Causal inference to assess the effect of a treatment on survival

MAP 573 - Group 10
E. Askinazi, J. Feitz, C. Lescure, M. Resplandy, H. Zylberajch

Plan de la présentation

I. Définitions ATE et HTE
II. Traumabase et preprocessing
III. Stabilité de l'ATE
IV. ATE sur des clusters et HTE

I. Définitions ATE et HTE

Définitions

Individual Treatment Effect : $Y_i(1) - Y_i(0)$

Average Treatment Effect : $\tau := E[Y_i(1) - Y_i(0)]$

Conditional Average treatment Effect : $\mathbb{E}[Y(1) - Y(0) | X \in J]$

Heterogeneous Treatment Effect : $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$

Objectif : trouver des clusters au sein desquels l'ATE est négatif

II. Traumabase et preprocessing

II. Traumabase et preprocessing

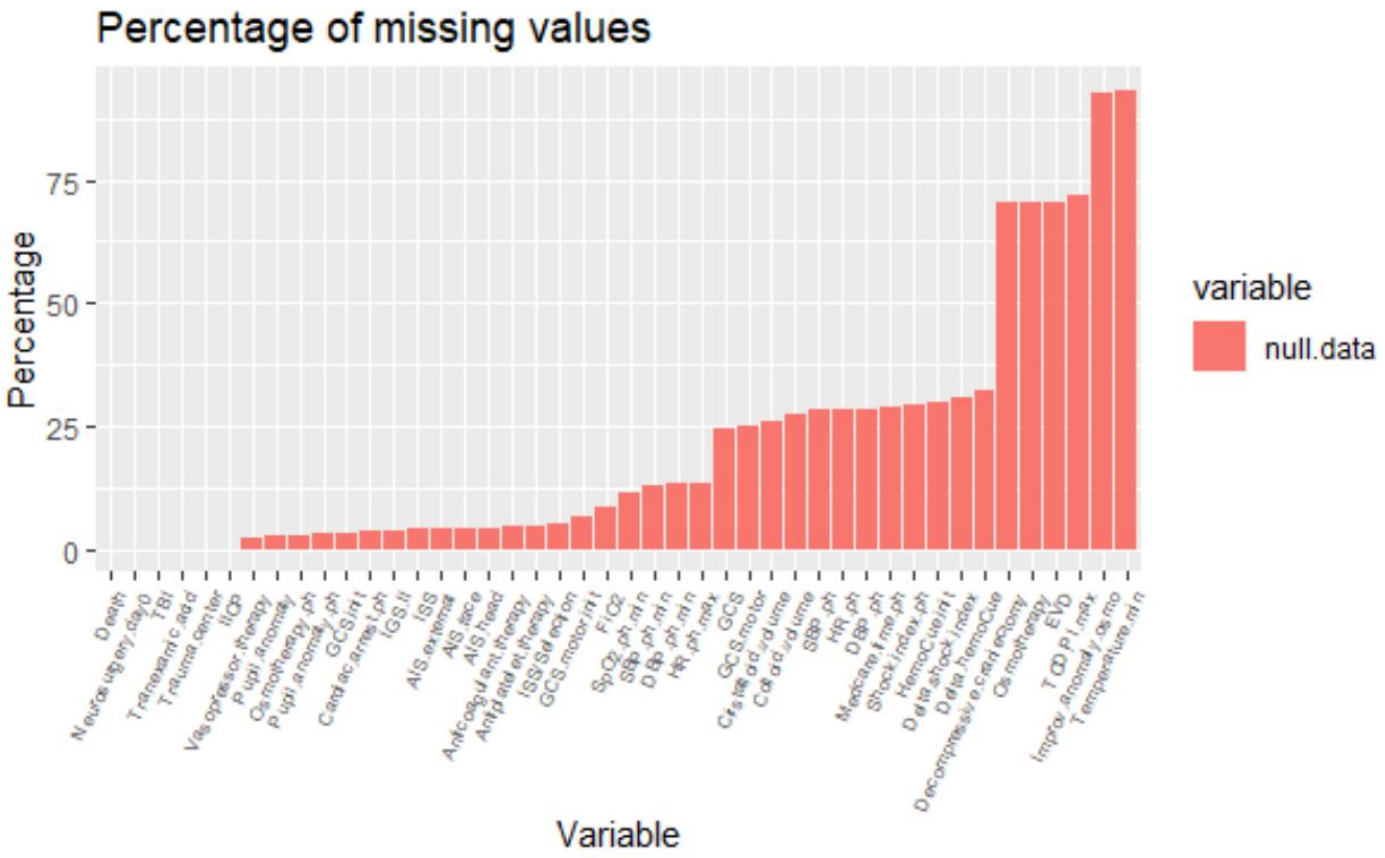
Deux types de valeurs manquantes :

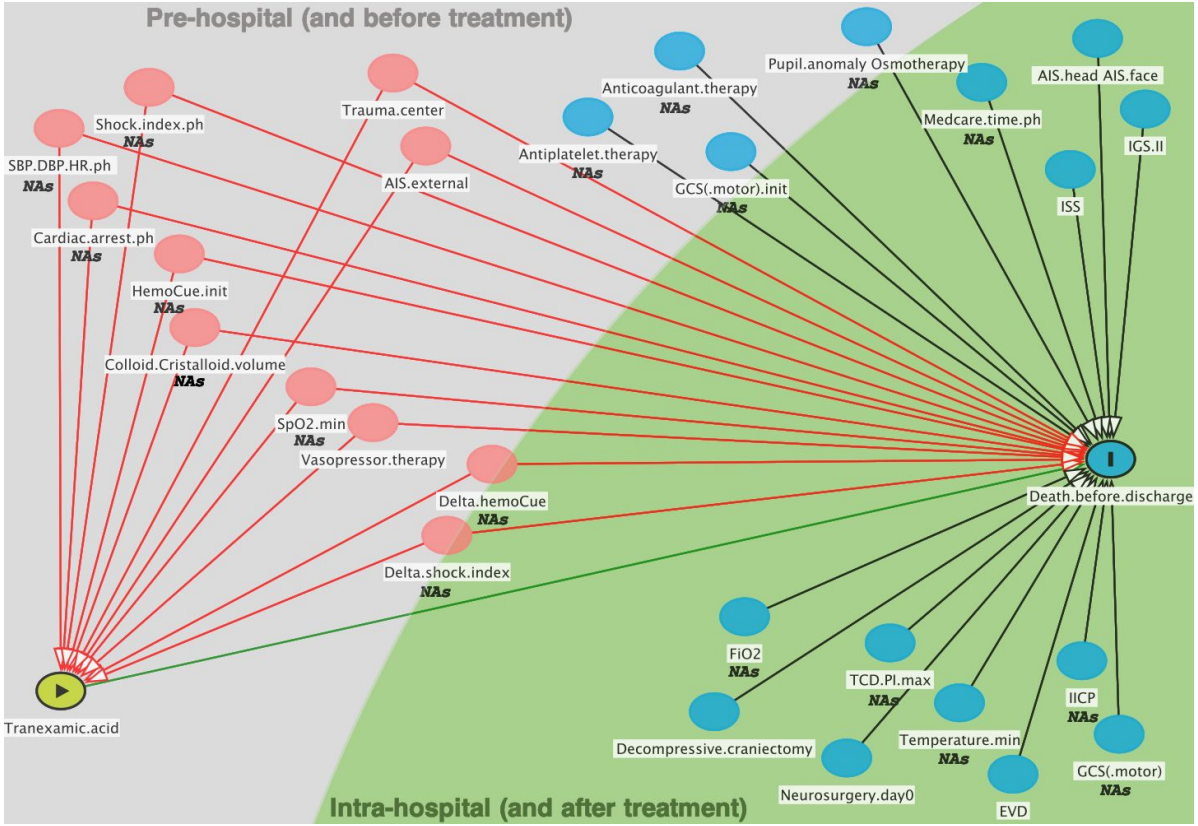
- MCAR : Missing Completely At Random.

  L’absence de la donnée n’a pas de cause particulière. Elle n’affecte pas les valeurs de la table.

- MNAR : Missing Not At Random

  La donnée manque pour une raison particulière. Cette absence peut être corrélée au contenu de la table.


Percentage of missing values
Percentage
75
50
25
0
Variable
variable
null.data

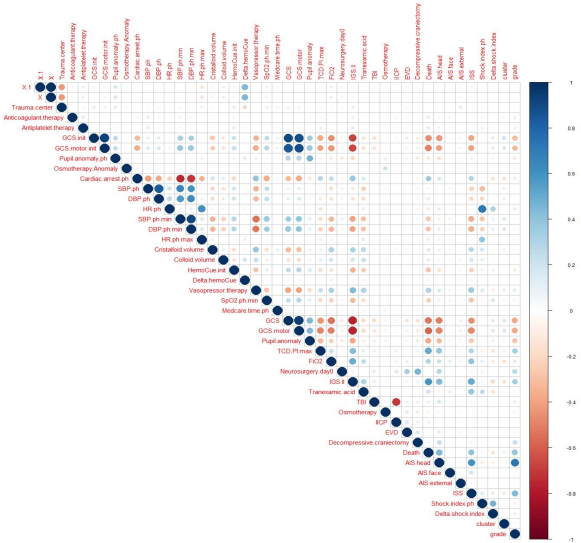

Pre-hospital (and before treatment)
Shock.index.ph
NAs
SBP.DBP.HR.ph
NAs
Cardiac.arrest.ph
NAs
HemoCue.init
NAs
Colloid.Cristalloid.volume
NAs
SpO2.min
NAs
Vasopressor.therapy
Delta.hemoCue
NAs
Delta.shock.index
NAs
Tranexamic.acid
Trauma.center
AIS.external
Anticoagulant.therapy
NAs
Antiplatelet.therapy
NAs
GCS(.motor).init
NAs
Pupil.anomaly Osmotherapy
NAs
Medcare.time.ph
NAs
AIS.head AIS.face
IGS.II
ISS
Death.before.discharge
FiO2
NAs
Decompressive.craniectomy
TCD.PI.max
NAs
Neurosurgery.day0
Temperature.min
NAs
EVD
IICP
NAs
GCS(.motor)
NAs
Intra-hospital (and after treatment)

Preprocessing

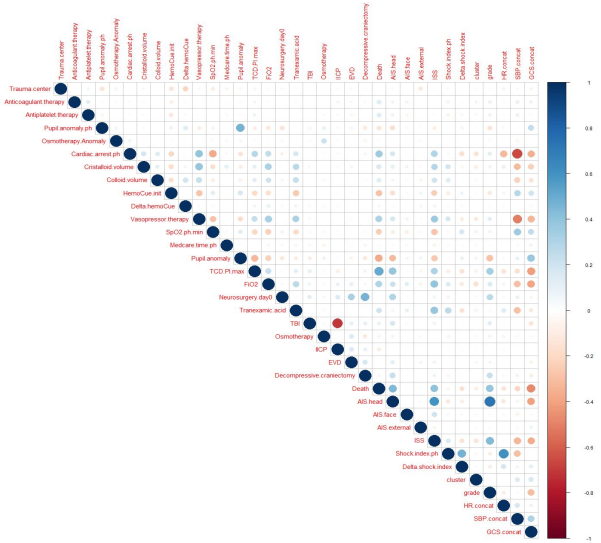
- Garder uniquement les covariables liées à Tranexamic.acid et Death → 39 covariables
- Garder uniquement les patients dont Trauma_cranien = 1 et AIS.head > 2
- Filtrer sur ISS/Selection → 5337 patients
- Fusion des colonnes Osmotherapy.ph et Improv.anomaly.osmo
- Suppression de Temperature.min (>90% de données manquantes)

Preprocessing

- Fusionner les colonnes très corrélées



Première matrice de corrélation
42 variables



Matrice de corrélation après fusion
33 variables

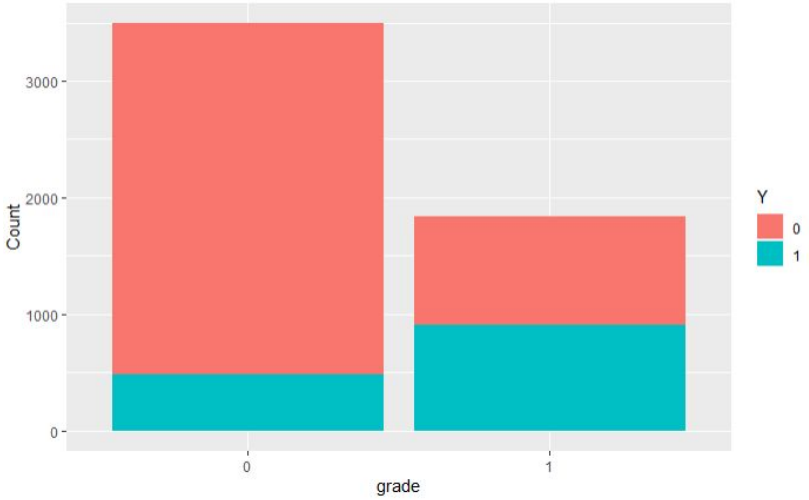
Data imputation

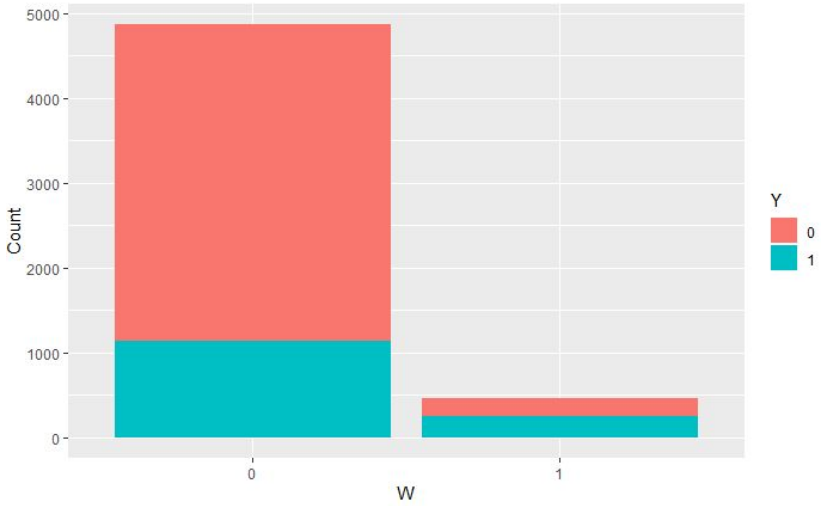
4 méthodes d'imputations :

- Simple Imputation : Imputation par moyenne
- MissForest : Imputation par Random Forest
- Mice : Imputation multiples avec plusieurs méthodes : pmm, logreg, polyreg
- FAMD : Imputed dataset en utilisant missMDA (ACP qui supporte des données catégorielles)

→ 8 datasets pour travailler

Contenu du dataset





III. Stabilité de l’ATE

III. Stabilité de l'ATE

Plusieurs décisions influent sur la valeur de l'ATE estimée :

- La méthode de calcul de l'ATE (OLS, IPW, AIPW…)
- Méthode d'imputation de la base
- Les méthodes d'estimation des paramètres nécessaires au calcul de l'ATE (propensity score par exemple)

Méthode de calcul de l'ATE

Nous avons utilisé trois méthodes :

- Ordinary Least Square Regression Adjustment (généralisation)
- Inverse-Propensity-Weighting:
- Augmented-Inverse-Propensity-Weighting:

Ordinary Least Square Regression Adjustment

$$
\begin{aligned}
\tau(x) &= E[\tau_i \mid X_i = x] \\
&= E[Y_i(1) - Y_i(0) \mid X_i = x] \\
&= E[Y_i(1)|X] - E[Y_i(0) \mid X_i = x] && \because \text{Linearity of expectations} \\
&= E[Y_i(1) \mid W_i = 1, X_i = x] - E[Y_i(0) \mid W_i = 0, X_i = x] && \because \text{Unconfoundedness} \\
&=: \mu(1, x) - \mu(0, x)
\end{aligned}
$$

L'ATE est obtenu en moyennant cette valeur sur l'échantillon

Inverse-Propensity-Weighting

On pondère chaque entrée par la probabilité qu'un patient a d'être traité : $e(x) = P(W_i = 1 \mid X_i = x)$

$$\tau = \mathbb{E}\left[\frac{Y_i W_i}{e(X_i)} - \frac{Y_i(1 - W_i)}{1 - e(X_i)}\right]$$

Augmented-Inverse-Propensity-Weighting

$$\tau = \mathbb{E}\left[W_i \frac{Y_i - \tau(1, X_i)}{e(X_i)} + (1 - W_i)\frac{Y_i - \tau(0, X_i)}{(1 - e(X_i))} + \tau(1, X_i) - \tau(0, X_i)\right]$$

Combinaison des deux estimateurs précédents

Doublement robuste

ATE en fonction de la base

On calcule l'ATE (méthode AIPW) avec les 5 bases de données créées par Mice

Base	ATE	std err.
MissForest	0.06201875	0.03421128
Mice 1	0.06599013	0.03428097
Mice 2	0.06591918	0.03269830
Mice 3	0.06515854	0.03049809
Mice 4	0.04747659	0.02389296
Mice 5	0.06897057	0.03042385

ATE en fonction de l'estimateur

	IPW	OLS	AIPW
Logistic regression	-1,827	-0.761	0.056
LGBoost	-0.209	-0.010	0.051
XGBoost	-0.054	0.094	0.060

L'AIPW est effectivement plus robuste : le résultat obtenu dépend moins de la méthode de régression utilisée

ATE en fonction des modèles de prédiction

tau\propensity	Logistic regression	LGBoost	XGBoost
Logistic regression	0.066 +/- 0.011	0.047 +/- 0.002	0.012 +/- 0.004
LGBoost	0.091 +/- 0.006	0.052 +/- 0.001	0.037 +/- 0.001
XGBoost	-0.011 +/- 0.013	0.091 +/- 0.001	0.062 +/- 0.004

Même l'AIPW est relativement instable.

IV. ATE sur clusters

ATE sur cluster des médecins

tau\propensity	Logistic regression	LGBoost	XGBoost
Logistic regression	-0.046 +/- 0.078	-0.001 +/- 0.004	0.0016 +/- 0.002
LGBoost	-0.174 +/- 0.036	0.011 +/- 0.002	0.011 +/- 0.002
XGBoost	-0.587 +/- 0.098	0.0347 +/- 0.007	0.055 +/- 0.004

ATE sur cluster des médecins

On peut calculer l'ATE sur les clusters établis par les médecins mais les résultats sont peu probants

Lésion axonale diffuse

estimate	std.err
0.04986106	0.03906732

Lésion extra-axiale

estimate	std.err
0.03626871	0.05964241

Lésion intra-axiale

estimate	std.err
0.1532967	0.1183543

Clustering

Données catégorielles → FAMD pour passer dans un espace continu

On utilise kmeans pour former différents clusters de patients qui ont des caractéristiques similaires

On calcule l'ATE sur ces différents clusters

ATE sur cluster

Cluster	Nombre d'individus	ATE	std. err
0	304	0.44	0.22
1	122	0.083	0.074
2	80	0.062	0.075
3	182	0.58	0.39
4	327	-0.20	0.20
...	...	...	...
15	92	-0.11	0.06

Etude des clusters intéressants

Le cluster n° 15 est particulièrement intéressant pour son ATE négatif

On s'intéresse aux caractéristiques de ce cluster

Covariable	GCS.motor.init	Cristalloid.volume	Vasopressor.therapy	GCS	GCS.motor	FiO2
Moyenne sur le cluster	3.478	1789	0.47	6.80	3.18	0.87
Moyenne globale	4.243	793	0.19	9.11	4.12	0.63

Forêts causales

Entrainement

Les noeuds sont splittés de façon à maximiser l'hétérogénéité (ie. la différence de CATE) entre les 2 nouveaux noeuds enfants

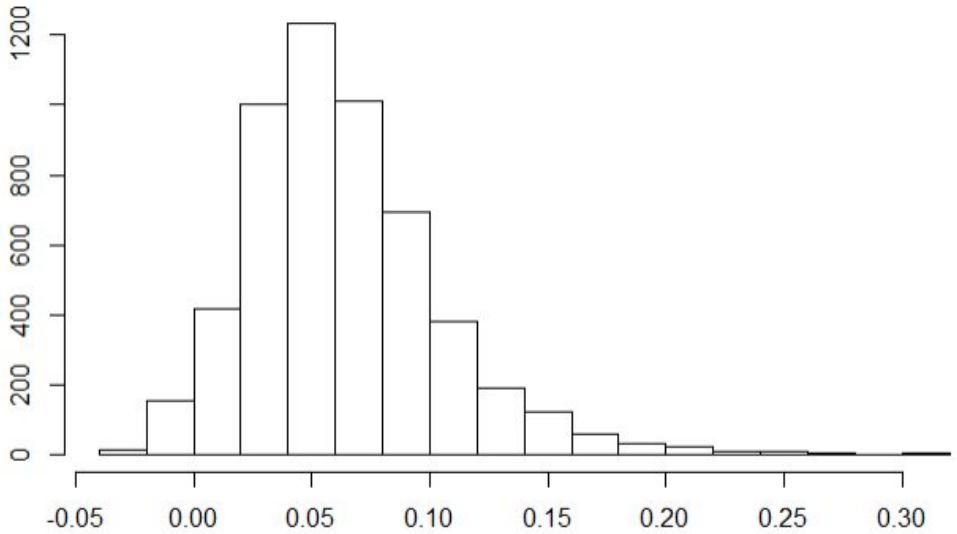
Les arbres sont **honnêtes** : les samples des feuilles ne sont pas ceux utilisés pour décider les splits

Calcul de l'ATE : on regarde la valeur moyenne pour W=1 et on soustrait la prédiction pour W=0

Sortie : on obtient une estimation de HTE

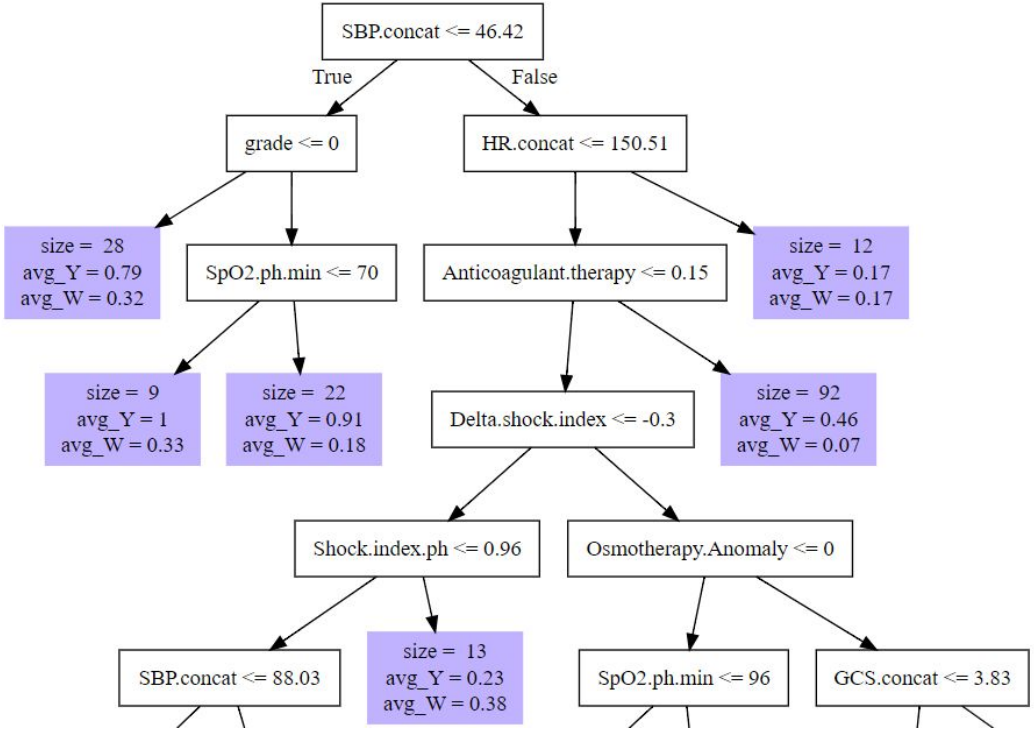
Implémentation : package `grf` (generalized random forests) en R

Forêts causales



Histogramme de l'HTE

Forêts causales



Forêts causales

Une donnée exploitable est les variables les plus utilisées lors des splits des noeuds des arbres :

Shock.index.ph	Cristalloid.volume	Delta.shock.index	HR.ph	Delta.hemoCue
0.1078471864	0.0925255985	0.0903593147	0.0822154104	0.0774661518

ISS	IGS.II	HemoCue.init	HR.ph.max	SpO2.ph.min
0.0657048374	0.0627114365	0.0528062748	0.0511535471	0.0295432506

Forêts causales

On peut également regarder l'effet des variables importantes dans le clustering, afin de voir des corrélations entre une covariable et le HTE