LA GESTION PAR PARTAGE DES POIDS DES CHANGEMENTS DE CONTOUR DES ENTREPRISES DANS L'ENQUÊTE SECTORIELLE ANNUELLE

Arnaud FIZZALA

Insee, Direction de la méthodologie et de la coordination statistique et internationale, Département des méthodes statistiques, Division Sondages.

arnaud.fizzala@insee.fr

Mots-clés : partage des poids, entreprises, profilage

Résumé

Depuis le millésime 2016, le tirage des échantillons de l'enquête sectorielle annuelle (ESA) et de l'enquête annuelle de production (EAP), qui font partie du système d'élaboration des statistiques annuelles d'entreprises (ESANE) [1] est réalisé au niveau des entreprises profilées (EP). Lorsqu'une entreprise profilée est tirée, toutes[1] 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 peut être utilisée. La méthode généralisée de partage des poids, décrite dans l'ouvrage Indirect Sampling de Pierre Lavallée, semble toute indiquée pour gérer cette mise à jour de contours, avec la règle suivante : l'échantillon d'EP est constitué de l'ensemble des EP dont au moins une[2] UL du contour mis à jour appartient à l'échantillon initial d'UL.

De fait, les probabilités d'inclusion des EP dans l'échantillon mis à jour suivant cette règle sont complexes à déterminer. La méthode généralisée de partage des poids permet par contre d'associer aux EP de cet échantillon des poids d'estimation ayant de bonnes propriétés, i.e. permettant de construire des estimateurs sans biais.

Deux versions de la méthode généralisée de partage des poids ont été envisagées : la version classique qui est habituellement utilisée dans les enquêtes ménages et une version consistant à pondérer les liens par le chiffre d'affaires des UL. Cette deuxième version, moins habituelle, est a priori davantage adaptée aux statistiques d'entreprises, car elle permet de mieux tenir compte de « l'importance économique » des unités constituant les liens.

Une comparaison de ces deux versions de partage des poids a pu être réalisée en s'appuyant sur des variables fiscales disponibles pour l'ensemble des unités légales, sur des simulations de tirage d'échantillons d'entreprises selon le nouveau plan de sondage de l'ESA et de l'EAP, et sur les mises à jour de contours qui auraient été utilisées si des résultats 2016 avaient été produits au niveau EP.

[1] En pratique, toutes les unités légales ne sont pas forcément interrogées et on procède alors par imputation pour celles contribuant le moins au chiffre d'affaires de l'entreprise profilée.

[2] Les UL à qui on n'a pas envoyé de questionnaire mais qui se retrouvent, suite à la mise à jour, dans le contour d'une EP de l'échantillon sont traitées par imputation pour constituer une réponse au niveau de l'EP.

L'étude permet de confirmer que la version de partage des poids consistant à pondérer les liens par le chiffre d'affaires des UL aboutit aux meilleurs résultats.

Abstract

The French Structural Business Statistics (SBS) production system, Esane, has two main uses :
- the answer to the SBS European regulation ;
- the estimations of businesses contribution to GDP for the national accounts.

It is based on a mix of exhaustive administrative fiscal data and data obtained on a random sample of the population.

Esane is currently changing to produce estimates based on profiled units or enterprises and no longer on legal units. Starting at the reference year 2016, the sampling design of SBS surveys selects profiled units, but information is still collected on the cluster sample of legal units belonging to the sampled enterprises.

The paper focus on the management of the changes in profiled units definition, that is changes in the list of legal units they are made of. The Generalized weight share method, described in Indirect Sampling by Pierre Lavallée allows to deal with these changes with the following rule : a profiled unit is in the sample if at least one legal unit in the original sample belongs to it.

The aim of the study is to compare two variants of the Generalized weight share method :
- The "classical" Generalized weight share method (each legal unit "equally" contributes to the weight of the enterprise, whatever its economic characteristics);
- The Generalized weight share method with weighted links (the contribution to the weight of the enterprise depends on the economic characteristics (for example the turnover) of the legal unit).

The comparisons are based on simulations of sampling, thanks to administrative data where variables with strong correlation with the SBS variables are available to each legal unit.

The study shows that the Generalized weight share method with links weighted by turnover perfoms better than the "classical" Generalized weight share method.

1. Cadre de l'étude

Depuis le millésime 2016, le tirage des échantillons de l'enquête sectorielle annuelle (ESA) et de l'enquête annuelle de production (EAP), qui font partie du système d'élaboration des statistiques annuelles d'entreprises (ESANE) [1] est réalisé au niveau des entreprises profilées (EP) [3]. Lorsqu'une EP est tirée, toutes[3] les unités légales (UL) du sous-champ 1 (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.

Ces échantillons sont tirés en novembre N avec des contours N « initiaux » des EP issus du dispositif Lifi N-2 « définitif »[4] qui sont les plus récents disponibles à cette période. Au printemps de l'année N+1, ces contours N des EP peuvent être actualisés à partir de Lifi N-1 « définitif » – qui donne donc

[3] En pratique, toutes les UL ne sont pas forcément interrogées notamment lorsque l'EP appartient à la partie exhaustive de l'échantillon. Les détails concernant cette question sont à consulter dans la note de tirage des échantillons.
[4] À terme, l'idée est d'utiliser en novembre N, une version « provisoire » de LIFI N-1.

une image plus à jour des contours des EP –, et il est naturel de souhaiter les utiliser pour élaborer les résultats au niveau EP qui concernent l'année N[5].

La méthode généralisée de partage des poids (MGPP), décrite dans *Indirect Sampling de Pierre Lavallée* [4], semble toute indiquée pour gérer cette mise à jour de contours, 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.

De fait, les probabilités d'inclusion des EP dans l'échantillon mis à jour suivant cette règle sont complexes à déterminer. La MGPP permet par contre d'associer aux EP de cet échantillon des poids d'estimation ayant de bonnes propriétés, i.e. permettant de construire des estimateurs sans biais.

Deux versions de la méthode sont testées dans l'étude qui suit :

– MGPP avec liens classiques ;

– MGPP avec liens pondérés par le chiffre d'affaires (CA).

2. Méthodologie

Dans notre cadre, l'application de la MGPP consiste[6] à recalculer le poids des EP en utilisant la formule suivante :

$$w_i = \sum_{k \in i \cap U^A} \tilde{\theta}_{k,i} w_{ik}$$

Avec :

- w_i : le poids final de l'EP *i* ;

- w_{ik} : le poids initial de l'UL *k* rattachée (contours mis à jour) à l'EP *i*, égal à 0 pour les unités légales non échantillonnées ;

- U^A : l'ensemble des UL de la base de sondage (UL du sous-champ 1 et rattachées – contours au moment du tirage - à une EP de la base de sondage) ;

- $\tilde{\theta}_{k,i}$: pondération du lien entre l'EP *i* et l'UL *k* qui lui est rattachée.

Dans le cadre général de la MGPP, les pondérations $\tilde{\theta}_{k,i}$ peuvent prendre n'importe quelle valeur dès lors qu'elles sont toutes positives ou nulles et que, pour une EP *i* (considérée après mise à jour des contours), leur somme sur l'ensemble des unités légales *k* de la base de sondage est égale à 1.

Pour notre étude, nous avons étudié deux versions d'estimateurs :

- La version *CL* (liens classiques), qui correspond aux liens que l'on rencontre le plus souvent lorsque la méthode est appliquée aux enquêtes auprès des ménages. Pour cette version, on a $\tilde{\theta}_{k,i} = \frac{1}{M_i^{AB}}$

[5]On pourrait même envisager à terme dôutiliser la version définitive de LIFI N-1 pour définir les contours des EP de la chaîne de production dôESANE N et la version définitive de LIFI N pour la chaîne de diffusion des résultats dôESANE. Mais ce point ne sera envisagé quôultérieurement et il ne fait donc pas partie du champ dôétude de cette note.
[6]Davantage de détails sont présentés en annexe.

si l'UL k appartient à l'EP i, avec M_i^{AB} le nombre d'UL rattachées à l'EP *i* (contours actualisés) appartenant à la base de sondage (qu'elles soient dans l'échantillon d'UL initial ou non), et 0 sinon.

- La version *CA* (liens pondérés par le CA). Dans cette version on a $\tilde{\theta}_{k,i} = \frac{CA_k}{\sum_{j \in i \cap U^A} CA_j}$ avec CA_k le CA de l'UL *k*, si l'UL k appartient à l'EP i, et 0 sinon. *Remarque : si* $\sum_{j \in i \cap U^A} CA_j = 0$ *alors on utilise la formule classique. Ce cas se produit 556 fois parmi les 81 997 EP de cible 1 ou 2 dans la base de sondage mise à jour.*

L'estimation du total d'une variable Y est ensuite tout simplement :

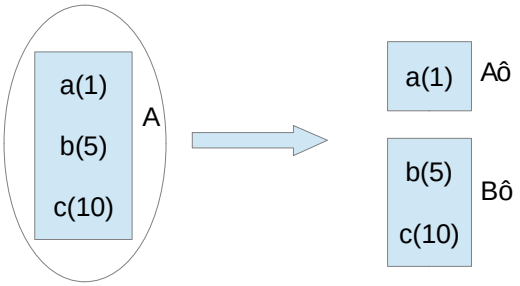
$$\hat{T}_Y = \sum_{i \in s} w_i Y_i$$

Dans nos simulations, nous comparons les estimateurs obtenus selon les deux scénarios de partage de poids avec les estimateurs issus de l'échantillon d'EP initial, sans mise à jour de contour. Comme ces derniers estimateurs ne tiennent pas compte de l'évolution du champ, des définitions des entreprises et des incidences de ces changements sur leurs caractéristiques (notamment leur classement sectoriel), ils sont a priori biaisés. Mais leur variance d'estimation peut être plus faible, du fait d'une dispersion des poids d'estimation a priori plus faible.

3. Cas pratiques

Avant de présenter l'étude et les simulations, nous illustrons la méthode à partir de quelques cas pratiques. Pour chaque exemple, on se restreint à une situation où il n'y aurait que trois unités légales (a, b, c) avec des CA respectifs de 1 M€, 5 M€ et 10 M€. Suivant les exemples, on supposera que certains liens existaient au moment du tirage entre ces trois unités légales et qu'ils sont modifiés au moment de la mise à jour. Les EP au moment du tirage sont identifiées par des lettres majuscules, et les EP au moment de la mise à jour sont identifiées par des lettres majuscules avec un prime. Les EP tirées dans l'échantillon initial sont entourées.

3.1. Division d'une EP



On suppose que l'EP A, composée des trois UL a,b,c est sélectionnée dans l'échantillon initial avec un poids $w_A = 10$.

On rappelle que le plan de sondage est assimilable à un tirage d'UL en grappes, les grappes correspondant aux contours des EP au moment du tirage. De ce fait le poids d'une UL correspond au poids de l'EP à laquelle elle est rattachée au moment du tirage.

On a donc $w_a = w_b = w_c = w_A$.

On suppose qu'au moment de la mise à jour des contours l'EP A se divise en deux EP : A' qui contient l'UL a et B' qui contient les UL b et c.

D'après la formule générale de la MGPP (voir partie *Méthodologie*), on a :

$$w_A' = \tilde{\theta}_a w_a = \tilde{\theta}_a w_A \quad \text{et} \quad w_B' = \tilde{\theta}_b w_b + \tilde{\theta}_c w_c = (\tilde{\theta}_b + \tilde{\theta}_c) w_A$$

Avec les liens classiques, on a :

$$\theta_a^{\tilde{CL}} = 1 \quad \text{et} \quad \theta_b^{\tilde{CL}} = \theta_c^{\tilde{CL}} = \frac{1}{2}$$

Donc $w_A'^{CL} = w_B'^{CL} = w_A$

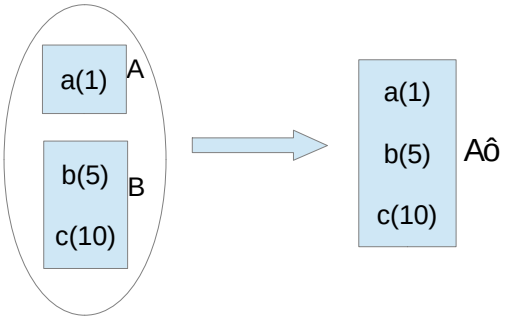
Avec les liens pondérés par le CA, on a :

$$\theta_a^{\tilde{CA}} = 1 \quad ; \quad \theta_b^{\tilde{CA}} = \frac{CA_b}{CA_b + CA_c} = \frac{5}{15} \quad ; \quad \theta_c^{\tilde{CA}} = \frac{CA_c}{CA_b + CA_c} = \frac{10}{15}$$

Donc $w_A'^{CA} = w_B'^{CA} = w_A$

Lorsqu'une EP se divise, les EP résultantes héritent du poids de tirage de l'EP initiale quelle que soit la version de la MGPP mise en oeuvre. Donc ici, A' et B' se retrouvent avec un poids de 10.

3.2. Fusion d'une indépendante échantillonnée et d'une non indépendante échantillonnée



Ici, on suppose que l'EP A qui est constituée de la seule UL a, et l'EP B constituée des UL b et c ont été tirées dans l'échantillon avec un poids respectivement de 10 et 1.

On a donc $w_a = w_A = 10$ et $w_b = w_c = w_B = 1$

On suppose qu'au moment de la mise à jour des contours l'EP A et l'EP B fusionnent en l'EP A' qui contient les UL a,b,c.

D'après la formule générale de la MGPP (voir partie méthodologie), on a :

$$w_A' = \tilde{\theta}_a w_a + \tilde{\theta}_b w_b + \tilde{\theta}_c w_c = \tilde{\theta}_a w_A + (\tilde{\theta}_b + \tilde{\theta}_c) w_B$$

Avec les liens classiques, on a : $\tilde{\theta}_a^{CL} = \tilde{\theta}_b^{CL} = \tilde{\theta}_c^{CL} = \frac{1}{3}$

Donc $w_A'^{CL} = \frac{1}{3} w_A + \frac{2}{3} w_B$

Avec les valeurs de w_A et w_B que l'on a supposées, on obtient : $w_A'^{CL} = \frac{1}{3} \times 10 + \frac{2}{3} \times 1 = 4$

Avec les liens pondérés par le CA, on a : $\tilde{\theta}_a^{CA} = \frac{CA_a}{CA_a + CA_b + CA_c} = \frac{1}{16}$;

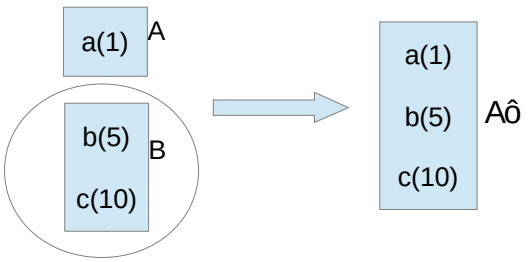
$\tilde{\theta}_b^{CA} = \frac{CA_b}{CA_a + CA_b + CA_c} = \frac{5}{16}$; $\tilde{\theta}_c^{CA} = \frac{CA_c}{CA_a + CA_b + CA_c} = \frac{10}{16}$.

Donc $w_A'^{CA} = \frac{1}{16} w_A + \frac{15}{16} w_B$

Avec les valeurs de w_A et w_B que l'on a supposées, on obtient : $w_A'^{CL} = \frac{1}{16} \times 10 + \frac{15}{16} \times 1 \simeq 1{,}6$

On voit qu'ici le partage des poids avec les liens pondérés par le CA aboutit à un poids de l'EP A' plus proche du poids de l'EP B, qui était la plus importante en termes de CA au moment du tirage.

3.3. Fusion d'une indépendante non échantillonnée et d'une non indépendante échantillonnée



Ce cas est similaire au cas ii, sauf qu'ici on suppose que A n'a pas été échantillonnée (mais appartient tout de même à la base de sondage initiale). On suppose, comme dans le cas ii, que B a un poids de 1.

On a donc $w_a = w_A = 0$ et $w_b = w_c = w_B = 1$.

Les formules de partage des poids sont les mêmes que dans le cas ii, c'est simplement la valeur de w_A qui a changé.

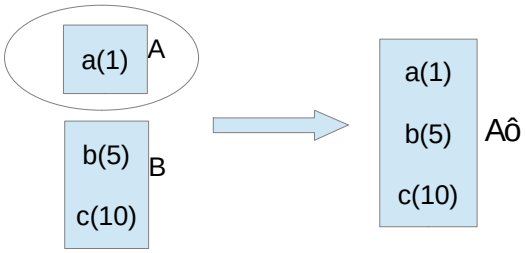
Avec les liens classiques, on a donc : $w_A{}'^{CL} = \frac{1}{3} \times 0 + \frac{2}{3} \times 1 \simeq 0{,}66$

Et avec les liens pondérés par le CA : $w_A{}'^{CA} = \frac{1}{16} \times 0 + \frac{15}{16} \times 1 \simeq 0{,}94$

On voit ici aussi que le partage des poids avec liens pondérés permet d'obtenir un poids final plus proche du poids initial de B.

On remarquera aussi que dans les deux versions, les poids finaux sont inférieurs à 1, il s'agit là d'une caractéristique du partage des poids : les poids inférieurs à 1 sont tout à fait possibles, que l'on pondère les liens par le CA ou non.

3.4. Fusion d'une indépendante échantillonnée et d'une non indépendante non échantillonnée



Ce cas est similaire au cas iii, sauf qu'ici on suppose que c'est B (et non A) qui n'a pas été échantillonnée (mais appartient tout de même à la base de sondage initiale). On suppose que A a un poids de 10.

On a donc $w_a = w_A = 10$ et $w_b = w_c = w_B = 0$

Les formules de partage des poids sont les mêmes que dans le cas ii, c'est simplement la valeur de w_B qui a changé.

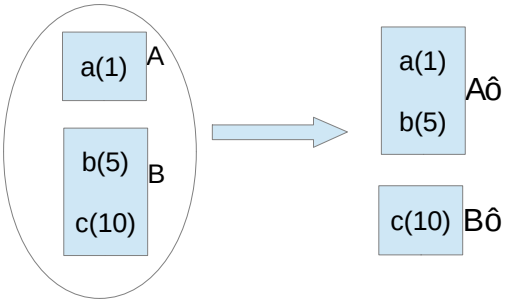
Avec les liens classiques, on a donc : $w_A{}'^{CL} = \frac{1}{3} \times 10 + \frac{2}{3} \times 0 \simeq 3{,}33$

Et avec les liens pondérés par le CA : $w_A{}'^{CA} = \frac{1}{16} \times 10 + \frac{15}{16} \times 0 \simeq 0{,}63$

On voit ici que le partage des poids avec liens pondérés par le CA est le seul qui conduit à un poids inférieur à 1. Cependant, étant donné le contexte (CA après mise à jour des contours 16 fois

supérieur au CA initial), ce poids inférieur à 1 aboutira probablement à des estimateurs plus stables que le poids supérieur à 3 obtenu avec la version classique du partage des poids.

3.5. Transfert d'UL entre EP



Dans ce dernier exemple, on suppose que l'EP A qui est constituée de la seule UL a, et l'EP B constituée des UL b et c ont été tirées dans l'échantillon avec un poids respectivement de 10 et 1.

On a donc $w_a = w_A = 10$ et $w_b = w_c = w_B = 1$

On suppose qu'au moment de la mise à jour des contours l'UL b « change » d'EP. Les EP résultantes sont donc A' qui contient les UL a et b, et B' qui contient l'UL c.

D'après la formule générale de la MGPP (voir partie méthodologie), on a :

$$w_A' = \widetilde{\theta}_a w_a + \widetilde{\theta}_b w_b = \widetilde{\theta}_a w_A + \widetilde{\theta}_b w_B$$
$$w_B' = \widetilde{\theta}_c w_c = \widetilde{\theta}_c w_B$$

Avec les liens classiques, on a :

$$\widetilde{\theta}_a^{CL} = \widetilde{\theta}_b^{CL} = \frac{1}{2} \quad \text{et} \quad \widetilde{\theta}_c^{CL} = 1$$

Donc $w_A'^{CL} = \frac{1}{2} w_A + \frac{1}{2} w_B = 5{,}5$ et $w_B'^{CL} = w_B = 1$

Avec les liens pondérés par le CA, on a :

$$\widetilde{\theta}_a^{CA} = \frac{CA_a}{CA_a + CA_b} = \frac{1}{6} \quad ; \quad \widetilde{\theta}_b^{CA} = \frac{CA_b}{CA_a + CA_b} = \frac{5}{6} \quad ; \quad \widetilde{\theta}_c^{CA} = 1$$

Donc $w_A'^{CA} = \frac{1}{6} w_A + \frac{5}{6} w_B = 2{,}5$ et $w_B'^{CA} = w_B = 1$

Cet exemple illustre à nouveau le fait que pondérer les liens par le CA permet de « privilégier » le poids de l'unité la plus importante économiquement (1/6 pour l'UL a et 5/6 pour l'UL b dans l'exemple), tandis que la version classique ne tient pas compte de l'importance économique des UL (1/2 pour l'UL a comme pour l'UL b).

Après avoir présenté quelques illustrations sur des cas pratiques des différentes versions du partage des poids, nous avons réalisé une étude basée sur des simulations pour comparer leurs performances. Les parties qui suivent présentent cette étude.

4. Données

L'étude se base sur les données suivantes :

– La base de sondage ESANE 2016 ;

– Les contours des EP utilisés pour le tirage ;

– Les contours des EP 2017 ;

- Les variables R310 (chiffre d'affaires), R003 (valeur ajoutée), B300 (total de passif au bilan), I009 (investissement corporel hors apport net d'amortissements), pour l'ensemble des unités légales.

Pour réaliser l'étude, on utilise donc des contours à jour qui ont un décalage temporel d'une année avec les contours utilisés pour le tirage, ce qui est plus que le décalage temporel entre les contours qui seront utilisés en mars pour mettre à jour l'échantillon tiré en novembre de l'année précédente en régime courant. Les raisons motivant ce choix sont les suivantes :

- Une nouvelle source issue de la Direction générale des finances publiques (DGFIP) a été intégrée pour les contours 2017, qui sont donc plus « complets » que les contours 2016 : il aurait été dommage de réaliser l'étude sur une base moins « complète » ;

- Des agrégats N à contours N+1 devront de toute façon être utilisés pour la production d'Esane N+1 à des fins de contrôle notamment : la méthode doit donc pouvoir fonctionner (aussi) dans ce cadre ;

- Si la méthode donne de bons résultats avec « beaucoup » de mouvements de contours, elle devrait aussi bien fonctionner lorsque les contours sont plus stables, ce qui sera vraisemblablement le cas lorsque les mises à jour de contours seront plus « proches » dans le temps.

L'ensemble des données listées ci-dessus permet de :

- Constituer une base d'EP dont les contours sont mis à jour qui permet notamment de calculer des totaux de référence que nous comparerons aux différentes versions d'estimateurs ;

- Mettre à jour l'échantillon d'EP selon la règle suivante : l'EP est dans l'échantillon si au moins une de ses UL est dans l'échantillon initial.

Dans le cadre de l'étude, les variables au niveau EP sont obtenues en sommant simplement[7] les valeurs des UL rattachées selon les contours à jour.

5. Simulations

Pour évaluer la qualité des différents estimateurs, nous avons réalisé 30 000 tirages d'échantillons selon le nouveau plan de sondage de l'ESA EAP et calculé pour chacun des secteurs d'activité étudiés

[7] Nous n'avons donc pas utilisé d'algorithmes de consolidation.

(niveau A10 et groupe de la NAF) les différentes versions (liens classiques, liens pondérés par le CA, poids de tirage) de l'estimateur du total des quatre variables étudiées. Ces estimateurs sont ensuite comparés au total dans la population que nous connaissons puisque l'étude se base sur des variables disponibles pour l'ensemble des unités légales[8].

On se place donc dans cette étude en absence de non-réponse. La forme composite des estimateurs Esane [2], la winsorisation et le calage ne sont pas pris en compte car cela complexifierait grandement l'étude avec impact a priori faible sur les objectifs visés : si une méthode de partage des poids est « meilleure » qu'une autre avec des estimateurs « simples », elle devrait rester « meilleure » avec des estimateurs plus « évolués ».

Dans les tableaux à suivre, nous avons reporté les valeurs des indicateurs suivants :

Le coefficient de variation :

$$Cv = \frac{\frac{1}{30000}\sqrt{\sum_{r=1}^{30000}(\hat{T}_y^m(r) - T_y)^2}}{T_y}$$

Avec :

- T_Y le vrai total de la variable *Y* ;
- $\hat{T}_Y^m(r)$: l'estimateur du total de la variable *Y* selon la méthode *m* (liens classiques, liens pondérés par le chiffre d'affaires, poids de tirage) pour la réplication *r*.

Note : On remarquera que le numérateur du Cv correspond ici à l'erreur quadratique moyenne de l'estimateur du total (et non à sa variance).

Le biais relatif :

$$Br = \frac{\frac{1}{30000}\sum_{r=1}^{30000}(\hat{T}_y^m(r) - T_y)}{T_y}$$

Le rapport entre le *CV* avec les liens pondérés par le chiffre d'affaires et le *CV* avec les liens classiques :

$$RCV = \frac{CV(\hat{T}_Y^{CA})}{CV(\hat{T}_Y^{CL})}$$

Le *RCV* permet de se donner une idée du gain (en % de *CV*) à utiliser l'estimateur avec les liens pondérés plutôt que l'estimateur avec les liens classiques.

5.1. Résultats par A10

On constate que pour les quatre variables étudiées, les résultats obtenus en pondérant les liens par le chiffre d'affaires des entreprises sont les meilleurs : les Cv sont les plus faibles et il n'y a pas de biais, ou alors ce dernier est négligeable. Les gains obtenus pour un estimateur en pondérant les liens

[8]Voir annexe pour le détail de la constitution des fichiers permettant le calcul des totaux de référence.

par le chiffre d'affaires des entreprises sont d'autant plus forts que la corrélation[9] de la variable d'intérêt avec le chiffre d'affaires de la base de sondage[10] est élevée.

Tableau 1a : Indicateurs par A10 pour la variable R310 (Chiffre d'affaires)

A10	Total de référence	CV			BR			RCV
		CL	CA	TIR	CL	CA	TIR	
AZ	1 656 871	10,6%	8,0%	5,4%	0,1%	0,0%	-0,5%	75,3%
BE	1 118 300 000	1,2%	0,1%	0,9%	0,0%	0,0%	-0,8%	10,0%
FZ	272 610 000	1,5%	0,9%	4,5%	0,0%	0,0%	-4,4%	59,5%
GI	1 567 300 000	1,7%	0,5%	0,7%	0,0%	0,0%	-0,4%	28,9%
JZ	190 400 000	3,9%	0,7%	5,0%	0,0%	0,0%	-5,0%	18,8%
LZ	61 895 925	4,6%	2,6%	3,9%	0,0%	0,0%	-2,8%	56,3%
MN	302 770 000	2,2%	0,9%	3,3%	0,0%	0,0%	-3,1%	42,3%
RU	28 722 805	4,9%	1,7%	12,6%	0,0%	0,0%	-12,5%	35,6%
TT	**3 543 700 000**	**0,9%**	**0,3%**	**1,5%**	**0,0%**	**0,0%**	**-1,5%**	**28,1%**

Tableau 1b : Indicateurs par A10 pour la variable R003 (Valeur ajoutée)

A10	Total de référence	CV			BR			RCV
		CL	CA	TIR	CL	CA	TIR	
AZ	408 548	7,9%	6,8%	5,3%	0,0%	0,0%	0,0%	85,6%
BE	287 840 000	1,0%	0,2%	1,5%	0,0%	0,0%	-1,5%	15,4%
FZ	89 571 067	1,5%	0,8%	3,1%	0,0%	0,0%	-3,0%	53,4%
GI	311 360 000	2,9%	0,4%	1,3%	0,0%	0,0%	-1,2%	12,7%
JZ	88 288 391	3,1%	0,7%	3,5%	0,0%	0,0%	-3,4%	22,3%
LZ	33 055 113	4,5%	2,1%	3,3%	0,0%	0,0%	-2,6%	46,7%
MN	151 370 000	2,4%	1,1%	2,1%	0,0%	0,0%	-1,8%	45,0%
RU	13 117 861	5,3%	1,9%	11,0%	0,0%	0,0%	-10,9%	36,0%
TT	**975 010 000**	**1,1%**	**0,2%**	**1,9%**	**0,0%**	**0,0%**	**-1,9%**	**22,0%**

Tableau 1c : Indicateurs par A10 pour la variable B300 (total de passif au bilan)

A10	Total de référence	CV			BR			RCV
		CL	CA	TIR	CL	CA	TIR	
AZ	1 587 280	22,0%	31,2%	13,3%	0,2%	0,3%	-10,5%	142,0%
BE	2 203 800 000	2,3%	0,5%	17,4%	0,0%	0,0%	-17,4%	20,7%
FZ	426 530 000	3,8%	2,2%	24,9%	0,0%	0,0%	-24,8%	56,3%
GI	1 555 900 000	6,0%	2,9%	20,5%	0,0%	0,0%	-20,4%	48,7%
JZ	557 170 000	6,8%	3,8%	34,5%	-0,1%	0,0%	-34,5%	56,3%
LZ	502 570 000	22,8%	22,4%	8,1%	0,1%	0,1%	-6,8%	98,4%
MN	1 270 800 000	6,8%	5,7%	22,5%	0,1%	0,1%	-21,4%	82,6%
RU	128 660 000	14,1%	3,2%	77,4%	0,0%	0,0%	-77,4%	22,8%
TT	**6 647 000 000**	**2,8%**	**2,2%**	**21,2%**	**0,0%**	**0,0%**	**-21,1%**	**77,8%**

[9] Les coefficients de corrélation sont de 0,93 pour le R310, 0,78 pour R003, 0,55 pour B300, 0,37 pour I009.

[10] Les liens sont pondérés par le chiffre d'affaires issu de la base de sondage, ce dernier diffère du R310 étudié qui concerne l'année 2016 contre 2015 pour la plupart des cas dans la base de sondage.

Tableau 1d : Indicateurs par A10 pour la variable I009 (investissement corporel hors apport net d'amortissements)

A10	Total de référence	CV			BR			RCV
		CL	CA	TIR	CL	CA	TIR	
AZ	19 712	77,0%	83,4%	74,1%	0,4%	0,5%	-3,3%	108,3%
BE	13 267 333	8,6%	4,6%	8,9%	0,0%	0,0%	-7,5%	53,2%
FZ	4 244 315	30,6%	28,9%	32,2%	0,0%	0,1%	-3,9%	94,6%
GI	7 706 599	28,4%	27,2%	30,3%	-0,2%	-0,1%	-3,1%	95,8%
JZ	2 718 219	35,9%	34,8%	39,0%	-0,2%	-0,1%	-17,3%	96,9%
LZ	13 264 655	20,7%	19,4%	24,7%	0,2%	0,1%	2,7%	93,9%
MN	4 768 198	70,9%	71,3%	75,9%	0,1%	0,1%	-8,8%	100,6%
RU	876 819	16,7%	9,0%	103,7%	0,0%	0,0%	-97,0%	53,8%
TT	**46 865 850**	**11,2%**	**10,7%**	**13,4%**	**0,0%**	**0,0%**	**-5,9%**	**95,6%**

5.2. Résultats par groupe[11]

Afin que les résultats soient lisibles, les groupes concernant moins de 30 EP ont été supprimés des analyses qui suivent[12]. Les indicateurs ont été calculés dans chacun des 195 groupes restants. On présente dans la suite la distribution[13] de ces indicateurs sous forme de boîtes à moustaches.

L'estimateur avec les liens pondérés par le CA est en général plus précis que les autres estimateurs : le nombre de groupes pour lesquels le *CV* de l'estimateur avec les liens pondérés par le CA est le plus faible est donné pour chaque variable dans le tableau 2.

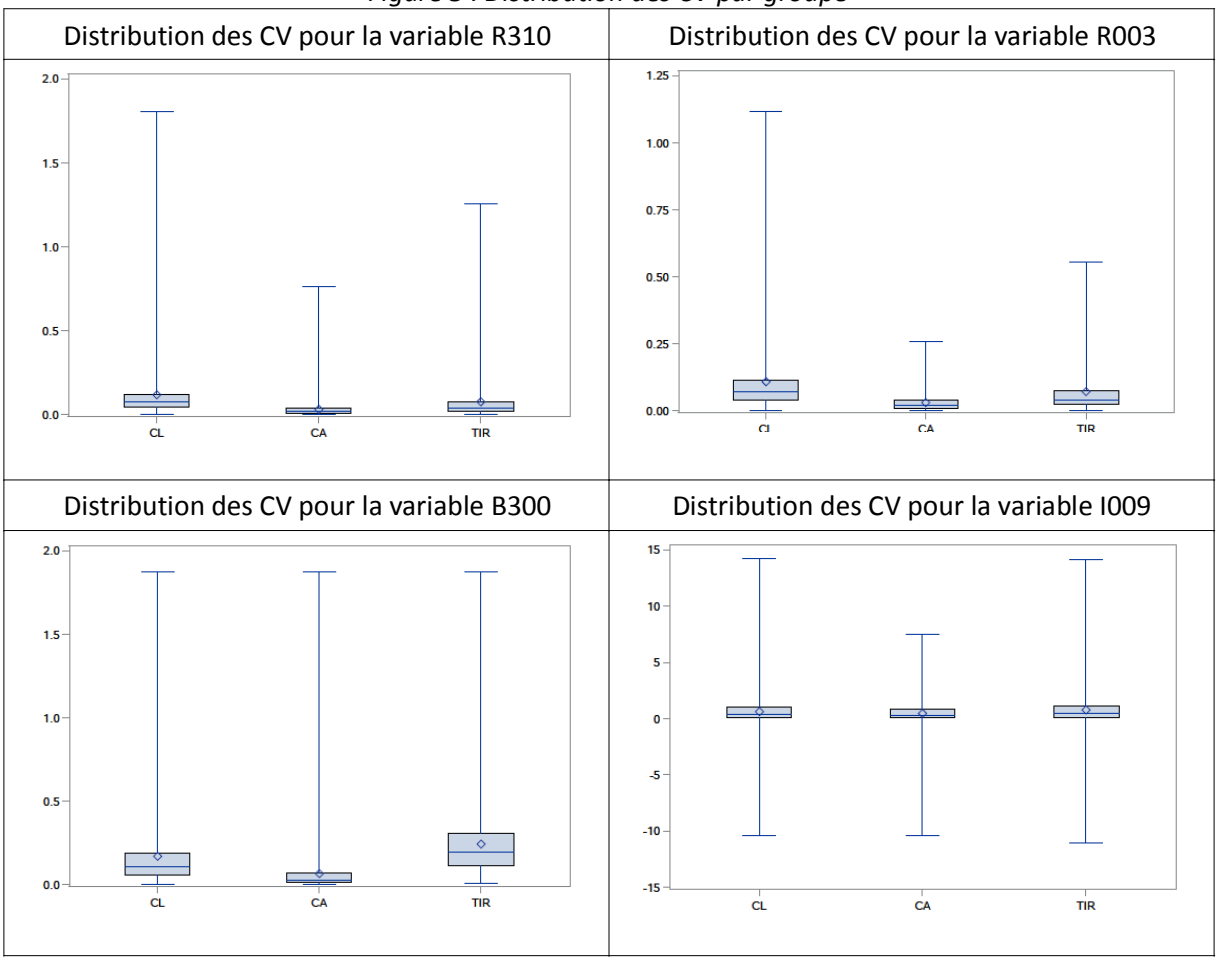
Tableau 2 : Nombre de groupes pour lesquels le CV le plus faible est obtenu pour l'estimateur avec les liens pondérés par le CA (il y a 195 groupes au total dans l'étude)

Variable	Nombre de groupes
R310	164
R003	173
B300	177
I009	155

Pour les quatre variables étudiées, les *CV* sont globalement plus concentrés et le « cœur » de la boite à moustache correspond à des valeurs plus basses pour l'estimateur « CA », c'est-à-dire avec les liens pondérés par le CA (figure 3).

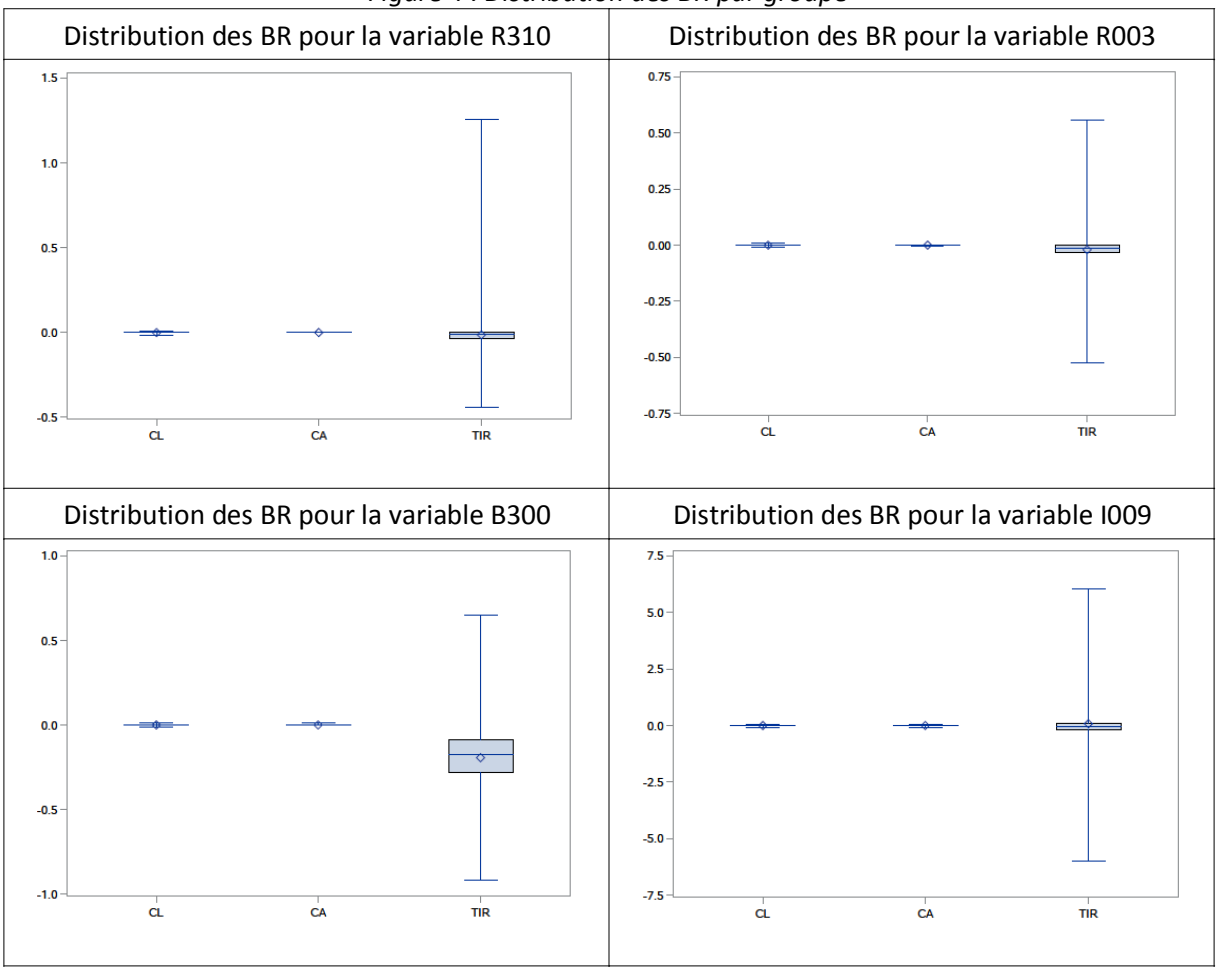
[11]Le terme groupe désigne ici le niveau de la nomenclature d'activité correspondant aux trois premières positions du code d'activité principale exercée.
[12]Il s'agit des groupes 051, 061, 062, 071, 072, 099, 120, 192, 206, 235, 253, 268, 272, 304, 491, 492, 495, 512.
[13]Calculée comme si une observation correspondait à un groupe

Figure 3 : Distribution des CV par groupe



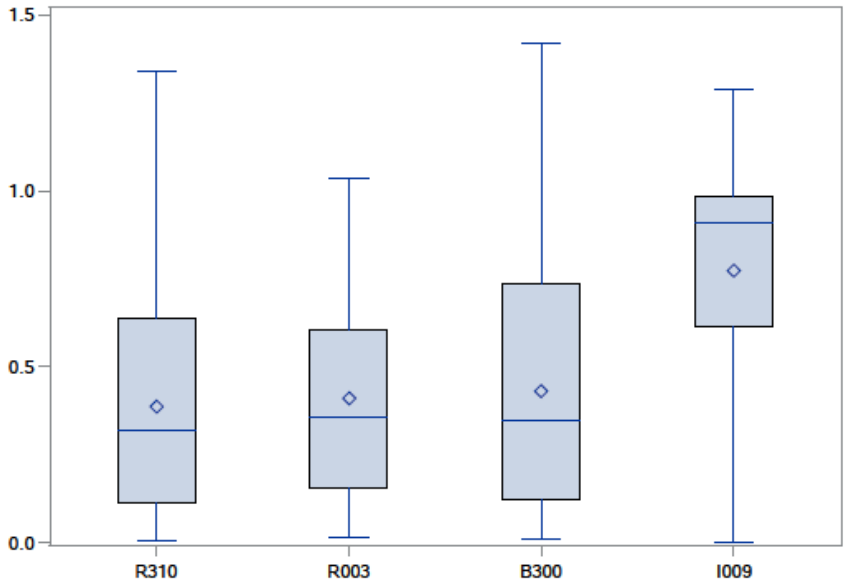
Les biais relatifs sont nuls (ou quasi-nul) pour l'ensemble des groupes étudiés avec les liens classiques ou les liens pondérés par le CA (figure 4). L'estimateur TIR, c'est-à-dire utilisant les poids de tirage sans mise à jour de l'échantillon d'entreprise est biaisé au sens où il ne tient pas compte de la mise à jour du champ obtenue par mise à jour des contours. Si le champ restait celui de la base de sondage, il ne serait pas biaisé.

Figure 4 : Distribution des BR par groupe



Le CV de l'estimateur avec les liens pondérés par le CA est inférieur au CV de l'estimateur avec les liens classiques (RCV inférieur à 1) dans la majorité des groupes (Figure 5) pour toutes les variables, mais de façon plus prononcée lorsque la variable est très corrélée au chiffre d'affaires (R310 et R003).

Figure 5 : Distribution du RCV par groupe, pour chaque variable



5.3. Distribution des poids

Dans cette partie, on s'intéresse à la distribution des poids des EP après partage des poids classique ou avec liens pondérés par le CA. La distribution des poids est propre à chaque réplication d'échantillon. Nous analysons dans la suite la distribution des poids de la réplication d'échantillon produisant à partir des poids de tirage l'estimation du R310 total médiane parmi les 1 000[14] dernières réplications d'échantillons.

Sur les 106 297 EP de l'échantillon, seules 11 107 ont des poids différents selon que le partage des poids est effectué de façon classique ou en pondérant les liens avec le CA. Globalement, les distributions de poids sont très proches comme on peut le constater dans le tableau ci-dessous (Tableau 3).

Tableau 3 : Distribution des poids de l'ensemble de l'échantillon

Version	max	P99	P95	P90	Q3	Q2	Q1	P10	P5	P1	min
Liens classiques	494,26	197,80	98,43	64,88	24,15	2,81	1,00	1,00	1,00	0,50	0,08
Liens pondérés par le CA	494,26	197,80	98,43	64,88	24,15	2,67	1,00	1,00	1,00	0,78	0,00

Les poids sont compris entre 0 et 500 et ont globalement des distributions très proches selon que le partage des poids est effectué de façon classique ou en pondérant les liens avec le CA.

Pour analyser l'impact de la méthode de partage des poids sur les plus grosses unités, nous avons isolé les EP de la partie exhaustive, c'est-à-dire dont au moins une UL a un poids de tirage égal à 1. Conformément à l'intuition que l'on pouvait avoir (cf. exemple iii de la partie Cas pratiques), on constate (Tableau 4) que les poids inférieurs à 1 sont généralement plus proches de 1 lorsque le partage des poids est réalisé en pondérant les liens par le CA. Cela explique probablement en grande partie que cette méthode produit des estimateurs plus stables et plus précis.

[14] Pour des raisons d'espace disque, les simulations sont réalisées par paquet de 1 000 échantillons effacés au fur et à mesure. Lorsque les simulations sont terminées, seuls les 1 000 derniers échantillons sont encore en mémoire.

Tableau 4 : Distribution des poids des EP de la partie exhaustive

Version	max	P99	P95	P90	Q3	Q2	Q1	P10	P5	P1	min
Liens classiques	117,67	1,00	1,00	1,00	1,00	1,00	1,00	0,94	0,67	0,50	0,08
Liens pondérés par le CA	158,43	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,68	0,00

6. Conclusion

Cette étude permet de conforter l'intuition que la méthode de partage des poids dont les liens sont pondérés par le chiffre d'affaires des unités légales est une meilleure option que le partage des poids « classique » pour gérer les changements de contour des EP dans Esane. Les estimateurs obtenus sont plus précis comme on a pu le constater lors des comparaisons des coefficients de variation, et bien que la méthode aboutisse parfois à des poids nuls[15], les estimateurs sont sans biais comme nous l'avons constaté dans nos simulations et comme cela est d'ailleurs décrit dans l'ouvrage de Pierre Lavallée faisant référence sur le sujet du partage des poids.

Cette méthode est d'ailleurs illustrée dans l'ouvrage de Pierre Lavallée par un exemple décrivant des liens entre des « établissements » et des « entreprises » et décrite comme une méthode permettant de mieux tenir compte de « l'importance économique » des unités constituant les liens, ce qui est évidemment un élément important dans la statistique d'entreprise.

Même si, par souci de simplification, les estimateurs analysés dans cette étude ne sont pas ceux directement[16] utilisés dans Esane, il n'y a pas de raison apparente pour que les conclusions soient différentes lorsque le processus complet d'estimation est pris en compte. Les étapes de winsorisation et de calage devraient d'ailleurs se dérouler plus facilement, grâce à la plus grande stabilité des estimateurs et la meilleure conservation de la corrélation entre poids des EP et importance économique permise par le partage des poids avec liens pondérés par le chiffre d'affaires.

Par ailleurs, les données utilisées pour cette étude sont particulières pour deux raisons :

- il y a une année de décalage entre les contours des EP utilisés pour le tirage et les contours mis à jour ;

- une nouvelle source a été intégrée au processus permettant de définir les contours.

Aussi, en régime courant, le choix de la méthode de partage des poids ne devrait pas avoir autant d'impact que ce qui a été vu dans cette étude.

7. Perspectives

Plusieurs points pourraient être ou seront approfondis dans des études ultérieures :

- La méthode pourrait être étudiée dans un cadre plus « général » de passage d'un échantillon d'UL (qui serait par exemple tiré sans tenir compte de la dimension EP) à un échantillon d'EP.

- Pourrait-on utiliser des méthodes similaires pour tenir compte des restructurations d'UL ?

- Comment adapter la correction de la non-réponse, la winsorisation et le calage sur marges de l'échantillon d'EP mis à jour ?

[15] Cela est le cas lorsque la ou les seules UL tirées dans l'échantillon initial ont un CA nul.

[16] On ne tient pas compte dans cette étude notamment de la partie composite des estimateurs, ainsi que du traitement de la non-réponse, de la winsorisation et du calage.

- L'impact sur les calculs de précision reste à documenter[17].

8. Bibliographie

[1] P. Brion, "*Esane, le dispositif rénové de production des statistiques structurelles d'entreprises*" Courrier des statistiques n°130, 2011 .

[2] E. Gros, "*Esane, ou les malheurs de l'estimation composite : comment gérer les valeurs négatives d'estimateurs par différence*", Actes des Journées de Méthodologie Statistique, 2012

[3] E. Gros, R. Le Gleut "The impact of profiling on sampling", presentation à l'European Establishment Statistics Workshop, 2017.

[4] P. Lavallée, *"Indirect sampling"* Springer Series in Statistics, 2007.

[17] Il y a a priori les différents éléments nécessaires dans l'ouvrage de Pierre Lavallée.

9. Annexes

9.1. Taille d'échantillons

Les tailles d'échantillon obtenues sur les 30 000 itérations varient assez peu, seules les valeurs minimales et maximales observées sont donc reportées.

Tableau 5 : Tailles d'échantillons d'EP (en nombre d'EP)

	minimum	maximum
Échantillon tiré (contours initiaux)	108 354	108 354
Échantillon avec contours mis à jour	113 237	113 485
- Dont EP dans le champ	106 019	106 612
- Dont cible 1 ou 2	27 669	28 088
- Dont non répondantes d'office	2 135	2 383

Note : les EP « non répondants d'office » sont les EP pour lesquelles les UL interrogées représentent moins de 70 % de la somme des CA de toutes les UL rattachées à l'EP. En effet, ces EP seraient forcément non répondantes puisqu'il est prévu de classer les EP répondantes si les UL ayant répondu représentent au moins 70 % de la somme des CA de toutes les UL rattachées à l'EP. Néanmoins, ces EP ont été traités comme des EP répondantes dans l'étude car la non-réponse n'a pas été prise en compte dans les simulations.

9.2. Constitution des données

a) Constitution de la base d'UL dont les contours sont mis à jour

On part de la base constituée des unités légales rattachées (contours au moment du tirage) aux EP de la base de tirage ESANE 2016.

On apparie cette base aux contours à jour des EP, ce qui permet de :

- Identifier les UL indépendantes (UL absentes de la base de contours à jour, même si cette UL appartenait à un contour au moment du tirage) et non indépendantes (UL présentes dans la base de contours à jour);
- Calculer les ϑ selon la méthode de partage des poids envisagée en tenant compte de la présence ou non de chaque UL dans la base initiale d'UL ;
- Ajouter les UL rattachées à une EP (contours à jour) et qui n'étaient pas dans la base initiale d'UL.

Cette nouvelle base d'UL sert de point de départ à la constitution de deux autres bases utiles pour l'étude :

- En se limitant aux seules UL qui étaient présentes dans la base initiale d'UL, on constitue une base d'UL recensant les ϑ et qui servira à construire les échantillons d'EP ;
- En sommant par identifiant d'EP (lorsqu'une UL est indépendante son identifiant EP correspond à son identifiant UL) les variables R310, R003, B300 et I009, on constitue une base d'EP (correspondant aux contours à jour et qui diffère donc de la base d'EP de tirage) qui servira à :
  - Calculer les totaux de référence ;
  - Répertorier les valeurs des variables au niveau EP.

b) Constitution d'un échantillon d'EP à partir de l'échantillon d'UL

On part de l'échantillon d'UL tiré avec les poids de tirage des UL[18].

On l'apparie à la base d'UL recensant les ϑ, le statut indépendant/non indépendant de l'UL et l'identifiant de l'EP à laquelle elle se rattache (voir a). Les EP de l'échantillon sont ainsi identifiées comme les EP contenant une UL de l'échantillon d'UL.

On agrège l'échantillon d'UL par identifiant d'EP : la présence des ϑ permet de calculer les poids.

On apparie l'échantillon d'EP obtenu à l'étape précédente à la base d'EP répertoriant les valeurs des variables R310, R003, B300 et I009 au niveau EP (voir a).

9.3. Développement permettant d'aboutir aux formules de la partie *Méthodologie*

On note :

U^A la population des M^A UL j de la base de sondage ;

U^B la population des M^B UL k après mise à jour des EP i ; les EP i considérées sont celles qui, après mise à jour des contours contiennent au moins une UL de la population U^A ; les unités légales de la population U^B sont donc les unités légales appartenant aux EP après mise à jour des contours ; certaines UL peuvent appartenir à U^B sans appartenir à U^A et vice-versa. On note U^{AB} l'intersection de U^A et U^B ;

$l_{j,ik}$: indicatrice de lien entre l'UL j et l'UL k appartenant à l'EP i ;

t_j : indicatrice d'appartenance à l'échantillon initial d'UL;

π_j^A : probabilité d'inclusion de l'UL j dans l'échantillon initial d'UL;

w_i le poids de l'EP i ;

w_{ik} le poids de sondage de l'UL k rattachée à l'EP i (contours actualisés) ;

M_i^B le nombre d'UL rattachées à l'EP i (contours actualisés) ;

M_i^{AB} le nombre d'UL rattachées à l'EP i (contours actualisés) appartenant à la base de sondage (qu'elles soient dans l'échantillon d'UL initial ou non).

$\tilde{\theta}_{j,ik}$ La fonction de lien pour le partage de poids pondéré et on pose :

$$\tilde{\theta}_{j,ik}=0 \text{ si } j\neq k$$

$$\tilde{\theta}_{j,ik}=\tilde{\theta}_{k,i}=\frac{CA_k}{\sum_{j\in i\cap U^A} CA_j} \text{ si j=k}$$

Pour retrouver les formules énoncées dans la partie *Méthodologie* à partir des formules exposées dans *Indirect Sampling*, on utilise des particularités de notre cadre :

i) Les liens entre les UL de U^A et de U^B sont bijectifs : il y a un lien uniquement si l'UL est la même.

ii) (qui découle de i) Une UL de U^A n'appartient qu'à une seule EP (contours actualisés) de U^B.

[18] Comme il s'agit dans le cas de l'Esane d'un tirage en grappes, le poids de l'UL est égal au poids de son EP (contours au moment du tirage).

De cette façon les formules (2.2) (2.3) et (2.4) de Indirect Sampling (pondération après partage de poids, liens non pondérés) se simplifient en la formule :

$$w_i = \sum_{k \in i \cap U^A} \tilde{\theta}_{k,i} w_{ik} \quad \text{avec ici} \quad \tilde{\theta}_{k,i} = \frac{1}{M_i^{AB}}$$

D'après (i) : $l_{i,jk}$ vaut 1 si j=k et 0 sinon. On en déduit :

$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A} = \frac{t_k}{\pi_k^A} = w_{ik}$ (2.2 : le poids « initial » de l'UL k correspond au poids de sondage de l'UL k (0 si k n'appartient pas à la base de sondage))

$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik} = 1$ si $k \in U^A$ ou 0 si $k \notin U^A$ (2.3 : nombre de liens entre l'UL k et la base de sondage)

$L_i^B = \sum_{k \in i} L_{ik}^B = \sum_{k \in i \text{ et } k \in U^A} 1 + \sum_{k \in i \text{ et } k \notin U^A} 0 = M_i^{AB}$ (nombre de liens entre l'EP i et la base de sondage)

d'où $w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B} = \frac{\sum_{k \in i} w_{ik}}{M_i^{AB}}$ (2.4 : poids final de l'EP i)

De même les étapes détaillées en page 58 de Indirect Sampling (pondération après partage de poids, liens pondérés) aboutissent à la formule :

$$w_i = \sum_{k \in i \cap U^A} \tilde{\theta}_{k,i} w_{ik} \quad \text{avec ici} \quad \tilde{\theta}_{k,i} = \frac{CA_k}{\sum_{j \in i \cap U^A} CA_j}$$

En utilisant les propriétés ci-dessus :

Si $k \in U^A$ alors $w'^{\theta}_{ik} = \sum_{j=1}^{M^A} \tilde{\theta}_{j,ik} \frac{t_j}{\pi_j^A} = \tilde{\theta}_{k,i} \frac{t_k}{\pi_k^A} = \tilde{\theta}_{k,i} w_{ik}$;

Si $k \notin U^A$ alors $w'^{\theta}_{ik} = 0$

$w_i^{\theta} = \sum_{k=1}^{M_i^B} w'^{\theta}_{ik} = \sum_{k=1}^{M_i^{AB}} w'^{\theta}_{ik} = \sum_{k \in i \cap U^A} \tilde{\theta}_{ki} w_{ik}$ (poids final de l'EP i).