LES ENJEUX MÉTHODOLOGIQUES LIÉS À L'USAGE DE BASES DE SONDAGE IMPARFAITES

Olivier SAUTORY[1] (*)

(*) Insee, Département des méthodes statistiques

Résumé

Il est rare que les bases de sondage utilisées par les statisticiens des instituts nationaux possèdent toutes les "bonnes propriétés" requises, telles qu'elle sont présentées au début des ouvrages de sondages. Les imperfections des bases utilisées sont de différents types : différences avec la population-cible de l'enquête (défaut ou excès de couverture), présence de doubles comptes, mauvaise qualité des identifiants, manque d'actualisation des informations présentes dans la base, etc. Pour permettre d'obtenir des estimations fiables à partir des échantillons sélectionnés dans ces bases, ces défauts doivent être si possible corrigés : amélioration de la qualité des bases, mise en œuvre de techniques adaptées au moment de l'échantillonnage, de la collecte, de la phase d'estimation. Certaines enquêtes utilisent simultanément plusieurs bases de sondage, par exemple pour pallier le défaut de couverture. Ce papier présente les différentes questions posées par ces bases imparfaites, et les principales réponses pratiques ou méthodologiques qui y sont apportées.

Abstract

The frames used by statisticians in national agencies seldom have all of the required "good properties", as commonly described at the beginning of survey manuals. The imperfections in the frames used are of various types: differences in the survey's target population (undercoverage or overcoverage), duplication, poor-quality identifiers, out-of-date information, etc. These shortcomings must be rectified, if possible, to produce reliable estimates with the samples selected from these frames, by improving the quality of the frames and employing special techniques in sampling, collection and estimation. Some surveys use multiple frames simultaneously, for example, to compensate for undercoverage. This paper describes the various issues that these imperfect frames raise, and the principal practical or methodological responses to them.

Mots-clés

Base de sondage, sous-couverture, sur-couverture, doublons, bases multiples

Introduction

La base de sondage, telle qu'elle est définie au début des manuels de sondages, doit vérifier un certain nombre de bonnes propriétés, qui sont malheureusement rarement satisfaites dans la pratique, ce qui cause bien des difficultés dans la "vraie vie" des méthodologues d'enquête. Les imperfections des bases utilisées sont de différents types : différences avec la population-cible de l'enquête (défaut ou excès de couverture), présence de doublons, mauvaise qualité des identifiants, défaut de mises à jour, voire des erreurs, sur les informations présentes dans la base, etc.

Malgré la fréquence de ces problèmes rencontrés par les statisticiens responsables d'enquêtes, Särndal et Lundström (2005) soulignent dans le chapitre 14 de leur ouvrage sur la

[1] olivier.sautory@insee.fr

non-réponse qu'il n'existe pas de méthodologie fermement établie dans ce domaine, et que les pratiques des instituts varient beaucoup, en utilisant souvent des méthodes *ad hoc*.

Ce papier se propose de faire un survol sur les différents types d'imperfections des bases de sondage, en indiquant les conséquences de ces imperfections et les remèdes que l'on peut y apporter.

1. Définition et contenu d'une base de sondage

1.1. Définition d'une base de sondage

Lorsque l'on veut réaliser une enquête par sondage, il faut commencer par définir la population-cible, ou champ de l'enquête : c'est la population sur laquelle on cherche à obtenir de l'information et à estimer des paramètres d'intérêt. La base de sondage est une liste d'unités permettant d'identifier les éléments de la population-cible, à partir de laquelle on va pouvoir sélectionner un échantillon selon une méthode probabiliste. Dans l'idéal, population-cible et base de sondage coïncident ; mais dans la pratique c'est très rarement le cas…

Parfois, la base de sondage est obtenue à partir d'une liste d'unités correspondant à une population plus large que la population-cible : on sélectionne dans cette liste les unités présentant certaines valeurs pour des variables figurant dans la liste.

1.2. Contenu d'une base de sondage

Dans une base de sondage idéale, on trouve pour chaque unité :

- un identifiant, qui identifie sans ambiguïté l'unité (par exemple nom, prénom et adresse d'un individu, la dénomination de l'entreprise), ou mieux encore, un numéro d'identification unique, par exemple un numéro d'enregistrement dans un répertoire administratif ;
- des données permettant de repérer l'unité lors de la collecte, ou de la contacter : une adresse précise (adresse postale + étage …), un n° de téléphone, etc. ;
- des informations qui vont pouvoir être utilisées soit au moment de la définition du plan de sondage, soit lors de la phase d'estimation, par exemple pour corriger la non-réponse ou pour caler l'échantillon. Cette information auxiliaire va permettre d'améliorer la précision des estimations.

2. Imperfections d'une base de sondage

2.1. Sous-couverture

2.1.1. Définition, causes de la sous-couverture

On dit qu'il y a sous-couverture lorsque des unités de la population-cible sont absentes de la base de sondage.

Parmi les causes possibles de la sous-couverture, on peut mentionner le délai entre le moment où on a sélectionné l'échantillon dans la base, qui elle-même peut avoir une certaine ancienneté, et le moment où on réalise l'enquête : dans l'intervalle, des unités ont pu entrer dans le champ de l'enquête (par exemple des entreprises qui se sont créées pendant cette période).

On peut aussi avoir volontairement enlevé des unités de la base de sondage parce qu'elles ont déjà été enquêtées, et que l'on ne souhaite pas réinterroger trop tôt. Mais on sait que ceci n'est pas une façon très correcte de traiter la question de la coordination négative d'échantillons, en particulier dans le domaine des enquêtes auprès des entreprises.

Plus simplement, il peut arriver que la seule base de sondage dont on dispose ne couvre pas toute la population-cible.

2.1.2. Si on ignore la sous-couverture[2]

La population-cible U (de taille N), est plus grande que la population de la base de sondage U_F (de taille N_F), elle contient une population U_O (de taille $N_O = N - N_F$) qu'on appelle population omise.

Pour une variable d'intérêt ***Y***, son total dans U_F, noté Y_F, est donc inférieur à son total dans U, noté Y, que l'on cherche à estimer : $Y = Y_F + Y_O$. On note $r = \overline{Y}_O / \overline{Y}_F$ le rapport entre la moyenne de ***Y*** dans U_O et sa moyenne dans U_F, $\tau = N_O / N$ la proportion de la population U non couverte par la base de sondage.

Si on dispose d'un estimateur sans biais du total Y_F, c'est bien entendu un estimateur biaisé du total Y ; son biais relatif $\frac{Y_F - Y}{Y} = \frac{-r\tau}{r\tau + (1-\tau)}$ est toujours négatif, il est faible dès que r ou τ est petit.

Si on dispose d'un estimateur sans biais de la moyenne $\overline{Y}_F$, c'est en général un estimateur biaisé de la moyenne $\overline{Y}$ sur U ; son biais relatif $\frac{\overline{Y}_F - \overline{Y}}{\overline{Y}} = \frac{\tau(1-r)}{r\tau + (1-\tau)}$ est nul si r = 1, négligeable si r ≈1 ou τ petit.

2.1.3. Évaluation de la sous-couverture

On peut évaluer la sous-couverture lorsque l'on dispose de sources externes indépendantes et fiables, que l'on peut utiliser de deux façons.

À un niveau agrégé, on peut calculer des taux de couverture, sur la population ou sur des sous-populations. Un exemple classique est la comparaison de la structure par sexe et âge dans la base de sondage avec celle provenant d'un recensement de la population. Mais le taux de couverture ainsi calculé peut résulter à la fois de sous-couverture sur une catégorie de la population et de sur-couverture sur une autre catégorie, sans que l'on puisse séparer les deux effets.

On peut aussi réaliser un appariement entre la base de sondage U_F et la source externe U_{ext}, au niveau individuel (voir Wolter, 1983).

Tableau 1 : appariement entre U_F et U_{ext}

	présent U_{ext}	absent U_{ext}	
présent U_F	N_{11}	N_{12}	N_{1+}
absent U_F	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	N_{++}

N_{11} désigne le nombre d'éléments que l'on a réussi à apparier, c'est-à-dire présents dans les deux bases, N_{21} désigne le nombre d'éléments de la source externe absents de la base de sondage, etc., et N_{++} désigne la taille totale de la population, qui est inconnue.

Le principe de la méthode est d'estimer le taux de couverture de la base de sondage $N_{1+}/N{+}{+}$ par le rapport N_{11}/N_{+1}, que l'on peut interpréter comme le "taux de couverture" observé dans la source externe.

On peut d'ailleurs en déduire une estimation de la quantité inconnue N_{22}, et donc de la taille totale N_{++} de la population : $N_{11} + N_{12} + N_{21} + (N_{12} \times N_{21}) / N_{11}$.

Cette méthode est aussi connue sous le nom de méthode de capture-recapture, utilisée en particulier pour estimer des tailles de populations d'animaux. Elle peut être aussi utilisée en remplaçant la source externe par un échantillon s_{ext} tiré indépendamment de la base de sondage dans la population totale, par exemple en réalisant un sondage aréolaire. Les formules sont identiques, sauf que les effectifs du tableau sont remplacés par des effectifs estimés à partir de s_{ext}.

[2] Un certain nombre de résultats de ce paragraphe 3 sont empruntés à l'ouvrage de Lessler et Kalsbeek (1992)

2.1.4. Que faire en cas de sous-couverture ?

Une méthode simple consiste à redéfinir la population-cible, i.e. le champ de l'enquête, en disant que ce champ est celui couvert par la base de sondage ! C'est facile, pas coûteux, et finalement on le fait assez souvent…

Sinon, une méthode, appelée "linking procedure", consiste à définir une règle d'association entre tous les éléments omis et des éléments de la base de sondage, et, au moment de l'enquête, à interroger les éléments omis qui sont reliés à des éléments de l'échantillon qui a été sélectionné. Un exemple d'une telle procédure, proposée par Kish (1965), est appelé "half-open interval procedure" : dans le cas d'une enquête-ménages réalisée en face-à-face, on demande aux enquêteurs d'interroger les ménages résidant dans des logements qui ne sont pas dans la base de sondage, et qui sont situés dans l'intervalle séparant deux logements sélectionnés dans l'échantillon.

Les bases de sondage multiples (voir §5) peuvent permettent de résoudre les problèmes de sous-couverture.

On peut aussi mentionner les plans d'échantillonnage adaptatifs, utilisés lorsque la population-cible est constituée d'éléments rares, mal identifiés, ou même absents des bases de sondage dont on dispose (Thompson 1990, 2011).

2.2. Sur-couverture

2.2.1. Définition, causes de la sur-couverture

On dit qu'il y a sur-couverture lorsque des unités de la base de sondage ne font pas partie de la population-cible.

Les causes possibles de la sur-couverture sont similaires à celles qui peuvent expliquer la sur-couverture. Ainsi, en raison du délai entre la date de sélection de l'échantillon dans la base de sondage, qui peut elle-même avoir une certaine "ancienneté", et la date de collecte, des unités de la base ont pu "disparaître" du champ de l'enquête (par exemple, des entreprises qui ont cessé leur activité entre la date de dernière mise à jour de la base de sondage et la date de l'enquête).

Il peut arriver aussi que la seule base de sondage disponible contienne des éléments n'appartenant pas à la population-cible

2.2.2. Si on ignore la sur-couverture

La base de sondage U_F (de taille N_F), est plus grande que la population-cible U (de taille N), elle contient une population U_{HC} (de taille $N_{HC} = N_F - N$) qu'on appelle population hors champ.

Pour une variable d'intérêt ***Y***, son total dans U_F, noté Y_F, est donc supérieur à son total dans U, noté Y, que l'on cherche à estimer : $Y_F = Y + Y_{HC}$. On note $r = \overline{Y}_{HC} / \overline{Y}$ le rapport entre la moyenne de ***Y*** dans U_{HC} et sa moyenne dans U, $\tau = N_{HC} / N_F$ la proportion de hors champ dans la base de sondage.

Si on dispose d'un estimateur sans biais du total Y_F, c'est bien entendu un estimateur biaisé du total Y ; son biais relatif $\frac{Y_F - Y}{Y} = \frac{r\tau}{1-\tau}$ est toujours positif, il est faible dès que r ou τ est petit.

Si on dispose d'un estimateur sans biais de la moyenne $\overline{Y}_F$, c'est en général un estimateur biaisé de la moyenne $\overline{Y}$ sur U ; son biais relatif $\frac{\overline{Y}_F - \overline{Y}}{\overline{Y}} = \tau(r-1)$ est nul si r = 1, négligeable si $r \approx 1$ ou τ petit.

2.2.3. Évaluation de la sur-couverture

S'il existe des sources externes indépendantes et fiables, on peut, comme dans le cas de la sur-couverture, les utiliser à un niveau agrégé, en calculant des taux de couverture sur la population ou des sous-populations (exemple : comparaison de structures par sexe et âge).

On peut aussi réaliser une enquête spécifique sur le terrain, par exemple une enquête aréolaire, pour évaluer l'ampleur de la sur-couverture, du type des enquêtes post-censitaires ; ce type d'enquête permet d'ailleurs également de mesurer la sous-couverture. Pour améliorer la qualité d'un répertoire d'entreprises, on peut interroger un échantillon d'entreprises bien ciblées pour vérifier qu'elles sont toujours en activité.

2.2.4. Que faire en cas de sur-couverture ?

La population-cible U est une sous-population de la base de sondage U_F, ce que l'on appelle un domaine en théorie des sondages. Si on parvient à identifier parmi les unités de l'échantillon celles qui sont dans U et celles qui sont dans U_{HC}, on peut utiliser les résultats classiques de l'estimation d'un total, d'une moyenne, etc. sur un domaine.

Mais, il y a aussi en général des unités de l'échantillon qui sont non-répondantes. Il peut alors être difficile de savoir si les unités sont bien de "vraies" non-répondantes, ou bien si elles n'ont pas répondu parce qu'elles sont devenues hors champ : c'est le cas, par exemple, des entreprises qui ont cessé leur activité.

Pour bien identifier unités "vraies" non-répondantes et unités hors champ, on peut apparier l'échantillon avec la version la plus récente de la base de sondage, pour détecter d'éventuelles unités hors champ. On peut aussi mobiliser des sources externes, par exemple regarder dans des sources fiscales si l'entreprise non-répondante semble ou non être encore en activité.

2.3. Doublons

2.3.1. Définition, causes des doublons

On dit qu'il y a des doublons, ou des répétitions, lorsque des unités apparaissent plusieurs fois dans la base.

Parmi les causes possibles de l'existence de doublons, on peut citer le cas où on utilise plusieurs listes pour constituer la base de sondage, avec des unités qui appartiennent à différentes listes, ou bien le cas où plusieurs identifiants dans la base de sondage désignent en réalité la même unité (par exemple le cas où une entreprise serait identifiée une fois par sa raison sociale et une autre fois par son nom commercial).

Lorsque des doublons sont détectés sans ambiguïté dans la base de sondage, on enlève en général les unités qui sont en trop.

2.3.2. Si on ignore les doublons

On va supposer que la base de sondage $U_F = \{1...i...N_F\}$ contient toutes les unités de la population-cible $U = \{1...k...N\}$ (pas de sous-couverture), et qu'elle ne contient pas d'autres unités (pas de sur-couverture).

Pour toute unité k de U, on note L_k le nombre d'unités de U_F associées à k, toujours supérieur ou égal à 1, et parfois strictement supérieur à 1 (doublons).

On note N_a le nombre d'unités de U qui sont présentes *a* fois dans la base U_F, *a* variant de 1 à A.

Si on dispose d'un estimateur sans biais du total Y_F, c'est bien entendu un estimateur biaisé du total Y ; son biais relatif $\frac{Y_F - Y}{Y} = \frac{\sum_{k \in U} (L_k - 1) y_k}{\sum_{k \in U} y_k}$ est toujours positif, il dépend du nombre de doublons.

Si on dispose d'un estimateur sans biais de la moyenne $\overline{Y}_F$, c'est en général un estimateur biaisé de la moyenne $\overline{Y}$ sur U ; son biais relatif $\frac{\overline{Y}_F - \overline{Y}}{\overline{Y}} = \frac{\sum_{a=1}^{A} \left(\frac{a N_a}{N_F} - \frac{N_a}{N} \right) \overline{Y}_a}{\overline{Y}}$, avec $\overline{Y}_a = \frac{\sum_{k:L_k=a} y_k}{N_a}$, est nul si les moyennes $\overline{Y}_a$ sur les unités présentes une fois dans la base, 2 fois dans la base, 3 fois dans la base, etc., sont toutes égales.

2.3.3. Que faire en cas de doublons ?

On suppose que l'on a sélectionné un échantillon s_F dans la base de sondage, avec des probabilités d'inclusion π_i. L'estimateur d'Horvitz-Thompson $\hat{Y}_F = \sum_{i \in s_F} y_i / \pi_i$ est donc un estimateur biaisé du total Y dans la population-cible.

Si la collecte ne permet pas de connaître la "multiplicité" de chaque unités enquêtée, c'est-à-dire le nombre de fois où l'unité est présente dans la base de sondage, il existe des méthodes permettant d'obtenir des estimateurs sans biais, fondés uniquement sur la présence de doublons dans l'échantillon (voir Kish (1965), dans le cas où il y des unités répliquées deux fois, mais pas plus).

Si la collecte permet de connaître la multiplicité des unités enquêtées, il existe plusieurs méthodes permettant de construire des estimateurs sans biais prenant en compte la multiplicité.
On suppose que pour chaque unité i de l'échantillon s_F, associée à l'unité k de U, on connaît la multiplicité L_k. Si on choisit des poids ω_{ik} qui se somment à 1 pour toutes les unités i de la base associées à la même unité k, alors $\sum_{i \in s_F} \omega_{ik} y_i / \pi_i$ est un estimateur sans biais de Y. On peut choisir par exemple $\omega_{ik} = 1/L_k$, ce qui revient à diviser le poids de sondage de l'unité i par la valeur de sa multiplicité : cette technique est basée sur le principe du partage des poids (voir Lavallée, 2002), même si cette méthode du partage des poids a été introduite à l'origine dans le contexte des sondages indirects.

3. Utilisation d'information auxiliaire pour corriger les erreurs de couverture

Au moment de la phase d'estimation, on peut faire appel aux méthodes classiques d'utilisation d'une information auxiliaire pour corriger les erreurs de couverture. On en présente trois exemples.

3.1. Ajustement par le ratio (sous-couverture)

On se place dans le cas de sous-couverture. On note s_F un échantillon tiré dans la base de sondage U_F, avec des probabilités d'inclusion π_i.

Pour une variable d'intérêt ***Y***, l'estimateur d'Horvitz-Thompson $\hat{Y}_F = \sum_{s_F} \frac{y_i}{\pi_i}$ est un estimateur sans biais du total Y_F sur U_F, mais un estimateur biaisé négativement du total Y sur la population-cible U.

On suppose que l'on dispose d'une variable ***X*** sur l'échantillon dont le total X_U sur U est connu. On peut alors estimer Y par un ajustement de type ratio : $\hat{Y}_r = X_U \left(\sum_{s_F} y_i / \pi_i \right) / \left(\sum_{s_F} x_i / \pi_i \right)$. Le biais relatif de cet estimateur vaut $\frac{E(\hat{Y}_r) - Y}{Y} \approx \frac{R_{U_F}}{R_U} - 1$, avec $R_{U_F} = \frac{Y_{U_F}}{X_{U_F}}$ et $R_U = \frac{Y_U}{X_U}$. Il est négligeable si les rapports (total de Y / total de X) dans U et dans U_F sont proches, hypothèse qui peut être assez forte.

3.2. Ajustement par post-stratification

Särndal & Lundstrom (2005) traitent le cas où il y a à la fois non-réponse et erreur de couverture, par exemple avec un ajustement de type post-stratification dans le cas d'un sondage aléatoire simple stratifié.

On suppose que l'on connaît les effectifs N_h des strates dans la population-cible U. On note r_{Fh} l'échantillon des répondants de s_F dans le champ et dans la strate h, m_{Fh} la taille de cet échantillon. On peut alors estimer Y par un ajustement de type post-stratification : $\hat{Y}_{ps} = \sum_{h=1}^{H} \frac{N_h}{m_{F_h}} \sum_{r_{F_h}} y_i$.

Cet estimateur est sans biais sous l'hypothèse (forte) que les éléments de la strate h dans U ont la même "probabilité" d'appartenir à la base de sondage U_F, et ont la même probabilité de réponse.

3.3. Ajustement par calage (classique ou généralisé)

Plus généralement, on peut utiliser des méthodes de calage, dès que l'on dispose d'un vecteur X_U de totaux de variables connus sur la population-cible U, et que ces variables sont connues sur l'échantillon. Si r_F désigne l'échantillon des répondants dans le champ, l'estimateur par calage[3] du total d'une variable ***Y*** vaut :

$\hat{Y}_{cal} = \sum_{r_F} w_i y_i = \sum_{r_F} g_i d_i y_i$ avec $d_i = 1/\pi_i$, avec $g_i = 1 + \left(X_U - \sum_{r_F} d_i x_i \right)' \left(\sum_{r_F} d_i x_i x_i' \right)^{-1} x_i$ (calage classique), ou $g_i = 1 + \left(X_U - \sum_{r_F} d_i x_i \right)' \left(\sum_{r_F} d_i z_i x_i' \right)^{-1} x_i$ (calage généralisé, voir Deville, 1998 ; Estevao & Särndal, 2000 ; Sautory, 2003). Parmi les variables intervenant dans le vecteur z_i, il est intéressant de disposer de variables "explicatives" de la non-réponse mais aussi du défaut de couverture, sans qu'il soit nécessaire de connaître leurs totaux sur la population.

4. Bases de sondage multiples

4.1. Pourquoi utiliser des bases de sondage multiples ?

Ce mode d'échantillonnage, qui consiste à constituer un échantillon en tirant des sous-échantillons dans des bases de sondage distinctes, peut répondre à plusieurs objectifs :

- l'utilisation conjointe de plusieurs bases de sondage, dont aucune ne couvre la population-cible, peut permettre d'améliorer la couverture de cette population ;
- lorsque l'on étudie une population rare, l'utilisation d'une base de sondage, même incomplète, mais contenant une proportion d'individus appartenant à cette population beaucoup plus forte que dans la population générale, permet de réduire significativement les coûts de collecte ;

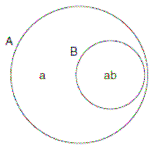
[3] avec la fonction de calage linéaire

– l'utilisation de modes de collecte différents selon la base de sondage peut permettre d'améliorer les taux de réponse.

Les développements suivants reposent beaucoup sur les articles de Sharon Lohr (2009, 2011).

4.2. Bases imbriquées

Figure 1: bases imbriquées



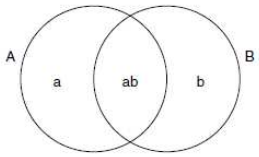
Les éléments de la base de sondage B constituent un sous-ensemble des éléments de la base de sondage A.
Un exemple classique est celui où A est une base "générale", avec peu d'informations (ou une base aréolaire), et où la base B est une liste d'unités plus adaptée à l'objectif de l'enquête, mais incomplète.

Par exemple, si l'on s'intéresse à la pratique du basket, la base B peut être constituée de tous les individus inscrits dans un club de basket. Mais si l'on veut aussi interroger les individus qui pratiquent le basket dans la rue, il est nécessaire de sélectionner un échantillon dans une base A qui couvre la population générale.

S'il est possible d'identifier, avant la collecte, quels sont les éléments de B qui appartiennent à A, on les enlève de A avant la sélection des échantillons: on est alors ramené au cas classique d'un sondage stratifié.
Sinon, on se trouve dans une situation similaire à celle des bases chevauchantes.

4.3. Bases chevauchantes

Figure 2 : bases chevauchantes



Les deux bases de sondage A et B se chevauchent, et, par ailleurs, leur réunion peut ne pas couvrir la totalité de la population-cible.
Par exemple, dans le cas d'une enquête téléphonique, A est une liste de numéros de téléphone fixe, B est une liste de numéros de téléphone mobile. La réunion de A et de B ne couvre pas les individus qui n'ont pas de téléphone.

Il existe 3 domaines dans $A \cup B$:

– le domaine *a* contient les unités qui sont dans A mais pas dans B ;

– le domaine *b* contient les unités qui sont dans B mais pas dans A ;

– le domaine *ab* contient les unités qui figurent dans les deux bases de sondage, par exemple les individus qui ont à la fois un téléphone fixe et un téléphone mobile.

Si l'appartenance aux domaines est connue à l'avance, on peut supprimer de la base B les unités de *ab* avant le tirage de l'échantillon, et on est ramené au cas d'un sondage stratifié. Mais cela est rarement possible.

Dans la suite, on supposera que l'on sait identifier, **au moment de la collecte**, les unités du domaine de *ab*. On pourrait alors, parmi ces unités, supprimer celles qui font partie de l'échantillon sélectionné à partir de la base B, mais en général on ne le fait pas, car on n'aime pas perdre des unités qui ont été enquêtées ! On suppose par la suite que l'on garde tous les éléments enquêtés.

Le total d'une variable d'intérêt ***Y*** est alors égal à la somme des totaux des variables sur chaque domaine :

$$Y = Y_a + Y_b + Y_{ab}$$

L'échantillon s_A tiré dans A permet d'obtenir des estimateurs, sans biais de préférence, de Y_a et de Y_{ab}, notés $\hat{Y}_a^A$ et $\hat{Y}_{ab}^A$. De même, un échantillon s_B tiré dans B permet d'obtenir des estimateurs de Y_b et de Y_{ab}, notés $\hat{Y}_b^B$ et $\hat{Y}_{ab}^B$.

Comment utiliser au mieux les deux estimateurs $\hat{Y}_{ab}^A$ et $\hat{Y}_{ab}^B$ pour estimer Y_{ab} ? De nombreuses méthodes ont été proposées dans la littérature.

Moyenne pondérée des estimateurs

Hartley (1962) propose de faire une moyenne pondérée des deux estimateurs, avec des pondérations θ et $1-\theta$: $\hat{Y}(\theta) = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B$

En termes de poids, ceci revient à ne pas modifier les poids des unités de *a* et de *b*, et à multiplier les poids des unités de *ab* par le coefficient θ si l'unité vient de la base A, et par le coefficient $1-\theta$ si elle vient de la base B.

Hartley propose de choisir comme valeur de θ celle qui minimise la variance de l'estimateur du total. On montre que plus $V\left(\hat{Y}_{ab}^B\right)/V\left(\hat{Y}_{ab}^A\right)$ est élevé, plus cette valeur optimale θ_{opt} est élevée.

Une variante a été proposée par Fuller et Burmeister (1972), qui utilisent deux estimateurs de la taille N_{ab} du domaine *ab* : $\hat{Y}_{FB}(\beta) = \hat{Y}_a^A + \beta_1 \hat{Y}_{ab}^A + (1-\beta_1)\hat{Y}_{ab}^B + \hat{Y}_b^B + \beta_2 (\hat{N}_{ab}^A - \hat{N}_{ab}^B)$

On peut choisir comme valeurs de β_1 et β_2 celles qui minimisent la variance de l'estimateur du total.

Mais pour cet estimateur comme celui de Hartley, si on utilise les pondérations optimales, on obtient pour les unités des poids aléatoires (i.e. qui dépendent des échantillons sélectionnés), et surtout qui dépendent de la variable d'intérêt : ceci n'est en général pas souhaitable, car cela entraîne une perte de la cohérence entre les variables d'intérêt (par exemple, si Z = X + Y, on n'aura pas $\hat{Z} = \hat{X} + \hat{Y}$).

Estimateur du pseudo-maximum de vraisemblance

Skinner et Rao (1996) ont proposé un estimateur du pseudo-maximum de vraisemblance, variante de l'estimateur de Fuller et Burmeister, utilisant des estimateurs des tailles N_a et N_b des domaines *a* et *b*, ainsi qu'un estimateur spécifique de N_{ab}. Les poids sont aléatoires, mais ne dépendent pas de la variable d'intérêt.

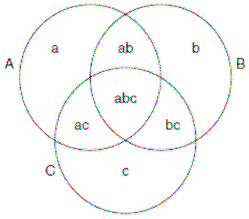
Lohr et Rao (2000, 2006) ont montré que cet estimateur est efficace sous de nombreux plans de sondage complexes.

Estimateurs à base de sondage unique

Dans cette approche, proposée par Bankier (1986) et Kalton et Anderson (1986), on fait comme si tous les éléments provenaient d'une seule base, et on ajuste les poids pour les éléments du domaine *ab* de façon à obtenir des estimateurs sans biais. Par exemple, si les poids utilisés sont les poids de sondage, inverses des probabilités d'inclusion, le poids d'un élément i du domaine *ab* est égal à l'inverse de la somme des probabilités d'inclusion, $1/(\pi_i^A + \pi_i^B)$, que cet élément provienne de la base A ou de la base B. Cela nécessite de connaître, pour chaque unité de *ab*, les probabilités d'inclusion dans s_A et dans s_B. Ces poids ne dépendent pas de la variable d'intérêt.

4.4. Cas de 3 bases de sondage et plus

Figure 3 : cas de 3 bases chevauchantes



La figure ci-contre donne un exemple de 3 bases de sondage chevauchantes, conduisant à 7 domaines. Plus généralement, on peut considérer le cas de Q bases de sondage A_1... A_q... A_Q, à partir desquelles on sélectionne Q échantillons s_1 ... s_q ... s_Q, ce qui conduit potentiellement à D (= 2Q - 1) domaines d.

Les estimateurs précédents et leurs propriétés s'étendent en général au cas d'un nombre quelconque de bases de sondage, mais ils peuvent prendre des formes un peu compliquées.

La forme générale d'un estimateur est $\hat{Y} = \sum_{q=1}^{Q} \sum_{i \in s_q} m_i^{(A_q,d)} w_i^{A_q} y_i$, où les m_i sont des coefficients qui modifient les poids initiaux w_i.

Un cas intéressant est celui où ces coefficients sont fixes pour toutes les unités d'un domaine donné et provenant d'une base donnée : $m_i^{(A_q,d)} = m^{(A_q,d)}$ avec $\sum_{q=1}^{Q} m^{(A_q,d)} = 1$ pour tout domaine d : on parle d'estimateur à poids fixes. Un cas particulier, appelé *estimateur fondé sur la multiplicité* (Lavallée 2002, Mecatti 2007), prend une forme très simple : on divise le poids d'une unité par le nombre de bases de sondage à laquelle elle appartient : $m^{(A_q,d)} = 1/$ (nombre de bases de sondage qui contiennent d).

Cette technique est utilisée couramment à l'Insee, à la suite des premiers travaux de Deville et Lavallée sur le partage des poids Dans le cas de deux bases de sondages, on retrouve l'estimateur de Hartley, avec θ=1/2.

4.5. Spécificités des bases de sondage multiples

Le développement des enquêtes à bases de sondage multiples, plus complexes que les enquêtes à base de sondage simple, oblige à traiter de nouvelles questions méthodologiques.

Doit-on traiter la non-réponse sur chacun des échantillons, ou après les avoir combinés ? En particulier, les caractéristiques de la non-réponse peuvent être très différentes d'une enquête à l'autre, de sorte que des corrections de la non-réponse distinctes soient nécessaires.

Si on veut mettre en œuvre des méthodes de calage, il y a beaucoup de variantes possibles : on peut le faire au niveau de chaque base de sondage, par exemple sur les effectifs des strates, mais aussi directement sur des effectifs connus dans la population totale.

Les techniques d'estimation de variance doivent être adaptées au contexte des bases multiples.

Si les modes ou protocoles de collecte diffèrent selon la base de sondage, il y a un risque accru d'augmentation d'erreurs non dues à l'échantillonnage, comme les "effets de mode", qui apparaissent lorsque les individus répondent différemment selon le mode de collecte (face à face, téléphone, web,...). Ils sont souvent très difficiles à mesurer. Les effets de sélection peuvent aussi varier selon le mode de collecte.

Enfin, les méthodes d'estimation reposent toutes sur l'hypothèse que l'on sait avec certitude à quel domaine appartient chaque unité échantillonnée. Il faut donc mettre en place un protocole de collecte permettant de recueillir cette information avec le moins d'erreur possible, sinon les estimateurs risquent d'être biaisés en cas d'erreur de classification dans les domaines.

5. Erreurs sur les variables de la base de sondage

Quelles peuvent être les conséquences des erreurs qui peuvent survenir sur les variables de la base de sondage ?

Il est clair que des erreurs sur les données d'identification ou de repérage rendent la collecte plus difficile et plus coûteuse.

S'il y des erreurs sur les variables auxiliaires, les méthodes d'échantillonnage, comme la stratification ou le sondage équilibré, deviennent moins efficaces. Il en est de même pour les méthodes de correction de la non-réponse ou les méthodes de calage.

Le manque de fraîcheur de la base, dû à des mises à jour pas assez fréquentes, est malheureusement assez fréquent. Par exemple, dans le cas d'une enquête-entreprises, si l'on veut réaliser une stratification selon l'effectif salarié, mais qui date de 2 ans, on aura une stratification moins efficace, c'est-à-dire une moins bonne précision. Mais on risque aussi d'être confronté au problème des "strata jumpers", c'est-à-dire des entreprises mal classées dans les strates, et qui ont un poids "inapproprié", par exemple trop élevé compte tenu de leur véritable effectif. Ces entreprises deviennent alors trop influentes lors du calcul des estimations.

Bibliographie

Bankier, M.D. (1986), "Estimators based on several stratified samples with applications to multiple frame surveys", *Journal of the American Statistical Association* 81, p.1074–1079.

Deville, J.-C. (1998), "La correction de la non-réponse par calage ou par échantillonnage équilibré", *Actes du colloque de la Société Statistique du Canada*, Sherbrooke, Canada, p. 103–110.

Estevao, V.M., Särndal, C.-E. (2000), "A functional form approach to calibration", *Journal of Official Statistics*, 16, p. 379-399.

Fuller, W.A., Burmeister, L.F. (1972), "Estimators for samples selected from two overlapping frames", *Proceedings of the Social Statistics Section*, *American Statistical Association*, Alexandria, VA, p. 245–249.

Hartley, H.O. (1962), "Multiple frame surveys", *Proceedings of the Social Statistics Section, American Statistical Association*, Alexandria, VA, p. 203–206.

Kalton, G., Anderson, D.W. (1986), "Sampling rare populations", *Journal of the Royal Statistical Society, Series A* 149, p. 65–82.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Lavallée, P. (2002), *Le sondage indirect ou la méthode généralisée du partage des poids*, Bruxelles: Éditions de l'Université de Bruxelles.

Lessler, J., Kalsbeek, W. (1992), *Nonsampling Error in Surveys*, New York: Wiley.

Lohr, S.L. (2009), "Multiple frame surveys", D. Pfeffermann & C.R. Rao (éd), *Handbook of Statistics*, *Sample Surveys*: *Design*, *Methods and Applications*, Amsterdam : North Holland, Vol. 29A, p. 71-88.

Lohr, S. (2011), "Alternative survey sample designs: Sampling with multiple overlapping frames", *Survey Methodology*. 37, p. 197-213.

Lohr, S., Rao, J.N.K. (2000), "Inference in dual frame surveys", *Journal of the American Statistical Association* 95, p. 271–280.

Lohr, S., Rao, J.N.K. (2006), "Estimation in multiple-frame surveys", *Journal of the American Statistical Association* 101, p.1019–1030.

Mecatti, F. (2007), "A single frame multiplicity estimator for multiple frame surveys", *Survey Methodology*, 33, p. 151-157.

Särndal, C.-E., Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, Chichester: Wiley.

Sautory, O. (2003), "CALMAR2: A new version of the CALMAR calibration adjustment program", *Proceedings of Statistics Canada's Symposium 2003*.

Skinner, C.J., Rao, J.N.K. (1996), "Estimation in dual frame surveys with complex designs", *Journal of the American Statistical Association,* 91, p. 349–356.

Thompson, S.K. (1990), "Adaptive cluster sampling", *Journal of American Statistical Association* 85, p. 1050–1059.

Thompson, S.K. (2011), "Adaptive network and spatial sampling", *Survey Methodology* 37, p. 183–196.

Wolter, K.M. (1986), "Some coverage error models for census data", *Journal of the American Statistical Association*, 81, p. 338-346.