Journées de méthodologie statistique de l'Insee 2022

Utilisation des probabilités d'inclusion exactes pour le sondage indirect en population asymétrique

Henri Bodet(*), Arnaud Fizzala (**)

(*) Insee, Pôle Ingénierie Statistique d'Enquête
(**) Insee, Division Sondages du département des Méthodes Statistiques

henri.bodet@insee.fr ,arnaud.fizzala@insee.fr

Mots-clés. : Enquêtes auprès des entreprises, sondage indirect, partage des poids.

Domaines. : Théorie des sondages aval, pondération et repondération, Intégration de données, appariement et fusion de sources

Résumé

La méthode généralisé de partage des poids (MGPP) mise au point par Lavallée [1] est la solution habituelle pour les situations relevant du sondage indirect. Sa principale motivation est qu'elle fournit une solution lorsqu'on n'est pas capable de calculer les probabilités d'inclusion. Son application pose toutefois des problèmes lorsque les poids sont très dispersés et peuvent prendre de petites valeurs. Plusieurs solutions ont été proposées par Lavallée et Labelle-Blanchet [2], et l'une d'elles - la version pondérée de la MGPP - est à présent mise en application dans l'enquête sectorielle annuelle de l'Insee [3]

Cependant les enquêtes auprès des entreprises menées à l'Insee présentent deux particularités : les poids sont très dispersés (ils varient typiquement de 1 à 60) et le plan de sondage est très simple (il s'agit d'un sondage stratifié à un seul degré).

Dans ce cadre, nous proposons d'approfondir l'approche, mentionnée par Lavallée et Labelle-Blanchet [2], qui consiste à utiliser comme poids l'inverse de la probabilité d'inclusion théorique. Nous détaillons le calcul des probabilités d'inclusion d'ordre 1 et 2. Cela permet notamment d'estimer la variance de l'estimateur obtenu. Enfin nous donnons le résultat de simulations que nous avons menées sur des population de type "entreprises".

La conclusion est que cette méthode reposant sur un calcul exact aboutit -lorsque les poids initiaux sont dispersés - à des estimateurs plus précis que ceux reposant sur la MGPP, même lorsque les liens sont pondérés par une variable auxiliaire. Dans ce dernier cas, il faut intégrer l'information provenant de cette variable auxiliaire dans le processus d'estimation, par un calage par exemple.

Abstract

The so-called *generalised weight share method* (GWSM) due to Lavalée ([1]) is the usual solution when facing an indirect sampling issue. Its main motivation is that it provides unbiaised weights when inclusion probabilities cannot be calculated. However, its application sometimes leads to some trouble when initials weighs are very spread. Labelle-Blanchet [2] proposed several solutions to this problem and one of them (weighted weight share method) is currently applied in the Annual Sectorial Business Survey by Insee ([3].

Nevertheless, Business Survey conducted by Insee have two characteristics : firstly, weights are spread (they typically range from 1 to 60) and, secondly, their samplig design is very simple (a one-degree stratified random sampling).

We propose to go deeper in a direction mentionned by Lavalée and Labelle-Blanchet [2] : using as weight the inverse of the inclusion probability. We detail the computation of order 1 and 2 inclusion probabilities. This enables estimation of the variance of the estimator. We also give results of simulations made on a business-like population

We conclude that using the estimation based on exact probabilities is more accurate than the GSWM even when the links are weighted by an auxilary variable. The latter being true only if the auxilary variable is taken into account - for example throught a calibration procedure. Nevertheless, when sampling weights are homogeneous, MGGP remains an efficient solution - even though we can compute true inclusion probabilities.

1 Contexte, notations et objectifs

1.1 Problème général du sondage indirect

Nous nous intéressons ici à l'application du sondage indirect dans un cadre particulier mais fréquent en statistique d'entreprises.

On dispose d'une base de sondage dans une population $\mathbb{U}^A$ dans laquelle on tire un échantillon mais ce qui nous intéresse vraiment ce sont des unités d'une population $\mathbb{U}^B$ qui sont liées à une ou plusieurs unités de la population $\mathbb{U}^A$. Seulement, on ne connaît pas les liens avant de faire l'enquête. Une unité "cible" de $\mathbb{U}^B$ est sélectionnée dès qu'une unité de $\mathbb{U}^A$ qui lui est liée est dans l'échantillon.

Nous reprendrons les notations du livre de Lavallée[1], livre qui est la référence sur le sondage indirect nous noterons s^A l'échantillon sélectionné dans la population $\mathbb{U}^A$ et s^B l'échantillon final dans la population $\mathbb{U}^B$.

La question qui se pose est de savoir quel poids affecter aux unités de s^B en sachant que leur probabilité d'être inclus dans l'échantillon dépend de l'ensemble des unités de U_A qui leur sont liées.

La méthode générale de partage des poids fournit une réponse pratique et une méthode pour obtenir des estimations non biaisées à partir du moment où :

— chaque unité de $\mathbb{U}^B$ est liée à au moins une unité de $\mathbb{U}^A$;
— chaque unité de $\mathbb{U}^A$ a une probabilité non-nulle d'être sélectionnée ;

— on connaît pour chaque unité de s^B indirectement interrogée l'ensemble des unités de $\mathbb{U}^A$ (et non pas seulement de s^A) qui lui sont liées et auraient pu conduire à la sélectionner indirectement.

Nous donnons en annexe (page 14) plusieurs exemples d'enquêtes auprès des entreprises où nous pouvons nous trouver dans cette situation.

1.2 Le sondage indirect vu comme un plan de sondage induit

Une façon de considérer le problème est de se dire que le processus d'échantillonnage au sein de $\mathbb{U}^A$ qui conduit indirectement à sélectionner des unités de la population $\mathbb{U}^B$ à travers des liens que l'on ne connaît pas mais qui existent est, finalement, un processus d'échantillonnage au sein de $\mathbb{U}^B$.
On a donc en fin de compte un plan de sondage qui conduit à sélectionner un échantillon s^B au sein de $\mathbb{U}^B$.

Il "suffit" donc de déterminer les caractéristiques de ce plan de sondage pour obtenir des estimateurs sans biais des totaux et des estimateurs de variance. Généralement, on considère que les calculs sont trop complexes pour pouvoir être menés – c'est l'une des motivations de la méthode généralisée du partage des poids qui permet d'obtenir des estimateurs sans biais sans réaliser ces calculs.

Il se trouve que, à l'Insee, les échantillons des enquêtes auprès des entreprises sont le plus souvent tirés selon des sondages aléatoires simples stratifiés. Dans ce cadre, les calculs sont réalisables comme mentionné dans Lavallée et Labelle-Blanchet [2]. Dans le présent article, nous détaillons le calcul de la probabilité d'inclusion d'ordre un, ainsi que le calcul de la probabilité d'inclusion d'ordre deux (mentionné mais non réalisé dans l'article [2]). Il en découle un estimateur de la variance de l'estimateur obtenu.

Nous reprendrons les notations classiques des ouvrages et articles de Pierre Lavallée : nous utiliserons la lettre A (en exposant) pour désigner ce qui est relatif à la population "source" A et la lettre B pour ce qui est relatif à la population "cible" B. Ainsi, avec les notations habituelles de la théorie des sondages, π_i^A désigne la probabilité d'inclusion d'une unité i de $\mathbb{U}^A$ dans le plan de sondage initial et π_α^B celle d'une unité α de $\mathbb{U}^B$ dans le plan de sondage induit. Nous utiliserons l'alphabet latin pour les indices dans la population $\mathbb{U}^A$ et l'alphabet grec pour ceux dans la population $\mathbb{U}^B$.

Dans toute la suite, nous supposerons que le plan de sondage sur $\mathbb{U}^A$ est un plan stratifié à un degré comprenant H strates et nous noterons N_h et n_h les effectifs de la population et de l'échantillon sur la strate h.

1.3 Objectifs de ce texte

Nous n'envisagerons que le cas des enquêtes auprès des entreprises (dont les associations font partie du point de vue de la statistique publique). C'est-à-dire des plans de sondages stratifiés à un seul degré, portant sur des populations asymétriques et pour lesquelles on dispose parfois d'une information auxiliaire via des registres administratifs (comme l'effectif salarié ou le chiffre d'affaires).

Nous proposons dans cette communication les éléments suivants :
— le calcul des probabilités d'inclusion exactes sur la population-cible ;
— le calcul des probabilités d'inclusion d'ordre deux qui permettent une estimation de la variance d'échantillonnage ;
— une simulation sur des données proches de données "entreprises" – cette simulation permet de comparer plusieurs méthodes :
  • le partage des poids "classique" et "pondéré"
  • une variante du partage des poids qui prend en compte les unités de la partie exhaustive
  • l'utilisation des probabilités d'inclusion exactes

2 Probabilités d'inclusion exactes s'il y a deux unités liées

Nous allons commencer par déterminer les probabilités d'inclusion dans le cas où deux unités de $\mathbb{U}^A$ sont liées à une unité de la population cible. Traiter ce cas est mathématiquement inutile parce qu'il est contenu dans le cas général décrit ci-dessous. Toutefois, il est plus simple à comprendre et, parfois, il se peut qu'il n'y ait que deux liens – comme par exemple le cas de l'enquête association décrit page 15.

On considère qu'il y a deux unités i et j de $\mathbb{U}^A$ qui sont liées à une unité α de $\mathbb{U}^B$. L'unité observée α est sélectionnée dans l'échantillon s^B dès que l'une des deux unités i ou j est dans l'échantillon s^A.

On se propose de déterminer ici π^B_α qui est la probabilité d'inclusion dans s^B de l'unité α. On a donc $\pi^B_\alpha = \mathbb{P}\left(\{i \in s^A\} \cup \{j \in s^A\}\right)$

De façon générale, on a

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \mathbb{P}\left(i \in s^A\right) + \mathbb{P}\left(j \in s^A\right) - \mathbb{P}\left(i \text{ et } j \in s^A\right) \tag{1}$$

2.1 Premier cas : i et j sont dans la même strate h

Si on note n_h et N_h la taille de l'échantillon et de la population de la strate h (avec, forcément, $2 \leq n_h \leq N_h$), on a $\pi^A_i = \pi^A_j = \dfrac{n_h}{N_h}$ et $\mathbb{P}\left(i \text{ et } j \in s\right) = \dfrac{n_h(n_h-1)}{N_h(N_h-1)}$.

Et, donc, d'après (1)

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \mathbb{P}\left(\alpha \in s^B\right) = \frac{n_h}{N_h}\left(2 - \frac{n_h - 1}{N_h - 1}\right) \tag{2}$$

En particulier :
— si la strate h est exhaustive : $n_h = N_h$ et $\pi^B_\alpha = 1$
— si le taux de sondage est négligeable :$\pi^B_\alpha \approx 2\dfrac{n_h}{N_h}$

2.2 Deuxième cas : i et j sont dans deux strates différentes h et k

Les tirages dans les strates étant indépendants, on a : $\mathbb{P}\left(i \text{ et } j \in s^A\right) = \pi^A_i \pi^A_j$. Et, ici, $\pi^A_i = \dfrac{n_h}{N_h}$ et $\pi^A_j = \dfrac{n_k}{N_k}$.

La relation (1) donne donc :

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \mathbb{P}\left(\alpha \in s^B\right) = \frac{n_h}{N_h} + \frac{n_k}{N_k} - \frac{n_h}{N_h}\frac{n_k}{N_k} \tag{3}$$

En particulier :
— si une des strates (mettons h) est exhaustive : $n_h = N_h$ et $\pi_\alpha^B = 1$
— si l'un des deux taux de sondage est négligeable (mettons celui de la strate k) : $\pi_\alpha^B \approx \frac{n_h}{N_h}$

2.3 Application au contexte de l'appariement imparfait de deux bases de sondages

Il s'agit de la situation de l'enquête Association décrite page 15 : on tire l'échantillon dans deux répertoires distincts mais qui comprennent des doublons que l'on identifie lors de la collecte. Ces doublons sont les associations inscrites dans les deux répertoires.

Dans ce cas, il ne peut y avoir que deux unités liées et qui sont forcément dans des strates distinctes car le répertoire est un critère de stratification.

Soient i et j deux unités de la base de sondages qui représentent la même unité économique α (dans notre exemple, α serait une association inscrite au répertoire Sirene avec l'identifiant i et au Répertoire Nationnal des Associations avec l'identifiant j).

Dans ce cas-là, on peut déterminer le poids w_α à affecter à partir des poids des unités liées - sans retourner au plan de sondage, en appliquant la formule (3) :

$$w_\alpha = \frac{w_i w_j}{w_i + w_j - 1} \tag{4}$$

Remarquons au passage que :
— si les poids sont proches et bien plus grands que l'unité ($w_i \approx w_j >> 1$), alors $w_\alpha \approx \frac{w_i}{2}$ *i.e.* le poids issu de l'application de la MGPP lorsqu'une seule unité est dans l'échantillon ;
— si l'un des poids est beaucoup pus grand que l'autre (mettons $w_i >> w_j$), alors le poids à retenir est proche du plus petit ($w_\alpha \approx w_j$)

3 Cas où plus de deux unités sont liées à l'unité cible

On suppose ici qu'il y a m unités de la population $\mathbb{U}^A$ liées à l'unité cible α de $\mathbb{U}^B$.

3.1 Cas où toutes les unités sont dans la même strate h

Notons m le nombre d'unités liées à l'unité α. On peut supposer que $1 \leq m \leq N_h - n_h$. En effet, il faut au moins une unité liée et s'il y a plus de $N_h - n_h$ unités liées, alors l'échantillon en comprend forcément une et la probabilité d'inclusion est 1.

L'unité cible α n'est pas sélectionnée si, et seulement si, aucune des m unités liées ne l'est. C'est-à-dire si l'échantillon est tiré dans les $N_h - m$ autres unités de la strate.

Comme les unités sont tirés dans la strate via un sondage aléatoire simple , il y a $\binom{N_h-m}{n_h}$ échantillons qui ne sélectionnent pas l'unité α parmi les $\binom{N_h}{n_h}$ échantillons possibles.

On en déduit que :

$$\mathbb{P}\left(\alpha \text{ non sélectionnée }\right) = \frac{\binom{N_h-m}{n_h}}{\binom{N_h}{n_h}} = \prod_{l=0}^{m-1} \frac{N_h - n_h - l}{N_h - l}$$

Finalement, on trouve l'expression suivante :

$$\mathbb{P}\left(\alpha \text{ sélectionnée }\right) = 1 - \prod_{l=0}^{m-1} \frac{N_h - n_h - l}{N_h - l} \tag{5}$$

On peut vérifier que dans les cas simples on retrouve les résultats connus :
— si $m = 1$, l'expression (5) devient $\dfrac{n_h}{N_h}$
— si $m = 2$, l'expression (5) devient $\dfrac{n_h}{N_h}\left(2 - \dfrac{n_h - 1}{N_h - 1}\right)$

3.2 Cas où les unités sont dans plusieurs strates

Notons m le nombre total d'unités liées et m_h le nombre figurant dans la strate h. On peut supposer que toutes les strates vérifient $1 \leq m_h \leq N_h - n_h$. Pour les mêmes raisons que ci-dessus : si $m_h = 0$, la strate n'intervient pas dans le calcul et si $m_h > N_h - n_h$, alors la probabilité d'inclusion est égale à l'unité.

Numérotons de 1 à H les strates impliquées et notons α_h les unités de la strate h appartenant à l'échantillon.

Avec cette notation, $\mathbb{P}\left(\alpha \text{ non sélectionnée}\right) = \mathbb{P}\left(\forall h \in \{1 \cdots H\} \text{ aucune unité de } \alpha_h \text{ n'est sélectionnée}\right)$. Comme les tirages entre les strates sont indépendants, on en déduit que :

$$\mathbb{P}\left(\alpha \text{ non sélectionnée}\right) = \prod_{h=1}^{H} \mathbb{P}\left(\text{ aucune unité de } \alpha_h \text{ n'est sélectionnée}\right)$$

Ceci s'écrit aussi :

$$\mathbb{P}\left(\alpha \text{ non sélectionnée}\right) = \prod_{h=1}^{H} \left(1 - \mathbb{P}\left(\text{ une unité de } \alpha_h \text{ est sélectionnée}\right)\right) \tag{6}$$

.

Or, la probabilité qu'une unité de α_h soit sélectionnée est donnée par le résultat (5). On notera $m_{\alpha,h}$ le nombre d'unités liées à α appartenant à la strate h

$$\mathbb{P}\left(\alpha \text{ non sélectionnée}\right) = \prod_{h \text{ tel que } m_{\alpha,h}>0} \prod_{l=0}^{m_{\alpha,h}-1} \frac{N_h - n_h - l}{N_h - l}$$

.

Et l'expression de la probabilité d'inclusion :

$$\mathbb{P}\left(\alpha \text{ sélectionnée}\right) = 1 - \prod_{h \text{ tel que } m_{\alpha,h}>0} \prod_{l=0}^{m_{\alpha,h}-1} \frac{N_h - n_h - l}{N_h - l} \tag{7}$$

.

Le poids à appliquer à l'unité α de s_B est finalement : $w_\alpha = \dfrac{1}{\mathbb{P}\left(\alpha \text{ sélectionnée}\right)}$.

L'expression générale ne se simplifie pas mais on peut expliciter le cas particulier où l'unité α est liée à à L_α unités de $\mathbb{U}_A$ appartenant à les strates distinctes. Dans ce cas, on a l'expression suivante :

$$w_\alpha = \frac{\prod\limits_{l=1,..,L_\alpha} w_l}{\prod\limits_{l=1,..,L_\alpha} w_l - \prod\limits_{l=1,..,L_\alpha} (w_l - 1)}$$

4 Comparaison avec le partage des poids

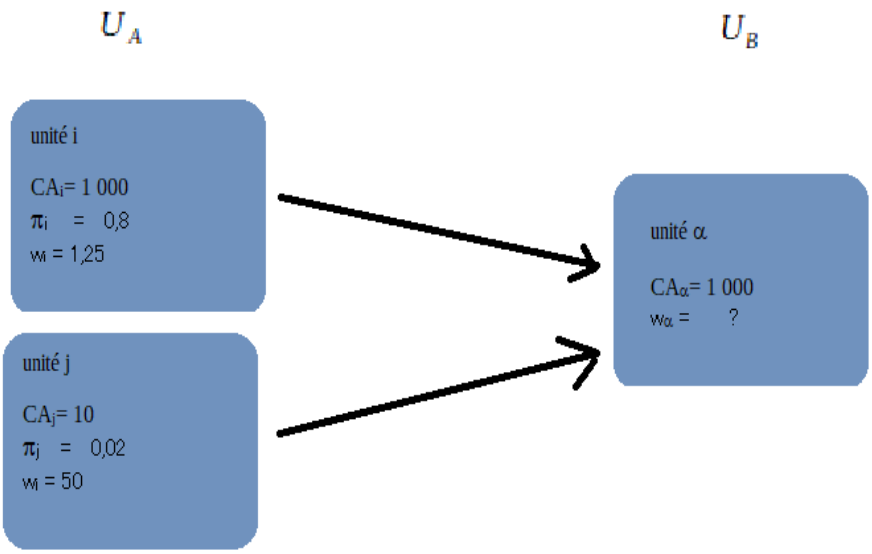
Les poids issus du calculs des probabilités exacts possèdent - comme ceux issus de la méthode de partage des poids - la propriété de fournir des estimateurs sans biais.

En revanche, ils possèdent d'autres propriétés qui rendent leur manipulation plus aisée :

- $w_\alpha \leq w_i$, pour tout i lié à α
- $w_\alpha \geq 1$
- Si pour un i lié à α, $w_i = 1$, alors $w_\alpha = 1$ (Les unités exhaustives restent exhaustives..)

Ces propriétés sont importantes dans le cas d'un sondage auprès des entreprises où les poids sont très dispersés et parfois proches de l'unité : affecter des trop grands poids à des unités dont les poids initiaux sont faibles pose problème. De même, l'application de la MGPP peut conduire à des poids inférieurs à l'unité.

Pour illustrer cela, on propose de se placer dans le cas où l'unité α de s_B est liée à deux unités i et j de s_A.



Situation	MGPP	MGPP Pondérée	Probabilités exactes
i et $j \in s^A$	25,63	1,73	1,24
$i \in s^A$ et $j \notin s^A$	0,63	1,23	1,24
$i \notin s^A$ et $j \in s^A$	25	0,50	1,24

Cette exemple illustre bien l'avantage de recourir aux probabilités exactes dans le cas d'une population asymétrique. Dans l'exemple ci-dessus, on n'accepterait ni le poids de 25,63 - qui semble trop élevé par rapport au chiffre d'affaires de l'unité liée α - ni les poids de 0,63 ou 0,50 - qui sont inférieurs à l'unité et difficiles à utiliser en pratique.

Si les poids n'avaient pas été aussi différents - ce qui est généralement adapté lorsque la variable étudiée est distribuée de façon asymétrique - le problème ne se serait pas posé. Or, les enquêtes auprès des entreprises sont souvent des enquêtes en population asymétriques. Il n'est pas rare que les poids varient de 1 à 60. Cette situation s'y produit donc régulièrement.

5 Probabilités d'inclusion doubles et variance de l'estimateur

5.1 Calcul des probabilités d'inclusion doubles

Le calcul des probabilités d'inclusion doubles de deux unités α et β de $\mathbb{U}^B$ peut se faire sans difficultés à l'aide de l'égalité suivante :

$$\mathbb{P}\left(\alpha \text{ et } \beta \in s^B\right) = \mathbb{P}\left(\alpha \in s^B\right) + \mathbb{P}\left(\beta \in s^B\right) - \mathbb{P}\left(\alpha \text{ ou } \beta \in s^B\right) \tag{8}$$

Les deux premiers termes du membre de droite peuvent se calculer à l'aide de la relation (7).

Le dernier terme également. En effet, si l_α désigne les unités de $\mathbb{U}^A$ liées à α, alors α ou $\beta \in s^B$ signifie que l'une des unités de l_α ou de l_β est dans l'échantillon s^A.

Ce terme $\mathbb{P}\left(\alpha \text{ ou } \beta \in s^B\right)$ est donc la probabilité d'inclusion d'une unité $\alpha \cup \beta$ qui serait liée à l'ensemble $l_{\alpha\cup\beta} = l_\alpha \cup l_\beta$ et peut donc se calculer aussi à l'aide de la relation (7).

Si, pour toute strate h, $m_{\alpha\cup\beta,h} < N_h + n_h$ alors :

$$\begin{aligned}\mathbb{P}\left(\alpha \text{ et } \beta \in s^B\right) = 1 + \prod_{h/m_{\alpha\cup\beta,h}>0} \prod_{l=0}^{m_{\alpha\cup\beta,h}-1} \frac{N_h - n_h - l}{N_h - l} - \prod_{h/m_{\alpha,h}>0} \prod_{l=0}^{m_{\alpha,h}-1} \frac{N_h - n_h - l}{N_h - l} \\ - \prod_{h/m_{\beta,h}>0} \prod_{l=0}^{m_{\beta,h}-1} \frac{N_h - n_h - l}{N_h - l}\end{aligned} \tag{9}$$

.

Sinon, c'est-à-dire s'il existe une strate h telle que $m_{\alpha\cup\beta,h} \geq N_h + n_h$,
on a $\mathbb{P}\left(\alpha \text{ ou } \beta \in s^B\right) = 1$ et

$$\mathbb{P}\left(\alpha \text{ et } \beta \in s^B\right) = 2 - \mathbb{P}\left(\alpha \in s^B\right) - \mathbb{P}\left(\beta \in s^B\right) \tag{10}$$

L'expression de $\mathbb{P}\left(\alpha \in s^B\right)$ et de $\mathbb{P}\left(\beta \in s^B\right)$ dépend alors du nombre d'unités liées dans chaque strate.

5.2 Expression des termes $\Delta^B_{\alpha,\beta}$

Le terme $\Delta^B_{\alpha,\beta} = \pi^B_{\alpha,\beta} - \pi^B_\alpha \pi^B_\beta$ intervient dans le calcul de la variance de l'estimateur de Horvitz-Thomson. Il est donc important de pouvoir le déterminer pour calculer puis, en pratique, estimer la précision du plan de sondage induit sur $\mathbb{U}^B$. Il a en outre une interprétation : s'il est nul, la sélection des unités α et β est indépendante, s'il est positif, il y a plus de chances de sélectionner α lorsque β est sélectionnée (et vice versa).

Une fois que l'on dispose des probabilités d'inclusion simples – données par (7) – et des probabilités d'inclusion doubles – données par (9), on obtient sans problèmes $\Delta^B_{\alpha,\beta}$.

De plus, afin que l'estimateur de variance d'Horvitz-Thompson soit sans biais ([6], p.139), il est nécessaire d'avoir $\pi^B_{\alpha,\beta} \neq 0$. Or, il n'est pas possible que l'on ait $\pi^B_{\alpha,\beta} = 0$ – sauf s'il y a une strate où $n_h = 1$ et $N_h > 1$ et que cette strate contient toutes les unités liées à α ou β. Ce cas n'ayant que peu de chances de se produire en pratique, on peut l'exclure sans vraie perte de généralité.

On peut donc supposer que la plupart du temps $\pi^B_{\alpha,\beta} > 0$

5.3 Calcul de la variance

On peut appliquer un théorème de Hortiz et Thomson (voir par exemple [6], p.136) qui donne, à partir des $\Delta^B_{\alpha,\beta}$, l'expression de la variance de l'estimateur $\hat{Y^B}$ du total d'une variable y définie sur $\mathbb{U}^B$:

$$\mathbb{V}\left(\hat{Y^B}\right) = \sum_{\alpha,\beta \in \mathbb{U}^B} \Delta^B_{\alpha,\beta} \frac{y_\alpha}{\pi^B_\alpha} \frac{y_\beta}{\pi^B_\beta} \tag{11}$$

Comme nous avons vu que l'on a toujours $\pi^B_{\alpha,\beta} > 0$, on a un estimateur sans biais de cette quantité ([6], p.139) grâce à :

$$\hat{\mathbb{V}}\left(\hat{Y^B}\right) = \sum_{\alpha,\beta \in s^B} \Delta^B_{\alpha,\beta} \frac{y_\alpha}{\pi^B_\alpha} \frac{y_\beta}{\pi^B_\beta} \frac{1}{\pi^B_{\alpha,\beta}} \tag{12}$$

Dans les situations où peu d'unités sont impliquées dans le sondage indirect (c'est par exemple le cas de l'enquête Associations décrite page 15 : seules 7 % des unités sont liées à une autre), s'il y a de nombreuses strates, il y a de forte chances pour que l'échantillon ne contienne, pour un bon nombre d'entre elles, que des unités de s^B liées à une seule unité de $\mathbb{U}^A$.

On peut donc partager les H strates en deux parties : un ensemble H_1 de strates où les unités de l'échantillon s^B_1 ne sont liées qu'à une unité de $\mathbb{U}^A$ et où l'on pourra utiliser les propriétés (poids, variance...) habituelles des sondages aléatoire simple stratifiés, et un ensemble H_2 contenant le reste des strates où l'on utilisera les propriétés développées dans cet article.

6 Simulations

6.1 Le contexte des simulations

Nous avons travaillé à partir de données simulées mais reproduisant le cadre des enquêtes auprès des entreprises, en particulier de l'enquête sectorielle annuelle (ESA) [1] : l'échantillon s^A

1. En pratique le cadre de l'ESA est plus complexe puisque le tirage des unités légales est effectué selon un tirage aléatoire simple de grappes d'unités légales (et non d'unités légales directement)

est composé d'unités légales permettant de constituer un échantillon d'entreprises s^B selon la règle suivante : une entreprise est dans l'échantillon si au moins une de ses unités légales a été tirée [3].

Les données sont constituées de cette façon : On génère une population $\mathbb{U}^A$ de 1600 unités légales réparties en 3 strates : la strate 1 contient les "petites entreprises" (taux de sondage faible), la strate 2 contient les entreprises de tailles moyennes (taux de sondage moyen), et la strate 3 contient les "grandes entreprises" (strate exhaustive).

Pour chaque unité légale j, on génère un chiffre d'affaires CA_j selon une loi normale dont les paramètres dépendent de la strate de tirage (voir table I - paramètres de la simulation). Ce chiffre d'affaires correspond à une variable auxiliaire : on suppose qu'elle serait disponible pour l'ensemble des unités de la base de sondage en situation réelle. Puis en s'appuyant sur ce chiffre d'affaires, on génère une variable d'intérêt : la valeur ajoutée VA_j en suivant le protocole suivant :

— génération d'un coefficient c_j selon une loi uniforme sur $[0; 1]$
— $VA_j = c_j \times CA_j$

Les 1600 unités légales sont ensuite regroupées aléatoirement en 500 entreprises. La valeur ajoutée d'une entreprise i correspond alors à la somme des valeurs ajoutés de ses unités légales.

On réalise ensuite 50000 tirages d'échantillons suivant les allocations indiquées table 4.

TABLE 1 – Paramètres de simulations

Strate	N^A	n^A	**Loi** CA_j
1	1000	30	$N(100, 50)$
2	500	50	$N(1000, 200)$
3	100	100	$N(2000, 500)$

Pour chaque échantillon, cinq pondérations sont calculées :

— L'application directe de la méthode généralisée de partage des poids ;
— Une variante de la méthode précédente qui traite séparément les unités exhaustives [2] ;
— L'application de la méthode généralisée de partage des poids en pondérant les liens par une variable auxiliaire (ici, le CA fortement corrélé à la variable d'intérêt) ;
— Une variante de la méthode précédente qui traite séparément les unités exhaustives ;
— L'utilisation des probabilités d'inclusion exactes données par la formule (7).

De plus, pour chaque pondération, deux méthodes d'estimation sont considérées :

— l'application "directe" des pondérations à l'échantillon pour avoir une estimation du total par expansion ($\hat{VA}_{exp} = \sum_{i \in s^B} w_i VA_i$) ;
— l'application d'un estimateur par le ratio : nous avons appliqué les mêmes pondérations pour avoir un estimateur($\hat{CA}_{exp} = \sum_{i \in s^B} w_i CA_i$) du total du chiffre d'affaires CA supposé connu et nous avons retenu comme estimation : $\hat{VA}_{ratio} = \hat{VA}_{exp} \dfrac{CA}{\hat{CA}_{exp}}$

Finalement, pour chacun des 50000 échantillons, nous produisons 10 estimations du total de la valeur ajoutée que nous pouvons comparer au vrai total $VA = \sum_{j \in U^B} VA_j$ connu ici puisque nous avons généré les données.

2. Plus précisément, si une entreprise a au moins un lien avec une unité légale de la strate exhaustive, alors on lui attribue un poids de 1, sinon on lui attribue le poids obtenu via la MGPP.

Afin d'évaluer la précision d'un estimateur e donné (parmi les 10 envisagés), nous calculons le coefficient de variation Monté-Carlo $CV_e = \frac{\sqrt{\frac{1}{50000}\sum_{r=1}^{50000}(\hat{VA}_{e,r} - VA)^2}}{VA}$

Des biais Monte-Carlo ont également été calculés à des fins de vérification des programmes, et confirment que les 10 procédures d'estimations sont bien sans biais.

Les résultats obtenus confirment que l'utilisation des probabilités exactes conduit à des estimateurs plus précis [3] :

TABLE 2 – CV (en %) des différents estimateurs de la valeur ajoutée - données simulées

	MGPP classique		**MGPP pondérée**		**Probabilité**
type d'estimateur	**standard**	**+ exhaustivité**	**standard**	**+ exhaustivité**	**exacte**
Application directe des pondérations	8.3	7.1	4.3	5.2	5.0
Estimateur par le ratio	5.3	4.0	3.7	3.4	3.2

En annexe (page 17), sont présentés :
- Une autre simulation basée sur des taux de sondage homogènes et des données moins asymétriques. Dans ce cadre, la MGPP semble aboutir à de meilleures performances que le recours aux probabilités exactes ;
- Un début de réflexion au sujet de la prise en compte de la non-réponse.

3. Nous considérons que la MGPP pondérée ne doit pas être comparée aux autres méthodes lorsque ces dernières n'utilisent pas la variable de pondération des liens comme variable auxiliaire dans le processus d'estimation. Autrement dit nous ne comparons pas "directement" 4.3 à 5 dans le tableau.

7 Conclusion

Nous avons donc vu qu'il est possible - dans le cas d'un sondage stratifié - de calculer de façon simple la probabilité d'inclusion exacte pour qu'une unité liée à d'autres soit indirectement sélectionnée. Pour cela, il suffit de connaître le nombre d'unités avec qui elle est liée dans chaque strate. Une fois ce calcul programmé [4], on peut l'utiliser pour obtenir les probabilités d'inclusion d'ordre 2 et estimer la variance.

Nous avons ensuite mis en oeuvre des simulations sur un jeu de données simple, afin de tester l'intuition que nous avions sur les bonnes performances des poids issus de la probabilité exacte. Sur la base de ces simulations, nous pouvons dire que :

— Lorsque les poids initiaux sont très dispersés - ce qui est le cas dans la plupart des enquêtes auprès des entreprises - cette méthode aboutit à des estimateurs plus précis que la MGPP et, surtout, fournit des poids plus simples à manipuler (plus petits et supérieurs à l'unité).

— Lorsqu'on dispose d'une variable auxiliaire quantitative, on peut utiliser la MGPP pondérée. Dans ce cas, les avantages d'utiliser les probabilités exactes demeurent mais seront moindres si la variable auxiliaire est fortement liée aux probabilités d'inclusion des unités initiales ou à la variable d'intérêt. Toutefois, on peut vraisemblablement dans cette situation recourir aux probabilités exactes et utiliser la variable auxiliaire par exemple lors d'un calage sur marges. Dans cette situation, les méthodes semblent équivalentes en termes de précision.

— Dans le cas où les poids sont dispersés et où on ne dispose pas de variable auxiliaire, l'utilisation des probabilités exactes semble clairement la meilleure option.

— Si les poids sont homogènes, les inconvénients de la MGPP s'amenuisent et cette méthode donne, dans notre simulation, de meilleurs résultats que l'utilisation des probabilités exactes.

Le mieux serait de confirmer ces intuitions en les démontrant mathématiquement. Cela devrait permettre de déterminer, pour une situation donnée, si l'utilisation des probabilités exactes est préférable à la MGPP. Ces travaux restent à mener.

4. Des fonctions R mettant en œuvre le calcul des probabilités exactes et également l'application de la MGPP seront mises à disposition sur github prochainement.

Bibliographie

Liste des documents cités

[1] Pierre Lavallée. *Indirect sampling* Springer Series in Statistics, 2007.

[2] Pierre Lavallée et Sébastien Labelle-Blanchet. *Le sondage indirect appliqué aux populations asymétriques* Techniques d'enquête, Vol. 39, No 1, pp. 207-241, juin 2013.

[3] Arnaud Fizzala. *La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle.* Acte des Journées de Méthodologie Statistique de l'Insee 2018.

[4] Ronan Le Gleut et Thomas Merly-Alpa. *L'impact du profilage sur la refonte du plan de sondage des Enquêtes Sectorielles Annuelles.* Acte des Journées de Méthodologie Statistique de l'Insee 2018.

[5] Camilia Coga *Cours de sondages dispensé à l'université de Besançon.* http ://-goga.perso.math.cnrs.fr/

[6] Pascal Ardilly *Les techniques de Sondage.* Editions TECHNIP, 2006.

[7] *Les entreprises en France* INSEE Références, Edition 2019.

8 Annexes

8.1 Exemples de cas concrets d'utilisation de sondage indirect pour des enquêtes auprès des entreprises

8.1.1 Enquête sectorielle annuelle (dite ESA)

Depuis le millésime 2017, et conformément au règlement européen encadrant les statistiques structurelles d'entreprises, les résultats issus du système d'élaboration des statistiques annuelles d'entreprises (ESANE) sont publiés au niveau des entreprises profilées (EP) [7]. ESANE se base, entre autres, sur l'enquête sectorielle annuelle (ESA) et l'enquête annuelle de production (EAP). Aussi, dès le millésime 2016, le tirage des échantillons de ces enquêtes est réalisé au niveau des entreprises profilées (EP). Lorsqu'une entreprise profilée est tirée, toutes les unités légales (UL) relevant du champ de l'enquête (en tant qu'UL) qui lui sont rattachées sont sélectionnées dans l'échantillon d'UL correspondant. On envoie alors un questionnaire aux UL de cet échantillon, et les réponses des EP sont ensuite « reconstituées » à partir des retours de questionnaires des UL.

Au moment du tirage, en novembre, les contours des EP sont provisoires, et c'est plus tard, en mars, que l'information à jour sur les contours est utilisée pour mettre à jour l'échantillon avec la règle suivante : l'échantillon d'EP est constitué de l'ensemble des EP dont au moins une UL du contour mis à jour appartient à l'échantillon initial d'UL.

Nous sommes donc typiquement dans une situation de sondage indirect avec les éléments suivant

- $\mathbb{U}^A$: UL dans la base de sondage ;
- s^A : UL échantillonnées ;
- $\mathbb{U}^B$: EP (nouveaux contours) dans le champ dans l'enquête et avec au moins une UL dans $\mathbb{U}^A$;
- s^B : EP de $\mathbb{U}^B$ avec au moins une UL dans s^A.

Jusqu'à présent, cette mise à jour de l'échantillon est réalisée via la méthode généralisée de partage des poids en pondérant les liens par le chiffre d'affaires des UL [3]. Utiliser les développements présentés dans la suite n'est pour l'instant pas envisageable du fait que l'échantillon d'UL s^A n'est pas issu d'un tirage aléatoire simple stratifié mais d'un tirage en grappes stratifié (les grappes étant les EP avec leurs anciens contours), mais si les calculs de probabilités d'inclusion d'unités liées présentés dans ce papier étaient étendus aux tirages en grappes, leur utilisation dans le cadre de l'ESA serait surement intéressante.

8.1.2 Enquête auprès des sous-traitants d'un secteur

Pour connaître le poids économique d'un secteur particulier (mettons la construction navale), on peut avoir besoin d'interroger les entreprises qui travaillent pour ce secteur. Seulement, l'information sur la nature de l'activité de l'entreprise n'est pas suffisante. Par exemple, il y a des entreprises de menuiserie qui n'interviennent que pour aménager les cabines des bateaux. Cette spécificité n'est pas prise en compte dans la description administrative de leur activité et on ne peut donc pas les repérer.

Une stratégie envisageable serait d'interroger les entreprises du secteur de la construction navale (population $\mathbb{U}^A$) et de leur demander la liste de leurs sous-traitants et prestataires (population $\mathbb{U}_B$). Dans un second temps, on interrogerait les entreprises de la population $\mathbb{U}^B$ en leur demandant la liste de leur clients du secteur de la construction navale - voire éventuellement le

montant des facturations (variable auxiliaire).

On voit bien, qu'en l'absence de traitement particulier, une entreprise qui a plusieurs clients aurait plus de chances d'être sélectionnée qu'une entreprise qui n'en a qu'un seul. Ne pas corriger ce biais serait d'autant plus contrariant qu'en général, un des objectifs de ces enquêtes est d'évaluer la dépendance des sous-traitants par rapport à leurs donneurs d'ordre. Une absence de correction conduirait donc à sous-estimer le nombre de sous-traitant et prestataires n'ayant qu'un seul ou deux clients.

En pratique, le plus simple pour prévenir l'absence de biais est d'interroger de façon exhaustive les donneurs d'ordre ou en interrogeant les "potentiels preneurs d'ordre" quitte à avoir beaucoup d'entreprises hors-champ. C'est ce qui est fait dans plusieurs enquêtes de l'Insee - qui se restreignent aux cas où la population $\mathbb{U}^A$ est petite - mais il serait intéressant de pouvoir n'interroger que partiellement les donneurs d'ordres. [5] D'autant plus qu'on dispose d'une information fiscale sur la sous-traitance confiée mais pas sur la sous-traitance reçue.

8.1.3 Enquête auprès des associations

Lorsqu'il s'agit d'enquêter les "associations" – la difficulté est qu'il n'y a pas de base de sondage "propre". Les associations peuvent être inscrites à deux répertoires :

— le répertoire national des associations (RNA) géré par le ministère de l'Intérieur ;
— le répertoire des entreprises et des établissements (Sirene) géré par le ministère de l'Économie.

Lors de la constitution de la base de sondage, des procédures sont mises en œuvre pour apparier les deux répertoires et faire en sorte que dans la base de sondage, chaque association ne soit présente qu'une seule fois. Malgré ces efforts, on sait bien que certaines associations sont présentes deux fois dans la base de sondage. C'est pourquoi on demande à chaque unité enquêtée au titre d'un répertoire si elle est immatriculée dans l'autre et sous quel numéro - ce qui fournit les "liens" de la MGPP.

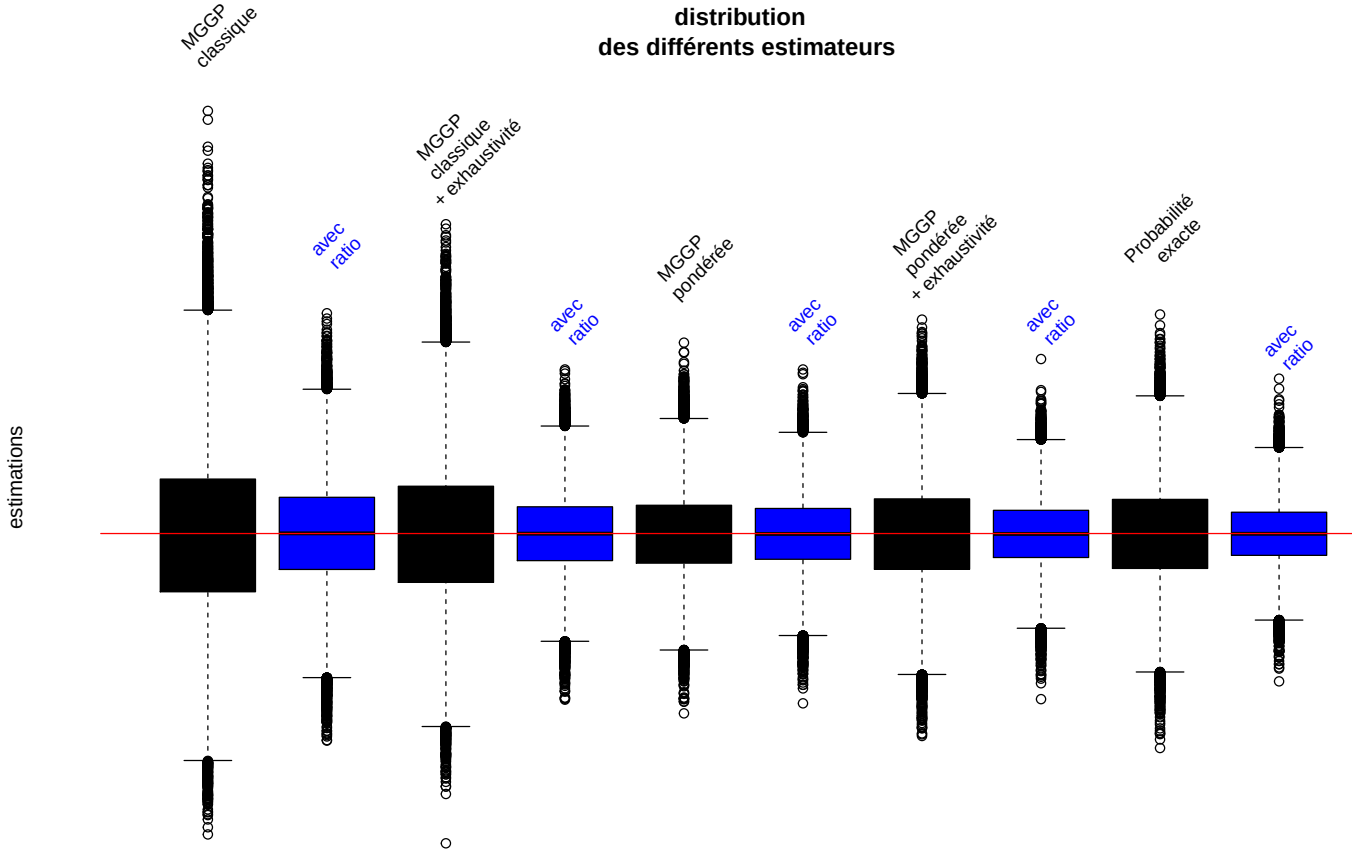
On voit bien que si une association interrogée est représentée par deux unités, sa probabilité d'inclusion sera plus élevée que celle de l'unité issue du répertoire qui a été sélectionnée et a conduit à l'interroger. Il faut donc effectuer un traitement spécifique. D'autant plus que ces cas ne sont pas rares : lors de l'enquête de 2015, 7 % des associations interrogées étaient représentées par plus d'une unité et 3.5 % lors de l'enquête de 2019.

Dans ce cas-là, la population "cible" $\mathbb{U}^B$ est constituée des associations "réelles" qui peuvent être liées à une ou deux unité de la population $\mathbb{U}^A$ constituée des "immatriculations après appariement". On dispose des liens (puisqu'on demande à chaque association par quelle autre unité elle est éventuellement représentée) mais il n'y a pas de variable auxiliaire qui permette de calculer des liens pondérés.

5. Lorsque ces enquêtes ont été envisagées, des problèmes juridiques sont aussi survenus : d'une part, le secret statistique (loi de 1951) interdit de divulguer à d'autres unités les réponses à une enquête, d'autre part la réglementation (loi Informatique et Liberté de 1978) oblige à communiquer à une unité les informations dont dispose sur elle.

8.2 Simulations - suite

Le graphique suivant donne la distribution des différents estimateurs. On note que la MGPP classique est plus susceptible de produire des valeurs extrêmes. Ceci est dû au fait que le poids des grandes unités peu augmenter énormément "par hasard".



Nous avons rajouté la comparaison avec l'estimateur par le ratio car, si on dispose d'une variable auxiliaire aussi efficace, il n'y a pas de raison qu'on ne l'utilise pas quelle que soit le jeu de pondération utilisé. De fait, il est plus pertinent de comparer l'estimateur par le ratio aux estimateurs qui utilisent aussi cette information auxilaire.

Rappelons que le contexte de cette simulation était celui d'une distribution fortement asymétrique et d'une variable d'intérêt bien corrélée à la variable auxiliaire. On trouvera les coefficients de corrélation dans la table 3.

TABLE 3 – Corrélation entre la variable auxiliaire (chiffre d'affaires) et la variable d'intérêt (valeur ajoutée)

Strate	coefficient de corrélation
S1	0.59
S2	0.40
S3 (exhaustive)	0.30
Ensemble	0.80

Ceci explique les gains que permet l'estimateur par le ratio quelque soit le jeu de pondération

utilisé.

On voit que, si on utilise l'estimateur par le ratio, les écarts entre les différentes pondérations ont moins forts.

— l'estimateur pondéré est plus efficace - surtout lorsqu'on n'utilise pas l'estimateur par le ratio ;
— en l'absence de variable auxiliaire - ce qui empêche d'utiliser l'estimateur pondéré et l'estimateur par le ratio - le calcul des probabilités d'inclusion exactes est plus efficace que l'application classique du partage des poids.

8.2.1 Autres simulations avec des taux de sondages plus homogènes

La situation précédente se rapproche de la pratique courante des statistiques d'entreprises – des taux de sondages très différents et une strate exhaustive qui regroupe les pus grandes unités. Toutefois, nous avons également simulé des situations avec des taux de sondages plus homogènes. Cela serait le cas pour des enquêtes portant uniquement auprès de petites entreprises ou auprès d'associations non employeuses. L'intérêt de ces simulations est surtout de bien comprendre comment se comportent les différents estimateurs.

Nous avons simulé une situation où les taux de sondages sont uniforme – c'est le cas quand on fait une allocation proportionnelle. Avec toujours le même contexte d'une sélection d'unités légales qui conduisent à interroger tout le contour d'une entreprise profilée.

Le taux de sondage a été uniformément fixé à 20 % et la variable auxiliaire (le chiffre d'affaires) est supposée avoir la même dispersion dans les trois strates. Il n'y a plus de strate exhaustive.

Les paramètres sont les suivants :

TABLE 4 – Paramètres de simulations - taux de sondage uniformes

Strate	N^A	n^A	**Loi** CA_j
1	1000	200	$\mathcal{N}(100, 50)$
2	500	100	$\mathcal{N}(100, 50)$
3	100	20	$\mathcal{N}(100, 50)$

Le coefficient de corrélation entre la variable auxiliaire et la variable d'intérêt est de 0.59 dans chaque strate. Cela est dû au fait qu'il n'y a pas de grandes strates.

Nous aboutissons aux résultats suivants :

Nous voyons que dans ce cas-là, l'estimateur issu de la MGPP classique se comporte très bien. Mieux que l'utilisation des probabilité exacte et, de façon surprenante, mieux que la version pondérée de la MGPP.

Comme dans la situation précédente, une fois que l'on recourt à un estimateur par le ratio, les différences sont amoindries – et l'ordre entre les estimateurs change. Toutefois, ces différences s'exprimant dixièmes de points de pourcentage sur une seule simulation, on ne peut pas en tirer de conclusions générales.

TABLE 5 – CV (en %) des différents estimateurs de la valeur ajoutée - données simulées - taux de sondage homogènes

	MGPP classique	**MGPP pondérée**	**Probabilités exactes**
Application directe des pondérations	2.2	3.1	3.0
Estimateur par le ratio	1.8	1.9	1.6

8.3 Réflexions sur la non-réponse

On suppose ici qu'il y a m unités de la population $\mathbb{U}^A$ liées à l'unité cible b de $\mathbb{U}^B$.

La prise en compte de la non-réponse dépend de la façon dont le "lien" qui existe entre les unités affecte leur comportement de réponse.
La méthode généralisée du partage des poids propose une alternative au calcul des probabilités d'inclusion. On trouvera dans [1] plusieurs façon de prendre en compte la non-réponse. Seulement, ici, l'objectif est de déterminer des probabilités d'inclusion "exactes" et donc de prendre en compte la non-réponse suivant la façon dont elle affecte la probabilité que l'unité soit dans l'échantillon "en bout de chaîne" (après sélection et réponse (ou non) des uns et des autres.

Ce point ne peut pas être traité de façon générale sans prendre en considération la nature du lien entre les unités. Pour expliquer cela prenons deux exemples :

La même unité est présente plusieurs fois dans la base de sondage :
C'est le cas par exemple dans l'enquête Associations : on tire l'échantillon dans deux ensembles disjoint (les immatriculations auprès du ministère de l'Intérieur (Répertoire National des Associations) et auprès du ministère de l'Économie (Répertoire Sirène). Certaines associations sont présentes dans les deux bases (on leur pose la question de l'immatriculation dans l'autre base lors de l'enquête). Toutefois, quelque soit l'unité sélectionné, c'est en fin de compte la même associations qui est interrogée. Dans ce cas, les deux unités ont le même comportement de réponse.

L'unité cible est constituée de plusieurs unités de la base de sondage :
C'est le cas par exemple dans l'Enquête Sectiorielle d'Activité (ESA) : on interroge des unités légales et on s'intéresse aux entreprises constituées de la réunion de ces unités légales. L'unité-cible (l'entreprise) n'est jamais directement interrogée et le fait qu'elle "réponde" ou pas dépend du comportement de l'ensemble des unités légales.

Nous présentons ci-dessous la façon dont la non-réponse peut être traitée dans le premier cas (la même unité est présente une ou deux fois dans la base de sondage).

Dans cette situation,lorsque l'unité est présente deux fois, on peut appliquer donc le résultat (4) donné ci-dessus.

Lorsque la non-réponse est traitée par repondération, on obtient un poids final $w_i' := c_i w_i$ pour l'unité i où c_i est l'inverse de la probabilité de réponse - en pratique l'inverse de l'estimation de cette probabilité. Cette technique est exposée dans les manuels de sondages - par exemple voir [6]

Or , le raisonnement qui a conduit à la formule (4) ci-dessus ne prend pas en compte la non-réponse assimilée à une deuxième phase de tirage mais uniquement le plan de sondage. Lorsqu'en pratique, on se trouve confronté au traitement de la non-réponse, il ne s'applique pas.

On peut suivre le raisonnement suivant : dans la mesure où i et j sont en définitive la même unité, elles ont la même probabilité de réponse. On a donc $c_i = c_j$ (notons c ce coefficient commun).
La probabilité que l'unité α liée au unités i et j réponde est donc celle qu'au moins l'une des deux unités soient interrogée et réponde. On peut être tenté d'appliquer le même résultat mais avec les poids corrigés de la non-réponse.

On obtiendrait alors :

$$w'_\alpha = \frac{w'_i w'_j}{w'_i + w'_j - 1}$$

Toutefois, les deux unités i et j ne répondent pas indépendamment l'une de l'autre -puisqu'il s'agit de la même unité en réalité et la formule (4) a été établie en supposant que les deux unités étaient sélectionnées indépendamment.

Il ne faut donc pas calculer le poids de sondage de l'unité liée corrigée de la non réponse ainsi. Pour coller à la réalité de l'échantillonnage, il faut plutôt le déterminer de la façon suivante :

$$\mathbb{P}(\alpha \in r) = \mathbb{P}(\alpha \text{ réponde } | i \text{ ou } j \in s)\, \mathbb{P}(i \text{ ou } j \in s \in s)$$

Comme i et j sont la même unité (α) qui a une probabilité de répondre $\mathbb{P}(\alpha \text{ réponde } | i \text{ ou } j \in s)\, \frac{1}{c}$ (c est le coefficient de repondération "classique"), on obtient :

$$\mathbb{P}(\alpha \in r) = \frac{1}{c}\mathbb{P}(i \text{ ou } j \in s)$$

Ce qui conduit à :

$$w'_\alpha = c \cdot w_\alpha \tag{13}$$

Dans cette expression, w_α est calculé avec la formule (4) qui s'applique aux poids avant correction de la non-réponse.