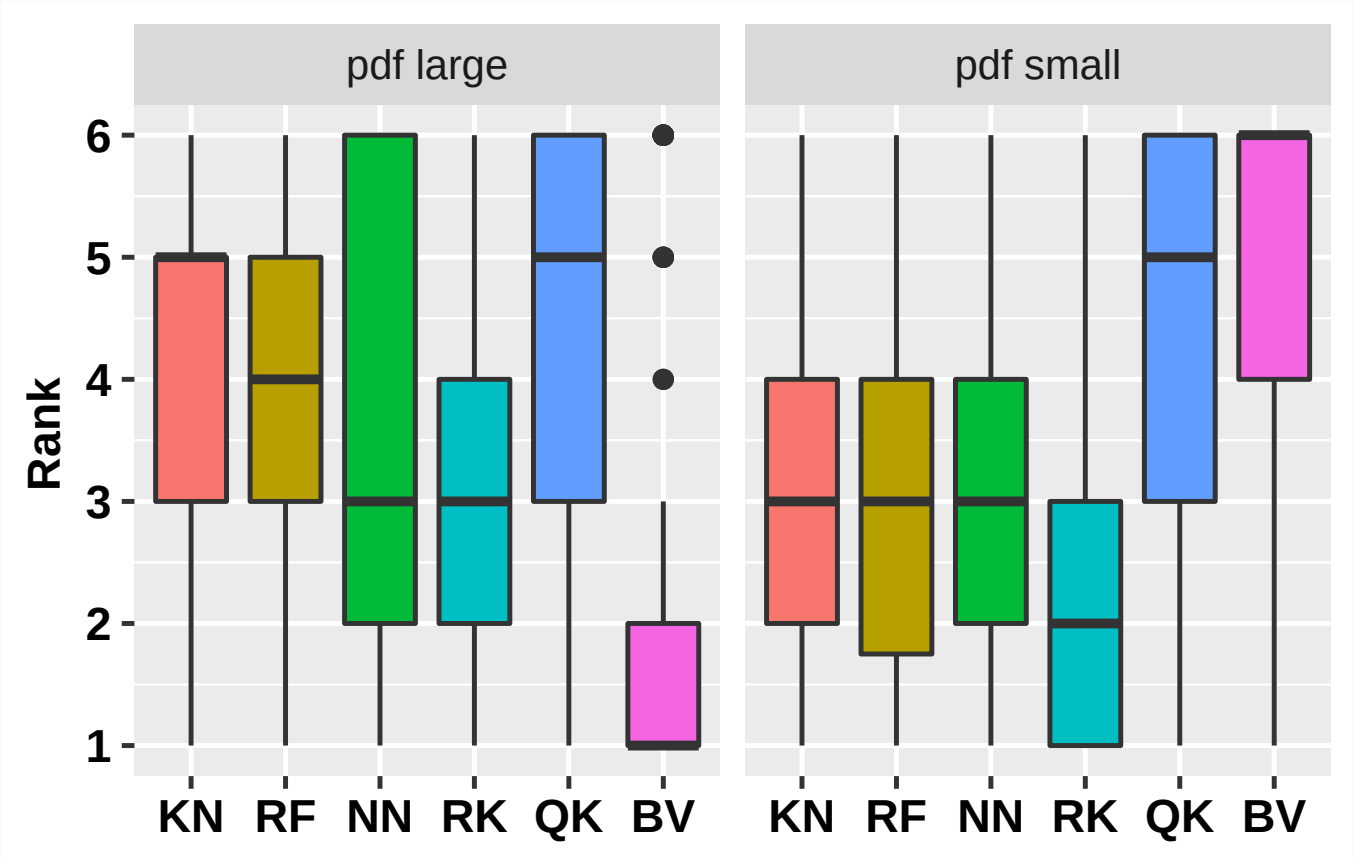
→ **Information plus concentrée autour de $q_{0.1}$ qu'autour de $q_{0.9}$**

Peut-on extraire des comportements propres à chaque méta-modèle ? Focus sur l'influence de la valeur de la densité



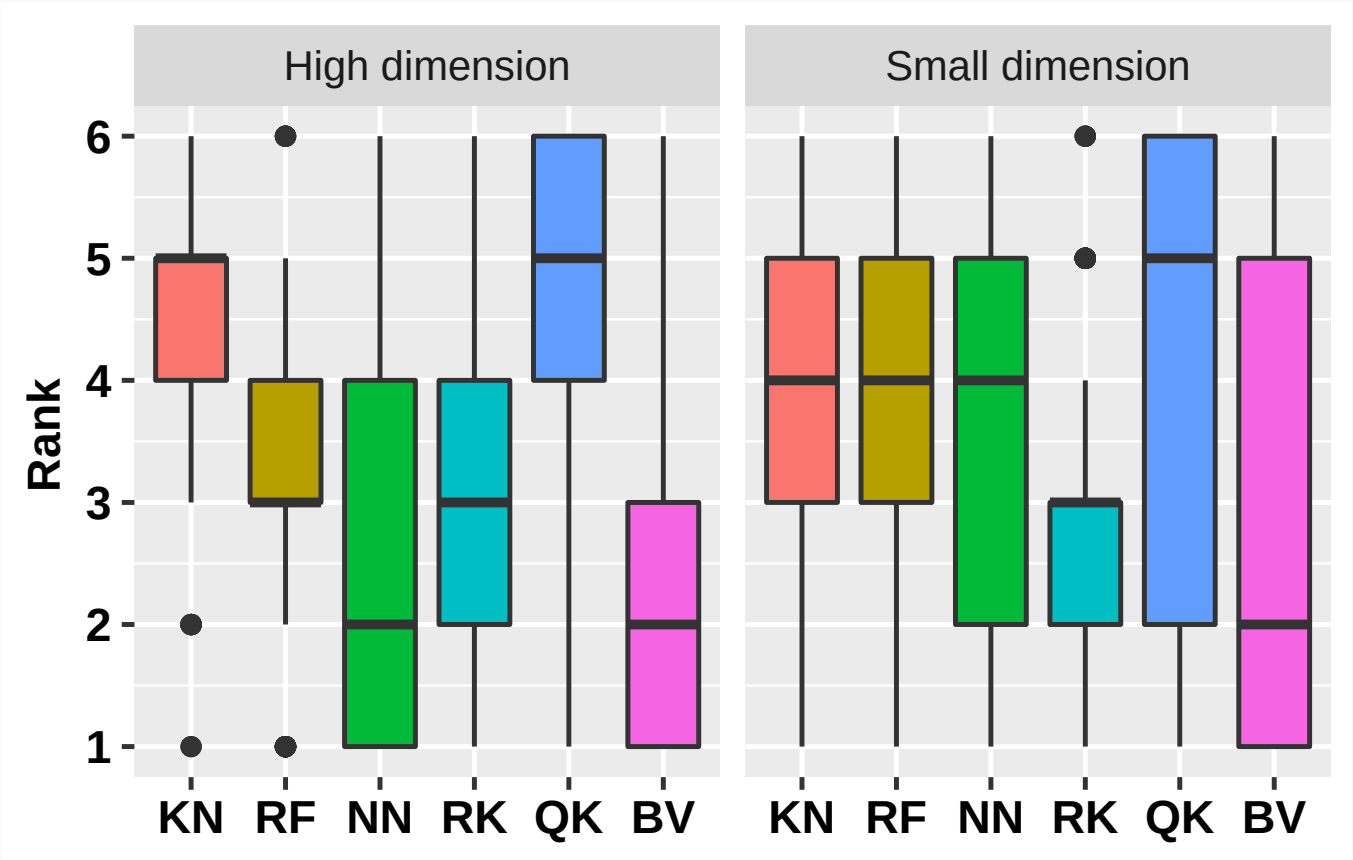
- Les variables *nombre de points* et *valeur de la densité* ont un impact fort sur E_{CQ}.

Peut-on extraire des comportements propres à chaque méta-modèle ? Focus sur l'influence de la valeur de la densité



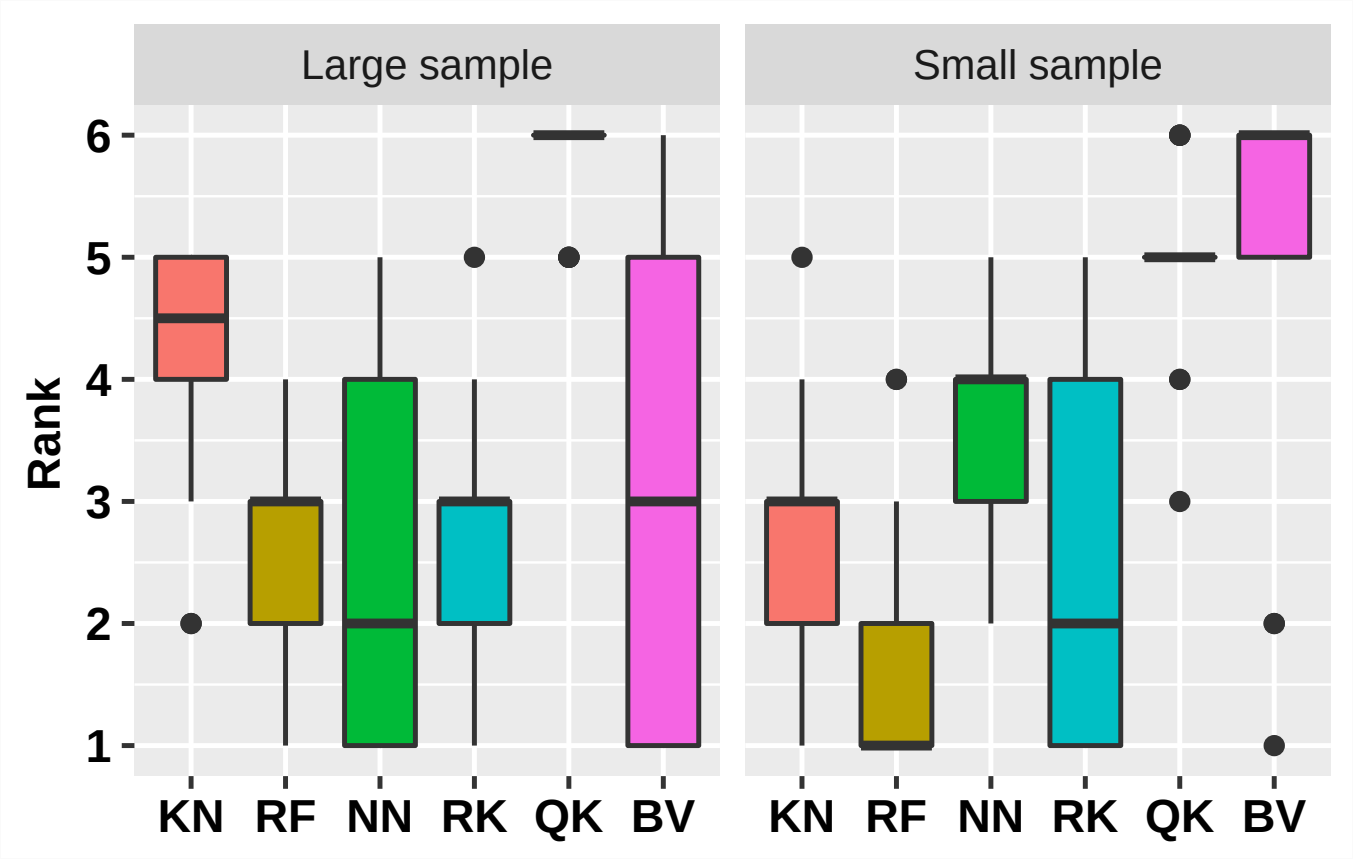
- BV est le métamodèle dépendant le plus de la valeur de la densité.

Peut-on extraire des comportements propres à chaque méta-modèle ? Focus sur l'influence de la dimension



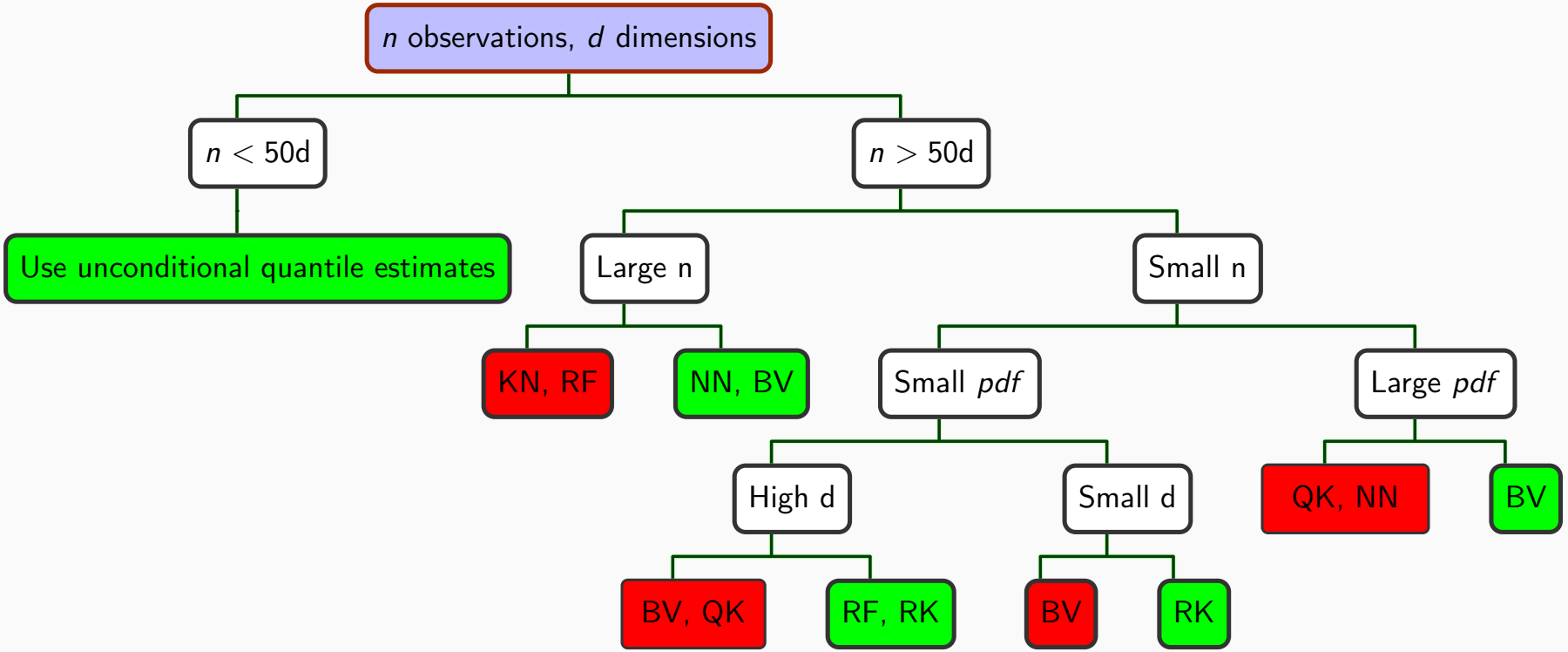
- NN et BV s'améliorent en grande dimension.
- RF en KN comparables en petite dimension mais RF meilleur en grande dimension.

Focus sur l'influence du nombre de points en grande dimension avec petite valeur de densité



- RF et RK sont en tête quand le problème est difficile, *i.e* petit nombre de points.
- NN et BV rattrapent quand le nombre de points augmente.

Synthèse


n observations, d dimensions
$n < 50d$
$n > 50d$
Use unconditional quantile estimates
Large n
Small n
KN, RF
NN, BV
Small *pdf*
Large *pdf*
High d
Small d
QK, NN
BV
BV, QK
RF, RK
BV
RK

References i


Abeywardana, S. and Ramos, F. (2015).
Variational inference for nonparametric bayesian quantile regression.
In *AAAI*, pages 1686–1692.


Bhattacharya, P. K. and Gangopadhyay, A. K. (1990).
Kernel and nearest-neighbor estimation of a conditional quantile.
The Annals of Statistics, pages 1400–1415.


Cannon, A. J. (2011).
Quantile regression neural networks: Implementation in r and application to precipitation downscaling.
Computers & geosciences, 37(9):1277–1284.

References ii


Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., and Debaeke, P. (2011).
Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments.
Agricultural and forest meteorology, 151(2):163–178.


Meinshausen, N. (2006).
Quantile regression forests.
Journal of Machine Learning Research, 7(Jun):983–999.

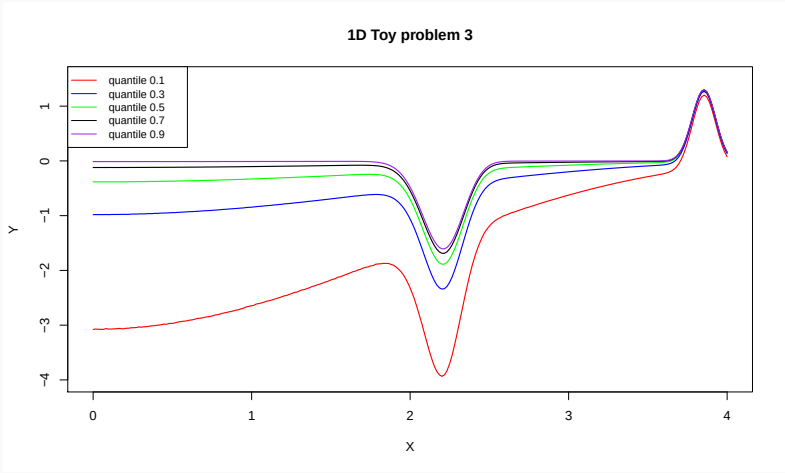

Plumlee, M. and Tuo, R. (2014).
Building accurate emulators for stochastic simulations via quantile kriging.
Technometrics, 56(4):466–473.

References iii


Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006).
Nonparametric quantile estimation.
Journal of Machine Learning Research, 7(Jul):1231–1264.

Merci de votre attention !

Ces métamodèles sont-ils capable d'estimer les quantiles d'une distribution dont la variance change fortement en espace ?



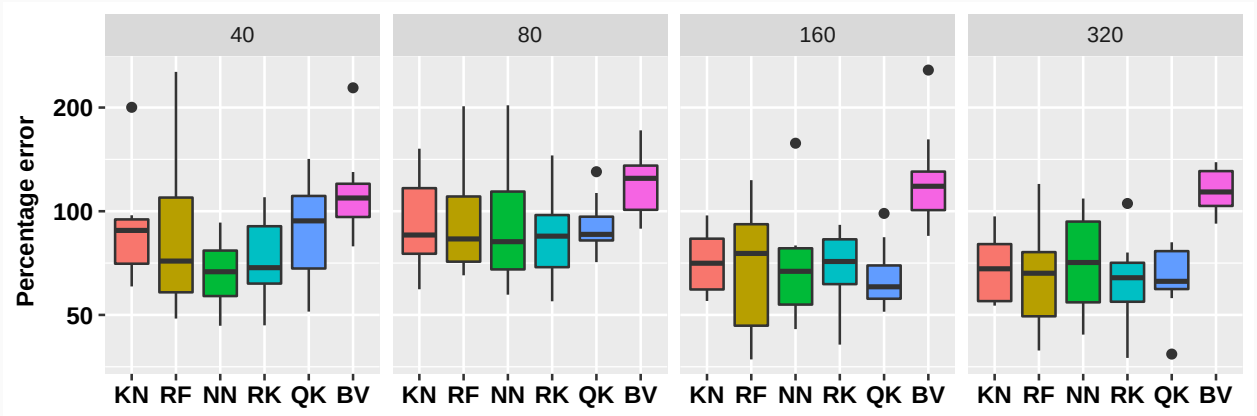
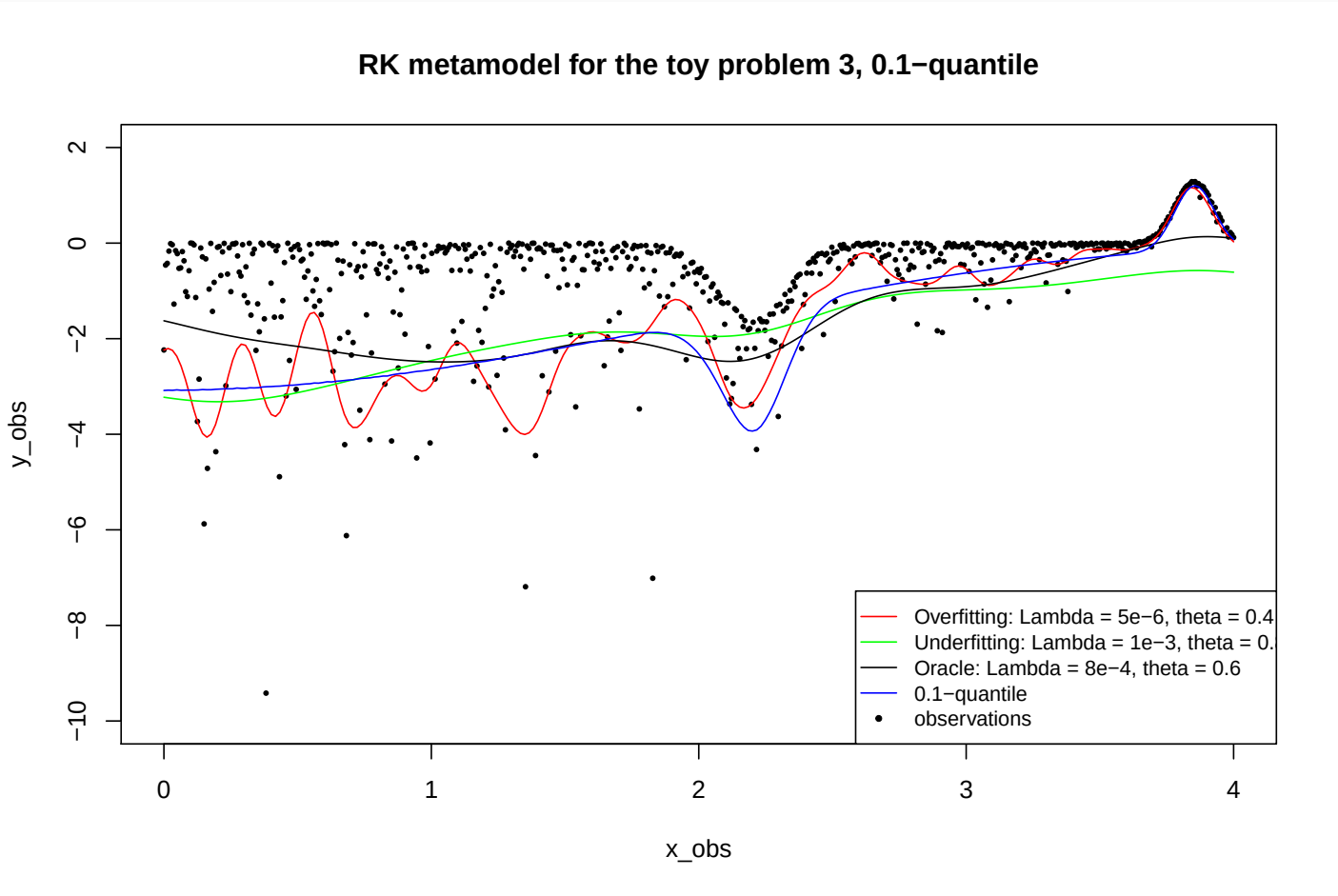


Figure 1: Erreur associée à l'estimation du quantile d'ordre 0.1

La performance plafonne !

Les métamodèles sont-ils capable d'estimer les quantiles d'une distribution dont la variance change fortement en espace ?



Le métamodèles étudiés ne semblent pas faits pour traiter des problèmes fortement hétérocédastiques.