Comment constituer des groupes de réponse homogène ?

Une comparaison de quelques méthodes appliquées aux enquêtes sectorielles annuelles en France

---

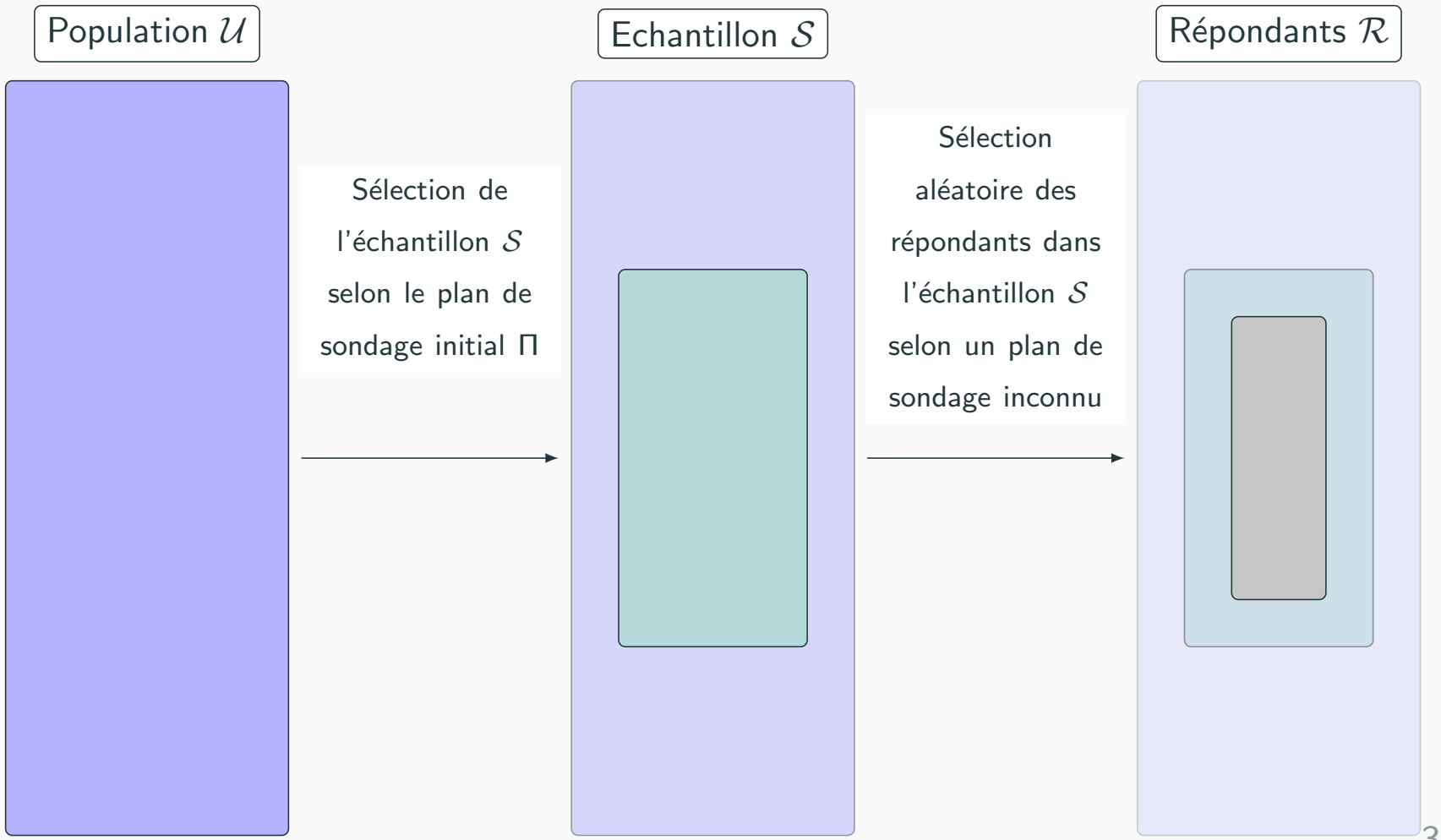
Thomas Deroyon - Insee

13 Octobre 2016

INSEE - Département des Méthodes Statistiques

Introduction

Repondération : la non-réponse comme phase additionnelle du plan de sondage 1/2



Repondération : la non-réponse comme phase additionnelle du plan de sondage 2/2

- $\pi_{i/S}$ probabilité de répondre ou propension à répondre : probabilité d'inclusion simple associé à la deuxième phase du plan de sondage
- $\pi_{i/S}$ connus $\Rightarrow \sum_{i \in R} \frac{y_i}{\pi \pi_{i/S}}$ estime sans biais le total de y
- $\pi_{i/S}$ inconnus $\Rightarrow$ objectif de la repondération : estimer $\pi_{i/S}$
- $\hat{\pi}_{i/S}$ estimateur convergent de $\pi_{i/S} \Rightarrow \sum_{i \in R} \frac{y_i}{\pi \hat{\pi}_{i/S}}$ estimateur (asymptotiquement) sans biais du total de y
- Hypothèses : comportements de réponse indépendants + non-réponse ignorable

La méthode des groupes de réponse homogène

PRINCIPE : l'échantillon peut être partitionné en groupes à l'intérieur desquels :

- les comportements de réponse sont indépendants
- toutes les observations ont la même probabilité de réponse
- la probabilité de réponse commune est estimée par la part des répondants dans chaque groupe

REMARQUES :

- GRH constitués à partir d'une première estimation de $\pi_{i/S}$: $<$ détérioration $>$ d'une estimation initiale ?
- méthode plus robuste : évite les $\hat{\pi}_{i/S}$ faibles
- biais faible ou nul si la corrélation dans chaque GRH entre probabilité de réponse et variable d'intérêt est faible ou nulle ([Bethlehem 1988])

Le problème du surapprentissage

PROBLÈME : repondération $\Leftrightarrow$ approche sous le plan de sondage :

- estimateur sans biais $\Leftrightarrow$ la moyenne de $\sum_{i \in R} \frac{Y_i}{\pi_i \ pi_{i/S}}$ sur l'ensemble des S et R possibles est égale à $\sum_{i \in U} y_i$
- **problème** : on ne dispose que d'une réalisation du comportement de réponse pour estimer les $\pi_{i/S}$
- risque de **surapprentissage** : les $\hat{\pi}_{i/S}$ sont trop adaptées au comportement observé dans R et pas au comportement de réponse sous-jacent

Principe de l'étude

Principe

IDÉE DE BASE : couper aléatoirement S en deux

- **échantillon d'apprentissage** S_a : chaque unité a 2 chances sur 3 d'appartenir à S_a
- **échantillon test** S_t : chaque unité a 1 chance sur 3 d'appartenir à S_t
- R_t : répondants de S_t

$\Rightarrow$ les comportements de réponse sont les mêmes sur S_a et S_t
$\Rightarrow$ calcul des $\hat{\pi}_{i/S}$ sur S_a et application sur R_t
$\Rightarrow$ calcul d'indicateurs sur R_t permettant de juger de la capacité des méthodes de CNR à **bien décrire le comportement de réponse** et **réduire le biais**
$\Rightarrow$ permet de contrôler le risque de surapprentissage

Critères pour la qualité de la description du comportement de réponse

- ERREUR ABSOLUE MOYENNE : $\frac{\sum_{i \in R_t} |R_i - \hat{\pi}_{i/S}|}{|R_t|}$
- ERREUR QUADRATIQUE MOYENNE : $\frac{\sum_{i \in R_t} (R_i - \hat{\pi}_{i/S})^2}{|R_t|}$
- SENSIBILITÉ (part de répondants détectés) : $\frac{\sum_{i \in R_t} R_i \, \hat{\pi}_{i/S}}{\sum_{i \in R_t} R_i}$
- SPÉCIFICITÉ (part de non-répondants détectés) : $\frac{\sum_{i \in R_t} (1 - R_i)(1 - \hat{\pi}_{i/S})}{\sum_{i \in R_t} 1 - R_i}$

Critères pour juger de la réduction du biais

IDÉE DE BASE : R_t échantillon sélectionné dans S par un plan de sondage composé de deux phases poisoniennes (non-réponse + sélection à probabilité égale de 1/3)

- $\sum_{i \in R_t} \frac{3\, z_i}{\hat{\pi}_{i/S}}$ estime sans biais $\sum_{i \in S} z_i$ si $\hat{\pi}_{i/S}$ sont corrects
- application à $z_i = \frac{y_i}{\pi}$
- comparaison de $\hat{T}_{y,S} = \sum_{i \in S} \frac{y_i}{\pi_i}$ et $\hat{T}_{y,R_e} = \sum_{i \in R_t} \frac{3\, y_i}{\pi_i\, \hat{\pi}_{i/S}}$ pour un ensemble de variables fiscales
- variance de $\hat{T}_{y,R_e}$ calculable $\Rightarrow$ l'écart entre $\hat{T}_{y,S}$ et $\hat{T}_{y,R_e}$ indique-t-il un biais résiduel ?

Les données

ESANE : estimer les statistiques structurelles d'entreprise (notamment les comptes agrégés par secteurs)

- données fiscales exhaustives (comptes de résultats + bilans)
- enquêtes sur un échantillon pour mesurer la ventilation du chiffre d'affaires par activité et réévaluer le secteur

ENQUÊTES SECTORIELLES ANNUELLES :

- échantillon de 150 000 entreprises, stratifié par secteur et taille
- 80 000 entreprises dans des strates exhaustives
- étude : strates non exhaustives du **secteur de la construction** (section F de la Nace) dans les ESA 2014

$\Rightarrow$ 5 803 entreprises sélectionnées parmi 387 500

Méthodes de correction de la non-réponse par groupes de réponse homogène

Méthodes de correction de la non-réponse par groupes de réponse homogène

Les méthodes directes

La méthode par croisements

PRINCIPE : identifier les variables auxiliaires (qualitatives) corrélées au fait d'être répondant et découper la population suivant leurs modalités

- identification et classement par régression logistique
- découpage de la population suivant les modalités de la variable auxiliaire la plus explicative
- itération du processus jusqu'à ce que plus aucune variable ne soit significative ou que les GRH contiennent moins de 50 observations

Application :

- variables utilisées : CA, investissement, effectif, indicatrice d'imputation des données fiscales, statut d'entrepreneur individuel, durée d'existence
- 38 GRH

Les méthodes par arbres

CHAID : voir [Kass 1980]

- même principe que la méthode par croisement, mais identification de la variable utilisée pour découper la population à chaque étape automatiquement par des tests du χ^2
- paramètre : seuil de significativité retenu pour les tests du χ^2
- choisi par validation croisée : 10 GRH

CART : voir [Breiman et al.. 1984]

- découpage itératif de la population
- à chaque étape, découpage en deux suivant la variable qui partage dans les deux groupe faisant le plus baisser l'EQM
- construction de l'arbre le plus long possible (taille des feuilles de 1 ou 2 unités)
- élagage de l'arbre par validation croisée : 8 GRH

Méthodes de correction de la non-réponse par groupes de réponse homogène

Les méthodes des scores

Principe et estimation des probabilités de réponse

PRINCIPE : méthodes en deux étapes

1. estimation d'une probabilité de réponse $\hat{\pi}_{i/S}$
2. regroupement des observations ayant des valeurs proches de $\hat{\pi}_{i/S}$

ESTIMATION DES PROBABILITÉS DE RÉPONSE : quatre méthodes utilisées

- **régression logistique**
- **bagging** : voir [Breiman 1994]
- **boosting** : voir [Freund et Shapire 1997]
- **forêts aléatoires** : voir [Breiman 2001]

Les méthodes de constitution de GRH à partir des scores

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE :
Nombre de GRH : 38 + nombre de GRH obtenu par des critères usuels de qualité d'une partition

QUANTILES : GRH définis par les quantiles de la distribution des $\hat{\pi}_{i/S}$
Nombre de GRH : 38 + méthode proposée par [Haziza et Beaumont 2007]

- en partant de 2 GRH, calcul du R^2 de la régression de $\hat{\pi}_{i/S}$ sur les indicatrices d'appartenance aux GRH
- arrêt dès que le R^2 dépasse un seuil fixé *a priori*
- choix du seuil par validation croisée

CENTRES MOBILES : application des centres mobiles aux centres des GRH obtenus par la méthode des quantiles
Nombre de GRH : 38 + utilisation de la méthode de Haziza et Beaumont avec choix du paramètre par validation croisée

Résultats

Comportement de réponse 1/2

Méthode	Nb GRH	EAM	EQM	Sensibilité	Specificité
Croisements	38	0.407	0.207	0.641	0.531
Cart	8	0.393	0.194	0.641	0.562
Chaid	10	0.402	0.199	0.633	0.553
Logit		0.412	0.203	0.623	0.542
Logit + CAH	9	0.412	0.204	0.627	0.537
Logit + CAH	38	0.412	0.209	0.625	0.539
Logit + Q	0.95/6	0.415	0.205	0.625	0.533
Logit + Q	38	0.414	0.208	0.621	0.541
Logit + CM	0.95/5	0.415	0.206	0.630	0.525
Logit + CM	38	0.412	0.207	0.623	0.543

Comportement de réponse 2/2

Méthode	Nb GRH	EAM	EQM	Sensibilité	Specificité
Croisements	38	0.407	0.207	0.641	0.531
Bagging		0.393	0.192	0.640	0.563
Bagging + CAH	3	0.400	0.193	0.633	0.558
Bagging + Q	0.95/6	0.388	0.195	0.663	0.547
Bagging + CM	0.99/8	0.388	0.195	0.659	0.550
Boosting		0.395	0.195	0.643	0.556
Boosting + CAH	6	0.400	0.197	0.640	0.549
Boosting + Q	0.98/9	0.397	0.195	0.642	0.553
Boosting + HB	0.98/8	0.397	0.196	0.643	0.551
Forêt aléatoire		0.399	0.201	0.634	0.558
Forêt aléatoire + CAH	3	0.348	0.256	0.679	0.616
Forêt aléatoire + Q	0.999/35	0.322	0.240	0.744	0.592
Forêt aléatoire + HB	0.999/30	0.323	0.241	0.746	0.588

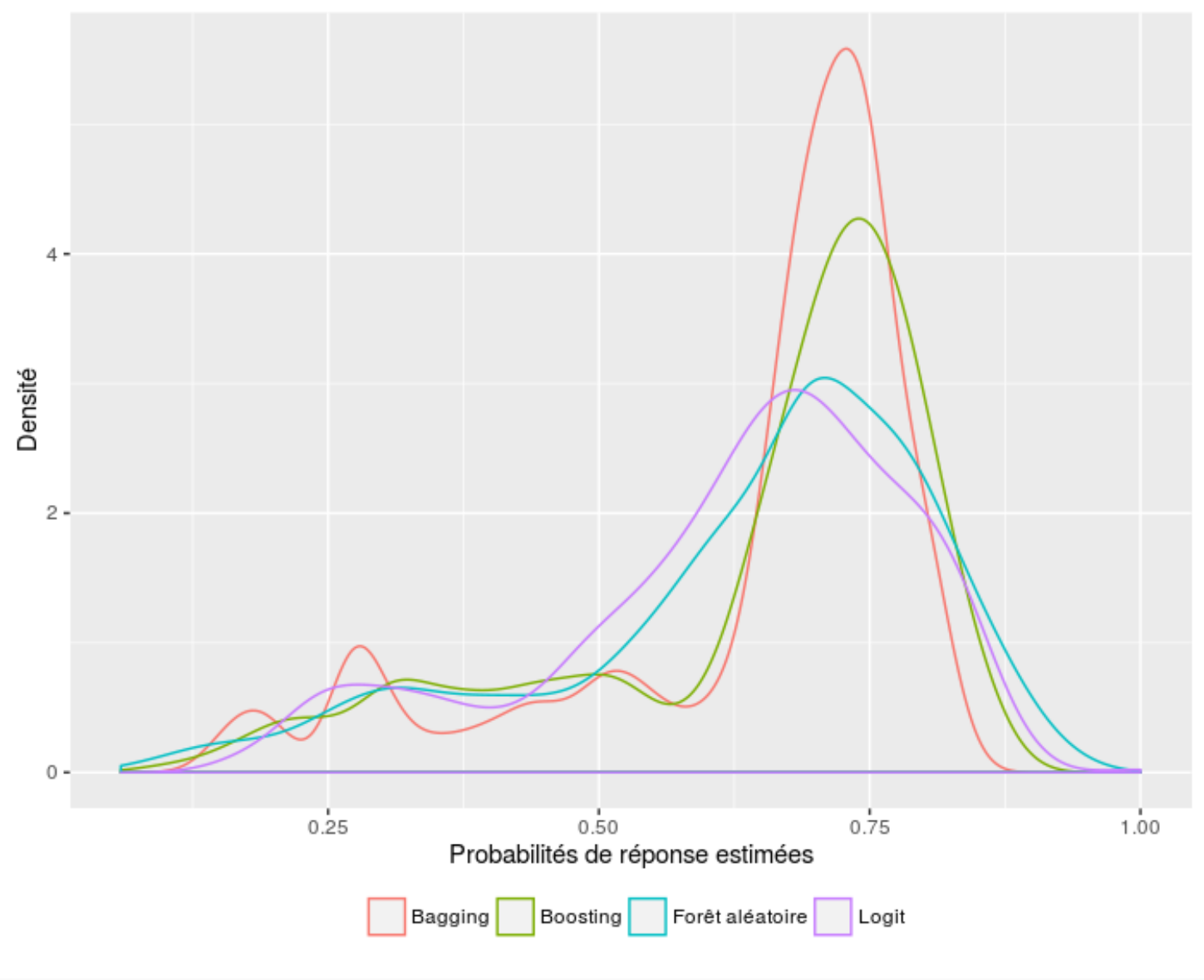
Biais 1/2

Méthode	Ecart relatif sur le CA	Ecart / CV sur le CA	Ecart relatif sur les immobilisations	Ecart / CV sur les immobilisations
Croisements	-11.430	2.165	-4.770	0.721
Cart	-11.070	1.970	-2.150	0.348
Chaid	-11.150	1.970	-4.030	0.632
Logit	-9.380	1.918	-5.200	0.720
Logit + CAH	-11.460	2.091	-6.460	0.858
Logit + CAH38	-12.340	2.396	-6.110	0.887
Logit + Q	-12.840	2.281	-6.630	0.896
Logit + Q38	-13.040	2.362	-7.480	1.007
Logit + CM	-11.740	2.119	-5.840	0.816
Logit + CM38	-12.530	2.447	-7.040	0.974

Biais 2/2

Méthode	Ecart CA	Ecart / CV CA	Ecart Imm.	Ecart / CV Imm.
Croisements	-11.430	2.165	-4.770	0.721
Bagging	-8.210	1.561	-2.020	0.315
Bagging + CAH	-12.380	2.153	-4.900	0.747
Bagging + Q	-3.820	0.775	1.940	0.306
Bagging + CM	-5.080	1.008	0.530	0.083
Boosting	-9.850	1.788	-3.040	0.450
Boosting + CAH	-10.650	1.922	-3.930	0.584
Boosting + Q	-10.090	1.828	-3.360	0.492
Boosting + CM	-10.660	1.910	-3.840	0.563
Forêt aléatoire	-12.830	2.153	-6.160	0.834
Forêt aléatoire + CAH	-55.620	4.658	-48.290	3.761
Forêt aléatoire + Q	-43.890	2.788	-41.410	2.850
Forêt aléatoire + HB	-47.140	2.959	-47.760	3.227

Probabilités de réponse en entrée de méthodes des scores



Conclusion

- Méthode des scores avec bagging présente de bonnes propriétés pour cette enquête
- Limites : choix des paramètres des bagging / boosting / forêts aléatoires
- Limites : dépendance au choix de $R_t \Rightarrow$ répéter le processus sur plusieurs tirages de R_a et R_t

Bibliographie

Bethlehem, J., *Reduction of nonreponse biais through regression estimation*, Journal of Official Statistics, 1988

Breiman, L., Friedman, J. et Ohlsen, R., *Classification and Regression Trees*, CRC Press, 1984

Breiman, L., *Bagging Predictors*, Technical Report of the Department of Statistics, University of California, 1994

Breiman, L., *Random Forests*, Machine Learning, 2001

Brick, J., *Unit non-response and weighting adjustment - a critical review*, Journal of Official Statistics, 2013

Bühlmann, P., *Bagging, boosting and ensemble methods*, *in* Handbook of Computational Statistics : Concepts and Methods, Springer Verlag, 2012

Dequidt, E., Sigler, N. et Buisson, B., *Comparaison de méthodes pour la correction de la non-réponse totale : méthode des scores et segmentation*, Actes du Septième Colloque Francophone sur les Sondages, 2012

Freund, Y., Shapire, R., *A decision-theoretic approach of on-line learning and an application to boosting*, Journal of Computer and System Science, 1997

Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, Springer Verlag, 2009

Haziza, D., Beaumont, J.-F., *On the construction of imputation classes in surveys*, International Statistical Review, 2007

Haziza, D., Lesage E., *A discussion of weighting procedures for unit non-response*, Journal of Official Statistics, 2016

Kass G., *An exploratory technique for investigating large quantities of categorical data*, Journal of Applied Statistics, 1980

Phipps, P. et Toth, D., *Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data*, The Annals of Applied Statistics, 2012