# Comment constituer des groupes de réponse homogène ?

## Une comparaison de quelques méthodes appliquées aux Enquêtes Sectorielles Annuelles en France

Thomas Deroyon [1]

[1] *Insee - 18 Boulevard Adolphe Pinard, 75014 Paris - thomas.deroyon@insee.fr*

**Résumé :**
La correction de la non-réponse totale est souvent traitée à l'Insee par repondération suivant la méthode des groupes de réponse homogène, pour les enquêtes auprès des entreprises comme auprès des ménages. Or, de nombreuses méthodes sont disponibles pour constituer ces groupes. La méthode par croisement, utilisée pour la correction de la non-réponse totale dans les enquêtes sectorielles annuelles permettant de produire les statistiques structurelles d'entreprise, est simple à comprendre et à mettre en œuvre, mais peut s'avérer lourde et nécessite beaucoup d'interventions de la personnes en charge des traitements. De nombreuses autres méthodes (méthodes de classification par arbre suivant les algorithmes CHAID ou CART, méthodes des scores), plus rapides, peuvent également être utilisées. Les performances de ces différentes méthodes peuvent être comparées dans des simulations où le mécanisme de réponse est supposé connu et répliqué. En situation réelle, le praticien est cependant confronté à un mécanisme de réponse inconnu qu'il doit essayer de décrire au mieux de façon à réduire le plus possible le biais de non-réponse. Cette étude a pour but de comparer les résultats d'un ensemble assez large de méthodes de constitution de groupes de réponse homogène appliquées aux données des Enquêtes Sectorielles Annuelles collectées en 2015 sur l'année 2014. Ces comparaisons portent à la fois sur un ensemble de mesures de qualité de la correction de la non-réponse et sur une série d'estimateurs de variables d'intérêt, dans des domaines de diffusion plus ou moins fins.

**Mots-clés :** Traitement de la non-réponse, Enquêtes entreprises, Répondération, Groupes de réponse homogène,...

La non-réponse altère la qualité des estimateurs construits à partir de données d'enquête. Elle augmente leur variance en réduisant le nombre d'observations exploitables ; elle introduit également un biais dans les estimations, qui tient surtout au fait que les caractéristiques des non-répondants, notamment celles qui intéressent l'enquête, sont susceptibles de différer de celles des répondants. Les traitements qui interviennent après la collecte des données et leurs contrôles et redressements ont parmi leurs objectifs principaux de réduire autant que possible le biais introduit par la non-réponse et d'améliorer

la précision des estimateurs construits sur l'échantillon des répondants.

De plus en plus, à l'Insee, ces traitements se font suivant une approche en deux étapes (Haziza et Lesage (2016)), pour les enquêtes auprès des entreprises comme auprès des ménages : d'abord, une correction de la non-réponse, par imputation pour la non-réponse partielle et le plus souvent par repondération pour la non-réponse totale, suivie d'un calage sur marges. Une troisième étape d'identification et de traitement des valeurs influentes est aussi parfois mise en œuvre, avant ou après le calage sur marges.

La correction de la non-réponse totale par repondération se place dans une approche sous le plan de sondage. En l'absence de non-réponse, l'estimateur $\sum_{i\in\mathcal{S}} \frac{Y_i}{\pi_i}$ du total de la variable $Y$ sur l'échantillon $\mathcal{S}$ obtenu avec les poids de sondage $\pi_i$ est en effet un estimateur sans biais : sa moyenne sur l'ensemble des échantillons possibles est égale au total de $Y$ sur la population. La non-réponse est alors assimilée à une phase additionnelle et non maîtrisée du plan de sondage, dans laquelle chaque observation échantillonnée à une certaine probabilité non-nulle mais inconnue $\pi^R_{i/\mathcal{S}}$ d'être répondante. Si $\mathcal{R}$ désigne le sous-échantillon des répondants, l'estimateur $\sum_{i\in\mathcal{R}} \frac{Y_i}{\pi_i \pi^R_{i/\mathcal{S}}}$ est également un estimateur sans biais du total de $Y$ dans la population. L'objectif de la correction de la non-réponse est alors de remplacer $\pi^R_{i/\mathcal{S}}$ dans cette formule par un estimateur $\hat{\pi}^R_{i/\mathcal{S}}$ construit sur la base d'un modèle de régression linéaire généralisée expliquant la probabilité d'être répondant par un ensemble de variables auxiliaires observées à la fois pour les répondants et les non-répondants. Kim et Kim (2007) ont montré que le biais de l'estimateur $\hat{Y}^{CNR} = \sum_{i\in\mathcal{R}} \frac{Y_i}{\pi_i \hat{\pi}^R_{i/\mathcal{S}}}$ ainsi obtenu tend vers 0 quand la taille de l'échantillon et de la population tendent vers l'infini, si le modèle est bien spécifié.

En pratique, les méthodes les plus souvent utilisés pour décrire le comportement de réponse reposent sur l'utilisation de groupes de réponse homogène (GRH, Brick (2013)). L'échantillon est partitionné à l'aide des variables auxiliaires en un ensemble de groupes à l'intérieur desquels nous supposons que toutes les unités, répondantes ou non-répondantes, ont le même comportement de réponse. Cela revient à estimer les probabilités de réponse $\pi^R_{i/\mathcal{S}}$ dans un modèle de régression linéaire généralisé dont les régresseurs sont les indicatrices d'appartenance aux groupes de réponse homogène. Cela revient également à décrire la phase additionnelle du plan de sondage représentant la non-réponse comme un sondage bernoullien stratifié, les strates étant égales aux groupes de réponse homogène.

Bethlehem (1988) et Little et Vartivarian (2005) ont montré que les variables auxiliaires utilisées pour construire les groupes de réponse homogène doivent être liées à la fois aux variables d'intérêt de l'enquête et au comportement de réponse. En particulier, si la corrélation dans chaque groupe de réponse homogène entre les probabilités de réponse et une variable d'intérêt $Y$ est nulle, alors l'estimateur corrigé de la non-réponse est

sans biais. Comme les variables d'intérêt dans les enquêtes sont nombreuses et que les poids corrigés de la non-réponse doivent permettre de réduire le biais des estimateurs pour toutes ces variables, les variables auxiliaires utilisées pour construire les GRH sont en général sélectionnées sur la base de leur capacité à prédire le comportement de réponse.

La méthode par croisement est la technique de base pour la constitution de groupes de réponse homogène et est utilisée notamment pour la correction de la non-réponse totale dans les enquêtes sectorielles annuelles. Dans un premier temps, nous identifions les variables auxiliaires qualitatives corrélées à la non-réponse à l'aide d'un modèle de régression logistique. Les régresseurs sont introduits sans interaction, et leurs modalités sont progressivement regroupées sur la base de tests de significativité ou de comparaisons de critères d'information. Les variables ainsi constituées sont ensuite classées de la plus à la moins corrélée au comportement de réponse.
Les groupes de réponse homogènes sont alors formés en découpant l'échantillon suivant les modalités de la variable la plus corrélée au comportement de réponse ; puis les groupes ainsi obtenus sont découpés suivant les modalités ou regroupements de modalités de la deuxième variable la plus corrélée au comportement de réponse. Ce processus est itéré jusqu'à ce que toutes les variables corrélées au comportement de réponse aient été utilisées, ou que les groupes constitués contiennent moins d'un nombre minimal d'observations, fixé arbitrairement, souvent à 100 ou 50 suivant les enquêtes. A chaque étape de découpage, il est possible de regrouper les modalités de la variable utilisée pour constituer les groupes si celles-ci conduisent à constituer des GRH de trop petite taille. Cette méthode permet en général de constituer des groupes interprétables aisément, mais leur constitution est longue et n'est pas aisément automatisable.

L'algorithme de *Chi-Squared Automatic Interaction Detection* (CHAID, Kass (1980)) est aussi très souvent utilisé. Le principe en est très proche de celui de la méthode par croisement présenté plus haut, mais l'identification des variables les plus corrélées au comportement de réponse et les regroupements de leurs modalités se font sur la base de tests de corrélation du $\chi^2$. L'algorithme CHAID et la méthode par croisement ne sont cependant pas du tout protégés contre le risque de sur-apprentissage, *i.e.* le risque d'identifier comme corrélées à l'indicatrice de réponse et participant de ce fait à la construction des GRH des variables qui, en réalité, n'expliquent pas la non-réponse, mais paraissent significatives du fait que la taille finie des échantillons sur lesquels les estimations des modèles et les tests sont conduits ne permet pas de séparer complètement les relations réelles entre variables du bruit. Le risque est alors d'augmenter inutilement la dispersion des poids corrigés de la non-réponse, ce qui peut se traduire par une détérioration de la précision des estimations (Little et Vartivarian (2005)). D'autres algorithmes de classification par arbre, notamment l'algorithme de *Classification And Regression Trees* (CART, Breiman, Friedman et Ohlsen (1984), utilisé par exemple par Phipps et Toth (2012)), permettent de construire des GRH moins sensibles au risque de sur-apprentissage.

D'autres méthodes, appelées méthodes des scores, sont également de plus en plus utilisées. Elles consistent à résumer d'abord l'information apportée par les variables auxiliaires sur le comportement de réponse dans une estimation $\hat{\pi}^R_{i/\mathcal{S}}$ de la probabilité de réponse. Cette estimation est souvent issue d'un modèle de régression linéaire généralisée, mais elle n'est pas utilisée directement pour calculer les poids corrigés de la non-réponse. En effet, d'une part, la forme du modèle, notamment quand il contient des variables auxiliaires continues, ne permet pas de maîtriser l'amplitude des valeurs de $\hat{\pi}^R_{i/\mathcal{S}}$, dont certaines peuvent tendre vers 0 et se traduiraient de ce fait par des ajustements très forts des poids et une dispersion élevée et sans doute excessive des poids corrigés de la non-réponse. D'autre part, la valeur exacte de $\hat{\pi}^R_{i/\mathcal{S}}$ peut dépendre de la forme du modèle retenu pour l'estimer et n'est de ce fait pas utilisable directement, mais son ordre de grandeur reflète correctement la probabilité de réponse sous-jacente $\pi^R_{i/\mathcal{S}}$ si le modèle est de qualité. Les groupes de réponse homogène sont dès lors constitués en regroupant les unités, répondantes ou non-répondantes, qui ont des valeurs proches des probabilités de réponse estimée. Plusieurs techniques de regroupement peuvent encore être utilisées. La plus simple consiste à découper l'échantillon suivant les quantiles de la distribution des $\hat{\pi}^R_{i/\mathcal{S}}$ (méthode des quantiles). Il est également possible de constituer les GRH à l'aide d'une classification ascendante hiérarchique dans laquelle la distance entre deux unités correspond à la différence entre leurs probabilités de réponse estimée. Haziza et Beaumont (2007) ont également proposé une méthode reposant sur l'utilisation de l'algorithme des centres-mobiles.

Les méthodes des scores reposent sur la qualité de l'estimation des probabilités de réponse utilisées pour constituer les GRH. Or, la sélection d'un modèle de régression linéaire généralisée de qualité pose de nombreuses questions pratiques : faut-il introduire les variables continues sous forme de polynômes ou en les ayant au préalable découpées en tranches ? Comment dans ce cas constituer ces tranches ? Comment sélectionner simplement les interactions entre variables et les regroupements de modalités pertinents ? D'autres méthodes que les régressions linéaires généralisées, dans laquelle la sélection des variables et de leurs interactions est traitée plus automatiquement, peuvent être proposées pour estimer les probabilités de réponse : *bagging* ou *boosting* d'arbres de classification, algorithmes de forêts aléatoires (Breiman (1994), Freund et Shapire (1997), Bühlmann (2012), Breiman (2001), Hastie, Tibshirani et Friedman (2009)) ... Plus généralement, toute méthode de classification permettant de prédire une variable à valeur dans $\{0, 1\}$ ou de décrire sa loi de probabilité conditionnelle à des variables auxiliaires peut être utilisée.

Il existe donc une très grande variété de méthodes pour constituer des groupes de réponse homogène. La méthode par croisement actuellement utilisée dans les enquêtes sectorielles annuelles aboutit à une description fine du comportement de réponse (en environ 500 groupes de réponse homogène) mais est assez lourde à mettre en œuvre. De plus, elle ne protège pas contre le risque de sur-apprentissage. Les autres méthodes testées aboutissent

à la formation d'un nombre beaucoup plus faible de groupes de réponse homogène (d'une trentaine à une centaine) et sont pour la plupart plus légères et rapides à mettre en œuvre car elles nécessitent moins d'interventions de l'utilisateur. Mais, avant d'abandonner la méthode par croisement, il faut s'assurer que les autres méthodes ont des performances équivalentes ou meilleures et ne conduisent pas à des changements trop importants des estimateurs.
Il est possible de comparer leurs performances dans des études par simulation. Dequidt, Sigler et Buisson (2012) ont par exemple comparé la méthode des quantiles avec l'algorithme CHAID et montré que les performances des deux méthodes étaient équivalentes. Dufour, Gagnon, Morin, Renaud et Särndal (2001) ont quant à eux comparé une méthode par croisement avec le même algorithme CHAID, auxquels était ajoutée une étape de calage sur marges : l'algorithme CHAID, pour un même nombre de groupes réponse homogène, conduit dans leurs simulation à des réductions du biais plus importantes. De plus, Dufour, Gagnon, Morin, Renaud et Särndal (2001) ont montré qu'il était possible de décomposer l'écart entre les poids de sondage et les poids après calage sur marges en la somme de quatre termes, représentant respectivement l'ajustement du niveau moyen des poids lié à la correction de la non-réponse, l'effet de la correction de la non-réponse d'un côté et du calage sur marges de l'autre part sur les poids individuels autour de cet ajustement moyen et enfin l'interaction des effets de la correction de la non-réponse et du calage sur marges sur les poids individuels. Dans leurs simulations, la qualité de la correction de la non-réponse, mesurée par la réduction du biais d'estimation, est corrélée positivement avec le terme mesurant l'effet de la correction de la non-réponse sur les modifications des poids individuels.

Les simulations s'appliquent cependant forcément à des situations théoriques, dans lesquelles le mécanisme générant la non-réponse est connu et peut de ce fait être répliqué. De plus, leurs conclusions sont valables pour les mécanismes de non-réponse utilisés pour les simulations, mais ne se généralisent pas forcément aux mécanismes inconnus rencontrés dans les situations concrètes. En situation réelle, face à une réalisation du mécanisme de réponse et à l'ensemble des méthodes disponibles pour constituer des groupes de réponse homogène, le praticien est ainsi confronté à deux questions : sur la base de quels critères choisir une méthode pour former des GRH ? Quelle incidence a le choix d'une méthode plutôt que d'une autre sur les estimateurs qui peuvent être calculés à partir des poids corrigés de la non-réponse ou des poids finaux après traitement des unités influentes et calage sur marges ?

Pour tenter de répondre à ces deux questions, nous avons testé de nombreuses méthodes de constitution de groupes de réponse homogène (méthode usuelle par croisement, algorithme CHAID, algorithme CART, méthodes des scores - quantiles, CAH, algorithme de Haziza et Beaumont - en partant des probabilités de réponse estimées issues de modèles de régression logistique ou d'algorithmes de *machine learning*) sur les données des Enquêtes

Sectorielles Annuelles portant sur l'année 2014 et collectées en 2015. Ces méthodes ont été appliquées à un sous-échantillon d'apprentissage sélectionné aléatoirement et représentant 2/3 de l'échantillon initial, et les GRH ont ensuite été appliqués au sous-échantillon de test restant. Nous avons ainsi pu mesurer leur capacité à rendre compte du comportement de réponse observé à l'aide d'indicateurs comme l'estimation des erreurs de classement, de la matrice de confusion,... Elles ont également été appliquées à l'échantillon complet, sur lequel nous avons pu calculer l'écart entre les totaux de variables de la base de sondage et leurs estimateurs à partir des poids corrigés de la non-réponse. Enfin, en appliquant un calage sur marges identique aux différents poids corrigés de la non-réponse, nous avons pu calculer l'indicateur de qualité de la correction de la non-réponse proposé par Dufour, Gagnon, Morin, Renaud et Särndal (2001). Nous avons également calculé une série d'estimateurs de grandeurs d'intérêt de l'enquête sur des domaines de diffusion plus ou moins fins pour chaque jeu de poids finaux construits. Ces comparaisons nous fournissent ainsi une analyse croisée de critères de qualité de la correction de la non-réponse et de l'impact des différentes méthodes sur les niveaux des estimations.

# Bibliographie

[1] Bethlehem, J. (1988), Reduction of nonreponse biais through regression estimation, Journal of Official Statistics.
[2] Breiman, L., Friedman, J. et Ohlsen, R. (1984), Classification and Regression Trees, CRC Press
[3] Breiman, L. (1994), Bagging Predictors, Technical Report of the Department of Statistics, University of California, Berkeley.
[4] Breiman, L. (2001), Random Forest, Machine Learning.
[5] Brick, J. (2013), Unit non-response and weighting adjustment - a critical review, Journal of Official Statistics.
[6] Bühlmann P. (2012), Bagging, boosting and ensemble methods, *in* Handbook of Computational Statistics : Concepts and Methods, Springer Verlag.
[7] Dequidt, E., Sigler, N. et Buisson, B. (2012), Comparaison de méthodes pour la correction de la non-réponse totale : méthode des scores et segmentation, Actes du Septième Colloque Francophone sur les Sondages, Rennes.
[8] Dufour, J., Gagnon, F., Morin, Y., Renaud, M. et Särndal C.E. (2001), Mieux comprendre la transformation des poids à l'aide d'une mesure du changement, Techniques d'Enquête.
[9] Freund, Y., Shapire, R. (1997), A decision-theoretic approach of on-line learning and an application to boosting, Journal of Computer and System Science.
[10] Hastie, T., Tibshirani, R., Friedman, J. (2009), The Elements of Statistical Learning, Springer Verlag.
[11] Haziza, D., Beaumont, J.-F. (2007), On the construction of imputation classes in

surveys, International Statistical Review.
[12] Haziza, D., Lesage E. (2016), A discussion of weighting procedures for unit non-response, Journal of Official Statistics.
[13] Kass, G. (1980), An exploratory technique for investigating large quantities of categorical data, Journal of Applied Statistics.
[14] Kim, J.K., Kim J. (2007), Non-response weighting adjustment using estimated response probability, The Canadian Journal of Statistics.
[15] Little, R. et Vartivarian, S. (2005), La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ?, Techniques d'Enquête.
[16] Phipps, P. et Toth, D. (2012), Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data, The Annals of Applied Statistics.