L'algorithme CURIOS pour l'optimisation du plan de sondage en fonction de la non-réponse

Thomas Merly-Alpa [1] & Antoine Rebecq [2]

[1] *thomas.merly-alpa@insee.fr. INSEE, 18 boulevard Adolphe Pinard, 75014 Paris.*
[2] *antoine.rebecq@insee.fr. INSEE, 18 boulevard Adolphe Pinard, 75014 Paris.*

Résumé. La non-réponse est un problème épineux en sondages, car la théorie a été construite sur l'hypothèse d'une participation totale de l'échantillon à l'enquête. Or les mécanismes de réponse sont mal connus, et les estimateurs corrigés de la non-réponse peuvent présenter de larges biais résiduels. Usuellement, la non-réponse est traitée en fin de collecte, en utilisant des techniques telles que le calage. Nous pensons qu'il est souhaitables de tenir compte des mécanismes de réponse de la population enquêtée dès la phase d'échantillonnage. C'est pourquoi nous présentons ici l'algorithme CURIOS (*Curios Uses Representativity Indicators to Optimize Samples*) qui vise à construire un meilleur échantillon en résolvant un problème d'optimisation. Celui-ci consiste en un compromis entre un indicateur de dispersion minimale des poids corrigés de la non-réponse, et un indicateur de similarité avec une allocation initiale, qu'on assimilera ici avec l'allocation de Neyman avec prise en compte de la non-réponse. Nous donnons ici une méthode basée sur l'étude de la variance d'un estimateur du total d'une variable d'enquête, méthode démontrée analytiquement sous de bonnes conditions, i.e dans le cas d'un problème classique rencontré par les instituts nationaux de statistique. Nous indiquons également une méthode numérique empirique permettant de tester les allocations obtenues sur différents scénarios. Enfin, nous réaliserons de telles simulations dans le cadre très simple d'un sondage stratifié et d'une non-réponse uniforme par strate afin d'étudier les résultats obtenus par l'algorithme.

Mots-clés. Sondages, échantillonnage adaptatif, non-réponse, optimisation.

Abstract. Non-response is one of the main issues in survey sampling, because the theory was built assuming every sampled individual would participate. However no such ideal conditions exist in reality. Estimators taking non-response into account are often biased, mainly because non-response mechanisms are not very well understood. Non-response is usually dealt with during the post-treatment phase of the survey, using methods such as calibration. We think that it is possible to take into account the response behaviour during the sampling phase. We present here the CURIOS (*Curios Uses Representativity Indicators to Optimize Samples*) algorithm which aims at solving an optimization problem which is a compromise between the dispersion of non-response adjusted weights and the distance to an initial sample, which was computed by Neyman allocation with non-response. We discuss here a method based on the study of the variance of the estimator

of an interest variable, for which we give a proof under some conditions, easily met in current scenarios encountered in national statistics. We also mention a numerical approach to the problem, which can be used to compute the algorithm. We then simulate a simple population and a stratified sample with uniform non-response in each strata, and analyse the results given by CURIOS.

Keywords. Survey sampling, adaptative sampling, non-response, optimization.

1 But 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.

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 (en particulier pour l'estimation de la précision, cf le chapitre V de [2]), qui suppose que 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. C'est déjà l'objectif principal de la procédure d'échantillonnage OCTOPUSSE [3].

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 [4] dans laquelle on considère des prévisions de comportement moyen de réponse $\tilde{\rho}$ dans les strates (cf équation 3). Or il est bien connu que l'optimum de Neyman est plat [2], on conserve donc ses bonnes propriétés de minimisation de la variance des estimateurs à distance faible de l'allocation initiale.

2 Algorithme CURIOS

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 :

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

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 maximisation 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 $(n^1_f, n^2_f) = (n^1_{\text{init}}, n^2_{\text{init}})$. 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.

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 du total de X, variable d'intérêt de l'enquête. Celle-ci dépend de la valeur de λ via les tailles d'échantillons n^i_f obtenues pour une telle valeur. On a le théorème suivant. L'obtention de la forme de l'hypothèse 2 est expliquée en Annexe 1.

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

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

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 1 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 1. 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 2 est vérifiée, on a :*

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

3 Simulations

3.1 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 2 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.

3.2 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} :

$$n^i_{\text{init}} = n \frac{\frac{N_i S_i}{\sqrt{\rho_i}}}{\sum_{i=1}^{2} \frac{N_i S_i}{\sqrt{\rho_i}}} \tag{3}$$

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

3.3 Résultats

En appliquant le programme de minimisation avec le $\lambda_{\text{num}} = 7352.131$ obtenu par la méthode numérique de calcul de la partie 2.2, on obtient les résultats suivants :

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

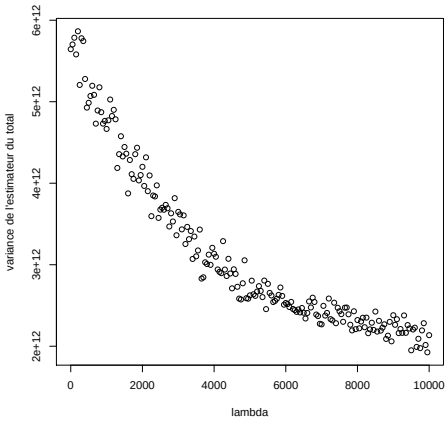

FIGURE 1 – Forme de $V(\lambda)$.

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 1. 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.

4 Conclusion

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.

L'étude menée ici gagnerait à être étendue à des cas plus généraux : nombre de strates supérieur à 2, comportement de réponse hétérogènes au sein des strates...

En pratique, cette procédure a été mise en place pour l'enquête Patrimoine 2014 de l'INSEE [6], en utilisant en lieu et place de la distance à l'allocation de Neyman les R-indicateurs définis par Schouten [7].

Références

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

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

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

[4] Neyman J. On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, pages 558–625, 1934.

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

[6] 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.

[7] 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.

Annexe 1 : Forme de l'équation 2

On s'intéresse à la fonction CURIOS($\mathrm{n}_\mathrm{f}^1, \mathrm{n}_\mathrm{f}^2$) que l'on cherche à maximiser en n_f^1 dans le programme de l'équation 1, c'est à dire :

$$\begin{aligned}
\mathrm{CURIOS}(\mathrm{n}_\mathrm{f}^1, \mathrm{n}_\mathrm{f}^2) &= \mathrm{Disp}(\mathrm{p}_\mathrm{CNR}^\mathrm{k}) + \lambda\ \mathrm{Dist}((\mathrm{n}_\mathrm{f}^1, \mathrm{n}_\mathrm{f}^2), (\mathrm{n}_\mathrm{init}^1, \mathrm{n}_\mathrm{init}^2)) \\
&= \frac{\rho_1 n_f^1}{N-1}\left(p_{\mathrm{CNR}}^1 - \bar{p}\right)^2 + \frac{\rho_2 n_f^2}{N-1}\left(p_{\mathrm{CNR}}^2 - \bar{p}\right)^2 + \lambda\sqrt{\left(n_f^1 - n_{\mathrm{init}}^1\right)^2 + (n_f^2 - n_{\mathrm{init}}^2)^2} \\
&= \frac{\rho_1 n_f^1}{N-1}\left(\frac{N_1}{n_f^1\rho_1} - \bar{p}\right)^2 + \frac{\rho_2 (n - n_f^1)}{N-1}\left(\frac{N_2}{(n-n_f^1)\rho_2} - \bar{p}\right)^2 \\
&\quad + \lambda\sqrt{\left(n_f^1 - n_{\mathrm{init}}^1\right)^2 + ((n - n_f^1) - n_{\mathrm{init}}^2)^2}
\end{aligned}$$

en utilisant le fait que $n_f^1 + n_f^2 = n$. D'autre part, on a :

$$\begin{aligned}
\bar{p} &= \frac{n_f^1\rho_1 p_{\mathrm{CNR}}^1 + n_f^2\rho_2 p_{\mathrm{CNR}}^2}{n_f^1\rho_1 + n_f^2\rho_2} \\
&= \frac{1}{n_f^1\rho_1 + n_f^2\rho_2}\left(n_f^1\rho_1\frac{N_1}{n_f^1\rho_1} + n_f^2\rho_2\frac{N_2}{n_f^2\rho_2}\right) \\
&= \frac{N}{n_f^1\rho_1 + n_f^2\rho_2}
\end{aligned}$$

et donc :

$$\begin{aligned}
\mathrm{CURIOS}(\mathrm{n}_\mathrm{f}^1, \mathrm{n}_\mathrm{f}^2) &= \frac{\rho_1 n_f^1}{N-1}\left(\frac{N_1}{n_f^1\rho_1} - \frac{n}{n_f^1\rho_1 + n_f^2\rho_2}\right)^2 + \frac{\rho_2 (n - n_f^1)}{N-1}\left(\frac{N_2}{(n-n_f^1)\rho_2} - \frac{n}{n_f^1\rho_1 + n_f^2\rho_2}\right)^2 \\
&\quad + \lambda\sqrt{\left(n_f^1 - n_{\mathrm{init}}^1\right)^2 + ((n - n_f^1) - (n - n_{\mathrm{init}}^1)^2} \\
&= \frac{\rho_1 n_f^1}{N-1}\left(\frac{N_1}{n_f^1\rho_1} - \frac{n}{n_f^1\rho_1 + (n-n_f^1)\rho_2}\right)^2 \\
&\quad + \frac{\rho_2 (n - n_f^1)}{N-1}\left(\frac{N_2}{(n-n_f^1)\rho_2} - \frac{n}{n_f^1\rho_1 + (n-n_f^1)\rho_2}\right)^2 + \lambda\sqrt{2}|n_f^1 - n_{\mathrm{init}}^1|
\end{aligned}$$

On considère donc la fonction suivante :

$$F(x) = Ax\left(\frac{C}{x} - \bar{p}\right)^2 + B(n-x)\left(\frac{D}{n-x} - \bar{p}\right)^2 + \lambda\sqrt{\left(x - n_{\mathrm{init}}^1\right)^2 + ((n-x) - n_{\mathrm{init}}^2)^2}$$

où :

$$A = \frac{\rho_1}{N-1} \quad C = \frac{N_1}{\rho_1}$$
$$B = \frac{\rho_2}{N-1} \quad D = \frac{N_2}{\rho_2}$$

On développe et on a :

$$F(x) = \frac{AC^2}{x} + \frac{BD^2}{n-x} + x(A\bar{p}^2 - B\bar{p}^2) - 2AC\bar{p} + Bn\bar{p} - 2BD\bar{p} + \lambda\sqrt{2}|n^1_{\text{init}} - x|$$

On se place dans le cas où $x \geq n^1_{\text{init}}$:

$$F(x) = \frac{AC^2}{x} + \frac{BD^2}{n-x} + x(A\bar{p}^2 - B\bar{p}^2 + \lambda\sqrt{2}) - 2AC\bar{p} + Bn\bar{p} - 2BD\bar{p} - \lambda\sqrt{2}n^1_{\text{init}}$$

On calcule la dérivée :

$$F'(x) = \frac{-AC^2}{x^2} + \frac{BD^2}{(n-x)^2} + \bar{p}^2(x)(A-B) - \bar{p}'(x)[2xB\bar{p}(x) - 2xA\bar{p}(x) + AC - Bx + 2BD] + \lambda\sqrt{2}$$

où $\bar{p}(x)$ et ses dérivées successives $\bar{p}'(x)$ et $\bar{p}''(x)$ valent respectivement :

$$\bar{p}(x) = \frac{N}{x\rho_1 + (n-x)\rho_2}$$
$$\bar{p}'(x) = \frac{N(\rho_2 - \rho_1)}{(x\rho_1 + (n-x)\rho_2)^2}$$
$$\bar{p}''(x) = \frac{2N(\rho_2 - \rho_1)^2}{(x\rho_1 + (n-x)\rho_2)^3}$$

Cette dérivée s'annule lorsque :

$$\lambda\sqrt{2} = \frac{AC^2}{x^2} - \frac{BD^2}{(n-x)^2} + \bar{p}^2(x)(B-A) + \bar{p}'(x)[2xB\bar{p}(x) - 2xA\bar{p}(x) + AC - Bx + 2BD]$$

On dérive des deux côtés par rapport à x, et on a :

$$\frac{d\lambda}{dx} = \frac{-2AC^2}{x^3} - \frac{2BD^2}{(n-x)^3} + 2\bar{p}'(x)(B-A)(2\bar{p}(x) + x\bar{p}'(x)) + \bar{p}''(x)(2x\bar{p}(x)(B-A) + 2AC - Bn + 2BD)$$
$$=: g(x) + h(x)$$

où :

$$g(x) = \frac{-2AC^2}{x^3} - \frac{2BD^2}{(n-x)^3}$$
$$h(x) = 2\bar{p}'(x)(B-A)(2\bar{p}(x) + x\bar{p}'(x)) + \bar{p}''(x)(2x\bar{p}(x)(B-A) + 2AC - Bn + 2BD)$$

La fonction g est strictement négative sur l'intervalle $[n^1_{\text{init}}; n]$. Le signe de la fonction h est plus compliqué à analyser : on va utiliser la condition 2 pour minorer h devant g. En effet, si l'on a $\forall x,\ |h(x)| \leq \min |g|$, alors F' sera du signe de g et donc négative sur l'intervalle $[n^1_{\text{init}}; n]$.

Quand est-ce que $|g| = -g$ est minimale ? On dérive g :

$$g'(x) = \frac{6AC^2}{x^4} - \frac{6BD^2}{(n-x)^4}$$

et donc $|g|$ est minimale (g maximale) en x_0 :

$$\begin{aligned}
g'(x_0) &= 0 \\
\frac{6AC^2}{x_0^4} &= \frac{6BD^2}{(n-x_0)^4} \\
\frac{(n-x_0)^4}{x_0^4} &= \frac{6BD^2}{6AC^2} \\
\frac{n}{x_0} - 1 &= \left(\frac{N_2^2\rho_1}{N_1^2\rho_2}\right)^{1/4} \\
\frac{1}{x_0} &= \frac{1}{n}\left[1 + \left(\frac{N_2^2\rho_1}{N_1^2\rho_2}\right)^{1/4}\right] \\
x_0 &= n\left[1 + \left(\frac{N_2^2\rho_1}{N_1^2\rho_2}\right)^{1/4}\right]^{-1}
\end{aligned}$$

La condition se réécrit donc $\forall x,\ |h(x)| \leq |g(x_0)|$. On remarque que le terme de droite correspond à celui de l'hypothèse 2 à un facteur $(N-1)$ près. Il ne reste donc plus qu'à majorer $h(x)$ par le terme de gauche (à un facteur $(N-1)$ près) pour tout x.