JMS

Journées de méthodologie statistique de l'Insee

2018

Pondération pour correction de la non-réponse totale par des méthodes de machine learning.

Brigitte GELEIN, Ensai,
David HAZIZA, Montréal University,
David CAUSEUR, Agrocampus Ouest.

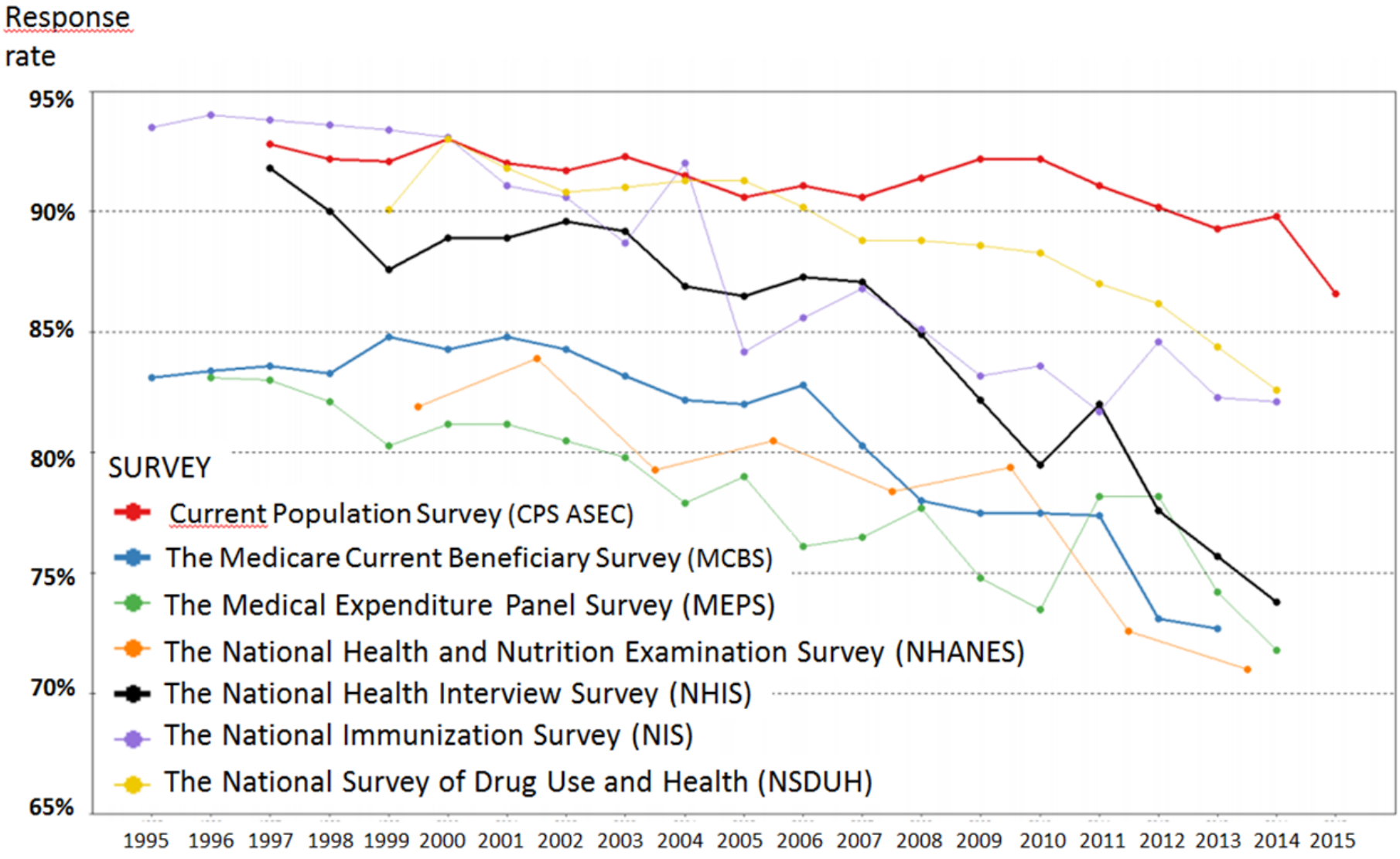
Introduction

Recensements, **enquêtes** ou encore **sources administratives** : peu importe l'origine des données, elles sont toutes susceptibles de présenter des **données manquantes.**

- **Biais des estimateurs ponctuels :**
  apparaît lorsque répondants et non répondants se comportent différemment au regard des variables d'intérêt.
- **Augmentation de la variance des estimateurs ponctuels :**
  se produit en raison d'une taille d'échantillon réduite de l'échantillon des répondants en comparaison avec l'échantillon initial.

Le traitement de la non-réponse est d'un intérêt pratique très important étant donnée la **baisse constante du taux de réponse** aux enquêtes depuis plusieurs décennies.

Declining Response Rates in Federal Surveys in the USA



Source : Czajka J.L., Beyler A. (2016). Declining Response Rates in Federal Surveys : Trends and Implications, *Mathematica Policy Research*.

Introduction

La non réponse totale :

- **aucune information utilisable** n'est disponible pour une observation de l'échantillon,
- souvent traitée par des procédures d'**ajustement des poids** :
  1. éliminer les non répondants du fichier,
  2. ajuster les poids des répondants par l'inverse des probabilités de réponse estimées.

Clé : la disponibilité de **variables auxiliaires** liées aux probabilités de réponse et aux variables d'intérêt (Little et Vartivarian, 2005 ; Haziza et Beaumont, 2017).

Introduction

Etude de **l'estimation de probabilités de réponse** dans un contexte de pondération pour traiter la non réponse totale.

1. **Cadre théorique** global et notations.
2. **Modélisation** de la non réponse par apprentissage supervisé.
3. **Modification** des probabilités brutes estimées.
4. Vaste étude par **simulation** pour comparer différentes méthodes de machine learning.

Cadre théorique et notation

- Population finie de taille N,
- $\mathbf{y}_U$ le vecteur des valeurs prises par la variable d'intérêt y,
- $\pi_i = P(i \in S)$ les probabilités d'inclusion de premier ordre pour chaque individu i de la population,
- r_i l'indicateur de réponse tel que $r_i = 1$ si l'individu i a répondu à la variable y, et $r_i = 0$ sinon.

Hypothèses :

- Chaque individu répond indépendamment des autres.
- Les données sont **Missing At Random (MAR)** :
  la non réponse peut être liée aux variables auxiliaires, mais conditionnellement à ces variables auxiliaires elle n'est pas liée aux variables d'intérêt.

Cadre théorique et notation

- **Paramètre d'intérêt** : **Total en population finie** $t_y = \sum_{i \in U} y_i$.
- **L'estimateur par expansion** :

$$\widehat{t}_{y,PSA} = \sum_{i \in S_r} \frac{1}{\pi_i \widehat{p}_i} y_i, \tag{1}$$

où $\widehat{p}_i$ est un estimateur de p_i.

- **L'estimateur de Hajek** est un estimateur alternatif de t_y :

$$\widehat{t}_{y,HAJ} = \frac{N}{\widehat{N}} \sum_{i \in S_r} \frac{1}{\pi_i \widehat{p}_i} y_i, \tag{2}$$

où $\widehat{N} = \sum_{i \in S_r} \frac{1}{\pi_i \widehat{p}_i}$ est un estimateur de N, basé sur les individus répondants.

Modélisation de la non réponse

Dans notre étude nous avons couvert un large éventail de méthodes **paramétriques ou non paramétriques, simples ou agrégées**, parmi lesquelles :

- Régression logistique
- Analyse discriminante non paramétrique
- Classification and Regression Tree (CART)
- Conditional Inference Trees (Ctree) pour cibles simples et multiples
- **Iterated Multivariate decision trees**
- Bagging et forêts aléatoires
- Gradient Boosting et Stochastic Gradient Boosting
- The Support Vector Machine (SVM)

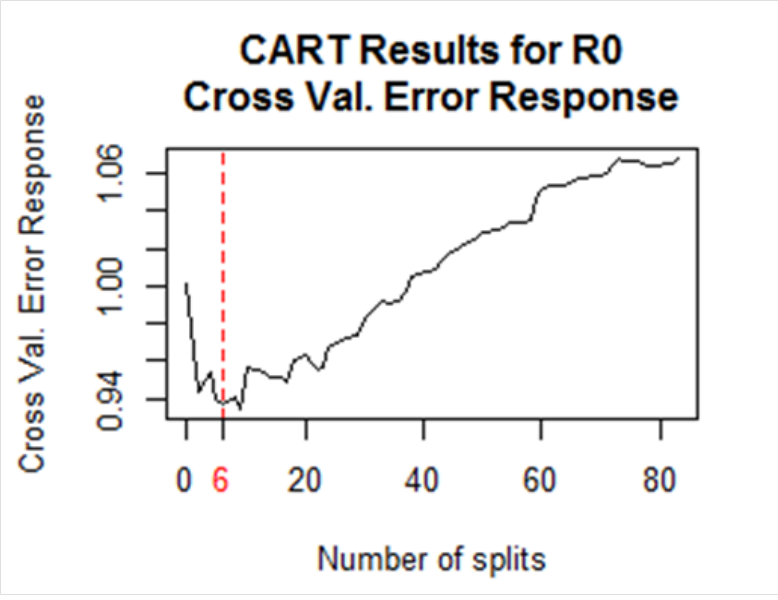
Modification des probabilités "brutes" estimées

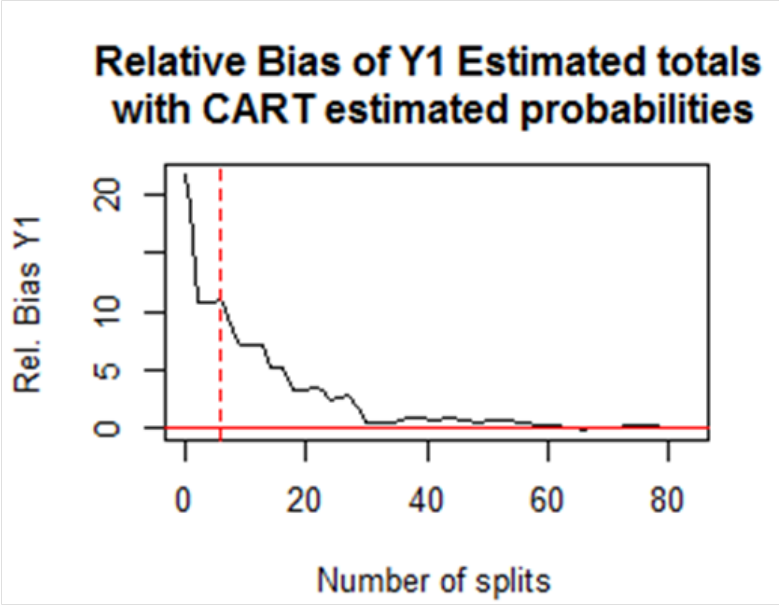
Les méthodes précédentes fournissent des probabilités de réponse estimées "brutes".

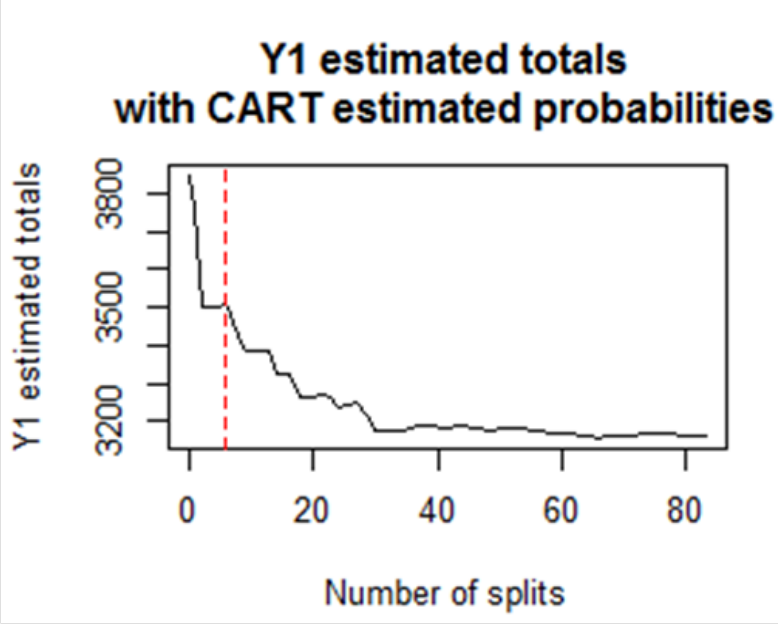
Problèmes si on utilise directement l'inverse de ces probabilités estimées "brutes" dans l'ajustement des poids des répondants :

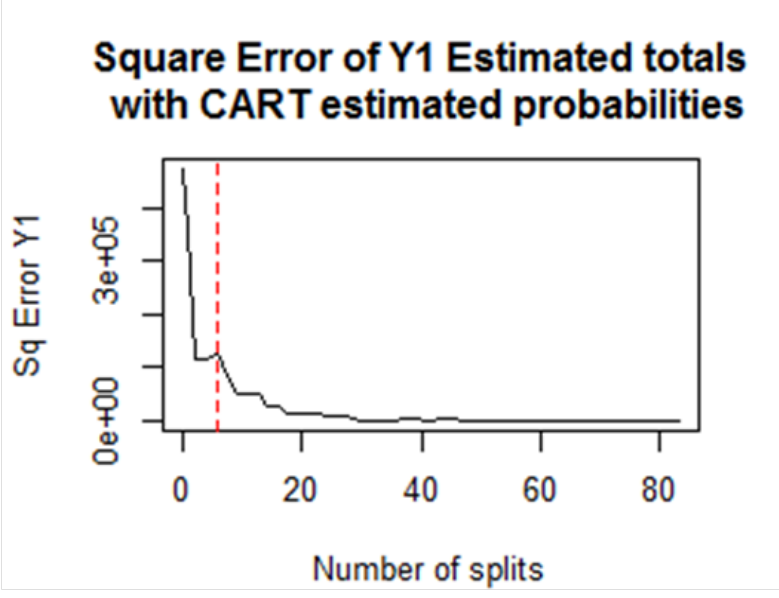
- **Biais** si le modèle de prédiction de l'indicateur de réponse est mal spécifié
  $\hookrightarrow$ **Groupes Homogènes de réponse (GRH)**,
- **Augmentation de la variance** si des probabilités estimées "brutes" sont proches de zéro :
  $\hookrightarrow$ **Tronquer** par valeur inférieure les probabilités estimées.

Exemple de l'utilité des GRH avec CART









Etude par simulations

De façon à se concentrer sur l'erreur de non réponse, nous nous plaçons dans le cadre d'un recensement de la population avec $n = N = 1500$.

Nous mettons en oeuvre $K = 1000$ itérations du processus suivant :

1. une population finie de taille $N = 1500$ est générée par dix modèles correspondant à 10 variables d'intérêt, y_j, $j = 1, ..., 10$ et 5 variables d'intérêt x_1-x_5,
2. à partir de la population générée, nous générons de la non réponse totale à partir de 7 mécanismes de non réponse différents.

Nous utilisons une troncature pour les $\hat{p}_i$ avec une limite inférieure de 0.02 pour toutes les méthodes - avec ou sans Groupes Homogènes de Réponse.

Etude par simulations

- Estimer les probabilités de réponse est typiquement un problème de **classification supervisée**, dans laquelle la variable à expliquer est la variable qualitative binaire indicatrice de réponse r.
- A noter : on ne se focalise pas sur l'optimisation de la performance de prédiction ($\hat{r}_i = r_i$?) mais sur l'**estimation de la probabilité a posteriori de répondre** $\hat{P}(r = 1|X = x)$
- Comparer un grand nombre de méthodes de **machine learning** pour estimer les probabilités de réponse.
- Pour chaque méthode, mesurer la performance de l'estimateur par expansion (1) et de l'estimateur de Hajek (2) en termes de **biais relatif et efficacité relative**.

Etude par simulations

- **Biais relatif de Monte Carlo** d'un estimateur $\hat{t}_y$ du paramètre de population finie t_y :

$$RB_{MC}(\hat{t}_y) = \frac{100}{K}\sum_{k=1}^{K}\frac{(\hat{t}_{y(k)} - t_y)}{t_y}, \quad (3)$$

où $\hat{t}_{y(k)}$ est l'estimateur calculé sur la k-ième itération.

- **Efficacité relative de Monte Carlo** de $\hat{t}_y$, avec $\hat{t}_{y\,GRH+reglog}$ comme référence :

$$RE_{MC}(\hat{t}_y) = \frac{MSE_{MC}(\hat{t}_y)}{MSE_{MC}(\hat{t}_{y\,GRH+reglog})},$$

avec

$$MSE_{MC}(\hat{t}_y) = \frac{1}{K}\sum_{k=1}^{K}(\hat{t}_{y(k)} - t_y)^2 \quad (4)$$

Etude par simulations

Utiliser $RB_{MC}(\hat{t}_{y_{(m)}})$ et $RE_{MC}(\hat{t}_{y_{(m)}})$ comme mesures de performance conduit à **42000 indicateurs de performance** car :

- 7 mécanismes de réponse $R_0, ...R_6$,
- 10 variables d'intérêt $y_1, ...y_{10}$,
- 30 méthodes (versions avec ou sans GRH de 15 méthodes de machine learning),
- 2 types d'estimateurs $\hat{t}_{y_{Exp}}$ et $\hat{t}_{y_{Haj}}$.

De façon à obtenir **un classement global** de ces 30 méthodes pour $\hat{t}_{y_{Exp}}$ et $\hat{t}_{y_{Haj}}$, nous construisons deux indicateurs globaux :

- un pour résumer les tableaux de RB_{MC},
- un pour résumer les tableaux de RE_{MC}

pour chacune des 30 méthodes de machine learning.

Etude par simulations

Pour chaque estimateur ($\hat{t}_{y_{Exp}}$ ou $\hat{t}_{y_{Haj}}$) et chaque méthode de machine learning, nous avons :

- deux tableaux RB_{MC} et RE_{MC}
- contenant chacun **70 indicateurs** (10 lignes pour les 10 variables d'intérêt et 7 colonnes pour les 7 mécanismes de réponse).

Chaque tableau T peut être résumé par un nombre unique, sa **norme de Frobenius** :

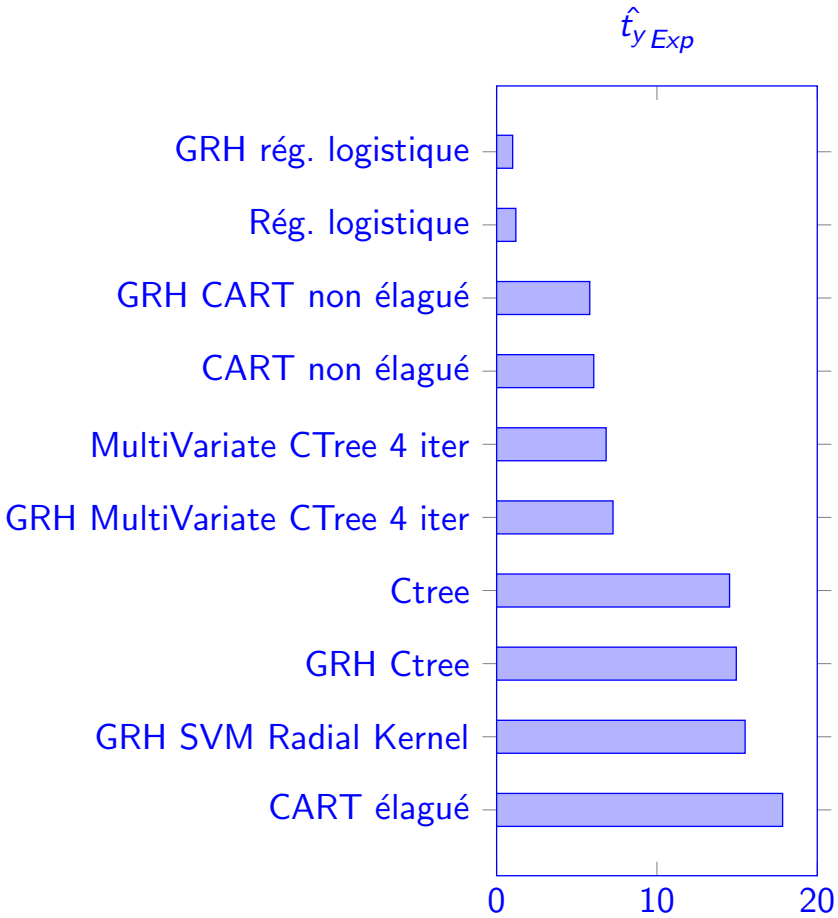
$$\|T\|_F = \sqrt{trace(T^* T)}$$

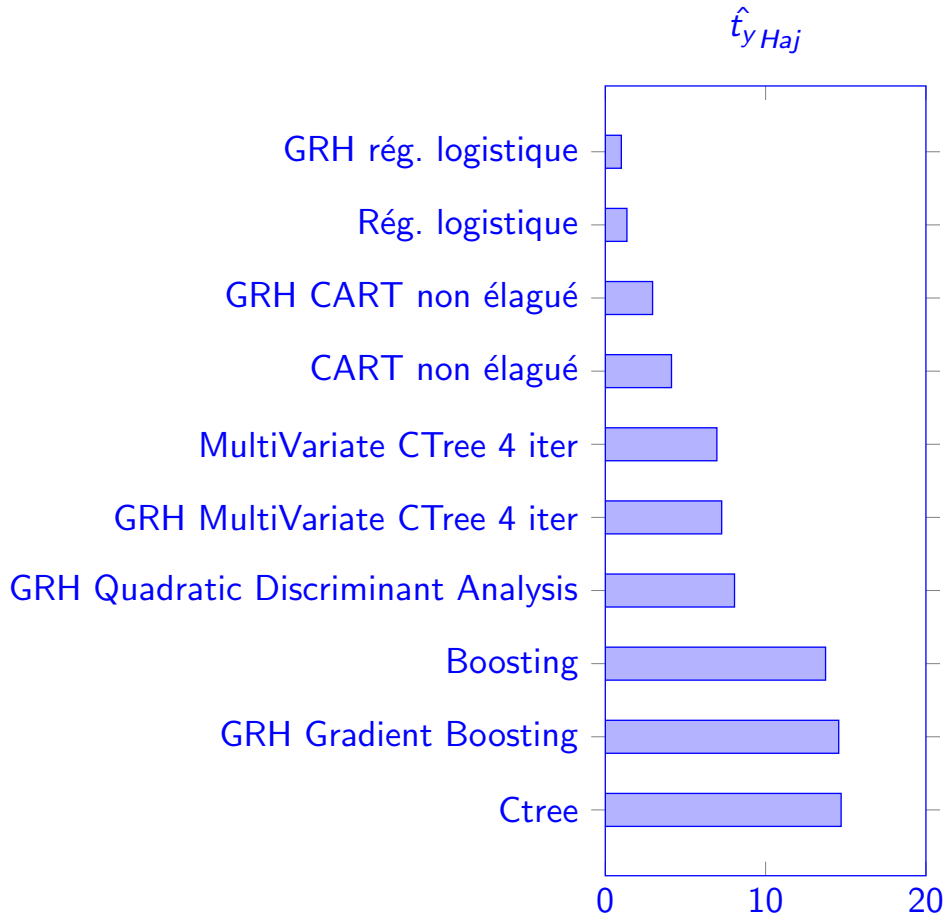
où T^* est la matrice adjointe de T.

- Identifier les méthodes les plus performantes : celles dont $\|RB_{MC}\|_F$ et $\|RE_{MC}\|_F$ sont les plus faibles.
- Etudier ensuite plus en détail ces méthodes.

Résultats pour l'efficacité relative

Norme de Frobenius pour les tableaux d'efficacité relative





Discussion

Recherches futures :

- Etude approfondie de notre **version itérée de l'arbre à cibles multiples Ctree** dont les performances sont assez bonnes. Cette méthode pourrait s'avérer utile dans le cadre de l'imputation.
- Evaluer la performance de différentes méthodes de machines learning lorsqu'il y a des **données manquantes parmi les régresseurs utilisés pour prédire** r_i.
- Etudier l'agrégation de modèles avec le **stacking** (Wolpert 1992, Breiman 1996, Nocairi et al. 2016).
- Evaluer les méthodes d'apprentissage lorsqu'elles sont appliquées après des **plans de sondages complexes**.

Merci pour votre attention

AGRO CAMPUS OUEST

ENSAI École nationale de la statistique et de l'analyse de l'information

IRMAR

Université de Montréal