Treatment of unit nonresponse in surveys through machine learning methods : an empirical comparison

Khaled Larbi - David Haziza - Mehdi Dadgoug

30 mars 2022

Ensae - University of Ottawa - Université de Bourgogne Franche-Comté

- Soit une population $\mathcal{U}$ finie de taille N telle que $\mathcal{U} = \{1, \ldots, N\}$.

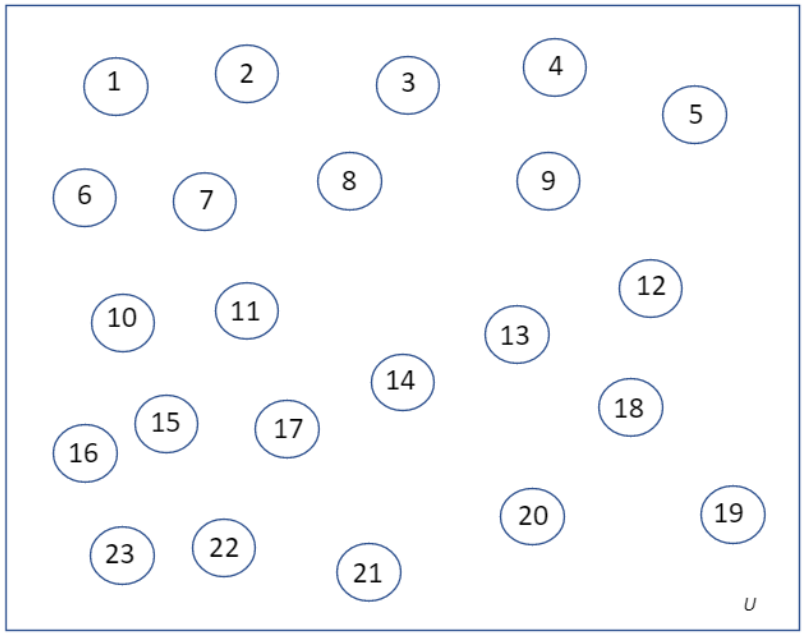

Figure 1: $\mathcal{U} = \{1, \ldots, 23\}$

- Soit une population $\mathcal{U}$ finie de taille N telle que $\mathcal{U} = \{1, \ldots, N\}$.
- Pour chaque individu de la population, :
  - y, la variable d'intérêt ($\rightarrow$ inconnue).

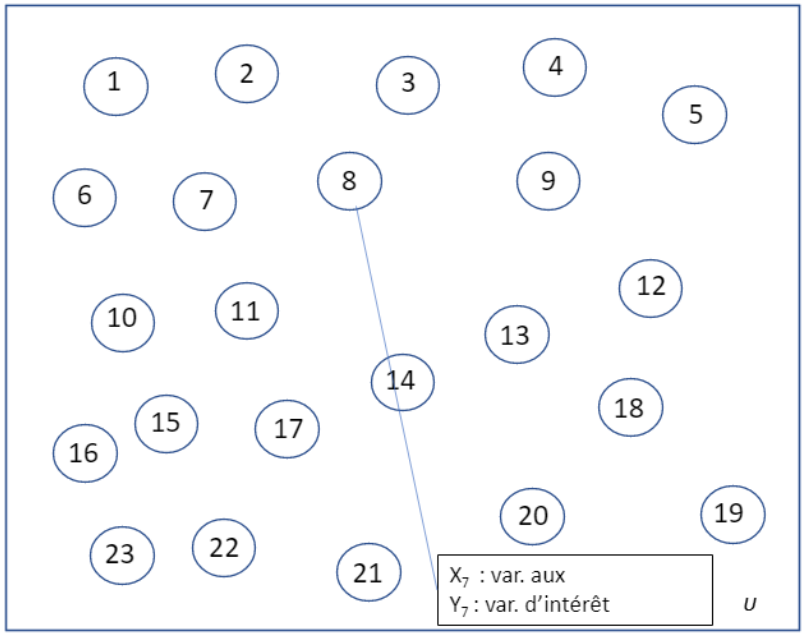

Figure 2: $\mathcal{U} = \{1, \ldots, 23\}$

- Soit une population $\mathcal{U}$ finie de taille N telle que $\mathcal{U} = \{1, \ldots, N\}$.
- Pour chaque individu de la population :
  - y, la variable d'intérêt ($\rightarrow$ inconnue).
- But : Estimer le total $t_y := \sum_{k \in \mathcal{U}} y_k$.

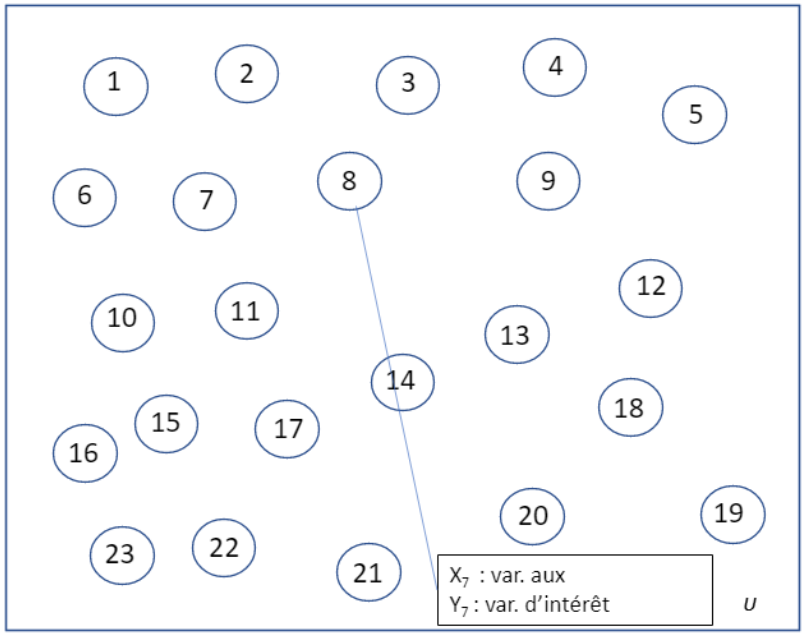

Figure 3: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

- Soit une population $\mathcal{U}$ finie de taille N telle que $\mathcal{U} = \{1, \ldots, N\}$.
- Pour chaque individu de la population :
  - y, la variable d'intérêt ($\rightarrow$ inconnue).
- But : Estimer le total $t_y := \sum_{k \in \mathcal{U}} y_k$.
- $\mathcal{S}$ tiré via P.

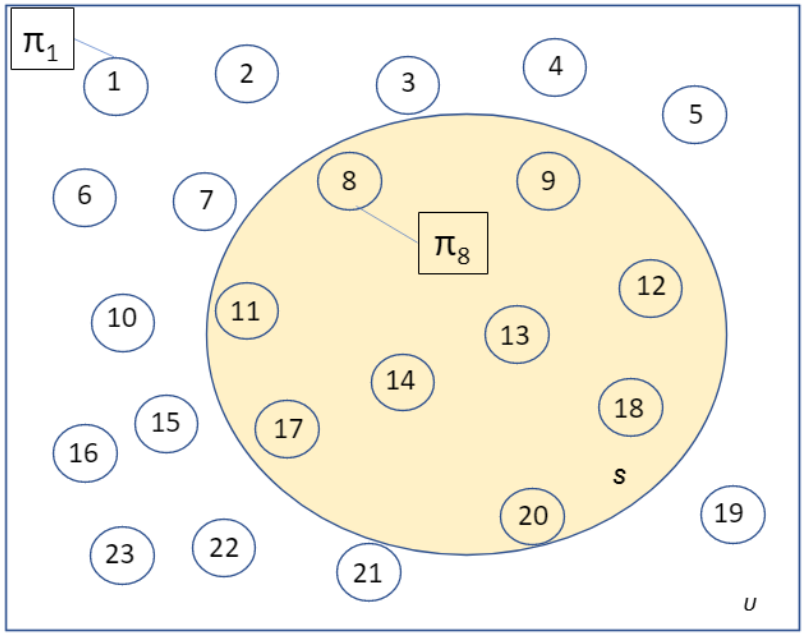

Figure 4: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

- Soit une population $\mathcal{U}$ finie de taille N telle que $\mathcal{U} = \{1, \ldots, N\}$.
- Pour chaque individu de la population :
  - y, la variable d'intérêt ($\rightarrow$ inconnue).
- But : Estimer le total $t_y := \sum_{k \in \mathcal{U}} y_k$.
- $\mathcal{S}$ tiré via P.
- Probabilité d'inclusion d'ordre 1 $\pi_k := \mathbb{E}_{S \sim P}(I_k(S))$

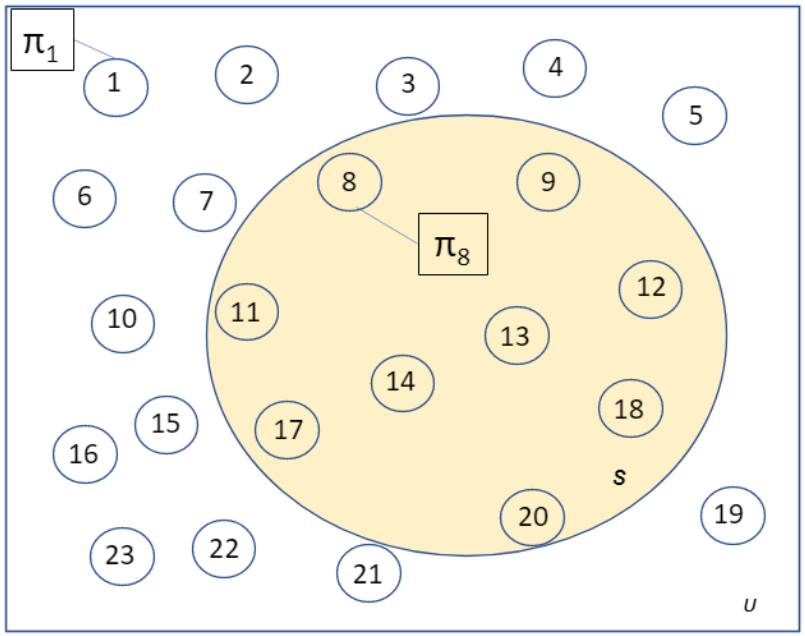

Figure 5: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

- Probabilité d'inclusion d'ordre 1 :

$$\pi_k := \mathbb{E}_{S\sim P}(I_k(S))$$

- Un estimateur possible : l'estimateur de Narain-Horvitz-Thompson :

$$\hat{t}_{y\pi} := \sum_{k\in\mathcal{S}} \frac{y_k}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k} I_k$$

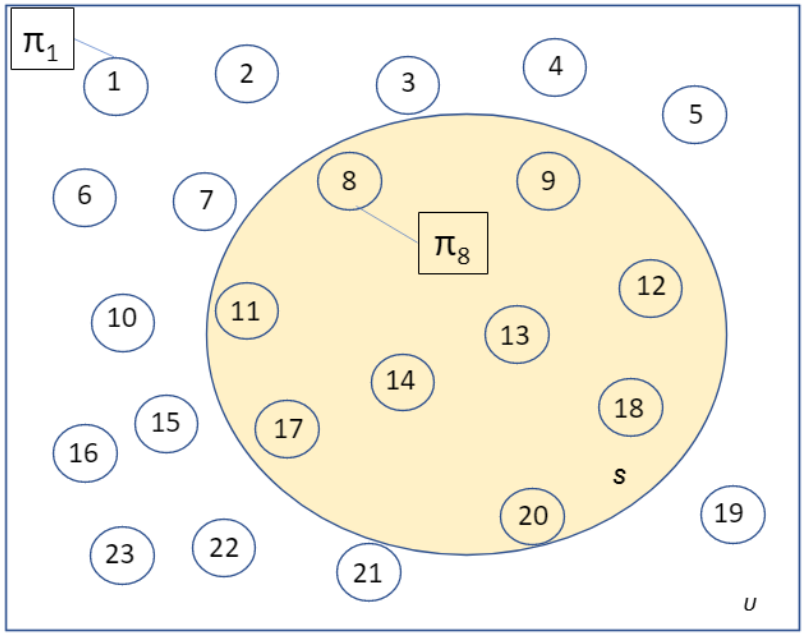

Figure 6: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

- Probabilité d'inclusion d'ordre 1 :

$$\pi_k := \mathbb{E}_{S\sim P}(I_k(S))$$

- Un estimateur possible : l'estimateur de Narain-Horvitz-Thompson :

$$\hat{t}_{y\pi} := \sum_{k\in\mathcal{S}} \frac{y_k}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k} I_k$$

- Quid si non réponse ?

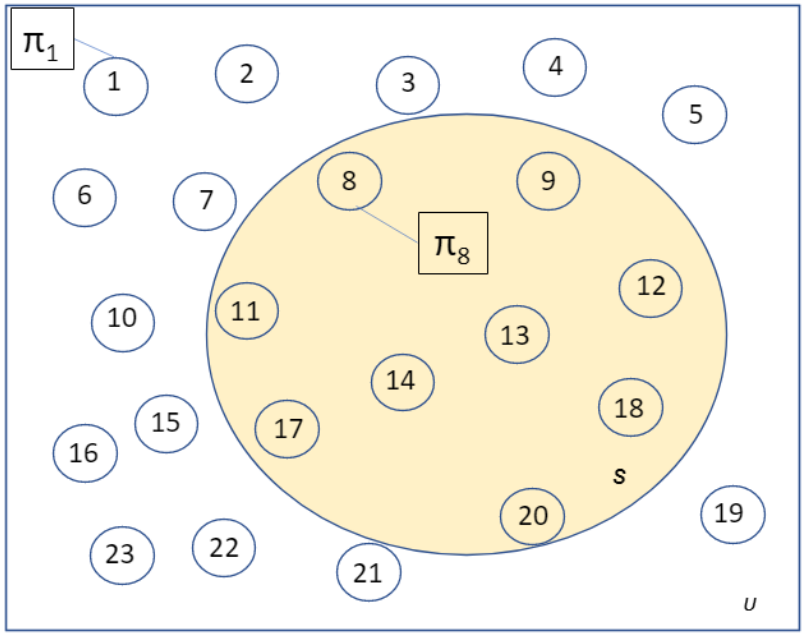

Figure 7: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

- Probabilité d'inclusion d'ordre 1 :

$$\pi_k := \mathbb{E}_{S\sim P}(I_k(S))$$

- Un estimateur possible : l'estimateur de Narain-Horvitz-Thompson :

$$\hat{t}_{y\pi} := \sum_{k\in\mathcal{S}} \frac{y_k}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k} I_k$$

- Quid si non réponse ?

$$\hat{t}_{y\pi,\mathcal{S}_r} := \sum_{k\in\mathcal{S}_r} \frac{y_k}{\pi_k}$$

n'est plus sans biais.

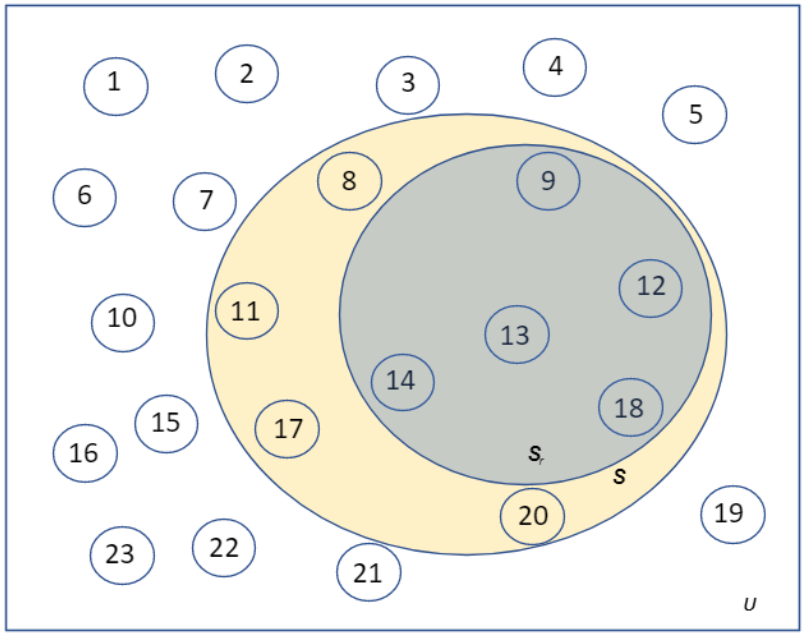

Figure 8: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

- Quid si non réponse ?
- Seuls les individus $\mathcal{S}_r \subset \mathcal{S}$ répondent.
- Pour tout $k \in \mathcal{U}$,

$$\begin{aligned} p_k :=& \mathbb{P}(R_k = 1 | \mathbf{x}_k, y_k) \\ =& \mathbb{P}(R_k = 1 | \mathbf{x}_k) = p_k(\mathbf{x}_k) \end{aligned}$$

$\rightarrow$ MAR.

- Hypothèses :
  - Chaque individu k a une probabilité $p_k > 0$ de répondre.
  - Les individus répondent indépendamment les uns des autres.

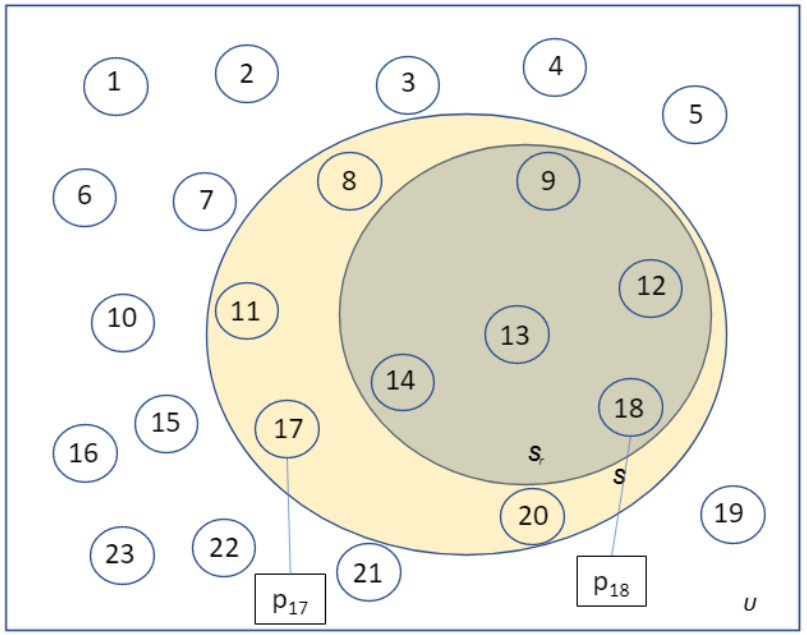

Figure 9: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

- Seuls les individus $\mathcal{S}_r \subset \mathcal{S}$ répondent $\rightarrow R_k = 1_{\{k \in \mathcal{S}_r\}}$.
- Pour tout $k \in \mathcal{U}$,

$$\begin{aligned} p_k :=& \mathbb{P}(R_k = 1 | \mathbf{x}_k, y_k) \\ =& \mathbb{P}(R_k = 1 | \mathbf{x}_k) = p_k(\mathbf{x}_k) \end{aligned}$$

$\rightarrow$ MAR.

- Hypothèses :
  - Chaque individu k a une probabilité $p_k > 0$ de répondre.
  - Les individus répondent indépendamment les uns des autres.
  - Estimateur par double expansion :

$$\hat{t}_{y,DE} = \sum_{k \in \mathcal{S}_r} \frac{y_k}{\pi_k p_k}$$

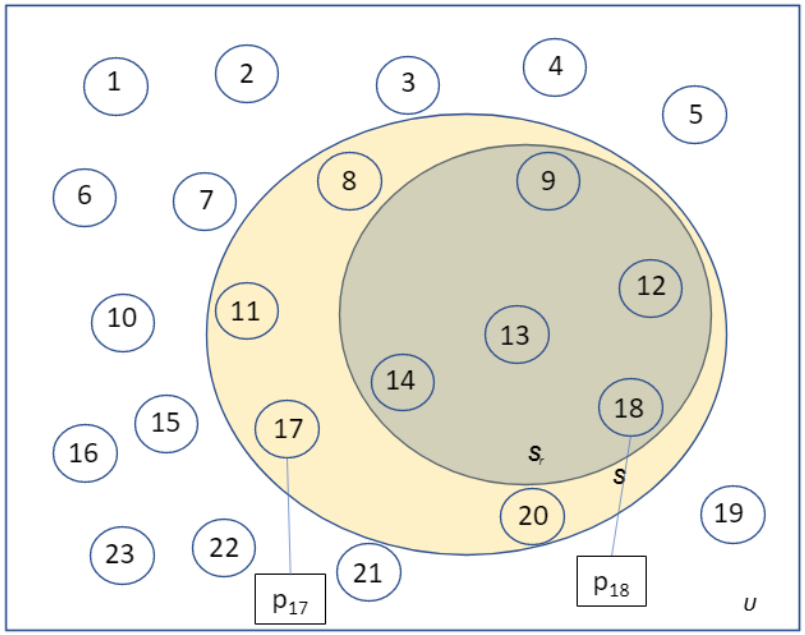

Figure 10: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

$$\hat{t}_{y,DE} = \frac{y_9}{\pi_9 p_9} + \frac{y_{12}}{\pi_{12} p_{12}} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

- Seuls les individus $\mathcal{S}_r \subset \mathcal{S}$ répondent.
- Pour tout $k \in \mathcal{U}$,

$$\begin{aligned} p_k :=& \mathbb{P}(R_k = 1 | \mathbf{x}_k, y_k) \\ =& \mathbb{P}(R_k = 1 | \mathbf{x}_k) = p_k(\mathbf{x}_k) \end{aligned}$$

- Hypothèses :
  - Chaque individu k a une probabilité $p_k > 0$ de répondre.
  - Les individus répondent indépendamment les uns des autres.
  - Estimateur par double expansion :

$$\hat{t}_{y,DE} = \sum_{k \in \mathcal{S}_r} \frac{y_k}{\pi_k p_k}$$

- Problème : $\{p_k\}_{k \in \mathcal{U}}$ inconnu.

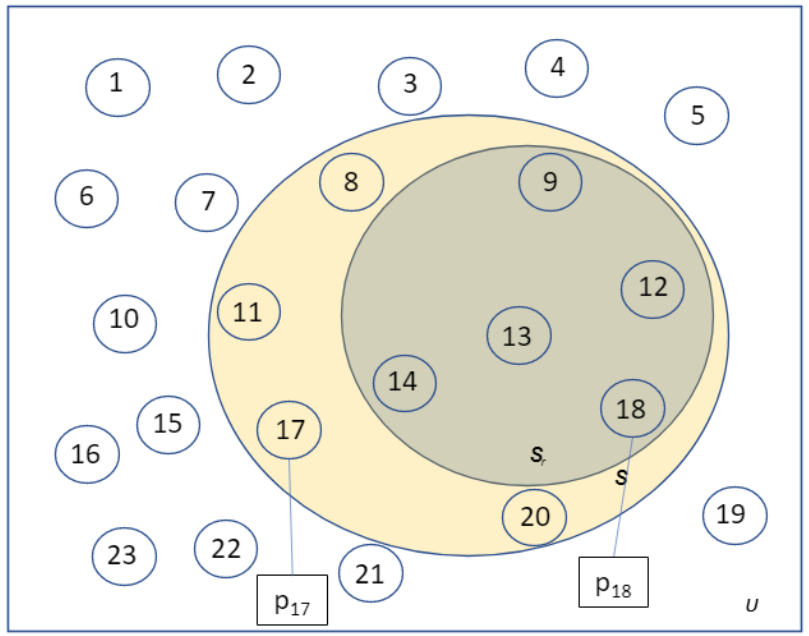

Figure 11: $\mathcal{U} = \{1, \ldots, 23\}$

$$t_y = \sum_{k=1}^{23} y_k$$

$$\mathcal{S} = \{8, 9, 11, 12, 13, 14, 17, 18, 29\}$$

$$\hat{t}_{y\pi} = \frac{y_8}{\pi_8} + \frac{y_9}{\pi_9} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

$$\hat{t}_{y,DE} = \frac{y_9}{\pi_9 p_9} + \frac{y_{12}}{\pi_{12} p_{12}} + \ldots + \frac{y_{18}}{\pi_{18}} + \frac{y_{20}}{\pi_{20}}$$

But

- Autres estimateurs inspirés :
  - ajusté par les scores de propension (ou PSA) :

$$\hat{t}_{y,PSA} := \sum_{k \in \mathcal{S}_r} \frac{y_k}{\pi_k \hat{p}_k} \tag{1}$$

    où $\{\hat{p}_k\}_{k \in \mathcal{S}_r}$ sont des estimations des probabilités de réponse.
  - estimateur d'Hajek :

$$\hat{t}_{y,H} := \frac{N}{\hat{N}_{PSA}} \sum_{k \in \mathcal{S}_r} \frac{y_k}{\pi_k \hat{p}_k} \tag{2}$$

    où $\hat{N}_{PSA} := \sum_{k \in \mathcal{S}_r} \frac{1}{\pi_k \hat{p}_k}$. Il s'agit d'un estimateur calé sur la taille de la population.

But

Pour les estimateurs PSA et d'Hajek, il faut donc estimer les probabilités de réponse $\{\hat{p}_k\}$ à partir d'un variable indicatrice de réponse R et de variables auxilaires x.

$\rightarrow$ Utilisation de méthodes d'apprentissage.

But de ces travaux :

- comparer empiriquement différentes méthodes d'apprentissage appliquées à l'estimation des probabilités de réponse afin d'obtenir les estimateurs PSA et d'Hajek du total.
- proposer des méthodes permettant de combiner les estimations de plusieurs méthodes d'apprentissage.

Méthodes d'apprentissage utilisées

- Méthodes basées sur la régression logistique :
  - Régression logistique : [Hastie et al., 2001]
  - Régression logistique avec pénalisation lasso : [Hastie et al., 2001]
- Méthodes basées sur des arbres de décision :
  - CART : [Breiman et al., 1983]
  - Forêts aléatoires (bagging) : [Breiman, 2004]
  - XGBoost (boosting) : [Chen and Guestrin, 2016]
  - BART (approche bayésienne) : [Chipman et al., 2010]
  - Cubist : [Quinlan, 1992] [Quinlan, 1993]
- Autres méthodes :
  - SVM : [Cortes and Vapnik, 1995]
  - K-plus proche voisin : [Hastie et al., 2001]
  - MOB : [Zeileis and Hornik, 2007] [Zeileis et al., 2008]

Ces méthodes utilisent des hyperparamètres $\rightarrow$ plusieurs jeux d'hyperparamètres testés.

Exemple : `rf1` pour forêts aléatoires avec 200 arbres et 30 observations par feuille min.

Méthode des scores (ou groupe homogène de réponse)

Supposons que nous disposons pour chaque individu $i \in \mathcal{S}$ d'une estimation $\hat{p}_i$ des probabilités de réponse.

Il est possible d'obtenir une autre estimation en utilisant la méthode des scores ([Little, 1986] [Haziza and Beaumont, 2007] [Gelein et al., 2018]) :

1. Découper les individus de l'échantillon en K classes en se basant sur $\{\hat{p}_i\}$ (en utilisant les quantiles empiriques par exemple).
2. La probabilité de réponse d'un individu i sera la proportion de répondants dans sa classe.

Cette méthode permet d'être plus robuste à des problèmes de mauvaise spécification.

Méthodes d'agrégation : COMPRESS

Supposons qu'on dispose pour chaque individu $i \in \mathcal{S}$ d'un vecteur $\hat{\mathbf{p}}_i := (\hat{p}_i^{(1)}, \hat{p}_i^{(2)}, \ldots, \hat{p}_i^{(d)}) \rightarrow$ estimation des probabilités de réponse de l'individu i en utilisant d méthodes différentes.

Il est possible d'obtenir des nouvelles estimations à partir de celle ci :

- En faisant la régression linéaire multiple de R_k sur $\hat{\mathbf{p}}_k$ sans constante $\rightarrow$ obtention d'un vecteur $\hat{\beta}$ puis normalisation $\tilde{\beta} := \frac{1}{||\beta||_2^2}(\beta^{(1)^2}, \ldots, \beta^{(d)^2})$
- Les nouvelles estimations des probabilités sont données par

$$\hat{p}_i^{\text{com}} := < \tilde{\beta}, \hat{\mathbf{p}}_i > \tag{3}$$

$\rightarrow$ Dans notre cas, $d = 4$ avec logistique score, forêts aléatoires, BART et Cubist.

Estimateur obtenu : $\hat{t}_{y,com} = \sum_{k \in \mathcal{S}_r} \frac{y_k}{\pi_k \hat{p}_k^{\text{com}}}$.

Méthodes d'agrégation : calage

Supposons qu'on dispose pour chaque individu $i \in \mathcal{S}$ d'un vecteur $\hat{\mathbf{p}}_i := (\hat{p}_i^{(1)}, \hat{p}_i^{(2)}, \ldots, \hat{p}_i^{(d)}) \rightarrow$ estimation des probabilités de réponse de l'individu i en utilisant d méthodes différentes.

Il est possible d'obtenir des nouvelles estimations à partir de celle ci :

- En calant les probabilités de réponse avec les contraintes suivantes :

$$\sum_{k \in S_r} w_k \hat{\mathbf{p}}_k = \sum_{k \in S} d_k \hat{\mathbf{p}}_k \text{ et } \sum_{k \in S_r} w_k = \sum_{k \in S} d_k$$

  et en utilisant une fonction de calage exponentielle.

$\rightarrow$ Dans notre cas, $d = 4$ avec logistique score, forêts aléatoires, BART et Cubist.

Estimateur obtenu : $\hat{t}_{y,cal} = \sum_{k \in \mathcal{S}_r} w_k y_k$.

Méthodes d'agrégation : COMPRESS + calage

Il est possible de combiner les deux approches précédentes

- Calcul pour tout $k \in \mathcal{S}$ des estimations des probabilités de réponse $\hat{p}_k^{\text{com}}$.
- Calage de ces estimations : $\sum_{k \in \mathcal{S}_r} w_k \log(p_k^{\text{com}}) = \sum_{k \in \mathcal{S}} d_k \log(p_k^{\text{com}})$ et $\sum_{k \in \mathcal{S}_r} w_k = \sum_{k \in \mathcal{S}} d_k$ avec une fonction de calage exponentielle.

Scénario

- $N = 50000$ et $n = 1000$
- Génération des variables auxiliaires :
  - Stratification $X^{(s)} \sim \Gamma(3, 2)$.
  - Autres :

$$(\mathbf{X^{(c_1)}}, X^{(c_2)}, X^{(c_3)}, X^{(d_1)}, X^{(d_2)}, X^{(d_3)}) \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \otimes \Gamma(\mathbf{3}, \mathbf{2}) \otimes \Gamma(\mathbf{3}, \mathbf{2}) \otimes \mathbb{P}^{(d_1)} \otimes \mathcal{B}(\mathbf{0.5}) \otimes \mathcal{U}_{\{\mathbf{1};\ldots;\mathbf{5}\}}$$

    - Cas indépendant + cas avec lien entre variables mais même lois marginales.
- Génération de la variable d'intérêt y.
  - y linéaire et autre cas où y non linéaire.
- Génération des probabilités de réponse
  - Utilisation de six mécanismes.
- Tirage de l'échantillon (tirage stratifié selon SASSR avec allocation de Neyman selon $X^{(c2)}$) puis plan poissonien avec les probabilités de réponses générées.
- Estimation des probabilités de réponse.
  - À partir de la variable indicatrice de réponse et des variables auxiliaires.
- Calcul des estimations PSA et Hajek.

Probabilités de réponse

Utilisation de six mécanismes de génération des probabilités de réponse :

1. $p_k^{(1)} = \text{logit}^{-1}(-0.8 - 0.05X_{1k}^{(s)} + 0.2X_{1k}^{(c)} + 0.5X_{2k}^{(c)} - 0.05X_{3k}^{(c)} + \sum_{k=2}^{5} 0.2(1_{\{X_{\mathbf{1}k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} + \sum_{k=2}^{5} 0.3(1_{\{X_{\mathbf{3}k}^{(d)}=k\}}))$.
2. $p_k^{(2)} = 0.1 + 0.9\,\text{logit}^{-1}(0.5 + 0.3X_{1k}^{(s)} - 1.1X_{1k}^{(c)} - 1.1X_{2k}^{(c)} - 1.1X_{3k}^{(c)} + \sum_{k=2}^{5} 0.8(1_{\{X_{\mathbf{1}k}^{(c)}=k\}}) + 0.8X_{2k}^{(d)} + \sum_{k=2}^{5} 0.8(1_{\{X_{\mathbf{3}k}^{(d)}=k\}}))$.
3. $p_k^{(3)} = 0.1 + 0.9\,\text{logit}^{-1}\left(-1 + \text{sgn}\,(X_{1k}^{c})\,(X_{1k}^{c})^2 + 3 \times 1_{\left\{X_{\mathbf{1}k}^{(d)}<4\right\}\cap\left\{X_{\mathbf{2}k}^{(d)}=1\right\}}\right)$.
4. $p_k^{(4)} = 0.55 + 0.45\tanh(0.05y_k - 0.5)$.
5. $p_k^{(5)} = 0.1 + 0.9\,\text{logit}^{-1}(0.2y_k - 1.2)$.
6. $p_k^{(6)} = 0.1 + 0.6\,\text{logit}^{-1}(0.85X_{1k}^{(s)} + 0.85X_{2k}^{(c)} - 0.85X_{3k}^{(c)} - \sum_{k=2}^{5} 0.2(1_{\{X_{\mathbf{1}k}^{(c)}=k\}}) + 0.2X_{2k}^{(d)} - \sum_{k=2}^{5} 0.3(1_{\{X_{\mathbf{3}k}^{(d)}=k\}}))$.

Modèles génératifs pour y

$$y_k^{(1)} = \gamma_0 + \gamma_1^{(s)} X_{1k}^{(s)} + \gamma_1^{(c)} X_{1k}^{(c)} + \gamma_2^{(c)} X_{2k}^{(c)} + \gamma_3^{(c)} X_{3k}^{(c)} + \sum_{j=2}^{5} \gamma_{1j}^{(d)} (1_{\{X_{\mathbf{1}k}^{(d)}=j\}})$$

$$+ \gamma_2^{(d)} X_{2k}^{(d)} + \sum_{k=2}^{5} \gamma_{3j}^{(d)} (1_{\{X_{\mathbf{3}k}^{(d)}=j\}}) + \varepsilon_k \tag{4}$$

et

$$y_k^{(2)} = \delta_1 X_{2k}^{(c)} + \delta_2 (X_{2k}^{(c)})^2 (1 - 1_{\{X_{\mathbf{3}k}^{(d)}=2\} \cup \{X_{\mathbf{3}k}^{(d)}=3\}}) + \log(1 + \delta_3 X_{2k}^{(c)}) (1_{\{X_{\mathbf{3}k}^{(d)}=2\} \cup \{X_{\mathbf{3}k}^{(d)}=3\}}) + \varepsilon_k, \tag{5}$$

où $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$

Estimation de la performance

Pour une méthode donnée, la performance de la méthode sera estimée à l'aide du :

- biais relatif Monte-Carlo :

$$\mathbb{B}_{MC}(\hat{t}_y) = \frac{100}{B}\sum_{k=1}^{B}\frac{\hat{t}_{y,k} - t_{y,k}}{t_{y,k}}$$

- efficacité Monte-Carlo :

$$\mathrm{Eff}_{MC}(\hat{t}_y) = 100\frac{\mathrm{MSE}_{MC}(\hat{t}_y)}{\mathrm{MSE}_{MC}(\hat{t}_{y,\pi})}$$

où

$$\mathrm{MSE}_{MC}(\hat{t}_y) = \frac{1}{B}\sum_{k=1}^{B}\left(\hat{t}_{y,k} - t_{y,k}\right)^2$$

$\rightarrow B = 1000$ simulations. Donc pour chaque méthode (et chaque estimateur) : 42 RMSE et biais $\rightarrow$ Comparaison des méthodes en utilisant la médiane sur les 42 RMSE.

Algorithm	Min	Q1	Med	Q3	Max	Mean
xgb1	15	9	1	10	21	19
COMPRESS_CAL	7	4	2	2	2	1
xgb4	13	8	3	12	18	16
xgb3	9	11	4	7	17	11
cart3	19	14	5	20	9	10
cart2	20	13	6	18	7	9
COMPRESS	6	1	7	5	12	6
CART_reg	16	21	8	14	8	4
cart1	18	15	9	17	6	8
xgb2	14	6	10	8	15	12
cart4	12	17	11	15	3	3
bart	3	2	12	3	13	5
knn	17	26	13	6	16	17
logit and score	5	7	14	13	11	7
svm1	4	25	15	1	20	18
knn_reg	11	16	16	9	19	20
rf4	22	10	17	11	5	13
cb4	25	18	18	21	24	21
cb5	26	20	19	26	25	25
calibration	29	27	20	4	1	2
rf2	27	24	21	19	10	14
rf5	28	22	22	28	28	26
rf3	23	19	23	16	4	15
cb1	24	23	24	23	22	22
bart_reg	10	3	25	27	33	33
cb2	21	12	26	29	27	28
logit_lasso	8	28	27	22	23	27
rf_reg	30	29	28	24	26	23
logit	2	5	29	30	33	29
rf1	31	30	30	25	14	24
mob	1	31	31	31	33	33
cb3	33	33	32	33	33	33
svm2	32	32	33	33	33	33

Algorithm	Min	Q1	Med	Q3	Max	Mean
xgb1	21	17	1	21	27	25
COMPRESS	14	2	2	12	15	13
xgb4	20	15	3	20	21	23
bart	16	5	4	10	16	14
xgb3	7	4	5	13	17	18
logit and score	3	12	6	3	14	8
xgb2	8	8	7	15	18	19
COMPRESS_CAL	5	9	8	1	2	1
CART_reg	17	26	9	19	9	9
cb4	19	18	10	7	23	20
cb5	18	19	11	8	25	21
cart4	6	21	12	9	3	3
cb1	26	20	13	6	26	22
cb2	4	10	14	5	20	7
cart1	23	24	15	22	6	15
cart2	25	22	16	23	8	16
cart3	24	23	17	24	11	17
rf4	13	1	18	17	12	12
knn	30	27	19	26	19	24
calibration	31	30	20	2	1	2
rf2	12	6	21	14	7	6
knn_reg	28	25	22	27	22	26
rf3	15	3	23	16	10	11
rf5	11	7	24	4	13	4
svm1	27	28	25	28	24	27
rf_reg	10	11	26	11	5	5
rf1	9	16	27	18	4	10
logit	2	14	28	25	30	28
logit_lasso	29	29	29	29	28	29
bart_reg	22	13	30	31	33	33
svm2	33	31	31	30	29	30
cb3	32	33	32	32	31	31
mob	1	32	33	33	33	33

Table 1: Rang en efficacité (PSA à gauche / Hajek à droite)

Gain à l'utilisation de la méthode des scores

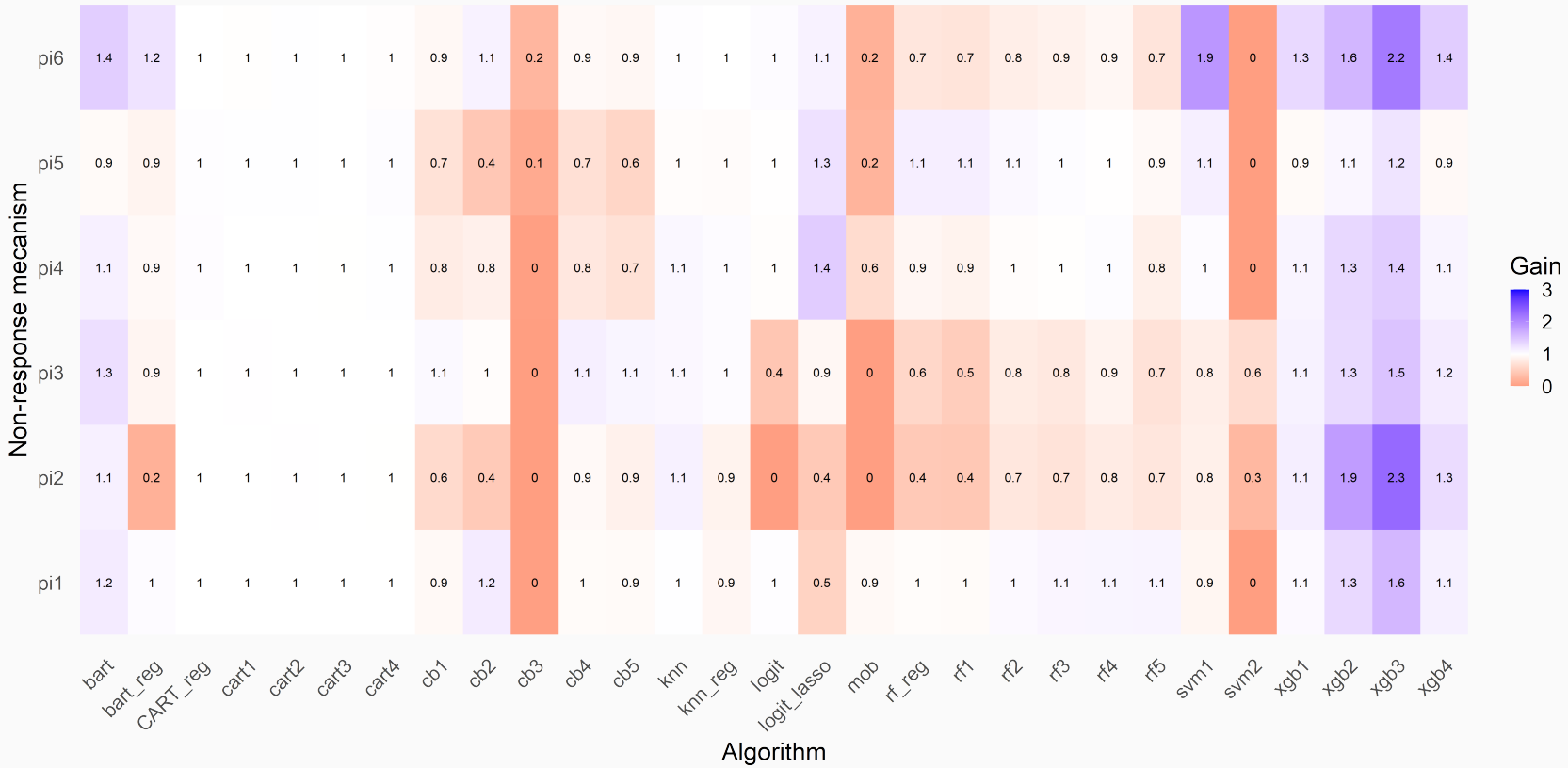

Figure 12: Gain (en efficacité) à l'utilisation de la méthode des scores.

Le gain est le rapport entre le RMSE avec la méthode des scores en 10 classes et le RMSE sans la méthode des scores : si le gain < 1 (rouge), la méthode des scores est plus efficace.

Méthode d'agrégation et nombre de méthodes

Est-ce que le nombre de méthodes utilisées permet d'accroître l'efficacité des estimateurs ?

Sélection de six méthodes : `rf4`, `bart`, `xgb4`, `svm1`, `cb1` et `logit` en 10 classes.

Calcul des performances (RMSE, biais) des estimateurs pour chacun des $\binom{6}{k}$ ensembles de méthodes.

Exemple : (`rf4`,`bart`), (`svm1`,`bart`), (`svm1`,`bart`, `cb1`, `xgb4`)

Il est possible de calculer la médiane des indicateurs de performances lorsqu'on utilise une méthode, puis deux etc.

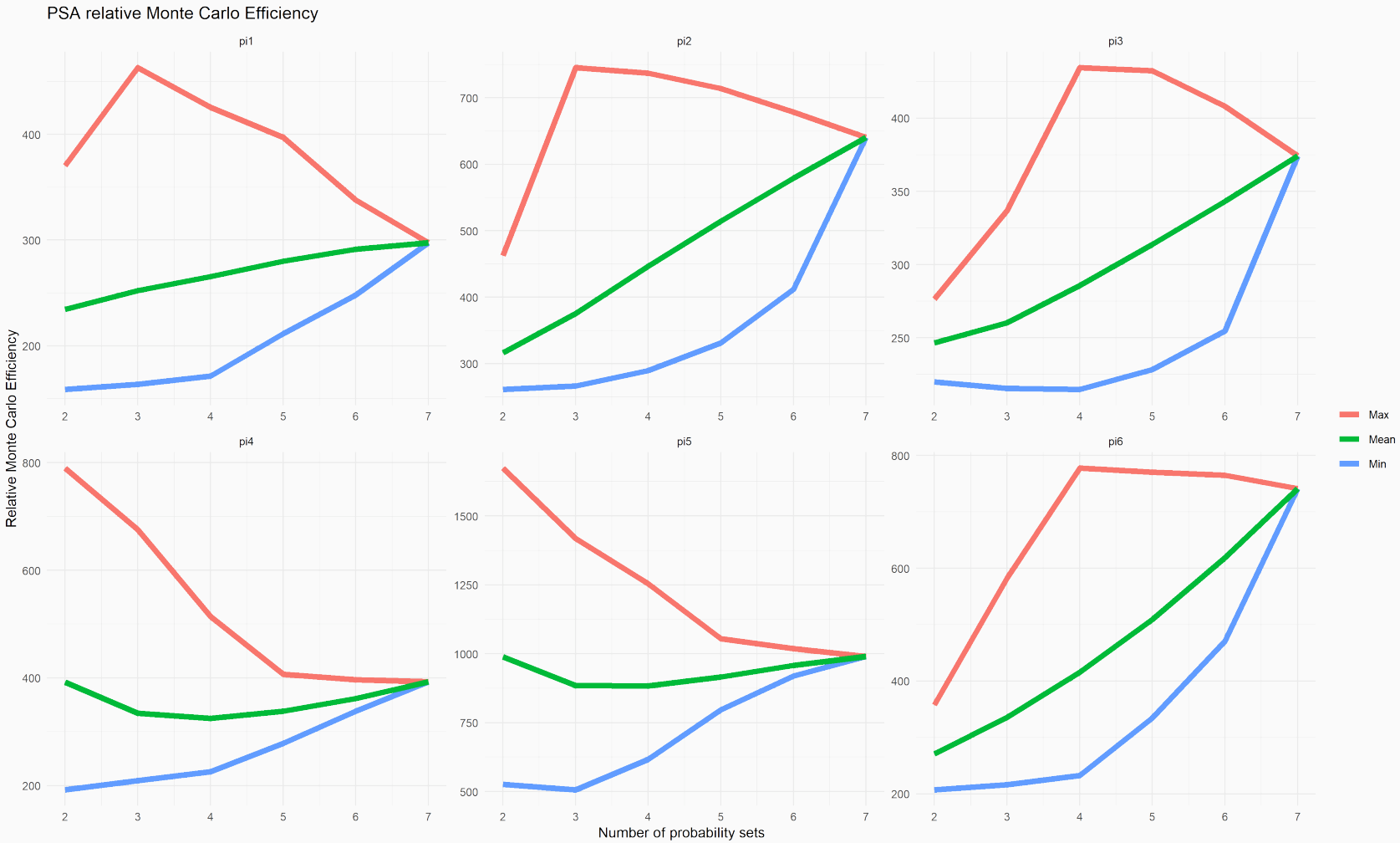

Figure 13: Évolution du RMSE pour l'estimateur PSA basé sur le calage

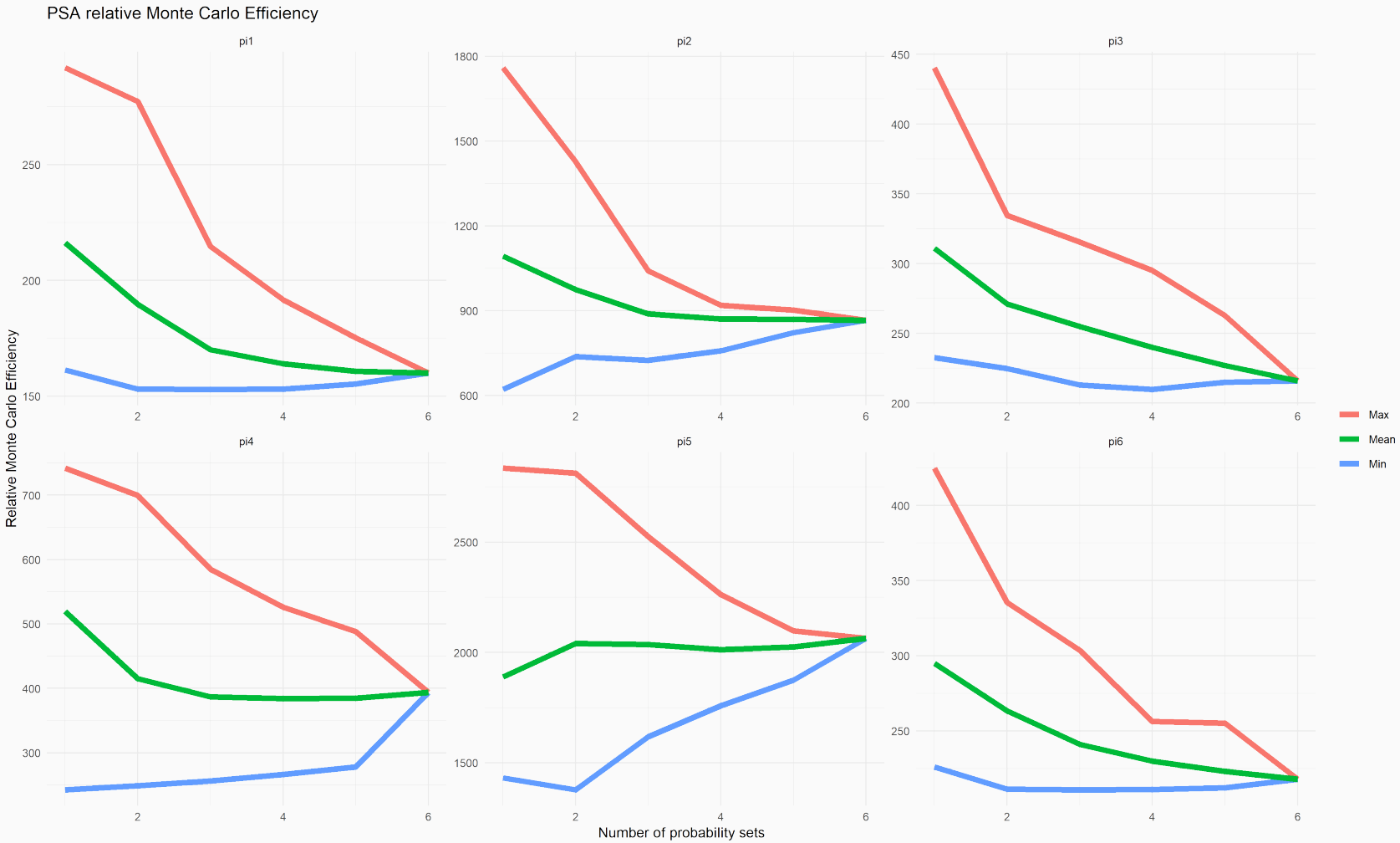

Figure 14: Évolution du RMSE pour l'estimateur PSA basé sur COMPRESS

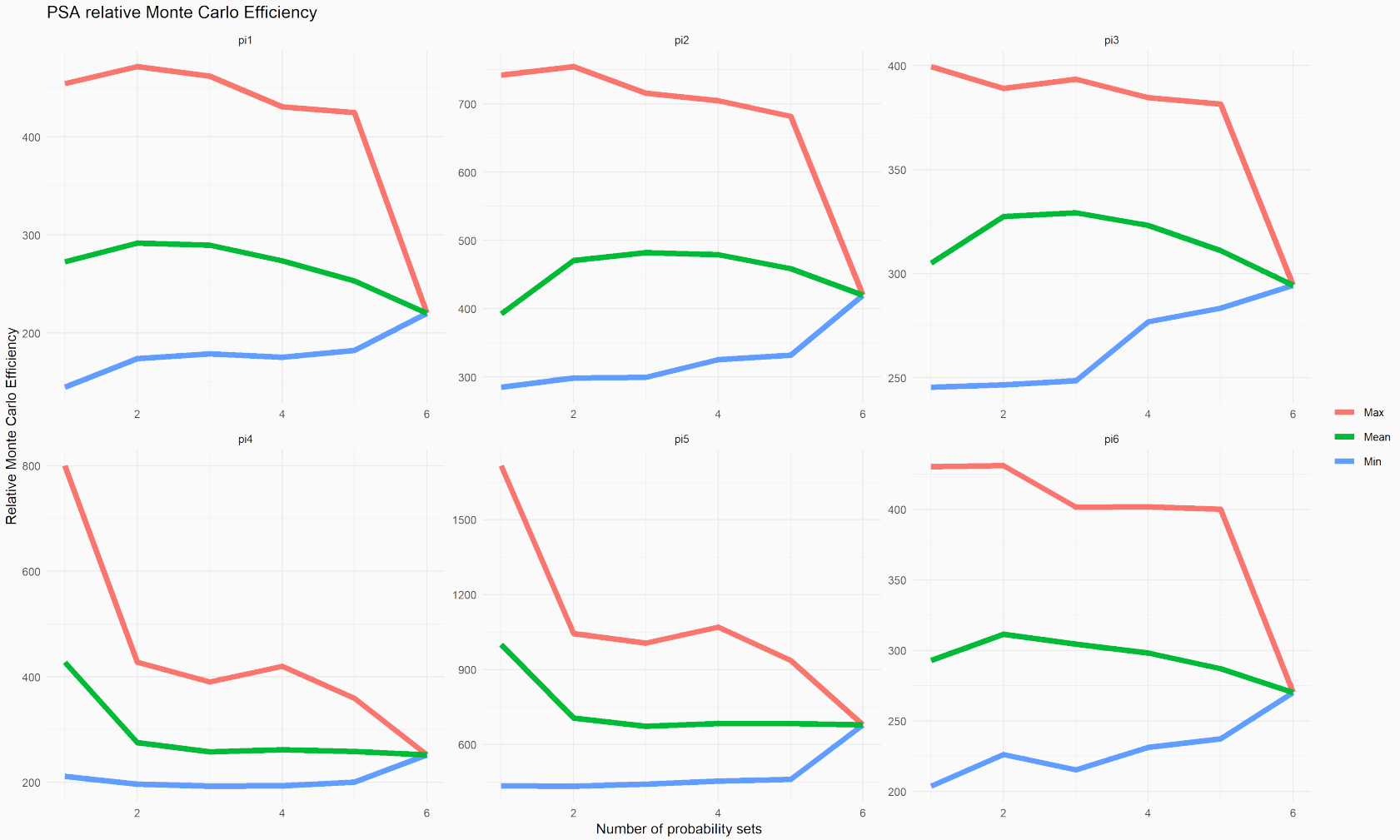

Figure 15: Évolution du RMSE pour l'estimateur PSA basé sur COMPRESS + calage

Conclusions

Il semblerait que :

- les méthodes agrégées permettent d'obtenir de meilleurs résultats dans les situations les moins favorables.
- la méthode des scores apporte de l'efficacité sur les modèles mal spécifiés et ne change pas les performances pour lorsque la méthode prédit des probabilités dans un ensemble fini (et de faible cardinal ?).
- l'augmentation du nombre de méthodes déteriore la qualité des estimateurs lorsqu'on utilise la méthode par calage.

References i


Breiman, L. (2004).
Random forests.
Machine Learning, 45:5–32.


Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983).
Classification and regression trees.


Chen, T. and Guestrin, C. (2016).
Xgboost: A scalable tree boosting system.
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.


Chipman, H. A., George, E. I., and McCulloch, R. E. (2010).
Bart: Bayesian additive regression trees.
The Annals of Applied Statistics, 4(1).

References ii


Cortes, C. and Vapnik, V. (1995).
Support vector networks.
Machine Learning, 20:273–297.


Gelein, B., Haziza, D., and Causeur, D. (2018).
Pondération pour correction de la non-réponse totale et machine learning.
Acte des JMS 2018, 1.


Hastie, T., Tibshirani, R., and Friedman, J. (2001).
***The Elements of Statistical Learning*.**
Springer Series in Statistics. Springer New York Inc., New York, NY, USA.


Haziza, D. and Beaumont, J.-F. (2007).
On the construction of imputation classes in surveys.
International Statistical Review, 75:25–43.

References iii


Little, R. J. A. (1986).
Survey nonresponse adjustments for estimates of means.
International Statistical Review, 54:139–157.


Quinlan, J. R. (1992).
Learning with continuous classes.


Quinlan, J. R. (1993).
Combining instance-based and model-based learning.
In *ICML*.


Zeileis, A. and Hornik, K. (2007).
Generalized m-fluctuation tests for parameter instability.
Statistica Neerlandica, 61:488–508.


Zeileis, A., Hothorn, T., and Hornik, K. (2008).
Model-based recursive partitioning.
Journal of Computational and Graphical Statistics, 17:492 – 514.

Merci de votre attention.