L'utilisation des R-indicateurs pour prioriser la collecte des enquête ménages: une application à l'enquête Patrimoine 2010

Thomas Merly-Alpa, Antoine Rebecq

L'utilisation des r-indicateurs pour prioriser la collecte des enquetes ménages : une application a l'enquete patrimoine 2010

Thomas MERLY-ALPA[1]*(*), Antoine REBECQ*[2]*(*)*
(*) DMCSI, INSEE

Résumé

Pour réaliser ses enquêtes ménages, qui se font principalement en face-à-face, l'INSEE échantillonne les individus interrogés par un sondage à deux degrés. Chaque unité primaire est une zone d'action enquêteur (ZAE), affectée à un ou plusieurs enquêteurs. L'INSEE peut rencontrer des problèmes logistiques tels que l'absence d'un enquêteur ou la baisse des taux de réponses dans une ZAE en particulier. Il est alors possible de réattribuer d'autres enquêteurs à la ZAE affectée, mais uniquement pour y réaliser un nombre restreint d'enquêtes. Nous avons donc besoin de techniques pour choisir les ménages à interroger en priorité, afin de réduire autant que possible le biais de non-réponse et la perte de précision induite par la baisse du taux de réponse.

Pour ce faire il est possible de s'appuyer sur le concept de R-indicateur, ou indicateur de Représentativité, introduit par Shouten & al. en 2009 [3]. Le R-indicateur permet de mesurer, sans utiliser la variable d'intérêt ou des covariables, le degré de similarité d'un échantillon par rapport à la population de base. Il est basé sur la dispersion des probabilités des ménages échantillonnés à répondre, et, suivant [6], se décline en R-indicateurs partiels permettant d'étudier cette représentativité variable par variable. Ces R-indicateurs sont des outils permettant d'analyser la collecte en isolant des groupes de populations sous-représentées ; de façon analogue à [13], on cible les populations ayant les R-indicateurs les plus grands. Cela permet de lancer une procédure dite de « priorisation » qui consiste à intensifier les efforts de collecte sur les groupes précédemment identifiés.

Cette procédure a été testée sur les données de l'enquête Patrimoine 2010 afin d'une part d'observer l'évolution de la représentativité au cours de la collecte, et d'autre part d'étudier la possibilité de compenser la perte de précision induite par une baisse générale ou localisée des taux de réponse. La procédure étant souple, plusieurs scénarios de priorisation ont été simulées sous des hypothèses de dégradation des taux de collecte, ce qui

1. thomas.merly-alpa@insee.fr
2. antoine.rebecq@insee.fr

a permis d'évaluer leurs effets sur la variance de certains indicateurs issus de l'enquête : patrimoine brut moyen, patrimoine net moyen...

Abstract

The French National Institute for Statistics and Economic Studies (INSEE) realizes lots of face to face surveys and may face logistical problems such as a loss of response rate. We need prioritization techniques to select the households that need to be investigated in order to reduce as far as possible the non-response bias. To fulfill this goal, we use R-indicators to analyse the representativeness of the sample during the survey. These R-indicators offer us a way to differentiate population in terms of representativeness in order to prioritize the collect of data. The purpose of this paper will be to apply these methods to the data which were collected for the 2010 Household Wealth survey and to evaluate their effects on the quality of estimation of the distribution of wealth in France

Mots-clés

Sondages, Enquêtes ménages, Non-réponse, Collecte adaptative, R-indicateurs.

Table des matières

1 Les R-indicateurs

1.1 Définitions

1.1.1 Le R-indicateur total

Le **R-indicateur** [3] est une mesure du manque d'association entre réponse et variables auxiliaires (la situation est d'autant meilleure que l'association est faible), ainsi qu'une mesure de similitude entre répondants et population totale. Cette représentativité, ici l'absence de forces sélectives, n'est pas définie en fonction du sujet étudié ou de l'estimateur utilisé et se concentre sur la qualité de la collecte des données plutôt que sur l'estimation. Le but est de juger la composition de l'échantillon des répondants en fonction de variables auxiliaires observées en dehors du cadre de l'enquête afin que la sélection des répondants soit similaire à un tirage aléatoire simple suivant ces variables. On est ainsi proche d'un mécanisme CMAR (Completly Missing At Random) comme défini par Rubin dans [7].

Considérons une population de taille N pour laquelle chaque individu possède une propension à répondre $\theta_i = \mathbb{P}[r_i = 1 \mid s_i = 1]$. On note la propension à répondre moyenne $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N} \theta_i$. On a alors la définition suivante.

Définition 1. *Le **R-indicateur** est une mesure du manque d'association entre réponse et variables auxiliaires :*

$$R(\theta) = 1 - 2S(\theta)$$

avec :

$$S(\theta) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}$$

La dispersion des propensions à répondre $S(\theta)$ vérifie l'inégalité suivante :

$$S(\theta) \leq \sqrt{\bar{\theta}(1-\bar{\theta})} \leq \frac{1}{2}$$

ce qui fait que le R-indicateur est un indicateur compris entre 0 et 1, où 1 signifie une représentativité parfaite (dispersion des propensions à répondre nulle).

Dans les faits, les θ_i ne sont pas connus. Il faut donc les estimer par des $\hat{\theta}_i$. Pour cela, on peut envisager plusieurs méthodes, par exemple utiliser des données d'une enquête préalable dans le cas d'une enquête récurrente, ou plus habituellement utiliser des variables auxiliaires et un modèle de type logit/probit. D'autre part, on ne possède de renseignements annexes que sur les unités qui sont échantillonnées. On note $s_i = 1$ lorsque que l'individu i a été échantillonnée, 0 sinon. On note alors :

$$\hat{\bar{\theta}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_i \frac{s_i}{\Pi_i}$$

et le R-indicateur utilisé dans la pratique est alors :

$$\hat{R}(\theta) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\Pi_i}(\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

Le R-indicateur utilisé ici permet d'obtenir une borne sur le biais de non-réponse (par Cauchy-Schwarz). Notons Y la variable d'intérêt que l'on cherche à estimer via l'enquête, et $\hat{\bar{Y}}_{HT}$ l'estimateur d'Horvitz-Thompson de sa moyenne. On a alors que :

$$\left|B(\hat{\bar{Y}}_{HT})\right| \leq \frac{1 - \hat{R}(\theta)S(Y)}{2\bar{\theta}}$$

Ainsi, si $\hat{R}(\theta)$ vérifie une inéquation de la forme $\hat{R}(\theta) \geq 1 - 4\hat{\bar{\theta}}\gamma$, alors le biais sera inférieur à une certaine valeur fixée γ.

1.1.2 Les R-indicateurs partiels

Les **R-indicateurs partiels** (cf [6], [13]) sont des mesures de l'impact d'une variable en particulier sur la représentativité. On parle de **R-indicateur partiel inconditionnel** pour mesurer la distance à une réponse représentative liée à la valeur d'une variable auxiliaire Z à H modalités, et de **R-indicateur partiel conditionnel** pour mesurer l'impact spécifique sur la représentativité d'une variable auxiliaire dans chacun des groupes d'individus similaires au sens des autres variables utilisées pour le traitement de la non-réponse, notées X : par exemple, si celles-ci sont le sexe et l'âge, le R-indicateur conditionnel de l'âge consiste en la mesure de l'impact d'une part de l'âge chez les hommes, d'autre part chez les femmes, tandis que le R-indicateur partiel inconditionnel mesure cet impact de façon globale sur la population entière.

Définition 2. *Le* ***R-indicateur partiel inconditionnel*** *mesure la distance à une réponse représentative pour une variable auxiliaire, et est basé sur la variance extérieure d'une stratification :*

$$R_U(Z) = \sqrt{\sum_{h=1}^{H}\frac{N_h}{N-1}(\bar{\theta}_h - \bar{\theta})^2}$$

où N_h est la taille de la strate h.

On l'estime par :

$$\hat{R}_U(Z) = \sqrt{\sum_{h=1}^{H}\frac{\hat{N}_h}{N}(\hat{\bar{\theta}}_h - \hat{\bar{\theta}})^2}$$

avec $\hat{N}_h = \sum_{\text{strate k}}\frac{s_i}{\Pi_i}$. On peut également s'intéresser au R-indicateur partiel inconditionnel relatif à une modalité h de la variable Z, dont un estimateur est :

$$\hat{R}_U(Z,h) = \sqrt{\frac{\hat{N}_h}{N}}(\hat{\bar{\theta}}_h - \hat{\bar{\theta}})$$

Cet indicateur peut être positif ou négatif.

Définition 3. *Le **R-indicateur partiel conditionnel** mesure la variance restante dûe à la variable Z dans un sous-groupe formé par toutes les autres variables restantes, notées X^-. Notons $j = 1 \dots J$ un découpage en classes de X^-. Le R-indicateur partiel conditionnel est basé sur la variance intérieure :*

$$R_C(Z) = \sqrt{\frac{1}{N-1} \sum_{j=1}^{J} \sum_{i \in U_j} (\theta_i - \bar{\theta}_j)^2}$$

On l'estime par :

$$\hat{R}_C(Z) = \sqrt{\frac{1}{N-1} \sum_{j=1}^{J} \sum_{i \in U_j} \frac{s_i}{\Pi_i} (\hat{\theta}_i - \hat{\bar{\theta}}_j)^2}$$

On peut ici aussi s'intéresser au R-indicateur partiel conditionnel relatif à une modalité h de la variable Z, dont un estimateur est :

$$\hat{R}_C(Z, h) = \sqrt{\frac{1}{N-1} \sum_{j=1}^{J} \sum_{i \in U_j} \frac{s_i}{\Pi_i} \mathbf{1}_{z_i = h} (\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

Cet indicateur est toujours positif.

Ces R-indicateurs partiels servent à déterminer si une variable est pertinente dans l'étude de la représentativité. En effet, via le calcul du R-indicateurs partiels inconditionnel relatifs à une variable, nous pouvons savoir si cette variable joue un rôle dans la représentativité de l'échantillon, c'est à dire si des modalités de cette variables sont sous-représentées, lorsque leur R-indicateur partiel inconditionnel est négatif, ou sur-représentées à l'inverse. Le R-indicateur partiel conditionnel sert quant à lui à mettre en lumière de possibles effets structurels qui conduiraient à considérer une variable comme pertinente alors qu'elle n'est qu'une conséquence de l'influence des autres variables.

1.2 Précision des R-indicateurs

1.2.1 Biais de sélection

L'estimateur utilisé ici pour le R-indicateur est biaisé : lorsque la taille de l'échantillon diminue, ce biais augmente. Pour remédier à ce problème, nous pouvons modifier la définition du R-indicateur utilisé et lui préférer :

$$\hat{R}(\theta) = 1 - 2\sqrt{\left(1 - \frac{1}{n} - \frac{1}{N-1}\right) \sum_{i=1}^{N} \frac{s_i}{\Pi_i} (\hat{\theta}_i - \hat{\bar{\theta}})^2 - \frac{1}{n} \sum_{i=1}^{N} s_i z_i (\sum_{j=1}^{N} s_j z_j x_j^T)^{-1} z_i}$$

où $z_i = \nabla h(x_i^T \hat{\beta}) x_i$, h étant la fonction *logit* et $\hat{\beta}$ le paramètre estimé du modèle logistique.

Ce biais existe également pour les R-indicateurs partiels, car ils sont construits à partir du R-indicateur et sont estimés par la même méthode. Il n'existe pas de formule directe permettant de résoudre ce problème, mais deux approches sont introduites dans [12] :

1. On peut répartir par une heuristique de prorata l'ajustument effectué ci-dessus entre R-indicateur partiel conditionnel et R-indicateur partiel inconditionnel. Pour cela, on utilise la formule de décomposition de la variance.
2. On peut aussi utiliser l'ajustement ci-dessus en ne considérant qu'une seule variable Z pour le modèle logistique (ou uniquement X^- pour le R-indicateur partiel conditionnel).

Pour les R-indicateurs partiels relatifs aux modalités, on ne peut utiliser que la première méthode.

Ces deux méthodes ont été testées sur des échantillons de différentes tailles issus de s données du 95 Israël Census. On remarque alors que le biais augmente bien lorsque la taille de l'échantillon diminue, mais qu'aucune des deux approches ne permet de corriger entièrement le biais : pour les R-indicateurs partiels inconditionnels, la méthode 2 semble meilleure, mais c'est l'inverse pour les R-indicateurs partiels conditionnels. On en conclut deux résultats :

— Lorsque l'échantillon est suffisamment grand (une borne de 15.000 foyers a minima est postulée par l'article), il n'y a pas besoin d'utiliser de méthodes de correction du biais pour les R-indicateurs total et partiels.
— En revanche, lorsque l'échantillon est de petite taille, on favorisera la méthode 1 qui permet d'obtenir des résultats concernant les modalités des variables.

Ces conclusions permettent de considérer que le biais dans l'estimation des différents R-indicateurs n'affecte pas les résultats de l'étude effectuée ici, en tout cas en ce qui concerne l'échantillon standard. Il serait peut-être à étudier d'implémenter l'estimateur non-biaisé pour l'échantillon non-standard.

1.2.2 Variance des estimateurs

R-indicateur total. Il est possible de construire des intervalles de confiance pour le R-indicateur [15] en se basant sur un estimateur de la variance de $\hat{S}^2$ que l'on notera $v(\hat{S}^2)$. Pour cela, en utilisant la relation $\hat{R} = 1 - 2\hat{S}$ et en linéarisant la variance, on obtient que :

$$\mathrm{Var}(\hat{\mathrm{R}}) \approx \hat{\mathrm{S}}^{-2}\mathrm{Var}(\hat{\mathrm{S}}^2)$$

On estime ensuite $\mathrm{Var}(\hat{\mathrm{S}}^2)$ en le décomposant suivant deux effets :

— Le terme provenant de l'erreur d'échantillonnage et donc du plan du sondage, à une valeur $\hat{\beta}$ du paramètre du modèle logistique fixée.
— Le terme provenant de l'erreur d'estimation de $\hat{\beta}$ dans le modèle logistique.

En faisant l'hypothèse que $\hat{\beta}$ suit une loi normale centrée en la vraie valeur β, on arrive à donner une formule pour un intervalle de confiance de niveau $1 - \alpha$ de $\hat{R}$:

$$1 - 2\sqrt{\hat{S}^2 \pm z_{\alpha/2}\sqrt{v(\hat{S}^2)}}$$

avec $z_{\alpha/2}$ quantile de niveau $\alpha/2$ de la Gaussienne centrée réduite.

R-indicateurs partiels. Il est également possible de s'intéresser à la variance des R-indicateurs partiels [12]. La variance des R-indicateurs partiels conditionnels s'estime de la

même façon que celle du R-indicateur, en utilisant un développement au second degré. En revanche il existe une méthode plus simple pour les R-indicateurs partiels inconditionnels. On a par définition, pour une modalité h de la variable Z :

$$\begin{aligned}\mathrm{Var}(\hat{\mathrm{R}}_{\mathrm{U}}(\mathrm{Z},\mathrm{h})) &= \frac{\hat{N}_h}{N}\mathrm{Var}(\hat{\bar{\theta}}_{\mathrm{h}} - \hat{\bar{\theta}}) \\ &= \frac{\hat{N}_h}{N}[\mathrm{Var}(\hat{\bar{\theta}}_{\mathrm{h}}) + \mathrm{Var}(\hat{\bar{\theta}}) - 2\mathrm{Cov}(\hat{\bar{\theta}}_{\mathrm{h}}, \hat{\bar{\theta}})]\end{aligned}$$

Or, si on pose la notation suivante :

$$\hat{\bar{\theta}}_{h^C} = \frac{1}{N - \hat{N}_h}\sum_{i=1}^{N}\frac{s_i}{\Pi_i}\hat{\theta}_i \mathbf{1}_{z_i \neq h}$$

on a que :

$$\hat{\bar{\theta}} = \frac{\hat{N}_h}{N}\hat{\bar{\theta}}_h + \left(1 - \frac{\hat{N}_h}{N}\right)\hat{\bar{\theta}}_{h^C}$$

et donc par définition de la covariance et par un jeu de réécriture :

$$\begin{aligned}\mathrm{Var}(\hat{\mathrm{R}}_{\mathrm{U}}(\mathrm{Z},\mathrm{h})) &= \frac{\hat{N}_h}{N}\left[\left(1 - \frac{2\hat{N}_h}{N}\right)\mathrm{Var}(\hat{\bar{\theta}}_{\mathrm{h}}) + \mathrm{Var}(\hat{\bar{\theta}})\right] \\ &= \frac{\hat{N}_h}{N}\left(1 - \frac{\hat{N}_h}{N}\right)^2[\mathrm{Var}(\hat{\bar{\theta}}_{\mathrm{h}}) + \mathrm{Var}(\hat{\bar{\theta}}_{\mathrm{h^C}})]\end{aligned}$$

On obtient donc, en développant les variances au premier ordre, une estimation de la variance du R-indicateur partiel inconditionnel de la modalité h de la variable Z :

$$\mathrm{Var}(\hat{\mathrm{R}}_{\mathrm{U}}(\mathrm{Z},\mathrm{h})) \approx \frac{\hat{\mathrm{N}}_{\mathrm{h}}}{\mathrm{N}}\left(1 - \frac{\hat{\mathrm{N}}_{\mathrm{h}}}{\mathrm{N}}\right)^2\left[\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{\mathrm{s_i}}{\Pi_{\mathrm{i}}}\left(\frac{\hat{\theta}_{\mathrm{i}}}{\hat{\mathrm{N}}_{\mathrm{h}}}\mathbf{1}_{\mathrm{z_i=h}} + \frac{\hat{\theta}_{\mathrm{i}}}{\mathrm{N} - \hat{\mathrm{N}}_{\mathrm{h}}}\mathbf{1}_{\mathrm{z_i \neq h}}\right)\right]$$

R-indicateurs partiels des variables. La variance du R-indicateur partiel associé à une variable est plus compliquée à calculer car elle fait intervenir des covariances difficiles à estimer. Il n'existe pas pour l'instant de résultat théorique donnant la variance de tels R-indicateurs partiels.

1.3 Controverses et limites

1.3.1 Comparaison avec le taux de réponse partiel

Le taux de réponse total de l'enquête n'est pas lui-même un indicateur de représentativité de l'échantillon, n'étant pas lié au biais de non-réponse. En revanche, les taux de réponse partiels dans différents sous-groupes liés aux modalités des variables auxiliaires sont des indicateurs utiles. Comme nous l'avons vu au paragraphe 2.2, il y a un lien assez direct entre R-indicateurs partiels inconditionnels et taux de réponse partiels. On pourrait donc se demander pourquoi utiliser les R-indicateurs partiels en lieu et place des taux de réponse partiels. Comme l'explique Schouten dans [5], ils ont 4 avantages :

1. Ils sont directement liés au R-indicateur global, qui est un indicateur de représentativité, contrairement au taux de réponse.
2. Ils sont disponibles au niveau de la variable et non seulement au niveau de ses modalités.
3. Ils sont à la fois conditionnels et inconditionnels, tandis que le taux de réponse partiel ne prend pas en compte les effets conditionnels des autres variables.
4. Ils prennent en compte la taille du sous-groupe considéré.

1.3.2 Traitement a posteriori de la non-réponse

Une limite de la méthode est que, selon Haziza, réduire directement le biais de non-réponse lors de la phase de collecte en utilisant les variables auxiliaires n'apporte rien de plus que d'utiliser ces variables lors de la phase de correction de la non-réponse. Il propose donc de plutôt s'intéresser à la minimisation de la variance de la non-réponse [4].

On peut pourtant trouver quelques avantages à une priorisation et donc une gestion de la collecte "au fil de l'eau". En effet, cela permet pour les groupes très sous-représentés d'augmenter le nombre de répondants, ce qui permet de réduire les poids corrigés pour traiter la non-réponse et donc de réduire un possible impact d'une valeur atypique. D'autre part, cela offre un moyen d'agir sur la collecte en cas de problème (comme celui évoqué ici avec les NCEE et l'absence d'un enquêteur), ce qui n'est pas possible avec le calage qui est effectué a posteriori. Un autre avantage est d'augmenter la fiabilité des méthodes d'imputation *hot deck* utilisés pour gérer la non-réponse partielle, en augmentant la taille de la base dans laquelle la valeur imputée est tirée. Il est même possible d'intégrer la problématique du traitement de la non-réponse et de la minimisation de la dispersion des poids corrigés de la non-réponse dans la procédure de priorisation, en utilisant l'algorithme CURIOS [2].

Enfin, la correction a posteriori de la non-réponse suppose que le mécanisme de non-réponse soit MAR pour que les techniques de calage fonctionnent. Or cette hypothèse est très forte. Les R-indicateurs sont justement un moyen de valider cette hypothèse en concevant un mécanisme d'enquête qui soit tel que la non-réponse soit, à défaut d'être uniforme, explicable par les variables auxiliaires.

2 La méthode de priorisation

2.1 Principe de la méthode

Ainsi, il s'agit de s'intéresser aux variables dont les R-indicateur partiel conditionnel et R-indicateur partiel inconditionnel sont élevés, puis de se concentrer sur les modalités de ces variables dont les R-indicateurs partiels inconditionnels associés sont négatifs.

2.2 Un exemple simple

Considérons un exemple simplifié afin d'étudier l'évolution du R-indicateur et des R-indicateurs partiels au long d'une collecte. On prend un échantillon de 10000 personnes au sein d'une population plus grande par sondage aléatoire simple, et l'on suppose que cette population est équitablement répartie entre hommes et femmes, ainsi qu'entre trois groupes d'âge. On fait l'hypothèse que les hommes ont une propension à répondre plus faible que les femmes, mais que la propension à répondre ne varie pas en fonction de l'âge.

On simule une collecte pour laquelle on a des informations en début, en milieu et en fin de collecte. Dans un premier temps, on considère que l'on n'a pas cherché à prioriser les enquêtes. Dans ce cas, la probabilité de réponse est de 60% chez les femmes et de 20% chez les hommes pendant le début de la collecte, puis on recontacte tous les non-répondants et on obtient de même 60% de réponses chez les femmes, et 3 fois moins chez les hommes, et on reproduit ce mécanisme une troisième fois.

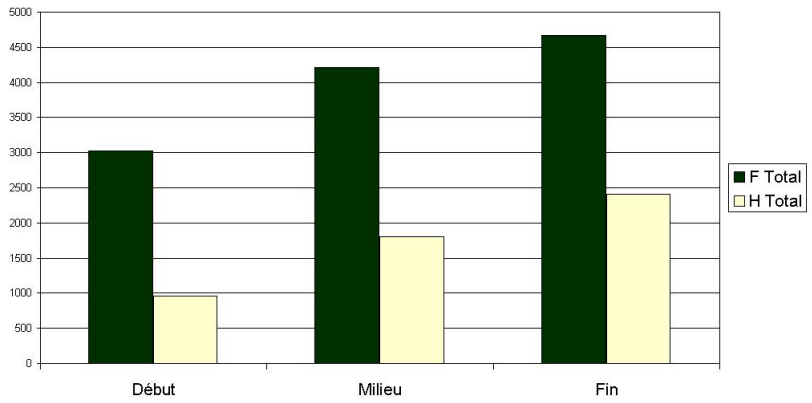

FIGURE 1 – Répartition hommes (en blanc)/femmes (en noir) des répondants dans une enquête sans priorisation

La Figure 1 montre bien qu'entre le début de l'enquête et le milieu, le nombre de femmes ayant répondu (déjà supérieur au nombre d'hommes parmi les répondants) a augmenté plus rapidement que celui d'hommes, ce qui s'explique par leur plus grande propension à répondre, mais qu'entre le milieu de l'enquête et sa fin, c'est le nombre d'hommes qui a augmenté le plus, par manque de femmes dans l'échantillon ; en effet, le taux de réponse des femmes était déjà à 80% en milieu de collecte. On imagine ainsi que la représentativité de l'échantillon diminue entre le début et le milieu de l'enquête

R-indicateur	Début	Milieu	Fin
Sans priorisation	0.588	0.518	0.547

FIGURE 2 – R-indicateurs au cours de la collecte.

(les hommes et les femmes ont des propensions à répondre [3] très éloignées) tandis qu'elle augmente entre le milieu et la fin de la collecte.

Calculons les R-indicateurs pour voir si nos premières conclusions coïncident avec ceux-ci. Pour cela, il faut estimer des propensions à répondre $\hat{\theta}_i$, ce que l'on fait à l'aide d'un modèle logistique faisant intervenir l'âge et le sexe. On obtient alors en Figure 2 les valeurs des R-indicateurs aux différents moments de la collecte.

Ces valeurs confirment bien les hypothèses que nous avions faites à la lecture de la répartition entre hommes et femmes des répondants. Comment faire pour améliorer la collecte pendant son déroulement ? Eh bien, nous allons calculer les R-indicateurs partiels pour chacune des variables et des modalités afin de savoir quel groupe [4] enquêter en priorité.

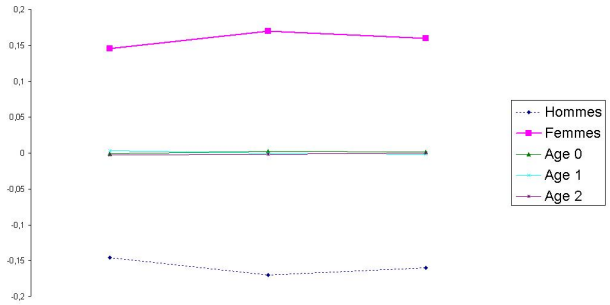

FIGURE 3 – R-indicateurs partiels inconditionnels

Comme le montre la Figure 3, les R-indicateurs partiels inconditionnels relatifs aux modalités concernant l'âge sont proches de 0, tandis que celui des hommes est très fortement négatif, ce qui montre un groupe sous-représenté, et l'inverse évidemment pour les femmes. Si l'on s'intéresse aux R-indicateurs partiels conditionnels, on trouve le même schéma, c'est à dire que le R-indicateur partiel conditionnel de la variable "âge" est très proche de 0 alors que celui de la variable "sexe" est important. Cela montre que :

— Tous les groupes d'âge sont bien représentés, il n'y a pas lieu de prioriser un groupe d'âge.
— Les hommes sont sous-représentés et doivent être priorisés.

3. Il est important de noter ici qu'il s'agit de propensions à répondre sachant le mode de collecte. Par exemple, en imaginant que les probabilités de réponse augmentent proportionnellement au nombre de contacts, si on contactait 3 fois les hommes et une seule fois les femmes, on aurait la même propension à répondre pour les deux, et un R-indicateur de 0. Cela ne signifie pas pour autant qu'à procédé égal, les hommes et les femmes ont autant de probabilités de répondre.

4. même si la question est triviale ici...

Étudions maintenant 3 situations possibles d'évolution du mécanisme de réponse des hommes, en supposant que les enquêteurs aient pour consigne de n'interroger plus que des hommes une fois le début de la collecte effectuée :

1. On suppose que l'effort plus important ne permet pas d'augmenter la probabilité de réponse, et ainsi d'avoir 20% de répondants parmi les hommes interrogés.
2. On suppose que doubler l'effort de collecte sur les hommes augmente un peu la probabilité de réponse, et ainsi que l'on a 30% de répondants parmi les hommes interrogés.
3. On suppose que doubler l'effort de collecte sur les hommes double la probabilité de réponse, et ainsi que l'on a 40% de répondants parmi les hommes interrogés.

On calcule alors dans ces 3 cas les R-indicateurs, et on obtient les résultats de la Figure 4.

R-indicateur	Début	Milieu	Fin	**Taux de réponse**
Sans priorisation	0.588	0.518	0.547	0.71
Hypothèse 1	0.588	0.747	0.873	0.55
Hypothèse 2	0.588	0.817	0.982	0.61
Hypothèse 3	0.588	0.909	0.906	0.65

FIGURE 4 – R-indicateurs pour chacune des hypothèses.

On déduit donc de la Figure 4 que la priorisation réduit le taux de réponse (moins dans les hypothèses 2 et 3, qui supposent une augmentation de probabilité de réponse chez les hommes) mais qu'en revanche, la représentativité mesurée par le R-indicateur augmente. Cela s'explique par le fait que se concentrer sur les hommes pendant la collecte augmente leur propension à répondre et ainsi elle est plus proche de celle des femmes, ce qui réduit mécaniquement la dispersion. Mais on voit que dans l'hypothèse 3, cette représentativité recommence à décroître entre le milieu et la fin de l'enquête. Cela est dû au fait que les hommes deviennent un groupe sur-représenté, ce que l'on voit si l'on s'intéresse aux R-indicateurs partiels inconditionnels (Figure 5).

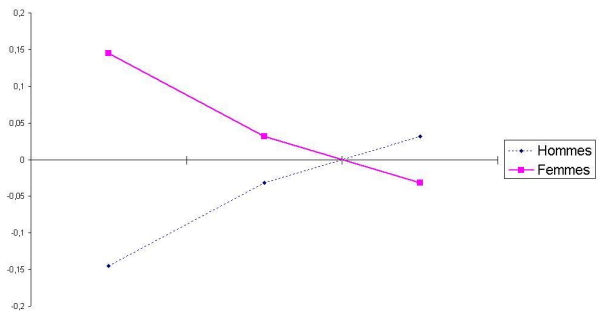

FIGURE 5 – R-indicateurs partiels inconditionnels sous l'Hypothèse 3

Ainsi, le R-indicateur partiel inconditionnel relatif à la modalité "hommes" (courbe en pointillés) est devenu positif, ce qui implique que si la collecte devait continuer il ne

faudrait plus prioriser ce groupe. Plus largement, on peut imaginer que dès le milieu de la collecte, le fait que les R-indicateurs partiels inconditionnels soient proches de 0 pourrait conduire à abandonner la stratégie de priorisation.

2.3 Dans la littérature

Dans [5], article récapitulant les utilisations des R-indicateurs et des R-indicateurs partiels pour étudier la représentativité, on trouve de nombreux exemples de leur utilisation. Par exemple, une comparaison entre différentes études européennes est faite, malgré l'utilisation de trop peu de variables auxiliaires (qui devaient être disponibles dans tous les pays) ce qui empêche d'en tirer des analyses probantes. Une analyse plus fine est possible lorsque l'on s'intéresse à l'enquête VAT [5] 07 uniquement. Cette enquête est menée tous les mois, et on remarque d'une part de grandes disparités selon les mois, ce qui s'explique par la variation de l'information à collecter, par exemple les résultats annuels, d'autre part, ce qui est intéressant, que continuer la collecte les 5 derniers jours n'apporte rien au niveau représentativité.

Enfin un exemple plus important est celui de l'enquête SCC [6] 05. Le calcul des R-indicateurs partiels est fait ici à la fois pour la phase de contact et pour celle de réponse, ce qui permet de différentier les motifs de non-réponse. On remarque alors que pour le contact, la seule variable pertinente est l'âge, et que ce sont les plus âgés qui sont sous-représentés. Pour la phase de réponse, l'étude des R-indicateurs partiels inconditionnels conduit à considérer les variables âge et ethnie [7], et en particulier les modalités suivantes : âges entre 25 et 40 ans ; non-natifs non-occidentaux, non-natifs occidentaux et ethnie inconnue. L'étude des R-indicateurs partiels conditionnels permet de se limiter aux sous-groupes suivants : âges entre 25 et 40 ans, non-natifs occidentaux, ethnie inconnue.

Une expérience approfondie dans [1] lors d'un pilote concernant l'adaptation multimodale de SCC a permis d'étudier l'impact des méthodes utilisant les R-indicateurs. En effet, le but était de maximiser le R-indicateur tout en respectant deux contraintes, qui étaient de garder le même taux de réponse (ou un meilleur) pour le même coût de collecte. Pour cela, on utilise des stratégies adaptées aux groupes déterminés au paragraphe précédent, soit par exemple appeler en journée les personnes âgées plutôt qu'en soirée, et plus généralement on applique les principes suivants :

— On augmente le nombre de contacts pour les unités à faible propension à répondre aux appels, et on diminue le nombre de contacts pour les unités à forte propension à répondre aux appels.
— On stimule la coopération (cadeaux, etc) pour les unités à faible propension à coopérer, et on décourage la coopération (pas de rappels par mail...) pour les unités à forte propension à coopérer.

On voit alors que le R-indicateur est de 0.85, alors que celui d'un groupe témoin n'est que de 0.77. De même, les R-indicateurs partiels ont vu leur valeur (absolue) baisser entre le groupe témoin et le groupe priorisé. Cela montre bien l'impact sur la représentativité que peut avoir une priorisation.

5. Enquête néerlandaise concernant le chiffre d'affaire des entreprises.
6. Enquête néerlandaise téléphonique, *Survey of Consumer Confidence*.
7. Qui n'existerait pas en France.

Une expérience menée dans [13] montre bien l'intérêt des R-indicateurs partiels même lorsque que seule une petite fraction de l'enquête peut être réalisée. Pour cela, on considère un échantillon de 1% d'une base de donnée de plusieurs dizaines de milliers d'unités issues du *1995 Israel Census Sample of Individuals*. On modélise parmi eux de la non-réponse suivant le modèle utilisé pour l'enquête d'origine. On souhaite sélectionner quelques individus à enquêter en plus. Pour cela, on considère une strate associée à une modalité pour laquelle le R-indicateur partiel inconditionnel est négatif, et les R-indicateurs partiels conditionnel et inconditionnel sont tous deux éloignés de 0. En concentrant les efforts sur quelques uns de ces individus (on fait l'hypothèse ici qu'on obtient un taux de réponse de 100% pour la vingtaine d'individus concernés), on n'augmente pas beaucoup le taux de réponse qui passe de 69.8% à 70.7%, mais en revanche le R-indicateur de l'ensemble de l'échantillon augmente de 3%, passant de 0.859 à 0.884.

2.4 Application à l'enquête Patrimoine 2010

2.4.1 Description de l'enquête

L'enquête Patrimoine est une enquête répétée régulièrement depuis 1986 [8] qui vise à étudier le patrimoine moyen des français, leur comportement vis à vis de ce patrimoine (transmissions, achats...) en lien avec leur situation personnelle et professionnelle. La dernière enquête dont les données sont disponibles date de 2010 et sera celle utilisée dans le cadre de ce travail, l'enquête 2014 n'étant pas terminée au moment de l'étude.

Plan de sondage. Le plan de sondage utilisé pour les enquêtes Patrimoine consiste à sélectionner aléatoirement au sein de chacune des ZAE un certain nombre de ménages, déterminé selon la taille de la ZAE. Or, en plus de cet échantillon dit **standard**, les responsables de l'enquête ont souhaité innover et suivre les recommandations de la BCE en sur représentant le haut de la distribution des patrimoines, comme le fait déjà la Fed. Cette sur représentation permet de tenir compte de la variation très importante du haut de la courbe de distribution des patrimoines. En outre, à cette variation s'ajoute un phénomène de non-réponse élevée dans les très hauts patrimoines, dûe à des difficultés d'accès, des réticences ou une plus faible disponibilité. Pour ces raisons, un second échantillon dit **non-standard** a été tiré en utilisant les sources fiscales parmi les individus ayant des patrimoines plus élevés.

Dans chacun de ces échantillons, les individus ont été stratifiés selon des caractéristiques professionnelles, sociales... :

— Pour l'échantillon standard, les 6 strates sont : les agriculteurs, les indépendants à hauts revenus, les cadres, les personnes possédant un revenu du patrimoine, les personnes âgées, et le reste de la population. Cette stratification est usuelle dans les enquêtes Patrimoine.

— Pour l'échantillon non-standard, les 4 strates sont : les riches urbains, les personnes possédant un patrimoine élevé à dominante mobilière, les personnes possédant un patrimoine élevé à dominante immobilière, et les patrimoines plus faibles. Cette

8. Tous les 6 ans jusqu'en 2010, le rythme a changé depuis.

stratification a été déterminée à partir de simulations et d'un algorithme de classification.

Cette stratification a permis une sur représentation de certains groupes (tels que les indépendants pour l'échantillon standard) qui sont connus pour avoir un patrimoine moyen plus important. Finalement, le plan de sondage consistait à réaliser deux échantillons, en tirant à chaque fois dans chaque ZAE et chaque strate un nombre de FA dépendant de la taille de la ZAE et de l'importance supposée de la strate.

Traitement post-collecte. L'échantillon standard comportait environ 17000 FA, et le taux de réponse a été de 70%, tandis que l'échantillon non-standard de 3000 FA avait un taux de réponse plus faible de 50%, soit un taux de réponse total aux alentours de 68%.

Pour corriger la non-réponse totale, une variante [8] de la méthode des scores ordinaires a été utilisée. Pour cela, il convient tout d'abord d'estimer les probabilités de réponse à l'aide d'un modèle logistique sur les variables usuelles (âge, région, strate, type de ménage, type de logement) mais également sur des variables fiscales (revenus d'activités, patrimoine). Les groupes de réponse homogène ont été constitués en regroupant ces probas estimées $\hat{p}$ en un nombre limité de groupes, qui sont difficilement interprétables, et que l'on considère comme ayant le même mécanisme de non-réponse.

Le calage a été fait selon la méthode classique en utilisant les sources fiscales et les données du TCM [9].

Enfin, l'imputation mise en œuvre pour résoudre le problème de la non-réponse partielle a dû faire face à deux problèmes principaux :

— L'imputation d'actifs financiers par exemple suppose de prendre en compte leur forte corrélation, et il a fallu utiliser des techniques de *hot deck* pour simuler les valeurs manquantes en respectant cette structure.

— La plupart des questions demandaient une réponse en tranche (par exemple, un montant épargné compris entre 1 500 euros et 4 500 euros) et il était nécessaire pour l'analyse statistique de simuler une valeur comprise dans cet intervalle. Pour cela, des modèles économétriques ont été utilisés en respectant les deux contraintes d'appartenance à l'intervalle et de cohérence avec le patrimoine total donné par le ménage.

2.4.2 Calcul de la représentativité

Nous souhaitons calculer les R-indicateurs et les R-indicateurs partiels relatifs à l'avancement de la collecte afin d'étudier l'évolution de la représentativité au cours de l'avancement de l'enquête. Le calcul des R-indicateurs et R-indicateurs partiels reposant sur la connaissance des probabilités de réponse, qui ne sont évidemment pas connues, il faut donc les estimer. Pour cela, on utilise des variables auxiliaires (âge, type de ménage, type

9. Tronc Commun des Ménages, un questionnaire posé en début de la plupart des enquêtes de l'INSEE visant à connaître la structure du ménage, les caractéristiques des personnes y habitant de manière permanente ou occasionnelle et les liens entre ces personnes.

de logement...) et un modèle de type logit, de manière similaire à celui mis en œuvre lors du traitement de la non-réponse de l'enquête. Notons bien ici que, du fait du plan de sondage de l'enquête, il faut bien considérer les deux échantillons séparément, car la correction de la non-réponse n'est pas la même et n'implique pas les mêmes variables.

R-indicateurs. On souhaite ici calculer les R-indicateurs tout au long de la collecte, tous les mois, afin d'étudier l'avancement de la représentativité. Cela nous servira également de base de comparaison pour des simulations ultérieures. Nous obtenons les résultats suivants pour les R-indicateurs totaux (Figure 6).

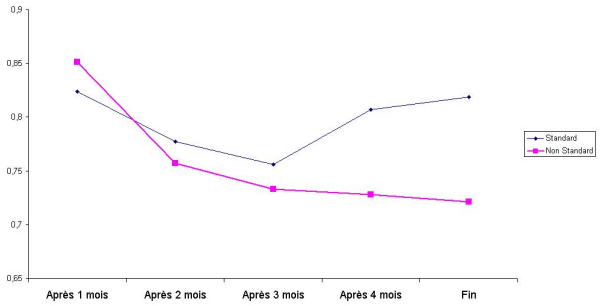

FIGURE 6 – R-indicateurs totaux dans l'enquête Patrimoine 2010

On remarque alors que la courbe des R-indicateurs pour l'échantillon non-standard (en gras) est décroissante, ce qui est comportement "classique" observé dans la plupart de la littérature : les inégalités de collecte entre les groupes ont tendance à croître au fur et à mesure de la progression de l'enquête. En revanche, celle relative à l'échantillon standard (en pointillés) décroît en début de collecte pour redevenir croissante plus tard. L'analyse de cette tendance n'est pas tout à fait claire, cependant un phénomène qui pourrait expliquer ce point d'inflexion de la courbe du R-indicateur de l'échantillon standard est la libération d'un échantillon de réserve qui a eu lieu aux alentours du troisième mois.

Calcul des R-indicateurs partiels. On doit donc essayer d'interpréter cette évolution (ainsi que celle concernant l'échantillon non-standard) en étudiant les R-indicateurs partiels pour déterminer les groupes sous-représentés en milieu de collecte afin de voir si cette sous-représentativité se résorbe, ou au contraire s'amplifie, en cours de collecte. Le tableau en Figure 7 compile les valeurs des R-indicateurs partiels inconditionnels pour les variables du modèle de non-réponse et leurs modalités.

On remarque donc, pour une grande partie des modalités, que le R-indicateur partiel conditionnel croît puis décroît en valeur absolue tout au long de l'enquête, ce en quoi il suit le R-indicateur. Pour certaines variables, par exemple la présence de revenus importants lié au patrimoine, la valeur du R-indicateur partiel conditionnel est très faible ce qui signifie que cette variable ne joue pas un rôle important dans le phénomène de représentativité et n'est donc pas à étudier dans le cadre de la priorisation. Considérons des variables pertinentes (pour ce critère). Comment expliquer l'évolution des R-indicateurs partiels inconditionnels ?

Variable	1 mois	2 mois	3 mois	4 mois	Fin
Strate	3.52	6.85	8.65	4.48	3.05
Agriculteurs	-0.14	1.07	1.40	1.12	0.92
Ages	-0.74	-3.03	-4.61	-1.78	-1.55
Indépendants	-1.67	-2.55	-2.60	-1.60	-0.89
Revenus pat.	-1.00	-1.49	-1.67	-0.36	0.07
Cadres	-2.01	-3.01	-3.16	-2.55	-1.50
Autres	2.00	4.34	5.67	2.53	1.734
Surface	1.40	1.99	1.94	3.44	4.16
Très petite surface	-0.65	-0.86	-0.82	-1.56	-2.19
Petite surface	-0.71	-1.21	-1.20	-2.00	-2.10
Surface moyenne	-0.31	-0.70	-0.70	-1.40	-1.74
Grande surface	0.79	0.88	0.82	1.05	1.22
Très grande surface	0.15	0.66	0.68	1.42	1.85
Plus grande surface	-0.55	-0.30	-0.21	0.54	0.43
Type de logement	1.07	1.95	1.93	3.14	3.17
Appartement	-0.82	-1.48	-1.47	-2.39	-2.42
Maison	0.69	1.26	1.25	2.03	2.05
Revenus Patrimoine	2.29	3.34	3.70	1.70	1.31
Revenus très faibles	0.91	1.31	1.55	0.50	0.22
Revenus faibles	-0.12	-0.12	-0.25	0.35	0.56
Revenus modérés	-1.29	-1.64	-2.17	-0.70	-0.45
Revenus importants	-0.99	-1.73	-1.85	-1.02	-0.65
Revenus très importants	-1.33	-1.94	-1.76	-1.00	-0.85
Revenus Activités	1.81	2.63	2.78	2.91	2.65
Revenus très faibles	0.26	-0.68	-1.07	-1.75	-1.95
Revenus faibles	0.29	0.35	0.03	0.14	-0.18
Revenus modérés	0.31	1.39	1.54	1.61	1.37
Revenus importants	0.32	0.33	0.79	0.937	1.05
Revenus très importants	-1.71	-2.07	-1.90	-1.40	-0.44
HLM	1.49	1.41	1.69	0.90	0.50
Autre	-0.59	-0.56	-0.66	-0.36	-0.20
HLM	1.37	1.30	1.55	0.83	0.46
Gros Revenus Act	1.57	2.40	2.41	1.76	1.37
Non	0.20	0.30	0.30	0.22	0.17
Oui	-1.56	-2.39	-2.39	-1.74	-1.36
Âge	0.19	1.83	3.17	1.50	1.52
18-35	-0.09	0.47	1.05	-0.29	-0.45
35-60	-0.12	1.34	2.20	1.27	1.29
60+	0.13	-1.15	-2.03	-0.73	-0.66
Type de ménage	1.48	3.13	4.12	3.86	3.89
Complet	0.72	0.02	0.19	0.23	0.36
Couple	-0.36	-0.52	-1.09	0.47	0.66
Famille	0.85	2.40	3.19	2.47	2.36
Parent seul	0.10	0.25	0.38	0.00	-0.11
Seul	-0.89	-1.93	-2.32	-2.92	-3.00

FIGURE 7 – R-indicateurs partiels inconditionnels de l'échantillon standard

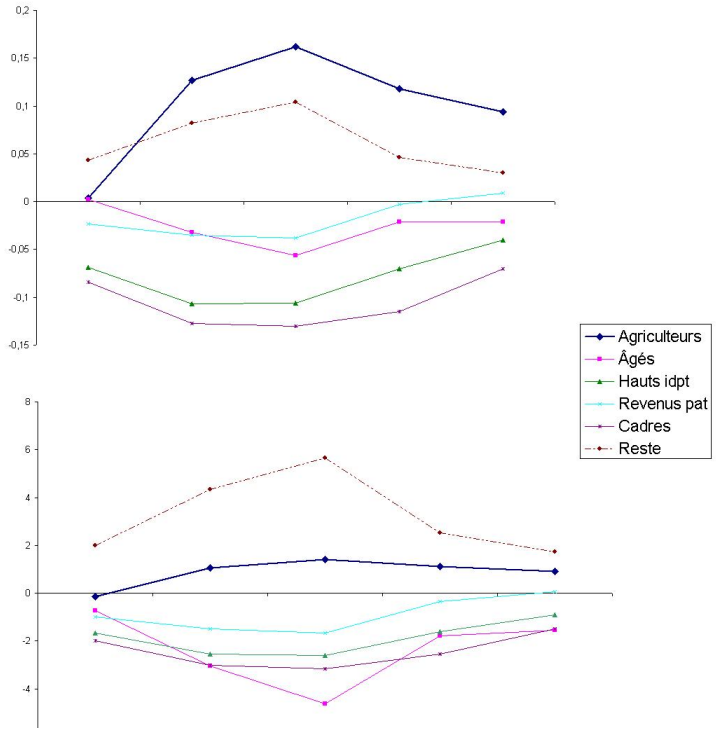

FIGURE 8 – Comparaison entre l'écart au taux de réponse moyen (en haut) et le R-indicateur partiel inconditionnel.

Pour cela, nous allons comparer l'évolution des R-indicateurs partiels inconditionnels et des taux de réponse de la population. Par exemple intéressons nous à la variable *strate* ayant servi à réaliser le plan de sondage et qui est très explicative de la représentativité. Lorsque l'on trace, en Figure 8, la courbe d'écart au taux de réponse moyen dans chaque strate, on remarque que :

- Les courbes sont assez proches en ce qui concerne les écarts positifs et négatifs (cohérence avec la notion de sous et sur-représentativité) et en ce qui concerne les variations.
- En revanche, certains groupes ayant un écart marqué de taux de réponse n'ont pas forcément un R-indicateur partiel inconditionnel très éloigné de 0. Par exemple, la courbe bleue (en gras) de la Figure 8 qui correspond à la strate des agriculteurs montre un fort écart positif du taux de réponse mais le R-indicateur partiel inconditionnel reste plus faible que celui de la catégorie "autres" (courbe en pointillés), qui a un taux de réponse plus faible. [10]

Calcul des R-indicateurs partiels conditionnels. Intéressons-nous aux R-indicateurs partiels conditionnels qui modélisent les effets de structures existant entre les variables et pouvant avoir un impact indirect sur la représentativité. Or, pour effectuer les calculs, il faut considérer toutes les modalités possibles de l'ensemble des variables exceptée celle considérée, ce qui peut représenter plus d'un million de cas. Étudier la dispersion des propensions à répondre dans chacun de ces sous-groupes n'a que de peu de sens, c'est

10. Dans l'exemple simple du paragraphe précédent, le taux de réponse et le R-indicateur partiel inconditionnel étaient directement liés par une relation simple, le modèle de non-réponse n'ayant que 2 variables explicatives. Ce n'est évidemment pas le cas ici.

pourquoi nous allons participer à diminuer le degré de liberté du modèle et donc le nombre de modalités par deux moyens :

1. Nous allons regrouper les modalités des variables "Taille d'unité urbaine" et "Région de gestion", qui avaient respectivement 9 et 21 modalités, de telle sorte que le nombre de modalité devienne de 5 et de 7.
2. Lors du calcul des probabilités de réponse à l'instant t, nous allons utiliser une approche *stepwise* de la régression logistique afin de ne sélectionner que les variables les plus pertinentes. Cela a pour effet de ne pas avoir à s'intéresser à certaines variables et donc diminuer le nombre de modalités.

Cette méthode a été appliquée aux résultats obtenus après 3 mois de collecte uniquement et nous obtenons les résultats présentés dans le tableau en Figure 9 (avec normalisation afin d'avoir des R-indicateurs partiels compris entre 0.1 et 10, pour rendre le tableau plus facilement lisible).

Remarquons tout d'abord que les deux modifications apportées pour permettre le calcul des R-indicateurs partiels conditionnels n'ont pas ou peu modifié les valeurs des R-indicateurs partiels inconditionnels, car l'écart est toujours de l'ordre de quelques %. Cette remarque permet de supposer que le calcul des R-indicateurs est robuste vis-à-vis du modèle utilisé pour l'estimation des propensions à répondre.

Pour qu'une variable soit considérée comme pertinente pour l'analyse de la représentativité, il convient que ses R-indicateur partiel conditionnel et inconditionnel soient grands. Ainsi tant la strate que le type de ménage peuvent être considérés comme importants, alors que par exemple le fait d'avoir de gros revenus d'activité, ou même l'âge, le sont moins. Leur impact sur la représentativité est dû à un effet structurel qui s'explique assez bien dans le cas de l'âge : en effet, prioriser les personnes de plus de 60 ans comme cela est conseillé par la valeur négative du R-indicateur partiel inconditionnel de cette modalité revient finalement à se concentrer sur la strate "Âgés".

Échantillon non standard. Si l'on s'intéresse également à l'échantillon non-standard de l'enquête Patrimoine, i.e celui des hauts patrimoines qui sont sur-représentés, on peut renouveler le calcul après 3 mois de collecte en utilisant les mêmes techniques de réduction du nombre de modalités. On obtient alors les résultats compilés dans le tableau en Figure 10 (avec normalisation).

On remarque que concernant l'échantillon non standard la strate joue un rôle moins primordial que dans l'autre échantillon, les variables explicatives strate et présence de très gros revenus du patrimoine étant structurellement liées comme le montre le faible R-indicateur partiel conditionnel. Néanmoins vu le faible nombre de variables explicatives de la représentativité (en excluant toujours la région), on pourra prioriser des ménages de personnes seules ou de parents seuls, ceux ayant de très hauts revenus du patrimoine et les ménages riches vivant en zone urbaine ; l'alternative étant de n'avoir une stratégie de priorisation uniquement pour les ménages constitués de personnes seules et de parents seuls.

En conclusion, les modalités correspondant aux sous-groupes à prioriser à 3 mois de collecte sont à choisir parmi celles de la Figure 11.

Variable	Inconditionnel	Conditionnel
Strate	8.67	1.88
Agriculteurs	1.42	
Ages	-4.62	
Indépendants	-2.60	
Revenus pat.	-1.66	
Cadres	-3.15	
Autres	5.67	
Type de ménage	4.08	0.52
Complet	0.19	
Couple	-1.06	
Famille	3.16	
Parent seul	0.39	
Seul	-2.31	
Gros Revenus Act.	2.40	0.18
Non	0.30	
Oui	-2.38	
Âge	3.12	0.25
18-35	1.04	
35-60	2.16	
60+	-1.99	
HLM	1.67	0.26
Autre	-0.66	
HLM	1.54	
Revenus Patrimoine	3.71	0.30
Revenus très faibles	1.56	
Revenus faibles	-0.27	
Revenus modérés	-2.19	
Revenus importants	-1.85	
Revenus très importants	-1.74	
Surface	1.78	0.21
Très petite surface	-0.85	
Petite surface	-1.04	
Surface moyenne	-0.64	
Grande surface	0.63	
Très grande surface	0.74	
Plus grande surface	-0.13	

FIGURE 9 – R-indicateurs partiels de l'échantillon standard à 3 mois.

2.4.3 Modélisation d'une collecte type CVS13.

Impact sur la représentativité On souhaite s'intéresser au comportement des méthodes de priorisation lorsque des problèmes de collecte surviennent. Pour cela, nous utiliserons les données de l'enquête CVS [11] 2013 comme exemple d'une telle collecte. L'application directe qui aurait consisté à faire en sorte que les taux de collecte dans chaque ZAE de

11. Cadre de Vie et Sécurité.

Variable	Inconditionnel	Conditionnel
Strate	7.39	0.275
Riches urbains	-0.98	
Pat. mobilier	0.52	
Pat. immobilier	0.30	
Autres	7.30	
Type de ménage	7.89	0.337
Complet	3.22	
Couple	7.13	
Famille	0.50	
Parent seul	-0.29	
Seul	-0.85	
Très Gros Revenus Pat.	6.66	0.256
Non	6.25	
Oui	-2.29	

FIGURE 10 – R-indicateurs partiels de l'échantillon non standard à 3 mois.

Échantillon standard	Échantillon non-standard
Âgés	Individus seuls
Indépendants	Parents seuls
Cadres	Riches urbains
Revenus du patrimoine	Très gros revenus pat.
Individus seuls	
Couples sans enfants	

FIGURE 11 – Modalités à prioriser.

notre enquête Patrimoine modifiée soient égaux à ceux de CVS13 nécessitant la création de plus de 2500 FAs, nous utilisons une autre méthode. Celle-ci consiste à étudier le taux de réponse moyen sur les enquêtes CVS11 et CVS12 et à calculer le ratio d'évolution entre celles-ci et CVS13, afin d'appliquer cette évolution à l'enquête Patrimoine en milieu de collecte. Ainsi, nous gardons à chaque fois un certain pourcentage p des FA enquêtées, les cas où p dépasse 100% étant ignorés car marginaux. La dispersion des p dans les ZAE est représentée par la Figure 2.4.3 et montre bien que la plupart des ZAE voient leur effectif diminuer de moins de 33% ou rester le même, les cas de fortes diminutions étant restreints à environ une quinzaine de ZAE.

Une fois ces nouvelles données de collecte simulées, on recalcule les R-indicateurs totaux et partiels sur cette nouvelle base. On obtient un R-indicateur total de 0.7610 pour l'échantillon standard, et un R-indicateur de 0.6925 pour l'échantillon non-standard. Les R-indicateurs partiels conditionnels et inconditionnels sont assez proches de ceux obtenus dans le cadre de la véritable collecte, ce qui est logique vu que le mécanisme simulé est uniforme. On en déduit que les groupes à prioriser sont assez similaires à ceux précédemment. La sélection se fait en regroupant tous les ménages correspondant à l'une au moins des modalités de la Figure 13.

On considère ensuite plusieurs scénarios.

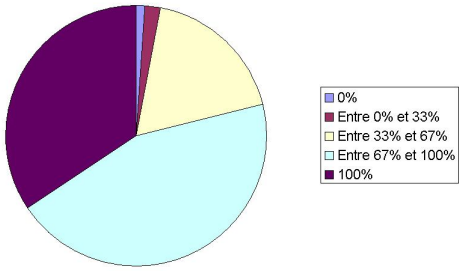

FIGURE 12 – Dispersion des ratios de sélection des FAs dans les ZAE.

Échantillon standard	**Échantillon non-standard**
Âgés	Riches urbains
Cadres	Appartements
Individus seuls	

FIGURE 13 – Modalités à prioriser après les problèmes de collecte.

1. On peut réaliser la priorisation dans la totalité des ZAE.
2. On peut se limiter à celles dont l'impact (selon le modèle) des NCEE a conduit à une perte d'au moins 33% des FA, ce qui fait se concentrer sur environ 21% des ZAE.
3. On peut aussi se limiter à celles qui ont été les plus atteintes et dont la perte est au moins de 66%, ce qui restreint le champ à environ 3% des ZAE.

Nous générons 100 scénarios de baisse de taux de réponse et considérons pour chacun des cas l'étude du R-indicateur moyen après enquête de k fiches-adresses supplémentaires, qu'elles soient choisies par la méthode de priorisation ou aléatoirement. On suppose que ce choix de FA n'entraîne pas leur réponse automatique, aucun effort particulier n'étant fait dans ce but : la modélisation consiste à les considérer comme répondantes si c'était le cas à la fin de l'enquête Patrimoine 2010. Intéressons-nous tout d'abord uniquement à l'échantillon standard et aux deux premiers scénarios uniquement. Dans chacun des tableaux de la Figure 14, pour un nombre de FA supplémentaires donné, la valeur maximale du R-indicateur a été mise en gras pour permettre une lecture plus rapide.

On constate tout d'abord que la priorisation permet d'obtenir dans tous les cas, hormis celui atypique de l'enquête de 10 FA dans toutes les ZAE, un meilleur R-indicateur que la sélection aléatoire, ce qui cohérent avec le but souhaité. On remarque également que le R-indicateur augmente lorsque le nombre de FA supplémentaires augmente, toujours en excluant le scénario atypique. De même, l'écart entre R-indicateur du cas priorisé et R-indicateur du cas aléatoire augmente, passant de le scénario 2 de 0.2% d'augmentation pour 1 FA supplémentaire à 1.8% pour 10. Enfin, lorsque l'on compare les deux scénarios, le deuxième scénario permet toujours d'obtenir un meilleur R-indicateur, alors que le nombre de FA à réallouer est nettement inférieur : en ne travaillant que sur les 20% des ZAE les plus atteintes par la baisse de taux de réponse, on divise par 5 la charge de travail des enquêteurs.

Scénario 1.

Nombre de FA	Priorisé	Aléatoire
5	**0.7768**	0.7686
10	0.7585	**0.7594**

Scénario 2.

Nombre de FA	Priorisé	Aléatoire
5	**0.7812**	0.7731
10	**0.7995**	0.7851

FIGURE 14 – R-indicateurs de l'échantillon standard après priorisation.

Si l'on inclut maintenant l'échantillon non-standard à l'étude, il convient de le prioriser en priorité : en effet celui-ci consistant en une sur-représentation des hauts patrimoines, il apporte une meilleure précision aux estimations des variables d'intérêt. On va utiliser une méthode de priorisation en deux temps : on commence par prioriser au sein de l'échantillon non-standard les ménages sélectionnés puis s'il reste des FA supplémentaires disponibles, on les affecte à des ménages sélectionnés parmi les ménages de l'échantillon standard. Les résultats obtenus sont en Figure 15.

Scénario 1.

	Échantillon std		**Échantillon non-std**	
Nombre de FA	Priorisé	Aléatoire	Priorisé	Aléatoire
5	0.7666	**0.7672**	**0.7015**	0.6840
10	0.7495	**0.7696**	**0.7034**	0.6770

Scénario 2.

	Échantillon std		**Échantillon non-std**	
Nombre de FA	Priorisé	Aléatoire	Priorisé	Aléatoire
5	**0.7756**	0.7706	**0.7118**	0.6972
10	**0.7906**	0.7809	**0.7255**	0.7027

Scénario 3.

	Échantillon std		**Échantillon non-std**	
Nombre de FA	Priorisé	Aléatoire	Priorisé	Aléatoire
5	0.7649	**0.7667**	**0.7112**	0.6973
10	**0.7725**	0.7706	**0.7200**	0.7040

FIGURE 15 – R-indicateurs après différents scénarios de priorisation.

Intéressons-nous tout d'abord à l'échantillon non-standard, qui est celui qui est favorisé par les deux mécanismes de sélection des FA. On remarque déjà que la priorisation

permet d'obtenir dans tous les cas un meilleur R-indicateur que la sélection aléatoire, ce qui cohérent avec le but souhaité. On remarque également que dans les 3 scénarios considérés, augmenter le nombre de FA augmente le R-indicateur, sans effet de saturation, i.e de sur représentation des populations priorisées : pour la sélection aléatoire cet effet est moins clair, en particulier dans le scénario 1. Lorsque l'on compare les différents scénarios, on remarque que le scénario 2 est le plus favorable à l'augmentation du R-indicateur de l'échantillon non-standard.

En ce qui concerne l'échantillon standard, l'analyse est plus complexe car, n'étant qu'un échantillon de secours pour la sélection des nouvelles FA, l'effet est forcément limité. Ainsi, la sélection par priorisation n'offre pas toujours un meilleur R-indicateur que la sélection aléatoire (mais jamais très inférieur en revanche), et augmenter le nombre de FA à prioriser n'améliore pas forcément le R-indicateur : on constate même une baisse nette de celui-ci par effet de saturation lorsque l'on priorise 10 FA dans toutes les ZAE (cas hautement improbable). Lorsque l'on compare les différents scénarios, on voit alors que le scénario 2 est également le plus favorable à l'augmentation du R-indicateur de l'échantillon non-standard.

On peut donc dire en conclusion que le scénario 2 consistant à prioriser des FA dans les 21% des ZAE les plus affectées est celui à privilégier. On remarque alors que plus on peut y réaliser de FA, plus les R-indicateurs des deux échantillons seront grands : il reste à réaliser un compromis coût/représentativité, en particulier en prenant en compte les problématiques de quotité de travail limitée et de réaffectation des enquêteurs.

Impact sur les variables d'intérêt. Nous allons reproduire pour chacun des scénarios, et dans le cadre de la priorisation et de la sélection aléatoire, le traitement des données afin d'obtenir les valeurs des variables d'intérêt y (à sélectionner). Pour cela, il faut renouveler le traitement de la non-réponse totale ainsi que le calage. Nous allons étudier la dispersion des $\hat{y}$ obtenus en reproduisant l'expérience 100 fois et déterminer si la priorisation permet une variabilité plus faible des résultats. Ensuite, nous procéderons en deux temps :

1. On considèrera tout d'abord que l'on ne peut intervenir qu'en fin de collecte, et que la priorisation a lieu après la collecte normale.
2. Ensuite, on s'intéressera à un scénario de priorisation en deux vagues, avec poursuite de la collecte usuelle dans les ZAE non concernées.

Nous nous intéressons aux variables d'intérêt suivantes : patrimoine brut du ménage, patrimoine net, patrimoine financier, patrimoine immobilier et patrimoine professionnel.

Les premiers résultats sont assez peu encourageants : une explication possible est que trop peu de modalités sont choisies pour la priorisation. En effet, rajouter des modalités permet de réduire le nombre de ZAE dans laquelle la priorisation entraînera un défaut de collecte, le nombre de FA disponible étant inférieur à celui souhaité. On considère alors les modalités à prioriser de la Figure 16.

Priorisation en fin de collecte. En utilisant ce nouveau jeu de modalités, et après utilisation de la méthode des scores (cf le traitement post-collecte de l'enquête) sur les probabilités estimées par un modèle logistique pour créer des groupes de repondération.

Échantillon standard	Échantillon non-standard
Âgés	Riches urbains
Cadres	Appartements
Indépendants à hauts revenus	Individus seuls
Individus seuls	
Appartements	

FIGURE 16 – Modalités finales à prioriser.

Ajout de 5 FA.	Brut	Net	Financier	Immobilier	Professionnel
Sans traitement	256890	226930	51452	162186	30820
Écart-type	*3160*	*3082*	*618*	*928*	*2779*
Aléatoire	257183	227210	51350	162282	31119
Écart-type	*2708*	*2644*	*571*	*855*	*2425*
Priorisation	257366	227445	51501	162146	31286
Écart-type	*2795*	*2695*	*650*	*876*	*2359*
Données Pat10	259000	229300	50800	160500	35300

Ajout de 10 FA.	Brut	Net	Financier	Immobilier	Professionnel
Sans traitement	256890	226930	51452	162186	30820
Écart-type	*3160*	*3082*	*618*	*928*	*2779*
Aléatoire	257561	227603	51446	162269	31396
Écart-type	*2323*	*2136*	*568*	*922*	*1947*
Priorisation	257533	227652	51596	162225	31292
Écart-type	*2175*	*2056*	*615*	*861*	*1732*
Données Pat10	259000	229300	50800	160500	35300

FIGURE 17 – Variances des estimateurs des variables d'intérêt avec priorisation en fin de collecte.

Ensuite, on applique un calage. Nous obtenons les résultats de la Figure 17.

On remarque que le biais estimé ici à partir des résultats publiés de l'enquête Patrimoine 2010 a diminué par rapport aux simulations précédentes. En ce qui concerne l'impact de la priorisation sur la variance, on ne remarque pas d'influence significative lorsque le nombre de FA rajoutées est de 5. En revanche, la méthode est légèrement meilleure pour la plupart des variables d'intérêt lorsque que l'on rajoute 10 FA.

Priorisation en cours de collecte. Une possibilité est, certes de ne prioriser que dans les 20 % de ZAE les plus affectées par les NCEE, mais de considérer que la collecte continue normalement dans les autres ZAE. On conserve les modalités à prioriser de la Figure 16. Les résultats obtenus sont alors meilleurs, ce qui s'explique principalement par le fait que la collecte est poursuivie dans la plupart des zones, ce qui augmente le taux de réponse moyen. Notons qu'ici la collecte sans traitement n'est plus la même qu'auparavant : il s'agit de comparer cette priorisation avec collecte continuée à la collecte finale de l'enquête, afin de voir si la priorisation dans certaines zones offre une meilleure précision

que de ne pas intervenir. On constate que c'est bien le cas, et qu'augmenter le nombre de FA priorisées réduit la variance des estimations.

Considérer que la collecte continue normalement dans les autres ZAE est une hypothèse un peu forte ; en effet, les enquêteurs devant se reporter sur les ZAE prioritaires, ils ne pourront pas réaliser la même charge de travail que dans la situation classique. On peut alors faire l'hypothèse que l'enquêteur réalise 75% de la collecte usuelle dans sa zone. Ce ratio de 75% peut s'expliquer par le fait que étant donné que l'on priorise 20% des ZAE, il faut qu'environ un enquêteur sur 4 parmi ceux rattachés aux autres ZAE, moins affectées, intervienne sur une autre ZAE : c'est suffisant car l'effort de priorisation concerne une dizaine de FA, sachant qu'il se fait sur un quart de la période de collecte et que chaque enquêteur a une quarantaine de fiches-adresse qui lui sont attribuées. Prendre en compte les données géographiques (i.e déterminer quels enquêteurs seraient à mobiliser) étant trop complexe pour une première simulation, on se limite à cette hypothèse. Les résultats de la Figure 18 sont logiquement moins bons que dans le cas improbable d'une collecte complète.

Ajout de 5 FA.	Brut	Net	Financier	Immobilier	Professionnel
Sans traitement	259843	229484	51314	163671	32456
Écart-type	*2175*	*2179*	*536*	*699*	*1904*
Aléatoire	260200	229796	51400	163696	32653
Écart-type	*2436*	*2270*	*539*	*766*	*1983*
Priorisation	260114	229784	51360	163531	32765
Écart-type	*2521*	*2337*	*552*	*741*	*2059*
Données Pat10	259000	229300	50800	160500	35300

Ajout de 10 FA.	Brut	Net	Financier	Immobilier	Professionnel
Sans traitement	259843	229484	51314	163671	32456
Écart-type	*2175*	*2179*	*536*	*699*	*1904*
Aléatoire	260089	229698	51309	163641	32676
Écart-type	*2327*	*2132*	*589*	*731*	*1870*
Priorisation	260377	230038	51451	163544	32930
Écart-type	*1967*	*1771*	*592*	*676*	*1504*
Données Pat10	259000	229300	50800	160500	35300

FIGURE 18 – Variances des estimateurs des variables d'intérêt avec priorisation en cours de collecte.

On remarque que lorsque l'on rajoute uniquement 5 FA, on ne fait pas mieux qu'une collecte complète ; par contre pour 10 FA, on obtient des résultats concluants. Cela peut s'expliquer par l'évolution des taux de réponse. On note X le taux de poursuite de la collecte dans les autres ZAE, et N le nombre de FA priorisées. Le tableau de la Figure 19 répertorie le nombre de répondants dans différents scénarios.

On voit bien que pour ce qui concerne la collecte poursuivie à 75%, lorsque l'on rajoute uniquement 5 FA la baisse du taux de réponse est conséquente (de l'ordre de 3%), ce qui

X	N	Collecte complète	Priorisation
75%	5	10600	10300
	10	10600	10500
100%	5	10600	11000
	10	10600	11200

FIGURE 19 – Nombre de répondants selon le scénario retenu.

n'est plus le cas pour un rajout de 10 FA. Cela peut expliquer les résultats obtenus. En résumé, les écarts-type des scénarios de priorisation retenus sont compilés en Figure 20.

Écarts types	Brut	Net	Fin	Immo	Prof
5 FA en fin de collecte	*2795*	*2695*	*650*	*876*	*2359*
10 FA en fin de collecte	*2175*	*2056*	*615*	*861*	*1732*
10 FA - collecte continuée à 75%	*1967*	*1771*	*592*	*676*	*1504*

FIGURE 20 – Écarts-types obtenus pour les différents scénarios de priorisation.

3 Comment prioriser en pratique ?

3.1 Une priorisation par vagues

3.1.1 Différences CAPI/CATI

Les travaux de validation du principe de priorisation par simulation effectués dans les deux parties précédentes supposent qu'il est possible d'effectuer la priorisation "à la volée". Il s'agirait en effet de pouvoir à tout instant signaler aux enquêteurs quelle fiche doit être enquêtée en priorité. Cela serait possible pour une collecte téléphonique : la méthode de priorisation pour les collectes CATI est appliquée par Statistique Canada dans le cadre de l'IBSP [12] [11],[16],[17].

Une expérience a toutefois été menée dans le cadre d'une collecte CAPI [13], lors de l'enquête Logement 2013 (voir [14]), et qui nous a permis de mieux comprendre les mécanismes à l'œuvre lors de l'organisation d'une collecte sur ce mode. Dans le contexte d'une enquête où le questionnaire est posé en face-à-face, les enquêteurs doivent effectuer un repérage de l'adresse et plusieurs tentatives de contact avant l'entretien. Le repérage permet d'identifier les éventuels hors-champ (logement identifié comme résidence prinicpale dans la base de sondage, mais dont le statut a changé entre-temps), de signaler au ménage enquêté le passage de l'enquêteur ainsi que le caractère obligatoire (ou non) de l'enquête. Cette phase de l'enquête doit être effectuée en début de période de collecte, et précède donc généralement la collecte de la première FA. La gestion par l'enquêteur des relances à effectuer est plus délicate (un enquêteur CAPI gère beaucoup moins d'enquêtés qu'un enquêteur CATI qui peut puiser à volonté dans le logiciel de gestion), et la propension à répondre dépend moins de l'effort de relance que par téléphone de toute façon.

L'organisation de la priorisation "à la volée" sur un mode similaire au CATI est sujet à un dernier écueil : le calcul effectif des probabilités de sélection, et donc des poids de sondage associés. Dans le cadre d'une collecte par téléphone, la propension à répondre de chaque unité dépend clairement de l'effort de relance qui est appliqué. En effet, on peut parfaitement conceptualiser la réponse au questionnaire de l'unité i comme la réalisation d'une expérience de Bernoulli à probabilité (cachée) p_i, répétée autant de fois que l'enquêteur tente de joindre l'enquêté par téléphone. L'utilisation des modèles classiques de correction de la non-réponse est parfaitement justifié dans ce cadre, car il s'agit de mesurer au mieux le paramètre p_i pour chaque unité ou groupe d'unité pour lesquels les efforts de relance ont été similaires.

Dans le cas d'une collecte CAPI, où l'enquêteur se rend au domicile de l'enquêté, afin de lui administrer le questionnaire en face-à-face, la propension à répondre ne semble pas répondre à la même logique. Le ménage échantillonné est averti du passage de l'enquêteur par une lettre-avis officielle, mentionnant parfois le caractère obligatoire au regard de la loi de réponse à l'enquête. La prise de rendez-vous peut s'effectuer à l'initiative de l'enquêteur ou de l'enquêté, mais elle est libre tout au long de la collecte, laissant ainsi quelques semaines, voire mois pour trouver une date propice. Si la probabilité de réponse

12. Integrated Business Statistics Program
13. Collecte Assistée par Informatique.

dépend bien in fine de l'effort via le nombre de relances ([9]), on ne peut considérer qu'il peut s'agir du meilleur prédicteur univarié.

On voit que la probabilité p_i de réponse de chaque unité échantillonnée ne dépend vraisemblablement pas aussi directement d'un effort de relance de la part de l'enquêteur que dans le cas du CAPI. Ainsi, se reposer uniquement sur les efforts de relance pour construire les poids de sondage pourrait amener à construire des estimateurs présentant les mêmes biais que les estimateurs typiques de correction de la non-réponse reposant sur un modèle mal spécifié.

3.1.2 Analyse d'une collecte CAPI

Afin de mieux comprendre le déroulement typique d'une collecte CAPI à l'INSEE, on étudie sommairement le déroulement de la collecte de l'enquête EPIC [14] en Bourgogne, sur une durée de 100 jours entre fin septembre et fin décembre 2013. Le graphe en figure 21 montre la progression de la collecte (courbe lissée, et en pointillés, droites des régressions linéaires pour trois phases de collecte).

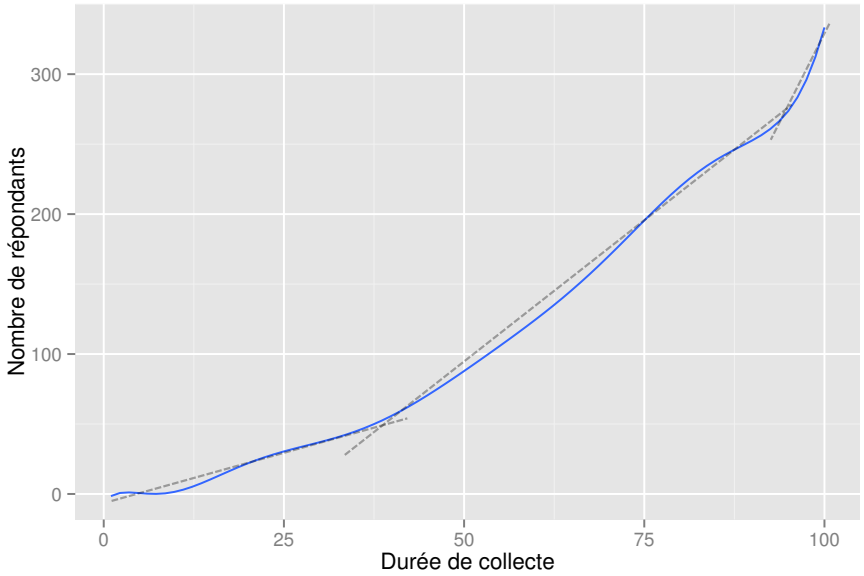

FIGURE 21 – Progression de la collecte de l'enquête EPIC en Bourgogne.

Le nombre de FA réalisées (et donc de répondants) n'est pas linéaire en fonction du temps. On peut identifier trois phases :

1. **1ère phase, "repérage" :** Les enquêteurs effectuent le repérage des logements échantillonnés et la distribution des lettres-avis. Quelques entretiens sont réalisés, mais le taux d'avancement reste faible.
2. **2ème phase, "régime permanent" :** Période la plus longue proportionnellement au déroulement de la collecte. Le taux d'entretiens réalisés est sensiblement constant, la prise de rendez-vous semble s'organiser de façon assez régulière.

14. Étude des Parcours Individuels et Conjugaux

3. **3ème phase, "accélération" :** Les rendez-vous pris avec les ménages les plus difficiles à contacter s'accumulent à proximité de la date de fin de collecte. Le taux d'entretiens réalisés augmente à l'approche de la date de clôture.

Il est également à noter que la fin de la collecte étudiée ici coïncide avec les vacances d'hiver, période pendant laquelle une grande majorité de ménages sera indisponible. En pratique, il faut souvent composer avec des périodes de l'année moins propices à la collecte d'une enquête CAPI.

3.1.3 Plusieurs vagues de collecte

Une approche possible consiste donc à réaliser la procédure de priorisation (calcul des R-indicateurs, sélection des groupes à prioriser) à la fin de chacune des vagues. Il ne s'agit pas alors de relance à proprement parler, mais d'une adaptation du plan d'échantillonnage au cours de la collecte : les fiches-adresses des ménages sous-représentées sont en nombre supérieur à ce qui était prévu, les autres étant en plus petit nombre. Dans le cadre d'une enquête de longue durée en plusieurs vagues, il est ainsi possible de répéter plusieurs fois ce schéma de priorisation.

L'organisation en vagues nécessite de prendre en compte l'analyse de la collecte réalisée en 3.1.2. Construire la collecte priorisée en respectant le déroulement usuel d'une collecte CAPI semble essentiel au succès de l'opération. Il faut s'assurer en particulier que les trois phases de la collecte peuvent se dérouler normalement afin de maximiser la qualité comme la quantité des questionnaires administrés.

À l'INSEE, les DEM (Direction Enquêtes Ménages) pilotent la collecte effectuée par les enquêteurs et disposent d'une expérience terrain particulière, qu'il est important de valoriser. L'intégration des échantillons priorisés doit se faire en étroite collaboration avec ces experts, qui doivent être intégrés dans toutes les discussions concernant la priorisation d'enquêtes ménages

En particulier, il s'agit de se reposer sur la connaissance de ces experts pour organiser convenablement les vagues de collecte. Ainsi, selon les DEM consultées, le déroulement de toutes les collectes d'enquêtes ménages semblent correspondre au schéma décrit en 3.1.2, et ce quelle que soit leur durée du moment qu'elle dépasse un seuil minimum. En-deçà de ce seuil minimum, les trois phases de collecte n'ont pas le temps de se mettre en place, et les taux de collecte observés sont nettement inférieurs aux attentes. Il est finalement retenu à l'INSEE que **la durée minimale d'une vague doit être de 6 semaines**.

L'organisation actuelle des vagues priorisées à l'INSEE repose entièrement sur le dire d'expert, et l'expérience de management de la collecte sur le terrain. Il est toutefois tout-à-fait possible d'imaginer se doter d'un logiciel recueillant nombre de données utiles pour l'analyse de la collecte, et d'affiner à partir des données recueillies l'organisation de la collecte de manière à optimiser l'opération de priorisation. La logique de l'organisation pratique répondrait alors aux principes de l'analyse de données massives (big data), concordant avec les récents développements de la littérature à ce sujet (voir notamment [10] et 3.3).

3.2 À quel moment effectuer la priorisation ?

La technique de priorisation et l'algorithme CURIOS (voir l'article associé [2] pour une description précise de l'algorithme de priorisation utilisé dans le cadre des enquêtes ménages de l'INSEE) peuvent être mis en œuvre dès qu'une quantité suffisante d'informations est disponible pour obtenir un gain anticipé significatif. En théorie, même un taux d'avancement faible pourrait convenir, et toutes les collectes pourraient être découpées en vagues de 6 semaines, un échantillon adapté étant proposé à chaque fois. On pourrait également imaginer se servir de CURIOS avant même le début de la collecte, juste après le tirage de l'échantillon de première vague, en utilisant les données issues d'une précédente mouture de la même enquête.

Toutefois, la méthode de priorisation étant très récente, il convient d'adopter une attitude prudente à cet égard. C'est pourquoi un taux d'avancement de 75% est pour l'instant recommandé à l'INSEE avant de mettre en œuvre la procédure de priorisation. Il s'agit surtout d'éviter de déséquilibrer des échantillons sur la base de données dont l'évolution dans le temps n'est pas clairement anticipée. Dans le pire des cas, on peut imaginer un signal bruité qui se manifesterait incorrectement jusqu'à un taux d'avancement élevé, et auquel l'algorithme de tirage CURIOS serait très sensible. Les échantillons priorisés dans ce cas seraient déséquilibrés à mauvais escient, et l'erreur commise serait difficile à rattraper sur les dernières vagues de collecte, particulièrement si celles-ci coïncident avec des périodes moins propices (vacances, etc.). Il faudrait donc conduire une analyse détaillée des indicateurs utilisés par CURIOS, en particulier leur évolution temporelle, et également déterminer à quel point les échantillons priorisés sont sensibles au modèle utilisé. Ces deux axes constituent les sujets de recherche futurs autour de la priorisation à l'INSEE.

3.3 Utilisation des paradonnées

Les paradonnées sont des données relatives au déroulement de la collecte dans le cadre d'une enquête : identité de l'enquêteur, heure de passage, jour de la semaine, durée de l'entretien, etc. Ces données permettent une modélisation fine des comportements de réponse (voir par exemple [10]).

L'utilisation des paradonnées pour réaliser la priorisation permettrait une meilleure sélection des groupes à enquêter en priorité. Cependant, ces données sont encore difficilement accessibles pour les enquêtes ménages réalisées par l'INSEE ; les projets d'amélioration des processus de collecte et de remontée des questionnaires devraient permettres d'utiliser les paradonnées pour la priorisation.

Conclusion

Les R-indicateurs global et partiels présentés dans ce document sont des indicateurs de représentativité qui permettent d'agir pendant la collecte pour évaluer la représentativité. Ils peuvent être estimés facilement à l'aide des données d'enquête et permettent ainsi de sélectionner des groupes de ménages sous-représentés.

L'application de ces méthodes à l'enquête Patrimoine 2010 a permis de fournir un profil d'évolution de la représentativité au cours de la collecte, et en particulier d'étudier les différents groupes après 3 mois de collecte. Cette analyse a été utilisée dans les simulations qui ont suivi :

— Tout d'abord simuler l'abandon de la collecte dans certaines ZAE, pour simuler une priorisation dans celles-ci et étudier l'effet sur la représentativité.
— Ensuite, pour étudier l'impact des méthodes de priorisation dans le cadre d'une collecte dégradée telle celle de CVS13, que ce soit au niveau des R-indicateurs ou de la précision des estimations.
— Enfin, étudier un scénario de diminution de la taille des échantillons de 5% pour voir s'il est possible de compenser la baisse probable de précision en utilisant ces méthodes.

Ces trois jeux de simulations permettent plusieurs conclusions :

1. **La priorisation semble utile**. Lorsque l'on compare celle-ci avec une sélection aléatoire, elle permet d'obtenir une meilleure représentativité ainsi que, pour peu qu'elle concerne suffisamment de FA, une meilleure précision.
2. **Utiliser des méthodes de priorisation dans toutes les zones est contre-productif**, du moins tant qu'il est pas possible d'avoir une connaissance en direct permettant de modifier les groupes priorisables au fur et à mesure de la collecte.
3. Il est possible et probablement pertinent de prioriser tant **à la fin de la collecte**, si on trouve que certaines zones ont un taux de réponse final trop faible, que **pendant la collecte**, en continuant la collecte normalement (bien que nécessairement un peu moins efficacement) dans les autres zones.

Références

[1] Luiten A. and Schouten B. Tailored fieldwork design to increase representative household survey response : an experiment in the survey of consumer satisfaction. *Journal of the Royal Statistical Society*, 176 :169–189, 2013.

[2] Rebecq A. and Merly-Alpa T. Algorithme CURIOS et méthode de "priorisation" pour les enquêtes en face-à-face. Application à l'enquête Patrimoine 2014. *JMS*, 2015.

[3] Schouten B., Cobben F., and Bethlehem J. Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35(1) :107–121, juin 2009.

[4] Schouten B., Cobben F., and Bethlehem J. An adaptive data collection procedure for call prioritization. *submitted*, 2012.

[5] Schouten B., Bethlehem J.G., Beullens K.and Kleven O., Loosveldt G., Luiten A., Rutar K., Shlomo N., and Skinner C. Comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80 :382–399, 2012.

[6] Schouten B., Shlomo N., and Skinner C. Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27(2) :231–253, 2011.

[7] Rubin D. Inference and missing data. *Biometrika*, 63(3) :581–592, Dec. 1976.

[8] David Haziza and Jean-François Beaumont. On the construction of imputation classes in surveys. *International Statistical Review*, 75(1) :25–43, 2007.

[9] Steve Jakoubovitch. Estimation a priori de l'impact du nombre de contacts sur les taux de réponse. 2014.

[10] Frauke Kreuter. *Improving surveys with paradata : Analytic uses of process information*, volume 581. John Wiley & Sons, 2013.

[11] Fraser Mills, Serge Godbout, Keven Bosa, and Claude Turmelle. Multivariate selective editing in the integrated business statistics program. 2013.

[12] Shlomo N. and Schouten B. Theoretical properties of partial indicators for representative response. *Technical Report*, 176 :169–189, 2013.

[13] Shlomo N., Schouten B., and de Heij V. Designing adaptative survey design with R-indicators. *New Techniques and Technologies for Statistics Conference*, 2013.

[14] Antoine Rebecq. Heuristique branch-and-bound pour la sous-allocation et la réallocation. 2014.

[15] Natalie Shlomo, Chris Skinner, and Barry Schouten. Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142(1) :201 – 211, 2012.

[16] C Turmelle, S Godbout, and K Bosa. Methodological challenges in the development of statistics canada's new integrated business statistics program. 2012.

[17] Lingyun Zhu and Serge Godbout. Using quality indicators to manage collection and editing in business surveys. 2011.