REDRESSEMENT DE LA NON-RÉPONSE ET CALAGE DANS LES ENQUÊTES COUPLÉES

Marie CORDIER-VILLOING (*), Olivier SAUTORY (**)

(*) Insee, Département des statistiques de court-terme
(**) Insee, Unité méthodologie statistique entreprises

Introduction

Un dispositif d'enquêtes est dit couplé lorsqu'il interroge plusieurs unités statistiques différentes mais entre lesquelles existe un lien. Suivant les dispositifs d'enquêtes, différentes stratégies d'échantillonnages des enquêtes couplées sont utilisées [1]. Les stratégies « ascendantes » sont assimilables à des échantillonnages indirects, les unités intermédiaires (ex : ménages) échantillonnées pointant ensuite vers les unités finales (ex : entreprises dans lesquelles travaillent les personnes de référence des ménages). Les stratégies « descendantes » sont assimilables à des échantillonnages en plusieurs degrés avec dans un premier temps l'échantillonnage des unités primaires (ex : ménages) et dans un deuxième temps celui d'unités secondaires au sein de chacune des unités primaires (ex : individus).

L'approchante ascendante d'échantillonnage des enquêtes couplées est plutôt adaptée aux enquêtes orientées sur les unités intermédiaires. Prenons l'exemple d'une enquête couplée où les salariés sont d'abord échantillonnés et pointent ensuite vers l'échantillon des entreprises dans lesquelles ils travaillent. Cette méthode de sondage n'assure pas que toutes les unités finales d'une sous-population particulière soient exhaustivement interrogées, comme on pourrait le souhaiter dans le cas d'enquêtes portant sur les entreprises où l'on assure l'interrogation exhaustive des entreprises les plus importantes. C'est pourquoi nous nous sommes, dans cet article, centrés sur les enquêtes couplées avec stratégies d'échantillonnages « descendantes », pour lesquelles l'échantillon des unités primaires comme celui des unités secondaires pourront être utilisés indépendamment l'un de l'autre pour l'élaboration de statistiques.

Cet article a pour objectif d'étudier l'intérêt de calages simultanés des différents volets d'une enquête couplée, par rapport à des calages indépendants : nous mettrons en avant les propriétés de ces calages, ainsi que l'impact sur ces propriétés de la présence de non-réponse au niveau de chacun des volets. Plusieurs méthodes de calages simultanés sont décrites par Estevao et Särndal [2]. L'une d'entre elles est utilisée à l'Insee depuis les années 90 [3], et a été programmée dans une deuxième version de la macro Calmar [4].

Dans un premier temps, nous présenterons l'apport d'un calage simultané tel qu'il est réalisé par Calmar 2. Dans le cas d'un sondage en grappes, plusieurs propriétés se dégagent : la possibilité d'établir des statistiques sur les unités primaires à partir du fichier des unités secondaires ; des estimations sur des effectifs d'unités secondaires identiques à partir des échantillons des unités primaires et secondaires ; des estimations sur des effectifs d'unités primaires identiques à partir des deux échantillons. Pour un sondage en deux degrés, seule la dernière propriété est obtenue. Les conditions nécessaires à ces propriétés portent sur l'échantillonnage des unités secondaires, qui doit être équilibré sur la taille des unités primaires (cas des sondages aléatoires simples au deuxième degré ou de SAS stratifiés au deuxième degré).

Dans un deuxième temps, nous prendrons en compte la présence de non-réponse au niveau des unités primaires comme des unités secondaires. Deux modes de correction de la non-réponse sont possibles : une correction en amont du calage, ou faire la correction de la non-réponse en même temps que le calage en introduisant dans les variables de calage les variables explicatives de la non-réponse. Nous verrons que la non-réponse des unités secondaires a un impact sur les propriétés de cohérence des estimations, qui ne sont alors plus respectées. La non-réponse des unités primaires, quel que soit le mode de correction choisi, n'a au contraire pas d'impact sur ces propriétés.

Enfin, nous étudierons deux autres méthodes de calages simultanés présentées par Estevao et Särndal, qui assurent la cohérence des estimations portant sur les effectifs d'unités primaires obtenues à partir des deux échantillons, en l'absence ou en présence de non-réponse.

1. Le calage simultané pratiqué à l'Insee et ses propriétés en l'absence de non-réponse

Pour une enquête multi-niveaux (« ménages / individus », « entreprises / salariés »), des calages indépendants des ménages et des individus, ou des entreprises et des salariés, n'assurent pas que des statistiques provenant de chacun de ces niveaux soient cohérentes.

Par exemple, pour une enquête « ménages / individus », avec un plan de sondage en grappes, les estimations après calage de la répartition des chefs de ménage par catégories socioprofessionnelles, qui peuvent être établies à partir des échantillons de ménages ou d'individus, n'ont aucune raison de coïncider. Autre exemple, pour une enquête « entreprises / salariés » avec sondage en deux degrés, les estimations du nombre de salariés par secteur d'activité calculées à partir des deux volets ne coïncideraient pas.

Dans le cas de calages indépendants de différents niveaux d'une enquête couplée, ces cohérences entre les estimations provenant de chacun des volets pourront être assurées si les totaux des variables concernées sont connus par ailleurs sur la population entière : on peut alors utiliser ces variables comme variables de calage de chacun des niveaux, ce qui assure ipso facto l'égalité des estimations avec les vrais totaux.

1.1. Quelques exemples d'enquêtes couplées

Les enquêtes « COI » (« Changements Organisationnels et Informatisation » [5]) sont des exemples d'enquêtes couplées avec stratégie « descendante » : le plan de sondage est un tirage à deux degrés, le premier degré de tirage concerne les employeurs (unités primaires, ou UP) et le deuxième, les salariés (unités secondaires, ou US).

L'enquête sur le coût de la main d'œuvre et la structure des salaires (ECMOSS) interroge 14 000 établissements employeurs appartenant à des entreprises de 10 salariés ou plus, et recueille ensuite des informations pour un échantillon de 120 000 salariés sélectionnés au sein de ces établissements.

Dans l'enquête sur les recours urgents ou non programmés à la médecine générale de ville, le premier volet de l'enquête s'adressait aux médecins. Les généralistes enquêtés étaient invités à remplir un questionnaire sur leur activité et autant de questionnaires sur les consultations que de recours urgents ou non programmés pris en charge. Le second volet s'adressait aux patients examinés en « urgence » par ces médecins, et qui n'ont pas été hospitalisés immédiatement après la consultation.

1.2. Notations

Ces notations seront utilisées tout au long de cet article.

- ω_j : pondération initiale[1] de l'unité primaire j

- N_j : taille (en nombre d'unités secondaires) de l'unité primaire j, que l'on appellera par la suite **effectif** de l'UP,

- s_I : échantillon des unités primaires j,

- $\omega_{i,j}$: pondération initiale de l'unité secondaire i, appartenant à l'unité primaire j,

- $\omega_{i/j} = \frac{\omega_{i,j}}{\omega_j}$: pondération initiale de l'unité secondaire i conditionnelle à l'unité primaire j,

- s_j : échantillon des unités secondaires i appartenant à l'unité primaire j,

[1] ou poids de sondage

- on notera $\mathbf{X_p}$ une variable auxiliaire de calage définie sur les unités primaires, où p=1…P, $x_{j,p}$ la valeur de cette variable sur l'unité primaire j, et X_p son total sur la population entière des UP.

- on notera $\mathbf{X_s}$ une variable auxiliaire de calage définie sur les unités secondaires, où s=1…S, $x_{ij,s}$ la valeur de cette variable sur l'unité secondaire i (de l'UP j), et X_s son total sur la population entière des US.

1.3. Les calages indépendants

Le principe général d'un calage est le suivant : les nouvelles pondérations calées des unités échantillonnées sont calculées aussi proches que possible des pondérations initiales, et avec ces pondérations l'échantillon aura la même structure que l'ensemble de la population étudiée sur certaines variables[2]. Les calages des unités primaires et des unités secondaires sont réalisés indépendamment l'un de l'autre.

Premier niveau de calage : unités primaires

On cherche des pondérations w_j qui sont à la fois proches des pondérations initiales ω_j de l'échantillon s_I et qui vérifient pour toutes les variables auxiliaires de calage $\mathbf{X_p}$, avec p=1…P, les équations de calage suivantes :

$$(1) : \forall p \in \{1...P\} : \; X_p = \sum_{j \in s_I} w_j \; x_{j,p}$$

Par exemple, pour des unités primaires « entreprises », on peut caler sur la structure par secteur d'activité, tranche d'effectifs, tranche d'unité urbaine, etc.

Deuxième niveau de calage : unités secondaires

Les équations de calage suivantes seront vérifiées par les nouvelles pondérations $w_{i,j,}$ pour toutes les variables auxiliaires de calage $\mathbf{X_s}$, avec s=1…S,

$$(2) : \forall s \in \{1...S\} : \; X_s = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \; x_{ij,s}$$

Par exemple, pour des unités secondaires « salariés », on peut caler sur le nombre de salariés par tranche d'âge, genre, catégorie socioprofessionnelle, tranche d'unité urbaine, etc.

Estimations portant sur des variables présentes à la fois dans les volets « unités primaires » et « unités secondaires ».

Prenons l'exemple d'une variable **Y** présente au niveau des unités secondaires, comme le revenu d'un individu, ou une indicatrice « être une femme » pour un salarié, ou encore une indicatrice d'appartenance d'un salarié à une entreprise d'un secteur d'activité donné.

On notera Y le total de la variable sur la population des US (exemples : revenu total des individus dans la population, nombre total de femmes salariées dans la population, nombre total de salariés d'un secteur d'activité).

On notera $y_{i,j}$ la valeur de **Y** sur l'US i de l'UP j, et $\hat{Y}_s$ son total estimé à partir de l'échantillon des US.

[2] en termes de distributions de variables qualitatives, ou de totaux de variables quantitatives.

On suppose que cette variable est également disponible au niveau des unités primaires, comme par exemple le revenu du ménage j, ou l'effectif salarié féminin de l'entreprise j, ou encore l'effectif salarié de l'entreprise × indicatrice d'appartenance à un secteur d'activité. Elle se calcule comme la somme des valeurs de la variable sur toutes les unités secondaires de l'unité primaire, échantillonnées ou non : $y_j = \sum_{i \in j} y_{i,j}$

Le total de cette variable sur la population des UP est encore égal à Y.

On notera $\hat{Y}_p$ son total estimé à partir de l'échantillon des UP.

Les deux estimateurs de Y se calculent comme suit :

- Via l'échantillon « unités primaires » : $\hat{Y}_p = \sum_{j \in s_I} w_j \, y_j$

- Via l'échantillon « unités secondaires » :

$$\hat{Y}_s = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, y_{i,j} = \sum_{j \in s_I} w_j \sum_{i \in s_j} w_{i/j} \, y_{i,j} = \sum_{j \in s_I} w_j \, \hat{y}_j$$

où $w_{i/j} = w_{i,j} / w_j$ désigne la nouvelle pondération conditionnelle de l'US j dans l'UP i, et $\hat{y}_j = \sum_{i \in s_j} w_{i/j} \, y_{i,j}$ désigne l'estimation du total de **Y** au sein de l'UP j.

Les estimateurs $\hat{Y}_p$ et $\hat{Y}_s$ sont sans biais, d'espérance Y, mais sont en général différents, car $\hat{y}_j \neq y_j$.

Dans le cas de calages sur marges indépendants, les estimations de totaux de variables connues à la fois via l'échantillon « unités primaires » et l'échantillon « unités secondaires » sont donc en général différentes, sauf si les totaux de ces variables sont connus par ailleurs sur la population totale, et introduits comme information auxiliaire dans le calage.

Un des objectifs que nous chercherons à atteindre en utilisant diverses méthodes de calage simultané est d'assurer la cohérence entre des estimateurs qui peuvent être calculés à partir des unités primaires ou des unités secondaires.

1.4. La méthode de calage simultané via les « unités primaires » : méthode (A)

La méthode de calage simultané communément utilisée à l'Insee et programmée dans la deuxième version de la macro Calmar assure (par construction même) que les poids conditionnels des unités secondaires (US) ne sont pas modifiés par le calage.

Cette méthode de calage simultané est une méthode de calage effectuée au niveau des unités primaires :

- les variables de calage des unités secondaires sont « remontées » au niveau des unités primaires,
- le calage est réalisé au niveau des unités primaires,
- les variables de calage portent à la fois sur les unités secondaires et les unités primaires.

On réalise le calage au niveau unité primaire, **les poids calés au niveau unités primaires** w_j **vérifiant deux systèmes d'équations de calage :**

$$(1): \forall p \in \{1...P\}: X_p = \sum_{j \in s_I} w_j \, x_{j,p}$$

$$(2'): \forall s \in \{1...S\}: X_s = \sum_{j \in s_I} w_j \, \hat{x}_{j,s}$$

avec $\hat{x}_{j,s} = \sum_{i \in s_j} \omega_{i/j} \, x_{ij,s}$ l'estimateur sans biais du total de la variable de calage $\mathbf{X_s}$ pour une UP j.

Les poids des unités secondaires sont ensuite calculés tels que :

$$(3): w_{i,j} = \omega_{i/j} \times w_j$$

Cette méthode assure donc (par construction même) que les poids conditionnels des US dans les UP ne sont pas modifiés par le calage.

Vérifions que les équations (2') assurent que les poids des unités secondaires sont calés sur les totaux des variables auxiliaires de calage des unités secondaires, soit les équations (2).

$$(2'): \forall s \in \{1...S\}: X_s = \sum_{j \in s_I} w_j \, \hat{x}_{j,s} = \sum_{j \in s_I} w_j \sum_{i \in s_j} \omega_{i/j} \, x_{ij,s} = \sum_{j \in s_I} \sum_{i \in s_j} w_j \, \omega_{i/j} \, x_{ij,s} = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, x_{ij,s} \Rightarrow (2)$$

Les poids calculés après calage simultané respectent plus de contraintes que pour les calages indépendants. Les rapports de poids après calage indépendants seront donc moins dispersés qu'avec le calage simultané. Cependant, plusieurs propriétés de cohérence, présentées ci-dessous, sont respectées après un calage simultané conservant les poids conditionnels de tirage des unités secondaires.

1.4.1. Cas d'un sondage en grappes

Dans le cas d'un sondage en grappes, les poids conditionnels sont égaux à 1, par conséquent les poids d'échantillonnage des unités secondaires appartenant à la même unité primaire sont égaux au poids d'échantillonnage de cette unité primaire.

Conserver cette égalité des poids **après calage**, ce que n'autorisent pas des calages indépendants des deux volets de l'enquête, permet - dans le cas d'un sondage en grappes - d'obtenir trois propriétés intéressantes.

- **Établir des statistiques sur les UP à partir du fichier d'US (prop 1)**

L'égalité entre les poids des US qui appartiennent à une même UP (indépendamment d'ailleurs de l'existence d'un fichier contenant des informations collectées sur les UP), permet - si l'on dispose de variables sur les UP au niveau des US - d'établir des statistiques sur les UP à partir du fichier d'US. Le poids d'une UP est alors le poids d'une US quelconque appartenant à cette UP.

De plus, l'égalité entre poids d'une UP et poids des US appartenant à l'UP assure que les statistiques sur les UP établies à partir du fichier d'US sont alors bien cohérentes avec les statistiques établies à partir du fichier d'UP.

Par exemple, dans le cas d'une enquête « ménages / individus », si on dispose dans le fichier des US des caractéristiques du ménage ou du logement (par exemple taille du ménage, nombre de pièces du logement), les répartitions des ménages par taille estimées à partir du fichier des ménages ou du fichier des individus coïncideront, de même que les répartitions des logements selon le nombre de pièces.

De plus, si on connaît pour chaque ménage la catégorie socioprofessionnelle (CS) de la personne de référence du ménage, et pour chaque individu sa CS et son lien avec la personne de référence du ménage, la répartition des ménages selon la CS de la personne de référence (estimée à partir du fichier des ménages) coïncidera avec la répartition des personnes de référence des ménages par CS (estimée à partir du fichier des individus).

- **Obtenir des estimations sur des nombres d'US identiques à partir des échantillons d'UP et d'US (prop 2)**

Dans le cas d'une enquête « ménages / individus » où on collecte des informations sur tous les individus du ménage (exemple fréquent d'un sondage en grappes), on « remonte » dans l'échantillon de ménages les informations portant sur la composition des ménages en termes d'individus (nombre d'individus par genre, âge, CS).

Prenons l'exemple de l'estimation du nombre de femmes de professions intermédiaires dans la population (noté Y), via l'échantillon des « unités primaires » $\hat{Y}_p$ (ménages) et via celui des unités secondaires $\hat{Y}_s$ (individus). On note $y_{i,j}$ l'indicatrice "être une femme de professions intermédiaires" pour l'individu i du ménage j, et y_j le nombre de femmes de professions intermédiaires du ménage j.

$$\hat{Y}_s = \sum_{j\in s_I}\sum_{i\in s_j} w_{i,j}\, y_{i,j} = \sum_{j\in s_I} w_j \sum_{i\in s_j} \omega_{i/j} y_{i,j} = \sum_{j\in s_I} w_j \sum_{i\in s_j} y_{i,j} = \sum_{j\in s_I} w_j\, y_j = \hat{Y}_p$$

Dans le cas d'un sondage en grappes, et sous l'hypothèse d'absence de non-réponse au deuxième degré, l'égalité entre les poids après calage des US d'une même UP et le poids après calage de l'UP assure que les estimations portant sur des effectifs d'unités secondaires établies à partir de l'échantillon des UP coïncident avec celles établies à partir de l'échantillon des US.

Cette cohérence demeure pour une variable **Y** quantitative, comme le revenu : $y_{i,j}$ désigne le revenu de l'individu i du ménage j, y_j désigne le revenu total du ménage j, et Y le revenu total dans la population.

- **Obtenir des estimations sur des effectifs de regroupements d'UP identiques à partir des échantillons d'UP et d'US (prop 3)**

Toujours dans le cas d'un sondage en grappes et en absence de non-réponse au deuxième degré, estimons un nombre d'US d'un regroupement G quelconque d'UP, en supposant connus dans le fichier des UP les effectifs N_j des UP :

- Via l'échantillon des « unités primaires » : $\hat{N}_{G,p} = \sum_{j\in s_I} w_j\, N_j\, \mathbb{1}_{j\in G}$

- Via l'échantillon des « unités secondaires » :

$$\hat{N}_{G,s} = \sum_{j\in s_I}\sum_{i\in s_j} w_{i,j}\, \mathbb{1}_{j\in G} = \sum_{j\in s_I}\sum_{i\in s_j} \omega_{i/j}\, w_j\, \mathbb{1}_{j\in G} = \sum_{j\in s_I} w_j\, \mathbb{1}_{j\in G} \sum_{i\in s_j} 1 = \sum_{j\in s_I} w_j\, \mathbb{1}_{j\in G}\, N_j$$

Par exemple, dans le cas d'une enquête « ménages / individus », les estimations du nombre d'individus appartenant à des ménages habitant un logement de 3 pièces établies à partir du fichier des ménages et du fichier des individus coïncideront.

On peut noter qu'un cas particulier de "regroupement d'UP" est le regroupement de toutes les UP, i.e. la population totale des US : dans ce cas, c'est l'effectif total de la population des US qui est estimé de façon identique à partir des deux échantillons, ce qui est une propriété en général souhaitable…

1.4.2. Cas d'un échantillonnage à deux degrés

Les stratégies de sondage des enquêtes couplées « descendantes » sont assimilables à des échantillonnages en deux degrés. **Les deux premières propriétés présentées au § 1.4.1. ne sont valables que dans le cas de sondages en grappes, car elles reposent sur l'égalité des poids des US d'une même UP avec le poids d'une UP, et sur le fait que toutes les US d'une UP sont sélectionnées.**

En revanche, la troisième propriété est respectée sous condition que le tirage du deuxième degré soit équilibré sur la taille des unités primaires.

En effet, estimons un nombre d'US d'un regroupement quelconque d'UP,

- Via l'échantillon des « unités primaires » : $\hat{N}_{G,p} = \sum_{j \in s_I} w_j \, N_j \, \mathbb{1}_{j \in G}$

- Via l'échantillon des « unités secondaires » :

$$\hat{N}_{G,s} = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, \mathbb{1}_{j \in G} = \sum_{j \in s_I} \sum_{i \in s_j} \omega_{i/j} \, w_j \, \mathbb{1}_{j \in G} = \sum_{j \in s_I} w_j \, \mathbb{1}_{j \in G} \sum_{i \in s_j} \omega_{i/j}$$

Or, dans une hypothèse de tirage SAS au 2ème degré, SAS stratifié au 2ème degré, ou plus généralement équilibré sur la taille des unités primaires, on a :

$\forall \, j \quad \sum_{i \in s_j} \omega_{i/j} = N_j$, soit $\hat{N}_{G,p} = \hat{N}_{G,s}$ et l**a cohérence entre les deux volets des estimations des effectifs d'un regroupement quelconque d'unités primaires est assurée.**

2. Effets de la non-réponse sur le calage simultané par la méthode (A)

Pour prendre en compte la non-réponse, deux méthodes sont possibles pour la corriger et caler les poids sur une information exhaustive :

- Méthode en deux temps - en commençant par corriger de la non-réponse et en calant ensuite les poids corrigés de la non-réponse sur des totaux connus sur la population entière.

- **Méthode en un temps** - en introduisant directement dans le calage les variables de correction de non-réponse, et le calage porte donc sur les poids de tirage.

Ces deux méthodes sont équivalentes lorsque la fonction de calage et la forme fonctionnelle du modèle de réponse sont exponentielles (cf. Dupont [6]).

Dans le cas d'un sondage en deux degrés, nous avons vu que la condition de cohérence entre les volets d'une même enquête, pour estimer des effectifs de regroupements d'UP (en particulier l'effectif total de la population d'US) est assurée lorsque l'on a l'égalité $\sum_{i \in s_j} \omega_{i/j} = N_j$ [3]. Elle dépend donc des poids de tirage conditionnels des unités secondaires et de l'échantillon des unités secondaires s_j. Nous supposons par la suite cette égalité sur les poids de tirage vérifiée.

En présence de non-réponse des unités secondaires, cet échantillon est réduit aux seules unités secondaires répondantes. Par ailleurs, l'introduction de la non-réponse au niveau des unités secondaires modifie les poids conditionnels des unités secondaires dans les unités primaires dans le cas où la correction de la non-réponse se fait en amont du calage sur marges : les poids conditionnels sont alors recalculés en fonction des poids des unités secondaires corrigés de la non-réponse.

La non-réponse des seules unités secondaires a un impact sur la condition de cohérence entre des estimations portant sur les volets d'une même enquête. En revanche, la non-réponse des unités primaires, quelle que soit la correction de la non-réponse effectuée, n'influe pas.

2.1. Notations

On note :

- s_I^r l'échantillon des unités primaires répondantes.

- s_j^r l'échantillon des unités secondaires répondantes dans l'unité primaire j.

[3] i.e. les tirages du 2ème degré sont équilibrés sur la taille des UP

- $\omega'_{i,j}$: les poids des US après correction de la non-réponse (CNR) des US, si CNR il y a

- ω'_j : les poids des UP après CNR des UP, si CNR il y a.

- $\omega'_{i/j}$: les poids des US conditionnels aux UP, égaux à $\omega_{i/j}$ sauf dans le cas de CNR des US, où ils sont recalculés avec les poids CNR $\omega'_{i,j}$.

- $\hat{N}_j = \sum_{i \in s_j^r} \omega'_{i/j}$ l'estimation de l'effectif de l'unité primaire à partir des poids conditionnels des unités secondaires corrigés de la non-réponse.

- N_s l'effectif total de la population des unités secondaires, $\hat{N}_s$ son estimation via l'échantillon des unités secondaires et $\hat{N}_p$ via celui des unités primaires.

2.2. Introduction de la non-réponse

2.2.1. Cas particulier où des effectifs de regroupements d'UP sont utilisés dans le calage des poids des unités secondaires

Si pour les unités secondaires, la correction du biais de non-réponse est effectuée en amont du calage :

Examinons par exemple le cas où le calage des US utilise la connaissance de l'effectif total N_s des US (ce qui est le cas dès qu'une variable qualitative intervient dans le calage).

Les poids des unités secondaires vérifient l'équation :

$$\hat{N}_s = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} = \sum_{j \in s_I^r} w_j \sum_{i \in s_j^r} \omega'_{i/j} = \sum_{j \in s_I^r} w_j \hat{N}_j = N_s$$

la dernière égalité étant obtenue grâce au calage.

Lorsque l'on estime l'effectif total des unités secondaires via les unités primaires, on obtient :

$$\hat{N}_p = \sum_{j \in s_I^r} N_j w_j = \sum_{j \in s_I^r} \hat{N}_j w_j * \frac{N_j}{\hat{N}_j} \neq \hat{N}_s = N_s$$

On obtient un résultat similaire si l'on estime un effectif de regroupement d'UP, utilisé lui-même dans le calage des US (il suffit d'ajouter l'indicatrice d'appartenance à ce groupe dans les égalités ci-dessus).

Dans le cas d'un calage simultané, la cohérence dans les estimations d'effectifs de regroupements d'UP, lorsqu'ils sont utilisés dans le calage des US, ne peut être obtenue qu'en utilisant également ces effectifs dans le calage des UP.

Si pour les unités secondaires la correction du biais de non-réponse est effectuée en même temps que le calage :

Examinons à nouveau le cas où le calage des US utilise la connaissance de l'effectif total N_s des US.

Les poids des unités secondaires vérifient l'équation :

$$\hat{N}_s = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} = N_s$$

On surestime alors l'effectif total des unités secondaires lorsqu'on l'estime via les unités primaires. En effet, si après calage les poids sont positifs, on a :

$N_s = \hat{N}_s = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} \leq \sum_{j \in s_I^r} \sum_{i \in s_j} w_{i,j} = \sum_{j \in s_I^r} w_j \sum_{i \in s_j} \omega_{i/j} = \sum_{j \in s_I^r} w_j \, N_j = \hat{N}_p$. L'égalité n'est obtenue qu'en cas d'absence de non-réponse ($s_j^r = s_j \ \ \forall j$) .

Dans ce cas, seule la présence de non-réponse au niveau des unités secondaires a un impact.

La conclusion est alors la même que dans le cas où les poids de tirage des unités secondaires ont déjà été corrigés de la non-réponse : l'introduction de l'effectif total des salariés doit se faire non seulement dans le calage des unités secondaires, mais également dans le calage des unités primaires.

Comme le montre l'inégalité ci-dessus, il ne peut y avoir de solutions au système d'équations de calage que si l'on utilise la méthode de calage "linéaire", qui autorise les poids négatifs. Les autres méthodes ne pourront converger.

2.2.2. Cas où les effectifs de regroupements d'UP ne sont pas utilisés dans le calage des poids des unités secondaires

Dans le cas où la non-réponse des unités primaires et/ou unités secondaires a été corrigée séparément du calage, on a $\sum_{i \in s_j^r} \omega'_{i/j} \neq N_j$, rien ne permettant d'assurer l'égalité après correction des poids d'échantillonnage.

Si la non-réponse a été corrigée simultanément au calage, on a $\sum_{i \in s_j^r} \omega_{i/j} \leq N_j$, l'égalité n'étant assurée que s'il n'y a pas de non-réponse dans l'unité primaire j.

La condition suffisante à la cohérence des estimations d'effectifs de regroupements d'unités primaires n'est plus respectée en présence de non-réponse au niveau des unités secondaires, quel que soit le mode de correction de la non-réponse. La non-réponse des unités primaires n'a en revanche aucun impact, et ce quel que soit le mode de correction.

2.2.3. Cas particulier du sondage en grappes

Les conclusions du cas du sondage en deux degrés s'adaptent au cas particulier du sondage en grappes.

Examinons si les deux autres propriétés **des sondages en grappe (cf. 1.3) continuent d'être respectées en cas d'introduction de la non-réponse.**

- **Établir des statistiques sur les UP à partir du fichier d'US (prop 1)**

L'égalité des poids des US au sein d'une même UP **permet d'établir des statistiques sur les UP à partir du fichier d'US** : le poids d'une UP est le poids d'une US quelconque appartenant à l'UP. De plus, les statistiques sur les UP établies à partir du fichier d'US sont alors bien cohérentes avec les statistiques établies à partir du fichier d'UP. Cette condition n'est vérifiée que dans le cas où la non-réponse est corrigée en même temps que le calage.

- **Récupérer des estimations sur des effectifs d'US identiques à partir des échantillons d'UP et d'US (prop 2)**

Reprenons l'exemple de l'estimation du nombre de femmes de professions intermédiaires dans la population, via l'échantillon des « unités primaires » $\hat{Y}_p$ (ménages) et via celui des unités secondaires $\hat{Y}_s$ (individus) (on utilise les mêmes notations qu'au § 1.4.1.).

Si la correction du biais de non-réponse est effectuée en amont du calage :

$$\hat{Y}_s = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j}\, y_{i,j} = \sum_{j \in s_I^r} w_j \sum_{i \in s_j^r} \omega'_{i/j}\, y_{i,j} = \sum_{j \in s_I^r} w_j\, \hat{y}_j \neq \sum_{j \in s_I^r} w_j\, y_j = \hat{Y}_p$$

avec $\hat{y}_j$ estimateur du nombre de femmes de professions intermédiaires par ménage.

Si la correction du biais de non-réponse est effectuée en même temps que le calage :

$$\hat{Y}_s = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j}\, y_{i,j} = \sum_{j \in s_I^r} w_j \sum_{i \in s_j^r} y_{i,j} \leq \sum_{j \in s_I} w_j\, y_j = \hat{Y}_p \quad \text{avec } y_j = \sum_{i \in s_j} y_{i,j} \geq \sum_{i \in s_j^r} y_{i,j}$$

De même, ces estimations sont différentes.

2.3. Conclusion

En cas de présence de non-réponse au niveau des unités secondaires (quel que soit le mode de correction de la non-réponse effectué), le calage simultané effectué entre les volets d'une enquête couplée et conservant les poids de tirage conditionnels perd ses propriétés.

Dans la partie suivante, nous verrons d'autres calages simultanés avec d'autres contraintes sur les poids conditionnels qui permettent de retrouver ces propriétés de cohérence même en présence de non-réponse.

3. Présentation et propriétés de deux méthodes de calages simultanés

Plusieurs méthodes de calage simultané sont proposées par Estevao et Särndal [2].

La méthode que nous noterons (A), dénommée dans [2] **Single step calibration with integration – 1**, est la méthode « via unités primaires » utilisée à l'Insee ; on l'appellera "calage simultané en une étape au niveau des UP" (cf. § 1.2.).

La deuxième méthode, que nous noterons (B), dénommée dans [2] **Single step calibration with integration – 2,** est une méthode « via unités secondaires » **:**

- les variables de calage des unités primaires sont « descendues » au niveau des unités secondaires,
- le calage est alors réalisé au niveau des unités secondaires,
- les poids des unités primaires sont calculés à partir des poids des unités secondaires les constituant.

La troisième méthode (C), dénommée dans [2] **Two step calibration with integration**, est également une méthode « via unités secondaires », mais réalisée en deux temps **:**

- dans un premier temps les unités primaires sont calées,
- les poids de chacune des unités primaires sont « descendus » au niveau des unités secondaires,
- un nouveau calage est alors réalisé au niveau des unités secondaires, en incluant comme contraintes de calage supplémentaires des relations entre les poids des UP et les poids des US (voir § 3.2.).

3.1. Méthode (B) : Calage simultané en une étape au niveau des US

Cette méthode assure que la « taille » (en unités secondaires) estimée de toute réunion d'unités primaires (i.e., l'estimation du nombre d'US appartenant à une quelconque sous-population d'UP) est la même, qu'on l'estime à partir de l'échantillon d'UP ou de l'échantillon d'US.

3.1.1. Principe général

On suppose qu'il n'y a pas de non-réponse, au niveau des UP comme au niveau des US.

On réalise le calage au niveau des unités secondaires en utilisant :

- les variables de calage des unités secondaires et leurs totaux X_s sur la population entière,
- les variables de calage des unités primaires et leurs totaux X_p. Elles sont intégrées au niveau des unités secondaires, en divisant les variables par l'effectif N_j de l'unité primaire.

Les poids calés au niveau unités secondaires $w_{i,j}$ vérifient les deux systèmes d'équations de calage :

$$(2) : \forall s \in \{1...S\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j}\ x_{ij,s} = X_s$$

$$(1') : \forall p \in \{1...P\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \frac{x_{j,p}}{N_j} = X_p$$

On obtient ainsi les poids des unités secondaires $w_{i,j}$. **Les poids des unités primaires** sont ensuite calculés tels que :

$$(3') : w_j = \frac{\sum_{i \in s_j} w_{i,j}}{N_j}$$

Dans le cas d'un sondage en grappes cela revient à calculer les poids des unités primaires comme la moyenne des poids des unités secondaires après calage.

3.1.2. Vérification du calage des unités primaires et secondaires

Vérifions que les équations (1') assurent que les poids des unités primaires sont calés sur les totaux des variables auxiliaires de calage des unités primaires, soit les équations (1).

$$(1') : \forall p \in \{1...P\} : X_p = \sum_{j \in s_I} \sum_{i \in\ s_j} w_{i,j} \frac{x_{j,p}}{N_j} = \sum_{j \in s_I} x_{j,p} \sum_{i \in s_j} \frac{w_{i,j}}{N_j} = \sum_{j \in s_I} w_j\ x_{j,p} \Rightarrow (1)$$

Les poids des unités primaires sont bien calés sur les totaux des variables auxiliaires de calage.

Cette méthode est moins contraignante sur les pondérations que la méthode (A) développée dans Calmar2. En effet, en l'absence de non-réponse des unités secondaires, la contrainte sur les pondérations pour la méthode (A) est :

(A) : $w_{i,j} = \omega_{i/j} * w_j$

et pour la méthode (B) :

(B) : $\sum_{i \in s_j} w_{i,j} = w_j\ N_j$

Donc, si (A) est vérifiée : $\sum_{i \in s_j} w_{i,j} = w_j \sum_{i \in s_j} \omega_{i/j} = w_j\, N_j$ pour un tirage au second degré équilibré sur les tailles des US, i.e. (B) est vérifiée.

3.1.3. Cohérence assurée entre les volets en absence de non-réponse

Dans le cas d'un sondage en grappe, les propriétés (1) et (2) présentées au § 1.4.1. ne sont pas vérifiées (en particulier les poids des US d'une même UP n'ont pas de raison d'être égaux).

En revanche, la propriété (3) (voir § 1.4.2.), dans le cas d'un sondage en grappes ou à deux degrés, est vérifiée grâce au principe même de la méthode, qui assure que $\sum_{i \in s_j} w_{i,j} = w_j \; N_j$

En effet, en reprenant les notations du § 1.4.2, les estimations de l'effectif d'un regroupement quelconque d'unités primaires G via chacun des deux volets s'écrivent :

$$\hat{N}_{G,s} = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \; \mathbb{1}_{j \in G} = \sum_{j \in s_I} \mathbb{1}_{j \in G} \sum_{i \in s_j} w_{i,j} = \sum_{j \in s_I} \mathbb{1}_{j \in G} \; w_j \; N_j = \; \hat{N}_{G,p}$$

3.1.4. Introduction de la non-réponse

Vérifions que la propriété (3) est conservée, et ce même en présence de non-réponse au niveau des unités primaires comme des unités secondaires.

- Via l'échantillon des « unités secondaires » : $\hat{N}_{G,s} = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} \; \mathbb{1}_{j \in G}$

- Via l'échantillon des « unités primaires » : $\hat{N}_{G,p} = \sum_{j \in s_I^r} N_j \; w_j \; \mathbb{1}_{j \in G} = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} \; \mathbb{1}_{j \in G} = \hat{N}_{G,s}$

La propriété (3) est conservée, le mode de correction de la non-réponse, en amont du calage ou simultanément à celui-ci, n'ayant aucun impact.

3.2. Méthode (C) : Calage simultané en deux étapes

Variante de la méthode (B), cette méthode assure la même propriété, mais avec un calage en deux étapes.

3.2.1. Principe général

On suppose qu'il n'y a pas de non-réponse, au niveau des UP comme au niveau des US.

Le principe est le suivant :

- on réalise le calage au niveau des unités primaires en utilisant les seules variables de calage des unités primaires et leurs totaux X_p ,

- on obtient ainsi les poids des unités primaires w_j ,

- on réalise ensuite un calage au niveau des unités secondaires en ajoutant aux variables de calage des unités secondaires les indicatrices d'appartenance d'une unité secondaire à une unité primaire, soit autant de variables de calage que d'unités primaires, avec les contraintes de calage (2") précisées ci-dessous.

Les poids calés w_j au niveau des unités primaires vérifient les équations :

$(1) : \forall p \in \{1...P\} : X_p = \sum_{j \in s_I} w_j \; x_{j,p}$: on obtient ainsi les poids des unités primaires w_j .

Les poids calés $w_{i,j}$ au niveau unités secondaires vérifient les deux systèmes d'équations de calage :

$$(2) : \forall s \in \{1...S\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \; x_{i,s} = X_s$$

$$(2'') : \forall j \in s_I : \sum_{i \in s_j} w_{i,j} = N_j * w_j$$

On obtient ainsi les poids des unités secondaires $w_{i,j}$. Les poids des unités primaires et des unités secondaires sont bien calés sur les totaux des variables auxiliaires de calage respectivement des unités primaires et des unités secondaires.

3.2.2. Cohérence assurée entre les volets en absence de non-réponse

Dans le cas d'un sondage en grappes, les propriétés (1) et (2) présentées au § 1.4.1. ne sont pas vérifiées (en particulier les poids des US d'une même UP n'ont pas de raison d'être égaux).

En revanche, la propriété (3) (voir § 1.4.2.), dans le cas d'un sondage en grappes ou à deux degrés, est vérifiée grâce aux équations de calage (2'') (même démonstration qu'au § 3.1.3.).

3.2.3. Introduction de la non-réponse

Vérifions que la propriété (3) est conservée, et ce même en présence de non-réponse au niveau des unités primaires comme des unités secondaires.

- Via l'échantillon des « unités secondaires » : $\hat{N}_{G,s} = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} \, \mathbb{1}_{j \in G}$

- Via l'échantillon des « unités primaires » : $\hat{N}_{G,p} = \sum_{j \in s_I^r} N_j \, w_j \, \mathbb{1}_{j \in G} = \sum_{j \in s_I^r} \sum_{i \in s_j^r} w_{i,j} \, \mathbb{1}_{j \in G} = \hat{N}_{G,s}$

La propriété (3) est conservée, le mode de correction de la non-réponse, en amont du calage ou simultanément à celui-ci, n'ayant encore aucun impact.

3.3. Tableau comparatif des contraintes portant sur les poids

	Contraintes sur les unités primaires	Contraintes sur les unités secondaires
Calages indépendants	$\forall p \in \{1...P\} : X_p = \sum_{j \in s_I} w_j \, x_{j,p}$	$\forall s \in \{1...S\} : X_s = \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, x_{ij,s}$
Méthode (A)	$\forall p \in \{1...P\} : X_p = \sum_{j \in s_I} w_j \, x_{j,p}$ $\forall s \in \{1...S\} : X_s = \sum_{j \in s_I} w_j \sum_{i \in s_j} \omega_{i/j} \, x_{ij,s}$	$\forall j \in s_I \ \forall i \in s_j : w_{i,j} = \omega_{i/j} * w_j$
Méthode (B)	$\forall j \in s_I : w_j = \frac{\sum_{i \in s_j} w_{i,j}}{N_j}$	$\forall s \in \{1...S\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, x_{ij,s} = X_s$ $\forall p \in \{1...P\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, \frac{x_{j,p}}{N_j} = X_p$
Méthode (C)	$\forall p \in \{1...P\} : X_p = \sum_{j \in s_I} w_j \, x_{j,p}$	$\forall s \in \{1...S\} : \sum_{j \in s_I} \sum_{i \in s_j} w_{i,j} \, x_{ij,s} = X_s$ $\forall j \in s_I : \sum_{i \in s_j} w_{i,j} = N_j * w_j$

Les distributions des poids des unités primaires après calages indépendants ou après calage simultané par la méthode (C) vérifient les mêmes contraintes, les poids après calage auront dont la même distribution (si la même méthode de calage est choisie, logit, linéaire, etc.).

Les poids des unités secondaires après calages indépendants seront moins dispersés que pour les calages simultanés par la méthode (A) et la méthode (C), les poids respectant les mêmes contraintes que pour un calage indépendant additionnées d'autres équations.

Bibliographie

[1] S. Hallépée, « Stratégies d'échantillonnage pour les enquêtes couplées », JMS, 2009

[2] Victor M. Estevao, Carl-Erik Särndal, « Survey Estimates by Calibration on Complex Auxiliary Information », *International Statistical Review*, volume 74, n°2 (2006), 127-147.

[3] N. Caron et O. Sautory, « Calages simultanés pour différentes unités d'une même enquête », Document de travail interne Insee (1998)

[4] O. Sautory, « Calmar 2 : une nouvelle version du programme Calmar de redressement d'échantillon par calage », Recueil du Symposium de Statistique Canada, 2003

[5] www.enquetecoi.net

[6] F. Dupont, « Calage et redressement de la non-réponse totale », JMS, 1993