Direction générale
Direction de la méthodologie et de la coordination statistique et internationale
Département « méthodes statistiques »
Division « sondages »

Note à l'attention
d'Emmanuel Gros (DSE)

Dossier Suivi par :
FIZZALA Arnaud
Tél : 0187695523
Mèl : arnaud.fizzala@insee.fr

Montrouge, le 16 mars 2020
N°2020_9486_DG75-L110

Objet : Estimation de Variance pour les estimateurs Esane en entreprises profilées

Cette note présente la méthodologie d'estimation de variance proposée par la division Sondages pour les estimateurs d'Esane en entreprises profilées (EP). Cette méthodologie s'applique aux estimateurs de variables fiscales à un niveau d'activité groupe ou plus agrégé.

La méthodologie présentée dans cette note se distingue de calculs de variances plus « habituels » par trois aspects :

- Un plan de sondage qui n'est pas aléatoire simple stratifié (cas le plus courant dans les enquêtes entreprises) ;

- La gestion d'une non-réponse après un partage des poids au niveau des « grappes » (ici les EP avec leurs nouveaux contours) ;

- La gestion de l'estimateur par différence pour les EP indépendantes : avec cet estimateur la valeur associée à une unité pour le calcul de variance peut être non nulle même lorsque cette unité n'est pas dans le domaine concerné par l'estimateur. C'est en particulier le cas pour les unités qui ont une APE obtenue via l'enquête différente de l'APE connue dans le répertoire.

Le chef de la division « sondages »

Signé : Sébastien Faivre

Pour information :
Laurent Leveillé, Claire Jacod, Claire Bidault, Olivier Haag (DSE),
Sylvain Quenum, Nicolas Paliod, Caroline Imberti, Maud Romani, Lionel Delta, Olivier Guin, Thomas Sauvaget, (DMCSI – Division Sondages),
Henri Bodet (DR 44).

Table des matières

Le plan de sondage et les traitements post-collecte

Depuis l'année de référence 2016, l'échantillon ESA/EAP est tiré au niveau des EP. Lorsqu'une EP est tirée, toutes[1] les UL qui lui sont rattachées au moment du tirage sont intégrées à l'échantillon d'UL. L'échantillon d'UL est donc un tirage aléatoire simple stratifié de grappes : les grappes sont les contours[2] des EP au moment du tirage et les strates sont construites à partir des caractéristiques (activité x taille) des EP au moment du tirage.

La probabilité d'inclusion simple d'une UL est la probabilité d'inclusion de l'EP à laquelle elle est rattachée au moment du tirage. Les probabilités d'inclusion double des UL sont calculables (fonction de l'appartenance à la même strate et/ou la même EP).

Quelques mois après l'envoi des questionnaires, de nouveaux contours d'EP sont disponibles et on utilise les réponses des UL aux questionnaires pour constituer une réponse au niveau de l'EP avec son nouveau contour. On considère qu'une EP (nouveau contour) est dans l'échantillon si au moins une UL qui lui est rattachée est dans l'échantillon d'UL initial (que cette UL soit dans les faits répondante ou non). La probabilité d'inclusion simple d'une EP (nouveau contour) correspond alors à la probabilité d'avoir sélectionné dans l'échantillon initial au moins une UL qui lui est rattachée. Lorsque les situations sont complexes avec des UL nombreuses et venant de différentes EP, le calcul de la probabilité d'inclusion simple de l'EP (nouveau contour) nous a paru hors de portée. C'est pourquoi une méthodologie spécifique[3] basée sur un partage de poids avec des liens pondérés par le chiffre d'affaires des UL a été mise en place pour pondérer les EP.

La constitution des réponses au niveau EP à partir des questionnaires des UL est réalisée par les équipes d'Esane. In fine, nous récupérons pour les traitements post-collecte l'échantillon d'EP avec les nouveaux contours et une indicatrice de réponse au niveau des EP, les variables d'intérêt constituées au niveau des EP à partir des réponses des UL sont considérées comme des déclarations[4] des EP.

Des traitements post-collecte sont appliqués à l'échantillon d'EP obtenu :

- Correction de la non-réponse par repondération dans des GRH ;
- Winsorisation sur la variable CA fiscal au niveau groupe ;
- Calage sur le CA fiscal au niveau groupe et sur le nombre d'entreprises au niveau division.

In fine, pour réaliser une estimation du total d'une variable fiscale sur un domaine (niveau d'activité groupe ou plus agrégé), on utilise :

- Un estimateur par expansion (avec les poids calés) pour les EP non indépendantes ;
- Un estimateur par différence pour les UL indépendantes.

Si le domaine est plus fin que le groupe, on utilise une clé de répartition appliquée à l'estimation au niveau groupe.

1 Cela est effectivement le cas sur la partie non-exhaustive. Sur la partie exhaustive de l'échantillon, un système de cut-off est mis en place pour n'interroger que les UL "importantes" économiquement et imputer les autres, mais nous ferons comme si toutes les UL étaient interrogées dans le cadre de cette note.

2 Le contour d'une EP désigne ici la liste des UL qui lui sont rattachées. Une UL n'est rattachée qu'à une seule EP (en pratique, nous observons quelques rares cas d'UL rattachées à plusieurs EP, on considère alors que l'UL n'est rattachée qu'à l'EP qui détient la plus grande part de son capital social).

3 Voir acte des JMS 2018 d'Arnaud Fizzala *La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle.*

4 En pratique il est possible que certaines UL rattachées à l'EP aient répondu et d'autres non. Dans le second cas des imputations sont réalisées au niveau UL pour construire les réponses au niveau EP. En toute rigueur, on pourrait essayer de tenir compte de l'aléa apporté par ces imputations dans nos calculs de précision, mais cela parait trop complexe. Aussi nous ferons comme si les données EP (lorsqu'elles sont considérées répondantes par Esane) étaient des déclarations directement réalisées par les EP.

Méthodologie d'estimation de variance

Dans cette partie nous allons décrire la méthodologie d'estimation de variance que nous proposons. Comme les probabilités d'inclusion double des UL sont calculables, et que la taille d'échantillon en UL n'est pas fixe[5], nous nous sommes orientés vers l'estimateur de variance d'Horvitz-Thompson.

La description se fera par étape en partant d'une situation simplifiée où on ne tient pas compte de :

- la non-réponse ;
- la partie exhaustive ;
- l'estimateur par différence pour les unités indépendantes ;
- la winsorisation ;
- le calage.

La prise en compte de ces différents éléments sera introduite progressivement ensuite.

a) Cas « simple »

Dans ce cas "simple", on considère une variable d'intérêt y_i au niveau des EP[6] i, et le paramètre d'intérêt est le total de cette variable sur la population U^B des EP : $Y = \sum_{i \in U^B} y_i$.

On suppose que toutes les EP ont une chance d'appartenir à l'échantillon, ce qui revient à se limiter aux EP dont au moins une UL j figurait dans la base de sondage initiale notée U^A. On s'intéresse à l'estimateur suivant[7] à partir de l'échantillon d'EP s^B :

$\hat{Y}_B = \sum_{i \in s^B} W_i y_i$ où le poids d'une EP i est $W_i = \sum_{j \in U^A} \theta_{i,j} w_j t_j$ avec :

$\theta_{i,j} = \frac{ca_j}{CA_i} l_{i,j}$ et[8] $CA_i = \sum_{j \in i} ca_j$

ca_j chiffre d'affaires de l'UL j ;

$l_{i,j}$ indicatrice de lien entre l'UL j et l'EP i ;

w_j le poids de sondage de l'UL j ;

t_j l'indicatrice d'appartenance à l'échantillon initial d'UL.

D'après Indirect Sampling (Pierre Lavallée, chapitre 4.1), cet estimateur peut s'écrire[9] comme un estimateur du total d'une variable sur l'échantillon initial d'UL correspondant, c'est-à-dire :

$\hat{Y}_B = \hat{Z}_A = \sum_{j \in s^A} w_j z_j$ Avec $z_j = \sum_{i \in U^B} \theta_{i,j} y_i$

5 Comme la taille d'échantillon en UL n'est pas fixe, nous ne pouvons pas utiliser l'estimateur de Sen-Yates-Grundy.
6 Il s'agit ici des EP après mise à jour des contours, donc des entreprises sur lesquelles sont calculés les résultats.
7 Il s'agit de pondérations EP calculées via un partage de poids avec liens pondérés par le CA des unités légales. Pour une description plus détaillée, voir acte des JMS 2018 d'Arnaud Fizzala *La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle.*
8 Remarque : on suppose que toutes les UL rattachées à une EP sont dans U^A. Si certaines UL rattachées à une EP n'appartiennent pas à U^A il faut faire, pour le calcul des pondérations, comme si elles n'existaient pas.
9 Démonstration en annexe.

Cette écriture donne l'impression que z_j n'est pas calculable car on ne dispose des y_i que sur s^B (et non sur U^B). En fait, il n'est nécessaire de connaître y_i que sur s^B.

En effet, comme une EP finale appartient à s^B seulement si au moins une de ses UL appartient à l'échantillon initial s^A, on a :

si $j \in s^A$ et $i \notin s^B$ alors $l_{i,j} = 0$

Donc pour $j \in s^A$ $z_j = \sum_{i \in U^B} \theta_{i,j} y_i = \sum_{i \in U^B} \frac{ca_j}{CA_i} l_{i,j} y_i = \sum_{i \in s^B} \frac{ca_j}{CA_i} l_{i,j} y_i = \sum_{i \in s^B} \theta_{i,j} y_i$

Sans non-réponse au niveau EP, on est en mesure de calculer les z_j pour toutes les UL de s^A, et on a l'estimateur de variance d'Horvitz-Thompson[10] (voir *Les techniques de Sondages*, Pascal Ardilly, p.139) suivant (valable quel que soit le plan de sondage des UL, dès lors qu'il n'y a pas de probabilités d'inclusion double nulles) :

$$\hat{V}_1 = \sum_{j \in s^A} \frac{1-\pi_j}{\pi_j^2} z_j^2 + \sum_{j \in s^A} \sum_{k \in s^A, k \neq j} \frac{\pi_{j,k} - \pi_j \pi_k}{\pi_{j,k} \pi_j \pi_k} z_j z_k$$

Cet estimateur peut s'écrire sous la forme matricielle suivante :

$\hat{V}_1 = Z' Q Z$ avec :

Z : vecteur colonne de dimension n^A (taille de l'échantillon initial d'UL) et de terme z_j.

Q : matrice de dimension n^A x n^A et de terme[11] si $j \neq k$ $q_{j,k} = \frac{\pi_{j,k} - \pi_j \pi_k}{\pi_{j,k} \pi_j \pi_k}$ et sur la diagonale

$$q_{j,j} = \frac{1-\pi_j}{\pi_j^2}$$

Avec un langage informatique capable d'effectuer du calcul matriciel il est donc possible, sans non-réponse au niveau EP, de calculer la variance de $\hat{Y}_B$ en :

- Constituant le vecteur Z en calculant les variables z_j pour chaque UL dans l'échantillon initial[12] d'UL ;
- Calculant les termes de la matrice Q (en fonction du plan de sondage qui a été appliqué au niveau UL) ;
- Appliquant la formule matricielle ci-dessus.

Pour calculer la variance de l'estimateur du total d'une autre variable collectée au niveau des EP, ce protocole oblige à recalculer un nouveau vecteur Z. Cependant, il est possible de « simplifier » la procédure en exprimant la formule en fonction de la variable Y directement.

10 Notre échantillon d'UL n'étant pas de taille fixe (voir *l'impact du profilage sur la refonte du plan de sondage des enquêtes sectorielles annuelles*, Ronan Le Gleut, Thomas Merly-Alpa, JMS 2018), on ne peut pas utiliser l'estimateur de Sen-Yates-Grundy.

11 Avec le plan de sondage d'Esane, les termes π_j et $\pi_{j,k}$ sont tout à fait calculables au niveau des UL (sondage aléatoire simple stratifié de grappes). Le calcul est détaillé plus tard.

12 Il s'agit bien de l'échantillon d'UL issue du tirage, les UL éventuellement « ajoutées » lors de la mise à jour des contours ne sont pas à considérer dans s^A.

Soit la matrice L de dimension n^A x n^B de terme[13] $L_{j,i}=l_{j,i}\theta_{j,i}$ alors on a $Z=LY$ Avec Y le vecteur colonne de dimension n^B ayant pour terme y_i.

D'où :

$$\hat{V}_1=Y'L'QLY=Y'Q_2Y \quad \text{avec} \quad Q_2=L'QL$$

Q_2 est alors une matrice de dimension n^B x n^B et, une fois que cette matrice a été calculée (sa valeur ne dépend pas de la variable y), le calcul de variance peut se faire « directement » pour toute variable y_i collectée au niveau des EP avec la formule matricielle :

$$\hat{V}_1=Y'Q_2Y$$

Un autre avantage de cette écriture est qu'elle correspond à une double somme sur les EP (et plus sur les UL). Aussi, la prise en compte de la non-réponse, qui est observée au niveau EP, sera plus facile avec cette écriture.

Remarque : dans le cas d'Esane, et sans non-réponse au niveau EP, on pourrait "facilement" estimer la variance de $\hat{Y}_B$ *en appliquant la formule de variance « non matricielle » associée à l'estimateur Horvitz-Thompson du total de la variable* z_j *avec un tirage aléatoire simple stratifié de grappes. Mais quand la non-réponse apparaît, cela devient moins « facile », et l'écriture matricielle facilite les choses…*

b) prise en compte de la non-réponse des EP

On intègre à présent de la non-réponse au niveau des EP finales.

On modélise la non-réponse comme une phase de tirage supplémentaire : tirage de Poisson des EP répondantes r^B parmi s^B avec la probabilité π_i^B . On suppose que π_i^B est connue[14].

On est donc dans une configuration de tirage en deux phases :

Phase 1 : tirage de l'échantillon d'UL s^A (sondage stratifié de grappes d'UL : les grappes étant les EP avec les contours au moment du tirage).

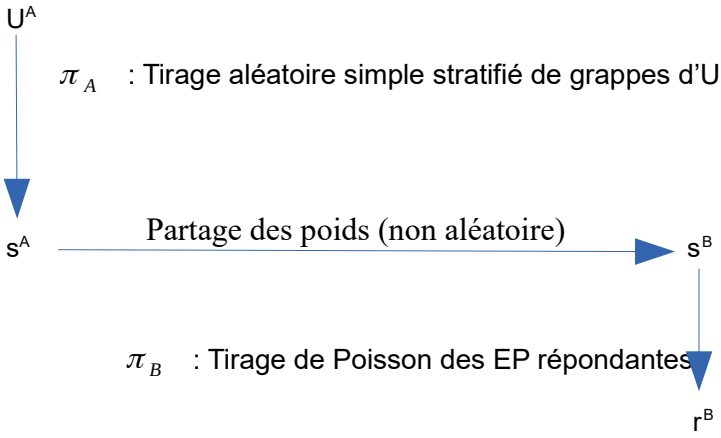
Phase 1bis : obtention de s^B par application de la méthode de partage des poids : cette phase est déterministe car la connaissance des liens n'est pas liée à l'enquête. Aussi connaître s^A implique de connaître s^B (la réciproque n'est pas vraie car plusieurs s^A peuvent conduire au même s^B).

Phase 2 : « Tirage » des EP répondantes r^B via un tirage de Poisson.

Cette configuration peut se schématiser de la façon suivante :

13 On préfère décrire les termes avec les indices j,i (dans cet ordre) pour garder à l'esprit que les lignes de la matrice L représentent les UL échantillonnées et les colonnes les EP (nouveaux contours) échantillonnées.

14 Dans le chapitre 8.5 de Indirect Sampling, il est tenu compte du fait que les probabilités de réponse au niveau EP sont estimées, mais cela complique beaucoup les calculs… Nous avons donc opté pour une version plus simple, et qui correspond à ce qui est fait habituellement à l'Insee.



Dans cette configuration, la variable et le paramètre d'intérêt restent les mêmes, mais l'estimateur auquel on s'intéresse change car il doit s'exprimer à partir des seules EP répondantes. Cet estimateur est :

$\widetilde{Y}_B = \sum_{i \in r^B} W_i \frac{y_i}{\pi_i^B}$ avec : π_i^B : la probabilité de réponse de l'EP i.

$\widetilde{Y}_B$ est un estimateur sans biais de Y (démonstration en annexe).

On cherche à présent à estimer sa variance :

$$V_{\pi_A \pi_B}(\widetilde{Y}_B) = \underbrace{V_{\pi_A}[E_{\pi_B}(\widetilde{Y}_B|s^A)]}_{V_1} + \underbrace{E_{\pi_A}[V_{\pi_B}(\widetilde{Y}_B|s^A)]}_{V_2}$$

b.1) Estimation de V_2

Comme s^B est déterminé par s^A, On a : $V_{\pi_B}(\widetilde{Y}_B|s^A) = V_{\pi_B}(\widetilde{Y}_B|s^A, s^B) = V_{\pi_B}(\widetilde{Y}_B|s^B)$

Comme on modélise la non réponse des EP par un tirage Poissonien de probabilité π_i^B (que l'on suppose connue), on a : $V_{\pi_B}(\widetilde{Y}_B|s^B) = \sum_{i \in s^B} \frac{1-\pi_i^B}{\pi_i^B} W_i^2 y_i^2$

$\sum_{i \in s^B} \frac{1-\pi_i^B}{\pi_i^B} W_i^2 y_i^2$ n'est pas calculable car on ne dispose des y_i que sur la population des EP répondantes. Cependant, à partir de r^B, on dispose de l'estimateur d'Horvitz-Thompson (sans biais) de $\sum_{i \in s^B} \frac{1-\pi_i^B}{\pi_i^B} W_i^2 y_i^2$ suivant :

$$\widetilde{V}_2 = \sum_{i \in r^B} \frac{1-\pi_i^B}{(\pi_i^B)^2} W_i^2 y_i^2$$

$\widetilde{V}_2$ estime sans biais $V_{\pi_B}(\widetilde{Y}_B|s^A)$ (voir *Les techniques de Sondage*, Pascal Ardilly, p.162).

C'est donc aussi un estimateur sans biais de $E_{\pi_A}[V_{\pi_B}(\widetilde{Y}_B|s^A)]$, donc de V_2.

b.2) Estimation de $V_1=V_{\pi_A}[E_{\pi_B}(\widetilde{Y}_B|s^A)]$

On a : $E_{\pi_B}(\widetilde{Y}_B|s^A)=\hat{Y}_B=\sum_{i\in s^B} W_i y_i$ (Voir annexe)

On se retrouve alors dans un cadre ressemblant à la partie précédente, et finalement $V_1=V_{\pi_A}[\hat{Y}_B]$. On pense alors à l'estimateur proposé dans la partie précédente :

$\hat{V}_1=Y'Q_2Y$

Problème : Cet estimateur s'exprime sur s^B et non sur r^B.

$\hat{V}_1=Y'Q_2Y$ correspond en fait à une double somme sur l'échantillon d'EP. On peut donc réaliser une estimation Horvitz-Thompson de cette double somme en réalisant la double somme sur les EP répondantes et en divisant chaque terme par la probabilité d'inclusion double du terme : $\pi^B_{i,t}$

Comme on modélise la non-réponse des EP comme un tirage de Poisson de probabilité π^B_i , on a :

si $t\neq i$ $\pi^B_{i,t}=\pi^B_i\pi^B_t$ et si $t=i$ $\pi^B_{i,t}=\pi^B_i$

On obtient alors $\widetilde{V}_1=Y_r'Q_3Y_r$ avec :

- Y_r : le vecteur colonne avec les y_i des répondants ;

- Q_3 : une matrice de taille n_r^B x n_r^B se limitant aux EP répondantes et de terme

$$Q_{3i,t}=\frac{Q_{2i,t}}{\pi^B_{i,t}}$$

Finalement, on est en mesure de calculer $\widetilde{V}_1=Y_r'Q_3Y_r$.

Et notre estimation de variance sera $\widetilde{V}=\widetilde{V}_1+\widetilde{V}_2$

c) Prise en compte de la partie exhaustive

La partie exhaustive de l'échantillon correspond à deux ensembles :
- Dans U^A, il s'agit des UL avec w_j=1. On le note E^A.

- Dans U^B, il s'agit des EP dont au moins une UL appartient à E^A. On le note E^B.

On considère qu'il n'y a pas[15] de non-réponse dans E^B.

15 S'il y a de la non-réponse en pratique, on la traite par imputation et on fait comme s'il s'agissait d'une vraie réponse dans nos calculs de précision.

On montre ci-dessous que le raisonnement des parties précédentes fonctionne en se limitant aux univers A et B sans l'exhaustif.

En tenant compte de l'exhaustif, notre paramètre d'intérêt s'écrit :

$$Y = \sum_{i \in U^B - E^B} y_i + \sum_{i \in E^B} y_i$$

Et l'estimateur devient :

$$\hat{Y}_B = \sum_{i \in s^B - E^B} W_i y_i + \sum_{i \in E^B} y_i$$

(*Attention dans les parties précédentes il était possible d'avoir* $i \in E^B$ et $W_i \neq 1$ *, il ne s'agit donc plus du même estimateur ! En pratique, nous utilisons bien l'estimateur ci-dessus : le poids d'une EP est 1 dès lors qu'une UL a un poids initial de 1*).

La partie exhaustive $\sum_{i \in E^B} y_i$ ne génère pas de variance et on a : $V(\hat{Y}_B) = V\left(\sum_{i \in s^B - E^B} W_i y_i\right)$

Avec $W_i = \sum_{j \in U^A - E^A} \theta_{i,j} w_j t_j$ car si $j \in E^A$ et $l_{i,j} = 1$ alors $i \in E^B$ (par définition de E^B).

Cela implique que seules les UL hors de l'exhaustif interviennent dans le calcul des poids des EP non exhaustives.

De plus $\hat{Y}_{B-E^B} = \sum_{i \in s^B - E^B} W_i y_i = \sum_{i \in s^B - E^B} y_i \sum_{j \in s^A - E^A} \theta_{i,j} w_j = \sum_{j \in s^A - E^A} w_j \sum_{i \in s^B - E^B} \theta_{i,j} y_i = \hat{Z}_{A-E^A}$

L'estimateur du total de Y sur la partie non exhaustive coïncide avec l'estimateur d'Horvitz-Thompson du total des z_j sur la partie non exhaustive.

Ainsi, on peut exclure les parties exhaustives de U^A et U^B et appliquer les raisonnements des parties précédentes en considérant que les univers U^A et U^B sont en fait les parties non exhaustives.

Dans la suite, on considère que les univers U^A et U^B sont en fait les parties non exhaustives.

d) Distinction entre les indépendantes et les non indépendantes

Une EP indépendante[16] correspond à une EP rattachée à une seule UL.

Soient i une EP indépendante et j l'UL correspondante, on a : $y_i = y_j$ et $W_i = w_j$

On note I^B l'ensemble des EP indépendantes et I^A l'ensemble des UL correspondantes.

L'ensemble des EP non indépendantes est noté[17] I_C^B et I_C^A l'ensemble des UL correspondantes.

16 Habituellement, on parle plutôt d'UL indépendante, mais ici pour bien distinguer U^A et U^B, on préfère la terminologie EP indépendante.

17 *C* comme complémentaire

Notre paramètre d'intérêt peut s'écrire : $Y=\underbrace{\sum_{i\in I^B} y_i}_{Y_I}+\underbrace{\sum_{i\in I_C^B} y_i}_{Y_{I_C}}$

Il est envisagé d'utiliser des estimateurs différents pour les EP indépendantes et les EP non indépendantes.

Pour les EP non-indépendantes, on utiliserait les estimateurs vus précédemment :
$\hat{Y}_{I_C}=\sum_{i\in s_{I_C}^B} W_i y_i$

Avec : $W_i=\sum_{j\in I_C^A} \theta_{i,j} w_j t_j$ car si $j\in I^A$ et $l_{i,j}=1$ alors $i\in I^B$

On remarque que : $\hat{Y}_{I_C}=\sum_{i\in s_{I_C}^B} W_i y_i=\sum_{i\in s_{I_C}^B} y_i \sum_{j\in s_{I_C}^A} \theta_{i,j} w_j=\sum_{j\in s_{I_C}^A} w_j \sum_{i\in s_{I_C}^B} y_i \theta_{i,j}=\sum_{j\in s_{I_C}^A} w_j z_j=\hat{Z}_{I_C}^A$

Pour les indépendantes, on utiliserait l'estimateur par différence qui est utilisé au niveau UL :

$\hat{Y}_I=\sum_{i\in U_I^B} Y_i+\sum_{i\in s_I^B} W_i[y_i-Y_i]=\sum_{j\in U_I^A} Y_j+\sum_{j\in s_I^A} w_j[y_j-Y_j]$ avec Y_i la valeur de la variable dans les fichiers fiscaux[18] et comme précédemment y_i la valeur relevée lors de l'enquête.

Ainsi, on a : $\hat{Y}=\hat{Y}_{I_C}+\hat{Y}_I=\hat{Y}_{I_C}+\sum_{i\in U_I^B} Y_i+\sum_{i\in s_I^B} W_i[y_i-Y_i]$

Comme $\sum_{i\in U_I^B} Y_i$ est une constante, on a :

$V(\hat{Y})=V(\hat{Y}_{I_C}+\sum_{i\in s_I^B} W_i[y_i-Y_i])$

Soit $x_i=y_i 1_{i\in I_C}+(y_i-Y_i)1_{i\in I}$

Alors $V(\hat{Y})=V(\sum_{i\in r^B} W_i x_i)$

On peut alors reprendre les raisonnements précédents en remplaçant Y par X.

e) Winsorisation

Winsoriser une unité i revient à multiplier son poids par un coefficient de winsorisation : $c_i^w=\frac{X_i^w}{X_i}$ avec X_i la variable de winsorisation (le chiffre d'affaires fiscal dans le cas d'Esane).

Aussi, avec la convention $c_i^w=1$ lorsque i n'est pas winsorisée, on a : $W_i^w=c_i^w W_i$.

Dans Esane, la winsorisation intervient après la correction de la non-réponse (voir note N°2019_2544_DG75-L110).

18 Rappelons ici que Y_i ne tient compte d'aucune information issue de l'enquête, on pense en particulier ici à l'APE et à l'appartenance au champ. Il est donc possible que Y_i soit non nul bien que l'enquête révèle que i n'est pas dans le champ où a finalement une APE hors du domaine étudié.

Après winsorisation, l'estimateur est : $\widetilde{Y}_B^w = \sum_{i \in s^B} W_i^w \frac{y_i r_i^B}{\pi_i^B} = \sum_{i \in s^B} W_i \frac{c_i^w y_i r_i^B}{\pi_i^B}$. On se retrouve alors dans le cas des parties précédentes mais avec la variable $c_i^w y_i$ à la place de y_i.

Remarque : Ce traitement permet de tenir compte de la winsorisation dans le calcul de la variance, mais ne tient pas compte du biais introduit...

f) Calage sur marges

Un calage est réalisé au niveau EP, séparément entre EP indépendantes et EP non indépendantes. Les poids en entrée du calage sont les poids après winsorisation.

L'estimateur devient alors : $\widetilde{Y}_B^c = \sum_{i \in s^B} W_i^c \frac{y_i r_i^B}{\pi_i^B}$

On a alors (*Les techniques de Sondage, Pascal Ardilly, p. 360)* : $V(\widetilde{Y}_B^c) = V(\widetilde{\epsilon}_B^w)$ où ϵ_i correspond au résidu de la régression linéaire pondérée par les poids w_i^w de y_i sur les variables de calage.

On peut donc se ramener au cas des parties précédentes en utilisant ϵ_i à la place de y_i.

Implémentation informatique, étape par étape

Pour faire les calculs de précisions, on peut finalement distinguer deux types d'étapes pour prendre en compte les traitements :

* Des modifications des matrices Q (correspondant à la prise en compte de la non-réponse et du partage des poids)

* Des modifications de la variable d'intérêt Y (correspondant à la prise en compte des indépendantes, de la winsorisation et du calage).

Aussi, on peut voir le calcul en trois grandes étapes :

1 - Constitution des matrices de « base » Q, L, Π^B (matrice de terme $1/\pi_{i,t}^B$).

2 - Calcul d'une matrice Q_{3r} qui pourra être utilisée pour l'échantillon d'EP répondantes obtenues et pour différentes variables Y : Utilisée « seule » la matrice Q_{3r} aboutira aux variances obtenues avec un estimateur « classique », c'est-à-dire sans winsorisation, sans calage et sans estimateur par différence pour les indépendantes.

3 - Calcul d'une matrice X_r (Modification de la variable Y_r) : afin de tenir compte de la winsorisation, du calage et de l'estimateur par différence pour les indépendantes.

Lors des 3 étapes, **on se limite aux UL et EP non-exhaustives.**

Étape 1 : Constitution des matrices de « base »

Calcul de Q

On a :

si $j \neq k$ $q_{j,k} = \frac{\pi_{j,k} - \pi_j \pi_k}{\pi_{j,k} \pi_j \pi_k}$ et sinon $q_{j,j} = \frac{1 - \pi_j}{\pi_j^2}$

Le calcul des probabilités d'inclusion simple π_j ne pose pas de problème : il s'agit du taux de sondage appliqué à l'EP à laquelle appartenait l'UL au moment du tirage : $\pi_j = \frac{m_h}{M_h}$ avec m_h le nombre d'EP (contours au moment du tirage) tirée et M_h le nombre d'EP (contours au moment du tirage) dans la strate h de la BdS.

Il faut distinguer 4 sous-cas pour évaluer les probabilités d'inclusion double $\pi_{j,k}$:

a) $j = k$ (il s'agit de la même UL) alors ; $\pi_{jk} = \pi_j = \frac{m_h}{M_h}$ et $q_{j,j} = \frac{1 - \frac{m_h}{M_h}}{\left(\frac{m_h}{M_h}\right)^2}$

b) $j \neq k \; et \; j,k \in E$ (les UL j et k appartiennent à la même EP *E* au moment du tirage, et tirée dans la strate h) alors ; $\pi_{jk} = \pi_E = \frac{m_h}{M_h}$ et $q_{j,k} = \frac{1 - \frac{m_h}{M_h}}{\left(\frac{m_h}{M_h}\right)^2}$

c) $j \neq k \; et \; j \notin E_k et \; E_j \in h$ (les UL j et k appartiennent à deux EP (au moment du tirage) différentes, mais tirées dans la même strate h) alors $\pi_{jk} = \frac{m_h}{M_h} \frac{(m_h - 1)}{(M_h - 1)}$ et

$$q_{j,k} = \frac{\frac{m_h - 1}{M_h - 1} - \frac{m_h}{M_h}}{\left(\frac{m_h}{M_h}\right)^2 \frac{m_h - 1}{M_h - 1}}$$

d) $j \neq k \; et \; j \notin E_k et \; E_j \notin h$ (les UL j et k appartiennent à deux EP (au moment du tirage) différentes, et tirées dans deux strates différentes) alors $\pi_{j,k} = \pi_j \pi_k$ et $q_{j,k} = 0$

En pratique Q est une matrice de $(n^{A}_{r})^{219}$ éléments et peut s'avérer trop volumineuse en deux sens :

- Il est impossible de calculer la matrice terme à terme avec une double boucle sur les UL[20] ;
- Il est peu envisageable de stocker une telle matrice (plusieurs Go).

Il est néanmoins possible de calculer cette matrice en s'aidant de deux constats :

a) De nombreuses valeurs sont égales à 0 (dès lors que deux UL ne sont pas dans la même strate). Il est alors possible d'utiliser des matrices dites Sparse. Ce type de matrice est stockée en retenant uniquement la dimension de la matrice, les valeurs non nulles et leurs positions. La taille de l'objet est beaucoup moins grande (environ 70 Mo pour l'ESA 2017) et le problème de stockage mentionné plus haut est alors résolu.

b) Si on trie les UL par strate puis identifiant d'EP, alors on s'aperçoit que Q est une matrice diagonale par blocs, les blocs étant constitués des UL appartenant aux mêmes strates. De plus, dans un bloc donné il n'y a que deux possibilités :

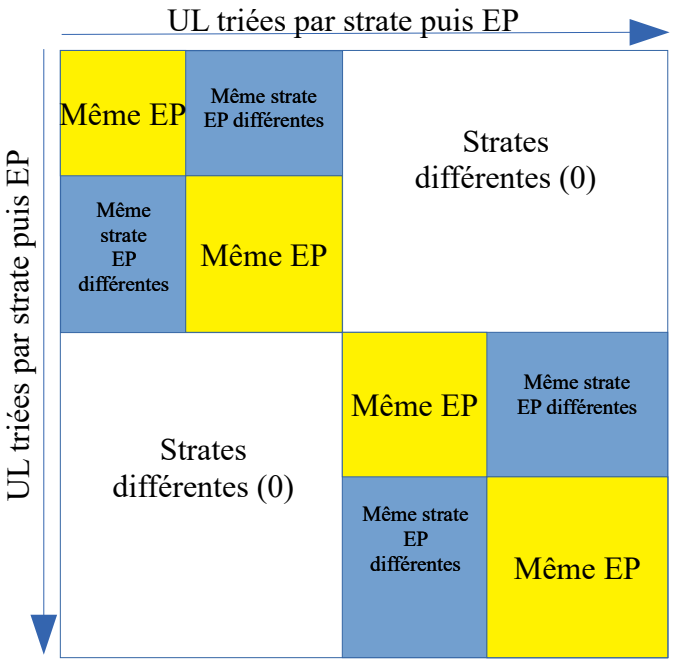
- j et k sont dans la même EP (y compris j=k) : on a alors une valeur q_{jk} associée à l'EP ;
- j et k ne sont pas dans la même EP : on a alors une valeur associée à la strate.

Grâce à la fonction R *bdiag*, il est assez rapide de calculer la matrice Q :

- On constitue une matrice Sparse diagonale par blocs S où un bloc est une strate (et chaque bloc est rempli avec la valeur q_{jk} correspondant à j et k dans la même strate mais pas dans la même EP : cas *c* ci-dessus) ;
- On constitue une matrice Sparse diagonale par blocs E où un bloc est une EP (et chaque bloc est rempli avec la valeur q_{jk} correspondant à j et k dans la même EP : cas *a* et *b* ci dessus) ;

Q est alors la matrice E, complétée, uniquement pour les termes où $E_{j,k}=0$, par S.

Schéma de la matrice Q



19 n^{a}_{r} correspond au nombre d'UL dans s^{A} (l'échantillon initial d'UL, sans les UL éventuellement « ajoutées » au moment du partage des poids) et rattachées à une EP finale répondante.

20 Il n'est donc pas possible d'aller voir pour chaque paire (j,k), si les UL sont dans le cas a, b ,c ou d mentionné précédemment et associer le bon terme.

Remarque : Dans le plan de sondage actuel, on tire une seule unité dans de nombreuses strates (674 strates sur 2 772 pour Esane 2017). Le calcul de $q_{j,k}$ dans le cas c pose problème pour ces strates. Pour le moment on a effectué des regroupements de strates pour contourner le problème[21].

Calcul de L

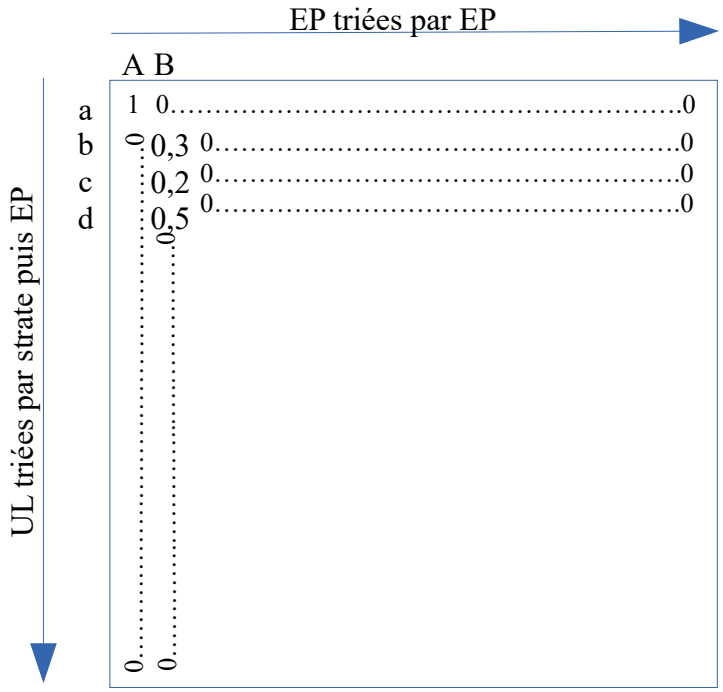
On utilise la formule : $L_{j,i}=l_{j,i}\theta_{j,i}$

En pratique, on part d'une table avec trois variables : identifiant de l'UL, identifiant de l'EP finale, valeur de $\theta_{j,i}$ associée. Grâce à la fonction block_matrix (fonction codée et utilisée du coté de la section ménages de la division Sondages qui décompose à partir d'une « clé » une matrice en une matrice par blocs), il est facile et immédiat de transformer la table en matrice L (c'est en fait une question de mise en forme).

Afin de tenir compte de la non-réponse, et pour que L soit compatible avec Q on se limite aux mêmes n^A_r UL que Q et aux r^B EP finales répondantes.

Là encore, on utilise des matrices Sparse pour économiser de la place : la plupart des $L_{j,i}$ sont nuls.

Schéma de la matrice L



Note : Dans cet exemple, on suppose que l'EP A est indépendante, composée donc uniquement de l'UL a, avec $\theta_{a,A}=1$, et que l'EP B est composée des UL b,c et d avec $\theta_{b,B}=0{,}3$, $\theta_{c,B}=0{,}2$, $\theta_{d,B}=0{,}5$.

Calcul de Π^B

21 Une méthode plus rigoureuse serait d'utiliser des techniques de collapse, mais nous avons jugé cela trop complexe pour une première implémentation. A étudier pour de futures versions.

Pour toute EP, on estime la probabilité de réponse dans chaque GRH *g* par $\hat{\pi}_i^B = \frac{\sum_{i \in g} W_i r_i}{\sum_{i \in g} W_i}$,

puis on utilise la formule (en considérant que $\pi_i^B = \hat{\pi}_i^B$) :

si $t \neq i$ $\pi_{i,t}^B = \pi_i^B \pi_t^B$ et si $t = i$ $\pi_{i,t}^B = \pi_i^B$

En pratique, la matrice est trop volumineuse (plusieurs Go) pour effectuer un calcul terme par terme. De plus, ce n'est pas une matrice sparse (il n'y a a priori pas de 0).

Aussi, la matrice Π^B n'est pas « calculée », on utilise à la place des astuces permettant de ne manipuler que des objets Sparse (voir étape 2).

Étape 2 : calcul de Q_{3r}

On calcule Q_{2r} : $Q_{2r} = L_r' Q_r L_r$ (produit matriciel)

Puis Q_{3r} : $Q_{3r} = Q_{2r} \cdot \Pi_r^B$ (produit terme à terme)

En pratique, Π^B est trop volumineuse pour être calculée, on remplace donc l'opération $Q_{2r} \cdot \Pi_r^B$ par la suite d'opérations suivantes, utilisant le vecteur π_r^B de terme $\frac{1}{\pi_i^B}$:

Opération 1 (multiplication des lignes de Q_{2r} par $\frac{1}{\pi_i^B}$) :

$Q_{2r1} = Q_{2r} . \pi_r^B$

*Remarque : Dans R la multiplication standard (signe *) entre une matrice et un vecteur donne une matrice où pour chaque ligne, les éléments ont été multipliés le terme correspondant dans le vecteur.*

Opération 2 (multiplication des colonnes de Q_{2r1} par $\frac{1}{\pi_i^B}$) :

$Q_{3r} = t(\pi_r^B . t(Q_{2r1}))$ (avec t() l'opérateur effectuant la transposition)

Opération 3 (correction des éléments de la diagonale) :

A la suite de l'opération 2, on a sur la diagonale $q_{3r}[i,i] = \frac{q_{2r}[i,i]}{(\pi_i^B)^2}$, ce qui n'est pas correct. On corrige les éléments de la diagonale pour obtenir $q_{3r}[i,i] = \frac{q_{2r}[i,i]}{\pi_i^B}$

Q_{3r} est ensuite stockée et on pourra l'utiliser pour plusieurs variables pour lesquelles on souhaiterait calculer une précision.

Étape 3 : Transformation de Y_r

La succession d'étapes est décrite comme un programme informatique avec une variable y qui évolue à chaque étape.

Y_r doit au préalable être trié par EP (comme Q_{3r}).

0 – Prise en compte du domaine de diffusion de l'estimateur

On note d l'ensemble des EP appartenant au domaine de diffusion de l'estimateur (par exemple un groupe (APE sur 3 positions) donné).

On effectue l'opération suivante :

$$y_i \leftarrow y_i 1_{i \in d}$$

Cela revient à mettre à 0 les y_i des EP qui ne sont pas dans le domaine *d* d'après l'enquête.

Remarque : il s'agit bien d'une mise à 0 et pas une suppression des observations qui ne sont pas dans le domaine d d'après l'enquête, sinon l'étape suivante ne fonctionnera pas.

1 – Prise en compte des indépendantes

On effectue l'opération suivante :

$$si\ i \in I \text{ alors } y_i \leftarrow y_i - Y_i$$

Remarque : il est important d'effectuer cette opération pour l'ensemble des EP répondantes, en particulier il ne faut pas oublier celles qui ne sont pas dans le domaine d d'après l'enquête ($y_i = 0$) mais qui le sont d'après les sources administratives ($Y_i \neq 0$).

2 – Prise en compte du calage

On effectue une régression pondérée par les poids après winsorisation de y_i (il s'agit ici du y_i en sortie de l'étape précédente) sur les variables de calage.
On récupère les résidus ϵ_i de cette régression et on effectue l'opération suivante :

$$y_i \leftarrow \epsilon_i$$

3 – Prise en compte de la winsorisation

On calcule les coefficients de winsorisation : $c_i^w = \frac{W_i^w}{W_i^{CNR}}$

Puis on effectue l'opération suivante (le yi de droite correspond donc au ϵ_i de l'étape précédente) :

$$y_i \leftarrow c_i^w y_i$$

Étape 4 : Estimation de variance

Finalement l'estimation de variance s'obtient en effectuant le calcul suivant :

$\widetilde{V} = Y_r' Q_{3r} Y_r + \sum_{i \in r^B} \frac{1-\pi_i^B}{(\pi_i^B)^2} w_i^2 y_i^2$ avec y_i transformé comme indiqué dans les étapes précédentes.

Élargissement aux domaines infra-groupes de la NAF

Lorsque le domaine d_G sur lequel est produit l'estimation est inférieur au niveau groupe, compris dans le groupe *G*, et dans le sous-champ 1, on utilise l'estimateur suivant[22] :

$$\hat{Y}_{d_G} = Y_{d_G}^{Redi} + \left(\hat{Y}_G^{Dif} - Y_G^{Redi}\right) \frac{Y_{d_G}^{Redi}}{\sum_{d \subset G} Y_d^{Redi}}$$

Avec :

$$Y_a^{Redi} = \sum_{i \in U_1} Y_i^{Redi} 1_{APE_{diff} \subset a}$$

En supposant que Y_a^{Redi} ne dépend pas de l'échantillon[23], on a :

$$V\left(\hat{Y}_{d_G}\right) = V\left(\hat{Y}_G^{Dif} \frac{Y_{d_G}^{Redi}}{\sum_{d \subset G} Y_d^{Redi}}\right) = \left(\frac{Y_{d_G}^{Redi}}{\sum_{d \subset G} Y_d^{Redi}}\right)^2 V\left(\hat{Y}_G^{Dif}\right)$$

Finalement, pour un domaine d_G infra au groupe *G*, la variance de l'estimateur du total sur d_G se déduit de la variance de l'estimateur du total sur *G,* en multipliant par le facteur

$$C_{d_G} = \left(\frac{Y_{d_G}^{Redi}}{\sum_{d \subset G} Y_d^{Redi}}\right)^2$$

Validation du programme informatique par des jeux d'essai

Afin de valider le programme informatique, nous avons élaboré plusieurs jeux d'essai simples dans le but de vérifier que le programme aboutissait à des résultats corrects dans des cas simples connus.

1) Sondage aléatoire simple d'UL indépendantes

Pour ce jeu d'essai on se place dans la situation la plus simple :

- Il n'y a que des UL indépendantes dans la base de sondage ;
- Le taux de sondage est de 10 % appliqué dans toute la base (pas de stratification) ;
- Toutes les unités sont répondantes (pas de non-réponse) ;
- Pas de changements de contours.

22 Voir document pdf *Estimateurs infra-groupe nouvelle méthodologie*.

23 En toute rigueur, comme il s'agit de données Redi (c'est-à-dire issues de la confrontation de données administratives et de données de l'enquête) la valeur dépend de l'échantillon, mais on fait ici comme si ce n'était pas le cas (à voir plus tard si étudier ce point paraît nécessaire).

Constitution du jeu d'essai : On génère N=1 000 observations, et pour chaque observation on génère trois variables :

$CA \sim |(N(\mu = 100, \sigma = 50))|$;

$c \sim U(0,1)$;

$VA = c \times CA$.

C'est la variable VA qui sera notre variable d'intérêt ici. Cette variable étant générée sur l'ensemble des 1 000 observations, on peut calculer son « vrai » total T_{VA} = 49 863.

On réalise ensuite 10 000 tirages d'échantillons de taille n=100 selon un sondage aléatoire simple.

Pour chacun des échantillons s, on calcule :

- l'estimateur du total de VA : $\hat{T_{VA}}(s) = \sum_{i \in s} w_i VA_i$ avec $w_i = \frac{N}{n} = 10$;

- le CV[24] associé à l'échantillon pour l'estimation de la VA totale : $\hat{CV}(s) = \frac{\hat{V}(s)}{T_{VA}}$ en utilisant le programme détaillé dans cette note pour calculer $\hat{V}(s)$.

Par ailleurs on est capable de calculer le CV empirique pour l'estimation de la VA totale :

$$CV = \frac{\sqrt{\left(\frac{1}{10000}\sum_{s=1}^{10000}\left(\hat{T_{VA}}(s) - T_{VA}\right)^2\right)}}{T_{VA}}$$

En comparant le CV empirique avec la moyenne des CV associés à chaque échantillon, on est en mesure de vérifier (voir tableau dans la partie *Résultats de la validation*) que les estimations de CV convergent vers la bonne valeur.

Remarque : on peut aussi vérifier dans ce cas simple que pour chaque échantillon s $\hat{V}(s)$ *obtenu avec le programme correspond à l'estimateur d'Horvitz-Thompson de la variance d'un total pour un Sondage aléatoire simple qu'on obtient la formule « connue » :*

$$\hat{V}(s) = N^2(1-f)\frac{s_{VA}^2}{n} \quad \text{où} \quad s_{VA}^2 = \frac{1}{n-1}\sum_{i \in s}(VA_i - \bar{VA})^2 \quad \text{et} \quad \bar{VA} = \frac{1}{n}\sum_{i \in s} VA_i$$

24 On met le vrai total au dénominateur du CV.

2) Sondage aléatoire simple d'EP sans changement de contours

Ce jeu d'essai ressemble au précédent, mais on regroupe aléatoirement les 1 000 UL en M=300 EP avant le tirage. La valeur ajoutée d'une EP est calculée comme la somme des VA des UL rattachées.

Puis on effectue 10 000 tirages par sondage aléatoire simple de m=70 EP parmi les 300[25].

L'estimateur du total de VA devient $\hat{T_{VA}}(s_{EP}) = \sum_{i \in s_{EP}} w_i VA_i$ avec $w_i = \frac{M}{m} = 4{,}13$;

De façon similaire au jeu d'essai précédent, on compare le CV empirique obtenu avec les 10 000 estimations de CV.

Remarque : on peut aussi vérifier dans ce cas simple que pour chaque échantillon s, l'estimateur de variance $\hat{V}(s)$ *obtenu avec le programme détaillé dans cette note correspond à la formule « connue »[26] :* $\hat{V}(s) = M^2(1-f)\frac{s_{VA}^2}{m}$ *.avec*

$s_{VA}^2 = \frac{1}{m-1}\sum_{i \in s}(VA_i - \bar{VA})^2$ et $\bar{VA} = \frac{1}{m}\sum_{i \in s} VA_i$ *où i représente ici les EP contrairement au jeu d'essai précédent où il représentait les UL)*

3) Sondage aléatoire simple d'EP sans changement de contours avec non-réponse

Ce jeu d'essai ressemble au précédent, mais on ajoute une phase de non-réponse. Pour simuler la non-réponse on applique un tirage de poisson à chaque échantillon d'EP avec la probabilité de tirage 60 % pour tout le monde.

L'estimateur du total de VA devient[27]

$\hat{T_{VA}}(r_{EP}) = \sum_{i \in r_{EP}} \frac{w_i}{p_i} VA_i$ avec $p_i = 0{,}6$ et $w_i = \frac{M}{m} = 4{,}13$;

Remarque : que pour chaque échantillon s, l'estimateur de variance $\hat{V}(s)$ obtenu avec le programme détaillé dans cette note correspond à la formule « connue »[28] :

$\hat{V}(s) = M^2(1-f)\frac{\tilde{s}_{VA}^2}{m} + \sum_{i \in r}\frac{1-p_i}{p_i}VA_i^2$ Avec

$\tilde{s}_{VA}^2 = \frac{1}{m-1}\sum_{i \in s}\left(\frac{VA_i}{p_i} - \bar{VA}\right)^2$ et $\bar{VA} = \frac{1}{m}\sum_{i \in s}\frac{VA_i}{p_i}$

25 En fait on génère les EP en tirant aléatoirement pour les 1000 UL un nombre (qui sera l'identifiant de l'EP) entre 1 et 300. Ainsi certains nombre ne sont jamais tirés et en pratique, l'aléa a conduit à obtenir 289 EP.

26 Estimateur d'Horvitz-Thompson de la variance d'un total pour un Sondage en grappes.

27 Pour que les simulations convergent vers le bon CV, il est necessaire de prendre le vrai p_i (et pas une estimation comme le nombre d'EP répondantes sur le nombre d'EP tirées).

28 Formule de Rao pour un tirage avec 1er degré : Sondage en grappes, 2ème degré : tirage de Poisson.

4) Sondage aléatoire simple d'EP avec non-réponse et changements de contours

Par rapport au jeu d'essai précédent, on ajoute des changements de contours des EP. Il y a donc une phase de partage des poids.

Pour créer les contours des EP finales, on procède comme pour les EP de tirage : on regroupe aléatoirement les 1 000 UL de la base UL en M=300 EP[29]. La valeur ajoutée d'une EP « finale » est calculée comme la somme des VA des UL rattachées.

On procède toujours au tirage des EP initiales, ce qui permet d'obtenir un échantillon d'UL rattachées et d'en déduire un échantillon d'EP finales avec la règle qu'une EP finale est dans l'échantillon si au moins une de ses UL est dans l'échantillon d'UL.

L'estimateur du total de VA est alors : $\hat{T_{VA}}(r_{EP}) = \sum_{i \in r_{EP}} \frac{w_i}{p_i} VA_i$ avec $p_i = 0{,}6$ et w_i issue d'un partage des poids des UL rattachées à i pondéré par le CA des UL.

5) Sondage stratifié d'EP avec non-réponse et changements de contours

Par rapport au jeu d'essai précédent, on crée deux strates lors de la génération des données avec les caractéristiques suivantes :

Strate	N (nombre d'UL)	M (nombre d'EP_TIR dans la BdS)	m (nombre d'EP_TIR dans l'échantillon)	CA (Loi pour générer le CA)
S1	1000	300	30	N(100,50)
S2	500	100	50	N(1000,200)

Les EP finales sont générées ensuite en regroupant aléatoirement les 1500 UL en 400 EP finales. En particulier, une même EP finale peut regrouper à la fois des UL de la strate 1 et de la strate 2.

Pour simuler la non-réponse on applique toujours un tirage de poisson à chaque échantillon d'EP finales avec la probabilité de tirage 60 % pour tout le monde.

6) Sondage stratifié d'EP avec changements de contours et non-réponse dans deux GRH

Ce jeu d'essai est le même que le précédent, mais au lieu d'appliquer une probabilité de réponse homogène de 60 %, on sépare aléatoirement les EP finales en deux GRH :

GRH	M	p_i
A	100	50 %
B	300	80 %

29 Comme les regroupements sont effectués aléatoirement, les EP finales sont différentes des EP de tirage.

7) Sondage stratifié d'EP avec strate exhaustive, changements de contours et non-réponse dans deux GRH

Ce jeu d'essai est le même que le précédent, mais on ajoute une strate exhaustive (S3).

Strate	N (nombre d'UL)	M (nombre d'EP_TIR dans la BdS)	m (nombre d'EP_TIR dans l'échantillon)	CA (Loi pour générer le CA)
S1	1000	300	30	N(100,50)
S2	500	100	50	N(1000,200)
S3	100	10	10	N(2000,500)

Les EP finales exhaustives sont ensuite détectées selon la règle qu'il suffit qu'une UL rattachée à l'EP finale soit dans l'exhaustif « initial » pour que l'EP se retrouve dans l'exhaustif.

Par ailleurs, il n'y a pas de non-réponse dans cette strate exhaustive. Le nombre d'EP « finales » est fixé à 500.

8/ Sondage stratifié d'EP avec strate exhaustive, changements de contours, non-réponse dans deux GRH, et domaine

Dans ce jeu d'essai on ajoute un domaine de diffusion d : on tire au hasard 40 % des EP de la base finale pour les considérer dans le domaine. La variable d'intérêt devient $VA.1_{i\in d}$.

Le nombre d'EP dans la première strate a été augmenté à 800 (au lieu de 300) afin d'obtenir davantage d'indépendantes.

9/ Sondage stratifié d'EP avec strate exhaustive, changements de contours, non-réponse dans deux GRH, et calage

Ce jeu d'essai correspond au jeu d'essai numéro 7 mais l'échantillon hors exhaustif est calé sur deux variables :

- Le CA total de la partie non exhaustive
- Le nombre finale d'EP non exhaustives.

Comme dans le jeu d'essai précédent, le nombre d'EP dans la première strate a été augmenté à 800 (au lieu de 300 dans le jeu 7).

10/ Sondage stratifié d'EP avec strate exhaustive, changements de contours, non-réponse dans deux GRH, calage et domaines

Ce jeu d'essai est le même que le numéro 9 mais l'estimation finale porte sur un domaine plutôt que sur le champ complet. Le domaine est créé de la même façon que dans le jeu d'essai 8.

11/ Sondage stratifié d'EP avec strate exhaustive, changements de contours, non-réponse dans deux GRH, calage, domaines et winsorisation

Ce jeu d'essai est le même que le précédent, mais on introduit une étape de winsorisation du CA. Par commodité et pour que l'effet de la winsorisation soit accentué, on fixe le seuil de winsorisation au 70ème percentile de la distribution de CA et on effectue une winsorisation de type 1[30], c'est à dire que si le CA d'une unité dépasse le seuil K, alors son nouveau poids est $w_i^w = w_i^{CNR} \frac{K}{CA}$. Seule la partie non-exhaustive est winsorisée. Un calage est effectué ensuite sur le CA (non winsorisé) avec en entrée les poids winsorisé.

Il est à noter que d'après les simulations, ce calage réduit en grande partie le biais introduit, mais nos estimations de variance ne convergent plus vers la variance empirique[31] . On pouvait s'attendre à un résultat de ce type car l'équivalence entre la variance d'un estimateur calé et la variance de l'estimateur du total des résidus de la régression de la variable d'intérêt sur les variables de calage n'est valide que lorsque les poids correspondent à un estimateur sans biais de la marge de calage, ce qui n'est pas le cas ici à cause de la winsorisation.

En pratique, dans l'ESA, ce phénomène est probablement beaucoup moins prononcé car les seuils de winsorisation sont a priori plus haut (peu d'unités sont winsorisées).

12/ Sondage stratifié d'EP avec strate exhaustive, changements de contours, non-réponse dans deux GRH, calage, domaines et estimateur par différence

Ce jeu d'essai est le même que le jeu d'essai 10, avec en plus une étape d'estimateur par différence pour les unités indépendantes.

En pratique on génère :

- Un domaine « administratif » valant 9 fois sur 10 le « vrai » domaine.
- Une VA « administrative » valant entre 0,75 et 1,25 fois la « vraie » VA.

On ne fait pas de winsorisation pour ce jeu d'essai afin de ne pas perturber les résultats (voir jeu d'essai 11).

30 En pratique, dans Esane et dans la plupart des enquêtes c'est une winsorisation de type 2 qui est effectuée, mais la winsorisation de type 1 a un effet plus marqué, c'est pourquoi elle a été privilégiée pour ces jeux d'essai.

31 En réalité c'est une erreur quadratique moyenne qui est calculée (voir formule p.19), et comme le biais introduit par la winsorisation n'est pas pris en compte par notre estimateur de variance, on corrige cette erreur quadratique moyenne en retirant le biais empirique au carré. C'est sur la base de ce calcul de variance empirique que sont présentés les CV de référence.

Résultats de la validation

Voici les résultats des simulations sur les différents jeux d'essai.

Jeu d'essai	CV empirique	$\hat{CV}_{100}$	$\hat{CV}_{1000}$	$\hat{CV}_{10000}$
1	7,903%	7,755%	7,869%	7,870%
2	6,851%	6,878%	6,916%	6,899%
3	13,413%	13,178%	13,232%	13,219%
4	10,039%	9,941%	10,018%	10,039%
5	7,701%	7,805%	7,737%	7,743%
6	6,906%	6,974%	6,904%	6,914%
7	4,545%	4,505%	4,503%	4,504%
8	7,302%	7,352%	7,272%	7,234%
9	1,987%	1,949%	1,934%	1,937%
10	5,617%	5,455%	5,506%	5,505%
11	*6,661%*	*5,550%*	*5,579%*	*5,566%*
12	5,762%	5,662%	5,664%	5,642%

Note de lecture : Pour le jeu d'essai n°1 (Sondage aléatoire simple d'UL indépendantes) le CV empirique (basé sur les 10 000 itérations) est de 7,9 %. La moyenne des estimations de CV avec le programme détaillé dans cette note est de 7,755 % sur les 100 premiers échantillons, 7,869 % sur les 1 000 premiers échantillons, 7,870 % sur les 10 000 échantillons.

On remarque que pour tous les jeux d'essai sauf le 11 (winsorisation + calage), l'estimateur du CV construit avec notre programme converge[32] bien vers le CV empirique.

Le fait que le jeu d'essai 11 ne converge pas vers la « bonne » valeur était prévisible car la winsorisation fait que le calage sort ensuite de son cadre théorique. Néanmoins, en pratique, la winsorisation touche moins d'unité et on suppose donc que l'écart avec le cadre théorique (le biais introduit) sera négligeable.

32 L'écart restant est probablement dû au nombre de simulations (10 000) qu'il faudrait augmenter. Cela peut aussi venir d'approximations réalisées : par exemple pour le calage, on sait que la méthode d'estimation (remplacer la variable d'intérêt par le résidu de la régression pondérée de la variable d'intérêt sur les variables de calage) n'est qu' asymptotiquement sans biais.

Annexes

Démonstration de $\hat{Y}_B = \hat{Z}_A = \sum_{j \in s^A} w_j z_j$ **Avec** $z_j = \sum_{i \in U^B} \theta_{i,j} y_i$

Soit q_i l'indicatrice d'appartenance à l'échantillon d'EP.

$$\hat{Y}_B = \sum_{i \in s^B} W_i y_i = \sum_{i \in U^B} W_i y_i q_i = \sum_{i \in U^B} \left(\sum_{j \in U^A} \theta_{i,j} w_j t_j \right) y_i q_i = \sum_{j \in U^A} w_j t_j \left(\sum_{i \in U^B} \theta_{i,j} y_i q_i \right)$$
$$= \sum_{j \in s^A} w_j \left(\sum_{i \in U^B} \theta_{i,j} y_i q_i \right) = \sum_{j \in s^A} w_j \left(\sum_{i \in s^B} \theta_{i,j} y_i \right)$$

Comme (voir corps de la note) $\sum_{i \in s^B} \theta_{i,j} y_i = \sum_{i \in U^B} \theta_{i,j} y_i = z_j$

Alors $\hat{Y}_B = \sum_{j \in s^A} w_j z_j = \hat{Z}_A$

Démonstration de $\tilde{Y}_B$ **est un estimateur sans biais de Y**

On a : $E_{\pi_A \pi_B}(\tilde{Y}_B) = E_{\pi_A}[E_{\pi_B}[\tilde{Y}_B | s^A]]$

Or : $E_{\pi^B}[\tilde{Y}_B | s^A] = E_{\pi^B}[\sum_{i \in r^B} W_i \frac{y_i}{\pi_i^B} | s^A] = E_{\pi^B}[\sum_{i \in s^B} W_i \frac{y_i r_i}{\pi_i^B} | s^A]$ avec r_i l'indicatrice de réponse de l'EP i.

Comme l'application du partage des poids est déterministe (la connaissance des liens ne vient pas de l'échantillon), la connaissance de s^A implique la connaissance de s^B.

Donc $E_{\pi^B}[\tilde{Y}_B | s^A] = E_{\pi^B}[\tilde{Y}_B | s^B] = \sum_{i \in s^B} W_i \frac{y_i}{\pi_i^B} E_{\pi^B}[r_i | s^B]$

Comme π^B représente un tirage de Poisson des EP de s^B dans r^B, on a : $E_{\pi^B}[r_i | s^B] = \pi_i^B$

Donc $E_{\pi^B}[\tilde{Y}_B | s^A] = \sum_{i \in s^B} W_i y_i = \hat{Y}_B$

D'où : $E_{\pi_A \pi_B}(\tilde{Y}_B) = E_{\pi_A}[\hat{Y}_B]$

Comme $\hat{Y}_B = \hat{Z}_A = \sum_{j \in s^A} w_j z_j = \sum_{j \in U^A} w_j z_j t_j$ et que $E_{\pi_A}[t_j] = \frac{1}{w_j}$ (wj est le poids de sondage de l'UL j), on a : $E_{\pi_A}[\hat{Y}_B] = \sum_{j \in U^A} z_j = Z_A$

Or $Z_A = \sum_{j \in U^A} z_j = \sum_{j \in U^A} \sum_{i \in U^B} \theta_{i,j} y_i = \sum_{i \in U^B} \left(\sum_{j \in U^A} \theta_{i,j} \right) y_i$

Comme $\sum_{j \in U^A} \theta_{i,j} = 1$ (propriété imposée aux $\theta_{i,j}$)

Alors $Z_A = Y_B$

Et Finalement : $E_{\pi_A \pi_B}(\tilde{Y}_B) = Y_B$ donc $\tilde{Y}_B$ est sans biais.