Enquête santé européenne (EHIS) 2019 : Calcul des pondérations

France métropolitaine

Thomas Deroyon

INTRODUCTION

L'échantillon S_0 de l'EHIS est composé de 27 600 individus sélectionnés suivant un plan de sondage à deux phases et auxquels est associé un jeu de pondérations w_k rendant cet échantillon représentatif de la base de sondage : l'estimateur $\sum_{k \in S_0} w_k y_k$ du total de la variable y_k sur la population est sans biais sous le plan de sondage, i.e. sa moyenne sur l'ensemble des échantillons possibles est bien égale au vrai total de y. Le plan de sondage et le calcul des pondérations w_k sont détaillés dans la section « Plan de sondage » du DREES Méthodes décrivant la méthodologie de l'EHIS 2019.

Ce système de pondération n'est cependant pas utilisable car une partie de l'échantillon initial n'a pas répondu à l'enquête : les estimateurs calculés sur les seuls répondants avec les pondérations w_k ne sont plus sans biais sous le plan de sondage. De plus, les poids de sondage n'intègrent pas toute l'information auxiliaire disponible au moment des redressements de l'enquête, notamment les informations sur la population au moment de la collecte que peuvent apporter des sources annexes comme l'enquête Emploi. Le calcul des pondérations détaillé dans cette note a pour but de supprimer ou limiter les biais introduits par la non réponse et d'améliorer la précision des pondérations de l'échantillon[1] en mobilisant au mieux l'information auxiliaire disponible via un calage sur marges.

[1] Plus précisément, d'améliorer la précision des estimateurs qui peuvent être calculés avec les pondérations et les réponses à l'enquête.

PRINCIPE GÉNÉRAL DU CALCUL DES PONDÉRATIONS

Comme précisé dans le DREES Méthodes décrivant la méthodologie de l'EHIS 2019, l'enquête santé métropole a mobilisé une collecte par téléphone (mode CATI) et une collecte en face-à-face (CAPI) selon le protocole suivant :

- les individus pour lesquels aucun numéro de téléphone n'était disponible dans la base de sondage et n'a été retrouvé par le prestataire de collecte étaient orientés directement vers la collecte en face-à-face ;
- les autres personnes échantillonnées, pour lesquelles au moins un numéro de téléphone était disponible, étaient d'abord contactées par téléphone. Au terme de cette collecte téléphonique, l'échantillon se partageait en trois groupes :
  - les personnes contactées par téléphone et ayant répondu à l'enquête, y compris celles dont l'enquêteur a pu déterminer qu'elles étaient hors du champ de la collecte ;
  - les personnes avec lesquelles les enquêteurs avaient réussi à avoir un contact au téléphone, mais qui n'avaient pas répondu à l'enquête ;
  - les personnes avec lesquelles les enquêteurs n'avaient pas réussi à avoir des contacts au téléphone, appelés par la suite « impossibles à joindre » (IAJ).

Pour les deux premiers groupes, la collecte s'achevait une fois tous les appels prévus pour chaque individu passés. Une moitié, sélectionnée aléatoirement, des IAJ était transmise aux enquêteurs chargés de la collecte en face-à-face. Par ailleurs, une petite partie, sélectionnée aléatoirement, des personnes pour lesquelles des numéros de téléphone étaient disponibles était orientée directement vers la collecte en face à face, de façon à constituer un échantillon méthodologique permettant d'identifier les éventuels effets de mode.

Le calcul des pondérations tient compte de ce protocole de collecte.

La correction de la non réponse pourrait en effet se faire en une seule étape : il s'agirait de modéliser, à partir des informations disponibles dans la base de sondage, la probabilité d'être au final répondant, quelle que soit l'étape du protocole de collecte au cours de laquelle ou le mode avec lequel la réponse a été obtenue.

Cette approche ne tient cependant pas compte de la complexité du protocole de collecte, qui combine collectes au téléphone ou en face à face appliquées à des sous-populations aux caractéristiques spécifiques (personnes n'ayant aucun numéro de téléphone dans les sources fiscales ou impossibles à joindre au téléphone pour la collecte en face à face). Aussi, il a paru préférable de détailler davantage la construction des poids corrigés de la non réponse de façon à mieux tenir compte des différentes étapes par lesquelles l'échantillon est collecté.

Plusieurs solutions étaient envisageables pour tenir compte dans le calcul des probabilités de réponse des différentes étapes de sélection induites par le protocole de collecte. Pour présenter l'approche qui a été retenue, supposons que nous cherchions à estimer le total d'une variable y sur la population. S est l'échantillon des répondants à l'enquête, quels que soient les modes de collecte par lesquels ont été obtenues les réponses. Nous cherchons donc à estimer $\sum_{k \in U} y_k$, avec U le champ de l'enquête, i.e. les personnes de 15 ans ou plus résidant au moment de la collecte en France métropolitaine dans un logement ordinaire à titre de résidence principale.

Si l'ensemble des individus échantillonnées avaient répondu, on disposerait directement d'un estimateur sans biais : $\sum_{k \in S_0} w_k y_k$ [2]. La non réponse rend cet estimateur inutilisable, mais on va chercher, en s'appuyant sur les répondants à l'enquête, à reconstituer cet estimateur, i.e. à estimer sa valeur ; l'estimateur obtenu devant être à son tour sans biais sous le plan de sondage. Pour ce faire, on suppose que la non réponse, quelle que soit l'étape du protocole de collecte à laquelle elle se produit, est un phénomène aléatoire, dont les probabilités sont inconnues mais toujours strictement positives [3].

---

[2] On néglige ici le défaut de couverture induit par la limitation de la base de sondage aux résidences principales et le biais qu'il cause. La section Calage sur marges revient sur ce point.

[3] Cela revient à supposer que toutes les personnes échantillonnées avaient une chance de répondre à l'enquête.

S_0 peut se découper en deux sous-échantillons : S^0_{CATI} désigne les individus intégrés à la collecte téléphonique et S^0_{CAPI} à l'inverse les personnes directement collectées en face à face. Ainsi,

$$\sum_{k \in S_0} w_k \, y_k = \sum_{k \in S^0_{CATI}} w_k \, y_k + \sum_{k \in S^0_{CAPI}} w_k \, y_k$$

Rappelons que si on s'intéresse à l'estimation du total d'une variable z sur une population P, et qu'on dispose pour ce faire d'un échantillon Σ tiré dans cette population suivant un plan de sondage associant des probabilités d'inclusion p_i à chaque membre de la population, alors $\sum_{i \in \Sigma} \frac{z_i}{p_i}$ estime sans biais $\sum_{i \in P} z_i$ [4].

Appliquons ce principe à $\sum_{k \in S^0_{CAPI}} w_k \, y_k$. Cette somme peut en effet être vue comme le total de la variable $w_k \, y_k$ sur la population S^0_{CAPI}.

S^0_{CAPI} est collecté en une seule étape, en face à face. Si S^R_{CAPI} désigne le sous-échantillon des répondants de S^0_{CAPI} et $\rho^{(1)}_k$ leurs probabilités de réponse, alors S^R_{CAPI} est un échantillon tiré dans S^0_{CAPI} suivant un plan de sondage qui associe à chaque répondant une probabilité d'inclusion égale à $\rho^{(1)}_k$, ce plan de sondage correspondant au comportement de réponse supposé aléatoire. Dès lors, $\sum_{k \in S^0_{CAPI}} w_k \, y_k$ peut être estimé sans biais par $\sum_{k \in S^R_{CAPI}} w_k / \rho^{(1)}_k \, y_k$.

Les probabilités de réponse $\rho^{(1)}_k$ à la collecte en face à face étant inconnues, elles doivent être estimées, ce qui est l'objectif de la correction de la non réponse à la collecte CAPI directe (sans collecte CATI et bascule préalable). Si ces probabilités de réponse sont estimées correctement, i.e. via des estimateurs asymptotiquement sans biais[5], alors $\sum_{k \in S^R_{CAPI}} w_k / \hat{\rho}^{(1)}_k \, y_k$ estime également le total de la variable d'intérêt asymptotiquement[6] sans biais sous le plan de sondage[7].

Pour estimer $\sum_{k \in S^0_{CATI}} w_k \, y_k$, on dispose des répondants directs à la collecte téléphone, S^R_{CATI}, mais aussi de la partie des non répondants au téléphone qui ont été basculés en face en face et ont alors répondu à l'enquête.

Plus précisément, au terme de la collecte téléphonique, S^0_{CATI} peut être divisé en quatre parties :

- S^R_{CATI}, i.e. l'ensemble des répondants par téléphone à l'enquête. ;
- S^{HC}_{CATI}, les personnes dont les enquêteurs ont réussi à déterminer qu'elles n'appartenaient pas au champ de l'enquête ;
- S^{NR}_{CATI}, les personnes que les enquêteurs ont réussi à contacter, dont ils ont pu déterminer qu'ils appartenaient au champ de l'enquête mais dont ils n'ont pas obtenu les réponses ;
- enfin, S^{NC}_{CATI}, les personnes que les enquêteurs n'ont pas réussi à contacter.

La moitié de ces dernières, sélectionnée aléatoirement, a été basculée vers la collecte en face à face.

Supposons que l'on dispose des réponses des individus de S^R_{CATI} et S^{NC}_{CATI} et que la probabilité $\rho^{(2)}_k$ avec laquelle chaque individu de S^0_{CATI} appartient à l'un de ces deux sous-échantillons soit connue. Alors on peut constituer $\sum_{k \in S^R_{CATI}} \left(w_k / \rho^{(2)}_k \right) y_k + \sum_{k \in S^{NC}_{CATI}} \left(w_k / \rho^{(2)}_k \right) y_k$, qui est un estimateur sans biais de $\sum_{k \in S^0_{CATI}} w_k \, y_k$.

Après estimation de $\rho^{(2)}_k$, $\sum_{k \in S^R_{CATI}} \left(w_k / \hat{\rho}^{(2)}_k \right) y_k$ peut être calculé, mais pas $\sum_{k \in S^{NC}_{CATI}} \left(w_k / \hat{\rho}^{(2)}_k \right) y_k$, car les réponses à l'enquête ne sont pas disponibles pour tous les membres de cet échantillon. Cependant, la moitié de S^{NC}_{CATI} est basculée sur la collecte en face à face. Si l'on note $S^R_{CATI,CAPI}$ l'échantillon des répondants au face à face

---

[4] pour que cette propriété soit vraie, il faut que les probabilités d'inclusion p_i soient toutes strictement positives, i.e. que chaque individu de la population ait une chance d'être échantillonné.

[5] i.e. avec un biais à distance finie qui décroît avec la taille de l'échantillon.

[6] asymptotiquement sans biais, i.e. dont le biais à distance finie décroît avec la taille de la population et de l'échantillon, suivant le cadre asymptotique décrit par Isaki et Fuller dans leur article de référence (voir C. Isaki, W. Fuller, *Survey design under the regression superpopulation model*, Journal of the American Statistical Association, 1982)

[7] voir J.K. Kim, J. J. Kim, *Nonreponse weighting adjustment using estimated response probability*, The Canadian Journal of Statistics, 2007.

issus de S_{CATI}^{NC}, et $\rho_k^{(3)}$ leurs probabilités de réponse à la collecte en face à face, alors $\sum_{k \in S_{CATI,CAPI}^{R}} \frac{2\,w_k}{\rho^{(2)_k}\,\rho^{(3)_k}}\,y_k$ estime sans biais $\sum_{k \in S_{CATI}^{NC}} \frac{w_k}{\rho^{(2)_k}}\,y_k$ [8].

Au final, on peut constituer $\sum_{k \in S_{CATI}^{R}} \frac{w_k}{\rho_k^{(2)}}\,y_k + \sum_{k \in S_{CATI,CAPI}^{R}} \frac{2\,w_k}{\rho^{(2)_k}\,\rho^{(3)_k}}\,y_k$ qui estime sans biais $\sum_{k \in S_{CATI}^{0}} w_k\,y_k$.

Les probabilités $\rho_k^{(2)}$ et $\rho^{(3)}$ sont inconnues mais peuvent être estimées, c'est à nouveau l'objet des travaux de correction de la non réponse qui seront décrits ultérieurement.

Une fois mis en oeuvre les travaux de correction de la non réponse qui permettent d'estimer $\rho_k^{(1)}$, $\rho_k^{(2)}$ et $\rho_k^{(3)}$, on peut donc constituer l'estimateur :

$$\hat{Y} = \sum_{k \in S_{CAPI}^{R}} \frac{w_k}{\hat{\rho}_k^{(1)}}\,y_k + \sum_{k \in S_{CATI}^{R}} \frac{w_k}{\hat{\rho}_k^{(2)}}\,y_k + \sum_{k \in S_{CATI,CAPI}^{R}} \frac{2\,w_k}{\hat{\rho}_k^{(2)}\,\hat{\rho}_k^{(3)}}\,y_k$$

qui estime sans biais le vrai total de y dans la population en mobilisant l'ensemble des réponses collectées, si les probabilités de réponse sont correctement estimées.

Ainsi, les pondérations corrigées de la non réponse des répondants à l'enquête sont égales à :

- $\frac{w_k}{\hat{\rho}_k^{(1)}}$ pour les individus mis en collecte directement en face à face et ayant répondu à l'enquête ;
- $\frac{w_k}{\hat{\rho}_k^{(2)}}$ pour les personnes mises en collecte et répondantes par téléphone ;
- $\frac{2\,w_k}{\hat{\rho}_k^{(2)}\,\hat{\rho}_k^{(3)}}$ pour les personnes mises en collecte au téléphone, impossibles à joindre, basculées vers la collecte en face à face et finalement répondantes.

Le système de pondération ainsi constitué permet à nouveau de constituer des estimateurs sans biais sous le plan de sondage, sous l'hypothèse que les techniques utilisées pour estimer les différentes probabilités $\rho_k^{(1)}$, $\rho_k^{(2)}$ et $\rho_k^{(3)}$ permettent d'obtenir des estimations convergentes. La dernière étape du calcul des poids consiste à modifier ces pondérations corrigées de la non réponse par un calage sur marges qui permet de rendre les poids finaux de l'échantillon de répondants à l'EHIS 2019 cohérents avec certaines caractéristiques de la population connues par d'autre source.

Les sections suivantes détaillent les différentes étapes de calcul de ces pondérations.

[8] Le facteur 2 dans la pondération $\frac{2\,w_k}{\rho^{(2)_k}\,\rho^{(3)_k}}$ tient au fait que seule la moitié, sélectionnée aléatoirement, des membres de S_{CATI}^{NC} est basculée en face à face.

CORRECTION DE LA NON RÉPONSE

Méthodes utilisées

Trois probabilités différentes doivent être estimées, les probabilités de répondre à l'enquête en face à face, pour les personnes directement orientées vers la collecte en face à face d'une part et pour les personnes impossibles à joindre par téléphone d'autre part ; et les probabilités de répondre ou d'être impossible à joindre pour les personnes orientées vers la collecte téléphonique.

Pour estimer ces trois jeux de probabilité, les enjeux sont les mêmes et les méthodes retenues identiques : il s'agit d'obtenir des estimations des probabilités de réponse les plus proches possibles des probabilités réelles sous-jacentes, et pour ce faire, on mobilise différents modèles et algorithmes dont on compare l'efficacité avec les méthodes détaillées dans la section suivante (cf. section « Critères de choix des probabilités de réponse »).

Ces modèles et algorithmes sont les suivants :

- des modèles de régression logistique, avec ou sans sélection de variable. La sélection de variable est réalisée à l'aide du critère d'information d'Akaike, suivant une approche de type *stepwise* ;
- des arbres de régression construits avec l'algorithme des *classification and regression trees*[9] (CART) proposé par L. Breiman ;
- des forêts aléatoires[10].

Les modèles de régression logistique produisent en sortie une estimation directe des probabilités de réponse de chaque individu.

L'algorithme CART forme des arbres de régression binaires : ces arbres sont construits de manière itérative ; à chaque étape, l'algorithme forme des groupes en partitionnant en deux les groupes constitués à l'étape précédente. Les partitions sont constituées en choisissant la variable auxiliaire et le découpage en deux ensembles de ses modalités qui conduit à constituer les sous-groupes les plus homogènes. Dans le problème qui nous intéresse, l'algorithme choisira les partitions qui permettent de former les deux sous-groupes dans lesquels les différences de probabilités de réponse observées sont les plus importantes. L'algorithme CART conduit ainsi à former des groupes d'individus auxquels est associée ensuite la probabilité de réponse moyenne observée dans leur groupe[11].

Les forêts aléatoires sont une généralisation de l'algorithme CART. L'algorithme commence en tirant dans l'échantillon de départ un échantillon *bootstrap* avec remise. Un arbre de régression est ensuite construit sur cet échantillon pour prédire la probabilité de réponse avec un algorithme proche de CART ; la seule différence est qu'à chaque étape de découpage, l'algorithme ne teste pas les partitions formées par toutes les variables auxiliaires utilisées par la forêt aléatoire, mais un sous-échantillon aléatoire de celles-ci. Cet algorithme CART conduit à affecter chaque individu de l'échantillon à un groupe, et à lui associer la probabilité de réponse observée dans ce groupe. Ce processus est répété un grand nombre de fois, et la forêt aléatoire associe au final à chaque individu la moyenne des probabilités estimées par chacun des arbres formés sur chacun des échantillons *bootstrap*.

Les forêts aléatoires ont été introduites pour pallier certaines des limites de l'algorithme CART et des arbres de classification et de régression, notamment leur grande variance de prédiction[12] : les arbres constitués par ces

---

[9] voir L. Breiman et al., *Classification and Regression Trees*, CRC Press, 1999.

[10] voir L. Breiman, *Random Forests*, Machine Learning, 2001.

[11] Cette probabilité de réponse estimée est égale à la moyenne des indicatrices de réponse pondérée par les probabilités en entrée de la correction de la non réponse :

$$\rho_i = \frac{\sum_{j \in G} d_i R_i}{\sum_{j \in G} d_i}$$

où G désigne le groupe formé par l'algorithme CART et d_i le poids en entrée de la correction de la non-réponse.

[12] voir P. Bühlmann, B. Yu, *Analysing bagging*, The Annals of Statistics, 2002.

algorithmes peuvent varier fortement si les données en entrée sont très légèrement perturbées, si bien que les valeurs prédites à l'aide de ces arbres, si elles ne présentent pas de biais, peuvent varier beaucoup entre des échantillons distincts dont les données ont été générées par le même phénomène. Les forêts aléatoires permettent d'atténuer cette forte variance de prédiction en calculant les valeurs prédites comme des moyennes sur un grand nombre d'arbres de régression modélisant le même phénomène mais de manière indépendante.

Deux des méthodes utilisées conduisent ainsi à l'estimation directe d'une probabilité de réponse (modèles de régression logistique et forêts aléatoires) tandis que les arbres de régression conduisent à former une partition de l'échantillon, les probabilités de réponse étant ensuite estimées par les probabilités de réponse moyennes observées dans chaque groupe. Pour les modèles de régression logistique et les forêts aléatoires, nous testons, en plus de l'utilisation directe des probabilités de réponse estimées par les algorithmes, la constitution de groupes de réponse homogène[13] (GRH) à partir de ces probabilités. L'idée est de former une partition des échantillons en regroupant les individus ayant des probabilités de réponse prédites par les forêts aléatoires ou les modèles de régression logistique proches. La méthode des groupes de réponse homogène présente en effet de bonnes propriétés de robustesse qui font qu'elle est très souvent employée pour les redressements des enquêtes de la statistique publique. En effet, même si les probabilités de réponse estimées grâce aux groupes de réponse homogène ne sont pas identiques ou proches des probabilités de réponse réelles des membres de l'échantillon, une correction de la non réponse par GRH peut quand même annuler le biais de non réponse : il faut pour cela que les GRH soient constitués de telle manière qu'à l'intérieur d'eux, il n'y ait pas de corrélation entre les variables d'intérêt de l'enquête et les probabilités de réponse réelles[14]. Dans ce cas, il n'est pas nécessaire de tenir compte des différences de probabilités de réponse, puisqu'elles ne sont pas en lien avec les caractéristiques des individus, et les estimateurs que l'on obtient en supposant que tout le monde a la même probabilité de réponse au sein des GRH sont non biaisés.

Pour constituer les groupes de réponse homogène, on utilise l'algorithme proposé par D. Haziza et J.-F. Beaumont[15]. Les GRH ne doivent pas non plus être trop petits, pour que les probabilités de réponse soient estimées de manière robuste dans chacun d'entre eux. En pratique, on impose une taille minimale de 100 pour chaque GRH. L'algorithme de Haziza et Beaumont est adapté pour tenir compte de ce deuxième critère d'arrêt : l'algorithme est itératif et commence par constituer 2, puis 3 GRH et s'arrête dès que le nombre de GRH est suffisant pour rendre compte de la variabilité des probabilités de réponse issues du modèle de régression logistique ou des forêts aléatoires. Nous introduisons une deuxième condition à cet algorithme : dès que la taille minimale des GRH constitués est inférieure à 100, l'algorithme cesse et les GRH finaux sont ceux constitués à l'étape précédente. Si l'un des deux premiers GRH est de taille inférieure à 100 et que l'algorithme bute sur ce critère d'arrêt dès sa première itération, alors il renvoie les deux GRH qu'il vient de constituer.

Tous ces modèles sont également comparés à un modèle de base, dans lequel on fait l'hypothèse que tous les membres de l'échantillon ont la même probabilité de réponse. Au final, 8 modèles sont donc comparés :

- modèle de base ;
- modèle de régression logistique complet, i.e. avec l'ensemble des variables auxiliaires ;
- modèle de régression logistique complet et groupes de réponse homogène ;
- modèle de régression logistique avec sélection des variables auxiliaires suivant une procédure *stepwise* reposant sur l'AIC ;
- modèle de régression logistique avec sélection des variables auxiliaires et groupes de réponse homogène ;
- arbres de régression construits avec l'algorithme CART ;
- forêts aléatoires ;

---

[13] voir par exemple J.Eltinge, I. Yansaneh, *Diagnostics for the formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey*, Survey Methodology, 1997 et D. Haziza, J.F. Beaumont, *Construction of weights in surveys: a review* , Statistical Science, 2017. Voir également la fiche méthodologique du département des méthodes statistiques de l'Insee consacrée à la correction de la non réponse par repondération sur le site de l'institut.

[14] pour plus de détails, voir par exemple J. Bethlehem, *Reduction on nonresponse bias through regression estimation*, Journal of Official Statistics, 1988.

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

- forêts aléatoires avec groupes de réponse homogène.

Critère de choix des probabilités de réponse

Plusieurs modèles et algorithmes sont testés et comparés pour estimer les probabilités de réponse. Pour comparer ces modèles, on découpe l'échantillon en un échantillon d'apprentissage (représentant deux tiers de l'échantillon) et un échantillon de test (représentant le tiers restant). Ce découpage est effectué aléatoirement. Les modèles des algorithmes sont estimés sur l'échantillon d'apprentissage puis appliqués à l'échantillon test (i.e. à des données qui n'ont pas servi à estimer leurs paramètres). De même, les GRH sont constitués sur l'échantillon d'apprentissage et appliqués tels quels, avec les probabilités de réponse qui leur sont associées dans l'échantillon d'apprentissage, à l'échantillon test. La comparaison entre les probabilités prévues par les modèles sur l'échantillon test et les indicatrices de réponse permet de calculer des indicateurs de qualité des modèles de non réponse. Parmi les indicateurs fréquemment utilisés, on s'est appuyé sur :

* les erreurs quadratiques et absolues moyennes :

Il s'agit de la moyenne des écarts quadratiques ou absolus entre la probabilité de réponse prédite et l'indicatrice de réponse. Cet indicateur mesure la capacité qu'à un algorithme d'attribuer en moyenne des probabilités de réponse plus élevées aux répondants et plus faibles aux non répondants, l'erreur quadratique moyenne sanctionnant plus fortement le fait d'attribuer une probabilité de réponse très faible à un répondant, ou très élevée à un non répondant ;

$$EAM = \frac{\sum_{i \in S_{test}} |R_i - \hat{\rho}_i|}{|S_{test}|}$$

$$EQM = \frac{\sum_{i \in S_{test}} (R_i - \hat{\rho}_i)^2}{|S_{test}|}$$

* la sensibilité, ou rappel :

Dans un problème de classification classique, où l'on essaie de prédire si un individu a (positif) ou pas une maladie par exemple, la sensibilité représente la part d'individus que l'algorithme classe comme affectés par la maladie parmi l'ensemble des individus positifs à la maladie. Elle mesure donc la capacité de l'algorithme à détecter les individus positifs. Dans notre cas, où on ne cherche pas à prédire si les individus sont répondants mais à déterminer quelle est la probabilité sous-jacente avec laquelle chacun répond à l'enquête, l'indicateur est adapté : il est égal à la probabilité moyenne de réponse pour les répondants.

$$Sensibilite = \frac{\sum_{i \in S_{test}} R_i \hat{\rho}_i}{\sum_{i \in S_{test}} R_i}$$

* la précision ou valeur prédite positive :

La précision est un indicateur complémentaire de la sensibilité. Dans un problème de classification classique, elle indique la part d'individus effectivement positifs parmi toutes les personnes classées comme positives. Adapté au traitement de la non réponse, l'indicateur est calculé comme la somme des probabilités de réponse prédites des répondants divisée par la somme des probabilités de réponse prédites. Elle mesure donc la part des probabilités de réponse prédites par le modèle effectivement attribuées à des individus répondants.

$$Precision = \frac{\sum_{i \in S_{test}} R_i \hat{\rho}_i}{\sum_{i \in S_{test}} \hat{\rho}_i}$$

* la F-statistique :

Dans un problème usuel de classification, il faut trouver un équilibre entre optimisation de la sensibilité et de la précision. Une solution en effet pour optimiser la sensibilité serait d'attribuer une probabilité de réponse de 1 à tout l'échantillon. Dans ce cas, la sensibilité est égale à 1, sa valeur maximale, puisque tous les membres de l'échantillon, y compris les répondants, sont considérés par l'algorithme comme des répondants. Cela se fait cependant au détriment de la précision, car alors l'algorithme attribue une part importante des probabilités de réponse à des non répondants. Ainsi, optimiser la sensibilité peut se faire en créant plus de faux négatifs (individus négatifs prédits comme positifs), i.e. dans notre cas en attribuant des probabilités de réponse trop élevées à des non répondants, tandis qu'optimiser la précision peut se faire en augmentant le nombre de faux positifs, i.e. en affectant des probabilités de réponse faibles à des répondants, dès lors que le profil de ces derniers ne correspond pas à celui des personnes qui répondent systématiquement.

La F-statistique a pour but de résumer ces deux indicateurs en un seul de manière à éviter que l'optimisation de l'un se fasse au détriment de l'autre. Elle est égale à la moyenne harmonique de la précision et du rappel.

$$F = 2\frac{Sensibilite \times Precision}{Sensibilite + Precision}$$

L'opération de découpage entre apprentissage et test est répétée plusieurs fois et les moyennes des indicateurs calculées sur ces itérations. Le modèle présentant les meilleures performances sur les différents indicateurs est retenu.

Variables auxiliaires

Les variables auxiliaires utilisables pour la correction de la non réponse sont des variables nécessairement disponibles à la fois pour les répondants et les non répondants. Ces variables viennent principalement de la base de sondage, i.e. de Fideli 2018, et sont complétées par des variables associées à la localisation géographique du logement de la personne échantillonnée dans la base de sondage.

Comme précisé plus haut, les variables auxiliaires mobilisées dans la correction de la non réponse doivent permettre de lever la corrélation existant entre les variables d'intérêt de l'enquête et les probabilités de réponse à l'enquête. Pour ce faire, elles doivent être corrélées à la fois aux variables d'intérêt de l'enquête et à l'indicatrice de réponse, i.e. à la variable dont on cherche à estimer la probabilité d'occurence.

On a sélectionné parmi l'ensemble des variables auxiliaires possibles celles effectivement intégrées dans les modèles de non réponse de la manière suivante : on retient une variable candidate si elle est corrélée, sur la base d'un test de corrélation du χ^2 au seuil de 15 %, à au moins l'une des trois variables indicatrices qui font l'objet de la correction de la non réponse d'une part[16], et d'autre part à l'indicatrice de déclaration de limitations fortes ou modérées depuis au moins six mois pour des raisons de santé dans les activités de la vie quotidienne.

Dès qu'une variable auxiliaire vérifie ces deux conditions, elle participe aux trois étapes de correction de la non réponse, même si elle n'était pas corrélée à l'indicatrice modélisée à l'une de ces étapes[17]. Le test de corrélation du χ^2 est en effet un instrument imparfait pour détecter les variables auxiliaires potentiellement intéressantes, car il ne tient pas du tout compte des interactions entre variables auxiliaires, et de la possibilité qu'une variable auxiliaire globalement peu corrélée avec une indicatrice de réponse soit fortement corrélée avec celle-ci sur une partie de l'échantillon et donc intéressante à prendre en compte dans la modélisation de la non réponse. Les tests du χ^2 sont utilisés comme des instruments grossiers de filtrage des variables manifestement inutiles à prendre en compte dans les redressements, d'où également le choix d'un seuil de test assez élevé (15 %).

Ainsi, dès qu'une variable auxiliaire semble présenter un intérêt pour la correction de la non réponse d'une des trois étapes, elle est testée sur chacune d'entre elles, à charge aux algorithmes eux-mêmes ou aux procédures de sélection de modèle utilisées de l'éliminer si elle s'avère finalement peu pertinente.

Les variables auxiliaires introduites dans les différents algorithmes testés sont les suivantes :

- sexe ;
- âge au moment de la collecte, avec les tranches suivantes : (15-19 ans, 20-29 ans, 30-39 ans, 40-49 ans, 50-59 ans, 60-69 ans, 70-79 ans, 80 ans ou plus) ;
- présence d'une personne âgée de 0 à 14 ans dans le ménage de l'individu échantillonné dans la base de sondage au 1er janvier 2018 ;
- présence d'une personne âgée de 15 à 24 ans dans le ménage de l'individu échantillonné dans la base de sondage au 1er janvier 2018 ;
- présence d'une personne âgée de 25 à 64 ans dans le ménage de l'individu échantillonné dans la base de sondage au 1er janvier 2018 ;

[16] indicatrice de réponse à la collecte directe en face à face, indicatrice de réponse ou d'absence de contact à la collecte par téléphone, indicatrice de réponse à la collecte en face à face après bascule.

[17] Par exemple, une variable participe à la modélisation de la probabilité de répondre au face à face dans la collecte directe, même si elle n'apparaissait pas comme corrélée à l'indicatrice de réponse à la collecte CAPI directe.

- présence d'une personne âgée de 65 ans ou plus dans le ménage de l'individu échantillonné dans la base de sondage au 1er janvier 2018 ;
- décile de niveau de vie du ménage de la personne interrogée en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu un salaire en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu une allocation chômage en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu un bénéfice agricole, industriel ou commercial en 2017 ;
- nombre de membres du ménage de la personne interrogée dans la base de sondage au 1er janvier 2018 ;
- indicatrice identifiant que la personne échantillonnée a perçu un minimum social en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu une allocation logement en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu une pension de retraite en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu une pension alimentaire en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu des prestations familiales en 2017 ;
- indicatrice identifiant que la personne échantillonnée a perçu des revenus de valeurs mobilières en 2017 ;
- indicatrice identifiant les personnes échantillonnées vivant dans un quartier prioritaire de la politique de ville dans la base de sondage au 1er janvier 2018 ;
- indicatrice identifiant les personnes échantillonnées dont la résidence est une maison au 1er janvier 2018 ;
- indicatrice identifiant les personnes échantillonnées locataires de leur logement au 1er janvier 2018 ;
- région du logement de la personne échantillonnée dans la base de sondage au 1er janvier 2018 ;
- indicatrice identifiant les personnes échantillonnées dont le ménage est pauvre en 2017 ;
- tranche de taille d'unité urbaine de la commune de résidence dans la base de sondage au 1er janvier 2018 ;
- taux de chômage en 2017 de la commune de la base de sondages[18] ;
- taux d'activité en 2017 de la commune de la base de sondage ;
- part des logements en 2017 vacants parmi les logements de la commune de la base de sondage ;
- part des résidences secondaires en 2017 parmi les logements de la commune de la base de sondage ;
- part des 15-64 ans dans la population en 2017 de la commune de la base de sondage ;
- part des établissements agricoles en 2015 parmi l'ensemble des établissements de la commune de la base de sondage ;
- part des établissements industriels en 2015 parmi l'ensemble des établissements de la commune de la base de sondage ;
- part des établissements en 2015 du secteur de la construction parmi l'ensemble des établissements de la commune de la base de sondage ;
- part des établissements commerciaux en 2015 parmi l'ensemble des établissements de la commune de la base de sondage ;
- part des établissements industriels en 2015 parmi l'ensemble des établissements de la commune de la base de sondage ;
- nombre de salariés travaillant en 2017 dans la commune de la base de sondage ramené à la population de celle-ci ;
- nombre de salariés travaillant en 2017 dans un établissement de moins de 10 salariés de la commune de la base de sondage ramené à la population de celle-ci ;
- tranche de taille de l'aire urbaine 2015 à laquelle appartient la commune de résidence dans la base de sondage.

[18] La commune de la base de sondage est la commune de résidence de chaque individu au 1er janvier 2018 dans Fideli.

Estimation de la probabilité de répondre ou de ne pas être joint à la collecte au téléphone

Détermination des statuts de réponse

19 354 individus ont participé à la collecte au téléphone. La première étape pour réaliser l'estimation des probabilités de répondre ou de ne pas être contacté au téléphone est de caractériser le statut de chaque membre de l'échantillon après la collecte. Cette identification se fait à l'aide des codes résultats des différentes tentatives de contact réalisées par les enquêteurs et des réponses collectées auprès des personnes ayant accepté de participer à l'enquête.

Pour l'identification des individus répondants, on se base sur le fichier des personnes pour lesquelles des réponses ont été collectées par téléphone, transmis par le prestataire en charge de la collecte, duquel on retire les personnes identifiées comme hors champs (voir *infra*). Ce fichier contient des personnes ayant répondu à l'ensemble de l'enquête, mais également des répondants très partiels, dont le questionnaire est trop lacunaire pour que les informations qu'ils ont communiquées soient exploitables et qui doivent donc être considérés comme des non répondants. Pour faire le partage entre non réponse totale [19] et partielle [20], on a défini une liste de 97 variables d'intérêt centrales de l'enquête et considéré qu'une personne était répondante à l'enquête dès lors qu'elle avait répondu à au moins deux tiers de celles-ci. Ainsi, les participants à l'enquête n'ayant pas répondu à au moins 33 % des 97 variables sont considérés comme des non répondants totaux.

Les variables choisies représentent la majeure partie des modules et des sujets d'intérêt du questionnaire. Elles ont été choisies parmi les questions peu filtrées, afin de rendre plus simple l'identification des non répondants à chacune d'entre elles. Ces variables décrivent ainsi :

- l'état de santé du répondant (variables hs1, hs2, hs3, cd2, ac1a, ac1b, ac1c, pl1, pl2, pl3, pl4, pl5, pl6, pl7, pl8, pl9a, pl9b et pn1) ;
- sa littératie en santé (variables liter_1, liter_2, liter_3, liter_4 et liter_5) ;
- son recours aux soins (variables am10a, ho1a, ho2a, am1, am2, am4, am6a, am6b, am7, md1, md2, pa2_fr, pa3_fr, pa4_fr, pa5a et pa6a) ;
- son renoncement aux soins (variables un1a, un1b, un2a, un2b, un2c, un2d, un2e1 et un2e2) ;
- son indice de masse corporelle (variable bm1 et bm2) ;
- sa pratique sportive (variables pe1, pe2, pe4, pe6, pe et pe9) ;
- ses pratiques alimentaires (variables dh1, dh3, dh5 et dh6) ;
- le support social dont il peut bénéficier (variables ss1, ss2, ss3 et ic1_fr) ;
- sa santé mentale (variables cantril, mh1a, mh1b, mh1c, mh1d, mh1e, mh1f, mh1g, mh1h et mh1i) ;
- sa consommation de tabac (variables sk1_fr, sk5 et sk6) ;
- sa consommation d'alcool (variables al1_fr, al2, al3, al4, al5 et al6) ;
- ses limitations dans les activités de la vie quotidienne, mesurées via les modules de questions sur les *activities of daily life* et les *instrumental activities of daily life* [21] (variables pc1a, pc1b, pc1c, pc1d, pc1e, ha1a, ha1b, ha1c, ha1d, ha1e, ha1f et ha1g) ;
- ses caractéristiques socio-démographiques (variables de catégorie sociale, de diplôme et d'état matrimonial - il n'y a pas de non réponse partielle sur le sexe et l'âge).

Par ailleurs, pour quelques observations, les réponses ont manifestement été collectées auprès d'une autre personne que celle échantillonnée et qui aurait dû être interrogée. Ces cas sont identifiés en comparant les données disponibles dans la base de sondage et les réponses à l'enquête.

---

[19] i.e. les personnes ayant participé à l'enquête et appartenant au champ de celle-c et qui ont répondu suffisamment pour participer aux exploitations de l'enquête, avec correction des lacunes de leur questionnaire par imputation.

[20] i.e. les personnes ayant participé à l'enquête et appartenant au champ de celle-ci qui ont répondu trop peu pour que leurs réponses soient prises en compte.

[21] voir J.M. Robine et al., *Creating a coherent set of indicators to monitor health across Europe - the Euro-REVES 2 project*, European Journal of Public Health, 2003 et J.M. Robine et al., *selection of a coherent set of health indicators for the European Union*, Euro-REVES 2 project final report, 2002.

Au final et selon cette définition, 8 091 personnes sont répondantes dans la collecte par téléphone, 663 ont répondu à l'enquête mais trop peu pour que leurs réponses soient prises en compte et pour 26 individus, les réponses ont été obtenues auprès d'une autre personne que celle qui devait être interrogée.

Les codes résultats de la collecte permettent quant à eux d'identifier 1 217 personnes hors champ, du fait de décès, de départ à l'étranger ou dans un logement collectif (en maison de retraite ou en ehpad par exemple).

L'ensemble des individus qui ne sont ni répondants ni hors champ sont non répondants, mais, pour le calcul des poids corrigés de la non réponse, il est nécessaire de partager ce sous-échantillon entre personnes éligibles à la bascule en face à face et personnes non éligibles. Les personnes éligibles à la bascule vers la collecte en face à face sont celles que les enquêteurs n'ont pas pu contacter au téléphone. Parmi les situations déjà identifiées, les 663 non répondants partiels basculés en non réponse totale ne peuvent pas être éligibles à la bascule, puisqu'ils ont répondu a une partie du questionnaire de l'enquête et ont donc été contactés par un enquêteur.

Pour l'ensemble des autres situations, on a opéré la caractérisation de la manière suivante : la division Sondages avait dès le tirage de l'échantillon affecté à chaque individu une variable B_i égale à 1 avec une probabilité de 50 %, utilisée pour déterminer qui, parmi les personnes éligibles à la bascule vers la collecte en face à face, devaient effectivement être basculées ; ce qui permet de collecter de l'information sur une partie sélectionnée aléatoirement des personnes impossibles à joindre par téléphone, tout en limitant l'effort de collecte en le concentrant sur la moitié de cette sous-population.

On peut donc identifier d'une part les personnes initialement collectées par téléphone qui ont ensuite été basculées dans la collecte en face à face et d'autre part les personnes de la collecte CATI basculables (au sens où la variable B_i est égale à 1 pour eux) mais qui n'ont pas été basculées en face à face alors qu'elles n'étaient ni répondantes ni hors champ dans la collecte téléphonique. En listant les codes résultats de la collecte téléphonique présents dans la première population et absents de la seconde, on a pu partager les non répondants entre personnes ayant été contactées par un enquêteur et non éligibles à la bascule, et les personnes impossibles à joindre et éligibles à cette bascule.

Au final, 4 340 individus sont éligibles à la bascule et 5 677 sont non répondants. Le tableau 1 récapitule la manière dont se répartit suivant les différents statuts l'échantillon collecté par téléphone.

Tableau 1 : Répartition de l'échantillon collecté par téléphone selon le statut de réponse

Statut	Effectif	Part (en %)
Total	19 354	100.0
Répondant	8 091	41.8
Eligible à la bascule	4 340	22.4
Hors Champ	1 246	6.4
Non répondant	5 677	29.3
dont NRP basculé en NRT	663	3.4

Le taux de réponse[22] à la collecte téléphonique est donc de 45 %.

Estimation des probabilités de répondre ou d'être impossibles à joindre

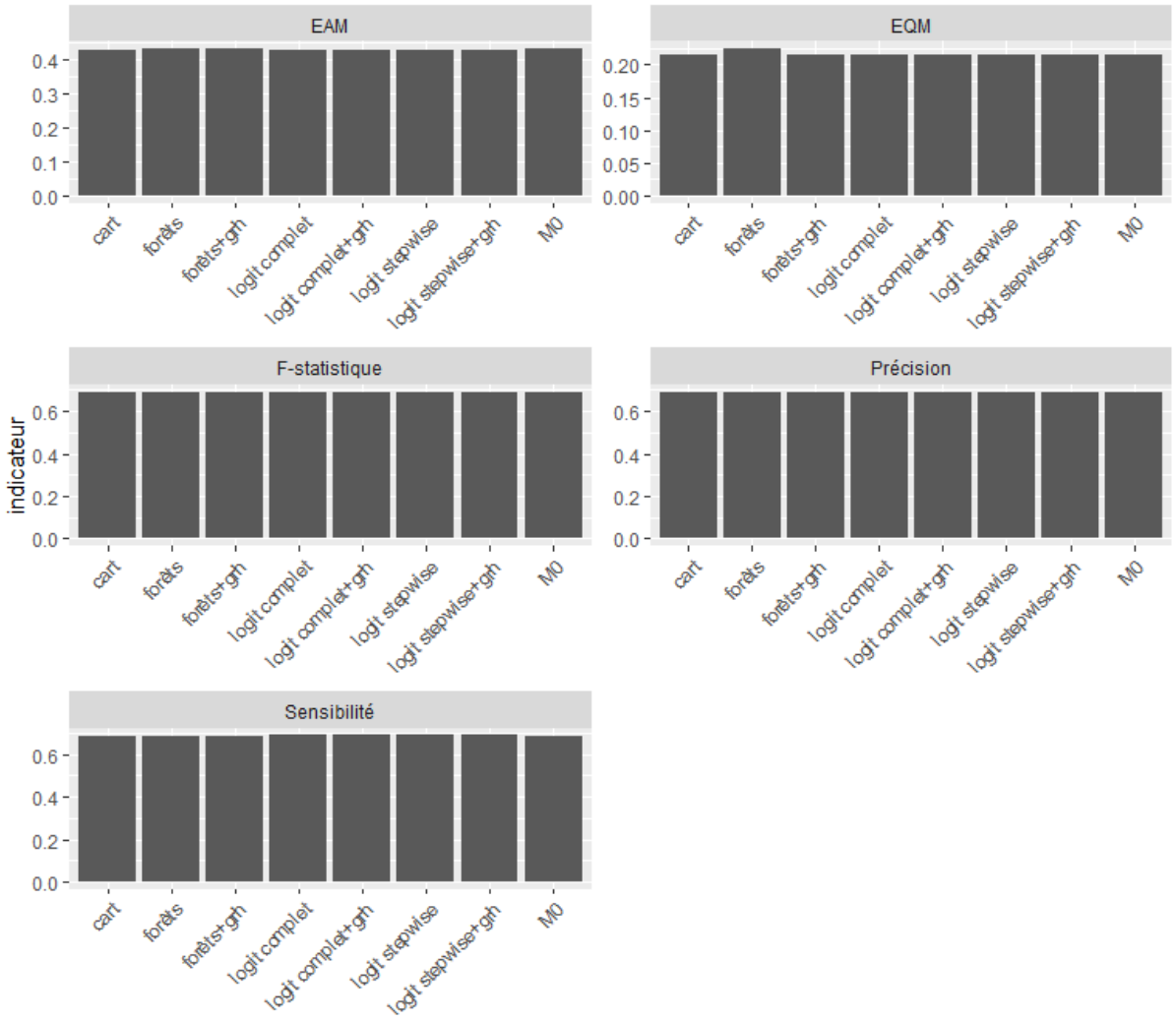
La modélisation vise à estimer la probabilité de répondre à l'enquête par téléphone ou de ne pas être contacté. Elle fait intervenir les répondants, les personnes impossibles à joindre et les non répondants. Les individus hors-champ sont mis de côté, ce qui conduit à poser l'hypothèse qu'il n'y a pas de hors-champ parmi les non répondants. Cette hypothèse paraît crédible dans la mesure où les non répondants sont tous des individus qui ont eu un échange ou dont un proche a eu un échange avec l'enquêteur, ce qui a pu permettre à ce dernier de vérifier son appartenance au champ de l'enquête.

L'estimation de la probabilité de répondre ou d'être impossible à joindre au téléphone fait donc intervenir 18 108 personnes. La variable modélisée pour estimer cette probabilité est égale à 1 pour 12 431 personnes et 0 pour 5 677.

[22] Le taux de réponse est défini ici comme le ratio du nombre de répondants sur la somme du nombre de répondants et de non répondants. Les hors champs ne sont pas pris en compte dans son calcul.

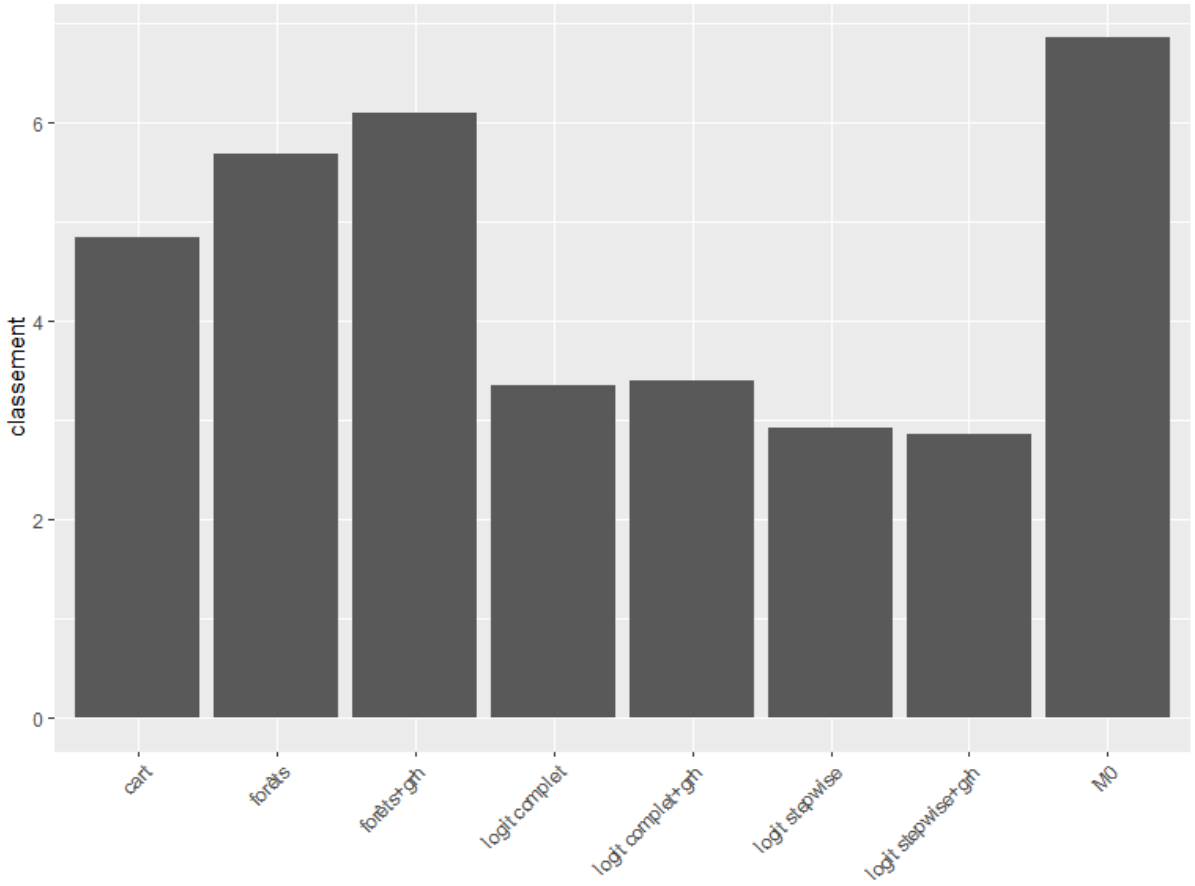
Les performances prédictives de l'ensemble des algorithmes et modèles testés sont proches de celles du modèle de base, ce qui veut dire que les variables auxiliaires mobilisées sont très peu prédictives du comportement de réponse à la collecte CATI. Le graphique 1 compare l'erreur quadratique moyenne (EQM), l'erreur absolue moyenne (EAM), la précision, la sensibilité et la F-statistique pour les différents modèles testés. Les statistiques présentées sur ces graphiques ont été estimées sur 50 itérations du processus de découpage entre échantillon d'apprentissage et échantillon de test.

Graphique 1 : Erreurs quadratique et absolue moyennes, sensibilité, précision et F-statistique des modèles appliqués aux résultats de la collecte CATI



Pour départager les différents modèles, on a classé, pour chacun des 50 découpages de l'échantillon d'origine entre apprentissage et test et pour chaque indicateur de performance, les modèles en compétition de 1 à 8, 1 désignant le modèle le plus performant pour l'indicateur et 8 le moins performant. On a ensuite calculé pour chaque découpage le classement moyen de chaque modèle sur tous les indicateurs de performance, puis la moyenne de ce classement sur les 50 découpages réalisés. Le graphique 2 présente les valeurs de cet indicateur sur les différents modèles.

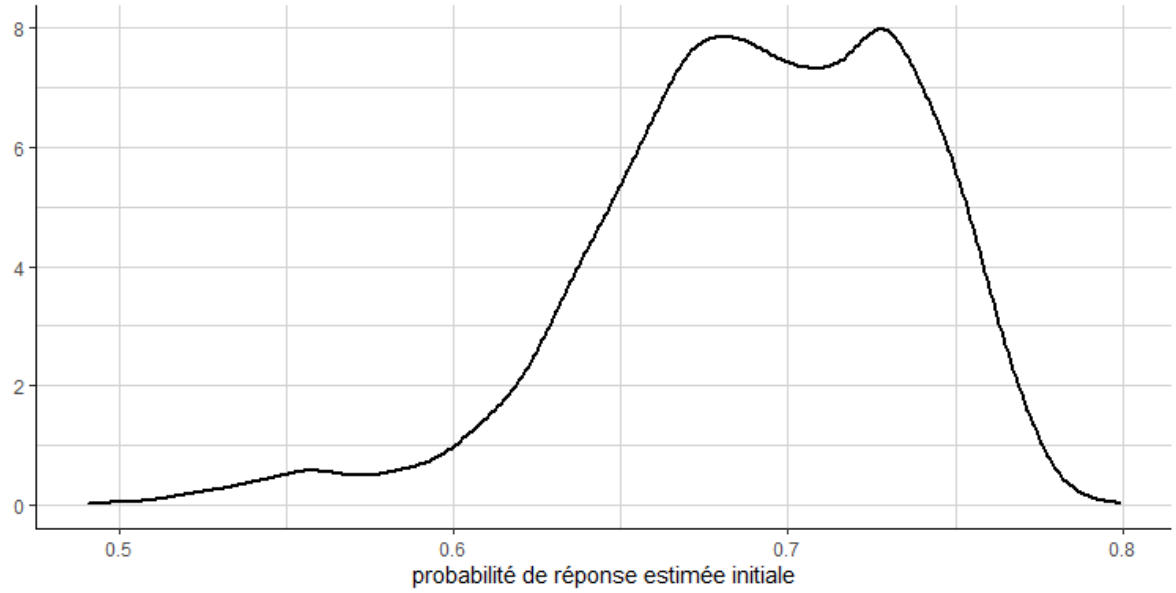
Graphique 2 : Classement moyen des différents modèles appliqués aux résultats de la collecte CATI



Les groupes de réponse homogène constitués avec les probabilités de réponse issues du modèle de régression logistique avec sélection *stepwise* ressortent comme l'algorithme le plus performant, en moyenne sur l'ensemble des indicateurs de performance et l'ensemble des 50 découpages de l'échantillon d'origine entre apprentissage et test. C'est ce modèle que l'on retient.

Le modèle de régression logistique avec sélection *stepwise*, appliqué à l'ensemble de l'échantillon conduit à des probabilités de réponse estimées dont la distribution est représentée dans le graphique 3. Compte-tenu du faible pouvoir explicatif du modèle, les probabilités de réponse estimées sont peu différenciées, même si le modèle sépare un groupe d'individus dont les probabilités de réponse sont légèrement supérieures à 50 %, alors que le reste de l'échantillon a des probabilités de réponse proches de 70 %.

Graphique 3 : Densité de la distribution des probabilités initiales de répondre ou de ne pas être contacté par téléphone estimées par le modèle de régression logistique retenu



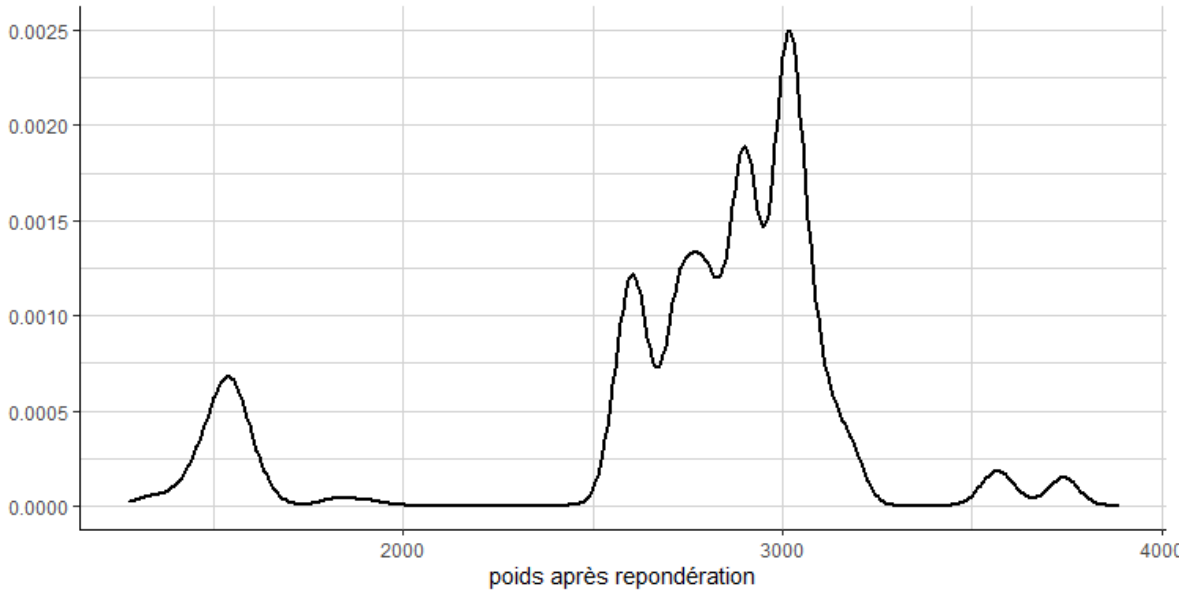
Les groupes de réponse homogène sont ensuite constitués à partir de ces probabilités en appliquant l'algorithme de Haziza et Beaumont à l'ensemble de l'échantillon. Celui-ci conduit à former 14 groupes de réponse homogène, dont les caractéristiques (effectif et probabilité de réponse moyenne associée) sont décrites dans le tableau 2.

Dans chaque GRH, les probabilités de répondre ou de ne pas être contactées au CATI sont estimées comme la part des répondants pondérée par les poids du plan de sondage. Les poids corrigés de la collecte CATI sont ensuite obtenus pour les individus répondants ou éligibles à la bascule en face à face comme le poids du plan de sondage divisé par la probabilité estimée dans le GRH. La distribution de ces poids est présentée dans le graphique 4.

Tableau 2 : Caractéristiques des groupes de réponse homogène pour l'estimation des probabilités de répondre ou de ne pas être contacté dans la collecte CATI

GRH	Effectif	Probabilité de réponse
4	418	53.71
5	177	55.91
9	354	56.48
3	751	63.30
11	1314	64.84
13	1656	66.23
2	1955	66.70
14	2001	66.71
1	1892	69.14
10	1799	69.33
12	1828	71.70
8	1785	73.62
6	1499	76.81
7	708	77.94

Graphique 4 : Distribution des poids corrigés de la collecte CATI



Estimation de la probabilité de réponse à la collecte directe en face-à-face

Détermination des statuts de réponse

La collecte directe en face à face concerne 8 241 personnes[23]. Les répondants ont été identifiés suivant le processus décrit pour la collecte par téléphone : en partant des individus ayant participé à l'enquête, présents dans les fichiers de réponse transmis par le prestataire de collecte, desquels on a retiré les personnes hors champ sur la base des codes résultats de la collecte en face à face et les personnes ayant répondu à quelques questions, mais pas assez cependant pour que leurs réponses soient exploitables.

Au final, 558 hors champ sont identifiés dans la collecte directe en face à face sur la base des codes résultats associés par les enquêteurs à leurs tentatives de contact, et 4 918 participants à cette collecte sont répondants. Les 2 765 individus restants sont tous considérés comme non répondants, i.e. dans le champ de l'enquête. Il est classique de supposer que, dans une collecte en face à face, les enquêteurs parviennent à qualifier l'appartenance au champ de l'ensemble de l'échantillon, y compris des personnes qu'ils n'arrivent pas à contacter, en s'appuyant sur les informations collectées auprès du voisinage.

Le taux de réponse à la collecte directe[24] en face à face est donc de 64 %.

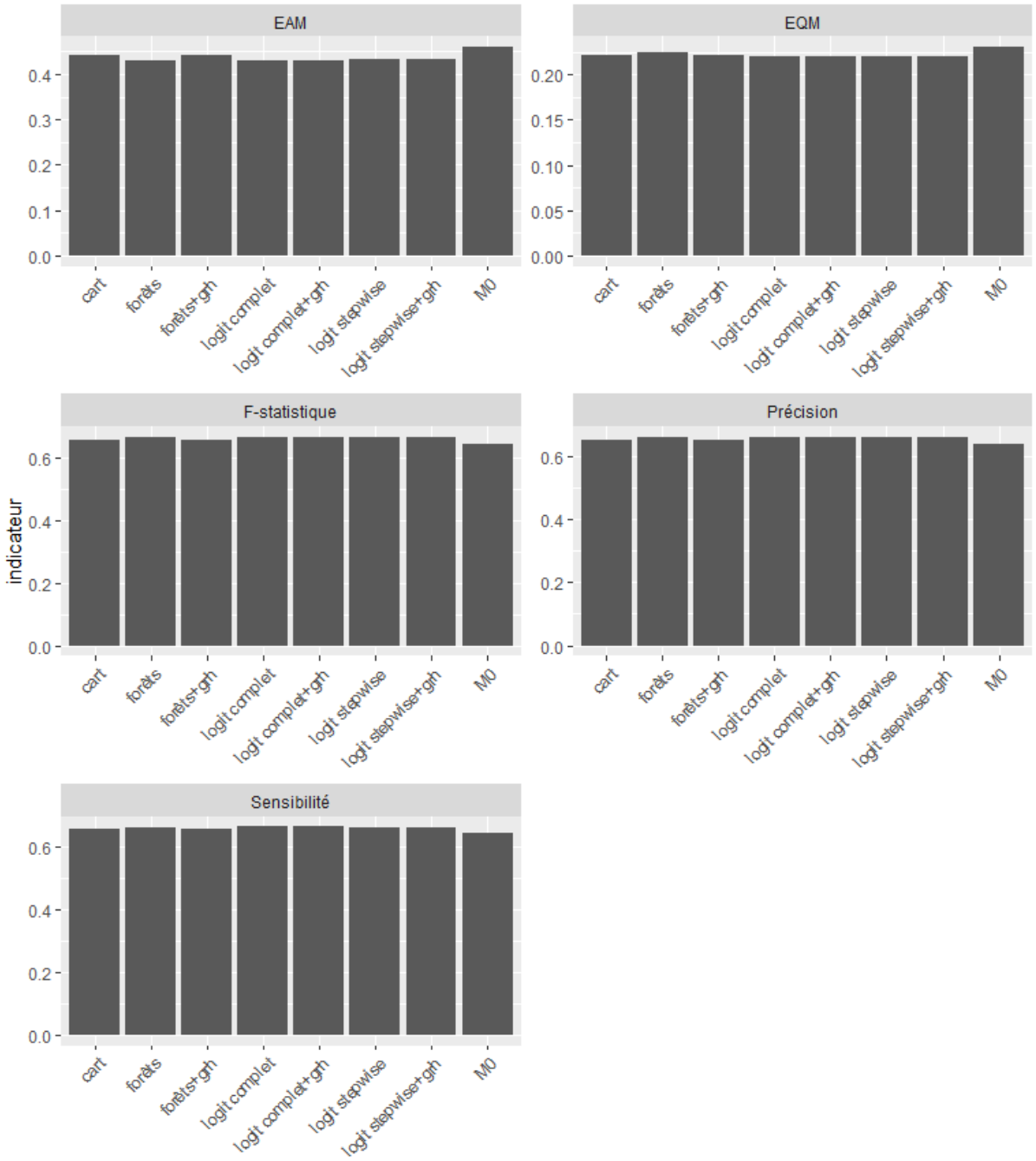
Estimation des probabilités de répondre

Les performances des différents modèles, dont les indicateurs de qualité sont estimés en moyenne sur 50 itérations du découpage de l'échantillon initial entre échantillon de test de d'apprentissage, sont à nouveau très proches, comme l'illustre le graphique 5. Néanmoins, le modèle de base est sensiblement moins performant que les autres, ce qui traduit le fait que les variables auxiliaires sont un peu plus discriminantes pour la prédiction du comportement de réponse à la collecte directe en face à face que pour la collecte CATI.

[23] L'échantillon complet comprend 27 600 individus mais les 8 241 participants à la collecte capi directe, ajoutés aux 19 354 personnes contactées par téléphone, ne représentent que 27 595 individus. Les 5 personnes manquantes correspondent à des individus échantillonnés mais dont le prestataire de collecte a pu déterminer avant le début de la collecte qu'ils n'appartenaient pas au champ de l'enquête.

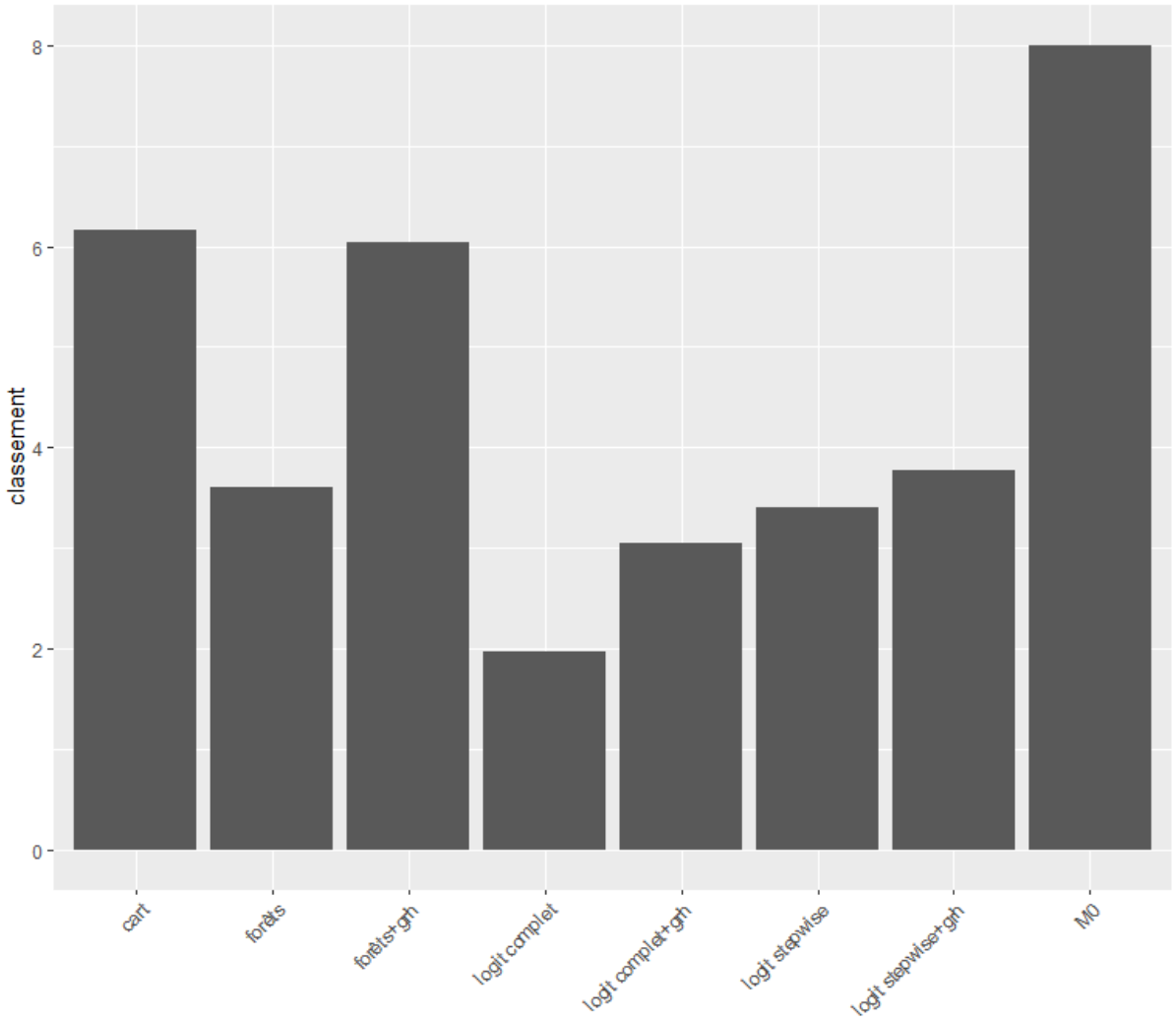
[24] Le taux de réponse est calculé comme la part des répondants dans l'ensemble des répondants et des non répondants. Les hors champs ne sont pas pris en compte dans son calcul.

Graphique 5 : Erreurs quadratique et absolue moyennes, précision, sensibilité et F-statistique pour les modèles de correction de la non réponse à la collecte directe en face à face



Le classement des différents modèles à chaque itération permet de faire ressortir plus clairement le modèle le plus performant : il s'agit du modèle de régression logistique mobilisant l'ensemble des variables auxiliaires (voir graphique 6) mais sans utilisation de groupes de réponse homogène.

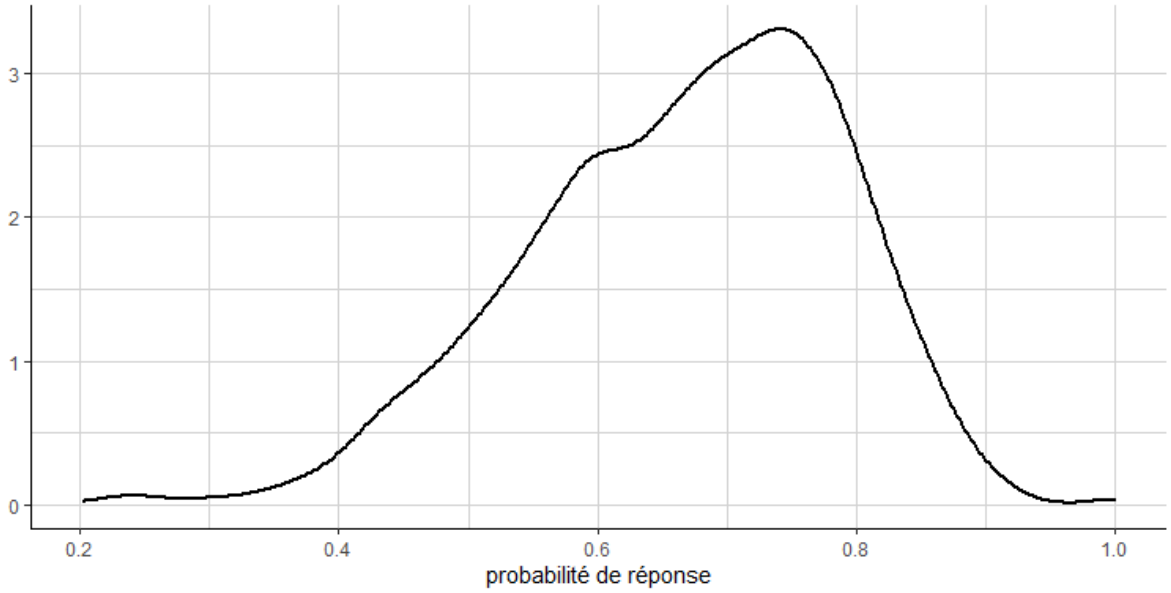
Graphique 6 : Classement de la performance des modèles de correction de la non réponse à la collecte directe en face à face



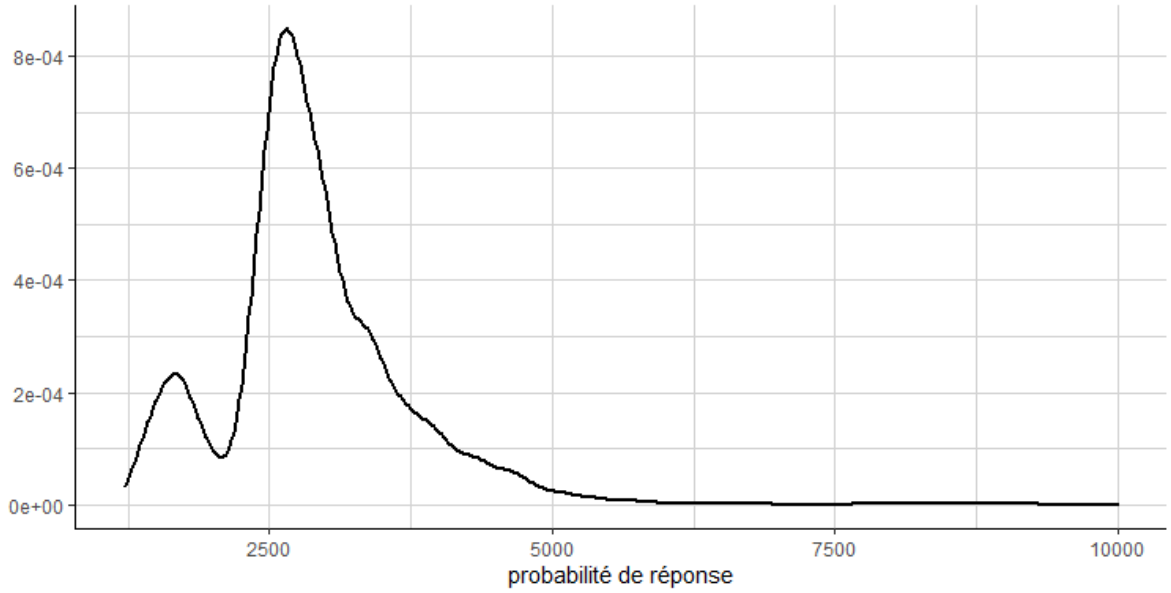
De fait, l'utilisation des groupes de réponse homogène après le modèle de régression logistique conduit à une détérioration des performances prédictives des modèles, que on n'observe pas pour la collecte CATI et qui est d'expérience rarement observée. Aussi les probabilités de réponse utilisées pour la correction de la non réponse à la collecte directe en face à face sont celles estimées par le modèle de régression logistique complet.

Le graphique 7 présente la distribution de ces probabilités de réponse pour les répondants, et le graphique 8 la distribution des poids corrigés de la non réponse des répondants à la collecte directe en face à face.

Graphique 7 : Densité de la distribution des probabilités de réponse des répondants utilisées pour la correction de la non réponse de l'échantillon de la collecte directe CAPI



Graphique 8 : Densité de la distribution des poids corrigés de la non réponse des répondants à la collecte directe en face à face



Estimation de la probabilité de répondre en face à face pour les individus basculés depuis la collecte téléphonique

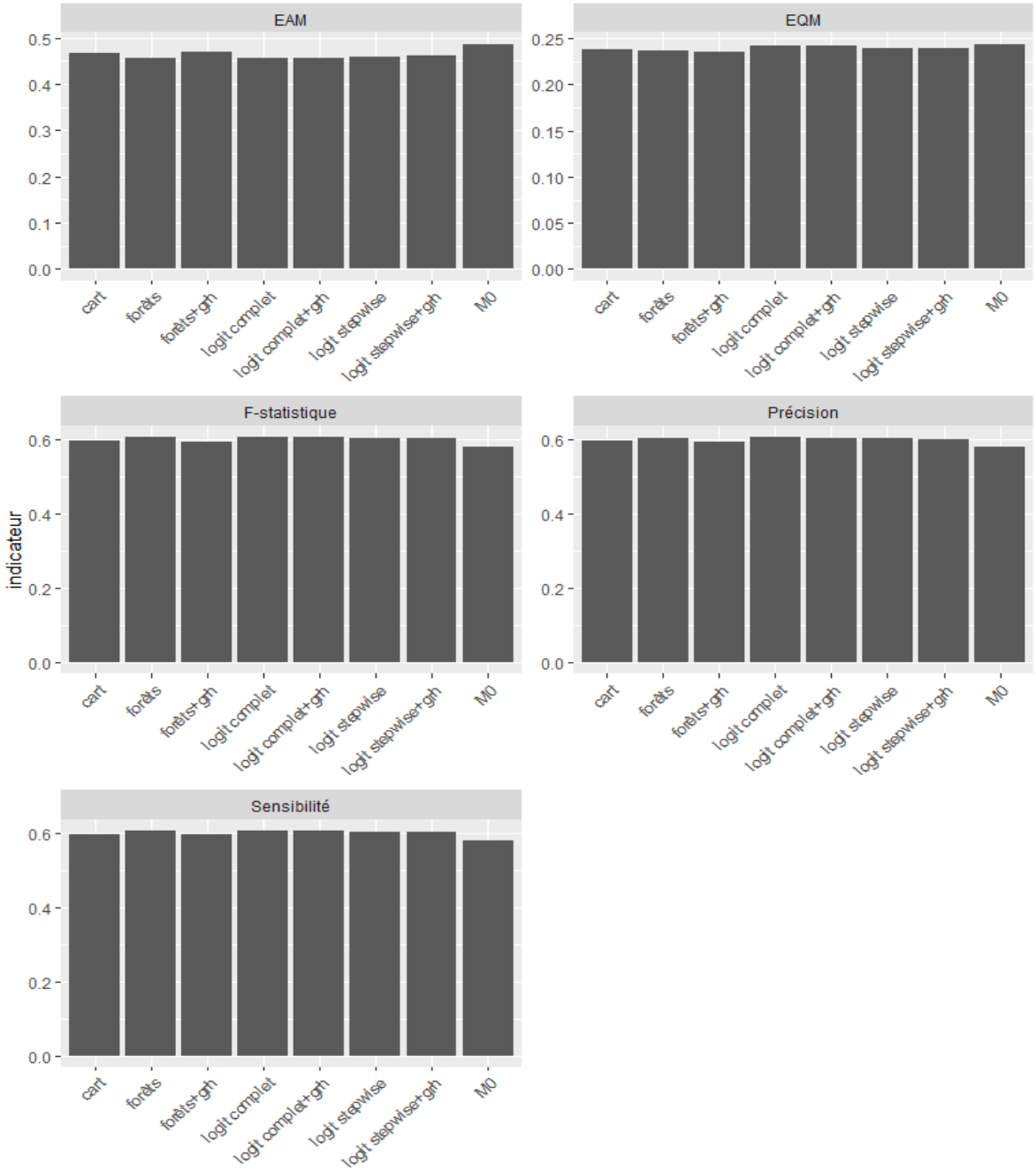
Détermination des statuts de réponse

La collecte en face à face après bascule depuis la collecte téléphonique concerne 2 216 personnes, dont 170 hors champs détectés lors de la collecte téléphonique basculés à tort. Si nous ne tenons pas compte de ces individus basculés à tort pour le calcul des pondérations, 2 046 personnes ont été basculées. Leurs statuts de réponse sont déterminés de façon similaire à ce que l'on a appliqué aux autres collectes. 4 basculés sont identifiés comme hors-champs et et 1 183 sont répondants, ce qui correspond à un taux de réponse de 58 %.

Estimation des probabilités de réponse

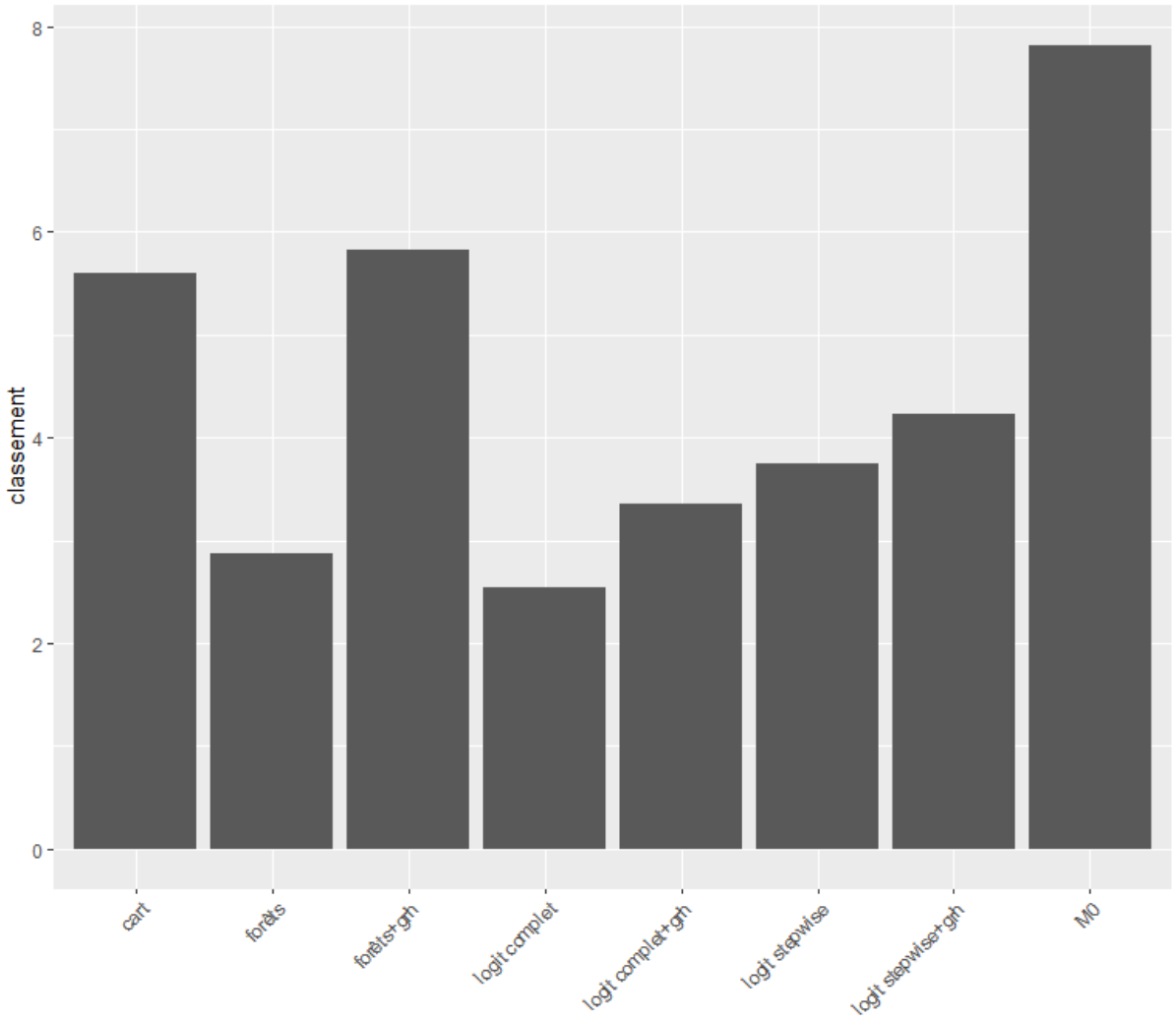
Les mêmes modèles et algorithmes utilisés pour les deux précédentes modélisations sont testés sur l'échantillon basculé en face à face[25]. Le graphique 9 présente les erreurs quadratique et absolue moyenne, les précision, sensibilité et F-statistique de ces différents algorithmes, estimées sur 50 itérations du processus de découpage de l'échantillon initial en échantillon d'apprentissage et échantillon de test. Le graphique 10 présente quant à lui leur classement moyen.

Graphique 9 : Erreurs quadratique et absolue moyennes, précision, sensibilité et F-statistique pour les modèles de correction de la non réponse à la collecte en face à face après bascule



[25] Les seules différences portent sur la variable de tranche de taille d'aire urbaine, dont deux modalités ont dû être regroupées pour cause d'effectifs insuffisants, et la variable de région, la Corse ayant dû être regroupée avec la région PACA pour le même motif.

Graphique 10 : Classement de la performance des modèles de correction de la non réponse à la collecte en face à face après bascule



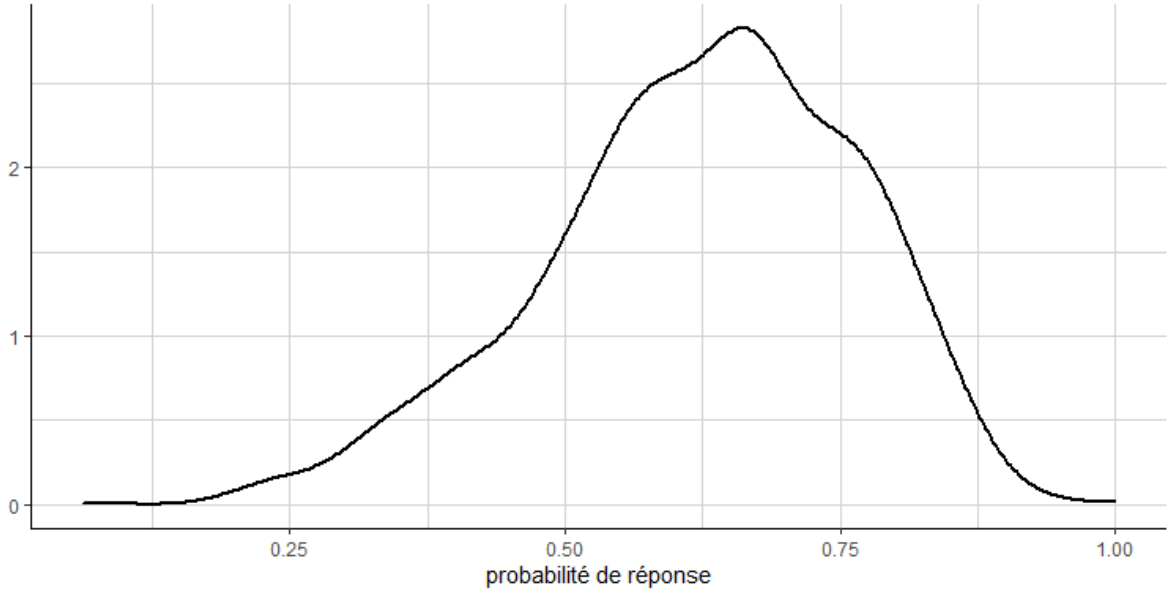
Les performances des différents modèles sont proches, même si les modèles de régression logistique complet et les forêts aléatoires se détachent des autres et notamment des performances du modèle de base. A nouveau, la constitution des groupes de réponse homogène se traduit par une dégradation des performances des modèles les plus efficaces, dégradation spectaculaire pour les forêts aléatoires et moins forte pour le modèle de régression logistique complet.

Une particularité de la correction de la non réponse sur l'échantillon des personnes basculées du téléphone vers le face à face tient à la taille de l'échantillon et aux limites qu'elle impose sur la constitution des groupes de réponse homogène. En effet, ceux-ci doivent contenir au moins 100 individus. Cela empêche les GRH constitués d'atteindre le seuil de 99 % fixé pour la constitution des GRH, et même le seuil de 95 % au delà duquel les performances des GRH sont le plus souvent suffisantes.

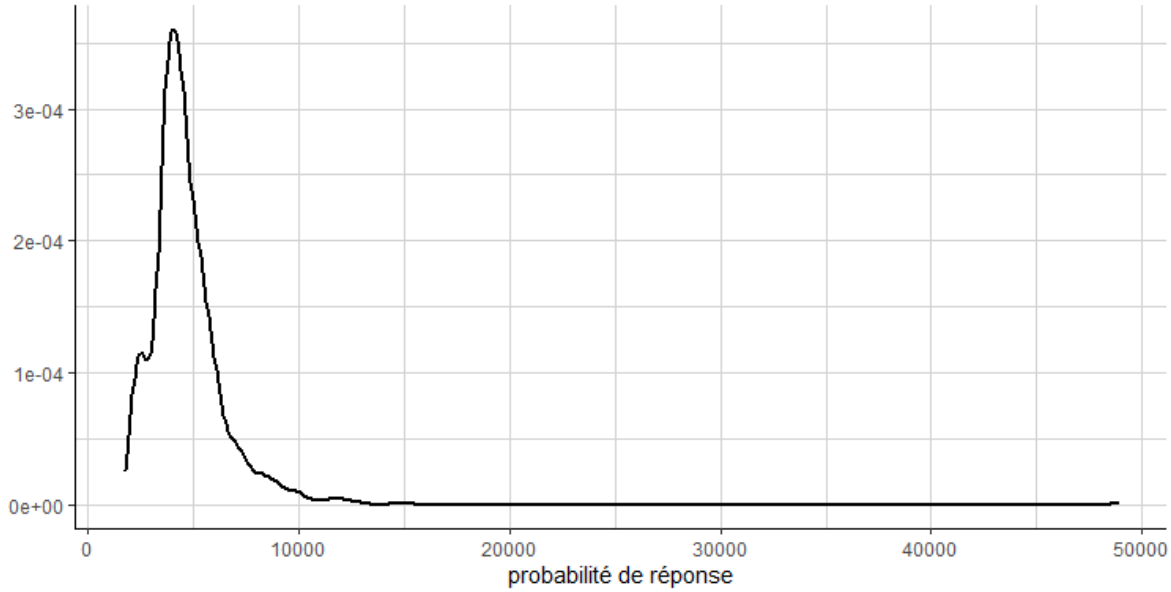
Pour ces différentes raisons, comme pour la correction de la non réponse à la collecte directe en face à face, on a choisi de ne pas utiliser les probabilités de réponse estimées avec les groupes de réponse homogène, mais les probabilités de réponse directement estimées par le meilleur modèle, i.e. le modèle de régression logistique mobilisant l'ensemble des variables auxiliaires. Cela se traduit par une modification plus importante des pondérations, car la probabilité de réponse minimale estimée sur les répondants est de 0.062, soit une multiplication maximale par 16.16 des poids par la correction de la non réponse. Cette dispersion accrue des poids n'induit cependant pas de perte de précision significative, l'échantillon de l'EHIS 2019 respectant les critères de précision européens.

Le graphique 11 présente la distribution des probabilités de réponse utilisées et le graphique 12 la distribution des poids corrigés de la non réponse de l'échantillon collecté en face à face après bascule depuis la collecte au téléphone.

Graphique 11 : Densité de la distribution des probabilités de réponse des répondants utilisées pour la correction de la non réponse de l'échantillon de la collecte CAPI après bascule de la collecte CATI



Graphique 12 : Densité de la distribution des poids corrigés de la non réponse des répondants à la collecte en face à face après bascule de la collecte au téléphone



CALAGE SUR MARGES

Les pondérations dont le calcul a été détaillé dans les sections précédentes permettent de calculer des estimateurs sans biais des caractéristiques de la population, notamment des totaux des variables d'intérêt de l'enquête sur la population. Certains de ces totaux sont connus par ailleurs, via d'autres sources que l'EHIS, et notamment via des sources exhaustives ou dont l'échantillon est beaucoup plus volumineux que celui de l'enquête santé en métropole. Il s'agit notamment de totaux issus du recensement de la population ou de l'Enquête Emploi en Continu (EEC). Aussi, les estimateurs des totaux de variables issues de ces autres sources sont beaucoup plus précis que ceux que permet de calculer l'EHIS.

L'opération avec laquelle on termine le calcul des pondérations de l'EHIS 2019 est un calage sur marges[26]. Son objectif est de modifier légèrement, en tout cas le moins possible, les pondérations corrigées de la non réponse de l'échantillon afin d'obtenir des poids calés sur des marges de référence, i.e. qui permettent d'estimer exactement les totaux d'un certain nombre de variables mesurées dans l'enquête et dont le total dans la population est connu par une source extérieure. L'échantillon est calé sur des marges de référence calculées sur les enquêtes Emploi. Ce calage a trois objectifs :

- améliorer la précision des indicateurs calculés sur l'échantillon des répondants à l'enquête ;
- garantir la cohérence de l'enquête avec des indicateurs de référence socio-démographiques ;
- redresser l'échantillon des défauts de couverture liés aux décalages entre la base de sondage et la population au moment de la collecte.

En effet, le champ de l'enquête est formé de la population de 15 ans ou plus vivant en logement ordinaire à titre de résidence principale au moment de sa collecte. La base de sondage de son côté a été restreinte aux résidences principales : la plupart des logements vacants ou des résidences secondaires au moment de la constitution de la base de sondage ne sont pas des résidences principales, 18 mois plus tard, au moment de la collecte de l'EHIS. Restreindre la base de sondage aux résidences principales permet d'éviter de consacrer des moyens de collecte à des logements dont une grande part est hors-champ. L'échantillon ne peut cependant représenter les logements devenus des résidences principales entre la base de sondage et le moment de la collecte. Par contre, l'échantillon de l'enquête Emploi, utilisée pour calculer les marges de calage des enquêtes auprès des ménages, ne souffre pas de cette limite, car il inclut des résidences non principales dans la base de sondage.

Le calage sur marges de ce fait permet d'assurer une forme de correction du défaut de couverture induit par les changements de statut des logements entre base de sondage et collecte. De fait, le biais lié au défaut de couverture est complètement supprimé pour les variables de calage. Ce faisant, on fait l'hypothèse qu'il est fortement diminué pour les variables d'intérêt [27].

Les marges de calage décrivent la population au moment de la collecte. Elles sont calculées à l'aide des échantillons trimestriels de l'enquête Emploi des quatre trimestres de l'année 2019.

Pour le calage de l'EHIS 2019, on a retenu les marges de calage suivantes, proches des marges de référence pour le calage des enquêtes auprès des ménages qu'a définies l'Insee et qu'il utilise pour ses enquêtes :

- le nombre de personnes par sexe et âge, en considérant les tranches d'âge suivantes : moins de 25 ans, de 25 à 44 ans, de 45 à 64 ans, de 65 à 74 ans, 75 ans ou plus ;
- le nombre de personnes par catégorie sociale, regroupées en 7 modalités :

---

[26] voir J.C. Deville, C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 1992 et J.C. Deville, C.E. Särndal, O. Sautory, *Generalized raking procedures in survey sampling*, Journal of the American Statistical Association, 1993. Voir également la fiche méthodologique sur le calage sur marges rédigée par le département des méthodes statistiques de l'Insee sur le site de l'Institut.

[27] La correction du biais de couverture se fait sous l'hypothèse qu'il n'est déterminé que par les variables de calage. En annulant le biais de défaut de couverture sur ces variables, on peut sous cette hypothèse considérer qu'il est également annulé sur les autres variables. Cette hypothèse est forte, mais elle est systématiquement posée pour les enquêtes auprès des ménages de la statistique publique, dans lesquelles il est rare d'échantillonner des logements vacants ou secondaires dans la base de sondage (l'EEC étant à ce titre une exception).

  - agriculteurs et anciens agriculteurs ;
  - Indépendants, chefs d'entreprise, anciens indépendants et anciens chefs d'entreprise ;
  - Cadres et professions intermédiaires ;
  - Anciens cadres et anciennes professions intermédiaires ;
  - Employés et ouvriers ;
  - Anciens employés et anciens ouvriers ;
  - Personnes sans catégorie sociale (notamment n'ayant jamais travaillé).
- le nombre de personnes par niveau de diplôme agrégé en quatre modalités, en tenant compte des regroupements de diplômes suivants :
  - niveau de diplôme supérieur strictement à BAC+2 ;
  - niveau de diplôme compris entre BAC et BAC+2 ;
  - diplômes de type CAP, BEP et équivalents ;
  - enfin personnes sans diplôme ou déclarant un diplôme inconnu du référentiel des diplômes ;
- le nombre de personnes par regroupements des anciennes régions, avec les regroupements suivants :
  - Île de France, Champagne Ardennes, Normandie, Bourgogne, Picardie, Centre Val de Loire ;
  - Nord Pas de Calais, Alsace, Lorraine, Franche Comté, Pays de Loire, Bretagne, Poitou-Charentes ;
  - Corse, PACA, Auvergne, Midi-Pyrénées, Languedoc-Roussillon, Rhône Alpes, Limousin, Aquitaine ;
- le nombre de personnes vivant dans un quartier prioritaire de la politique de la ville ;
- le nombre de personnes par nationalité en distinguant les personnes de nationalité française et les personnes étrangères ;
- le nombre de personnes suivant la tranche d'unité urbaine de leur commune de résidence, en distinguant les modalités suivantes :
  - communes rurales ;
  - communes de 2 000 à 19 999 habitants ;
  - communes de 20 000 à 99 999 habitants ;
  - communes de 100 000 à 1 999 999 habitants ;
  - Paris.

Le calage sur marges est précédé d'une mise à niveau générale des poids : ceux-ci sont tous multipliés par le ratio entre la taille de la population issue des marges de calage et la somme des poids corrigés de la non réponse. Cette opération ne modifie par les résultats du calage sur marge[28] et permet de calculer des ratios entre les poids après calage et les poids avant calage centrés sur 1.

Plusieurs méthodes ont été testées pour la mise en oeuvre du calage sur marges : méthode linéaire, méthode exponentielle, méthode logistique sur différents jeux de bornes. Le calage finalement retenu utilise la méthode logistique en contraignant les rapports entre les poids après et avant calage à être compris entre 0,6 et 1,7.

Le graphique 13 présente les graphiques des rapports de poids après calage / avant calage pour différentes méthodes testées. Il est complété par le tableau 3 qui décrit plus en détail la distribution de ces rapports de poids. Ces différents éléments montrent que le choix d'une méthode de calage bornée permet de restreindre les effets extrêmes du calage, tout en évitant de faire trop augmenter le nombre d'individus pour lesquels le calage se traduit par une modification forte des poids : il est possible de limiter l'intervalle dans lequel les rapports des poids prennent leur valeur sans que la distribution de ces derniers se concentre sur les bornes de l'intervalle.

[28] Les poids obtenus en ajustant avant calage le niveau global des poids sont identiques à ceux qui seraient obtenus si cette opération était réalisée par le calage sur marges lui-même. Ceci est dû au fait que les variables de calage sont toutes des variables qualitatives. Si le calage intégrait au moins une variable continue (un niveau de revenu par exemple), alors les jeux de poids avec et sans ajustement préalable pourraient être différents.

Graphique 13 : Distribution des rapports de poids après et avant calage pour différents calages

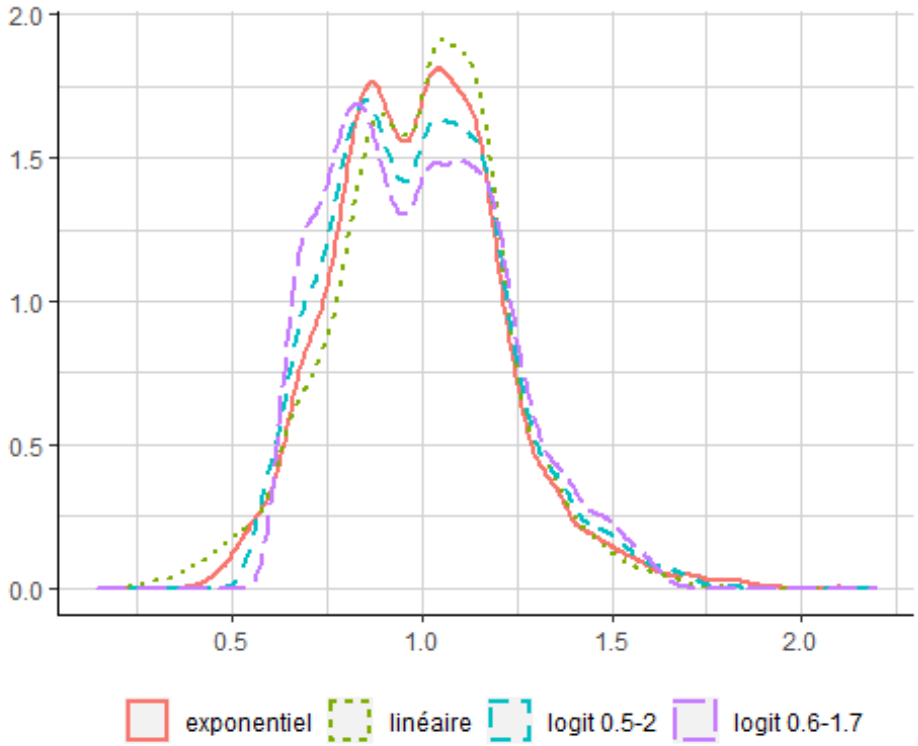


Tableau 3 : Statistiques sur la distribution des rapports de poids après et avant calage pour différents calages

Statistiques	Calage linéaire	Méthode exponentielle	Méthode logistique 0.5-2	Méthode Logistique 0.6-1.7
moyenne	0.99	0.99	0.99	0.99
minimim	0.15	0.41	0.54	0.61
P1	0.45	0.54	0.59	0.63
P5	1.01	0.99	0.99	0.98
P10	0.71	0.72	0.71	0.71
Q1	0.85	0.84	0.82	0.81
Médiane	1.01	0.99	0.99	0.98
Q3	1.14	1.13	1.14	1.15
P90	1.25	1.25	1.27	1.28
P95	1.34	1.36	1.38	1.39
P99	1.53	1.61	1.57	1.55
Maximum	1.85	2.19	1.82	1.67

Le tableau 4 résume l'évolution de la distribution des poids, en partant des poids définis par le plan de sondage jusqu'aux poids calés.

Tableau 4 : Distribution des poids de sondage, des poids corrigés de la non réponse et des poids calés finaux de l'EHIS 2019

Statistiques	Poids de sondage	Poids CNR	Poids calé
moyenne	1874.5	2967.0	3692.5
minimim	824.0	1222.6	992.2
P1	1011.4	1408.8	1254.9
P5	2006.4	2891.3	3481.0
P10	1031.2	2025.6	2196.3
Q1	1999.9	2618.0	2735.5
Médiane	2006.4	2891.3	3481.0
Q3	2012.5	3044.8	4329.0
P90	2016.3	3775.9	5307.1
P95	2017.5	4511.1	6245.9
P99	2021.7	6586.5	8969.9
Maximum	2187.5	48844.6	52701.1