Algorithme curios et méthode de priorisation pour les enquêtes en face-à-face - application à l'enquête Patrimoine 2014

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

(*) DMCSI, INSEE

Résumé

Dans un contexte de dégradation des taux de collecte dans les enquêtes ménages, l'INSEE cherche à utiliser au mieux les ressources disponibles. Il s'agit, étant donnés les moyens alloués à une enquête, d'obtenir l'échantillon collecté contenant le plus d'information possible (et conduisant ex post à la variance la plus faible possible). Un point de départ est le concept de R-indicateur, développé dès 2009 par Schouten, qui permet de construire dans des enquêtes par téléphone des échantillons de répondants représentatifs de la population.

Les R-indicateurs ont été initialement développés pour permettre la priorisation d'efforts de relance dans des enquêtes dont la collecte s'effectue par téléphone. Contrairement à ce cadre, la collecte en face-à-face ne permet pas un réajustement réactif des efforts de collecte, principalement car les enquêteurs organisent sur plusieurs semaines la collecte des unités qui leurs sont affectées. La solution choisie est de réaliser l'enquête en deux vagues. L'échantillon de vague 2 est tiré en prenant en compte le portrait de la collecte réalisée en vague 1. Il est tiré avec l'objectif d'équilibrer la collecte (et de minimiser la dispersion des poids) à la fin de la vague 2, en supposant que les conditions de collecte restent identiques entre les deux vagues. Ce principe a été mis en place pour l'enquête Patrimoine 2014 en région Île-de-France.

D'un point de vue technique, il s'agit d'un problème d'exercice optimal à la date 0 d'une stratégie dont les résultats sont attendus à la date T, la quantité optimisée pouvant évoluer entre les dates (ultérieures) 0 et T. L'idée est d'utiliser les R-indicateurs de manière à tirer l'échantillon donnant la prévision de collecte optimale, c'est-à-dire équilibrée au sens des R-indicateurs, et avec la dispersion des poids corrigés de la non-réponse la plus faible possible, en simulant la collecte avec les nouvelles probabilités de tirage. On intègre

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

à l'algorithme une phase de correction de la non-réponse par Groupes de Réponse Homogène de façon anticipée. La fonction définie par ces paramètres en fonction du vecteur de "sur-représentation" possède de bonne propriétés, et son optimum peut être déterminé par optimisation linéaire. L'algorithme est baptisé CURIOS (Curios Uses Representativity Indicators to Optimize Samples).

Cette méthode permet de s'affranchir du risque de "trous de collecte" (modalité ou zone géographique pour laquelle le taux de collecte est tellement bas que les risques de biais ne sont plus négligeables) et enrichit le monitoring de collecte par R-indicateur d'un contrôle de la dispersion des poids corrigés de la non-réponse, ce qui permet de diminuer la variance ex post.

Abstract

This paper presents the CURIOS algorithm used for the prioritization of CAPI surveys led at the French National Institute for Statistics and Economic Studies (INSEE). It is based on the minimization of a linear combination of several factors related to the quality of the sample, using Monte Carlo techniques to achieve the optimum. We explain how this algorithm functions and then present some results obtained for the 2014 Household Wealth survey, which second wave was prioritized.

Mots-clés

Sondages, Enquêtes ménages, Non-réponse, Collecte adaptative, Monte Carlo.

Table des matières

1 Problématique liée à l'enquête Patrimoine 2014

1.1 Pourquoi deux vagues ?

Les méthodes de priorisation évoquées dans [12] 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, ce qui est en fait uniquement possible pour une collecte téléphonique.

Le déroulement de la collecte d'une enquête CAPI [3] de l'INSEE se déroule en trois phases : la phase de repérage, la phase de régime permanent et la phase d'accélération. Le graphe en figure 1 montre la progression de la collecte pour une enquête CAPI spécifique, l'enquête EPIC [4] en Bourgogne (courbe lissée, et en pointillés, droites des régressions linéaires pour trois phases de collecte). On voit que le nombre de répondants n'est pas linéaire en fonction du temps, mais affine par morceaux.

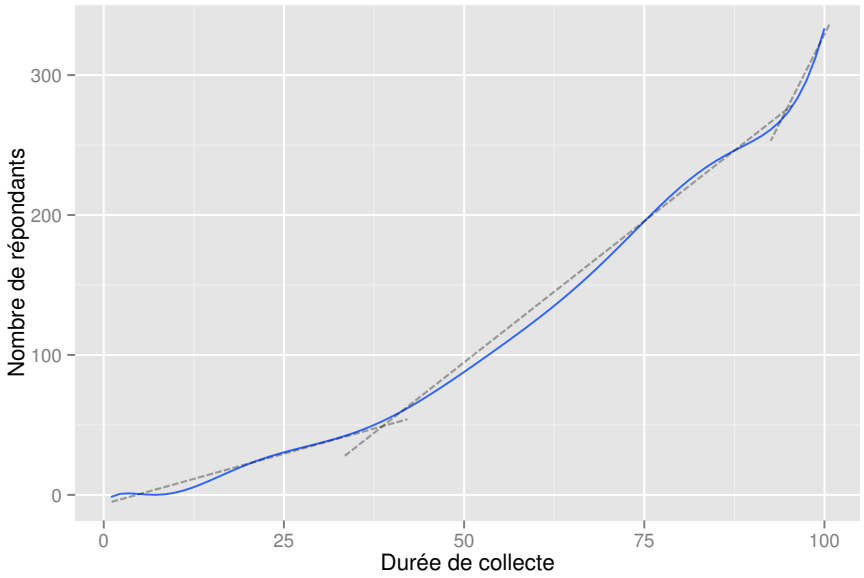

FIGURE 1 – Progression de la collecte de l'enquête EPIC en Bourgogne.

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.

Une approche possible consiste donc à réaliser la procédure de priorisation à la fin d'une vague comportant les trois phases, à condition qu'un taux d'avancement suffisant soit atteint, afin d'éviter de déséquilibrer des échantillons sur la base d'hypothèses prématurées.

3. *Computer-Assisted Personal Interview.*
4. Étude des Parcours Individuels et Conjugaux, plus connue sous le nom Enquête Couples.

1.2 Méthode de construction de l'échantillon de vague 2

1.2.1 Logique de l'algorithme CURIOS

L'algorithme CURIOS consiste en la mise en œuvre d'un compromis entre plusieurs facteurs pouvant définir le "bon" caractère d'un échantillon : pour l'instant, le modèle qui a été utilisé pour tirer la priorisation de la deuxième vague de l'enquête Patrimoine 2014 utilise deux facteurs, les R-indicateurs, introduits par Schouten dans [2], et la dispersion des poids.

Ces derniers sont des paramètres pertinents pour l'optimisation d'un échantillon. En effet, d'une part les R-indicateurs servent à étudier la représentativité de la collecte, afin d'obtenir des échantillons équilibrés selon un certain sens [11]. D'autre part, bien qu'une dispersion faible des poids de sondage corrigés de la non-réponse n'implique pas que les résidus obtenus lors de l'application de la méthode de calage soient faibles également, cela participe de la robustesse de la méthode, qui suppose tous les ménages sont équivalents au sein de l'échantillon obtenu : il n'y a pas de raisons qu'un ménage soit beaucoup plus influent qu'un autre (voir [1] pour une réflexion plus précise sur la question). C'est déjà l'objectif principal de la procédure d'échantillonnage OCTOPUSSE [8].

Le schéma en Figure 2 résume de manière graphique la logique d'opposition entre les deux objectifs poursuivis par l'algorithme CURIOS.

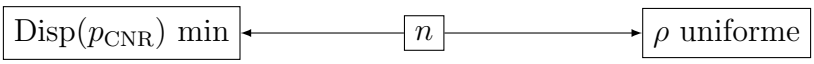

FIGURE 2 – Logique de l'algorithme CURIOS.

Une étude plus approfondie des différentes typologies des objectifs poursuivis se trouve en Partie 4.3.

1.2.2 Décision optimale à la date 0

Il nous faut obtenir l'échantillon **de répondants** pour lesquels la dispersion des poids **corrigés de la non-réponse** est minimal. Si dans l'exemple simple de la partie 4 le taux de réponse de chaque individu est parfaitement déterminé, il s'agit en réalité de prédire le comportement de réponse en vague 2, de manière à optimiser l'échantillon en fonction.

Finalement, le problème d'optimisation de l'échantillon priorisé s'apparente ainsi à un problème de décision optimale. On verra que les probabilités de réponse sont modélisées par des méthodes semblables aux méthodes usuelles de correction de la non-réponse. Les échantillons potentiels en fin de vague 2 sont explorés par méthode de Monte-Carlo, et l'adéquation de l'échantillon proposé à l'objectif est quantifié par une variable de $\mathbb{R}$. La recherche de l'optimalité se fait ensuite en utilisant un algorithme classique d'optimisation linéaire dans le cas bruité.

2 Construction de l'échantillon optimal de vague 2 par Monte-Carlo

2.1 Programme d'optimisation

2.1.1 Objectif

Notons S_1 et S_2 les échantillons de vague 1 et de vague 2. Nous prenons comme date de référence 0 la fin de la vague 1. À cette date, il s'agit de trouver l'échantillon S_2 qui optimise la dispersion des poids et l'équilibre de l'échantillon de répondants **pour l'échantillon total** $S = S_1 \cup S_2$. Ceci s'écrit :

$$\arg\min_{S_2} \ \mathbb{E}\left[\Sigma(w_{CNR}) + \lambda \cdot \Gamma(S)\right] \tag{1}$$

avec :

$\Sigma(X)$ = dispersion (variance empirique) du vecteur X
$S = S_1 \cup S_2$ = échantillon total
w_{CNR} = vecteur des poids corrigé de la non-réponse des unités de S
$\Gamma(S)$ = fonction de mesure de l'équilibre de l'échantillon de répondants (voir 2.1.2)
$\lambda \in [0, +\infty[$

2.1.2 Fonction Γ de quantification de l'équilibre de l'échantillon de répondants

La fonction Γ quantifie l'équilibre de l'échantillon de répondants. Comme précisé en 1.2.1, minimiser le déséquilibre de l'échantillon de répondants répond à un objectif de robustesse : un échantillon équilibré, relativement aux variables par rapport auxquelles cet équilibre est contrôlé, minimise le risque de biais par défaut de couverture.

Les différents R-indicateurs de Schouten (voir [2] et [12]) fournissent une mesure de l'équilibre de l'échantillon. Deux versions de Γ sont testées.

Le choix de Γ_1 fait intervenir le R-indicateur total. Si cet indicateur vaut 1, la collecte est censée être totalement équilibrée. La recherche de l'optimalité cherche à amener Γ le plus près possible de 0.

$$\Gamma_1 = 1 - R_{total}$$

Le choix de Γ_2 se concentre sur les R-indicateurs par modalité : il s'agit de corriger le plus possible les R-indicateurs déséquilibrés, en les amenant le plus proche possible de 0.

$$\Gamma_2 = \|R_{par\ modalite}\|_p$$
$$\text{où : } p \in]1; 2[$$

Il s'avère dans notre cas pratique que le choix de Γ_1 ou de Γ_2 n'influe que très peu sur les allocations obtenues. En général, le choix de Γ_2 semble plus logique, la valeur du R-indicateur global semblant plus indicative de la qualité discriminante du modèle que d'un véritable témoin de l'équilibre de la collecte.

2.2 Stock et vecteur de sur-représentation

2.2.1 Rappels sur les méthodes de tirage de l'INSEE

Dans le cadre des échantillons tirés pour les enquêtes ménages à l'INSEE, on dispose usuellement d'un échantillon de réserve, qui peut être déclenché suite à une décision de la maîtrise d'ouvrage d'une enquête ou sur requête du management terrain de la collecte (DEM [5]). Dans ce cadre, la procédure de tirage d'un échantillon S_1 de taille n_1 au sein de la population U peut être décrite ainsi :

1. Tirage d'un premier échantillon S_0 de taille $n_0 > n_1$ selon le plan de sondage choisi par la maîtrise d'ouvrage
2. Tirage de l'échantillon $S_1 \subset S_0$ de taille n_1. En pratique, il s'agit souvent d'un tirage systématique sur fichier trié sur une variable déjà concernée par le même type de tirage lors de la constitution de S_0.

Le schéma de la figure 3 illustre ce mécanisme.

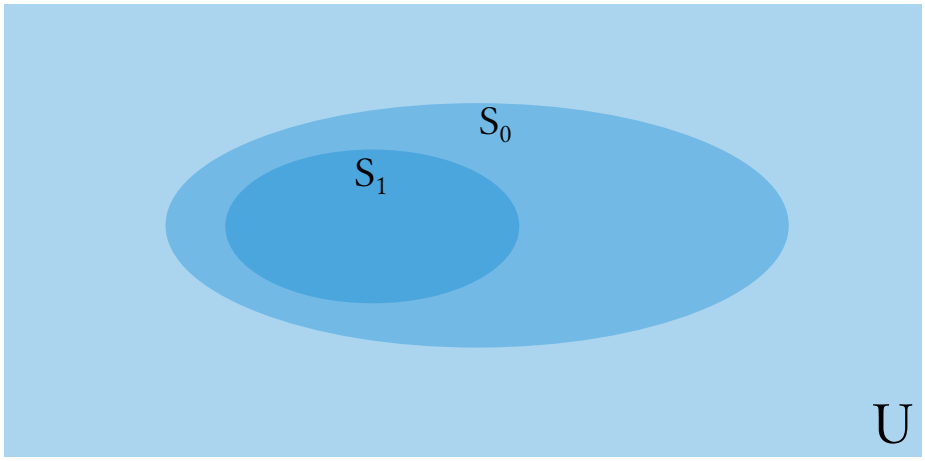

FIGURE 3 – Illustration du tirage en une vague

Les différents ensembles mis en jeu dans le tirage en une vague sont désignés par :

$$\begin{aligned} U &= \text{population} \\ S_1 &= \text{échantillon} \\ S_0 - S_1 &= \text{échantillon de réserve} \end{aligned}$$

Tirage en deux phases Cette méthode constitue un exemple de tirage en deux phases. On rappelle que dans ce cadre les poids définis par :

$$\begin{aligned} w_k &= w_k^0 \cdot w_k^{1/0} \text{ , où :} \\ w_k^0 &= \text{inverse des probabilités de tirage de } S_0 \text{ dans } U \\ w_k^{1/0} &= \text{inverse des probabilités de tirage de } S_1 \text{ dans } S_0 \end{aligned}$$

onduisent à un estimateur sans biais (estimateur en expansion, voir par exemple [10]).

5. Direction Enquêtes Ménages

2.2.2 Stock - tirage de la deuxième vague

Dans le cas d'un tirage en deux vagues tel que celui présenté ici, il s'agit de tirer deux échantillons : S_1 selon la méthode présentée au paragraphe 2.2.1, puis S_2 après la fin de la collecte de S_1. S_2 est tiré parmi les éléments de l'ensemble $S_0 - S_1$, qui appelé **stock** disponible pour la vague 2. La figure 4 illustre cette méthode de tirage.

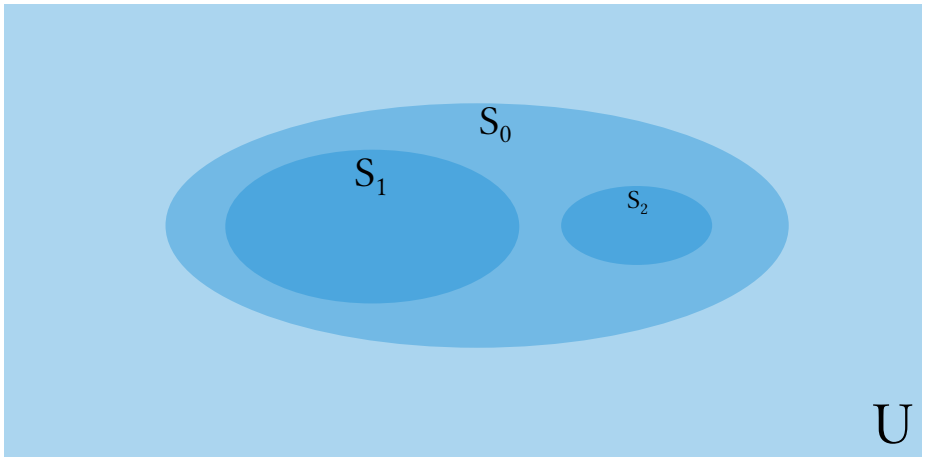

FIGURE 4 – Illustration du tirage en 2 vagues

Les différents ensembles mis en jeu dans le tirage en deux vagues sont désignés par :

$$\begin{aligned} U &= \text{population} \\ S_1 &= \text{échantillon de première vague} \\ S_0 - S_1 &= \text{stock disponible pour le tirage de deuxième vague} \\ S_2 &= \text{échantillon de deuxième vague} \\ S_1 \cup S_2 &= \text{échantillon total} \\ S_0 - S_1 - S_2 &= \text{échantillon de réserve} \end{aligned}$$

Le tirage de S_2 parmi le stock s'effectue de la même manière que le tirage de S_1 au sein de S_0 en première vague. Afin de disposer de la plus grande variété de profils pour un tirage optimal de l'échantillon priorisé S_2, on constitue le stock le plus large possible : $n_0 \gg n_1$.

2.2.3 Fonction $\mathcal{R}$

La fonction de "redimensionnement"

$$\mathcal{R} : \left\{ \begin{array}{rcl} (\mathbb{R}^n)^{+*} & \to & (\mathbb{R}^n)^{+*} \\ \epsilon & \to & \mathcal{R}(\epsilon) \end{array} \right.$$

permet de construire un vecteur de probabilités $\mathcal{R}(\epsilon)$ compatible avec les probabilités d'inclusion à l'ordre 1 d'un tirage proportionnel au vecteur ϵ. ϵ peut être vu comme un vecteur donnant la "taille" de chaque unité de la population : dans ce cas $\mathcal{R}$ permet de construire un sondage à probabilités inégales de taille fixe proportionnel à la taille. Il est à noter que la fonction $\mathcal{R}$ n'est pas bijective.

2.2.4 Vecteur ϵ de sur-représentation

Comme expliqué en 2.2.2, l'échantillon S_2 de deuxième vague est sélectionné par tirage systématique (et possède donc une taille fixe n_2). Les probabilités d'inclusion d'ordre 1 $\pi_k^{2/1}$ sont assignées aux $n_{Stock} = n_0 - n_1$ éléments du stock ainsi :

$$\pi^{2/1} = \mathcal{R}(\epsilon)$$

ϵ est donc un vecteur de taille n_2, nommé **vecteur de sur-représentation**. On peut donner un exemple d'initialisation d'utilisation du vecteur ϵ tel qu'il est spécifié pour initialiser l'algorithme de recherche de l'optimum (voir 2.4) : afin de se conformer à l'esprit des simulations effectués en phase préparatoire de l'opération, on fixe $\epsilon_k = 1$ pour les fiches adresses ne présentant pas de profil sur- ou sous-priorisé, et on ajuste les $\epsilon_k \geq 1$ (respectivement $\epsilon_k \leq 1$) pour les unités sous-représentées (respectivement sur-représentées) au sens de la fonction Γ.

De même qu'en 2.2.1, il s'agit donc d'un tirage en deux phases, de S_2 au sein de S_0. Les poids :

$$w_k = w_{0,k} \cdot \frac{1}{\pi_k^{2/1}}$$

où : $w_{0,k}$ = poids de tirage de première phase de l'unité k

constituent donc un estimateur sans biais pour le calcul des totaux sur la population U.

2.3 Simulation de la collecte

Notre approche consiste à simuler, à ϵ donné, la collecte de vague 2, et à recueillir une prédiction des paramètres de l'algorithme CURIOS pour le scénario donné.

2.3.1 Prédiction de la collecte

On utilise une modélisation logit de la non-réponse en vague 1, et l'on utilise la prédiction du modèle pour obtenir un vecteur de propensions à répondre $\hat{p}$. Les $\hat{p}_k$ sont utilisées pour la simulation (2.4).

L'utilisation du modèle logit coïncide avec les modèles de correction de la non-réponse classiquement utilisés à l'INSEE, incluant la correction de la non-réponse appliquée en 2.3.3. Ceci nous permet de comparer notre modélisation avec les standards INSEE en la matière, et on se satisfait de retrouver des déterminants classiques de la non-réponse.

Cependant, la performance de l'algorithme CURIOS est plus dépendante de la qualité prédictive du modèle que de sa qualité explicative. La régression logit propose un modèle facilement interprétable, mais peu prédictif. Fonder la prédiction en vague 2 sur des algorithmes d'apprentissage (arbres de régression, boosting, etc.) est une amélioration envisagée à moyen terme.

2.3.2 Simulation

Pour chaque unité k de vague 2, on réalise une expérience de Bernoulli de probabilité $\hat{p}_k$, afin de simuler sa réponse.

Le programme implémentant l'algorithme CURIOS est écrit en Python, et utilise la programmation orientée objet pour proposer une intégration modulaire des différentes typologies de variables (voir 4.3). Les algorithmes de tirage de S_0 sont vus comme des classes abstraites et ont vocation à être ré-écrits pour chaque enquête (bien que certaines méthodes classiques, telle OCTOPUSSE, aient déjà été implémentés). L'optimisation linéaire utilise les méthodes du package scipy ([6]), et le traitement de données le package pandas ([13]).

Le calcul est pour l'instant effectué sur un seul CPU, et via l'interpréteur Python. Il est envisagé à moyen terme d'écrire les algorithmes de calcul sous Cython [6] pour bénéficier des avantages du code directement compilé. Les méthodes de Monte-Carlo telles que celle utilisée pour la simulation de la collecte se prêtent généralement bien à la parallélisation. L'utilisation de calcul sur GPU (Graphical Processing Unit) permettrait d'augmenter drastiquement le nombre de simulations, et donc de diminuer la variance des estimateurs utiles donnés en 2.3.3.

2.3.3 Analyse de la collecte simulée

Parmi les scénarios possibles, on calcule les variables d'intérêt pour l'algorithme CURIOS à savoir : $\Sigma(w_{CNR})$, la dispersion des poids corrigés de la non-réponse, et $\Gamma(S)$, l'indicateur de représentativité, défini en 2.1.2. Pour chacune de ces valeurs, on calcule la variance empirique $\hat{\Sigma}$ qui permet de construire un intervalle de confiance sous l'hypothèse gaussienne asymptotique :

$$\hat{IC_{95\%}}(\hat{\Gamma}) \approx \left[\bar{\Gamma} - 2\hat{\Sigma}(\hat{\Gamma}); \bar{\Gamma} + 2\hat{\Sigma}(\hat{\Gamma})\right]$$

Et il en va de même pour les w_{CNR}. L'exactitude de l'hypothèse gaussienne importe peu, l'intervalle de confiance sert d'indicateur pour décider de déclencher ou non un échantillon priorisé en vague 2. En effet, on simule également la collecte dans le cas où S_2 est tiré avec le même plan que S_1 (c'est-à-dire non priorisé), et l'on observe $\Sigma(w_{CNR})$ et $\Gamma(S)$. Si ces valeurs ne sont pas contenues dans les intervalles de confiance correspondants, alors le scénario envisagé (c'est-à-dire le vecteur ϵ) est réputé permettre un gain significatif. Dans le cas contraire, le scénario ne permet pas d'envisager raisonnablement un gain sur les paramètres choisis.

Le calcul de la dispersion (anticipée) des poids corrigés de la non-réponse $\Sigma(w_{CNR})$ impose d'effectuer, pour chaque scénario simulé, une correction de la non-réponse. L'opération de correction de la non-réponse est appliquée aux poids w_{final} définis en 2.5, de manière à reproduire exactement la phase de post-traitement de l'enquête. La correction de la non-réponse doit répliquer le plus fidèlement possible la méthode qui sera effectivement appliquée lors du post-traitement de l'enquête. On utilise donc la méthode classique des

6. `http://cython.org`

GRH [7], fondés sur une modélisation par la régression logistique. Le différents paramètres utilisés pour la méthode des GRH correspondent à ceux proposés par Beaumont et Haziza ([4]). Les variables utilisées pour le modèle logit sont uniquement constituées des variables de la base de sondage. Aucune paradonnée n'est pour l'instant utilisable pour cette opération à l'INSEE, bien que l'utilisation de celles-ci permettraient d'améliorer la prédictivité du modèle, et par là même la qualité des simulations, et donc de l'échantillon priorisé ([12]). Le calcul de $\Gamma(S) = \Gamma(S_1 \cup S_2)$ ne pose pas de problème particulier.

2.4 Recherche de l'optimum

2.4.1 Algorithme de Nelder-Mead

La fonction d'optimisation $\mathbb{E}\left[\Sigma(w_{CNR}) + \lambda \cdot \Gamma(S)\right]$ est très bruitée, en particulier car l'espérance est estimée par méthode de Monte-Carlo. On choisit donc une méthode d'optimisation linéaire peu sensible au bruit et qui ne repose pas sur un calcul de gradient : l'algorithme de Nelder-Mead, connu aussi sous le nom d'algorithme du simplexe.

Groupes d'optimisation

Le temps de calcul pour évaluer la quantité à optimiser ainsi que le nombre d'appels à cette fonction d'évaluation par l'algorithme du simplexe augmentent avec la dimension du problème. Afin de ne pas faire exploser le temps de calcul, on effectue des regroupements de fiches adresses substituables. La valeur du ϵ_k est la même pour toutes les unités au sein d'un groupe, ce qui permet de nettement réduire la dimension du problème, au prix toutefois d'une perte d'optimalité qu'on se doit de contrôler.

Concrètement, toutes les unités appartenant à un même groupe d'optimisation seront considérées identiques. En termes de conception d'enquête, la question se pose de la **substituabilité** de fiches adresses en fonction de leurs caractéristiques. Il convient donc :

1. d'optimiser le programme de résolution de manière à effectuer le moins de regroupements possibles
2. d'effectuer les regroupements d'unités les plus semblables possibles en regard de l'objectif de l'enquête

2.4.2 Optimums locaux et scénarios

L'algorithme précédent converge vers un optimum local, et non un optimum global. L'optimum local vers lequel l'algorithme converge dépend du point à partir duquel la recherche est lancée. Afin de s'assurer de ne pas manquer les meilleurs optimums locaux existants, on relance l'algorithme de recherche plusieurs fois à partir de points parcourant à intervalles prédéfinis l'ensemble de l'espace du problème, selon la logique du "grid search". La taille de la grille parcourue est fixée selon des critères de temps machine.

L'algorithme d'optimisation linéaire est gourmand en nombre d'appels à la fonction d'évaluation. Chaque appel de la fonction d'évaluation demande un temps de calcul conséquent pour assurer une précision suffisante. Afin de diminuer le temps de calcul total, l'évaluation de la fonction d'optimisation est effectuée avec un nombre de simulations

7. Groupes de Réponse Homogènes

faible lors de l'étape de "grid search". On sélectionne les optimums locaux intéressants, et on recherche plus finement leur localisation avec un nombre de simulations plus important à partir des points où ils ont été détectés lors de la première étape.

La collection des optimums locaux obtenus permet de proposer plusieurs scénarios de priorisation. Plusieurs critères permettent ensuite de choisir le vecteur ϵ finalement implémenté. On regarde tout d'abord la significativité de l'amélioration attendue (au sens décrit en 2.3.3) sur les variables de contrôle. Le scénario présentant l'amélioration la plus significative n'est pas forcément à retenir, notamment s'il conduit à des allocations très déséquilibrées faisant courir le risque de variances très élevées sur des petits domaines. En effet, l'algorithme CURIOS n'intègre pas à ce stade de garde-fou permettant de se prémunir contre ce type d'écueil. Un autre aspect intéressant consiste à observer le nombre de FA anticipées, caluclé à partir des $\hat{p}$ constitués pour la simulation de la collecte de vague 2. Un scénario permettant un fort gain en dispersion des poids associé à une baisse drastique des taux de collecte pourrait conduire à une variance *in fine* plus élevée que sans mise en place de la priorisation. Le choix du scénario doit donc se faire avec un regard critique, et en collaboration avec l'équipe en charge de la conception de l'enquête (voir 3.4).

On remarque également qu'il n'existe pas un unique optimum global, car plusieurs ϵ_k peuvent conduire au même jeu de probabilités d'inclusion d'ordre 1 (non-injectivité de la fonction $\mathcal{R}$). La définition de ϵ pourrait être modifiée de façon à assurer cette propriété.

2.5 Poids concaténés des vagues

L'obtention du vecteur optimum ϵ donne accès à $w^{2/1} = \dfrac{1}{\pi^{2/1}}$, poids de tirage des unités de vague 2, conditionnels à la collecte de vague 1. On cherche ensuite à constituer w_{final}, les poids de l'échantillon total après la vague 2 à partir des $w^{2/1}$.

On utilise pour cela le principe de pondération optimale décrit dans Ardilly ([10]). Dans la même logique que la méthode de partage des poids, il consiste à repondérer chaque échantillon de manière à assurer une variance de l'échantillon concaténé minimale. On obtient :

$$w_{final} = \frac{n_1}{n_1 + n_2} \cdot w_1 + \frac{n_2}{n_1 + n_2} \cdot w^{2/1}$$

3 Résultats pour l'enquête Patrimoine 2014

3.1 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.

Le plan de sondage utilisé pour les enquêtes Patrimoine consiste à sélectionner aléatoirement au sein de chacune des ZAE [9] 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 l'hétérogénéité importante de la queue de la distribution des patrimoines. En outre, à cette variation s'ajoute un phénomène de non-réponse élevée dans les très hauts patrimoines, due à des difficultés d'accès, des réticences ou une plus faible disponibilité. Pour ces raisons, un second échantillon dit **"Hauts Patrimoines"** 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 5 strates sont : 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 "hauts patrimoines", les 3 strates sont : les riches urbains, les riches en zone rurale, et les patrimoines plus faibles.

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.

3.2 Particularités pratiques pour la priorisation

Le travail préparatoire pour la priorisation a été effectué par simulations ([12], [11]), sous l'hypothèse que l'opération serait menée sur les Fiches Adresse de la réserve, réparties dans tous les départements français.

Pour des raisons organisationnelles, la décision a été prise de ne déclencher l'opération que pour une partie de l'échantillon : celui dont la collecte est gérée par la DEM [10] Île-de-France, c'est-à-dire les départements de Paris, de la petite couronne (Hauts-de-Seine,

8. Tous les 6 ans jusqu'en 2010, le rythme a changé depuis.
9. Zone d'Action Enquêteur, unité primaire des plans de sondage à l'INSEE
10. Direction Enquêtes Ménages

Seine-Saint-Deins, Val-de-Marne) et des Yvelines. L'objectif de la priorisation consiste alors à maximiser l'information collectée à coût donné, dans le contexte de taux de réponse faibles dans ces départements.

Afin de permettre d'anticiper le calcul des allocations avant la fin effective de la collecte de la première vague, la DEM Île-de-France a organisé la remontée des informations de prises de contact par les enquêteurs. 25% de l'échantillon final est priorisé (appartient à la vague 2). Afin de satisfaire l'exigence de 6 semaines de durée de collecte expliquée dans [12], la fin de la collecte de l'enquête Patrimoine est repoussée. La collecte de la vague 2 s'effectue du 1/1/2015 au 14/2/2015.

Au moment d'effectuer les simulations pour le calcul des allocations, une partie des ménages contactés mais pas encore rencontrés est considérée comme répondante, de manière à augmenter la significativité des indicateurs décrits en 4.3. Cette anticipation permet de livrer les Fiches Adresses avant le début de la collecte de vague 2, ce qui permet aux enquêteurs de commencer le repérage pour la vague 2 en avance (la vague 1 de collecte de Patrimoine 2014 prenait fin juste avant les fêtes de fin d'années).

Enfin, afin d'assurer la couverture d'un maximum de communes des départements sélectionnés pour la vague 2, les enquêteurs sélectionnés ont accepté de couvrir certaines communes situées hors de leur zone d'action supposée.

L'échantillon final de vague 2 doit finalement avoir une taille de **820 fiches adresses**. La répartition entre les deux sous-échantillons au sens de l'allocation initiale est indiquée en table 1.

Échantillon	Taille
Total	820
Hauts Patrimoines	344
Patrimoines Standards	476

TABLE 1 – Taille des sous-échantillons pour la région de gestion Île-de-France, répartition selon l'allocation initiale

3.3 Statistiques descriptives

Le modèle logit choisi a pour équation :

$$\begin{aligned} \text{repondant} \sim\ & \text{typo commune} + \text{strate} + \text{statut occupation logement} \\ & + \text{type logement} + \text{sexe} + \text{indicatrice surface} \end{aligned}$$

3.3.1 Typologie de communes

La conception d'enquête indique que les fiches adresses des communes à revenu médian élevés ne sont pas substituables à des fiches adresses provenant d'autres communes présentant des caractéristiques socio-démographiques très différentes, particulièrement pour l'Île-de-France où les inégalités sont très marquées. Il nous faut donc séparer ces différents profils au sein des regroupements (voir 2.4.1).

Pour cela, on crée une typologie de communes de la région Île-de-France à l'aide d'une méthode de clustering (méthode des centres mobiles) fondée sur le revenu médian et la tranche d'unité urbaine de chaque commune. Il en résulte 4 modalités, que l'on peut décrire ainsi : Communes urbaines aux revenus élevés / Communes rurales aux revenus élevés / Communes urbaines aux revenus faibles / Communes rurales aux revenus faibles.

On remarque d'ailleurs par la suite que la typologie de communes n'est pas une variable à fort R-indicateur (voir 3.3.2), et elle n'aurait pas été retenue pour la constitution des groupes d'optimisation sans la consigne spécifique de la conception d'enquête.

3.3.2 R-indicateurs

Les tables 2 et 3 donnent les R-indicateurs par variable pour les variables du modèle 3.3 ainsi que les R-indicateurs partiels inconditionnels pour les modalités des variables aux R-indicateur par variable les plus forts.

R-indicateur global
0,9648

Variable	R-indicateur	Écart-type
Typologie commune	0,0066	0,0135
Strate	0,0083	0,0162
Statut occupation logement	0,0098	0,0126
Type logement	0,0116	0,0130
Sexe personne de référence	0,0005	0,0803
Surface logement (tranches)	0,0130	0,0121

Statut occupation	R-indicateur
Locataire	-0,0131
Propriétaire	0,0185

Type logement	R-indicateur
Appartement	-0,0100
Maison	0,0248

Surface logement	R-indicateur
1	-0,021
2	0,007
3	0,020

Table 2 – R-indicateurs pour l'échantillon Patrimoines Standards

Lien avec le taux de collecte Dans les tables 2 et 3, les modalités déficitaires possèdent un R-indicateur négatif. On peut légitimement s'attendre à ce que les allocations issues de CURIOS soient plus élevées sur ces modalités. On peut se demander s'il

R-indicateur global
0,8848

Variable	R-indicateur	Écart-type
Typologie commune	0,0157	0,0167
Strate	0,0090	0,0245
Statut occupation logement	0,0178	0,0162
Type logement	0,0291	0,0158
Sexe personne de référence	0,0457	0,0147
Surface logement (tranches)	0,0287	0,0154

Type logement	R-indicateur
Appartement	-0,0168
Maison	0,0306

Sexe personne référence	R-indicateur
Homme	0,0345
Femme	-0,0426

Surface logement	R-indicateur
1	-0,0222
2	-0,0036
3	0,0261

TABLE 3 – R-indicateurs pour l'échantillon Hauts Patrimoines

existe un lien direct entre taux de collecte et modalité à prioriser. On présente dans les tables 4 et 5 les taux de collecte pour les modalités de statut d'occupation du logement et de type de logement. Il ressort que les modalités déficitaires au sens des R-indicateurs sont également celles pour lequel le taux de collecte à la date d'étude est le plus faible (locataires, appartements). L'analyse est beaucoup moins aisée pour les modalités de la variable de surface en tranches (table 6) : si le taux de collecte des petites surfaces (modalité 1) est largement inférieur à celui des grandes surfaces (modalité 3), et qu'il se conçoit donc aisément que la première soit sous-représentée quand la seconde est sur-représentée ; la modalité 2 possède un taux de collecte intermédiaire, et il est difficile d'intuiter son caractère sous ou sur représenté (sous-représenté dans le cas Patrimoines Standards). On peut se référer à [12] pour une explication plus détaillée à ce sujet.

Statut occupation logement	Taux de réponse
Locataire	0.30
Propriétaire	0.38

TABLE 4 – Taux de collecte par statut d'occupation du logement

Type de logement	Taux de réponse
Appartement	0.34
Maison	0.41

TABLE 5 – Taux de collecte par type de logement

Tranche de surface de logement	Taux de réponse
1	0.31
2	0.34
3	0.37

TABLE 6 – Taux de collecte par tranche de surface du logement

Par ailleurs, on constate des R-indicateurs globaux de l'ordre de 0.9. Cela signifie que les indicateurs utilisés ne parviennent pas à bien discriminer les déterminants de la non-réponse à ce stade de l'enquête. En se référant à l'analyse effectuée dans [12], cela correspond aux R-indicateur des stades peu avancés de la collecte. Or, l'avancement est de l'ordre de 75% au moment où l'analyse des taux de collecte est effectuée. Il s'agit donc plus probablement d'un effet de la taille réduite de l'échantillon analysé, l'opération ayant été déclenchée uniquement dans certains départements de l'Île-de-France, alors que les simulations de [12] avaient été effectuées pour la France entière. On peut dès lors s'attendre à de moins bons résultats, notamment en termes de gains en variance, que ceux prévus par simulation.

3.4 Scénarios proposés

On effectue l'analyse de la sortie de CURIOS séparément pour les deux sous-échantillons "Hauts Patrimoines" et "Patrimoines Standards". Il aurait été tout à fait possible d'envisager d'effectuer une optimisation conjointe intégrant les deux sous-échantillons, mais on souhaite d'abord s'assurer que l'allocation priorisée permettrait un gain sur chaque sous-échantillon traité séparément. En l'occurence, aucun scénario permettant un gain significatif n'a été décelé pour le sous-échantillon "Hauts Patrimoines", et il est finalement décidé de **ne pas changer l'allocation pour l'échantillon "Hauts Patrimoines"**.

On se concentre donc sur le sous-échantillon "Patrimoines Standards". Comme indiqué en partie 2.4, les scénarios sont constitués par les différents optimums locaux trouvés. Dans notre cas, on trouve 3 optimums locaux dont seulement deux présentent un gain significatif (comme expliqué en partie 2.3.3). On présente les allocations correspondant à ces optimums locaux en table 7. Le Scénario 1 permet d'anticiper un gain de **3,5%** en dispersion des poids ; le scénario 2 un gain de **7,5%**.

3.5 Allocation finale

Le regroupement pour l'optimisation est effectué en croisant la typologie de communes et les strates (voir 2.4.1). On pourrait penser que le scénario 2, présentant un gain anticipé nettement plus fort, serait l'option la plus intéressante. Cependant, l'étude conjointe avec la conception d'enquête des allocations de la table 7 montre que le déséquilibre proposé est très important, et conduit notamment à déprioriser la strate "cadres", qui constitue pourtant un domaine d'intérêt pour l'enquête. D'un point de vue méthodologique, on

Typo. commune × strate	Sans priorisation	Scénario 1	Scénario 2
1 × ages	27	21	15
1 × independants	11	9	5
1 × revenus patrimoine	10	7	5
1 × cadres	75	62	45
1 × reste	49	69	86
2 × ages	1	0	0
2 × cadres	4	2	2
2 × reste	2	2	2
3 × ages	6	6	4
3 × independants	4	3	3
3 × revenus patrimoine	2	2	2
3 × cadres	21	19	10
3 × reste	7	10	16
4 × ages	47	38	25
4 × independants	19	13	8
4 × revenus patrimoine	8	5	3
4 × cadres	83	68	51
4 × reste	100	140	194

TABLE 7 – Allocations par groupes d'optimisation, sans priorisation et scénario retenu

cherche à s'assurer que les allocations ne dépendent pas trop fortement d'un modèle potentiellement mal spécifié, d'autant que le nombre de fiches adresses du problème laisse augurer d'une faible robustesse. Afin de limiter les risques, **on retient finalement le scénario 1**.

On s'attendait également à ce que les allocations priorisées favorisent les fiches adresses appartenant à des groupes pour lesquels les taux de collecte sont les plus faibles. La prévision du nombre de fiches adresses collectées par scénario en utilisant les $\hat{p}$ déterminés en 2.3 nous semblait devoir être plus faible pour les scénarios priorisés. En fait il n'en est rien, les prévisions de collecte s'établissent à 195, 188 et 191 fiches adresses respectivement pour le scénario non priorisé, le scénario 1, et le scénario 2. D'une part, la faible prédictivité du modèle logit suggère que qu'il est difficile d'anticiper ce que sera réellement le nombre final de fiches adresses collectées. D'autre part, cette estimation repose sur l'hypothèse que les conditions de collecte sont strictement équivalentes entre la vague 1 et la vague 2. Or en réalité, les enquêteurs participant à l'opération ne sont pas nécessairement les mêmes, et les consignes passées par les DEM sont potentiellement différentes.

On recense dans la table 9 la variation en pourcentage du nombre de fiches adresses présentes dans l'échantillon de vague 2 entre le scénario retenu et l'allocation initiale (non priorisée) pour les modalités "locataires", "appartement" et "petites surfaces". On constate une nette augmentation, notamment pour la strate "reste de la population", qui constitue le réservoir le plus important pour ces modalités (voir table 8). Pour schématiser, on peut considérer que la deuxième vague priorise les locataires urbains d'appartements ou de petites surfaces.

Typo. commune × strate	Locataires	Appartements	Petites surfaces
1 × ages	0.33	0.77	0.16
1 × independants	0.32	0.84	0.10
1 × revenus patrimoine	0.26	0.81	0.11
1 × cadres	0.36	0.84	0.17
1 × reste	0.78	0.92	0.38
(2,3) × ages	0.29	0.71	0.29
(2,3) × independants	0.38	0.85	0.15
(2,3) × revenus patrimoine	0.50	0.83	0
(2,3) × cadres	0.46	0.79	0.16
(2,3) × reste	0.64	0.56	0.48
4 × ages	0.40	0.79	0.18
4 × independants	0.27	0.63	0.10
4 × revenus patrimoine	0.32	0.73	0
4 × cadres	0.39	0.82	0.20
4 × reste	0.78	0.92	0.38

TABLE 8 – Part des locataires, appartements et petites surfaces (modalité 1) par groupe d'optimisation

Typo. commune × strate	Locataires	Appartements	Petites surfaces
1 × ages	-0.17	-0.06	-0.08
1 × independants	-0.18	0.01	-0.14
1 × revenus patrimoine	-0.24	-0.02	-0.13
1 × cadres	-0.14	0.01	-0.07
1 × reste	**0.28**	**0.09**	**0.14**
(2,3) × ages	-0.21	-0.12	0.05
(2,3) × independants	-0.12	0.02	-0.09
(2,3) × revenus patrimoine	0	0	-0.24
(2,3) × cadres	-0.04	-0.04	-0.08
(2,3) × reste	**0.14**	-0.27	**0.24**
4 × ages	-0.10	-0.04	-0.06
4 × independants	-0.23	-0.20	-0.14
4 × revenus patrimoine	-0.18	-0.10	-0.24
4 × cadres	-0.11	-0.01	-0.04
4 × reste	**0.28**	**0.09**	**0.14**

TABLE 9 – Part des locataires, appartements et petites surfaces (modalité 1) par groupe d'optimisation. Variation entre le scénario proposé et le scénario sans priorisation

Seules les variables de la base de sondages ont pu être utilisées pour construire les modèles de simulation et de correction de la non-réponse. On peut souhaiter que l'INSEE se dote à moyen terme d'outils permettant le recueil de paradonnées qui permettraient une analyse plus fine, et donc une priorisation plus optimale (voir [12]). Enfin, il faut souligner qu'on ne peut malheureusement pas établir un lien direct entre le gain anticipé en dispersion des poids et le gain potentiel en variance. On peut se référer à [1] pour une analyse plus détaillée.

4 L'algorithme CURIOS simplifié

Cette partie vise à poser les fondements d'un cadre plus théorique pour la compréhension du fonctionnement de l'algorithme CURIOS. Afin d'aboutir à une solution analytique du problème, de nombreuses simplifications sont proposées : en particulier, les R-indicateurs des parties précédentes sont remplacés dans l'analyse par une fonction de distance à l'allocation initiale.

4.1 Objectif de l'algorithme

L'algorithme CURIOS consiste en la mise en œuvre d'un compromis entre plusieurs facteurs pouvant définir le "bon" caractère d'un échantillon : il s'agit ici de deux facteurs, la dispersion des poids corrigés de la non-réponse (pour une discussion sur cet objectif, voir [1]) et la distance à une allocation initialement choisie pour l'échantillonnage.

La proximité avec l'allocation initiale vient contrebalancer l'effet de minimisation de la dispersion des poids afin de conserver une structure définie par les concepteurs de l'enquête. Afin de tenir compte de la non-réponse, et dans un souci de simplification, nous supposerons que cette allocation initiale est l'allocation de Neyman dans laquelle on considère des prévisions de comportement moyen de réponse $\tilde{\rho}$ dans les strates (cf équation 4). Or il est bien connu que l'optimum de Neyman est plat (voir par exemple [9]), on conserve donc ses bonnes propriétés de minimisation de la variance des estimateurs à distance faible de l'allocation initiale.

4.2 Un exemple simple

4.2.1 Principe de l'algorithme

L'algorithme CURIOS réalise un arbitrage entre dispersion des poids corrigés de la non-réponse et distance à l'échantillon initialement déterminé par l'allocation de Neyman afin de déterminer une nouvelle allocation. Usuellement, celle-ci ne peut être réalisée que dans un second temps, une fois une partie de la collecte réalisée ; on se place ici dans un exemple simple pour lequel on connaît déjà les caractéristiques de la population, et on peut ainsi intervenir sur l'allocation en début de collecte. La population est séparée en deux groupes $\mathcal{P}_i$ de taille N_i avec un taux de réponse uniforme ρ_i. On rappelle que les poids corrigés de la non-réponse p^k_{CNR} des n_i individus répondants de $\mathcal{P}_i$ sont :

$$p^k_{\text{CNR}} = \frac{N_i}{n_i \rho_i}$$

On souhaite tirer un échantillon de taille fixe n. On réalise donc le programme de minimisation suivant :

$$n^1_f = \text{argmin} \quad \text{Disp}(\mathrm{p^k_{CNR}}) + \lambda \, \text{Dist}((\mathrm{n^1_f}, \mathrm{n^2_f}), (\mathrm{n^1_{init}}, \mathrm{n^2_{init}})) \tag{2}$$

où Disp est l'opérateur de dispersion autour de leur moyenne des poids corrigés de la non-réponse p^k_{CNR}, Dist est la distance euclidienne dans $\mathbb{R}^2$ et $n^2_f = n - n^1_f$ entièrement défini par la donnée de n^1_f.

Ce programme de minimisation ne dépend que de la constante $\lambda \geq 0$ choisie. On remarque aisément que lorsque $\lambda \to +\infty$, le terme de distance devient prépondérant et on a $(n_f^1, n_f^2) = (n_{\text{init}}^1, n_{\text{init}}^2)$. Dans le cas inverse, i.e $\lambda \to 0$, on obtient une concentration de l'échantillon sur une des deux strates afin de limiter la dispersion des poids.

4.2.2 Choix du λ

Afin de pouvoir appliquer l'algorithme CURIOS, il nous faut choisir une valeur de λ. Une première approche consiste à s'intéresser à la variance d'un estimateur de Horvitz-Thompson corrigé de la non-réponse du total de X, variable d'intérêt de l'enquête. Celle-ci dépend de la valeur de λ via les tailles d'échantillons n_f^i obtenues pour une telle valeur. On a le théorème suivant :

Théorème 1. *Soit $V(\lambda)$ la fonction de variance d'un estimateur du total de X pour les tailles d'échantillons $n_f^i(\lambda)$. On suppose que l'on a l'hypothèse suivante :*

$$\frac{N(\rho_1 - \rho_2)^2}{(n\rho_2)^3}\left[4N + \frac{3N}{\rho_2} - n\rho_2\right] \leq \left| g\left(n\left[1 + \left(\frac{N_2^2\rho_1}{N_1^2\rho_2}\right)^{1/4}\right]^{-1}\right)\right| \tag{3}$$

où

$$g : x \to -\frac{2N_1^2}{\rho_1 x^3} - \frac{2N_2^2}{\rho_2(n-x)^3}$$

Alors $V(\lambda)$ est décroissante et sa dérivée seconde admet un maximum dans $]0, +\infty[$ qu'on appelle point de torsion de $V(\lambda)$.

On veut prendre λ au point de torsion de la courbe, qui est aussi un point d'inflexion de sa dérivée ; en effet, cela permet d'être suffisamment proche du plateau de variance dû à la proximité de l'allocation de Neyman, qui est un optimum plat, tout en limitant au maximum la valeur de λ et donc la dispersion des poids corrigés de la non-réponse.

La détection du point d'inflexion de la dérivée est un problème complexe numériquement. Il peut être souhaitable de rechercher une méthode *ad hoc* de calcul d'une valeur de λ "acceptable", au sens où celle-ci est à droite du coude, sur le plateau de variance. En effet, se trouver à gauche du coude induirait une variance de l'estimation du total de X bien supérieure, ce qui est à éviter, même pour gagner un peu en dispersion des poids.

On définit λ_{num} de telle sorte que chacun des termes de l'équation 2 participe de façon égale au terme à minimiser, les deux composantes - dispersion des poids CNR et écart à l'allocation de Neyman - étant également importantes dans le choix d'une nouvelle allocation. On écrit donc une procédure visant à égaliser les deux termes de l'équation 2. La conjecture suivante affirme que la valeur obtenue par la méthode numérique se situe bien sur le plateau obtenu à droite du coude.

Conjecture 1. *Si l'hypothèse (3) est vérifiée, on a :*

$$\lambda_{\text{num}} \geq \lambda_{\text{coude}}$$

4.2.3 Simulations

Définition de la population On s'intéresse à une variable X sur une population séparée en deux groupes distincts : les "patrimoines standards" qui sont nombreux ($N_1 = 10^5$) et qui sont plutôt bons répondants ($\rho_1 = 0.6$), mais qui ont des valeurs de X peu dispersées ($V_1 = 1$), et les "hauts patrimoines", qui sont moins nombreux ($N_2 = 10^4$), moins bons répondants ($\rho_2 = 0.4$), et avec une grande dispersion des valeurs de X ($V_2 = 100$). On supposera que X est gaussienne afin de simplifier les simulations.

Dans ce cadre, l'hypothèse (3) est vérifiée : en effet, le terme de gauche vaut environ 10890, tandis que la fonction g est toujours négative et atteint son maximum en -13750. On peut donc bien utiliser l'algorithme CURIOS dans ce cas.

Échantillonnage On réalise un sondage aléatoire simple stratifié sur les deux populations précédemment mises en avant. Pour cela, il nous faut définir n^1_{init} et n^2_{init} les tailles des échantillons sur chacune des deux populations dans le plan de sondage initial. On fixe la taille de l'échantillon total $n = 200$.

On réalise une allocation optimale de Neyman vis à vis de la variable X, avec prise en compte des taux de réponse anticipés par strates [5], pour déterminer n^1_{init} et n^2_{init} :

$$n^i_{\text{init}} = n \frac{\frac{N_i S_i}{\sqrt{\rho_i}}}{\sum_{i=1}^{2} \frac{N_i S_i}{\sqrt{\rho_i}}} \tag{4}$$

où S_i est la dispersion de X dans la population $\mathcal{P}_i$. On obtient $n^1_{\text{init}} = 90$ et $n^2_{\text{init}} = 110$.

Résultats En appliquant le programme de minimisation avec le $\lambda_{\text{num}} = 7352.131$ obtenu par la méthode numérique de calcul de la partie 4.2.2, on obtient les résultats suivants :

$$\begin{aligned} n^1_{\text{init}} &= 90 \qquad n^2_{\text{init}} = 110 \\ n^1_f &= 137 \qquad n^2_f = 63 \end{aligned}$$

On remarque que le nombre d'individus échantillonnés dans la population de "patrimoines standards" a augmenté par rapport à l'échantillon initial : cela est dû à l'effet de minimisation de dispersion des poids finaux, ceux-ci étant plus importants dans la population de "patrimoines standards", qui sont plus nombreux même si meilleurs répondants. On remarque également que la solution obtenue n'est pas extrême - ni (90,110) ni (200,0) - ce qui est un résultat intéressant.

La fonction $V(\lambda)$ obtenue par simulations et calcul de la variance empirique est en Figure 5. On remarque la décroissance et la présence d'un coude de la fonction, c'est à dire d'un point de torsion, situé à $\lambda_{\text{coude}} \approx 3000 \leq \lambda_{\text{num}}$, ce qui satisfait à la conjecture.

L'application de l'algorithme CURIOS à un exemple simple permet de constater qu'il a un comportement non trivial, différent de celui de l'allocation de Neyman.

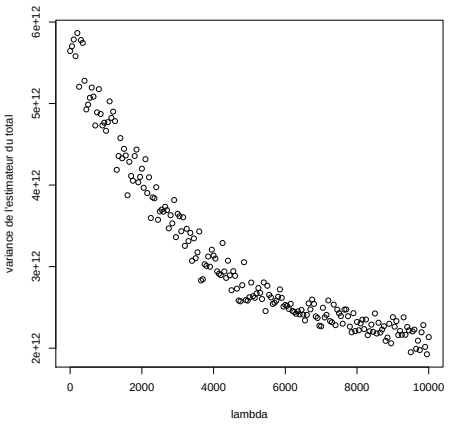

FIGURE 5 – Forme de $V(\lambda)$.

4.3 Typologie des critères

Les différentes contraintes intervenant dans le programme d'optimisation à la base de l'algorithme CURIOS peuvent être classées dans trois catégories. Le schéma de la Figure 6 résume graphiquement les trois types d'objectifs poursuivis.

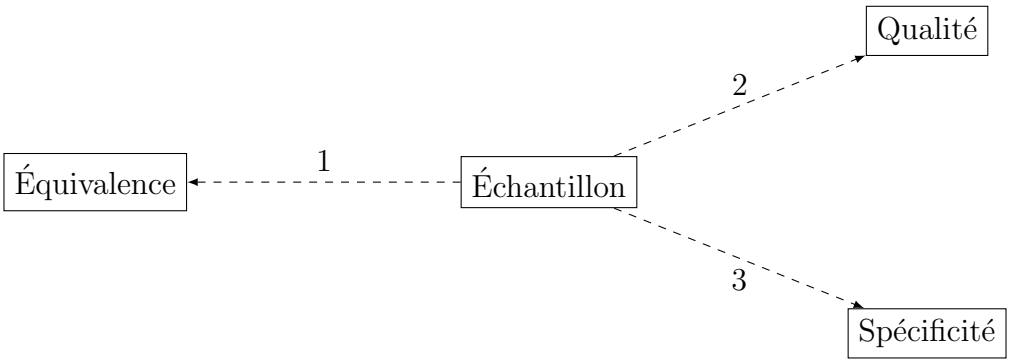

FIGURE 6 – Typologie des objectifs poursuivis par l'algorithme CURIOS.

Détaillons plus précisément chacun de ces trois objectifs :

1. **Équivalence :** Ici, tous les ménages sont supposés équivalents au sein de l'échantillon. Cela suppose en particulier que les poids de chacun des ménages, y compris après correction de la non-réponse, ne doivent pas être de variance trop grande, ce qui conduit à minimiser cette variance. De manière plus large, on peut également s'intéresser à une équivalence interne à un ensemble de strates (dans le cadre d'enquêtes entreprises par exemple), et donc chercher à minimiser la dispersion des poids corrigés de la non-réponse dans chacune des strates.
2. **Qualité :** On souhaite obtenir une composition de l'échantillon des répondants qui soit équilibrée 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 conditionnellement à ces variables et d'assurer la qualité des estimations. Le R-indicateur et ses déclinaisons partielles sont de telles mesures de similitude

entre répondants et population totale. Il est également possible de s'intéresser à des indicateurs de type *Balance Indicators* [3], ou plus directement à la variance liée à la non-réponse pour un estimateur corrigé de la non-réponse [7].

3. **Spécificité :** Le processus d'enquête a été décidé dans un objectif spécifique vis-à-vis des variables d'intérêt de celle-ci. On peut par exemple considérer l'allocation de Neyman dans le cadre d'une variable d'intérêt unique, ou plus simplement utiliser le plan de sondage initialement spécifié, basé sur les connaissances d'experts. On cherche donc à minimiser la distance à ce plan initial, garant des objectifs de précision fixés pour l'enquête.

Conclusion

L'algorithme CURIOS présenté dans cet article permet la mise en œuvre d'une méthode de priorisation entre plusieurs vagues d'une enquête en face-à-face réalisée par l'INSEE. À l'aide d'indicateurs de représentativité, il s'agit d'adapter l'échantillonnage à la fin d'une vague afin de permettre le rééquilibrage de la population des répondants.

Cette méthode a été testée dans le cadre de l'enquête Patrimoine 2014, et a permis d'adapter légèrement les échantillons pour la deuxième phase de collecte. Il conviendra d'étudier l'effet a posteriori, tout en sachant qu'il ne pourra être que minime compte tenu du faible nombre de ménages concernés.

À l'avenir, la méthode de priorisation pourra être généralisée à l'ensemble des enquêtes ménages réalisées en plusieurs phases, ou au moins sur une longue période. Cela demandera néanmoins une période d'adaptation à l'ensemble des acteurs de la chaîne de réalisation des enquêtes, afin de pouvoir effectuer de manière efficace l'algorithme CURIOS : remontée des informations de collecte, validation définitive des nouveaux échantillons, réimpression des fiches-adresses concernées.

Références

[1] Rebecq A. and Merly-Alpa T. Pourquoi minimiser la dispersion des poids en sondage ? *preprint*.

[2] 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.

[3] Särndal C-E. The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27(1) :1–21, 2011.

[4] Haziza D. and Beaumont J-F. On the construction of imputation classes in surveys. *International Statistical Review*, 75(1) :25–43, 2007.

[5] Gros E., Brion P., and Deroyon T. Formation aux méthodes de traitement d'enquêtes auprès des entreprises. 2014.

[6] Jones E., Oliphant T., Peterson P., et al. SciPy : Open source scientific tools for Python, 2001–. [Online ; accessed 2015-03-13].

[7] Beaumont J-F., Bocci C., and Haziza D. An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30(4) :607–621, 2014.

[8] Christine M. and Faivre S. Le projet OCTOPUSSE de nouvel Echantillon-Maître de l'INSEE. *JMS*, 2009 :24, 2009.

[9] Koubi M. and Mathern S. Résolution d'une des limites de l'allocation de Neyman. *JMS*, 2009 :1, 2009.

[10] Ardilly P. *Les techniques de sondage*. Editions Technip, 2006.

[11] Merly-Alpa T. L'utilisation de R-indicateurs pour la priorisation des enquêtes ménages en cours de collecte. 2014.

[12] Merly-Alpa T. and Rebecq A. L'utilisation des R-indicateurs pour "prioriser" la collecte des enquêtes ménages : une application à l'enquête Patrimoine 2010. *JMS*, 2015.

[13] McKinney W. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.