Méthodologie statistique

M 2015/03

Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse

Emmanuel Gros – Karim Moussallam

Document de travail

Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2015/03

Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse

Emmanuel Gros* – Karim Moussallam**

Les auteurs remercient Guillaume Chauvet – qui a initié les réflexions méthodologiques sur les calculs de précision dans le contexte d'Octopusse et a aimablement mis à disposition ses programmes d'estimation de probabilités d'inclusion double dans le cadre de tirages équilibrés –, ainsi que Sébastien Faivre – pour sa relecture attentive.

* Insee – DMCSI – Division Sondages
18 boulevard Adolphe Pinard - 75675 PARIS CEDEX 14

** Au moment de la rédaction de ce document de travail, Insee – DMCSI – Division Sondages
18 boulevard Adolphe Pinard - 75675 PARIS CEDEX 14

Direction de la méthodologie et de la coordination statistique et internationale – Département des Méthodes Statistiques – Timbre L101
18, bd Adolphe Pinard – 75675 PARIS CEDEX – France
Tél. : 33 (1) 41 17 66 01 – Fax : 33 (1) 41 17 61 97 – CEDEX – E-mail : DG75-L001@insee.fr – Site Web Insee : http://www.insee.fr

Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse

Résumé

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse, s'articule autour de deux concepts majeurs : d'une part le principe de système de tirage à deux degrés des échantillons-maîtres classiques, et d'autre part le recensement rotatif de la population qui permet l'apport d' « informations fraîches » concernant les logements à échantillonner. Cette interaction entre ces deux concepts conduit à un processus d'échantillonnage efficace mais singulièrement complexe, qui rend particulièrement ardus les calculs de précision portant sur des statistiques issues d'enquêtes tirées dans Octopusse. Ce document de travail représente l'aboutissement de plusieurs années de travail sur le sujet.
La première partie, intitulée « Compendium », revient sur le cadre méthodologique du système Octopusse et présente une synthèse du reste du document de travail. Les parties suivantes exposent beaucoup plus en détails les simulations et études effectuées, les différentes formules de variance et leur justification, ainsi que les résultats obtenus lors d'une première application de cette méthode d'estimation de variance dans le cadre de l'enquête AES 2012.

Mots clés : Enquêtes ménages, estimation de variance, Octopusse, échantillon-maître, recensement de la population, sondage en plusieurs phases, échantillons équilibrés, estimation de probabilités d'inclusion double, estimateur de variance de Yates-Grundy.

Abstract

The current sampling system for households surveys conducted by INSEE, named Octopusse, is based on two major concepts: on the one hand, the principle of two-stage sampling use for the traditional Master Samples, and on the other hand the new population census, based on annual census surveys which contains updated information. The interaction between these two concepts leads to an efficient but singularly complex sampling procedure, which makes variance estimation particularly difficult. This working paper constitutes the outcome of years of work on this subject.
The first part, titled "Compendium", draws the methodological framework of the sampling system Octopusse and presents a synthesis of the rest of the working paper. The following sections set out in details the simulations and studies carried out, the various variance formulas and their justification as well as the results obtained for a first application of this variance estimation method to the AES 2012 survey.

Keyword : Household survey, variance estimation, Octopusse, Master Sample, population census, multiphase sampling, balanced sampling, estimation of second-order inclusion probability, Yates-Grundy variance estimator.

Table des matières

Compendium

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse, s'articule autour de deux concepts majeurs : d'une part le principe de système de tirage à deux degrés des échantillons-maîtres classiques, et d'autre part le recensement rotatif de la population qui permet l'apport d' « informations fraîches » concernant les logements à échantillonner. Cette interaction entre ces deux concepts conduit à un processus d'échantillonnage plus efficace que dans le cadre du précédent système d'échantillon-maître, mais également nettement plus complexe : aux aléas de sondage « classiques » des échantillons-maîtres s'ajoutent les aléas de sondage relatifs aux enquêtes annuelles de recensement. On passe ainsi du cadre théorique d'un sondage à deux degrés pour les anciens échantillons-maîtres à celui d'un sondage en trois phases pour le système Octopusse, ce qui rend particulièrement ardus les calculs de variance.

En 2011, Guillaume Chauvet s'est intéressé au problème des calculs de précision dans le contexte d'Octopusse, et a proposé dans [2], dans un cadre légèrement simplifié[1], une méthode d'estimation de variance fondée sur l'utilisation d'estimateurs de variance de Yates-Grundy. Ces estimateurs s'appuient sur des probabilités d'inclusion double estimées par réplication à partir des propriétés de martingale de l'algorithme du Cube. Prolongeant ces travaux, et s'appuyant sur les programmes fournis par Guillaume Chauvet, Karim Moussallam a mené un travail complet et rigoureux d'élaboration et de mise en œuvre des formules de calcul de variance dans le contexte réel d'Octopusse. Ces travaux ont consisté :

- d'une part à calculer, via la méthode mise au point par Guillaume Chauvet, les probabilités d'inclusion double – des communes dans les groupes de rotation du recensement d'une part, des unités primaires dans l'échantillon-maître d'autre part –, à analyser la qualité de ces probabilités et celle des estimateurs de Yates-Grundy les utilisant. Les résultats de ces travaux sont détaillés dans les parties II et III, et une synthèse en est présentée dans la partie 2.2. de ce compendium ;
- d'autre part à prendre en compte les aspects du système Octopusse non traités dans le cadre simplifié des travaux de Guillaume Chauvet : sélection des logements au sein de la fraction des unités primaires recensée lors de la dernière enquête annuelle de recensement, prise en compte de la non-réponse, calage. C'est l'objet de la partie IV, qui se trouve synthétisée dans les parties 2.3. et 2.4 de ce compendium.

1 Le système d'échantillonnage des enquêtes ménages à l'Insee : Octopusse

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse[2] et présenté en détail dans [1], a été conçu pour répondre à un double objectif :

- d'une part conserver le principe de système de tirage à deux degrés des Échantillons-Maîtres (EM) classiques construits auparavant à partir des recensements exhaustifs : constitution et tirage d'unités primaires une fois pour toutes à l'initialisation du système, puis tirage pour chaque enquête d'un échantillon de logements au sein de chaque unité primaire. Cette ligne directrice permet en effet d'assurer une précision acceptable pour les enquêtes nationales tout en limitant les coûts d'enquête, notamment via la constitution d'un réseau d'enquêteurs fixe et pérenne et une limitation des coûts de déplacement ;

[1] En particulier, le dernier degré de sondage correspondant à la sélection des logements au sein de la fraction des unités primaires recensée lors de la dernière enquête annuelle de recensement n'était pas pris en compte.
[2] Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Échantillons.

- d'autre part pouvoir bénéficier de la « fraîcheur » des informations disponibles via le recensement rotatif continu de la population mis en place en 2004. Pour ce faire, la sélection des échantillons des enquêtes ménages est effectuée dans une base de sondage fraîche composée des logements recensés lors de l'Enquête Annuelle de Recensement (EAR) de l'année N-1.

Afin d'employer un réseau fixe d'enquêteurs tout en interrogeant des logements tirés dans la dernière EAR, les unités primaires du système Octopusse – renommées à cette occasion Zones d'Action Enquêteurs (ZAE) – ont été adaptées à cet objectif :

- **en grandes communes :** chaque grande commune[3] constitue une ZAE « Grande Commune » (ZAEGC) à elle seule. Les ZAEGC de plus de 40 000 résidences principales au recensement de 1999 sont « exhaustives » (i.e. sélectionnées d'office) ;
- **en petites communes :** chaque ZAE « Petites Communes » (ZAEPC) est constituée de petites communes appartenant aux cinq groupes de rotation de façon à avoir 300 résidences principales dans chacun des cinq groupes. Les ZAEPC ainsi constituées sont donc des objets aléatoires, construits conditionnellement à l'affectation aléatoire des petites communes en groupes de rotation effectuée par le recensement.

À l'initialisation du système, un échantillon-maître de 525 ZAE – 37 ZAEGC exhaustives, 202 ZAEGC non exhaustives et 286 ZAEPC – a été sélectionné pour la réalisation des enquêtes ménages nationales. Puis chaque année, la base de sondage annuelle d'Octopusse est constituée en chargeant les logements recensés lors de l'EAR de l'année N-1 situés dans cet échantillon-maître. Ensuite, deux opérations statistiques sont menées, avant de procéder au tirage des logements d'une enquête donnée dans cette base de sondage :

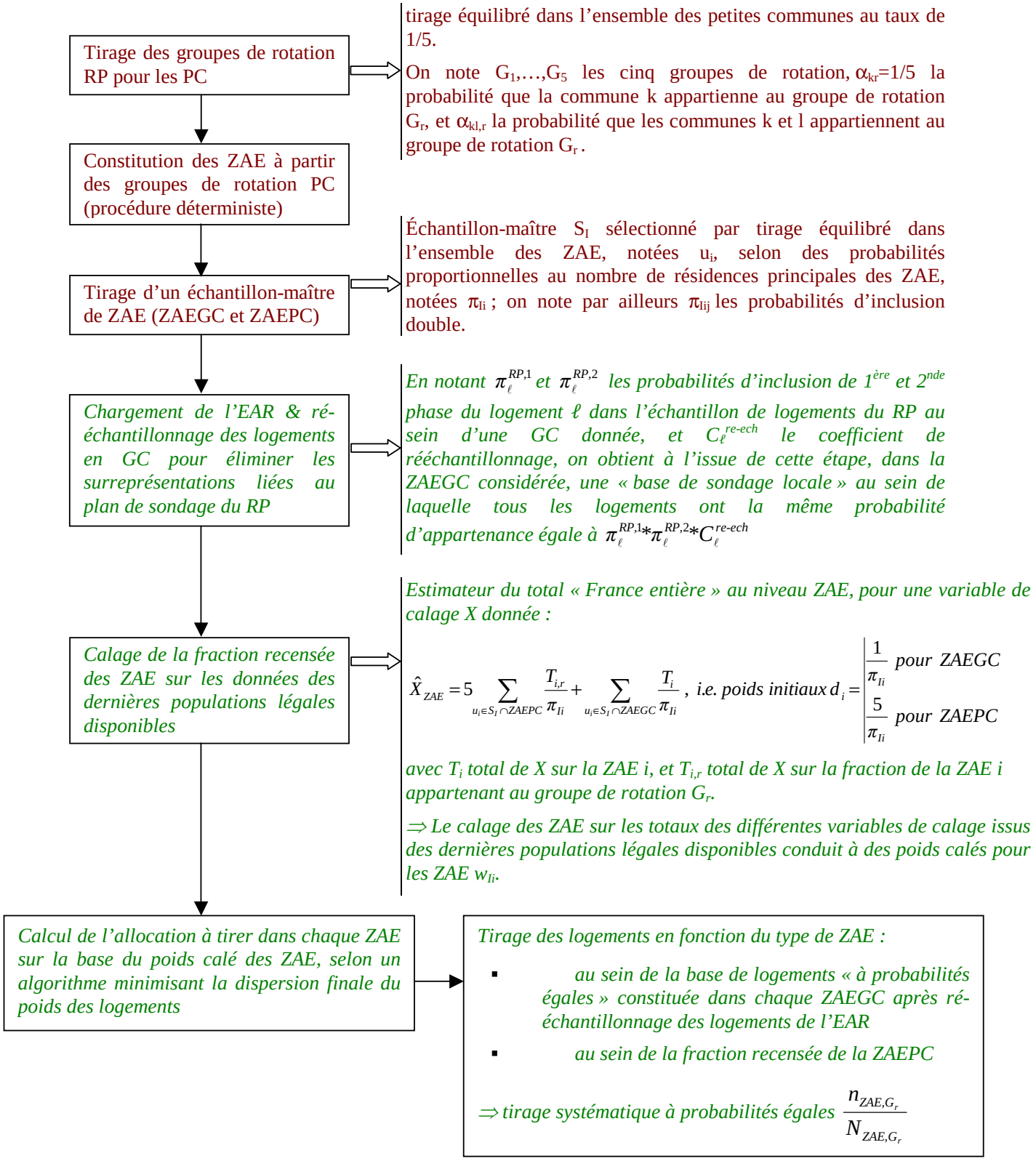
- d'une part, l'Enquête Annuelle de Recensement surreprésente certaines strates de logements en grandes communes : logements des grandes adresses[4] et des adresses neuves. Afin d'éliminer ces surreprésentations, une procédure de rééchantillonnage – qui consiste à ne conserver, par tirage aléatoire, qu'une fraction des logements recensés dans les strates surreprésentées – permet de constituer une base de sondage de logements « à probabilités égales » au sein de chaque ZAEGC ;
- d'autre part, l'analyse des bases de sondage annuelles – composées des communes des ZAE de l'échantillon-maître appartenant aux fractions recensées lors de la dernière EAR – a mis en évidence des problèmes de représentativité, notamment pour des variables de segmentation de l'espace urbain / périurbain / rural. Afin de pallier ce problème, un calage des ZAE sur différentes variables socio-démographiques issues des dernières populations légales disponibles est réalisé chaque année à l'initialisation de la campagne. Cette opération permet d'améliorer la représentativité des échantillons de logements, d'une part en assurant la représentativité de la « base de sondage annuelle » des unités primaires restreintes à la dernière EAR, et d'autre part en augmentant / diminuant les allocations de logements à tirer par ZAE dans les zones dont le profil est sous-représenté / surreprésenté.

Sur la base des poids calés des ZAE, on calcule une allocation à tirer dans chaque ZAE avec un algorithme de minimisation de la dispersion du poids final des logements. Enfin, le tirage des logements est effectué dans chaque ZAE par tirage au sein des logements chargés dans la ZAE lors du chargement de la dernière EAR disponible et conservés à l'issue de la phase de rééchantillonnage. Le tirage est un tirage systématique à probabilités égales au sein des logements

[3] Les grandes communes, au sens du recensement, sont les communes de 10 000 habitants ou plus.
[4] Est considérée comme grande adresse toute adresse dont le nombre de logements est au moins égal à 60 et qui est telle que l'ensemble des grandes adresses ne représente pas plus de 10% des logements de la commune.

encore disponibles (logements qui n'ont pas déjà été sélectionnés pour une enquête au cours des quatre années précédentes).

Au final, le tirage d'un échantillon dans le cadre du système Octopusse s'effectue donc selon le schéma suivant – en rouge plein, procédures réalisées une seule fois à l'initialisation d'Octopusse, *en vert italique, procédures réalisées tous les ans pour les tirages d'échantillons de la campagne en cours* :



Dans ce contexte, l'estimateur utilisé pour estimer le total d'une variable Y est le suivant[5] :

$$\hat{Y} = \sum_{\substack{\ell \in \text{échantillon} \\ \text{final de logements}}} w_\ell y_\ell, \text{ avec } w_\ell = \begin{cases} w_{Ii} \dfrac{N_{i,r}}{n_{i,r}} \text{ pour tout logement } \ell \in ZAEPC\, u_i \cap G_r \\ w_{Ii} \dfrac{1}{\pi_\ell^{RP,1} * \pi_\ell^{RP,2} * C_\ell^{re\text{-}ech}} \dfrac{N_{i,r}}{n_{i,r}} \text{ pour tout logement } \ell \in ZAEGC\, u_i \cap G_r \end{cases}$$

2 Les calculs de précision dans Octopusse

2.1. Cadre général, hypothèses et notations

Ainsi, Octopusse constitue un système d'échantillonnage complexe au sein duquel s'imbriquent plusieurs phases d'échantillonnage non indépendantes – en particulier la constitution des groupes de rotation du recensement et la sélection de l'échantillon-maître de ZAE – et qui conduit à pas moins de cinq niveaux d'aléa :

- le tirage des groupes de rotation en petites communes, qui détermine la constitution des ZAE et leurs probabilités de tirage – proportionnelles à la taille totale de la ZAE ;
- le tirage de l'échantillon-maître de ZAE ;
- le tirage des adresses de l'Enquête Annuelle de Recensement au sein des ZAEGC tirées pour Octopusse ;
- le tirage des logements conservés en ZAEGC pour la base de sondage annuelle d'Octopusse – processus de rééchantillonnage ;
- enfin, le tirage final des logements au sein de la fraction recensée des ZAE de l'EM.

Si l'on ajoute à cela l'opération de calage des ZAE, estimer la variance d'une enquête tirée dans Octopusse de manière exacte relève de la gageure, et il est impératif de procéder à un minimum d'hypothèses simplificatrices. En conséquence, les approximations suivantes ont été effectuées :

❶ l'impact du calage des ZAE sur la précision des estimations n'est pas pris en compte ;

❷ dans les ZAEGC, on assimile l'enchaînement des deux phases de sélection des adresses du recensement, de la phase de ré-échantillonnage et du tirage final de l'échantillon de logements à un unique tirage de n_k logements parmi les N_k logements de la ZAEGC k ;

❸ enfin, la variance intra-communale résultant de l'estimation de la variance liée à la constitution des groupes de rotation du recensement sur l'échantillon des logements finaux est négligée. Cette hypothèse signifie que, lorsque l'on estimera, à partir de l'échantillon de logements, la variance liée au tirage des groupes de rotation du RP, on ne prendra pas en compte pour cette composante de la variance l'aléa lié au tirage des logements au sein des communes[6].

Sous ces hypothèses, une formule de variance analytique « générique » – i.e. valable pour toute enquête standard tirée dans Octopusse – a pu être établie ; cette formule, exposée en détail avec son application à l'enquête AES dans la partie IV de ce document de travail, repose sur les deux principes centraux suivants :

[5] On ne prend pas en compte à ce stade la non-réponse observée lors de l'enquête, ni un éventuel calage final.
[6] Cf. point 3 du §2.3 de ce compendium pour plus de détails.

- d'une part, la variance liée aux tirages équilibrés des groupes du recensement et des ZAE de l'échantillon-maître est estimée en suivant la méthode proposée par Guillaume Chauvet dans [2]. Cette méthode – qui repose sur l'utilisation d'estimateurs de variance de Yates-Grundy s'appuyant sur des probabilités d'inclusion double estimées par réplication à partir des propriétés de martingale de l'algorithme du Cube – permet en effet, contrairement à la formule « usuelle » proposée par Deville et Tillé dans [3], de prendre en compte la variance liée à la phase d'atterrissage de l'algorithme du Cube, qui risque d'être importante, au moins pour la constitution de l'échantillon-maître, étant donné la taille relativement faible de ce dernier dans certaines régions et le nombre de contraintes d'équilibrage retenues. Cet estimateur, ainsi que ses principales propriétés, sont détaillées dans les parties II et III de ce document de travail, et une synthèse de ces travaux est présentée en partie 2.2. de ce compendium ;
- d'autre part, les degrés de sondage relatifs aux sélections de logements – tirage des logements au sein de la fraction recensée lors de la dernière EAR des ZAE de l'EM, non-réponse, etc. – sont pris en compte en appliquant la formule de décomposition de la variance ainsi que diverses formules relatives aux sondages à plusieurs degrés, du type formule de Rao. Les parties 2.3. et 2.4. explicitent ces formules ainsi que la façon dont elles sont appliquées dans le cadre de l'estimation de variance d'Octopusse, ce que l'on retrouve avec plus de détails dans la partie IV de ce document de travail.

Les notations utilisées dans la suite sont cohérentes avec celles de l'article de Guillaume Chauvet [2], ainsi qu'avec celles du schéma récapitulatif de la page 3. Plus précisément, on note :

- U la population des communes ;
- $G_1,\ldots,G_5$ les cinq groupes de rotation constitués dans le cadre du recensement ; par convention[7], on considère que toutes les grandes communes sont sélectionnées exhaustivement dans chaque groupe de rotation ;
- α_{kr} la probabilité que la commune k appartienne au groupe de rotation G_r – égale à 1/5 pour les petites communes, et à 1 par convention pour les grandes communes – et $\alpha_{kl,r}$ la probabilité que les communes k et l appartiennent au groupe de rotation G_r ;
- U_I la population des M ZAE, notées u_i, constituées conditionnellement aux groupes de rotation G_1 à G_5 selon un algorithme déterministe ; notons que, les ZAE étant constituées exclusivement de petites communes pour les ZAEPC ou d'une seule grande commune pour les ZAEGC, au sein d'une ZAE u_i donnée, toutes les communes k ont les mêmes probabilités α_{kr} (1/5 pour les communes des ZAEPC, 1 pour les ZAEGC) ; par convention, on notera α_{ir} cette probabilité commune à toutes les communes de la ZAE u_i , et $\alpha_{\ell r} = \alpha_{kr} = \alpha_{ir}$ pour tout logement $\ell \in$ commune $k \in$ ZAE u_i ;
- S_I l'échantillon-maître de m ZAE sélectionnées selon des probabilités (conditionnelles aux groupes de rotation) proportionnelles au nombre de résidences principales des ZAE, notées π_{Ii} ; on note par ailleurs π_{Iij} la probabilité (toujours conditionnelle) d'inclusion double des unités u_i et u_j au sein de S_I ; par convention, on notera $\pi_{Ik}=\pi_{Ii}$ pour toute commune k appartenant à la ZAE u_i, et $\pi_{I\ell} = \pi_{Ik} = \pi_{Ii}$ pour tout logement $\ell \in$ commune $k \in$ ZAE u_i ;
- S_r l'ensemble des n_c communes appartenant à la fois à l'échantillon-maître de ZAE S_I et au groupe de rotation G_r → il s'agit de la « base de sondage annuelle » au sein de laquelle les logements sont sélectionnés *in fine*. On note $N_{i,r}$ le nombre de logements appartenant à la fraction de la ZAE u_i incluse dans le groupe de rotation G_r ;

[7] Il s'agit d'une convention purement technique et propre à ce document de travail, qui permet d'écrire des formules de variance plus générales, sans avoir à distinguer grandes communes et petites communes dans certains termes de variance.

- w_{Ii} le poids calé de la ZAE u_i restreinte à la dernière EAR, w_{Ik}=w_{Ii} pour toute commune k appartenant à la ZAE u_i et $w_{I\ell} = w_{Ik} = w_{Ii}$ pour tout logement $\ell \in$ commune k $\in$ ZAE u_i ;
- S_ℓ l'échantillon final de L logements obtenu en sélectionnant $n_{i,r}$ logement parmi $N_{i,r}$ au sein des communes de chaque ZAE u_i de S_r selon des probabilités $\pi_{\ell|G_r,u_i}$;
- w_ℓ le poids d'échantillonnage du logement ℓ : $w_\ell = w_{I\ell} / \pi_{\ell|G_r,u_i}$;
- enfin, pour une variable d'intérêt Y donnée, y_ℓ désignera la valeur de Y pour le logement ℓ, y_k le total de Y sur la commune k et $y_{ui\cap Gr}$ le total de Y sur la fraction recensée lors de la dernière EAR de la ZAE u_i.

2.2. Variance liée aux tirages des groupes de rotation du recensement et de l'échantillon-maître

On raisonne ici en <u>supposant connus les totaux de la variable d'intérêt Y par commune</u> – i.e. en ne prenant pas en compte les degrés de sondage relatifs aux sélections de logements – et on s'intéresse donc, pour une variable Y donnée, à l'estimateur par expansion[8] suivant :

$$\hat{Y}_\pi^{S_r} = \sum_{k\in S_r} \frac{y_k}{\alpha_{kr}\pi_{Ik}}$$

Dans [2], Guillaume Chauvet, en supposant d'une part que les groupes de rotations du recensement sont de taille fixe et d'autre part que l'échantillon-maître n'est composé que de ZAEPC, propose d'estimer sans biais sur l'échantillon S_r la variance de $\hat{Y}_\pi^{S_r}$ en s'appuyant sur des estimateurs de variance de Yates-Grundy à chaque phase, via l'estimateur suivant :

$$\hat{V}^{S_r}(\hat{Y}_\pi^{S_r}) = -\frac{1}{2}\sum_{\substack{k,l\in S_r\\ k\neq l}} \frac{\hat{\alpha}_{kl,r}-\alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}} \underbrace{\hat{\pi}_{Iij}}_{\substack{(i,j)\,/\,k\in u_i\\ l\in u_j}} \left(\frac{y_k}{\alpha_{kr}}-\frac{y_l}{\alpha_{lr}}\right)^2 - \frac{1}{2}\sum_{\substack{u_i,u_j\in S_I\\ u_i\neq u_j}} \frac{\hat{\pi}_{Iij}-\pi_{Ii}\pi_{Ij}}{\hat{\pi}_{Iij}} \left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}}-\frac{\tilde{Y}_{jr}}{\pi_{Ij}}\right)^2$$

avec $\tilde{Y}_{ir} = \sum_{k\in u_i} \frac{y_k 1\!\!1_{k\in G_r}}{\alpha_{kr}}$ et où les probabilités d'inclusion double $\hat{\alpha}_{kl,r}$ et $\hat{\pi}_{Iij}$ sont estimées par réplication en s'appuyant sur les propriétés de martingale de l'algorithme de tirage équilibré Cube selon la méthode proposée par Breidt & Chauvet dans [4]. La première somme de cet estimateur correspond à l'estimateur de variance de Yates-Grundy, estimé sur l'échantillon de seconde phase S_r, de la variance liée à la constitution des groupes de rotation du recensement, tandis que la seconde somme correspond à l'estimateur de variance de Yates-Grundy de la variance liée à la sélection des ZAE de l'échantillon-maître.

L'adaptation de cet estimateur au contexte réel d'Octopusse ne pose pas de problème particulier. Il s'agit d'une part de prendre en compte le fait que les groupes de rotation du recensement ne sont pas de taille fixe via l'ajout d'un terme correctif à l'estimateur de Yates-Grundy ; et d'autre part de prendre en compte les ZAEGC, ce qui se fait de manière relativement transparente sous l'hypothèse simplificatrice ❷ : leur contribution au 1er terme de variance lié à la constitution des groupes de rotation du RP est alors nulle, et par ailleurs, chaque grande commune constituant une ZAE à elle seule et les totaux communaux étant ici supposés connus, on a dans le 2nd terme $\tilde{Y}_{ir} = y_i$.

[8] Estimateur sans biais usuel dans le contexte d'un échantillonnage en deux phases qui est le nôtre.

Ainsi, la variance de $\hat{Y}_{\pi}^{S_r}$, liée aux aléas de sondage relatifs à la constitution des groupes de rotation du recensement d'une part et à la sélection des ZAE de l'échantillon-maître d'autre part, s'estime sans biais sur l'échantillon S_r par :

$$\hat{V}^{S_r}(\hat{Y}_{\pi}^{S_r}) = \overbrace{-\frac{1}{2}\sum_{\substack{k,l\in S_r \\ k\neq l}} \frac{\hat{\alpha}_{kl,r}-\alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}\underbrace{\hat{\pi}_{Iij}}_{\substack{(i,j)/k\in u_i \\ l\in u_j}}}\left(\frac{y_k}{\alpha_{kr}}-\frac{y_l}{\alpha_{lr}}\right)^2 + \sum_{k\in S_r}\frac{\sum_{l\in U}\hat{\alpha}_{kl,r}-\alpha_{kr}\sum_{l\in U}\alpha_{l,r}}{\alpha_{kr}\underbrace{\pi_{Ii}}_{i/k\in u_i}}\left(\frac{y_k}{\alpha_{kr}}\right)^2}^{\hat{V}_{GR}^{S_r}=Q_{GR}(y_1,\ldots,y_{n_c})}$$
$$\underbrace{-\frac{1}{2}\sum_{\substack{u_i,u_j\in S_I \\ u_i\neq u_j}}\frac{\hat{\pi}_{Iij}-\pi_{Ii}\pi_{Ij}}{\hat{\pi}_{Iij}}\left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}}-\frac{\tilde{Y}_{jr}}{\pi_{Ij}}\right)^2}_{\hat{V}_{EM}^{S_r}=Q_{EM}(\frac{\tilde{Y}_{1r}}{\pi_{I1}},\ldots,\frac{\tilde{Y}_{mr}}{\pi_{Im}})}$$

Des études par simulation ont permis d'une part d'évaluer la qualité de l'estimation des probabilités d'inclusion double[9] estimées via la méthode de Breidt & Chauvet dans le contexte d'Octopusse et d'autre part et surtout d'analyser les qualités statistiques – biais, variance – des estimateurs de Yates-Grundy reposant sur ces probabilités estimées. Les principales conclusions de ces études – qui rejoignent celles obtenues par Guillaume Chauvet dans [2] – sont les suivantes :

- la qualité des probabilités d'inclusion double estimées – sur 430 000 réplications pour les communes et sur 2 300 000 pour les ZAE – est très satisfaisante : probabilités strictement positives et inférieures à 1, taux d'erreur faible sur l'estimation des probabilités d'inclusion simple et décroissant avec le nombre de réplications, distance entre matrices des probabilités d'inclusion calculées sur deux jeux de réplications indépendants qui diminue en fonction du nombre de réplications, etc. ;
- les estimateurs de variance de Yates-Grundy reposant sur ces probabilités estimées présentent de très bonnes propriétés statistiques – absence de biais, dispersion mesurée de l'estimateur de variance – et s'avèrent préférables[10] aux autres estimateurs de variance envisageables – estimateur d'Horvitz-Thompson, de Deville-Tillé (cf. [3]) ou de Deville (cf. [5]).

À ce stade, on dispose donc d'un estimateur $\hat{V}^{S_r}(\hat{Y}_{\pi}^{S_r})$ permettant d'estimer correctement, à partir de l'échantillon de communes S_r, la variance liée aux deux premières phases du plan de sondage d'Octopusse : d'une part la variance liée à la constitution des groupes de rotation du recensement estimée par la forme quadratique $Q_{GR}(y_1,\ldots,y_{n_c}) = \hat{V}_{GR}^{S_r}$; d'autre part la variance liée à la sélection des ZAE de l'échantillon-maître estimée par la forme quadratique $Q_{EM}(y_1,\ldots,y_{n_c}) = \hat{V}_{EM}^{S_r}$. Il reste alors d'une part à estimer cette composante de variance à partir de l'échantillon de logements final S_ℓ et d'autre part à prendre en compte la variance liée au degré de sondage supplémentaire relatif au tirage des logements au sein de S_r.

[9] Des communes dans les groupes de rotation du recensement d'une part, des ZAE dans l'échantillon-maître d'autre part.

[10] Sauf dans le cas de la Corse pour l'estimation de la variance liée au tirage des ZAE de l'EM, où l'existence de probabilités d'inclusion doubles très faibles conduit à un estimateur de Yates-Grundy très instable. En conséquence, l'estimation de variance retenue *in fine* pour la Corse est celle de Deville.

2.3. Prise en compte du degré de tirage des logements et estimation de la variance sur l'échantillon de logements final.

Lorsque l'on intègre le degré de sondage supplémentaire relatif au tirage des logements au sein de S_r ainsi que l'opération de calage des ZAE effectuée dans Octopusse, l'estimateur du total d'une variable Y à partir d'un échantillon de logements issu d'Octopusse devient :

$$\hat{Y}_w^{S_\ell} = \sum_{\ell \in S_\ell} w_\ell y_\ell = \sum_{\ell \in S_\ell} w_{I\ell} \frac{y_\ell}{\pi_{\ell|G_r,u_i}} = \sum_{u_i \in S_I} w_{Ii} \sum_{\ell \in S_I \cap u_i} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}$$

Cet estimateur intègre le calage effectué sur les ZAE dans Octopusse, au travers des poids w_{Ii}. Toutefois, du point de vue des calculs de variance, cette opération de calage est impossible à prendre en compte. Aussi, conformément à l'hypothèse simplificatrice ❶ précédemment énoncée, nous allons négliger l'impact de cette opération pour le calcul de variance et nous intéresser à la variance de l'estimateur en expansion suivant, qui correspond à l'estimateur que serait celui d'Octopusse en l'absence de calage. Cet estimateur peut s'écrire de deux façons différentes :

$$\hat{Y}_\pi^{S_\ell} = \sum_{\ell \in S_\ell} \frac{y_\ell}{\alpha_{\ell r} \pi_{I\ell} \pi_{\ell|G_r,u_i}} = \begin{vmatrix} \sum_{k \in S_r} \frac{1}{\alpha_{kr}\pi_{Ik}} \overbrace{\sum_{\ell \in S_\ell \cap k} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}^{\hat{y}_k} = \sum_{k \in S_r} \frac{\hat{y}_k}{\alpha_{kr}\pi_{Ik}} \\ \sum_{u_i \in S_I} \frac{1}{\alpha_{ir}\pi_{Ii}} \underbrace{\sum_{\ell \in S_\ell \cap u_i} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}_{\hat{y}_{u_i \cap G_r}} = \sum_{u_i \in S_I} \frac{\hat{y}_{u_i \cap G_r}}{\alpha_{ir}\pi_{Ii}} \end{vmatrix}$$

En appliquant la formule de décomposition de la variance, nous obtenons :

$$V^{S_\ell}(\hat{Y}_\pi^{S_\ell}) = V[E(\hat{Y}_\pi^{S_\ell}|S_r)] + E[V(\hat{Y}_\pi^{S_\ell}|S_r)]$$

$$= V\left[E\left(\sum_{k \in S_r} \frac{\hat{y}_k}{\alpha_{kr}\pi_{Ik}} \middle| S_r\right)\right] + E\left[V\left(\sum_{u_i \in S_I} \frac{\hat{y}_{u_i \cap G_r}}{\alpha_{ir}\pi_{Ii}} \middle| S_r\right)\right]$$

$$= V\left[\sum_{k \in S_r} \frac{\overbrace{E(\hat{y}_k|S_r)}^{y_k}}{\alpha_{kr}\pi_{Ik}}\right] + E\left[\sum_{u_i \in S_I} \frac{\overbrace{V(\hat{y}_{u_i \cap G_r}|S_r)}^{\text{car le tirage des logements se fait indépendamment d'une ZAE à l'autre}}}{(\alpha_{ir}\pi_{Ii})^2}\right]$$

$$= \underbrace{V\left[\hat{Y}_\pi^{S_r}\right]}_{V_{GR}+V_{EM}} + \underbrace{E\left[\sum_{u_i \in S_I} \frac{V(\hat{y}_{u_i \cap G_r}|S_r)}{(\alpha_{ir}\pi_{Ii})^2}\right]}_{V_{logement}}$$

Pour estimer cette variance, il convient donc d'une part d'estimer à partir de l'échantillon de logements la variance de $\hat{Y}_{\pi}^{S_r}$ – variance que l'on sait déjà estimer sur l'échantillon des communes S_r via la formule exposée au paragraphe 2.2 et qui se décompose en une variance V_{GR} liée à la constitution des groupes de rotation du recensement et une variance V_{EM} liée à la sélection des ZAE de l'échantillon-maître –, et d'autre part d'estimer le second terme correspondant à la variance liée au tirage des logements au sein de S_r.

- Sous l'hypothèse simplificatrice ❷, le tirage des logements au sein de la fraction recensée lors de la dernière EAR de chaque ZAE de l'échantillon-maître est assimilé à un tirage à probabilités inégales[11], et sa variance estimée sans biais à l'aide de la formule proposée par Deville dans [5] page 7 :

$$\hat{V}(\hat{y}_{u_i \cap G_r} | S_r) = \frac{n_{ir}}{n_{ir}-1} \sum_{\ell \in S_\ell \cap u_i} (1-\pi_{\ell|G_r,u_i}) \left(\frac{y_\ell}{\pi_{\ell|G_r,u_i}} - \frac{\sum_{\ell \in S_\ell \cap u_i} (1-\pi_{\ell|G_r,u_i}) \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}{\sum_{\ell \in S_\ell \cap u_i} (1-\pi_{\ell|G_r,u_i})} \right)^2$$

et $V_{logement}$ s'estime donc sans biais par :

$$\hat{V}_{logement}^{S_\ell} = \sum_{u_i \in S_I} \frac{n_{ir}}{n_{ir}-1} \sum_{\ell \in S_\ell \cap u_i} \frac{(1-\pi_{\ell|G_r,u_i})}{(\alpha_{ir}\pi_{Ii})^2} \left(\frac{y_\ell}{\pi_{\ell|G_r,u_i}} - \frac{\sum_{\ell \in S_\ell \cap u_i} (1-\pi_{\ell|G_r,u_i}) \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}{\sum_{\ell \in S_\ell \cap u_i} (1-\pi_{\ell|G_r,u_i})} \right)^2$$

- Pour l'estimation de la variance liée à la sélection des ZAE de l'échantillon-maître à partir de l'échantillon de logements, conditionnellement à la constitution des groupes de rotation du recensement, nous sommes dans le contexte d'un plan de sondage à deux degrés, et nous pouvons donc appliquer la propriété suivante, relative aux tirages à deux degrés avec tirages des unités secondaires indépendants entre les unités primaires, et démontrée par Ardilly dans [6], page 648 :

> Soit un plan de sondage à deux degrés, au sein duquel d'une part le plan de sondage conduit à un échantillon S_{UP} de m unités primaires et à des estimateurs sans biais $\hat{T}_i$ des vrais totaux T_i par unités primaires, et d'autre part les unités secondaires sont tirées indépendamment d'une unité primaire à l'autre. Si l'on dispose d'une forme quadratique $Q(T_1,...,T_m) = \sum_{i \in S_{UP}} q_i T_i^2 + \sum_{(i,j) \in S_{UP},\, i \neq j} q_{ij} T_i T_j$ permettant d'estimer sans biais sur l'échantillon S_{UP} la variance relative au premier degré de sondage en fonction des vrais totaux T_i par unité primaire, alors on montre que :
>
> $$E[Q(\hat{T}_1,...,\hat{T}_m)] = Q(T_1,...,T_m) + \sum_{i \in S_{UP}} q_i V(\hat{T}_i)$$
>
> Par conséquent, si l'on dispose d'un estimateur sans biais de la variance de $\hat{T}_i$ liée au

[11] En effet, malgré le rééchantillonnage effectué par Octopusse, la probabilité de tirage n'est pas rigoureusement constante au sein des ZAEGC.

second degré de sondage au sein de l'unité primaire i, on peut estimer sans biais à partir de l'échantillon final S_{US} la variance V_{UP} relative au premier degré de sondage grâce à la formule suivante :

$$\hat{V}_{UP}^{S_{US}}(\hat{T}) = Q(\hat{T}_1,...,\hat{T}_m) - \sum_{i\in S_{UP}} q_i \hat{V}(\hat{T}_i),$$

Pour l'estimation de V_{EM} à partir de l'échantillon S_ℓ, conditionnellement à la constitution des groupes de rotation du recensement, on a $S_{UP}=S_I$, $S_{US}=S_\ell$, $Q=Q_{EM}$,

$$T_i = \frac{\tilde{Y}_{ir}}{\pi_{Ii}} = \sum_{k\in u_i} \frac{y_k 1\!\!1_{k\in G_r}}{\alpha_{kr}\pi_{Ii}} \quad \text{et} \quad \hat{T}_i = \sum_{k\in u_i} \frac{\sum_{\ell\in S_\ell\cap k} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}{\alpha_{kr}\pi_{Ii}} = \frac{1}{\alpha_{ir}\pi_{Ii}} \underbrace{\sum_{\ell\in S_\ell\cap u_i} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}_{\hat{y}_{u_i\cap G_r}} = \frac{\hat{y}_{u_i\cap G_r}}{\alpha_{ir}\pi_{Ii}}.$$ En

notant q_i^{EM} le coefficient diagonal[12] de la forme quadratique Q_{EM} associé à la ZAE u_i, la composante V_{EM} s'estime donc sans biais à partir de l'échantillon par :

$$\hat{V}_{EM}^{S_\ell} = \underbrace{-\frac{1}{2}\sum_{\substack{u_i,u_j\in S_I \\ u_i\neq u_j}} \frac{\hat{\pi}_{Iij}-\pi_{Ii}\pi_{Ij}}{\hat{\pi}_{Iij}} \left(\frac{\hat{y}_{u_i\cap G_r}}{\alpha_{ir}\pi_{Ii}} - \frac{\hat{y}_{u_j\cap G_r}}{\alpha_{jr}\pi_{Ij}}\right)^2}_{Q_{EM}(\hat{T}_1,...,\hat{T}_m)} - \sum_{u_i\in S_I} q_i^{EM} \underbrace{\hat{V}\left(\frac{\hat{y}_{u_i\cap G_r}}{\alpha_{ir}\pi_{Ii}}\right)}_{\hat{V}_{logement}^{S_\ell}}$$

- Enfin, pour l'estimation de la variance liée à la constitution des groupes de rotation du recensement à partir de l'échantillon de logements, nous nous heurtons au problème suivant : la sélection des groupes de rotation du recensement constitue une phase et non pas un degré de sondage supplémentaire. En effet, les ZAE et les logements ne sont pas tirés de manière indépendante au sein des groupes de rotation du recensement, mais conditionnellement à ceux-ci. Faute de pouvoir appliquer la propriété utilisée au point précédent pour estimer V_{EM} à partir de S_l, nous allons procéder à l'hypothèse simplificatrice ❸. La variance liée à la constitution des groupes de rotation du recensement sera donc estimée à partir de l'échantillon de logements par « plug-in direct », en remplaçant directement dans la forme quadratique Q_{GR} les y_k inconnus par les $\hat{y}_k$ estimés à partir de l'échantillon de logement au sein de la commune k :

$$\hat{V}_{GR}^{S_l} = \underbrace{-\frac{1}{2}\sum_{\substack{k,l\in S_r \\ k\neq l}} \frac{\hat{\alpha}_{kl,r}-\alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}} \underbrace{\hat{\pi}_{Iij}}_{\substack{(i,j)/k\in u_i \\ l\in u_j}} \left(\frac{\hat{y}_k}{\alpha_{kr}} - \frac{\hat{y}_l}{\alpha_{lr}}\right)^2 + \sum_{k\in S_r} \frac{\sum_{l\in U}\hat{\alpha}_{kl,r} - \alpha_{kr}\sum_{l\in U}\alpha_{l,r}}{\alpha_{kr}\underbrace{\pi_{Ii}}_{i/k\in u_i}} \left(\frac{\hat{y}_k}{\alpha_{kr}}\right)^2}_{Q_{GR}(\hat{y}_1,...,\hat{y}_{n_c})}$$

Au final, la variance de l'estimateur du total d'une variable Y à partir d'un échantillon de logements issu d'Octopusse peut donc être estimée via la formule suivante :

[12] Ce coefficient vaut $\sum_{\substack{u_j\in S_I \\ u_j\neq u_i}} \frac{\pi_{Ii}\pi_{Ij}-\hat{\pi}_{Iij}}{\hat{\pi}_{Iij}}$

$$\hat{V}(\hat{Y}_w^{S_\ell}) = Q_L(y_1,...,y_L) = -\frac{1}{2}\sum_{\substack{k,l\in S_r \\ k\neq l}} \frac{\hat{\alpha}_{kl,r} - \alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}\ \underbrace{\hat{\pi}_{Iij}}_{\substack{(i,j)/k\in u_i \\ l\in u_j}}} \left(\frac{\hat{y}_k}{\alpha_{kr}} - \frac{\hat{y}_l}{\alpha_{lr}}\right)^2 + \sum_{k\in S_r} \frac{\sum_{l\in U}\hat{\alpha}_{kl,r} - \alpha_{kr}\sum_{l\in U}\alpha_{l,r}}{\alpha_{kr}\ \underbrace{\pi_{Ii}}_{i/k\in u_i}} \left(\frac{\hat{y}_k}{\alpha_{kr}}\right)^2$$

$$-\frac{1}{2}\sum_{\substack{u_i,u_j\in S_I \\ u_i\neq u_j}} \frac{\hat{\pi}_{Iij} - \pi_{Ii}\pi_{Ij}}{\hat{\pi}_{Iij}} \left(\frac{\hat{y}_{u_i\cap G_r}}{\alpha_{ir}\pi_{Ii}} - \frac{\hat{y}_{u_j\cap G_r}}{\alpha_{jr}\pi_{Ij}}\right)^2$$

$$+\sum_{u_i\in S_I}\left(1-q_i^{EM}\right)\frac{n_{ir}}{n_{ir}-1}\frac{1}{(\alpha_{ir}\pi_{Ii})^2}\sum_{\ell\in S_\ell\cap u_i}(1-\pi_{\ell|G_r,u_i})\left(\frac{y_\ell}{\pi_{\ell|G_r,u_i}} - \frac{\sum_{\ell\in S_\ell\cap u_i}(1-\pi_{\ell|G_r,u_i})\frac{y_\ell}{\pi_{\ell|G_r,u_i}}}{\sum_{\ell\in S_\ell\cap u_i}(1-\pi_{\ell|G_r,u_i})}\right)^2$$

avec $\hat{y}_k = \sum_{\ell\in S_\ell\cap k}\frac{y_\ell}{\pi_{\ell|G_r,u_i}}$ et $\hat{y}_{u_i\cap G_r} = \sum_{\ell\in S_\ell\cap u_i}\frac{y_\ell}{\pi_{\ell|G_r,u_i}}$.

2.4. Prise en compte de la non-réponse et du calage.

Dans les enquêtes auprès des ménages menées par l'Insee, la non-réponse totale est systématiquement corrigée par repondération, soit au travers d'une procédure spécifique de correction de la non-réponse – méthode des groupes de réponse homogène par exemple –, soit par calage direct de l'échantillon de répondants. En notant R l'échantillon de répondants et $\hat{p}_\ell$ la probabilité de réponse issue de la procédure de correction de la non-réponse pour le logement ℓ, l'estimateur corrigé de la non-réponse est l'estimateur suivant :

$$\hat{Y}_{CNR}^R = \sum_{\ell\in R}\frac{w_\ell}{\hat{p}_\ell}y_\ell$$

Dans le cadre du calcul de variance, la non-réponse va être modélisée par un mécanisme poissonnien : la réponse est indépendante entre les logements conditionnellement à l'échantillon de l'enquête. On peut donc voir la phase de non-réponse comme un degré de sondage supplémentaire – au sein de chaque logement de l'échantillon, tirage bernoullien de zéro ou une unité selon la probabilité de tirage $\hat{p}_l$, et tirages indépendants entre les logements – et s'appuyer sur la formule d'estimation de variance de Rao (cf. [5], page 40) :

Formule de Rao

> Soit un plan de sondage à deux degrés, au sein duquel d'une part le plan de sondage conduit à un échantillon S_{UP} de m unités primaires sélectionnées selon des probabilités d'inclusion γ_i et à des estimateurs sans biais $\hat{T}_i$ des vrais totaux T_i par unités primaires, et d'autre part les unités secondaires sont tirées indépendamment d'une unité primaire à l'autre. Si l'on dispose d'une forme quadratique $Q(T_1,...,T_m) = \sum_{i\in S_{UP}} q_i T_i^2 + \sum_{(i,j)\in S_{UP},\, i\neq j} q_{ij} T_i T_j$ permettant d'estimer sans biais sur l'échantillon S_{UP} la variance relative au premier degré de sondage en fonction des vrais totaux T_i par unité primaire, ainsi que d'estimateurs sans

biais des variances des $\hat{T}_i$ liées au second degré de sondage au sein de l'unité primaire i, on peut estimer sans biais à partir de l'échantillon final S_{US} la variance de l'estimateur d'Horvitz-Thompson $\hat{T}$ du total T par :

$$\hat{V}^{S_{US}}(\hat{T}) = Q(\hat{T}_1,..., \hat{T}_m) + \sum_{i \in S_{UP}} \left(\frac{1}{\gamma_i^2} - q_i \right) \hat{V}(\hat{T}_i)$$

Dans notre contexte, $S_{UP}=S_\ell$, $S_{US}=R$, Q est la forme quadratique Q^L associée à la formule d'estimation de variance $\hat{V}^{S_\ell}(\hat{Y}_w^{S_\ell})$ obtenue au paragraphe précédent, $T_\ell = y_\ell$, $\hat{T}_\ell = \frac{y_\ell \mathbb{1}_{\ell \in R}}{\hat{p}_\ell}$ et $\gamma_\ell = \frac{1}{w_\ell}$. Par ailleurs, comme, au sein d'un logement ℓ donné, la sélection de ce logement comme répondant ou non-répondant s'effectue par tirage bernoullien selon la probabilité de tirage $\hat{p}_\ell$, on a $\hat{V}\left(\hat{T}_\ell = \frac{y_\ell \mathbb{1}_{\ell \in R}}{\hat{p}_\ell} \right) = (1-\hat{p}_\ell)\left(\frac{y_\ell \mathbb{1}_{\ell \in R}}{\hat{p}_\ell} \right)^2$. Ainsi, en notant q_ℓ^L le coefficient diagonal de la forme quadratique Q_L associé au logement ℓ, la variance de l'estimateur $\hat{Y}_{CNR}^R$ s'estime donc à partir de l'échantillon de répondants R par :

$$\hat{V}(\hat{Y}_{CNR}^R) = Q_L\left(\frac{y_1 \mathbb{1}_{1 \in R}}{\hat{p}_1},..., \frac{y_L \mathbb{1}_{L \in R}}{\hat{p}_L} \right) + \sum_{\ell \in R} (w_\ell^2 - q_\ell^L)(1-\hat{p}_\ell)\left(\frac{y_\ell}{\hat{p}_\ell} \right)^2$$

Enfin, le calage usuellement mis en œuvre dans les enquêtes auprès des ménages sur l'échantillon de logements répondants est pris en compte de manière usuelle, en faisant porter le calcul de variance non pas sur la variable Y elle-même mais sur les résidus de sa régression, pondérée par les poids adéquats – poids calés en cas de calage direct de l'échantillon de répondants, poids corrigés de la non-réponse lorsque le calage intervient après une étape spécifique de correction de la non-réponse – sur les variables de calage. En notant ε_ℓ le résidu de cette régression pour le logement ℓ, la variance de l'estimateur final $\hat{Y}_{calé}^R$ s'estime donc à partir de l'échantillon de répondants R par :

$$\hat{V}(\hat{Y}_{calé}^R) = Q_L\left(\frac{\varepsilon_1 \mathbb{1}_{1 \in R}}{\hat{p}_1},..., \frac{\varepsilon_L \mathbb{1}_{L \in R}}{\hat{p}_L} \right) + \sum_{\ell \in R} (w_\ell^2 - q_\ell^L)(1-\hat{p}_\ell)\left(\frac{\varepsilon_\ell}{\hat{p}_\ell} \right)^2$$

Bibliographie

[1] Christine M. et Faivre S. (2009), Octopusse : un système d'Échantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement, *Actes des Journées de Méthodologie Statistique de 2009.*

[2] Chauvet G. (2011), On variance estimation for the French master sample, *Journal of Official Statistics, Vol. 27, No. 4, 2011, pp. 651–668.*

[3] Deville J.C. et Tillé Y. (2005), Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference, No. 128, pp. 569 – 591*.

[4] Breidt F.J. et Chauvet G. (2011), Improved variance estimation for balanced samples drawn via the cube method, *Journal of Statistical Planning and Inference, No. 141, pp. 479–487.*

[5] Caron N., Deville J.C. et Sautory O. (1998), Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE, *document de travail Insee M9806,* Insee.

[6] Ardilly P. (2006), Les techniques de sondage, *Éditions Technip*.

Partie I
méthodes et notations

- notations :
  - s désigne un échantillon. L'échantillon s est considéré indifféremment comme un sous-ensemble de l'univers $\mathscr{P}$, comme un vecteur de coordonnées dans $\{0,1\}$ ou comme une famille d'indicatrices sur les unités de l'univers : $s = (i \longmapsto s_i \in \{0,1\})$. De plus, toujours pour alléger les notations, le même symbole représente la variable aléatoire de valeur égale à l'échantillon tiré.
  - $\pi = \pi_1$ est la probabilité d'inclusion simple : $\pi_1(i) = p(i \in s)$.
  - π_2 est la probabilité d'inclusion double : $\pi_2(i,j) = p\{\{i,j\} \subset s\}$.

$$\Delta(i,j) = \pi_2(i,j) - \pi_1(i)\pi_1(j) = Cov(s_i, s_j)$$

$$\widehat{\Delta}(i,j) = \widehat{\pi}_2(i,j) - \pi_1(i)\pi_1(j)$$

$$\widehat{\widehat{\Delta}}(i,j) = \frac{\widehat{\pi}_2(i,j) - \pi_1(i)\pi_1(j)}{\widehat{\pi}_2(i,j)}$$

  - L'ensemble des échantillons simulés est noté S. Celui des échantillons répliqués est dénommé R.
  - $\widehat{\ }^s$ désigne un estimateur fonction de l'échantillon s [1].
  - $\widehat{Y} = \widehat{Y}^s$ est ici l'estimateur Horvitz-Thompson du total Y de la variable d'intérêt y.
  - Le produit terme à terme de deux matrices carrées ou d'une matrice et d'un vecteur colonne ou ligne approprié est distingué du produit matriciel par l'opérateur # (cf SAS/IML). Cette distinction est parfois omise pour alléger les notations. Par exemple, pour un vecteur x, x^2 désigne le vecteur dont les coordonnées sont les carrés de celles de x.
  - Pour un ensemble A :
    - $|A|$ désigne son cardinal.
    - $\sum_A f$ ou $\sum f(A)$ signifie $\sum_{a \in A} f(a)$. En particulier, $\sum A = \sum_{a \in A} a$.
    - x^A est une famille de vecteurs ou une matrice indicée en colonne par A
  - x désigne en général le vecteur d'équilibrage et $|x|$ est la dimension de l'espace d'équilibrage ($|x| = \text{rang}\left\{\left(\frac{x}{\pi}\right)^{\mathscr{P}}\right\}$)

- Pour limiter la confusion avec les réplications utilisées pour estimer les probabilités d'inclusion dans l'estimateur de variance effectif, fonction des valeurs de la variable observées sur un seul échantillon, cette étude adopte la convention de réserver le terme de simulations aux échantillons tirés dans la phase de validation pour estimer directement $\text{Var}\left(\widehat{Y}\right)$ et pour moyenner les estimations de variance sur ces échantillons, pour un calcul du biais.

I.1 estimateurs par simulation de la probabilité d'inclusion et de la variance

- La méthode de Breidt et Chauvet [BC] estime par réplication les probabilités d'inclusion double d'un tirage équilibré. La propriété de base est que la variance de l'échantillon tiré par Cube

1. ie calculable avec les informations observées sur l'échantillon

est estimée sans biais par la somme d'une fonction des vecteurs générés par cet algorithme pondérée par des coefficients également calculés par la procédure, grâce à l'équation (**eI.1.1**).

$$\mathrm{Var}\,(s) = \Delta = E\left[\sum_t \underline{\lambda}_t \overline{\lambda}_t u_t u_t{}'\right] \qquad \textbf{(eI.1.1)}$$

(t indice l'itération de l'algorithme Cube ; $u_t \in \mathrm{Ker}\left\{\left(\frac{x}{\pi}\right)^{\mathscr{P}}\right\}$, où x est le vecteur d'équilibrage de l'échantillon s tiré dans la population $\mathscr{P}$ avec la probabilité d'inclusion $\pi = \pi_1$; $\underline{\lambda}_t, \overline{\lambda_t} > 0$)

L'absence de biais implique que cet estimateur prend en compte l'atterrissage de la procédure de tirage équilibré. Des réplications du tirage permettent d'obtenir une estimation convergente des probabilités d'inclusion.

– Avec cette méthode, la matrice des probabilités d'inclusion double est estimée sans biais par la formule (**eI.1.2**).

$$\widehat{\pi_2}^{BC,R} = \overline{\sum_t \underline{\lambda}_t \overline{\lambda}_t u_t u_t{}'}^{R} + \pi_1 \pi_1{}' \qquad \textbf{(eI.1.2)}$$

(où R désigne l'ensemble des réplications ; π_1 est le vecteur des probabilités d'inclusion simple du sondage ; $\widehat{\ }^R$ désigne un estimateur en fonction de l'ensemble des réplications ; $\bar{}^R$ est la moyenne sur l'ensemble R)

– L'utilisation de cette estimation de la probabilité d'inclusion dans les estimateurs sans biais de la variance (Horvitz-Thompson et Yates-Grundy) fournit des estimateurs asymptotiquement sans biais, lorsque le nombre de réplications tend vers l'infini.

– Le programme de calcul par réplication des probabilités d'inclusion double **MacrosCube** a été fourni par Guillaume Chauvet.

• Pour la mesure du biais, la variance d'un estimateur de total $\widehat{Y} = \sum_s \frac{y}{\pi_1} = \left(\frac{y}{\pi_1}\right)' s$, pour une variable (scalaire) de total connu, peut être estimée sur un ensemble S de simulations de l'échantillonnage par la moyenne suivante :

$$\begin{aligned}
\overline{\mathrm{Var}\left(\widehat{Y}\right)}^{S} &= \frac{\sum_{s\in S}\left(\widehat{Y}^s - Y\right)^2}{|S|} = \frac{\sum_{s\in S}\left(\left(\frac{y}{\pi_1}\right)' s - \left(\frac{y}{\pi_1}\right)' \pi_1\right)^2}{|S|} \\
&= \frac{\sum_{s\in S}\left(\left(\frac{y}{\pi_1}\right)'(s-\pi_1)\right)^2}{|S|} = \frac{\sum_{s\in S}\left(\frac{y}{\pi_1}\right)'(s-\pi_1)(s-\pi_1)'\left(\frac{y}{\pi_1}\right)}{|S|} \\
&= \left(\frac{y}{\pi_1}\right)' \frac{\sum_{s\in S}(s-\pi_1)(s-\pi_1)'}{|S|}\left(\frac{y}{\pi_1}\right)
\end{aligned}$$

→ Il apparait ainsi un estimateur sans biais de la variance du vecteur-échantillon s sur les simulations [2] :

$$\overline{\mathrm{Var}\,(s)}^{S,f,1} = \frac{\sum_S (s-\pi_1)(s-\pi_1)'}{|S|} \qquad \textbf{(eI.1.3)}$$

2. L'exposant f se réfère à l'estimation de la probabilité d'inclusion double par la fréquence des couples sur les simulations, qui apparait dans l'expression de cet estimateur de la variance de l'échantillon.

Par conséquent, une autre expression de l'estimation sur les simulations de la variance est :

$$\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S} = \left(\frac{y}{\pi_1}\right)' \widehat{\mathrm{Var}\,(s)}^{S,f,1} \left(\frac{y}{\pi_1}\right)$$

Cette relation entre l'estimateur de variance d'un total estimé et celui de l'échantillon peut se déduire directement de l'expression de la variance du total estimé, qui est fonction linéaire de s :

$$\mathrm{Var}\left(\widehat{Y}\right) = \mathrm{Var}\left[\left(\frac{y}{\pi_1}\right)' s\right] = \left(\frac{y}{\pi_1}\right)' \mathrm{Var}\,(s) \left(\frac{y}{\pi_1}\right)$$

– Comme $E(s) = \pi_1$ et $E(ss') = \pi_2$, le premier estimateur par simulation de la variance de l'échantillon (**eI.1.3**) peut s'exprimer ainsi :

$$\widehat{\mathrm{Var}\,(s)}^{S,f,1} = \widehat{\pi}_2^{S,f} - \pi_1 \widehat{\pi}_1^{S,f\prime} - \widehat{\pi}_1^{S,f} \pi_1' + \pi_1 \pi_1' \qquad \textbf{(eI.1.4)}$$

$$\text{où :} \quad \widehat{\pi}_2^{S,f} = \overline{ss'}^{S} = \frac{\sum_S ss'}{|S|} \qquad \textbf{(eI.1.5)}$$

$$\widehat{\pi}_1^{S,f} = \bar{s}^{S} = \frac{\sum_S s}{|S|} \qquad \textbf{(eI.1.6)}$$

• L'expression (**eI.1.4**) présente l'intérêt d'utiliser l'estimateur des probabilités d'inclusion double (**eI.1.5**). D'une part, le calcul de celui-ci est plus aisé que celui de (**eI.1.3**), qui utilise une matrice des échantillons simulés (S) et son stockage moins coûteux que celui de S, d'autre part l'estimateur de la probabilité d'inclusion utilisé sert à d'autres calculs décrits dans la suite.

$\Longrightarrow$ Comme $\mathrm{Var}(s) = E(ss') - \pi_1 \pi_1'$, une première alternative à l'estimateur par simulation de la variance (**eI.1.3**) est donnée par (**eI.1.7**) [3].

$$\widehat{\mathrm{Var}\,(s)}^{S,f,2} = \overline{ss'}^{S} - \pi_1 \pi_1' = \widehat{\pi}_2^{S,f} - \pi_1 \pi_1' \qquad \textbf{(eI.1.7)}$$

3. L'estimateur (**eI.1.4**) est plus précis que (**eI.1.7**) lorsque $\mathrm{Var}\left\{(s-\pi_1)(s-\pi_1)'\right\} \leqslant \mathrm{Var}(ss')$, au sens où $\forall a \in \mathbb{R}^{\mathscr{P}}, \mathrm{Var}\left\{(a'(s-\pi_1))^2\right\} \leqslant \mathrm{Var}\left[(a's)^2\right]$. C'est le cas pour $a's \rightsquigarrow \mathcal{N}(m, \sigma^2)$. En effet, $\mathrm{Var}\left[\mathcal{N}(m,\sigma^2)^2\right] = 2\sigma^4 + 4\sigma^4 m^2$ alors que $\mathrm{Var}\left[\mathcal{N}(0,\sigma^2)^2\right] = 2\sigma^4$(voir note suivante). Plus généralement, si $E\left\{(x - E(x))^3\right\} = 0$ alors la fonction convexe $\tau \longmapsto \mathrm{Var}\left[(x-\tau)^2\right]$ admet un minimum en $\tau = E(x)$. Une condition nécessaire et suffisante dans le cas où $E(x) > 0$: $\mathrm{Var}\left\{(x - E(x))^2\right\} \leqslant \mathrm{Var}(x^2) \Longleftrightarrow \dfrac{Cov(x, x^2)}{\mathrm{Var}(x)} \geqslant E(x)$.

note : Si $\sqrt{|S|}\left(\dfrac{\sum_S(\widehat{Y}-Y)^2}{|S|}-\mathrm{Var}\left(\widehat{Y}\right)\right) \cong \mathcal{N}\left\{0,\mathrm{Var}\left[(\widehat{Y}-Y)^2\right]\right\}$ et que $\widehat{Y} \rightsquigarrow \mathcal{N}\left[Y,\mathrm{Var}\left(\widehat{Y}\right)\right]$ alors $\mathrm{Var}\left\{\dfrac{\left(\widehat{Y}-Y\right)^2}{\mathrm{Var}\left(\widehat{Y}\right)}\right\} = 2$ [4] et l'intervalle de confiance à 95% pour la variance est donné approximativement par :

$$\sqrt{|S|}\left|\sum_S(\widehat{Y}-Y)^2 \Big/ |S| - \mathrm{Var}\left(\widehat{Y}\right)\right| \leqslant 2\sqrt{2}\mathrm{Var}\left(\widehat{Y}\right)$$

Pour estimer l'écart-type $\sigma\left(\widehat{Y}\right)$ à 1% près, il faut approximativement estimer $\mathrm{Var}\left(\widehat{Y}\right)$ à 2% près. Avec les approximations normales précédentes, cela donne la condition $2\sqrt{2}\Big/\sqrt{|S|} \leqslant 0.02$ soit $|S| \geqslant 20\,000$.

- Comme la méthode de Breidt-Chauvet estime directement $\mathrm{Var}(s) = \pi_2 - \pi_1\pi_1'$, une troisième alternative pour estimer la variance est donnée par la formule (**eI.1.8**), où $\widehat{\pi}_2^{BC,S}$ désigne la probabilité d'inclusion double estimée par cette méthode sur les simulations.

$$\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{BC,S} = \left(\frac{y}{\pi_1}\right)'\left(\widehat{\pi}_2^{BC,S} - \pi_1\pi_1'\right)\frac{y}{\pi_1} \qquad \textbf{(eI.1.8)}$$

→ En résumé, les trois estimateurs par simulation de la variance [5] se formulent ainsi :

$$\widehat{V}^{S,f,1} = \left(\frac{y}{\pi_1}\right)'\left(\widehat{\pi}_2^{S,f} - 2\pi_1\widehat{\pi}_1^{S,f\prime} + \pi_1\pi_1'\right)\frac{y}{\pi_1} \qquad \textbf{(eI.1.4)}$$

$$\widehat{V}^{S,f,2} = \left(\frac{y}{\pi_1}\right)'\left(\widehat{\pi}_2^{S,f} - \pi_1\pi_1'\right)\frac{y}{\pi_1} \qquad \textbf{(eI.1.7)}$$

$$\widehat{V}^{S,BC} = \left(\frac{y}{\pi_1}\right)'\left(\widehat{\pi}_2^{BC,S} - \pi_1\pi_1'\right)\frac{y}{\pi_1} \qquad \textbf{(eI.1.8)}$$

– Le terme 'écart-type simulé' désigne $\widehat{\sigma}^S\left(\widehat{Y}\right) = \sqrt{\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^S}$ [6].

- La qualité des estimateurs de variance ci-dessus dépend de celle des probabilités d'inclusion (et du type d'estimateur choisi). Dans ce document, cette qualité est mesurée par la distance entre la diagonale de la matrice des probabilités d'inclusion double estimées et la probabilité d'inclusion simple.

---

4. Cette égalité découle de la propriété $E\left\{\mathcal{N}(0,1)^{2k}\right\} = \dfrac{(2k)!}{2^k k!}$, qui entraîne que $\mathrm{Var}\left[\chi^2(1)\right] = 2$.

5. Cette expression est utilisée pour distinguer $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^S$ de la variance estimée en fonction de l'échantillon $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^s$.

6. Avec le choix (**eI.1.8**), comme $M = \widehat{\pi}_2^{BC,S} - \pi_1\pi_1{}' = \overline{\sum_t \lambda_t\overline{\lambda_t}u_t u_t{}'}^S \geqslant 0$, la variance simulée ainsi est théoriquement toujours positive, comme pour (**eI.1.4**). En pratique, $\mathbb{1}'M\mathbb{1}$ est légèrement négative pour 10 strates ZAE. Ceci s'explique sans doute par l'imprécision du calcul numérique. Cependant, ce problème numérique n'est observé que pour la première variable d'équilibrage du tirage de l'échantillon maître. Il semble donc négligeable.

L'estimation des probabilités d'inclusion simple n'intervient pas dans la variance estimée par Yates-Grundy. Néanmoins, il parait pertinent d'utiliser la distance $(\pi_1, \widehat{\pi_1})$ comme indicateur de qualité de $\widehat{\pi_2}$, parce que π_1 est une donnée connue de la matrice π_2.

rappel : pour limiter les confusions

- Par convention, le terme de réplication renvoie ici à l'estimation des probabilités d'inclusion double pour estimer la variance en fonction des données observées sur un échantillon ($\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}$).
- En revanche, les simulations sont utilisées pour la validation, d'une part pour estimer directement $\mathrm{Var}\left(\widehat{Y}\right)$, d'autre part pour estimer $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)$, ainsi que $E\left\{\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s} - \mathrm{Var}\left(\widehat{Y}\right)\right)^2\right\}$.
- $\widehat{\pi_2}^S$ renvoie à un estimateur de la probabilité d'inclusion par les simulations, alors que $\widehat{\pi_2}^R$ fait référence à un estimateur calculé sur les réplications.
- Cette distinction est commode, mais la confusion reste possible lorsque la même méthode d'estimation de la probabilité d'inclusion est utilisée pour estimer la variance en fonction de l'échantillon et pour mesurer la qualité de l'estimateur correspondant.

I.2 formulation des estimateurs de variance fonctions de l'échantillon

– Les formules générales des estimateurs de variance de Yates-Grundy et Horvitz-Thompson sont fournies ci-dessous, avec des probabilités d'inclusion doubles estimées. Dans la suite, elles sont appliquées avec les probabilités d'inclusion estimées par la méthode de Breidt-Chauvet ($\widehat{\pi}_2 = \widehat{\pi}_2^{BC}$). Les termes de $\widehat{\pi}_2 < 10^{-8}$ sont omis de leurs calculs.

• La formulation (**eI.2.9**) de l'estimateur de variance de Yates-Grundy généralisé tient compte de la variabilité éventuelle de la taille d'échantillon $|s| = s_+$. Elle est justifiée ci-dessous.

$$\widehat{V}^{YG} = -\frac{1}{2}\sum_{i,j\in s}\left(1-\frac{\pi_1(i)\,\pi_1(j)}{\widehat{\pi}_2(i,j)}\right)\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2+\sum_{i\in s}\left(\frac{y}{\pi_1}\right)^2(i)\frac{\widehat{\pi}_2(i,+)-\pi_1(i)\,\pi_1(+)}{\pi_1(i)} \qquad \textbf{(eI.2.9)}$$

(La totalisation sur $\widehat{\pi}_2$, symbolisée par +, est effectuée sur l'univers, pas sur l'échantillon.)

• Pour un sondage de taille fixe, le second terme est nul (pour $\widehat{\pi}_2 = \pi_2$), et l'estimateur de Yates-Grungy se réduit à (**eI.2.10**).

$$\widehat{V}^{YG} = -\frac{1}{2}\sum_{i,j\in s}\left(1-\frac{\pi_1(i)\,\pi_1(j)}{\widehat{\pi}_2(i,j)}\right)\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2 \qquad \textbf{(eI.2.10)}$$

• La formule générale de l'estimateur de variance d'Horvitz-Thompson est (**eI.2.11**).

$$\widehat{V}^{HT} = \sum_{i,j\in s}\left(1-\frac{\pi_1(i)\,\pi_1(j)}{\widehat{\pi}_2(i,j)}\right)\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j) \qquad \textbf{(eI.2.11)}$$

– Matriciellement, l'estimateur de variance d'Horvitz-Thompson est calculé par la formule (**eI.2.12**), où # désigne le produit terme à terme et le symbole du produit matriciel est omis [7].

$$\widehat{V}^{HT} = \left\{\frac{y}{\pi_1}[s,]\#(\widehat{\widehat{\Delta}}(s,s)\frac{y}{\pi_1}[s,])\right\}[+,] \qquad \textbf{(eI.2.12)}$$

– Le calcul matriciel de l'estimateur de variance de Yates-Grundy (généralisé) s'en déduit (**eI.2.13**).

$$\widehat{V}^{YG} = \widehat{V}^{HT} - \left(\widehat{\widehat{\Delta}}(s,s)\left(\frac{y}{\pi_1}\right)^2[s,]\right)[+,]+\left\{\left(\frac{\widehat{\pi}_2(s,+)}{\pi_1(s)}-\pi_1(+)\right)\#\left(\frac{y}{\pi_1}\right)^2[s,]\right\}[+,] \qquad \textbf{(eI.2.13)}$$

– La justification de la formule de l'estimateur de Yates-Grundy généralisé (**eI.2.9**) repose sur l'expression suivante de la variance :

$$\begin{aligned}
\mathrm{Var}\left(\widehat{Y}\right) &= \sum_{i,j}(\pi_2(i,j)-\pi_1(i)\,\pi_1(j))\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j)\\
&= \sum_{i,j}(\pi_2(i,j)-\pi_1(i)\,\pi_1(j))\left\{-\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2+\left(\frac{y}{\pi_1}\right)^2(i)+\left(\frac{y}{\pi_1}\right)^2(j)\right\}\frac{1}{2}\\
&= -\frac{1}{2}\sum_{i,j}(\pi_2(i,j)-\pi_1(i)\,\pi_1(j))\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2+\sum_i\sum_j(\pi_2(i,j)-\pi_1(i)\,\pi_1(j))\left(\frac{y}{\pi_1}\right)^2(i)\\
&= -\frac{1}{2}\sum_{i,j}(\pi_2(i,j)-\pi_1(i)\,\pi_1(j))\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2+\sum_i(\pi_2(i,+)-\pi_1(i)\,\pi_1(+))\left(\frac{y}{\pi_1}\right)^2(i)
\end{aligned}$$

7. La totalisation + porte ici sur l'échantillon, comme les matrices de cette expression sont resteintes à s.

– Les deux premiers termes de la formule (**eI.2.13**) se déduisent de l'égalité :

$$-\frac{1}{2}\sum_{i,j\in s}\frac{\pi_2(i,j)-\pi_1(i)\pi_1(j)}{\pi_2(i,j)}\left(\frac{y}{\pi_1}(i)-\frac{y}{\pi_1}(j)\right)^2$$
$$=\sum_{i,j\in s}\frac{\pi_2(i,j)-\pi_1(i)\pi_1(j)}{\pi_2(i,j)}\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j)-\sum_{i\in s}\left(\sum_{j\in s}\frac{\pi_2(i,j)-\pi_1(i)\pi_1(j)}{\pi_2(i,j)}\right)\left(\frac{y}{\pi_1}\right)^2(i)$$

qui reste vraie lorsque π_2 est remplacé par $\widehat{\pi}_2$, pourvu que cet estimateur soit une matrice symétrique

• Il est prévisible que l'estimateur de Yates-Grundy [8] surestime en moyenne la variance et que cette surestimation diminue en fonction du nombre de réplications. En effet, compte tenu de la convexité de la fonction $x \longmapsto \frac{1}{x}$, l'espérance de l'inverse de la probabilité estimée est supérieure à l'inverse de son espérance ($E\left(\frac{1}{\widehat{\pi}_2}\right) \geqslant \frac{1}{\pi_2}$) [9]. Vu la formule (**eI.2.9**), l'estimateur de Yates-Grundy est la somme d'une fonction croissante de ces inverses de probabilité d'inclusion et d'un terme estimé sans biais. Donc à nombre de réplications fixé et fini, il doit surestimer la variance, en moyenne sur l'ensemble des simulations utilisées, si le nombre de ces simulations est suffisant.

• Cet effet de convexité joue à la sous-estimation de la variance par l'estimateur d'Horvitz-Thompson. C'est un avantage de l'estimateur de Yates-Grundy (à probabilité d'inclusion estimée sans biais).

8. Cette formulation est approximative : il s'agit de l'approximation de cet estimateur avec les probabilités d'inclusion double estimées par réplication.

9. Ce raisonnement ignore le fait que $p(\widehat{\pi}_2 \leqslant 0) > 0$. Cette omission peut modifier le sens de l'erreur d'estimation pour certains plans de sondage.

– estimateur de Deville :

La formule de Deville (**eI.2.14**) approxime la variance d'un sondage de taille fixe et de probabilité d'inclusion π_1.

$$\widehat{V}^{D} = \sum_s (1-\pi_1)\frac{|s|}{|s|-1}\left(\frac{y}{\pi_1} - \frac{\sum_s (1-\pi_1)\frac{y}{\pi_1}}{\sum_s (1-\pi_1)}\right)^2 \qquad \textbf{(eI.2.14)}$$

$$= \frac{|s|}{|s|-1}\left\{\sum_s (1-\pi_1)\left(\frac{y}{\pi_1}\right)^2 - \frac{\left(\sum_s (1-\pi_1)\frac{y}{\pi_1}\right)^2}{\sum_s (1-\pi_1)}\right\}$$

– estimateur de Deville-Tillé : La formule de l'estimateur de Deville-Tillé (**eI.2.16**) approxime la variance d'un sondage équilibré sur le vecteur x. Un avantage est qu'il donne des valeurs positives.

$$\widehat{V}^{DT} = \sum_s c\left(\frac{y}{\pi} - \sum_s c\frac{yx'}{\pi^2}\left[\sum_s c\frac{xx'}{\pi^2}\right]^{-}\frac{x}{\pi}\right)^2 = \sum_s c\frac{y^2}{\pi^2} - \sum_s c\frac{yx'}{\pi^2}\left[\sum_s c\frac{xx'}{\pi^2}\right]^{-}\sum_s c\frac{xy}{\pi^2} \qquad \textbf{(eI.2.15)}$$

$$\text{avec } c = \frac{|s|}{|s|-|x|}(1-\pi) \qquad \textbf{(eI.2.16)}$$

note : L'approximation de Deville peut s'en déduire en remplaçant x par π dans cette expression, et $|x|$ par 1. Pour contourner un problème numérique de calcul du rang et de l'inverse en entrée de l'approximation de Deville-Tillé, les variables d'équilibrage ont été divisées par leur maximum sur l'univers.

I.3 mesure du biais d'un estimateur de variance

- L'objectif de la mesure du biais est d'estimer l'écart entre $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)$ et $\mathrm{Var}\left(\widehat{Y}\right)$.
- Au préalable, la variance d'échantillonnage d'une variable connue sur l'univers ($\mathrm{Var}\left(\widehat{Y}\right)$) est estimée par simulation. Ensuite, cette variance sert de référence pour comparer la qualité des estimateurs de variance calculés sur l'échantillon, du type $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}$. L'expression de celui-ci fait intervenir $\widehat{\mathrm{Var}\,(s)}^{R}$, mais de manière non linéaire. Ainsi, l'estimateur d'Horvitz-Thompson calculé avec les probabilités d'inclusion estimées par réplications peut se formuler $\left(\frac{y}{\pi_1}\#s\right)'\widehat{\mathrm{Var}\,(s)}^{R}\left(\frac{y}{\pi_1}\#s\right)$, où $\widehat{\mathrm{Var}\,(s)}^{R}$ est une fonction non linéaire de $\widehat{\mathrm{Var}\,(s)}^{R}$ et de s. C'est un argument pour penser que le nombre de réplications nécessaires pour calculer de manière suffisamment précise ces estimateurs de variance par échantillon est beaucoup plus élevé que celui pour estimer adéquatement par simulation la variance d'une variable connue sur la base de sondage ($\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}$).
- La qualité d'un estimateur de la variance calculable sur l'échantillon sera jugée dans un premier temps par sa proximité à cette variance de référence, pour des variables connues sur l'univers.

- Si $\mathrm{Var}\left(\widehat{Y}\right)$ peut manifestement être estimée avec suffisamment de précision pour la mesure du biais, le point délicat pour ce calcul est d'estimer $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)$. Le calcul de cette moyenne est décrit dans l'Annexe A. Il s'avère que pour certaines variables cette moyenne appliquée à l'estimateur Horvitz-Thompson de la variance est très sensible à celle des probabilités d'inclusion 'simulées'.
- Le biais de l'estimateur de variance est défini dans ce document par :

$$\sqrt{E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)} - \sqrt{\mathrm{Var}\left(\widehat{Y}\right)}$$

  - Le terme de biais relatif désigne la statistique [10] :

$$\frac{\sqrt{\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}}^{s} - \sqrt{\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}}}{\sqrt{\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}}}$$

- Cependant, la mesure du biais est discutable. Pour l'estimateur de variance d'Horvitz-Thompson, l'espérance (sur les échantillons simulés, conditionnellement aux réplications) de l'estimateur de variance peut s'écrire :

$$E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right) = E\left\{\sum_{i,j\in s}\widehat{\widehat{\Delta}}(i,j)\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j)\right\} = E\left\{\sum_{i,j\in s}\widehat{\widehat{\Delta}}(i,j)\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j)s_i s_j\right\}$$
$$= \sum_{i,j}\widehat{\widehat{\Delta}}(i,j)\frac{y}{\pi_1}(i)\frac{y}{\pi_1}(j)\pi_2(i,j)$$

Or si d'une part $\pi_2(i,j)$ est remplacé par l'estimation $\widehat{\pi}_2(i,j)$ utilisée pour $\widehat{\widehat{\Delta}}$ et d'autre part $\mathrm{Var}\left(\widehat{Y}\right)$ est estimé par la formule (**eI.1.8**), alors le biais estimé d'Horvitz-Thompson est systématiquement nul [11], vu que $\widehat{\widehat{\Delta}}\#\widehat{\pi}_2 = \widehat{\pi}_2 - \pi_1\pi_1'$, alors que le vrai biais n'est pas nul pour $\widehat{\pi}_2 \neq \pi_2$.

10. Plus généralement, $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}$ peut être remplacé par $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{S}\right)$.

11. C'est le cas également pour l'estimateur de Yates-Grundy généralisé.

Alternativement, si des simulations annexes sont utilisées pour mesurer $\pi_2(i,j)$ dans l'estimation de $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)$ alors le biais mesuré dépend de ces simulations en même temps que de $\widehat{\widehat{\Delta}}$. L'impact de l'incertitude de l'estimation $\widehat{\pi}_2^{\,S}$ sur le biais mesuré représente un enjeu réel compte tenu que pour le tirage du groupe de rotation de petites communes, le Tableau 1 suggère que $\widehat{\pi}_2^{\,S,f,240\,000}$ est moins précis que $\widehat{\pi}_2^{\,BC,50\,000}$ [12]. Pour tenter de discerner le biais de l'estimateur de variance de celui dû à l'estimation par simulation des probabilités d'inclusion, le nombre de simulations a été augmenté autant que possible. Ceci permet de profiter de la convergence vers le vrai biais en fonction du nombre de simulations.

I.4 mesure de la dispersion de l'estimateur de variance

- L'approche de la qualité de l'estimateur de variance par le biais (estimé) est complétée par une analyse de la dispersion, pour laquelle le nombre de simulations est beaucoup plus contraint.
- La moyenne des estimations de variance est utile pour juger de la qualité de l'estimation des probabilités d'inclusion, à condition que le nombre de simulations utilisées d'une part pour estimer $\mathrm{Var}\left(\widehat{Y}\right)$ et d'autre part pour moyenner les estimations de variance soit suffisant.
- En revanche, selon l'idée qu'un petit biais vaut mieux qu'une grande variance, elle ne suffit pas au choix entre les deux estimateurs de variance. La comparaison a d'abord été complétée par l'estimation de variance sur le premier échantillon simulé.
- Une approche plus globale s'intéresse à la variance (ou plus rigoureusement l'écart quadratique moyen) de la variance estimée. L'indicateur de dispersion retenu ici est formulé par (**eI.4.17**), où $\widehat{V}^{S}$ désigne l'estimation de $\mathrm{Var}\left(\widehat{Y}\right)$ sur les simulations par la formule (**eI.1.4**).

$$100\frac{\left\{\dfrac{\sum_{s\in S_d}\left(\widehat{V}^{s}-\widehat{V}^{S}\right)^2}{|S_d|}\right\}^{1/4}}{\sqrt{\widehat{V}^{S}}} \qquad \textbf{(eI.4.17)}$$

Les simulations S_d utilisées pour le calcul de cet indicateur diffèrent de celles pour le biais (S). Le nombre de simulations pour calculer la dispersion est limité à $10\,000$. La contrainte supplémentaire est que cet indicateur ne peut pas être formulé en fonction des probabilités d'inclusion simulées.

note : Si $\forall s \in S_d$, $\widehat{V}^{s} = 0$ alors l'indicateur de dispersion vaut 100%. Cette valeur représente donc un plafond d'acceptabilité des estimateurs de variance. Mais il peut être opportun d'accepter un estimateur qui le transgresse pour certaines variables, s'il est moins dispersé pour les variables d'intérêt.

12. Cette inférence est sujette à caution, en l'absence d'information sur les termes non diagonaux. Pour des probabilités d'inclusion très faibles, l'estimateur par la fréquence présente l'avantage d'assurer une estimation positive.

Partie II
probabilités d'inclusion double des petites communes du recensement et variance d'un groupe de rotation

Cette partie présente une estimation des probabilités d'inclusion double des petites communes (au sens du recensement, notées PC) de métropole dans un groupe de rotation du recensement. Avec les probabilités ainsi estimées, les qualités des estimateurs de variance d'Horvitz-Thompson et de Yates-Grundy sont comparées, en biais et en dispersion. Le biais apparent est également utilisé pour décider du nombre de réplications nécessaires pour estimer les probabilités d'inclusion. Ces éléments étayent le choix de la méthode d'estimation de la variance du tirage du groupe de rotation des petites communes recensées une année. C'est une composante de la variance des enquêtes tirées par Octopusse dans une seule campagne du recensement.

II.1 méthode d'estimation des probabilités d'inclusion des petites communes

• Pour l'objectif de calculer la variance du tirage par le recensement d'un groupe de rotation de petites communes, il s'agit ici d'estimer la probabilité d'inclusion double des petites communes dans un groupe de rotation (noté gr).

$$\pi_{2,\mathrm{gr}}(c_1, c_2) = proba\,\{\{c_1, c_2\} \subset \mathrm{gr}\}$$

– Le rang du groupe de rotation n'est pas distingué. L'hypothèse implicite de ce choix [13] est que les probabilités d'inclusion double des petites communes dans un groupe de rotation ne dépendent pas (sensiblement) de ce rang.

– La probabilité d'inclusion simple d'une petite commune dans un groupe de rotation vaut :

$$\pi_{1,\mathrm{gr}} = 1 / 5$$

• La méthode appliquée ici est celle de Breidt et Chauvet [BC]. Elle fournit un estimateur sans biais des probabilités d'inclusion double, calculé sur un ensemble de réplications d'un tirage équilibré. Ces probabilités permettent d'estimer la variance due au tirage du groupe de rotation des petites communes, composante de la précision des enquêtes tirées par Octopusse dans une seule campagne de recensement, suivant la méthode proposée par G Chauvet [GC].

• Les variables d'équilibrage du tirage du groupe de rotation des petites communes d'une région sont les suivantes (extrait de 'Pour comprendre le recensement de la population' http://www.insee.fr/fr/ppp/sommaire/imeths01e.pdf) :

13. Il est possible d'estimer des probabilités d'inclusion dans chacun des 5 groupes de rotation des petites communes.

le nombre de logements ;
le nombre de logements en immeuble collectif ;
la population des personnes de moins de 20 ans ;
la population des personnes de 20 à 39 ans ;
la population des personnes de 40 à 59 ans ;
la population des personnes de 60 à 74 ans ;
la population des personnes de 75 ans ou plus ;
la population des femmes ;
la population des hommes ;
et, pour chacun des départements, la population totale.

• La probabilité d'inclusion a été rajoutée en première position de la liste des variables d'équilibrage, afin de contrôler la taille en nombre de communes du groupe de rotation. Vraisemblablement, le tirage effectif des groupes de rotation du recensement a été réalisé ainsi.

– L'ordre de cette liste est supposé identique à celui pris en compte par l'atterrissage de Cube. Les populations départementales sont déclarées dans l'ordre des codes de département.

• La table des petites communes utilisée par le recensement pour tirer les groupes de rotation n'a pas été retrouvée. En substitution, une exploitation du recensement de 1999 a été fournie par le pôle EDL, avec les totaux des variables d'équilibrage ci-dessus par commune [14]. Les petites communes ont ensuite été sélectionnées dans cette table avec la variable de population livrée (population<10 000).

→ Cette table constitue le référentiel pour le calcul et le stockage des probabilités d'inclusion double des petites communes. Elle est utilisée pour simuler le tirage des petites communes.

– Sur l'ensemble des petites communes utilisé ici, l'exécution de 10 000 simulations dure entre 24 et 26 heures [15]. Cette durée de traitement contraint le nombre de simulations réalisables.

• Le programme utilisé pour estimer les probabilités d'inclusion des petites communes dans un groupe de rotation par Breidt-Chauvet est ci-joint **pi2_hat_pc**. Il stocke les probabilités d'inclusion double des petites communes sous la forme d'une matrice carrée des probabilités d'inclusion double $\widehat{\pi_{2,\mathrm{gr}}}$ par région, sans correction de la diagonale, dans une table SAS nommée `pi2_hat_pc_c_&region._430000`. Les identifiants communaux ne sont pas stockés. L'ordre des lignes et des colonnes correspond à celui de la table `pc_donnees_rp99`. Ce stockage est adapté à un calcul matriciel de l'estimateur de variance [16].

Pour la validation, le programme utilisé pour estimer par la fréquence les probabilités d'inclusion double est **pi2_hat_pcf**.

• Les résultats présentés portent sur des probabilités d'inclusion doubles des petites communes estimées sur 230 000 réplications par la méthode de Breidt-Chauvet. Pour évaluer la qualité des estimations de variance, une deuxième estimation de ces probabilités a été réalisée par un calcul direct de la fréquence des couples de communes mesurée sur 240 000 simulations. En outre, une troisième

14. La géographie communale de cette table est sans doute celle du RP 1999.

15. Selon une table du site insee.fr en géographie du 1/1/2011, le nombre de communes de population 1999 <10 000 est 35 698.

16. L'estimation alternative par la fréquence des couples dans les échantillons simulés est plus rapide. Toutefois 10 000 simulations durent plus de 17 heures.

17. L'alternative d'un stockage en ligne des coefficients d'estimation de la variance $\widehat{\widehat{\Delta}}(c_1,c_2) = 1 - \dfrac{0.2^2}{\widehat{\pi_{2,\mathrm{gr}}}(c_1,c_2)}$, pour tous les couples régionaux de petites communes, a été abandonnée. Un inconvénient de cette table alternative est sa taille de 2 Go. Elle s'explique notamment parce que tous les couples régionaux (ordonnés) sont stockés. Une simple tabulation du nombre de couples par région sur une telle table dure plus de 17 minutes.

estimation des probabilités d'inclusion par la méthode de Breidt-Chauvet sur un lot supplémentaire de 200 000 simulations a été également utilisée pour la validation.

II.2 qualité de l'estimation des probabilités d'inclusion des PC

– La diagonale de la matrice des probabilités d'inclusion double des petites communes estimée par la méthode de Breidt-Chauvet sur 230 000 réplications est très proche de la probabilité d'inclusion simple : l'écart est inférieur à 1% de celle-ci (Tableau 1). Il n'apparait pas de valeurs aberrantes des probabilités d'inclusion double estimées. En particulier, elles sont toutes positives. Cette propriété n'est pas vérifiée avec un nombre plus réduit de réplications, même pour un lot de 50 000.

Tableau 1 – Indicateurs de qualité des probabilités d'inclusion double des petites communes estimées sur 230 000 réplications

région	nb petites communes	$\min(\widehat{\pi_2})$	$\max(\widehat{\pi_2})$	$\max[\lvert\widehat{\pi_1}-\pi_1\rvert]$	$\lVert\widehat{\pi_1}-\pi_1\rVert$	taux d'erreur maximal 230 000	100 000	50 000	10 000	5 000	$\overline{ss'}^{240\,000}$
11 Île-de-France	1 042	3,5E-04	0,201	1,5E-03	0,0126	0,7	1,3	1,3	2,8	3,4	1,7
21 Champagne-Ardenne	1 931	8,9E-03	0,201	9,9E-04	0,0137	0,5	0,8	1,1	2,6	3,7	1,3
22 Picardie	2 271	9,8E-03	0,201	1,3E-03	0,0151	0,7	0,8	1,2	2,7	3,9	1,9
23 Haute-Normandie	1 395	1,1E-02	0,201	1,1E-03	0,0117	0,6	0,9	1,0	2,7	3,7	1,5
24 Centre	1 810	1,0E-02	0,201	1,1E-03	0,0137	0,5	1,0	1,4	3,2	3,7	1,4
25 Basse-Normandie	1 799	6,8E-04	0,201	1,1E-03	0,0133	0,6	1,1	1,5	2,7	3,6	1,3
26 Bourgogne	2 029	6,2E-03	0,201	1,2E-03	0,0143	0,6	1,1	1,4	2,8	3,9	1,9
31 Nord-Pas-de-Calais	1 466	7,1E-03	0,201	1,2E-03	0,0119	0,6	0,8	1,1	2,9	3,7	1,3
41 Lorraine	2 301	7,7E-03	0,201	1,1E-03	0,0147	0,6	0,9	1,4	2,8	3,6	1,5
42 Alsace	880	2,0E-02	0,201	1,2E-03	0,0091	0,6	0,9	1,2	2,9	3,7	1,4
43 Franche-Comté	1 775	2,9E-03	0,201	9,5E-04	0,0132	0,5	0,7	1,0	2,3	3,4	1,7
52 Pays de la Loire	1 468	3,1E-03	0,201	1,1E-03	0,0127	0,5	0,8	1,2	3,1	3,5	1,4
53 Bretagne	1 237	7,0E-03	0,201	1,1E-03	0,0114	0,6	0,8	1,5	2,5	4,3	1,4
54 Poitou-Charentes	1 452	3,0E-03	0,201	1,2E-03	0,0119	0,6	0,8	1,3	3,2	3,8	1,6
72 Aquitaine	2 252	6,7E-03	0,201	1,1E-03	0,0149	0,6	1,0	1,1	2,6	4,0	1,6
73 Midi-Pyrénées	2 990	4,6E-03	0,201	1,2E-03	0,0174	0,6	0,9	1,2	2,7	3,9	1,5
74 Limousin	741	4,1E-03	0,201	1,2E-03	0,0083	0,6	0,8	1,2	2,4	3,4	1,2
82 Rhône-Alpes	2 806	1,7E-02	0,201	1,4E-03	0,0168	0,7	0,8	1,3	3,0	3,9	1,5
83 Auvergne	1 296	7,7E-03	0,201	9,8E-04	0,0112	0,5	0,8	1,2	3,0	3,9	1,3
91 Languedoc-Roussillon	1 523	2,6E-04	0,201	1,0E-03	0,0120	0,5	0,8	1,2	3,2	3,6	1,3
93 Provence-Alpes-Côte d'Azur	887	1,2E-02	0,201	1,3E-03	0,0095	0,6	1,0	1,1	2,0	4,8	1,3
94 Corse	357	2,8E-04	0,201	8,6E-04	0,0057	0,4	0,8	1,1	2,2	3,5	1,4
total	35 708	2,6E-04	0,201	1,5E-03		0,7	1,3	1,5	3,2	4,8	1,9

Notes :

– *Les deux premières colonnes, après le nombre de petites communes, fournissent le minimum et le maximum de l'estimation des probabilités d'inclusion doubles, sur l'ensemble des petites communes. Les deux colonnes suivantes donnent d'une part le maximum de l'écart absolu entre la probabilité d'inclusion simple (égale à 0.2) et son estimation d'autre part la distance euclidienne entre ces deux variables. Les 6 dernières colonnes décrivent le taux d'erreur maximal en pourcentage de la probabilité d'inclusion simple, selon le nombre de réplications utilisées pour l'estimer.*

– *Avec* 430 000 *réplications, le taux d'erreur maximal de* $\widehat{\pi_1}$ *est de 0.6%.*

– Cependant, la qualité d'estimation de la diagonale n'est pas complètement représentative. Par exemple, la probabilité d'inclusion double minimale du Languedoc-Roussillon augmente de 54% entre 230 000 et 430 000 réplications.

II.3 mesure par simulations de la variance d'un groupe de rotation PC

→ Il ressort du Tableau 2 que le premier estimateur de la variance des totaux estimés des variables d'équilibrage par la fréquence (**eI.1.4**) est très proche de celui par la méthode de Breidt et Chauvet. Ce n'est pas le cas pour le deuxième estimateur par la fréquence (**eI.1.7**). De plus, celui-ci parait se stabiliser plus lentement que le premier.

Tableau 2 – Taux d'écart entre les différentes estimations de l'écart-type simulé, pour les variables d'équilibrage, en %

simulations	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75 ans	feminin	masculin
f 240 000/40 000	0,0		-0,1	0,0	0,0	-0,1	-0,1	0,0	0,0	0,0			
f alt 240 000/40 000	10,4		2,7	-8,7	-6,9	-6,2	2,0	2,2	-4,7	-6,1			
BC 230 000/50 000	0,1		0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0			
BC 230 000/f 240 000	0,1		0,2	0,0	0,1	0,0	0,1	0,0	0,0	0,1			
BC 230 000/f alt 240 000	-7,8		-3,0	10,0	8,4	9,7	3,8	0,7	7,4	9,4			

Notes :

- *f désigne l'estimateur* (**eI.1.4**) *et 'f alt' représente* (**eI.1.7**).
- *En première colonne figure les nombres de simulations utilisées pour estimer les écarts-types comparés.*
- *Les variables d'équilibrage sont utilisées pour la validation parce que d'une part elles sont connues sur l'univers, d'autre part leur variance est faible et donc sans doute particulièrement difficile à estimer. Mais leur utilisation pour le choix de l'estimateur de variance présente des limites (cf dispersion).*

→ Sauf mention contraire, la suite de cette partie sur les petites communes retient pour la validation le premier estimateur par simulation de la variance (**eI.1.4**). Cette version présente l'avantage d'être directement interprétable comme une moyenne des écarts quadratiques entre le total et son estimation, comme vu précédemment. De plus, l'indépendance par rapport à la méthode Breidt-Chauvet d'estimation des probabilités d'inclusion peut consolider l'évaluation de l'estimateur de variance calculé avec celles-ci.

- La comparaison de l'écart-type simulé sur deux lots indépendants de simulations d'un groupe de rotation de petites communes donne un taux d'écart absolu de moins de 2%, dès 1 000 simulations, pour toutes les variables d'équilibrage (Tableau 3).

Tableau 3 – Taux d'écart entre les écarts-types mesurés sur deux lots indépendants de simulations du tirage d'un groupe de rotation, en %

nb simulations	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75 ans	feminin	masculin
1 000	-1,5		-0,8	0,6	0,5	0,4	-0,2	-1,1	0,2	0,5			
2 000	0,9		1,2	0,1	-0,1	0,5	0,6	0,4	0,3	0,2			
5 000	0,5		0,5	0,0	0,2	0,1	0,0	0,2	0,1	0,1			

note : Le taux d'écart entre la moyenne des estimations de totaux et les vrais totaux est inférieur à 0.03% pour toutes les variables d'équilibrage dès 1 000 simulations. De plus, ce maximum est observé pour le nombre de logements collectifs. Pour les autres variables, le taux d'écart est dix fois plus bas. La convergence de la moyenne des totaux estimés est donc sensiblement plus rapide que celle des estimateurs des écarts-quadratiques.

$\Longrightarrow$ Vu l'écart relatif très faible entre les versions (**eI.1.4**) et (**eI.1.8**), ainsi que l'effet du nombre de simulations pour ces deux estimateurs négligeable au delà de $40\,000$, la mesure de la variance $\mathrm{Var}\left(\widehat{Y}\right)$ utilisée dans cette étude peut sans doute être considérée comme une référence fiable.

• Dans la suite de cette partie, la validation utilise comme référence la variance estimée par la formule (**eI.1.4**) sur les $240\,000$ simulations réalisées. La mesure de la dispersion nécessite le recours à des échantillons simulés stockés, dont le nombre est limité à $10\,000$.

II.4 précision estimée pour les variables d'équilibrage des PC

La qualité de l'estimation de variance est dans un premier temps évaluée par l'analyse de la proximité moyenne entre la variance estimée pour les variables d'équilibrage et la vraie variance évaluée sur les $240\,000$ groupes de rotation simulés.

– Le biais relatif de l'estimateur de Yates-Grundy devient négligeable lorsque les probabilités d'inclusion sont estimées sur au moins $100\,000$ réplications (Tableau 4). Il diminue en fonction du nombre de réplications, ce qui suggère que le biais est correctement estimé. Les régions dont le nombre de petites communes est un multiple de 5 (en grisé) ne se distinguent pas sur ce tableau, ce qui conforte la formule de l'estimateur de Yates-Grundy généralisé.

Tableau 4 – Biais relatif de l'estimateur Yates-Grundy, en % de l'écart-type simulé des variables d'équilibrage des petites communes

région	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75 ans	feminin	masculin
11	-0,3		-0,1	0,7	0,2	-0,3	-0,2	-0,2	-0,2	0,2	0,2		
21	-0,3		-0,2	-0,2	-0,2	-0,1	-0,1	-0,4	-0,2	-0,2			
22	0,2		-0,2	0,3	0,1	0,3	0,4	0,3	0,3	0,2			
23	0,5		0,0	0,3	0,1	0,4	0,4	0,5	0,4	0,3			
24	-0,2		0,2	-0,6	-0,2	-0,7	-0,3	0,4	-0,5	-0,5			
25	0,1		0,2	-0,8	-0,6	-0,7	-0,9	-0,8	-0,8	-0,8			
26	0,2		0,1	0,1	0,1	0,2	0,4	0,1	0,2	0,2			
31	0,0		0,0	-0,3	-0,3	-0,5	-0,5	-0,4	-0,5	-0,4			
41	0,1		0,1	-0,1	-0,1	0,0	-0,1	0,3	0,0	-0,1			
42	-0,7		-0,2	-0,6	-0,5	-0,5	-0,6	-0,7	-0,5	-0,6			
43	0,1		-0,2	0,2	0,3	0,2	-0,1	-0,2	0,1	0,1			
52	0,0		-0,3	0,2	0,1	0,1	-0,2	-0,2	0,0	0,0			
53	0,6		0,0	0,4	0,2	0,2	0,6	0,5	0,3	0,4			
54	-0,4		-0,2	-0,5	-0,4	-0,2	-0,2	-0,1	-0,3	-0,3			
72	0,2		0,0	0,2	0,2	0,3	0,4	0,4	0,3	0,2			
73	0,3		-0,3	0,0	0,0	0,1	0,5	0,5	0,2	0,1			
74	-0,2		0,0	-0,5	-0,5	-0,4	-0,3	-0,2	-0,4	-0,4			
82	0,2		-0,1	1,4	0,9	1,3	1,2	0,6	1,2	1,3			
83	-0,1		-0,2	-0,1	-0,3	0,1	0,1	0,2	-0,1	-0,1			
91	0,2		0,1	-0,1	0,0	-0,3	0,0	0,1	-0,1	-0,1			
93	-0,2		-0,3	-0,1	-0,2	-0,1	0,0	0,1	-0,1	-0,1			
94	-0,2		-0,5	-0,2	-0,2	-0,1	-0,1	-0,8	-0,6	-0,1			
430 000 réplications	0,1		0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0			
100 000	0,1		0,0	0,1	0,0	0,1	0,1	0,1	0,1	0,1			
50 000	0,1		0,0	0,3	0,2	0,3	0,3	0,3	0,3	0,3			
10 000	0,3		-0,1	1,6	1,2	1,6	1,2	1,2	1,4	1,5			
5 000	1,1		0,8	3,0	2,2	3,0	2,6	2,2	2,8	2,7			
BC 200 000 simulations ($\widehat{\pi_2}^{S,BC}$)	0,0		0,0	0,1	0,1	0,1	0,1	0,0	0,1	0,1			

Notes :

– *Il s'agit du taux d'écart entre la moyenne sur les échantillons des estimations de l'écart-type et l'estimation par simulation de celui-ci, en pourcentage. Plus précisément, le numérateur est la racine carrée de la moyenne des estimations de variance sur les échantillons simulés, y compris celles négatives, selon la formule* (**eA.34**)*. Au dénominateur, l'écart-type 'vrai' est estimé par* (**eI.1.4**) *sur les* 240 000 *simulations. Les programmes sont ci-joints* ***biais_relatif*** *+* ***biais_relatif_equilib_yg****.*

– *La dernière ligne est calculée avec une matrice* $\widehat{\pi}_2^{BC,S}$ *estimée par Breidt-Chauvet sur un lot supplémentaire de* 200 000 *simulations, pour à la fois moyenner le numérateur (ie les estimations de variance) et estimer le dénominateur (ie la vraie variance de l'estimateur de total).*

– *Les 4 lignes précédentes sont calculées avec les probabilités d'inclusion estimées sur le nombre de réplications qui figure dans la première colonne. Le dénominateur reste le même.*

– *Le biais relatif de l'estimateur de Yates-Grundy non généralisé est compris entre -22.6% et -37.0% pour ces variables, ce qui le disqualifie manifestement pour le tirage des petites communes.*

– La mesure du biais par des probabilités d'inclusion estimées par la méthode Breidt-Chauvet est globalement très proche de celle mesurée avec les fréquences des couples de communes sur les simulations. Cependant, les biais régionaux sont compris entre -0.5 et 0.6%, contre -0.9 et 1.4% avec la pondération par la fréquence. Ceci peut plaider pour le biais estimé par la méthode de Breidt-Chauvet, avec la meilleure proximité à la diagonale des probabilités d'inclusion.

• La comparaison avec l'estimateur d'Horvitz-Thompson est motivée d'abord par la variabilité de la taille de l'échantillon des petites communes. Mais la quasi absence de biais de l'estimateur de Yates-Grundy généralisé (Tableau 4) répond à ce besoin. La deuxième motivation est que la comparaison entre les deux estimateurs est un moyen indirect de contrôler que le nombre de réplications est suffisant pour une estimation correcte des probabilités d'inclusion. En effet, lorsque ces dernières sont parfaitement estimées, les biais des deux estimateurs sont nuls. Enfin, la comparaison théorique de la précision de ces deux estimateurs de variance parait hors de portée. C'est pourquoi une évaluation empirique est effectuée ici sur plusieurs lots de variables et à l'aide de différentes statistiques.

– L'estimateur d'Horvitz-Thompson semble mesurer en moyenne moins précisément que Yates-Grundy la variance simulée des variables d'équilibrage (Tableau 5). De manière surprenante, le biais estimé ne décroit pas systématiquement en fonction du nombre de réplications. La réduction du biais apparent entre 100 000 et 230 000 réplications utilisées pour les probabilités d'inclusion ne bénéficie qu'à trois variables, alors que l'accroissement de 50 000 à 100 000 semble encore bénéfique à la précision de l'estimation de variance.

Tableau 5 – Biais relatif d'Horvitz-Thompson,% de l'écart-type simulé des variables d'équilibrage

région		totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans 39	40 ans 59	60 ans 74	plus 75 ans	feminin	masculin
11	Île-de-France	-25,2		-46,6	-23,6	-9,6	-17,2	-11,5	-6,2	-24,0	-12,5
21	Champagne-Ardenne	-1,1		-1,2	-0,7	-1,1	0,0	-0,3	-0,9	-0,6	-0,8
22	Picardie	-1,0		-1,0	-0,3	-0,6	-0,2	0,0	-0,4	-0,3	-0,4
23	Haute-Normandie	-0,1		0,3	0,0	0,0	0,0	-0,1	0,3	0,0	0,0
24	Centre	-2,1		-0,8	-2,6	-2,5	-3,1	-1,7	0,1	-2,6	-2,9
25	Basse-Normandie	7,2		6,9	1,0	1,6	1,5	1,8	1,6	1,6	1,5
26	Bourgogne	-1,8		-1,6	-1,8	-2,1	-1,8	-1,3	-1,4	-1,8	-2,0
31	Nord-Pas-de-Calais	0,6		0,3	0,5	0,3	0,6	0,0	-0,3	0,2	0,5
41	Lorraine	-0,3		-0,1	-0,2	-0,1	0,0	0,1	0,2	0,1	-0,1
42	Alsace	-1,6		-0,2	-1,8	-1,2	-2,1	-1,8	-1,0	-1,7	-1,8
43	Franche-Comté	-1,2		-2,0	1,2	1,4	1,8	1,4	1,1	1,6	1,6
52	Pays de la Loire	-0,5		-1,0	0,4	0,3	0,2	-0,3	-0,7	0,0	0,2
53	Bretagne	0,0		0,0	1,5	1,0	1,0	0,4	0,4	1,0	1,1
54	Poitou-Charentes	0,7		1,0	-2,8	-2,5	-2,2	-0,8	-0,2	-2,1	-2,3
72	Aquitaine	-2,0		-1,2	-0,3	-0,8	-0,5	-1,2	-1,1	-0,9	-0,8
73	Midi-Pyrénées	2,2		3,0	-2,9	-3,5	-2,1	-0,6	0,2	-2,2	-2,6
74	Limousin	1,2		1,4	0,6	0,8	0,9	1,1	1,1	1,0	0,9
82	Rhône-Alpes	0,4		-0,3	4,1	3,4	3,8	3,4	2,8	3,9	3,9
83	Auvergne	-1,0		0,0	-0,6	-0,4	-0,8	-1,1	-1,4	-0,9	-0,8
91	Languedoc-Roussillon	0,5		1,2	-1,2	-0,8	-0,8	0,1	0,7	-0,4	-0,6
93	Provence-Alpes-Côte d'Azur	-0,5		-0,3	-0,8	-1,0	-0,8	-0,9	-1,3	-1,0	-0,9
94	Corse	5,6		4,4	4,7	5,4	3,8	4,9	4,2	4,8	5,2
total	(230 000 réplications)	0,3		0,8	-1,8	-1,4	-1,2	-0,5	-0,1	-1,3	-1,3
430 000	réplications	3,0		5,2	-1,0	-0,6	-0,5	0,1	0,2	-0,5	-0,5
100 000		0,8		1,9	-0,3	-0,2	-0,5	-0,4	-0,4	-0,4	-0,4
50 000		3,5		7,0	-0,8	-0,8	-0,8	-0,6	-0,8	-0,9	-0,9
10 000		10,1		14,2	-4,9	-2,9	-3,6	-0,6	0,0	-3,3	-3,3
5 000		8,3		16,5	-7,5	-5,0	-9,2	-5,2	-3,6	-7,2	-7,2
140 000	simulations	1,0		2,4	-1,7	-1,3	-1,1	-0,5	-0,2	-1,2	-1,2
40 000		1,2		3,2	-2,5	-1,9	-1,6	-0,8	-0,6	-1,7	-1,8
BC 200 000	simulations ($\widehat{\pi}_2^{S,BC}$)	-10,3		-17,3	-1,4	-1,5	-1,6	-1,7	-1,0	-1,6	-1,6

Notes :

– *Pour la ligne commençant par* '100 000 *réplications' et les trois suivantes, voir le tableau précédent.*

– *La ligne débutant par* '140 000 *simulations' et la suivante sont calculées avec des estimations de variance (dont les probabilités d'inclusion sont approximées sur* 230 000 *réplications) moyennées sur le nombre de simulations de la première colonne. La variance simulée reste estimée avec les probabilités d'inclusion mesurées par la fréquence observée sur toutes les* 240 000 *simulations.*

– *Si d'une part il n'y avait pas d'incertitude sur l'estimation des probabilités d'inclusion* ($\widehat{\pi}_2^{R} = \pi_2$) *et d'autre part le nombre de simulations était infini à la fois pour estimer* $\mathrm{Var}\left(\widehat{Y}\right)$ *et pour moyenner* $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s,HT}$ *alors le taux d'erreur serait nul (parce que* $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s,HT}\right) = \mathrm{Var}\left(\widehat{Y}\right)$*), pourvu que* $\pi_{2,gr} > 0$*. Donc ces trois incertitudes interviennent dans le taux d'erreur observé, ainsi qu'éventuellement la nullité de probabilités d'inclusion double.*

– *La correction de la diagonale de* $\widehat{\pi}_2^{S}$ *par* $\pi_1 = 0.2$ *pour la moyenne sur les simulations des estimations de variance n'a quasiment pas d'incidence sur les lignes supra-régionales de ce tableau.*

• Cependant les résultats de ce tableau sont délicats à interpréter. Le biais apparent de l'estimateur Horvitz-Thompson de la variance pour les variables d'équilibrage fluctue assez largement (en valeur relative) selon le lot de réplications pris en compte pour estimer les probabilités d'inclusion double, y compris pour les deux lots de 100 000 réplications (Tableau 6). De plus, le nombre de simulations utilisées pour moyenner les estimations dans la mesure du biais intervient sensiblement sur celle-ci, cf les trois dernières lignes du Tableau 5. L'insuffisance du nombre de simulations pourrait expliquer la dégradation apparente du biais observée entre 100 000 et 230 000 réplications.

Tableau 6 – Biais relatif de l'estimateur d'Horvitz-Thompson selon le lot de réplications utilisé pour estimer les probabilités d'inclusion de Breidt-Chauvet

lot	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75 ans	feminin	masculin
50 000	3,5		7,0	-0,8	-0,8	-0,8	-0,6	-0,8	-0,9	-0,9			
b	-100,0		-100,0	-2,6	-2,8	-5,1	-7,6	-4,5	-4,7	-4,4			
c	1,4		2,6	-7,7	-7,7	-4,3	-1,6	-0,7	-5,5	-5,7			
d	2,7		5,9	-6,0	-4,4	-3,3	-0,9	0,0	-3,8	-3,8			
100 000	0,8		1,9	-0,3	-0,2	-0,5	-0,4	-0,4	-0,4	-0,4			
b	2,8		5,0	-6,3	-5,0	-3,3	-0,8	0,0	-4,0	-4,0			

Notes :

– *Le lot* 50 000*-b comporte une probabilité estimée* $\widehat{\pi_2} = 3.5\ 10^{-5}$*, pour la région Languedoc-Roussillon.*

– *Les deux lots de* 100 000 *réplications sont obtenus en agrégeant les deux premiers et les deux derniers lots de* 50 000.

• Par contre, les fluctuations entre ces lots de réplications du biais relatif de l'estimateur de Yates-Grundy sont petites (Tableau 7).

Tableau 7 – Fluctuations du biais relatif de l'estimateur de Yates-Grundy entre les lots de réplications

lot	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75 ans	feminin	masculin
50 000	0,1		0,0	0,3	0,2	0,3	0,3	0,3	0,3	0,3			
b	0,1		0,1	0,2	0,2	0,2	0,1	0,1	0,2	0,2			
c	0,2		0,2	0,6	0,4	0,3	0,3	0,4	0,5	0,4			
d	0,2		0,0	0,5	0,3	0,2	0,3	0,2	0,3	0,3			
100 000	0,1		0,0	0,1	0,0	0,1	0,1	0,1	0,1	0,1			
b	0,1		0,1	0,4	0,2	0,1	0,2	0,2	0,2	0,2			

note : L'Ile-de-France se distingue également lorsque la moyenne sur les simulations utilise l'estimateur Breidt-Chauvet. La sous-estimation est réduite avec 430 000 réplications, mais reste visible. La taille moyenne des petites communes est plus grande que les autres régions, mais ceci ne parait pas être un facteur déterminant du biais observé, de même que la variabilité de cette taille.

$\Longrightarrow$ L'estimateur d'Horvitz-Thompson semble nettement plus sensible à l'estimation des probabilités d'inclusion double que la version de Yates-Grundy. L'impact du passage de 5 000 à 230 000 réplications modifie jusqu'à 14 points la précision (apparente) relative d'Horvitz-Thompson contre moins de 3 points pour Yates-Grundy.

$\Longrightarrow$ Le choix d'un nombre de réplications suffisamment élevé est plus crucial pour la qualité de l'estimateur de variance d'Horvitz-Thompson que pour Yates-Grundy.

• Comme pour Yates-Grundy, le signe du biais de l'estimateur d'Horvitz-Thompson, négatif pour la plupart des variables, peut s'interpréter comme une conséquence de la convexité de la fonction

$x \longmapsto \frac{1}{x}$. Cet effet de biais doit diminuer en fonction du nombre de réplications. C'est ce qui est en général observé sur le Tableau 5. Cet élément plaide en faveur de la version de Yates-Grundy, pour laquelle cet effet joue à la hausse. L'autre explication possible d'une sous-estimation est l'existence de probabilités d'inclusion double nulles. Ceci induirait un biais négatif de l'estimateur d'Horvitz-Thompson $\left(- \sum_{\pi_{2,\mathrm{gr}}(c_1,c_2)=0} y\left(c_1\right) y\left(c_2\right) \right)$. Mais cette hypothèse parait contredite par le biais quasiment nul de l'estimateur de Yates-Grundy (et par le Tableau 1).

- Les deux premières variables se distinguent par le signe positif du biais estimé par Horvitz-Thompson. Cependant, le biais de ces deux variables est fortement réduit lorsque le nombre de simulations est accru à $240\,000$. De plus, il est largement négatif lorsque le biais est mesuré avec les probabilités d'inclusion estimées par Breidt-Chauvet. Ces résultats suggèrent que le nombre de simulations n'est pas suffisant pour estimer correctement le biais de l'estimateur d'Horvitz-Thompson.
- La raison pour laquelle l'augmentation du nombre de réplications entre $100\,000$ et $230\,000$ semble détériorer le biais de plusieurs variables est vraisemblablement l'insuffisance du nombre de simulations utilisées pour moyenner les estimations de variance. En effet, le taux d'erreur sur l'estimation de la variance de référence parait trop faible pour intervenir comme explication alternative.

→ Si le principe de choisir le nombre de réplications en fonction du biais mesuré pour des variables connues sur l'univers parait attractif, son application à l'estimateur de variance d'Horvitz-Thompson s'avère délicate. Pour certaines variables, l'incertitude sur la mesure du biais ne permet pas de discerner clairement l'effet du nombre de réplications sur la qualité de l'estimateur 'échantillon' de variance $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}$.

– Une augmentation sensible du nombre de simulations pourrait permettre d'éclaircir le diagnostic pour le deuxième estimateur de variance. Mais le coût des simulations de l'échantillonnage étudié n'a pas permis d'augmenter suffisamment leur nombre pour stabiliser la mesure du biais de l'estimateur de variance d'Horvitz-Thompson.

II.5 mesure de la dispersion des estimateurs de variance

– Sur le premier échantillon simulé, l'erreur absolue d'estimation de l'écart-type (simulé) est largement supérieure à 10% pour la majorité des variables d'équilibrage (Tableau 8). Ceci suggère que les dispersions des deux estimateurs de variance sont élevées.

Tableau 8 – Erreur relative des estimateurs de variance calculés sur le premier échantillon simulé

estimateur	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75	feminin	masculin
YG	-11,1		-46,7	-35,9	-47,9	-34,5	16,5	33,7	-25,5	-27,4			
HT	46,5		27,7	-100,0	-100,0	-42,7	-7,4	15,4	-60,3	-76,1			

→ La dispersion (**eI.4.17**) est calculée avec $\widehat{V}^{S}$ estimé via $\widehat{\pi}_2^{S,f}$ sur $240\,000$ simulations. Selon cet indicateur, la version de Yates-Grundy estime la variance simulée des variables d'équilibrage avec un écart quadratique beaucoup plus réduit que la version d'Horvitz-Thompson (Tableau 9, Tableau 10). En fait, l'estimateur d'Horvitz-Thompson parait inacceptable, comme son écart quadratique moyen est supérieur à celui de l'estimation de variance nulle. Mais ce diagnostic ne porte à ce

stade que sur les variables d'équilibrage.

Tableau 9 – Écart quadratique moyen de l'estimateur de variance de Yates-Grundy, en % de l'écart-type simulé

région	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans 39	40 ans 59	60 ans 74	plus 75 ans	feminin	masculin
11	201,3		204,5	199,8	144,8	203,0	195,5	190,1	202,9	162,0
21	147,5		139,4	156,5	149,4	154,9	159,9	149,1	153,4	155,7
22	171,9		162,5	165,6	159,0	168,9	176,0	176,5	166,9	165,9
23	196,3		170,4	195,6	190,6	201,4	198,1	188,3	194,9	201,1
24	211,0		132,6	222,7	185,6	236,0	238,4	220,2	221,9	225,3
25	118,3		111,3	181,9	171,5	182,4	183,4	167,3	177,4	180,4
26	172,8		155,8	156,5	151,5	162,7	181,0	183,8	166,3	164,9
31	137,9		145,3	253,3	229,5	252,6	207,4	175,8	240,4	244,7
41	172,9		158,0	185,7	186,8	194,8	195,1	183,2	189,6	192,2
42	222,4		175,0	232,3	222,6	237,8	228,9	200,0	230,9	233,0
43	199,9		173,8	174,9	185,4	186,0	178,3	158,6	180,8	182,1
52	122,6		124,8	177,1	166,7	163,2	146,5	150,6	157,2	162,2
53	152,0		130,2	206,9	185,8	200,7	196,7	185,6	191,4	198,6
54	136,4		126,0	185,7	164,3	177,4	166,5	165,6	174,7	177,8
72	138,3		144,3	188,9	179,1	182,3	164,5	172,2	178,1	178,8
73	203,8		147,3	233,8	237,1	231,1	234,0	224,3	244,2	241,3
74	154,3		147,6	150,7	151,0	148,6	158,0	171,5	151,8	150,7
82	177,3		166,8	240,3	214,6	228,3	207,0	193,6	218,2	226,7
83	177,4		151,7	183,2	153,7	179,7	172,7	175,1	175,5	170,0
91	118,5		117,5	203,7	180,6	182,7	169,7	165,1	184,2	185,2
93	158,9		160,8	200,8	194,6	193,9	174,3	160,5	188,7	192,7
94	128,8		152,1	135,1	115,0	130,6	130,4	211,9	170,9	116,5
total	78,3		88,7	93,9	83,9	91,3	87,5	86,2	89,3	88,5

Notes :

- *Les programmes sont ci-joints* **`dispersion`** + **`dispersion_equilib_yg`**.
- *L'omission du dernier terme de l'estimateur de Yates-Grundy donne une dispersion légèrement plus élevée.*

Tableau 10 – Écart quadratique moyen de l'estimateur de variance d'Horvitz-Thompson en pourcentage de l'écart-type simulé

région	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans 39	40 ans 59	60 ans 74	plus 75 ans	feminin	masculin
11	555,1		721,4	546,1	360,5	474,1	395,7	308,6	544,3	404,3
21	240,6		199,2	238,8	227,4	243,6	252,7	230,0	238,9	243,0
22	276,5		232,9	231,0	229,8	252,9	246,6	242,0	241,1	238,7
23	297,8		257,8	282,8	279,9	294,6	290,2	262,9	288,1	293,1
24	340,3		249,4	325,0	313,9	337,5	343,5	301,6	340,6	340,1
25	526,9		513,7	296,2	303,2	311,3	336,4	301,4	317,8	315,4
26	282,2		284,7	287,9	296,9	275,8	260,1	248,3	283,4	286,7
31	165,3		136,9	287,6	265,2	284,4	240,2	197,5	274,0	278,7
41	269,2		245,9	259,6	258,6	259,6	259,3	248,1	260,3	263,5
42	296,7		244,9	301,1	294,5	302,1	292,1	261,4	299,7	302,3
43	439,7		355,6	346,2	377,0	344,8	314,9	283,9	359,2	359,5
52	335,2		317,7	255,9	263,0	263,2	294,6	281,1	274,2	273,7
53	283,6		275,6	247,1	235,6	245,1	243,7	229,8	240,6	244,2
54	309,6		305,4	260,1	256,4	261,1	271,3	260,8	263,4	265,2
72	262,5		238,9	238,0	233,7	237,1	223,7	225,9	235,9	235,8
73	401,1		344,6	307,6	323,4	310,9	321,3	299,5	334,0	331,9
74	290,9		306,9	276,8	280,4	269,9	278,2	280,7	279,9	279,1
82	231,4		221,3	285,9	264,2	276,2	262,8	246,3	273,1	277,4
83	267,2		284,0	252,0	213,3	252,1	246,2	238,0	248,0	240,4
91	549,4		554,8	341,8	373,3	379,2	383,9	322,9	380,6	380,0
93	231,9		234,5	250,8	246,8	243,2	227,5	213,5	240,1	243,2
94	548,7		492,1	496,8	532,4	464,5	531,7	483,9	509,2	526,3
total	310,7		394,7	174,1	167,0	152,6	146,8	132,0	158,6	157,4

– Sur les 10 000 simulations stockées, l'estimateur d'Horvitz-Thompson donne des valeurs négatives pour la variance estimée des variables d'équilibrage dans un cinquième des cas (Tableau 11). C'est beaucoup moins fréquent pour Yates-Grundy [17].

Tableau 11 – Fréquence des estimations de variance négatives (en % des 10 000 simulations)

estimateur	totallog	Logt dans immeuble collectif	moins 20 ans	20 39 ans	40 59 ans	60 74 ans	plus 75 ans	feminin	masculin
HT	17,4	13,5	23,5	20,4	23,0	22,1	21,0	23,1	22,5
YG	4,5	6,4	13,0	7,7	11,5	9,4	8,6	10,8	10,1

II.6 troncature des petites probabilités d'inclusion des PC

• Une troncature des très petites estimations des probabilités d'inclusion peut paraître souhaitable si le nombre de réplications n'est pas suffisant. Ce traitement permet d'éviter que l'estimation de variance ne prenne des valeurs excessivement élevées en valeur absolue.

→ La troncature des probabilités d'inclusion estimées inférieures à 10^{-3} [18] pour l'estimateur de variance d'Horvitz-Thompson présente l'avantage, en moyenne, d'éviter les sous-estimations (Tableau 12). Mais en contrepartie, le biais positif apparent des deux premières variables est sensiblement accru.

Tableau 12 – Impact de la troncature des probabilités d'inclusion à 10^{-3} sur le biais relatif de l'écart-type estimé par Horvitz-Thompson, en % de l'écart-type simulé

traitement	totallog	Logt dans immeuble collectif	moins 20 ans	20 ans	39 ans	40 59 ans	60 74 ans	plus 75 ans	feminin	masculin
avec	6,9	11,4	0,9	1,2	0,7	0,9	0,7	0,9	1,0	
sans	0,3	0,8	-1,8	-1,4	-1,2	-0,5	-0,1	-1,3	-1,3	

Note : La troncature consiste à remplacer les $\widehat{\pi}_2 < 10^{-3}$ *par* 10^{-3}.

• Pour l'estimateur d'Horvitz-Thompson, la troncature présente l'avantage de réduire sensiblement la dispersion de toutes les variables d'équilibrage (Tableau 13). Cependant, la dispersion reste beaucoup plus élevée que celle de l'estimateur de Yates-Grundy.

Tableau 13 – Impact de la troncature sur la dispersion relative de l'estimateur d'Horvitz-Thompson

traitement	totallog	Logt dans immeuble collectif	moins 20 ans	20 ans	39 ans	40 59 ans	60 74 ans	plus 75 ans	feminin	masculin
avec	189,0	212,7	136,5	127,7	133,5	130,6	124,1	133,3	131,1	
sans	310,7	394,7	174,1	167,0	152,6	146,8	132,0	158,6	157,4	

18. Cependant un graphique de dispersion montre une très grande variabilité de l'estimateur de variance de ces variables.

19. Le Tableau 1 montre qu'une troncature à 10^{-4} serait sans incidence, et que seules 4 régions ont des $\widehat{\pi}_2^R < 10^{-3}$.

• Pour l'estimateur de Yates-Grundy, la troncature des probabilités d'inclusion double inférieures à 10^{-3} a un effet faible sur la moyenne des estimations de variance des variables d'équilibrage : moins de 0.2% de l'écart-type simulé (Tableau 14). Ceci illustre la faible sensibilité de cet estimateur aux petites probabilités d'inclusion. De plus, la troncature détériore le biais des deux premières variables.

Tableau 14 – Impact de la troncature des probabilités d'inclusion à 10^{-3} sur le biais relatif de l'écart-type estimé par Yates-Grundy

traitement	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus	75	feminin	masculin
avec	0,3		0,3	0,0		0,0		0,0		0,1		0,0	0,0	0,0
sans	0,1		0,0	0,1		0,0		0,0		0,1		0,1	0,0	0,0

• L'impact sur la dispersion n'apparait pas décisivement favorable pour cet estimateur, du moins au niveau national (Tableau 15).

Tableau 15 – Impact de la troncature des probabilités d'inclusion à 10^{-3} sur la dispersion relative de l'écart-type estimé par Yates-Grundy

traitement	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus	75	feminin	masculin
avec	78,4		88,9	93,6		83,7		91,3		87,5		86,0	89,1	88,4
sans	78,3		88,7	93,9		83,9		91,3		87,5		86,2	89,3	88,5

– La différence d'impact de la troncature illustre la sensibilité plus forte de l'estimateur d'Horvitz-Thompson aux plus petites probabilités d'inclusion.

→ L'option de tronquer les probabilités d'inclusion estimées dans l'estimateur de variance du groupe de rotation des petites communes n'a pas été retenue.

II.7 précision estimée pour des variables générées 'purement aléatoires'

– Un premier lot de variables a été généré sur l'univers des petites communes à l'aide de fonctions SAS de génération de nombres aléatoires :

- loi normale (Var = 1)
- loi normale sur 5% des petites communes [19] (Var = 0.05)
- loi normale sur 60% des petites communes (Var = 0.6)
- loi de Poisson de moyenne 1 (Var = $1; CV = 1$)
- loi uniforme sur $[0, 1]$ (Var = $\frac{1}{12}; CV = \frac{1}{4\sqrt{3}}$)
- loi gamma de paramètre 0.5 (Var = $0.5; CV = 1$)
- loi gamma de paramètre 2 (Var = $2; CV = 1$)
- loi 'exponentielle d'exponentielle' (Gumbel) (Var = $\frac{\pi^2}{6}; CV = \frac{\pi}{\sqrt{6}\gamma}$ avec $\gamma = -\int_0^\infty log\,(x)\, e^{-x} dx$)

Ces variables sont 'purement aléatoires' au sens où elles n'ont pas de lien avec les variables d'équilibrage.

20. Il s'agit plus précisément de la loi du produit de deux variables indépendantes $\mathcal{N}(0,1) \times \mathcal{B}(1,p)$ avec $p = 0.05$.

→ Sur le total des régions, l'estimateur Yates-Grundy de probabilités d'inclusion doubles approximées sur 230 000 réplications estime l'écart-type simulé à moins de 0.2% près en moyenne (Tableau 16). Comme précédemment, l'écart-type simulé est la racine carrée de la moyenne des estimateurs de variance sur les 240 000 simulations, y compris les valeurs négatives.

Tableau 16 – Taux d'erreur de la moyenne des estimateurs de variance de Yates-Grundy par rapport à l'écart-type simulé, en %

région	y norm	y norm 005	y norm 060	y poisson	y uniform	y gamma	y gamma 2	y extreme
11	0,0	0,0	-0,3	0,1	-0,1	-0,1	-0,1	0,0
21	0,0	0,0	0,1	0,1	0,0	0,1	-0,1	0,2
22	-0,2	-0,2	0,2	-0,1	-0,2	0,2	0,0	0,0
23	-0,1	-0,1	-0,1	0,1	0,0	-0,2	0,0	0,2
24	0,1	0,1	0,0	-0,1	-0,2	-0,1	-0,2	0,0
25	0,0	-0,1	0,1	0,0	0,2	0,2	0,0	0,3
26	-0,1	-0,2	-0,1	0,0	-0,2	-0,1	-0,2	0,2
31	-0,2	0,0	0,0	-0,1	-0,1	0,0	-0,1	-0,1
41	0,1	0,3	0,1	0,1	0,0	-0,1	-0,2	-0,2
42	0,1	-0,3	-0,2	0,0	-0,2	0,1	-0,1	0,2
43	0,2	0,0	0,1	-0,1	-0,1	0,0	0,2	0,0
52	-0,1	-0,1	0,1	-0,2	0,2	0,1	-0,1	-0,1
53	-0,2	0,1	-0,4	-0,1	0,0	0,1	0,0	0,2
54	0,2	0,0	-0,1	-0,1	0,1	-0,1	0,0	-0,1
72	-0,1	0,1	-0,3	-0,1	0,4	-0,3	0,2	0,0
73	-0,2	-0,1	0,1	0,0	0,3	0,1	0,2	0,1
74	0,0	-0,1	-0,2	0,0	0,2	0,0	0,0	0,0
82	-0,1	0,3	-0,1	0,1	-0,2	-0,3	0,0	0,0
83	0,1	0,2	-0,2	0,2	0,0	-0,1	-0,1	-0,2
91	-0,3	0,2	0,0	0,1	-0,3	0,1	0,1	-0,1
93	0,0	0,1	0,1	0,2	-0,2	0,0	0,0	0,2
94	-0,3	0,1	0,0	0,2	-0,2	-0,3	-0,1	0,1
total	-0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Note : Le programme de génération de ces variables est ci-joint ***pc_variables_generees****.*

→ La qualité de l'estimateur de variance d'Horvitz-Thompson est très similaire à celle de la version Yates-Grundy pour les variables générées 'purement aléatoires' (Tableau 17). Pour ces variables, le nombre de réplications utilisées pour estimer les probabilités d'inclusion n'influe pas sensiblement sur le biais.

Tableau 17 – Taux d'erreur de la moyenne de l'estimateur de variance d'Horvitz-Thompson, en % de l'écart-type simulé

région	y norm	y norm 005	y norm 060	y poisson	y uniform	y gamma	y gamma 2	y extreme
11	-0,1	-0,1	-0,3	0,0	-0,2	-0,3	-0,2	0,0
21	0,0	0,0	0,1	0,1	0,0	0,1	-0,1	0,2
22	-0,2	-0,1	0,1	0,0	-0,2	0,2	0,0	0,0
23	-0,1	-0,1	-0,1	0,1	0,0	-0,2	0,0	0,2
24	0,1	0,1	0,0	-0,1	-0,2	-0,1	-0,2	0,0
25	0,0	-0,1	0,1	0,0	0,2	0,1	0,0	0,3
26	-0,1	-0,2	-0,1	0,0	-0,2	-0,1	-0,2	0,2
31	-0,2	0,0	0,0	-0,1	-0,1	0,0	-0,1	-0,1
41	0,1	0,3	0,1	0,1	0,0	-0,1	-0,2	-0,2
42	0,1	-0,3	-0,2	0,0	-0,2	0,1	-0,1	0,2
43	0,2	0,0	0,1	-0,1	-0,1	0,0	0,3	0,0
52	-0,1	-0,1	0,1	-0,2	0,2	0,1	-0,1	-0,1
53	-0,2	0,1	-0,4	-0,1	0,0	0,1	0,0	0,2
54	0,1	0,1	-0,1	-0,1	0,1	-0,1	-0,1	-0,1
72	-0,1	0,1	-0,3	-0,1	0,4	-0,3	0,2	0,0
73	-0,2	-0,1	0,1	-0,1	0,3	0,1	0,2	0,1
74	0,0	-0,1	-0,2	0,0	0,2	0,0	0,0	0,0
82	-0,1	0,3	-0,1	0,1	-0,2	-0,3	0,0	0,0
83	0,1	0,3	-0,2	0,2	0,0	-0,2	-0,1	-0,2
91	-0,3	0,2	0,0	0,1	-0,3	0,1	0,1	-0,1
93	-0,1	0,2	0,1	0,2	-0,1	0,0	0,0	0,2
94	-0,3	0,1	0,1	0,2	-0,2	-0,2	-0,1	0,1
total	-0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
100 000 réplications	-0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
50 000	-0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
10 000	-0,1	0,0	0,0	0,0	-0,1	-0,1	-0,1	0,0

- Le même constat s'applique au premier échantillon généré : les deux estimations de variance des variables 'purement aléatoires' sont proches, d'erreur relative faible et comprise entre -2.2% et +1.3%.
- Le contraste entre la qualité de l'estimation de variance des variables d'équilibrage et celle des variables générées montre que cette qualité varie largement en fonction de la variable d'intérêt. Il faut en tenir compte dans le choix de l'estimateur de variance et en particulier du nombre de réplications utilisées pour les probabilités d'inclusion, afin de permettre un calcul de variance correct pour le maximum de variables d'intérêt.

II.8 précision estimée pour des variables générées en lien avec l'équilibrage des PC

• Une interprétation du contraste entre la comparaison des deux estimateurs de variance d'une part des variables d'équilibrage (de vecteur noté x) et d'autre part des variables générées est que le biais relatif de l'estimateur de Yates-Grundy dépendrait du lien entre la variable d'intérêt et les variables d'équilibrage.

→ Pour évaluer cette explication, un deuxième lot de variables a été généré de manière à disperser les ratios $\frac{\mathrm{Var}\left[E\left(y|x\right)\right]}{\mathrm{Var}\left(y\right)}$.

– Pour chacun des trois premiers vecteurs propres u de $\mathrm{Var}\left(x\right)$, mesuré sur l'univers des petites communes d'une région, de valeurs propres maximales $\lambda_1 > \lambda_2 > \lambda_3 > 0$, les variables suivantes ont été générées :

$$y_{u,\tau} = \sqrt{\frac{1}{2}\frac{\tau}{10}}\left\{\sqrt{\frac{3}{\lambda_+}}(x-\overline{x})'u + \epsilon\right\} + 1$$

avec : $\lambda_+ = \lambda_1 + \lambda_2 + \lambda_3$, τ allant de 1 à 10 et $\epsilon \rightsquigarrow \mathcal{N}(0,1)$

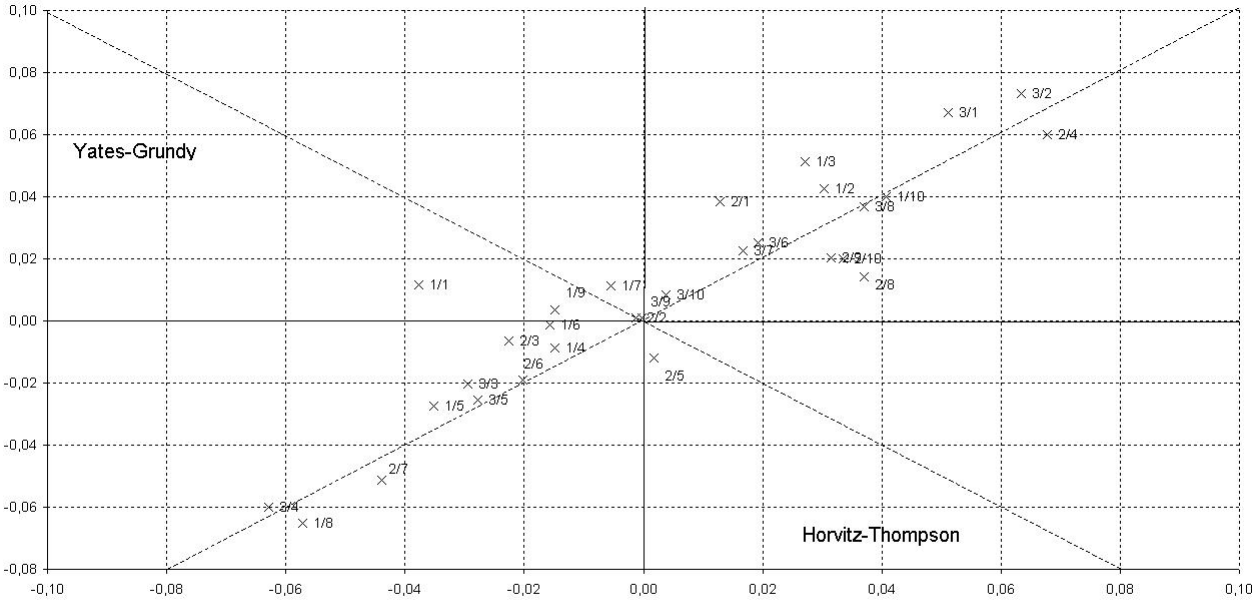
→ $\frac{\mathrm{Var}\left[E\left(y_{u,\tau}|x\right)\right]}{\mathrm{Var}\left(y_{u,\tau}\right)} = \frac{\frac{1}{2}\frac{\tau}{10}\frac{3}{\lambda_+}\lambda}{\frac{1}{2}\frac{\tau}{10}\frac{3}{\lambda_+}\lambda + \frac{1}{2}\frac{\tau}{10}} = \frac{3\frac{\lambda}{\lambda_+}}{3\frac{\lambda}{\lambda_+}+1} = \frac{1}{1+\frac{\lambda_+}{3\lambda}}$ Donc le lien entre y et x croit en fonction de la valeur propre. Ceci correspond à une décroissance en fonction du premier indice du nom de la variable générée ($y_{u_k,\tau}$ est notée k/τ sur le graphique ci-dessous, où k désigne le rang de la valeur propre, dans l'ordre décroissant).

→ $CV(y) = \frac{1}{2}\frac{\tau}{10}\left(3\frac{\lambda}{\lambda_+}+1\right)$ est décroissant en fonction du premier indice.

→ Le graphique **pc_variables_generees2.xls** montre que la variance des variables ($\mathrm{Var}\left(y\right)$ et non $\mathrm{Var}\left(\widehat{Y}\right)$) évolue en fonction des deux paramètres de leur construction comme prévu par ces deux relations. Le programme de construction de ces variables est **pc_variables_generees2.sas**.

• Sur cet ensemble de variables, il n'apparait pas de relation nette entre le lien de la variable avec le vecteur d'équilibrage et le biais apparent des deux estimateurs (Graphique 1). Pour la plupart des variables pour lesquelles l'estimateur de Yates-Grundy sous-estime apparemment la variance, la sous-estimation de l'estimateur d'Horvitz-Thompson est plus forte.

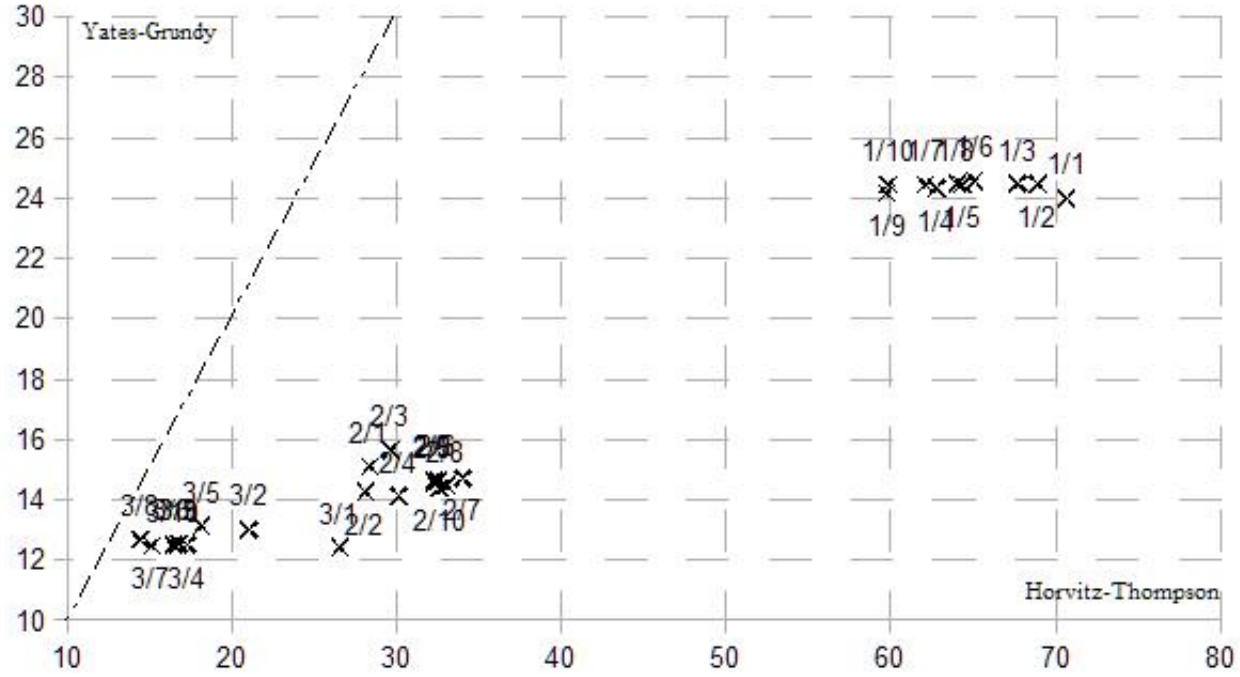
Graphique 1 – Biais relatif de l'estimation de l'écart-type simulé selon la méthode d'estimation, en % de l'écart-type simulé



Note : Ci-joints les résultats détaillés par région ***biais_relatif_gener2_YGc_230000.xls*** ***biais_relatif_gener2_HTc_230000.xls***.

- Pour toutes ces variables, la dispersion de l'estimateur de variance d'Horvitz-Thompson est plus grande que celle de Yates-Grundy (Graphique 2). La différence est particulièrement sensible pour les variables les plus liées à l'équilibrage (celles préfixées par 1/). Ce résultat est cohérent avec la comparaison précédente des dispersions pour les variables d'équilibrage.

Graphique 2 – Dispersion de l'estimateur de variance selon la méthode d'estimation, en % de l'écart-type simulé



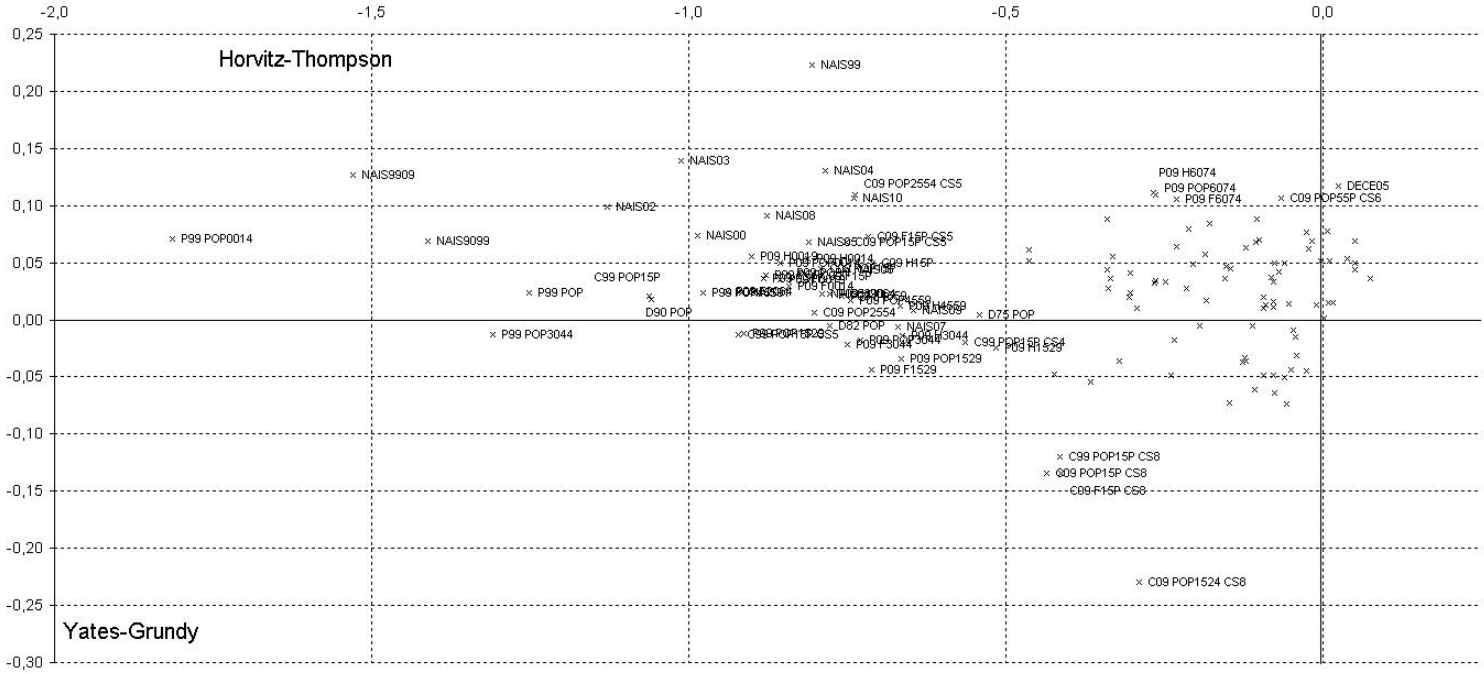
Note : Ci-joints les résultats détaillés par région ***dispersion_gener2_HTc_230000.ods*** ***dispersion_gener2_YGc_230000.ods****, ainsi que l'un des programmes* ***dispersion_gener2_HT.sas****.*

II.9 précision de la variance estimée de variables issues de recensements

La comparaison des biais mesurés pour les deux estimateurs de variance a été étendue à 138 variables socio-démographiques issues des recensements de 1968 à 2009.

→ Il apparait que pour ces variables 'naturelles', le biais de la variance estimée par Yates-Grundy est inférieur à 0.3% de l'écart-type simulé (Graphique 3), en valeur absolue. L'estimateur d'Horvitz-Thompson sous-estime la variance de la grande majorité de ces variables. Son biais apparent excède 1% pour 8 variables. Le signe prédominant du biais de chacun des deux estimateurs est conforme aux attentes.

Graphique 3 – Biais relatif des estimateurs d'Horvitz-Thompson et de Yates-Grundy pour des variables issues de recensements, en % de l'écart-type simulé



Notes :

– *P09=exploitation principale RP 2009, C09=exploitation complémentaire RP 2009*

– *CS1=Agriculteurs exploitants, CS2=Artisans, Commerçants, Chefs d'entreprise, CS3=Cadres et Professions intellectuelles supérieures, CS4=Professions intermédiaires, CS5=Employés, CS6=Ouvriers, CS7=Retraités, CS8=Autres sans activité professionnelle.*

– *H=homme, F=femme*

– *Ci-joints les résultats détaillés* ***biais_relatif_varrp_HTc_230000.xls*** ***biais_relatif_varrp_ygc_230000.xls****.*

– *Le tableau complémentaire ci-joint fournit la dispersion mesurée pour ces variables* ***dispersion_varrp_YGc_230000.ods****.*

II.10 un critère de classement des estimateurs de variance

– La proportion des 10 000 simulations stockées telles que l'estimateur de variance d'Horvitz-Thompson est préférable a été calculée par variable. L'estimateur de variance HT est jugé préférable, au sens large, à YG si l'une des trois conditions suivantes est vérifiée :

$$\widehat{V}^{YG} \geqslant \widehat{V}^{HT} \geqslant \widehat{V}^{S}$$
$$\widehat{V}^{HT} \geqslant \widehat{V}^{S} > \widehat{V}^{YG}$$
$$\widehat{V}^{YG} \leqslant \widehat{V}^{HT} < \widehat{V}^{S}$$

→ Pour les variables d'équilibrage, l'estimateur d'Horvitz-Thompson n'est pas préférable en ce sens, sauf pour la variable 'nombre de logements collectifs' (Tableau 18).

Tableau 18 – Taux de préférence pour l'estimateur d'Horvitz-Thompson, en pourcentage des 10 000 simulations des estimateurs de variance

totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39	40 ans	59	60 ans	74	plus ans	75	feminin	masculin
41,9		50,4	35,2	35,8		35,4		35,9		36,4		35,6	35,2

Note : Le programme est ***pc_var_equilib_prefg.sas****.*

→ Pour toutes les variables générées du deuxième lot (décrit dans II.8), l'estimateur de Yates-Grundy est préférable selon ce critère. La préférence est plus prononcée pour les variables les plus corrélées aux variables d'équilibrage. C'est également le cas pour les variables 'recensement', à 18 exceptions près.

$\Longrightarrow$ Cette approche confirme clairement celles du biais et de la dispersion, que l'estimateur de Yates-Grundy généralisé est préférable à celui d'Horvitz-Thompson pour la variance du groupe de rotation des petites communes..

II.11 estimation de la variance sur les groupes de rotation effectifs de PC

– Une table datée de la mi-2008 décrit l'affectation par groupe de rotation des petites communes [20].

• Au sein de chacun des 5 groupes de rotation effectifs, la probabilité d'inclusion double minimale est supérieure à $1.1\ 10^{-2}$, pour toutes les régions. Donc la troncature de la probabilité d'inclusion à 10^{-3} n'a aucune incidence pour l'estimation de variance sur ces groupes de rotation.

• Le calcul de variance des variables d'équilibrage a été effectué pour chacun de ces groupes de rotation, sans traiter les divergences de codes géographiques par rapport à la table des probabilités d'inclusion. La variance de référence est calculée de la même façon que précédemment, donc indépendamment des groupes de rotation effectifs.

21. Elle semble incomplète : elle ne contient que 35 684 communes. 35 petites communes du référentiel ne sont pas dans la table des groupes de rotation (Annexe B). Inversement, 11 de ses communes ne sont pas retrouvées dans la table des petites communes utilisée ici. Il s'agit d'une part des quatre communes inhabitées de la Meuse, d'autre part de communes qui n'existaient pas en 1999.

→ L'erreur observée sur les 5 groupes de rotation effectifs pour les variables d'équilibrage estimateur de Yates-Grundy est dans les deux tiers des cas d'ampleur inférieure à celle de l'estimateur d'Horvitz-Thompson (Tableau 19, Tableau 20).

Tableau 19 – Erreur relative de l'estimation de Yates-Grundy par groupe de rotation effectif, en % de l'écart-type simulé

GR	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75	feminin	masculin
1	-22,2		-2,7	39,4	28,9	38,3	17,9	1,9	33,8	33,5			
2	5,8		-18,1	-6,2	-58,1	-11,2	-9,7	-13,7	-30,2	-24,5			
3	39,2		16,6	39,2	35,4	44,1	60,3	46,7	51,2	46,3			
4	12,7		16,4	-67,4	-26,3	-100,0	-68,3	4,7	-57,7	-100,0			
5	-4,6		-3,1	6,6	18,1	-1,5	-23,7	-30,6	-9,8	0,9			

Note : -100% correspond à une estimation de variance négative.

Tableau 20 – Erreur relative de l'estimation d'Horvitz-Thompson par groupe de rotation effectif

GR	totallog	Logt dans collectif	immeuble	moins 20 ans	20 ans	39 ans	40 ans	59 ans	60 ans	74 ans	plus 75	feminin	masculin
1	-100,0		-12,7	31,3	22,4	40,8	23,5	-12,9	29,1	28,0			
2	43,2		-1,5	-28,0	-100,0	-30,9	-21,4	-43,2	-76,8	-55,6			
3	72,3		74,1	-26,4	1,7	-40,6	0,6	36,5	-2,8	-12,2			
4	10,6		44,7	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0			
5	64,5		41,6	44,7	34,3	56,6	55,6	41,7	49,7	43,5			

Note : Il s'agit du taux d'erreur observée de l'estimation de variance, en pourcentage de l'écart-type simulé.

• Pour le premier lot de variables générées, les taux d'erreurs sur les groupes de rotation effectifs sont nettement plus petits que pour les variables d'équilibrage (Tableau 21, Tableau 22). Les deux séries d'estimations ne sont pas clairement ordonnées : l'erreur est moins ample pour Yates-Grundy dans 55% des cas.

Tableau 21 – Erreur relative de l'estimation de Yates-Grundy pour le premier lot de variables générées, en % de l'écart-type simulé

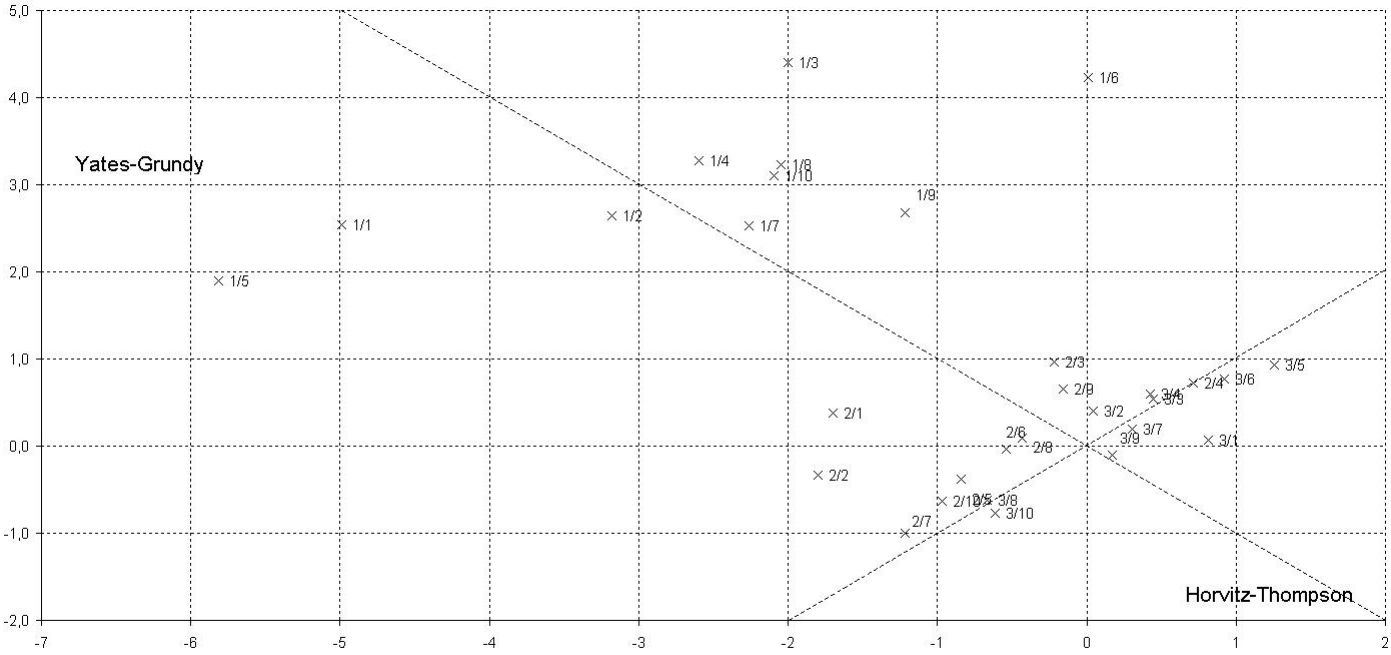
GR	y norm	y norm 005	y norm 060	y poisson	y uniform	y gamma	y gamma 2	y extreme
1	-0,3	-6,1	-2,4	-1,1	-0,7	0,0	1,3	-1,4
2	-1,4	6,6	0,0	0,0	0,1	1,2	0,0	0,5
3	0,3	-3,8	1,0	1,1	0,5	0,8	-0,6	2,6
4	0,6	5,6	1,9	1,6	0,2	-3,1	-1,2	-2,0
5	0,0	-0,5	-0,9	-1,7	-0,1	0,3	0,1	0,3

Tableau 22 – Erreur relative de l'estimation d'Horvitz-Thompson pour le premier lot de variables générées

GR	y norm	y norm 005	y norm 060	y poisson	y uniform	y gamma	y gamma 2	y extreme
1	-0,4	-7,5	-2,4	-1,1	-1,0	0,0	1,3	-1,3
2	-1,3	6,7	0,2	-0,1	0,0	1,3	0,1	0,6
3	0,3	-4,5	1,0	1,3	0,6	0,6	-0,4	2,5
4	0,5	5,4	1,9	1,4	0,1	-3,2	-1,7	-2,1
5	0,0	-0,9	-0,9	-1,5	-0,3	0,5	0,2	0,2

– Pour le deuxième lot de variables générées, le sous-groupe des variables les plus corrélées aux variables d'équilibrage se distingue par l'ampleur de la surestimation par l'estimateur de Yates-Grundy sur le premier groupe de rotation et, moins nettement, celle de la sous-estimation par Horvitz-Thompson (Graphique 4).

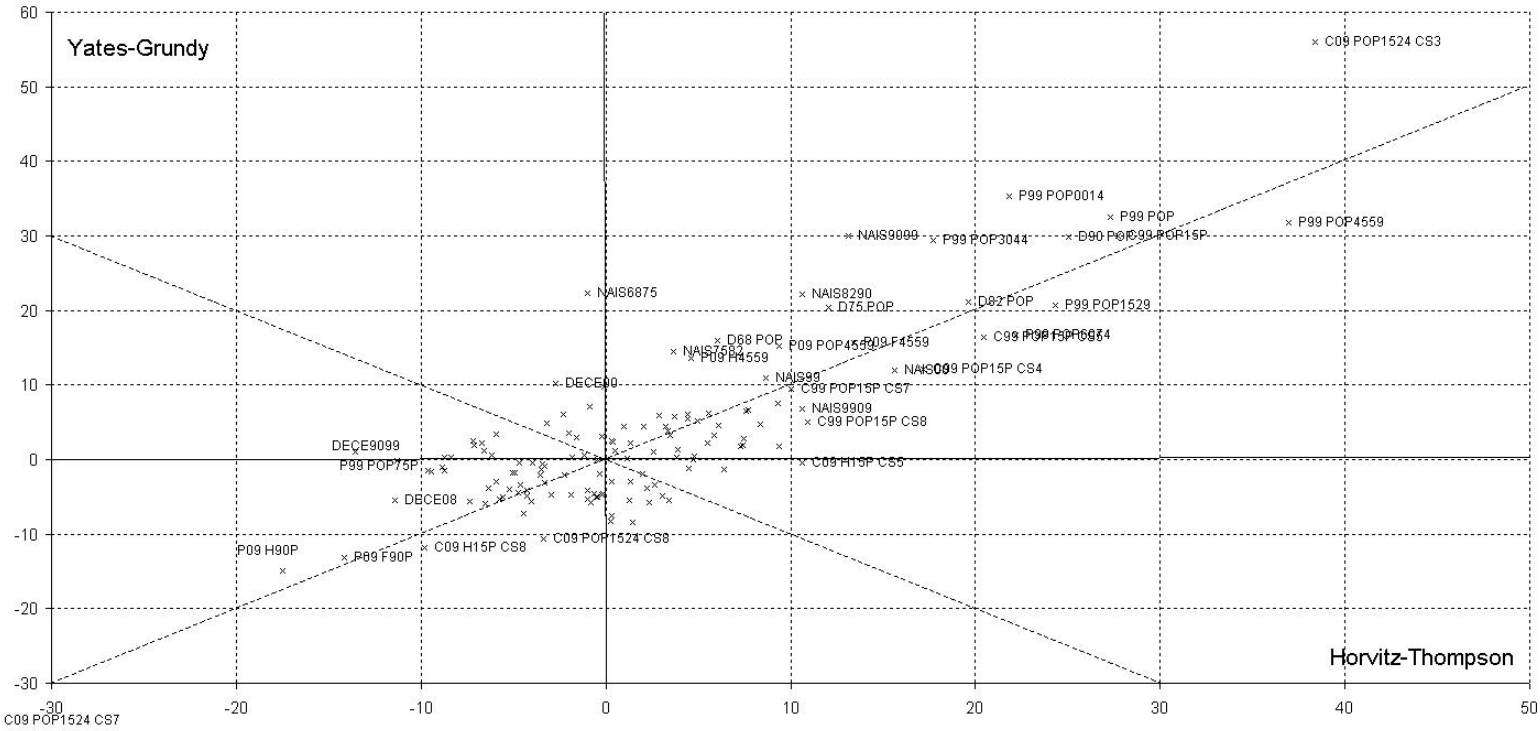
Graphique 4 – Taux d'erreur de l'estimation de variance sur le premier groupe de rotation effectif, en % de l'écart-type simulé



Note : Les résultats détaillés sont ci-joints ***erreur_par_gr_gener2_htc_230000.xls*** ***erreur_par_gr_gener2_ygc_230000.xls*** ***erreur_par_gr_gener2_graph.xls***.

• Pour les 138 variables issues des recensements, les erreurs d'estimation de la variance observées sur le premier groupe de rotation paraissent similaires entre les deux estimateurs (Graphique 5). Toutefois, sur le quatrième groupe de rotation l'estimateur d'Horvitz-Thompson donne une estimation négative de la variance pour 73 variables, contre 3 pour Yates-Grundy. Ces nombres n'excèdent pas 1 sur les autres groupes de rotation. Ceci pourrait illustrer l'instabilité plus forte de l'estimateur de variance d'Horvitz-Thompson.

Graphique 5 – Taux d'erreur de l'estimation de variance sur le premier groupe de rotation effectif, en % de l'écart-type simulé



Notes :

– *Ci-joints les tableaux détaillés* ***erreur_par_gr_varrp_ygc_230000.xls*** ***erreur_par_gr_varrp_htc_230000.xls****.*

– *26 communes présentes au RP 2009 et de moins de* 10 000 *habitants au RP 1999 ne sont pas appariés sur le code communal avec la table de référence des petites communes. Inversement 15 petites communes de ce référentiel ne sont pas dans les données communales du RP 2009 (Annexe C).*

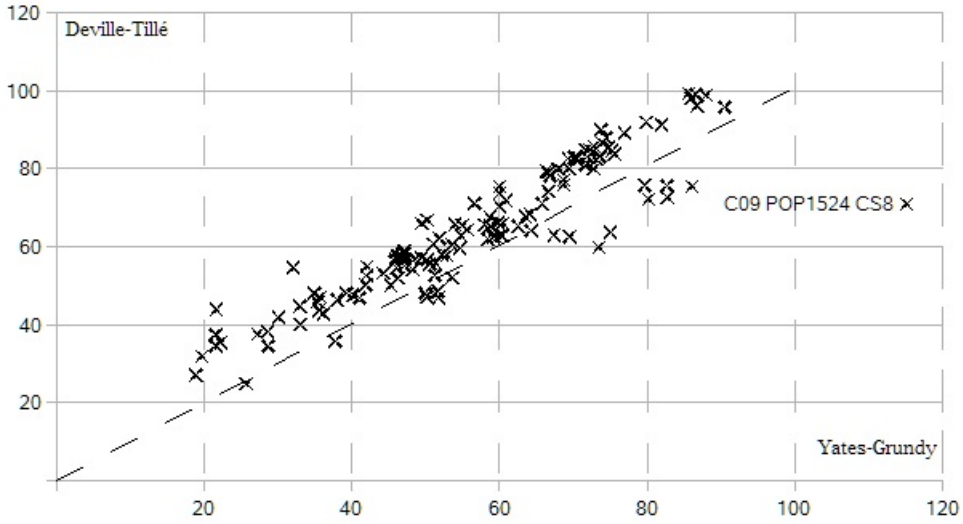
• Au total, la comparaison des erreurs observées pour les deux estimateurs sur le premier groupe de rotation tiré ne parait pas aussi défavorable à la version d'Horvitz-Thompson que l'étude du biais. Ceci suggère que le biais de cet estimateur n'est pas estimé avec suffisamment de précision par les simulations disponibles.

II.12 comparaison à l'estimateur Deville-Tillé

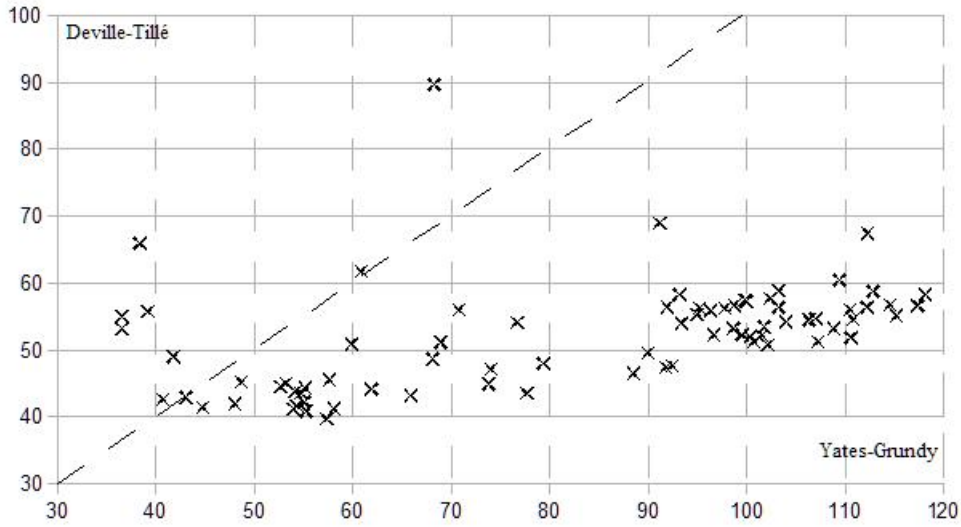
– L'estimateur Yates-Grundy présente l'avantage par rapport à celui de Deville-Tillé de prendre en compte l'atterrissage de l'équilibrage. Néanmoins, sa variabilité est forte, ce qui incite à vérifier qu'il est bien préférable à cette alternative plus simple.

– Pour la majorité des variables 'recensement', la dispersion de l'estimateur Yates-Grundy est inférieure à celle de Deville-Tillé au niveau national (Graphique 6). C'est le contraire au niveau régional.

Graphique 6 – Comparaison des dispersions pour les variables 'recensement'



Nord-Pas-de-Calais :



Note :

– L'estimateur Deville-Tillé est calculé ici sans prendre en compte les variables d'équilibrage départementales.

– Pour l'approximation de Deville, la dispersion de la plupart de ces variables (126/139) est largement supérieure à 100%. Elle n'est inférieure à celle de Yates-Grundy que dans un seul cas.

II.13 conclusions sur la variance du groupe de rotation des petites communes

– L'utilisation du biais et de la dispersion pour la validation de l'estimateur de variance basé sur les probabilités d'inclusion estimées parait globalement justifiée. Cependant, un certain manque de cohérence des résultats pour l'estimateur de variance d'Horvitz-Thompson suggère que le nombre de simulations nécessaire pour mesurer correctement la qualité de celui-ci est très élevé.

– La qualité des estimations de variance obtenues par Yates-Grundy avec les probabilités d'inclusion estimées sur les réplications produites parait acceptable. Pour la majorité des variables étudiées et des groupes de rotations, hormis les variables d'équilibrage, l'erreur d'estimation sur un seul groupe de rotation est inférieure à 10% de l'écart-type. Pour les variables d'équilibrage, la variance est faible, donc un taux d'erreur élevé peut être toléré.

– En conclusion opérationnelle, la variance due au tirage du groupe de rotation des petites communes recensées peut être calculée correctement par l'estimateur de Yates-Grundy généralisé, avec les probabilités d'inclusion double estimées par la méthode de Breidt-Chauvet sur les $430\,000$ réplications disponibles, sans troncature des petites probabilités d'inclusion.

– Selon les indicateurs de qualité appliqués aux variables étudiées, cet estimateur présente les avantages par rapport à celui d'Horvitz-Thompson d'un biais mesuré plus faible et d'une dispersion beaucoup plus petite. De plus, il semble nettement moins sensible à l'estimation des probabilités d'inclusion. Un dernier avantage de cette version est que l'estimation des probabilités d'inclusion joue à la surestimation de la variance.

– L'estimateur Breidt-Chauvet de la variance parait également préférable à celui de Deville-Tillé, bien que l'avantage soit moins ample et moins général.

– Les $230\,000$ ($+200\,000$) réplications réalisées paraissent suffire pour estimer les probabilités d'inclusion double. L'erreur sur les probabilités d'inclusion simple est de moins de 1%. De plus, il n'apparait plus d'effet sensible de l'augmentation du nombre de réplications sur la qualité de l'estimation de variance des variables étudiées. Le gain en biais attendu d'une augmentation du nombre de réplications serait très inférieur à 1% de l'écart-type des variables d'équilibrage. Toutefois, ce diagnostic repose sur une mesure du biais qui est incertaine.

– Pour le calcul de variance effectif, il semble difficile d'envisager un traitement systématique des discordances de la géographie communale entre les données d'enquête et la table des probabilités d'inclusion estimées des petites communes. Dans un premier temps au moins, il est proposé d'ignorer ce défaut d'appariement. Son impact sur la variance estimée est vraisemblablement négligeable au niveau national. Les résultats obtenus dans la présente étude ne semblent pas contredire cette hypothèse.

– L'incidence sur les probabilités d'inclusion estimées, et donc sur la variance estimée, de la différence entre la table des petites communes utilisée ici et celle du tirage effectif des groupes de rotation du recensement est inconnue. Vu le faible nombre des modifications géographiques communales intervenues depuis, elle devrait être négligeable, au moins au niveau national.

Partie III
probabilités d'inclusion doubles des ZAE dans l'échantillon maître

Cette partie estime les probabilités d'inclusion double des ZAE (zones d'action enquêteur [MCSF]) dans l'échantillon maître (EM) par la méthode de G.Chauvet [GC] sur l'ensemble des régions. Les résultats obtenus paraissent cohérents avec ce papier de référence. La méthode fournit un estimateur de la variance de l'échantillon maître qui semble clairement préférable aux options plus simples des approximations de Deville et de Deville-Tillé. Néanmoins, l'application de la formule de Deville à la Corse s'avère utile pour limiter la dispersion de l'estimation de variance. L'alternative d'une troncature des plus petites probabilités d'inclusion nécessiterait un seuil élevé (10^{-3}), avec l'inconvénient d'une sous-estimation de la variance sensible en moyenne.

Deux annexes (Annexe E) et (Annexe F) traitent des échantillons des extensions (EMEX).

III.1 méthode d'estimation des probabilités d'inclusion des ZAE

– La matrice des probabilités d'inclusion double de l'ensemble des ZAE d'une strate ZAE [21] est estimée par la méthode de Breidt-Chauvet.

– Les variables d'équilibrage prises en compte ici sont celles effectivement utilisées pour le tirage des ZAE, dans l'ordre de la déclaration à Cube. Les variables d'équilibrage des deux strates ZAE de la région Ile-de-France sont spécifiques, listées ci-dessous (et explicitées dans [MCSF] p.12).

```
%let liste_variables_equilibrage=totres99 nresgr1 nresgr2 nresgr3 nresgr4
 nres_rural nres_periurbain         revenufisc04_1 revenufisc04_2
 revenufisc04_3 revenufisc04_4 revenufisc04_5;
pour la petite couronne de l'Ile-de-France :
%let liste_variables_equilibrageidfPC=totres99 revenufisc04 age1 age2 age3
 etranger monoparental grande_taille proprietaire nb_hlm collectif;
pour la grande couronne de l'Ile-de-France :
%let liste_variables_equilibrageidfGC=totres99 nresgr1 nresgr2 nresgr3
 nresgr4 nres_periurbain revenufisc04_1 revenufisc04_2
 revenufisc04_3 revenufisc04_4 revenufisc04_5
 age1 age2 age3 etranger monoparental grande_taille proprietaire
 nb_hlm collectif;
```

– A noter que l'échantillon maître effectif résulte d'un choix régional entre deux échantillons indépendants. Ce traitement affecte les probabilités d'inclusion double. Il n'est pas pris en compte ici.

– Le programme ci-joint **pi2_hat_zae** tourne en 5h10 pour 50 000 simulations sur PC. L'exécution semble plus lente sur AUS. Le stockage de la matrice des probabilités d'inclusion double pour l'ensemble des 3 743 ZAE ne pose pas de difficultés : elle occupe moins de 7 Mo.

22. La région, sauf pour l'Ile-de-France qui est décomposée en deux strates ZAE : la petite et la grande couronne.

notes :

– Le programme d'estimation des probabilités d'inclusion exclut les deux ZAE exhaustives d'Île-de-France (Z75000 et Z92012 : Boulogne-Billancourt) [22]. Mais les ZAE exhaustives n'interviennent pas dans l'estimation de variance.

– L'estimation de $E\left(ss'\right)$ par la moyenne simple $\widehat{\pi}_2^{\,f} = \sum_R ss' \Big/ |R|$ (estimateur par la fréquence) est plus rapide d'environ 15% par rapport à la méthode de Breidt et Chauvet, sur 100 réplications. Donc la question se pose de la méthode préférable pour une durée des réplications fixée. Pour apporter un élément empirique de réponse, une estimation a été effectuée avec cette méthode alternative sur $1\,100\,000$ réplications du tirage des ZAE. Son exécution sur AUS a duré 5 jours. La comparaison des résultats obtenus a mené à écarter cette alternative.

• Une deuxième estimation Breidt-Chauvet des probabilités d'inclusion double a été réalisée sur un million de simulations. Celles-ci sont utilisées pour estimer la qualité de l'estimation de variance (biais et dispersion). Les probabilités d'inclusion finales ont été estimées sur toutes les $2\,300\,000$ réplications disponibles, qui intègrent les simulations utilisées pour mesurer la qualité de l'estimateur de variance.

23. Cette exclusion est la conséquence d'un détail technique, la codification de la strate ZAE dans la table utilisée.

III.2 qualité de l'estimation des probabilités d'inclusion des ZAE

– Pour l'estimation réalisée sur un million de réplications, l'erreur de l'estimateur Breidt-Chauvet de la probabilité d'inclusion simple est d'au plus 0.6% pour toutes les régions (Tableau 23). La probabilité d'inclusion double minimale est supérieure à 10^{-4} dans la majorité des régions. Cependant, elle pourrait être nulle pour la Corse, et très petite pour le Limousin. Les résultats obtenus avec l'estimateur des probabilités d'inclusion par la fréquence confirment ce diagnostic.

Tableau 23 – Indicateurs de qualité de l'estimation des probabilités d'inclusion des ZAE dans l'échantillon maître

région		nombre zae	$\|s\|$	taux d'erreur maximal selon nombre réplications 2 300 000	1 000 000	300 000	50 000	1 100 000 fréquence	sur 1 000 000 réplications $\min(\widehat{\pi}_2)$	$\max(\widehat{\pi}_2)$	$\min(\widehat{\pi}_2)$ fréquence
11_pc	Île-de-France/petite couronne	108	40	0,1	0,3	0,6	1,0	0,4	6,7E-03	0,96	6,6E-03
11_gc	Île-de-France/grande couronne	253	44	0,3	0,5	0,9	2,3	0,8	1,5E-03	0,90	1,4E-03
21	Champagne-Ardenne	115	13	0,4	0,5	0,7	2,1	0,9	9,5E-05	1,00	1,1E-04
22	Picardie	183	18	0,4	0,5	0,9	2,0	1,1	4,0E-04	1,00	3,7E-04
23	Haute-Normandie	141	16	0,4	0,4	0,9	2,5	1,2	4,5E-04	1,00	4,4E-04
24	Centre	194	24	0,3	0,5	1,0	2,8	1,2	7,7E-04	1,00	7,6E-04
25	Basse-Normandie	148	14	0,3	0,4	0,8	2,1	1,0	2,8E-04	1,00	2,5E-04
26	Bourgogne	144	16	0,3	0,6	1,1	2,3	0,8	1,9E-04	1,00	1,9E-04
31	Nord-Pas-de-Calais	235	36	0,3	0,4	0,9	2,7	0,8	5,4E-04	1,00	5,6E-04
41	Lorraine	181	23	0,4	0,5	0,9	2,3	1,1	1,0E-04	1,00	9,6E-05
42	Alsace	123	15	0,3	0,4	0,8	2,0	0,9	2,6E-04	1,00	2,5E-04
43	Franche-Comté	114	11	0,4	0,6	1,1	1,8	0,9	1,2E-04	1,00	1,2E-04
52	Pays de la Loire	198	28	0,3	0,5	0,8	1,9	0,9	1,7E-03	1,00	1,6E-03
53	Bretagne	188	28	0,2	0,3	0,6	2,2	0,8	2,1E-03	1,00	2,2E-03
54	Poitou-Charentes	138	17	0,3	0,5	0,7	2,5	1,4	9,4E-04	1,00	9,0E-04
72	Aquitaine	221	28	0,3	0,4	0,9	2,1	1,0	9,0E-04	1,00	8,9E-04
73	Midi-Pyrénées	213	23	0,4	0,6	1,5	2,9	1,3	2,4E-04	1,00	2,2E-04
74	Limousin	57	7	0,3	0,5	0,6	2,0	0,9	-6,8E-06	1,00	2,7E-06
82	Rhône-Alpes	363	51	0,3	0,5	0,9	2,3	1,3	1,4E-03	1,00	1,4E-03
83	Auvergne	115	13	0,3	0,4	0,9	2,2	0,8	1,5E-04	1,00	1,7E-04
91	Languedoc-Roussillon	140	21	0,3	0,4	0,6	1,6	1,2	1,5E-04	1,00	1,5E-04
93	Provence-Alpes-Côte d'Azur	150	34	0,3	0,5	0,5	1,4	0,7	2,4E-04	1,00	1,9E-04
94	Corse	19	3	0,2	0,4	0,5	1,0	0,4	-8,2E-06	0,63	0,0E+00
	total	3 741	523	0,4	0,6	1,5	2,9	1,4	-8,2E-06	1,00	0,0E+00

Notes :

– $|s|$ *est la taille de l'échantillon maître des ZAE (Paris et Boulogne-Billancourt sont exclues).*

– *Le taux d'erreur maximal est calculé par rapport aux probabilités d'inclusion simple :* $100\left|\widehat{\pi}_1^{\,R} \,/\, \pi_1 - 1\right|$.

– Avec une probabilité d'inclusion minimale de 10^{-3} et si la loi de l'estimateur de la probabilité d'inclusion double est approximée par une binomiale, le nombre de réplications nécessaire pour obtenir un CV $\dfrac{\sqrt{\mathrm{Var}\left(\widehat{p}^{R}\right)}}{p} = \dfrac{\sqrt{\dfrac{p(1-p)}{|R|}}}{p} = \sqrt{\dfrac{\dfrac{1}{p}-1}{|R|}}$ de 1% sur cette probabilité est de $\dfrac{1/10^{-3}-1}{0.01^2} \cong 10$ millions. La procédure utilisée apparait ainsi relativement performante.

– La convergence de l'estimateur des probabilités d'inclusion est retrouvée dans les résultats. Elle reste sensible entre 300 000 et un million de réplications, en pourcentage.

– L'estimation des probabilités d'inclusion des ZAE obtenue par la méthode de Breidt et Chauvet présente de bonnes propriétés. En particulier, les probabilités estimées sont toutes comprises entre 0 (à 10^{-5} près) et 1. La raison de cette performance reste à expliquer rigoureusement (Annexe D).

– L'estimateur par la fréquence calculée sur $1\,100\,000$ réplications donne une estimation des probabilités d'inclusion double qui n'est pas nettement meilleure que celle de Breidt-Chauvet sur $300\,000$, selon la proximité de la diagonale à la probabilité d'inclusion simple.

– La distance euclidienne entre la probabilité d'inclusion simple π_1 et son estimation $\widehat{\pi}_1^{\,R}$ est inférieure à 0.003 pour toutes les régions (0.010 avec $50\,000$ réplications).

• La distance entre les estimations de la matrice des probabilités d'inclusion double sur deux lots indépendants respectivement de $50\,000$ et $100\,000$ réplications parait faible (Tableau 24). Compte tenu de la dimension de la matrice, cette distance n'apparait pas excessive par rapport à celle entre l'estimation des probabilités d'inclusion simple et leurs vraies valeurs.

Tableau 24 – Distance euclidienne entre deux lots d'estimation des probabilités d'inclusion double

$\lvert R\rvert$	11 pc	11 gc	21	22	23	24	25	26	31	41	42	43	52	53	54	72	73	74	82	83	91	93	94
50 000	0,020	0,029	0,012	0,016	0,013	0,019	0,014	0,014	0,021	0,016	0,013	0,012	0,019	0,020	0,015	0,018	0,017	0,009	0,025	0,013	0,015	0,018	0,004
100 000	0,014	0,020	0,009	0,011	0,010	0,012	0,010	0,010	0,014	0,012	0,009	0,008	0,014	0,014	0,010	0,013	0,013	0,006	0,018	0,009	0,012	0,012	0,003

III.3 simulations du tirage de l'échantillon maître des ZAE

• L'estimateur de variance utilisé dans cette partie, sauf mention contraire, est celui de Yates-Grundy (**eI.2.10**). Il est approximativement sans biais [23], comme le tirage des ZAE est de taille fixe.

• Une seconde série S de $1\,300\,000$ simulations a été réalisée pour évaluer la qualité des estimateurs de variance (biais et dispersion), via des probabilités d'inclusion estimées également par Breidt-Chauvet. Les deux lots de simulations (R et S) sont indépendants.

• La validation utilise l'estimation des probabilités d'inclusion $\widehat{\pi}_2^{\,BC,S}$, notamment pour le calcul d'une variance de référence suivant la formule (**eI.1.8**). En effet, d'une part l'étude portant sur les petites communes a montré que cet estimateur conduisait à des résultats très proches de ceux obtenus avec l'estimateur de variance par simulation « classique » (**eI.1.4**), et d'autre part la précision de l'estimateur des probabilités d'inclusion par la méthode de Breidt-Chauvet semble meilleure que celle par la fréquence (cf. Tableau 23). Dès lors, il semble préférable de retenir l'estimateur (**eI.1.8**), qui permet en outre d'utiliser les simulations effectuées pour la validation conjointement avec les réplications pour procéder à l'estimation finale des probabilités d'inclusion doubles.

24. C'est-à-dire qu'il converge vers un estimateur sans biais lorsque le nombre de réplications tend vers l'infini.

III.4 qualité de l'estimation de variance des variables d'équilibrage des ZAE

• Le biais de l'estimation de variance semble encore diminuer légèrement lorsque le nombre de réplications augmente de 50 000 à 100 000 (Tableau 25). L'augmentation apparente du biais lorsque le nombre de réplications est accru au delà de 700 000 s'explique sans doute par le nombre limité de simulations utilisé pour calculer le biais. Même si le biais estimé est ici marginal (très inférieur à 1% de l'écart-type à estimer dès 50 000 réplications), l'approche conserve un intérêt pour choisir le nombre de réplications utilisées pour estimer les probabilités d'inclusion.

Tableau 25 – Biais relatif de l'estimateur de la variance des totaux d'équilibrage estimés, en % de l'écart-type simulé

nombre réplications	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
1 000 000	0,0	0,1	-0,1	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,1	
700 000	0,0	0,0	-0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	
300 000	0,0	0,0	0,0	-0,1	0,1	0,0	0,0	0,1	0,0	0,0	0,0	
100 000	0,0	0,1	0,0	-0,1	0,0	0,0	-0,1	0,1	0,0	0,0	0,0	
50 000	0,3	0,2	0,2	0,2	0,0	0,0	0,2	0,2	0,1	0,1	0,1	
10 000	1,2	0,6	1,1	1,2	0,8	1,3	1,1	0,8	1,0	0,9	1,2	

Notes :

– *Il s'agit du taux d'écart entre la racine carrée de la moyenne des estimations 'échantillon' de variance sur 1 300 000 simulations, par rapport à celle de la variance du total estimé calculée sur ces mêmes simulations.*

– *Le programme est ci-joint* ***biais_relatif*** + ***biais_relatif_equilib_yg***.

– La troncature à 10^{-3} des estimations des probabilités d'inclusion induit une sous-estimation de la variance (Tableau 26). C'est compréhensible, comme l'estimateur de Yates-Grundy est fonction décroissante des probabilités d'inclusion double. Ce traitement rend l'estimation de variance quasiment insensible au nombre de réplications utilisées pour estimer les probabilités d'inclusion, au delà de 50 000 [24].

Tableau 26 – Biais relatif de l'estimateur de la variance avec troncature à 10^{-3} des probabilités d'inclusion

nombre réplications	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
1 000 000	-0,5	-0,6	-0,4	-0,5	-1,3	-0,9	-0,5	-0,7	-0,4	-0,4	-0,4	
700 000	-0,5	-0,6	-0,4	-0,5	-1,3	-0,9	-0,5	-0,7	-0,4	-0,4	-0,4	
300 000	-0,5	-0,6	-0,4	-0,5	-1,3	-0,9	-0,5	-0,7	-0,4	-0,4	-0,4	
100 000	-0,5	-0,5	-0,3	-0,5	-1,2	-0,8	-0,5	-0,7	-0,3	-0,4	-0,4	
50 000	-0,3	-0,4	-0,2	-0,3	-1,2	-0,8	-0,3	-0,6	-0,3	-0,3	-0,3	

Note : Le biais est deux fois plus élevé si les termes de probabilités d'inclusion double inférieures au seuil sont annulés.

25. Ce résultat suggère que les nombres élevés de réplications n'ont d'effet sensible que sur les petites probabilités d'inclusion.

– Une troncature plus limitée des probabilités d'inclusion, à 10^{-4}, donne un biais apparent à la fois faible et stable en fonction du nombre de réplications (Tableau 27).

Tableau 27 – Biais relatif de l'estimateur de la variance avec troncature à 10^{-4} des probabilités d'inclusion

nombre réplications	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
1 000 000	-0,1	-0,1	-0,1	-0,1	0,0	-0,1	-0,1	0,0	-0,1	0,0	-0,1	
700 000	-0,1	-0,1	-0,1	-0,1	0,0	0,0	-0,1	0,0	0,0	0,0	-0,1	
300 000	0,0	-0,1	-0,1	-0,1	0,0	0,0	-0,1	0,0	0,0	0,0	-0,1	
100 000	0,1	0,1	0,0	-0,1	0,1	0,0	-0,1	0,1	0,0	0,0	-0,1	
50 000	0,2	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,0	

– La troncature à 10^{-4} impacte essentiellement la Corse. Pour cette région, la troncature mène à sous-estimer la variance des variables d'équilibrage, en contrepartie de l'élimination de surestimations (Tableau 28).

Tableau 28 – Biais relatif de la variance estimée pour la Corse, en % de l'écart-type simulé

troncature	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
avec	-2,8	-2,6	-6,3	-3,3	-0,4	-1,9	-2,7	-0,4	-1,5	-3,3	-4,4	
sans	1,1	4,0	-4,2	-0,9	1,7	-0,9	0,7	2,7	0,0	2,3	5,3	

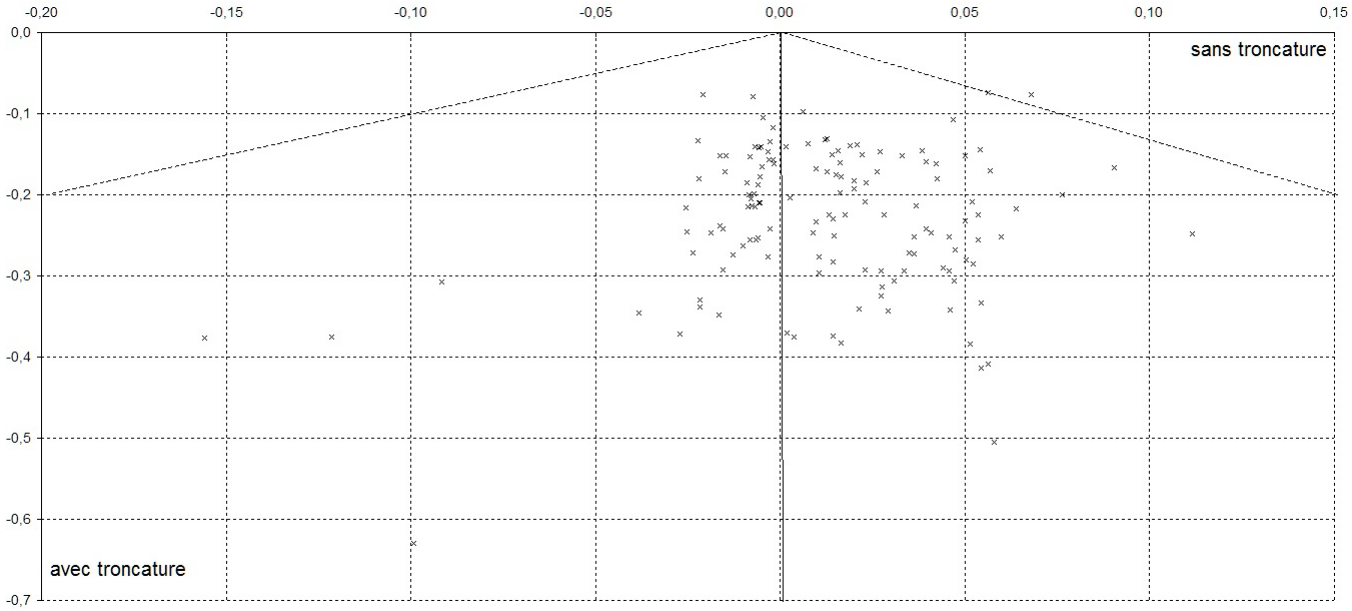
Note : Le biais mesuré pour la Corse sur 300 000 *simulations est beaucoup plus élevé, par exemple 17.6% pour* `nresgr2` *sans troncature.*

– La comparaison des deux troncatures est prolongée ci-après sur un ensemble de 138 variables socio-démographiques estimées par le recensement de 2009 [25].

26. Le cas de la ZAE-GC de Saint-Pol-sur-Mer (59540), commune rattachée en 2010 à Dunkerque (59183), a été traité ici par annulation des variables d'intérêt.

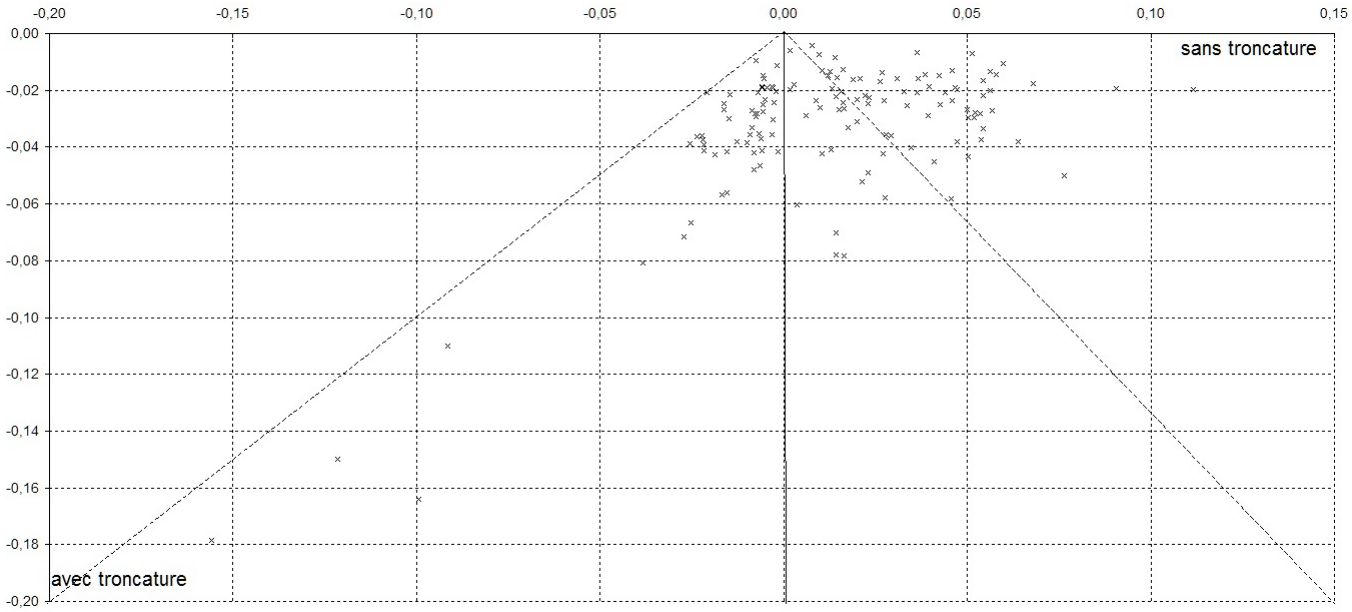
→ La troncature 'large' présente l'inconvénient d'induire une sous-estimation de la variance de toutes ces variables (Graphique 7).

Graphique 7 – Impact de la troncature à 10^{-3} sur le biais de l'estimateur de variance, en % de l'écart-type simulé



→ La troncature plus limitée apparait également comme une option non conservatrice, bien que la sous-estimation induite soit faible (Graphique 8).

Graphique 8 – Impact de la troncature à 10^{-4} sur le biais de l'estimateur de variance, en % de l'écart-type simulé



• Le biais de l'estimateur Horvitz-Thompson de la variance semble excéder celui de Yates-Grundy pour la grande majorité des variables d'équilibrage (Tableau 29). En fait cet estimateur est beaucoup plus sensible au nombre de réplications utilisées pour estimer les probabilités d'inclusion. La mesure de son biais est également plus dépendante du nombre et du lot de simulations utilisées.

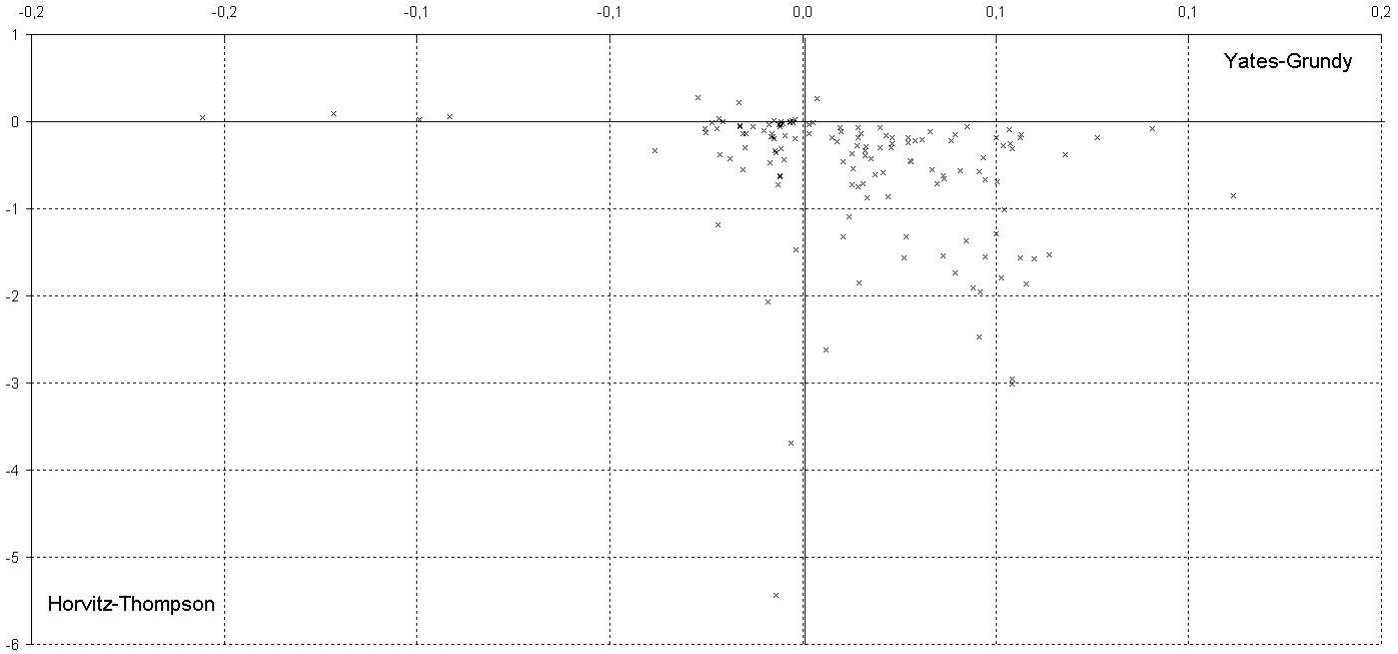
Tableau 29 – Comparaison du biais des estimateurs de variance de Yates-Grundy et Horvitz-Thompson

estimateur X réplications	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	0,0	0,1	-0,1	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,1	
HT 1 000 000	0,5	-0,2	-0,1	-1,3	-0,7	0,0	0,6	0,0	-0,1	-0,7	0,0	
700 000	0,7	0,5	-0,5	0,0	0,5	-0,2	0,3	0,3	-0,4	-0,1	-0,4	
300 000	1,3	-0,5	0,1	0,0	1,0	0,0	0,8	-0,3	0,1	-0,4	0,1	
100 000	7,9	0,1	-1,2	1,6	1,8	0,5	5,0	-0,2	-0,5	1,4	0,8	
50 000	0,5	-3,8	0,6	-0,1	-0,7	-0,6	0,3	-2,1	0,3	0,6	0,1	

Note : Les 5 dernières lignes montrent l'incidence du nombre de réplications sur le biais estimé de l'estimateur de variance d'Horvitz-Thompson. Le lot de 100 000 *réplications est ici disjoint de celui de* 50 000.

→ Pour la plupart des variables étudiées issues des recensements, l'estimateur d'Horvitz-Thompson semble sous-estimer la variance simulée (Graphique 9), selon la mesure du biais retenue ici.

Graphique 9 – Comparaison du biais estimé des deux estimateurs de variance



• La sous-estimation fréquente de l'estimateur de variance d'Horvitz-Thompson pourrait s'expliquer par l'insuffisance du nombre de réplications utilisées pour estimer les probabilités d'inclusion ainsi que par la convexité de la fonction $x \longmapsto \frac{1}{x}$, qui entraine que $E\left(\frac{1}{\widehat{\pi_2}^R}\right) > \frac{1}{\pi_2}$. Cette explication s'applique également au signe majoritairement positif du biais de l'estimateur de Yates-Grundy.

III.5 mesure de la dispersion des estimateurs de variance des ZAE

– Le lot S_d des simulations stockées pour la mesure de dispersion (**eI.4.17**) a été produit par le programme **zae_s_simules**.

– La dispersion de l'estimateur de variance est nettement plus élevée pour la Corse (94) et, dans une moindre mesure, le Limousin (74), ce qui s'explique sans doute par leur petit nombre de ZAE (Tableau 30). Mais cette taille ne suffit visiblement pas à expliquer les différences de dispersion de cet estimateur entre les régions, vu la dispersion relativement élevée de la région Rhône-Alpes (82).

Tableau 30 – Dispersion de l'estimateur de variance de Yates-Grundy mesuré sur 10 000 simulations, en % de l'écart-type simulé de l'EM

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	118,4	103,6	109,7	108,7	0,0	0,0		108,7	119,9	104,3	93,3	129,8
11_gc	189,9	188,4	166,2	175,5	169,6	182,4		186,5	146,7	170,1	171,9	165,9
21	152,3	158,0	149,6	146,3	223,6	178,8		160,9	158,6	147,5	149,9	141,1
22	155,8	148,7	146,5	176,1	161,2	155,1		146,6	148,2	136,2	163,5	146,5
23	190,0	169,8	173,7	159,1	174,0	177,7		178,0	170,1	149,8	156,1	154,2
24	159,8	170,2	181,8	159,9	150,1	160,7		161,7	172,9	176,7	149,7	165,5
25	150,9	147,6	157,1	157,9	191,9	168,8		148,7	152,1	152,9	158,4	156,3
26	162,6	180,0	166,8	160,8	157,9	178,4		161,9	156,7	166,3	154,1	156,5
31	197,6	194,7	184,0	172,4	131,3	164,1		185,2	164,7	163,0	149,5	168,7
41	185,3	181,3	179,1	168,5	159,5	158,5		169,8	166,0	177,6	168,3	147,4
42	174,5	166,4	199,6	174,8	171,7	167,0		172,6	157,1	190,7	159,7	170,0
43	170,6	141,5	158,8	159,7	311,9	186,4		163,1	142,5	159,5	159,0	160,5
52	195,1	181,6	166,8	159,0	156,7	146,5		190,6	182,5	164,2	157,2	159,9
53	176,9	177,6	190,6	153,3	157,6	148,5		154,0	173,9	177,1	156,0	170,5
54	172,2	181,5	186,1	165,1	151,9	151,8		166,4	178,2	178,5	149,7	175,0
72	168,8	177,2	164,1	157,2	157,2	161,7		146,9	179,9	156,4	155,2	148,0
73	164,8	174,8	166,9	160,2	153,1	168,7		163,4	159,2	159,1	148,5	146,7
74	204,7	214,2	222,8	183,4	175,5	209,1		203,4	198,4	213,5	171,3	197,2
82	213,4	189,1	206,4	200,6	162,7	176,8		194,2	187,7	178,9	169,5	180,2
83	174,3	165,9	159,6	172,1	161,9	176,3		179,3	157,4	160,6	172,9	154,7
91	145,0	177,6	160,6	159,2	156,6	154,6		150,8	156,5	156,6	150,2	151,7
93	170,7	164,5	174,4	158,9	147,4	157,5		173,6	156,3	170,0	134,9	159,9
94	309,0	316,9	331,5	381,7	250,9	288,8		238,9	398,8	230,5	313,1	267,8
total	81,8	82,6	81,5	88,2	89,5	81,8		78,9	89,3	76,4	74,7	77,0

Note :

– *La dispersion est calculée avec les probabilités d'inclusion estimées sur* 1 000 000 *réplications. Un calcul alternatif sur les* 2 300 000 *réplications donne des dispersions peu différentes pour ces variables, avec un écart entre les deux dispersions compris entre -0.4 et +1.9 point de pourcentage. Les programmes utilisés sont* **dispersion** + **dispersion_equilib_yg**.

– *Le taux d'écart entre la moyenne des estimations de variance sur les* 10 000 *échantillons simulés et la variance de référence simulée est inférieur à 1% (en racines carrées)*

– *L'indicateur décroit légèrement (2 points au niveau national) si les estimations de variance négatives sont annulées. (Le biais augmente très sensiblement.)*

→ La dispersion de l'estimateur d'Horvitz-Thompson est beaucoup plus élevée que celle de Yates-Grundy (Tableau 31), bien que dans une proportion moindre que pour les petites communes. L'incidence de la troncature sur la dispersion est plus forte que pour l'estimateur de Yates-Grundy, mais elle ne suffit pas à résorber l'écart entre leurs variabilités.

Tableau 31 – Dispersion relative des estimateurs de variance des variables d'équilibrage

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	81,8	82,6	81,5	88,2	89,5	81,8	78,9	89,3	76,4	74,7	77,0	
YGt	80,5	79,8	79,9	77,2	76,9	77,2	78,3	77,8	75,8	73,5	75,9	
YGt2	81,7	82,0	80,5	81,1	89,3	81,6	78,9	85,5	76,1	74,1	76,6	
HT	186,1	232,3	168,0	180,6	135,6	133,0	170,8	189,0	148,0	154,5	171,3	
HTt	169,8	174,5	161,2	154,4	116,1	107,0	158,4	160,4	142,6	140,5	143,3	
HTt2	184,6	231,6	165,1	170,0	126,9	114,7	169,1	188,3	146,5	149,6	149,1	

Note : Le suffixe t représente la troncature des probabilités d'inclusion à 10^{-3} *(qui donne* $\max\left[\widehat{\pi}_2, 10^{-3}\right]$*), et t2 la troncature à* 10^{-4}*. L'annulation réduit légèrement la dispersion par rapport à la troncature, mais augmente le biais. Pour les variables 'recensement', l'annulation réduit la dispersion par rapport à la troncature dans 98 cas, et l'augmente pour 40 variables.*

note : La dispersion plus faible de l'estimateur de Yates-Grundy par rapport à Horvitz-Thompson est observée ici pour des variables qui s'avèrent nettement corrélées à la probabilité d'inclusion, sauf `nres_rural` et `nres_periurbain` [26]. La comparaison sur 138 variables issues du recensement montre un écart faible entre les deux dispersions pour les variables de corrélation absolue avec π_1 inférieure à 0.2. Cependant, l'estimateur de variance d'Horvitz-Thompson reste plus dispersé pour la quasi-totalité de ces variables.

– Le nombre de réplications utilisées pour l'estimation 'échantillon' de variance n'a pas d'effet apparent sur la dispersion de la variance estimée par Yates-Grundy (Tableau 32). Une explication possible est que la dispersion est mesurée sur un nombre trop limité de simulations (10 000). Néanmoins, l'effet réel du nombre de réplications devrait être faible, compte tenu de ces résultats.

Tableau 32 – Effet du nombre de réplications sur la dispersion relative estimée par Yates-Grundy

$\lvert R\rvert$	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
1 000 000	81,8	82,6	81,5	88,2	89,5	81,8	78,9	89,3	76,4	74,7	77,0	
700 000	81,7	83,5	81,2	91,5	88,6	82,6	78,9	92,0	76,2	75,1	77,1	
300 000	81,9	81,8	81,1	84,0	94,0	81,6	79,0	87,5	76,2	74,4	76,4	
100 000	82,3	81,9	80,5	77,9	90,0	81,9	78,9	87,2	76,1	73,9	76,4	
50 000	81,9	81,9	80,8	79,1	112,9	82,0	79,1	85,7	76,4	73,9	76,5	

– En occultant l'incidence des simulations, la dispersion peut s'analyser par l'équation :

$$E\left[\left(\widehat{V}^{s,R} - V\right)^2 \middle| R\right] = E\left\{\left(\widehat{V}^{s,R} - E\left(\widehat{V}^{s,R}\middle| R\right)\right)^2 \middle| R\right\} + \left(E\left(\widehat{V}^{s,R}\middle| R\right) - V\right)^2$$

La faiblesse de l'effet du nombre de réplications sur la dispersion de Yates-Grundy (et, dans une moindre mesure, d'Horvitz-Thompson) pourrait s'expliquer par la prédominance de la première composante, qui représente l'effet de l'aléa de l'échantillonnage sur l'estimation de variance, par rapport à celle du biais induit par l'estimation des probabilités d'inclusion.

27. L'estimateur Yates-Grundy de la variance d'une variable fonction linéaire de la probabilité d'inclusion est nul et d'erreur quadratique moyenne nulle.

• Pour 7 des 11 variables d'équilibrage, la dispersion de l'estimateur Breidt-Chauvet de la variance est supérieure à celle d'un estimateur nul sur toutes les régions sauf la petite couronne d'Ile-de-France, la Franche-Comté, le Limousin et la Corse, pour lesquelles l'estimation de Deville est appliqué.

Ce résultat illustre les limites de l'approche du choix de l'estimateur de variance selon la dispersion mesurée sur les variables d'équilibrage. Plus généralement, compte tenu du nombre limité de simulations exploitées pour le calcul de dispersion, les petites différences observées pour cet indicateur sont à relativiser.

III.6 calcul de la variance sur l'échantillon effectif des ZAE-EM

• La variance estimée sur l'échantillon de ZAE effectivement tiré pour l'EM surestime considérablement la variance simulée de certaines variables d'équilibrage (Tableau 33). L'ampleur de cette surestimation dépasse largement celle de l'erreur observée sur deux échantillons simulés. Elle est sensiblement réduite lorsque la troncature de 10^{-3} est appliquée aux probabilités d'inclusion estimées.

Tableau 33 – Taux d'erreur de l'estimation de l'écart-type observée sur l'échantillon maître effectivement tiré et sur deux échantillons simulés, en % de l'écart-type simulé

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	10,8	-8,9	-74,1	4,0	-100,0	-100,0		21,8	40,0	-54,0	-33,3	15,1
11_gc	-100,0	87,5	44,3	71,3	-100,0	-100,0		-100,0	-2,5	-41,8	10,7	-22,5
21	8,5	28,6	-100,0	34,0	-100,0	-100,0		-10,5	-1,2	-94,8	16,5	148,3
22	52,6	80,0	42,8	26,3	1,5	-100,0		-100,0	65,8	8,1	-100,0	-42,6
23	-100,0	31,1	68,4	19,7	36,3	173,4		-100,0	83,2	12,0	3,1	-6,7
24	-25,2	-100,0	62,0	27,9	84,1	44,2		-27,8	-100,0	55,4	0,6	-100,0
25	20,1	-33,5	44,2	6,2	-100,0	72,7		23,5	17,9	7,6	10,8	-32,3
26	-100,0	-100,0	-100,0	-100,0	123,1	-39,0		-100,0	-100,0	-17,7	-100,0	2,4
31	135,2	8,4	162,8	114,6	10,6	-100,0		142,3	75,1	109,3	35,9	2,2
41	-25,1	-56,8	61,8	63,1	82,7	147,8		-10,4	72,7	68,2	55,1	56,0
42	-100,0	-100,0	-28,0	-100,0	-31,5	-100,0		-100,0	-100,0	-39,4	-65,9	-100,0
43	99,6	90,0	-100,0	117,6	203,4	99,8		144,3	8,8	-100,0	132,0	146,0
52	171,4	-100,0	108,2	106,1	56,7	-100,0		200,9	-100,0	36,0	152,8	180,2
53	-100,0	102,4	-100,0	-43,5	-100,0	-100,0		-100,0	84,0	-100,0	-20,3	-23,0
54	185,6	-100,0	-100,0	78,3	-100,0	-100,0		163,8	-100,0	51,2	44,0	84,1
72	114,3	57,6	-100,0	-100,0	140,7	104,7		71,0	71,0	-65,7	-0,8	70,2
73	55,5	-100,0	79,7	55,7	82,9	61,6		-10,1	-100,0	126,4	24,4	36,2
74	239,7	-13,4	83,8	11,3	19,1	46,7		331,6	95,4	215,3	-41,4	430,8
82	-100,0	-100,0	-100,0	-100,0	19,1	51,4		-100,0	-100,0	-100,0	-25,3	-28,4
83	-23,5	-62,4	12,4	-36,6	27,6	-37,0		36,9	-4,6	-83,0	-100,0	52,9
91	-100,0	11,7	2,5	-100,0	-100,0	-43,3		-100,0	15,5	-25,2	-100,0	-4,3
93	147,5	27,5	-23,1	-100,0	89,1	93,4		91,5	59,5	34,5	-53,3	108,1
94	1 820,9	1 158,4	291,4	73,4	1 122,3	1 343,4		1 019,7	1 967,3	140,4	53,7	257,0
total	178,8	123,6	1,0	8,4	179,3	179,8		95,0	295,5	-8,8	4,9	73,9
$\widehat{\pi_2} \geqslant 10^{-3}$	73,8	43,8	-6,5	4,9	92,9	71,2		40,8	126,4	-12,8	3,8	69,0
$\widehat{\pi_2} \geqslant 10^{-4}$	178,8	123,6	1,0	8,4	179,3	179,8		95,0	295,5	-8,8	4,9	73,9
s_1	36,6	-33,9	-15,0	-15,9	-10,3	-32,6		21,4	-26,2	-7,4	-11,3	-6,4
s_2	12,8	-2,8	-29,5	-23,8	35,7	15,4		14,9	5,5	-28,6	-8,9	-6,2

Notes :

– *Les échantillons s_1 et s_2 ont été tirés indépendamment des deux lots de simulations S et R.*

– *Le taux d'erreur vaut -100% lorsque la variance estimée est négative.*

– Il s'avère que l'essentiel de cette surestimation provient de la Corse. En effet, l'annulation de la variance estimée de cette région ramène le taux d'erreur entre -14.4% et +66.6%. Les trois ZAE tirées en Corse, `Z2A006`, `Z2B033` et `Z2B134` ne se retrouvent dans aucun des 10 000 échantillons simulés, ni dans un deuxième lot indépendant de 10 000 simulations. La probabilité d'inclusion double minimale observée sur l'échantillon maître est presque 10 fois inférieure à celle observée pour les autres régions (Tableau 34).

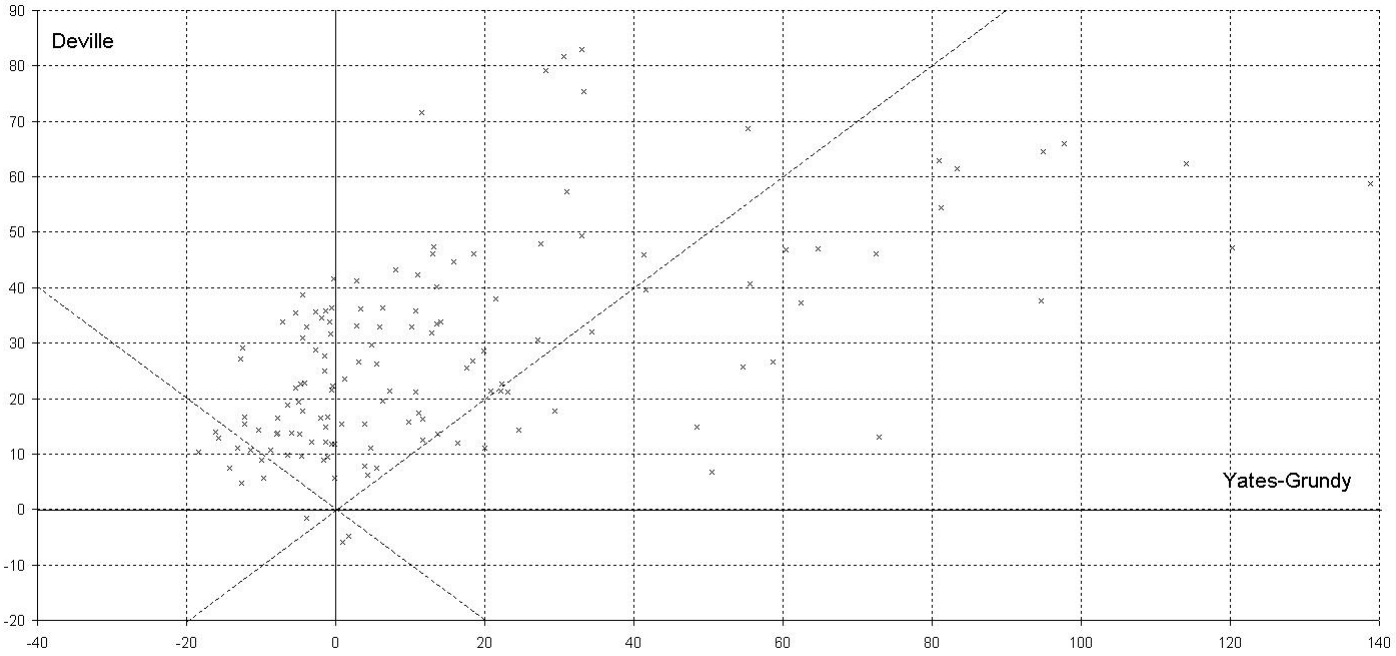
Tableau 34 – Probabilité d'inclusion double minimale, estimée sur 2 300 000 réplications

11 pc	11 gc	21	22	23	24	25	26	31	41	42	43	52	53	54	72	73	74	82	83	91	93	94
2,3E-02	3,9E-03	3,4E-03	2,1E-03	1,8E-03	3,0E-03	1,7E-03	1,8E-03	2,3E-03	1,0E-03	3,2E-03	2,1E-03	2,2E-03	3,9E-03	2,1E-03	3,0E-03	2,1E-03	2,7E-03	2,3E-03	2,1E-03	2,3E-03	6,8E-03	3,6E-04

– Cependant, le maximum du taux d'erreur sur l'estimation de variance de la Corse observé sur les 10 000 simulations est de 1 811% pour la variable `nresgr1`. Donc l'observation sur l'échantillon effectif des ZAE n'est pas incompatible avec le plan de sondage simulé.

– Sur l'échantillon effectif des ZAE, l'approximation de la variance par la formule de Deville surestime de plus de 10% l'écart-type simulé de 122 des variables issues des recensements, alors que l'erreur absolue de l'estimateur de Yates-Grundy est de moins de 10% dans 66 cas (Graphique 10) [27].

Graphique 10 – Comparaison du taux d'erreur de l'estimation sur l'échantillon effectif des ZAE, en % de l'écart-type simulé



note : Dans toute cette partie, l'estimation 'échantillon' de la variance est calculée avec les probabilités estimées sur les 1 000 000 réplications, et le calcul d'erreur est relatif à la variance estimée sur les 1 300 000 simulations. Le calcul sur l'ensemble des 2 300 000 réplications, pour les deux familles de probabilités d'inclusion double, ne modifie pas radicalement les résultats. (Cependant, les erreurs ainsi mesurées sont réduites, sauf pour deux des variables d'équilibrage.)

28. L'estimateur de Deville-Tillé sous-estime de plus de 10% la variance simulée de 111 de ces variables sur l'échantillon effectif, et pour 68 variables en moyenne sur 10 000 simulations.

III.7 comparaison aux approximations de Deville et Deville-Tillé

• Ces résultats incitent à retourner au choix de la méthode d'estimation de variance, pour explorer la possibilité de substituer partiellement l'approximation de Deville à l'estimation de Yates-Grundy.

– Au niveau national, la dispersion [28] de l'estimateur de Deville est largement plus élevée que celle de Yates-Grundy (Tableau 35). Toutefois, elle est beaucoup plus réduite pour la Corse. L'estimateur de Deville est également moins dispersé pour presque toutes les variables du Limousin,

Tableau 35 – Écart entre les dispersions relatives de l'estimateur de variance de Deville et de Yates-Grundy, en % de l'écart-type simulé, pour les variables d'équilibrage

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	-9,8	-11,5	-3,4	-7,1	0,0	0,0		122,2	136,6	58,3	91,6	165,9
11_gc	152,4	129,8	117,6	137,2	-14,0	312,5		170,6	100,2	144,0	157,5	139,7
21	50,8	68,9	48,8	56,4	15,5	52,5		37,8	69,7	46,4	44,1	45,4
22	99,9	84,0	96,7	98,0	117,1	136,0		75,2	69,1	40,4	103,5	87,0
23	69,9	67,4	80,2	72,9	39,7	113,5		78,9	76,7	58,7	73,9	68,2
24	134,2	136,2	143,2	128,1	184,6	163,7		118,7	134,7	143,4	89,9	127,5
25	85,2	67,4	86,8	64,4	80,3	79,6		80,9	68,4	83,7	58,7	78,0
26	103,4	90,1	104,8	98,6	127,9	85,5		101,8	65,6	105,7	90,0	67,4
31	157,7	139,0	139,4	118,7	110,6	205,5		169,6	139,0	141,2	122,6	156,5
41	136,3	136,8	119,6	153,4	138,2	151,0		113,3	121,9	125,2	145,1	103,7
42	46,8	66,4	74,3	58,9	17,1	95,2		57,9	65,7	73,1	62,1	68,9
43	45,0	32,1	31,1	47,8	-97,2	50,3		45,4	33,1	30,8	41,1	45,2
52	159,9	161,0	153,7	138,6	204,1	168,8		159,3	156,2	164,9	139,1	148,6
53	158,8	181,9	183,4	137,9	206,8	182,9		143,0	169,8	164,5	150,5	148,9
54	105,6	111,8	100,4	102,8	136,7	104,2		84,2	94,2	99,8	82,8	91,2
72	141,1	153,2	136,0	130,8	210,9	134,7		116,4	155,1	126,5	131,9	125,0
73	140,3	138,3	132,2	126,8	202,9	124,7		134,5	120,7	100,6	121,3	109,7
74	3,0	-17,5	-20,3	-23,8	-10,6	-55,5		-16,9	-21,9	-26,9	-22,4	-38,1
82	268,6	202,5	233,2	221,6	294,1	290,3		245,6	205,3	212,0	192,6	214,4
83	95,4	74,7	75,8	77,4	96,7	52,0		87,2	68,7	73,3	47,9	55,5
91	123,1	132,8	123,6	103,8	160,0	127,3		109,2	115,4	110,2	105,4	95,9
93	93,1	109,2	106,1	89,8	127,6	164,4		117,9	111,5	121,0	80,2	96,2
94	-161,6	-168,2	-193,2	-235,2	-140,6	-180,1		-99,3	-238,3	-96,1	-191,6	-150,2
total	192,2	181,8	189,9	167,0	199,4	218,7		190,2	163,1	180,5	170,4	178,9

→ Le remplacement de l'estimation de variance de la Corse par l'approximation de Deville fait baisser la dispersion mesurée de toutes les variables d'équilibrage (Tableau 36). Le traitement additionnel du Limousin a un effet plus mitigé et limité, donc il ne parait pas indispensable [29].

Tableau 36 – Incidence sur la dispersion des variables d'équilibrage du traitement de la Corse

traitement	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
avec	80,5	79,7	80,4	76,4	87,5	78,0		78,3	76,9	76,0	73,6	76,0
sans	81,8	82,6	81,5	88,2	89,5	81,8		78,9	89,3	76,4	74,7	77,0
variation	-1,3	-2,9	-1,0	-11,8	-1,9	-3,8		-0,6	-12,4	-0,4	-1,1	-1,0
+Limousin	0,4	-0,1	-0,3	-0,3	0,0	-1,1		-0,3	-0,1	-0,1	-0,7	-0,5

Note : La dernière ligne donne l'incidence supplémentaire sur la dispersion du remplacement de l'estimation de la variance du Limousin par l'approximation de Deville.

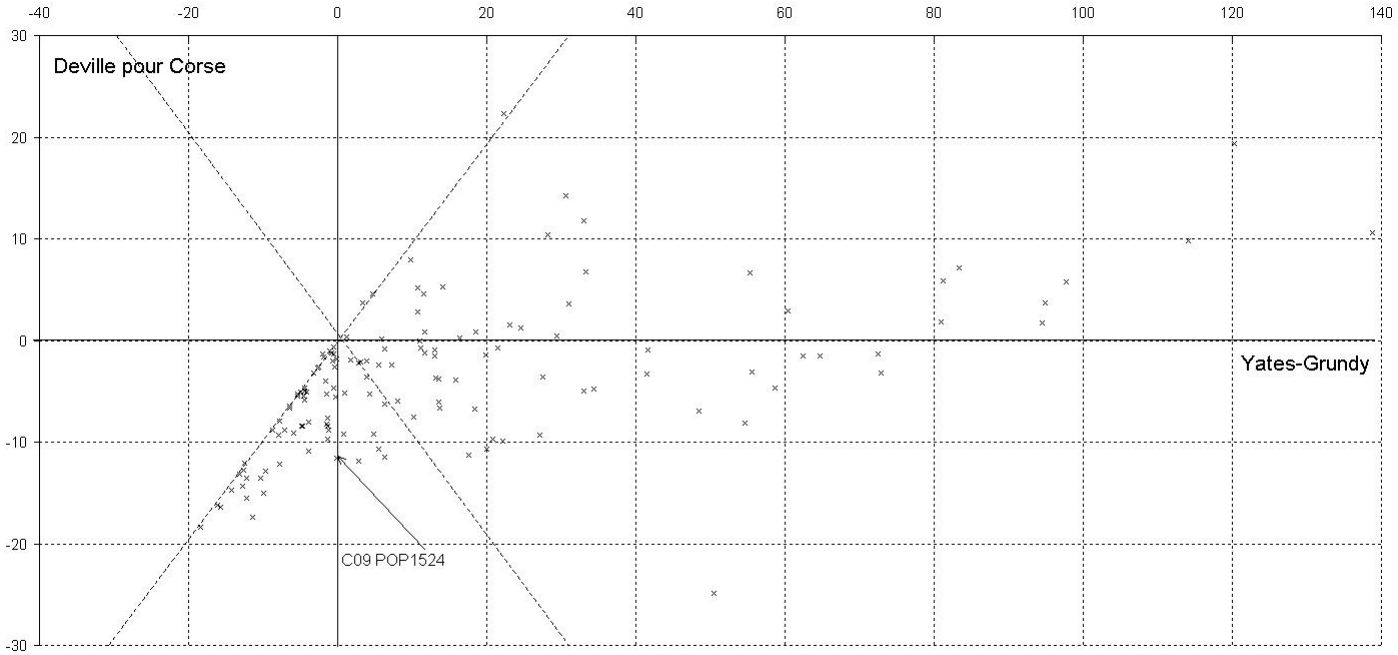
29. par rapport à la variance simulée, donc intégrant le biais

30. Cette conclusion est discutable : le traitement additionnel du Limousin abaisse la dispersion de 128 des 138 variables socio-démographiques étudiées, par rapport à celui limité à la Corse.

– Après le traitement de la Corse, l'effet de la troncature des probabilités d'inclusion à 10^{-3} sur la dispersion des variables d'équilibrage est négligeable, sauf pour `nres_rural` (-11.1 points) et `nres_periurbain` (-1.3). Une possibilité alternative pour réduire la dispersion pour la première de ces variables est d'appliquer l'approximation de Deville à la Franche-Comté (43), vu le Tableau 35. Mais cette extension accroit la dispersion d'autres variables.

→ Le traitement de la Corse permet d'éviter les très fortes surestimations observées sur l'échantillon effectif pour certaines variables issues des recensements (Graphique 11). En contrepartie, il induit dans certains cas des sous-estimations qui paraissent tolérables [30].

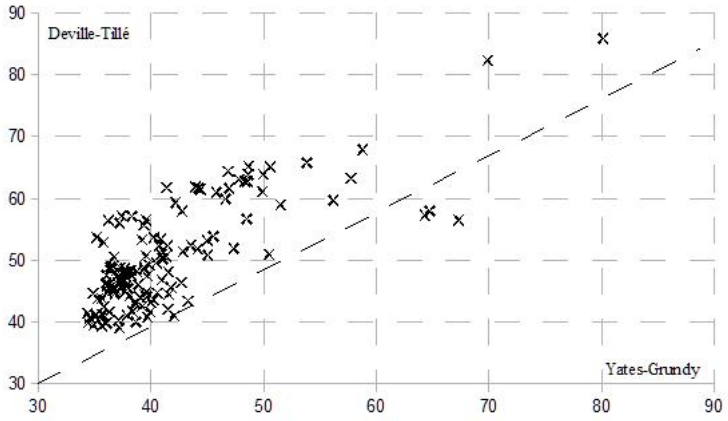
Graphique 11 – Comparaison du taux d'erreur de l'estimation sur l'échantillon effectif des ZAE, en % de l'écart-type simulé



31. Pour la variable `c09_pop1524`, l'erreur de l'estimation de Deville sur l'échantillon effectif de la Corse est de -54% de l'écart-type simulé, alors que celle de Yates-Grundy est de 314%. L'erreur globale de ce dernier estimateur est quasiment nulle du fait de la sous-estimation compensatrice d'autres régions, notamment la Champagne-Ardennes (-68.5%)

– L'estimateur de variance retenu limite la dispersion par rapport à Deville-Tillé pour la grande majorité des variables 'recensement' étudiées, au niveau national (Graphique 12). Ce n'est le cas que pour 7 régions.

Graphique 12 – Dispersion mesurée pour les variables 'recensement', en % de l'écart-type simulé



III.8 conclusions sur la variance de l'échantillon maître des ZAE

- L'estimation de variance de l'échantillon maître tiré dans les ZAE peut être correctement réalisée avec l'estimateur de Yates-Grundy appliqué aux probabilités d'inclusion double estimées par la méthode de Breidt-Chauvet sur un million de réplications.
  - Le biais estimé est quasiment nul pour les variables étudiées.
  - Comme pour le tirage du groupe de rotation PC-RP, la procédure de Breidt et Chauvet parait nettement plus performante que l'estimation par la fréquence des probabilités d'inclusion.
- Cependant, la dispersion semble excessive. Il parait utile de particulariser l'estimateur de variance de la Corse pour lui appliquer l'approximation de Deville.
- Une troncature des petites probabilités estimées stabilise l'estimation de variance, et contourne le problème précédent si le seuil est de 10^{-3}. Mais cette option présente l'inconvénient d'induire une sous-estimation de la variance, faible mais peut-être pas négligeable pour certaines variables d'intérêt. De toute manière, elle est sans incidence sur l'échantillon effectif des ZAE hors Corse.
  - Les simulations utilisées pour la validation peuvent être exploitées pour estimer les probabilités d'inclusion définitives. Celles-ci sont donc calculées sur $2\,300\,000$ réplications.

Partie IV
une application : estimation de la variance de l'enquête AES 2012

Cette partie décrit l'estimation de variance de l'enquête AES 2012 (Adult Education Survey), tirée par Octopusse. La variance due au tirage du groupe de rotation des petites communes (PC) et de l'échantillon maître des ZAE est estimée par les méthodes décrites précédemment.

IV.1 étapes du tirage d'AES et des redressements - notations spécifiques

① tirage Octopusse dans l'EAR 2010

– s désigne ici l'échantillon de logements (résidences principales pour l'enquête de recensement, avant éclatement des logements et séparation des budgets). Il est confondu avec l'indicatrice correspondante ($l \longmapsto s_l \in \{0,1\}$). s_B est l'échantillon des ménages correspondants.

– $l \in L$ indice les logements et $b \in B$ les ménages-budgets séparés

– w est le poids de tirage du logement [31]

– Octopusse fournit ce poids (variable nommée `poids_final_0` dans le fichier-échantillon d'AES 2012), au calage des ZAE près.

– Gr désigne le découpage en 5 groupes de rotation des petites communes (GR-PC), et gr est le groupe de rotation de la campagne utilisée. $\pi_{1,\text{gr}} = 1/5$ est la probabilité d'inclusion dans le GR-PC. (Le groupe de rotation des adresses recensées dans les grandes communes n'est pas pris en compte explicitement dans le calcul de variance [32].)

– Zae = s_{ZAE} est l'échantillon-maître des ZAE ; zae désigne une ZAE particulière. $\pi_{1,\text{Zae}}$ est la probabilité d'inclusion dans s_{ZAE}. A noter que les ZAE sont définies conditionnellement au tirage des 5 groupes de rotation de petites communes. AES 2012 n'est pas affectée par les modifications géographiques des ZAE.

– s^{ne}_{ZAE} est l'échantillon des ZAE non exhaustives ($\pi_{1,\text{Zae}} < 1$)

– Le tirage Octopusse est stratifié par ZAE.

– Pour AES, ce tirage est effectué dans la dernière campagne de recensement disponible (EAR 2010).

→ Le total Y d'une variable d'intérêt y est estimable sur l'échantillon s par :

$$\widehat{Y}^s = \sum_{l \in s} w_l y_l = \sum_{l \in s} \frac{1}{\pi_{1,\text{gr}} \pi_{1,\text{Zae}} \pi_s\,(l|\text{Gr},\text{Zae})}(l) y_l$$

y_l désigne le total par logement, y_b celui par ménage et y_i la valeur pour un individu. y_L est la fonction qui donne les totaux par logement. $\pi_{s|\text{Gr},\text{Zae}}[l] = p\,(l \in s|\text{Gr},\text{Zae})$

② modélisation logit de la probabilité de réponse au niveau ménage

– r est l'échantillon des ménages répondants (y compris logements et ménages hors champ)

– x désigne le vecteur de la régression logistique

→ L'estimateur sur l'échantillon des ménages répondants est donné par :

$$\widehat{Y}^r = \sum_r w_r y = \sum_{b \in r} \frac{w(b)}{\widehat{\pi}_r(b)} y(b)$$

③ sélection d'un habitant par ménage répondant

– s_I est l'échantillon des individus échantillonnés

– Les tirages d'individus sont indépendants entre les ménages conditionnellement à s.

→ L'estimateur sur l'échantillon des individus répondants vaut :

$$\widehat{Y}^{s_I} = \sum_{i \in s_I} w_i y_i = \sum_{b \in r} \frac{w_r}{\pi_{i|b}} y_i$$

32. Il est calculé par l'inverse d'un produit de probabilités conditionnelles, qui donne un estimateur par expansion.

33. Le signe de l'effet de cette omission sur l'estimation de variance n'est pas déterminé. D'un côté, l'équilibrage des groupes de rotation de petites adresses connues et le tirage systématique Octopusse limitent la variance, d'un autre côté l'effet de grappe des adresses, ainsi que l'hétérogénéité des groupes de rotation des grandes adresses l'augmentent.

④ calage

– x_{cal} est le vecteur de calage

– $\widehat{Y}^{rc}$ désigne l'estimation sur l'échantillon des répondants et calée du total d'une variable d'intérêt y, pondérée par w_{rc}. L'objet de la note est d'estimer $\text{Var}\left(\widehat{Y}^{rc}\right)$ en fonction des valeurs observées $y\,(r)$.

note : Les échantillons de réserve n'ont pas été mobilisés pour AES 2012.

IV.2 décomposition de la variance de niveau logement

• La variance de l'échantillon de logements (pondéré par w) se décompose en conditionnant par le découpage en GR-PC et par l'échantillon des ZAE selon (**eIV.2.18**).

$$\text{Var}\left(\widehat{Y}^{s}\right) = \text{Var}\left[E\left(\widehat{Y}^{s}\middle|\text{Gr},\text{Zae}\right)\right] + E\left[\text{Var}\left(\widehat{Y}^{s}\middle|\text{Gr},\text{Zae}\right)\right] \qquad \textbf{(eIV.2.18)}$$

$$\text{soit : } V_s = V_{\text{gr,Zae}} + V_{s|\text{Gr,Zae}}$$

▷ Le premier terme de (**eIV.2.18**), $V_{\text{gr,Zae}}$, représente la variance due aux tirages du groupe de rotation des petites communes et de l'échantillon des ZAE. Il peut à nouveau se décomposer en conditionnant par le découpage en GR-PC.

$$\begin{aligned} V_{\text{gr,Zae}} &= \text{Var}\left(\widehat{Y}^{\text{gr,Zae}}\right) = \text{Var}\left[E\left(\widehat{Y}^{\text{gr,Zae}}\middle|\text{Gr}\right)\right] + E\left[\text{Var}\left(\widehat{Y}^{\text{gr,Zae}}\middle|\text{Gr}\right)\right] \\ &= \text{Var}\left(\widehat{Y}^{\text{gr}}\right) + E\left[\text{Var}\left(\widehat{Y}^{\text{gr,Zae}}\middle|\text{Gr}\right)\right] \\ &= V_{\text{gr}} + V_{\text{Zae}} \end{aligned}$$

IV.2.1 estimation de la variance intra-ZAE ($V_{s|\text{Gr,Zae}}$)

▷ Comme le tirage Octopusse est stratifié par ZAE (avec distinction des arrondissements de PLM) et que $\widehat{Y}^{s}$ est un estimateur par expansion [33] :

$$\text{Var}\left(\widehat{Y}^{s}\middle|\text{Gr},\text{Zae}\right) = \sum_{\text{Zae}} \text{Var}\left(\frac{\widehat{y_{\text{Zae}}}^{s}}{\pi_{1,\text{Zae}}}\middle|\text{Gr},\text{Zae}\right)$$

▷ Vu que la taille de l'échantillon des ménages-RP par ZAE ($|s \cap \text{zae}|$) est fixe, conditionnellement aux groupes de rotation et aux ZAE, que $w = \dfrac{1}{\pi_{1,\text{Zae}}\pi_{1,\text{gr}}\pi_{s|\text{Gr,Zae}}}$, et que $\dfrac{\widehat{y_{\text{zae}}}^{s}}{\pi_{1,\text{Zae}}} = \sum_{s\cap\text{zae}} \dfrac{y_L}{\pi_{1,\text{Zae}}\pi_{1,\text{gr}}\pi_{s|\text{Gr,Zae}}} = \sum_{s\cap\text{zae}} \dfrac{y_L/(\pi_{1,\text{Zae}}\pi_{1,\text{gr}})}{\pi_{s|\text{Gr,Zae}}} = \sum_{s\cap\text{zae}} \dfrac{wy_L\pi_{s|\text{Gr,Zae}}}{\pi_{s|\text{Gr,Zae}}}$, $\text{Var}\left(\dfrac{\widehat{y_{\text{zae}}}^{s}}{\pi_{1,\text{Zae}}}\middle|\text{Gr},\text{Zae}\right)$ peut être approximée en fonction de l'échantillon par la formule de Deville (**eIV.2.19**).

$$\text{Var}\left(\frac{\widehat{\widehat{y_{\text{zae}}}^{s}}}{\pi_{1,\text{Zae}}}\middle|\text{Gr},\text{Zae}\right) = \frac{|s\cap\text{zae}|}{|s\cap\text{zae}|-1}\sum_{s\cap\text{zae}}(1-\pi_{s|\text{Gr,Zae}})\left(wy_L - \frac{\sum_{s\cap\text{zae}}(1-\pi_{s|\text{Gr,Zae}})wy_L}{\sum_{s\cap\text{zae}}(1-\pi_{s|\text{Gr,Zae}})}\right)^2 \qquad \textbf{(eIV.2.19)}$$

Ceci fournit la forme quadratique estimant le deuxième terme de (**eIV.2.18**), $V_{s|\text{Gr,Zae}}$.

34. Le calage du poids des ZAE n'est pas pris en compte dans le calcul de variance d'AES.

– Pour les ZAE-PC, une alternative est d'appliquer la formule d'approximation de la variance d'un tirage systématique.

IV.2.2 estimation de la variance du groupe de rotation de petites communes, par région

L'estimateur de Yates-Grundy généralisé permet de tenir compte de la variabilité de la taille de l'échantillon des petites communes, par le deuxième terme de l'équation (**eIV.2.20**) ci-après.

$$\widehat{V}_{\text{gr}}^{\text{gr}} = \left(-\frac{1}{2}\right) \sum_{c_1,c_2 \in \text{gr}} \widehat{\widehat{\Delta}}_{\text{gr}}(c_1,c_2) \left(\frac{y_C}{\pi_{1,\text{gr}}}(c_1) - \frac{y_C}{\pi_{1,\text{gr}}}(c_2)\right)^2 + \sum_{c \in \text{gr}} \left(\frac{y_c}{\pi_{1,\text{gr}}}\right)^2 \left(\frac{\widehat{\pi_{2,\text{gr}}}(c,+)}{\pi_{1,\text{gr}}(c)} - \pi_{1,\text{gr}}(+)\right) \quad \textbf{(eIV.2.20)}$$

$$\widehat{V}_{\text{gr}}^{\text{gr,Zae}} = \left(-\frac{1}{2}\right) \sum_{c_1,c_2 \in \text{gr} \cap \text{Zae}} \frac{\widehat{\widehat{\Delta}}_{\text{gr}}(c_1,c_2)}{\widehat{\pi_{2,\text{Zae}}}(z_1,z_2)} \left(\frac{y_C}{\pi_{1,\text{gr}}}(c_1) - \frac{y_C}{\pi_{1,\text{gr}}}(c_2)\right)^2 + \sum_{c \in \text{gr} \cap \text{Zae}} \left(\frac{y_c}{\pi_{1,\text{gr}}}\right)^2 \frac{1}{\pi_{1,\text{Zae}}(z)} \left(\frac{\widehat{\pi_{2,\text{gr}}}(c,+)}{\pi_{1,\text{gr}}(c)} - \pi_{1,\text{gr}}(+)\right) = Q_{\text{gr}}\left[\frac{y_C}{\pi_{1,\text{gr}}}(\text{gr} \cap \text{Zae})\right] \quad \textbf{(eIV.2.21)}$$

où $\pi_{1,\text{gr}} = 0.2$. $\widehat{\widehat{\Delta}}_{\text{gr}}(c_1,c_2) = \dfrac{\widehat{\pi_{2,\text{gr}}}(c_1,c_2) - \pi_{1,\text{gr}}(c_1)\,\pi_{1,\text{gr}}(c_2)}{\pi_{2,\text{gr}}(c_1,c_2)}$ et $\widehat{\pi_{2,\text{Zae}}}(z_1,z_2)$ ont été estimés par réplication, selon la méthode de Breidt-Chauvet et $c_1 \in z_1$, $c_2 \in z_2$, $c \in z$ [34].

– L'Ile-de-France comporte deux strates ZAE. Pour des ZAE de strates différentes, le calcul est réalisé avec la probabilité d'inclusion double donnée par $\pi_{2,\text{Zae}}(z_1,z_2) = \pi_{1,\text{Zae}}(z_1)\,\pi_{1,\text{Zae}}(z_2)$.

– Pour la Corse, la composante de la variance due au tirage du GR-PC, V_{gr}, est annulée [35]. En effet, le coefficient $\sum_{c_2 \in \text{gr} \cap \text{Zae}} \dfrac{\widehat{\widehat{\Delta}}_{\text{gr}}(c,c_2)}{\widehat{\pi_{2,\text{Zae}}}(z,z_2)}$ dépasse 100 en valeur absolue pour 8 des 13 petites communes corses de l'intersection entre le deuxième groupe de rotation et l'échantillon des ZAE, et va jusqu'à 1 352, contre au plus 37 pour les autres régions. Des tests ont montré que ceci pouvait produire au niveau national des contributions négatives de cette composante de la variance (voir la section 'résultats').

IV.2.3 estimateur de la variance de l'échantillon des ZAE, par strate ZAE

La variance du tirage des ZAE est estimée par (**eIV.2.22**), où l'estimation sur le groupe de rotation ne concerne que les ZAE-PC.

$$\widehat{V_{\text{Zae}}}^{\text{gr,Zae}} = \left(-\frac{1}{2}\right) \sum_{z_1,z_2 \in s_{\text{ZAE}}^{ne}} \widehat{\widehat{\Delta}}_{\text{Zae}}(z_1,z_2) \left(\frac{\widehat{y_{\text{Zae}}}^{\text{gr}}}{\pi_{1,\text{Zae}}}(z_1) - \frac{\widehat{y_{\text{Zae}}}^{\text{gr}}}{\pi_{1,\text{Zae}}}(z_2)\right)^2 = Q_{\text{Zae}}\left[\frac{\widehat{y_{\text{Zae}}}^{\text{gr}}}{\pi_{1,\text{Zae}}}(s_{\text{ZAE}}^{ne})\right] \quad \textbf{(eIV.2.22)}$$

(En effet, pour une ZAE-PC, $\widehat{y_{\text{zae}}}^{\text{gr}} = \sum_{c \in \text{zae}} \dfrac{y_c \mathbb{1}_{\text{gr}}}{\pi_{1,\text{gr}}}(c)$ peut être considéré comme le total sur la ZAE d'une variable bien définie (non aléatoire) conditionnellement au découpage en groupes de rotation RP-PC

35. Pour les ZAE exhaustives, $\widehat{\pi_{2,\text{Zae}}}$ est annulé.

36. Un autre traitement possible de cette région est d'utiliser un estimateur 'plug-in', où la probabilité d'inclusion double des ZAE n'intervient pas, en remplaçant y_c dans l'estimateur (**eIV.2.20**) par $\dfrac{y_c}{\pi_{1,\text{Zae}}}$, pour $c \in \text{gr} \cap \text{Zae}$.

et à s_{ZAE}, et $\widehat{\widehat{\Delta}}_{\text{Zae}}$ donne un estimateur de la variance de l'échantillon des ZAE conditionnellement à ces groupes. Donc (**eIV.2.22**) estime bien $\text{Var}\left(\widehat{Y}^{\text{gr,Zae}}\middle|\text{Gr}\right) = \text{Var}\left(\sum \frac{\widehat{y_{\text{Zae}}}^{\text{gr}}}{\pi_{1,\text{Zae}}}\middle|\text{Gr}\right)$ et ainsi V_{Zae}, qui est son espérance.)

– Pour la Corse, la formule de Deville est appliquée pour estimer la variance des ZAE. Ce traitement réduit la dispersion de l'estimation de variance de l'échantillon des ZAE, selon l'étude précédente dans ce document. La taille de l'échantillon corse est limitée à 3 ZAE sur 19. Il s'avère que la probabilité d'inclusion double des ZAE atteint des valeurs très faibles par rapport aux autres régions.

IV.2.4 estimation des variances GR-PC et ZAE sur l'échantillon enquêté

▷ A ce stade, un estimateur de la variance due aux échantillonnages Gr-PC et Zae est construit, par les formules (**eIV.2.21**) et (**eIV.2.22**) :

$$\widehat{V_{\text{gr,Zae}}}^{\text{gr,Zae}} = \widehat{V}_{\text{gr}}^{\text{gr,Zae}} + \widehat{V_{\text{Zae}}}^{\text{gr,Zae}}$$

- Il s'agit ensuite d'estimer cette composante $V_{\text{gr,Zae}}$ sur l'échantillon de logements s.

▷ Pour la variance du groupe de rotation des petites communes, l'absence de stratification des enquêtes Octopusse par la commune mène à choisir un estimateur par 'plug-in' simple, pour lequel y est remplacé par $\frac{ys}{\pi_{s|\text{Gr,Zae}}}$, faute de mieux [36]. Par exemple, le total communal y_c est remplacé par son estimateur $\widehat{y_c}^{s} = \sum_{c\cap s} \frac{y}{\pi_{s|\text{Gr,Zae}}}$.

▷ Le coefficient diagonal de la forme quadratique $Q_{\text{Zae}}\left(\frac{y_{\text{Zae}}}{\pi_{1,\text{Zae}}}\right) = \widehat{V_{\text{Zae}}}^{\text{Zae}}$ vaut (hors Corse) :

$$q_{\text{Zae}}(z) = -\sum_{z_2 \in s^{ne}_{\text{ZAE}} \neq z} \widehat{\widehat{\Delta}}_{\text{Zae}}(z, z_2) \quad [37]$$

→ Un estimateur approximatif de la variance due au groupe de rotation et à l'échantillon des ZAE est donné par l'équation (**eIV.2.23**).

$$\widehat{V_{\text{gr,Zae}}}^{s} = \widehat{V_{\text{gr,Zae}}}^{\text{gr,Zae}}\left(\frac{y_L s}{\pi_{s|\text{Gr,Zae}}}\right) - \sum_{z\in s^{ne}_{\text{ZAE}}} q_{\text{Zae}}(z)\,\widehat{\text{Var}}^{s}\left[\frac{\widehat{y_{\text{Zae}}}^{s}}{\pi_{1,\text{Zae}}}(z)\middle|\text{Gr,Zae}\right] \qquad \textbf{(eIV.2.23)}$$

⟹ Finalement, en intégrant la composante de variance propre à l'échantillon de l'enquête $V_{s|\text{Gr,Zae}}$, la variance de niveau logement s'estime sur son échantillon par la formule (**eIV.2.24**).

$$\begin{aligned}\widehat{V}_s^{s} &= \widehat{\text{Var}}^{s}\left(\widehat{Y}^{s}\right)\\ &= \widehat{V_{\text{gr,Zae}}}^{\text{gr,Zae}}\left(\frac{y_L s}{\pi_{s|\text{Gr,Zae}}}\right) + \sum_{z\in\text{Zae}}(1-q_{\text{Zae}}(z))\widehat{\text{Var}}^{s}\left[\frac{\widehat{y_{\text{Zae}}}^{s}}{\pi_{1,\text{Zae}}}(z)\middle|\text{Gr,Zae}\right] \qquad \textbf{(eIV.2.24)}\\ &= Q_L\left[w y_L(s)\right]\end{aligned}$$

---

37. A noter que $E\left[Q_C\left(\widehat{y_C}^{s}\right)|\text{Gr,Zae}\right] = Q_C(y_C) + \sum_{\text{zae}\in\text{Zae-Pc}} \text{trace}\left\{Q_C^{\text{zae}}\text{Var}\left[\widehat{y_C}^{s}(\text{gr}\cap\text{zae})|\text{Gr,Zae}\right]\right\}$, vu que les covariances conditionnelles inter-ZAE sont nulles. Par ailleurs le tri par identifiant RP et le tirage systématique par Octopusse assure une certaine représentativité communale.

38. Pour éviter les problèmes numériques, la somme est explicitement restreinte aux ZAE non exhaustives.

Notes :

– Pour une petite commune c, $\frac{\widehat{y_c}^s}{\pi_{1,\text{gr}}} = \frac{\sum_{l \in s \cap c} \frac{y_l}{\pi_{s|\text{Gr,Zae}}}}{\pi_{1,\text{gr}}} = \sum_{l \in s \cap c} w y_l \pi_{1,\text{Zae}}$ est bien une fonction de wy_l.

– Pour une ZAE, $\frac{\widehat{y_{\text{zae}}}^s}{\pi_{1,\text{Zae}}} = \frac{\sum_{l \in s \cap \text{zae}} \frac{y_l}{\pi_{s|\text{Gr,Zae}}} / \pi_{1,\text{gr}}}{\pi_{1,\text{Zae}}} = \sum_{l \in s \cap \text{zae}} w y_l$

• Plus explicitement, la forme quadratique (**eIV.2.25**) estime la variance de l'échantillon des logements avant non-réponse (hors Corse).

$$
\begin{aligned}
Q_L(wy_L) = & \left(-\frac{1}{2}\right) \sum_{c_1, c_2 \in \text{gr} \cap \text{Zae}} \frac{\widehat{\widehat{\Delta}}_{\text{gr}}(c_1, c_2)}{\widehat{\pi_{2,\text{Zae}}}(z_1, z_2)} \left(\frac{\widehat{y_C}^s}{\pi_{1,\text{gr}}}(c_1) - \frac{\widehat{y_C}^s}{\pi_{1,\text{gr}}}(c_2) \right)^2 \\
& + \sum_{c \in \text{gr} \cap \text{Zae}} \left(\frac{\widehat{y_C}^s}{\pi_{1,\text{gr}}}(c) \right)^2 \frac{1}{\pi_{1,\text{Zae}}(z)} \left(\frac{\widehat{\pi_{2,\text{gr}}}(c, +)}{\pi_{1,\text{gr}}(c)} - \pi_{1,\text{gr}}(+) \right) \\
& + \left(-\frac{1}{2}\right) \sum_{z_1, z_2 \in s^{ne}_{\text{ZAE}}} \widehat{\widehat{\Delta}}_{\text{Zae}}(z_1, z_2) \left(\frac{\widehat{y_{\text{Zae}}}^s}{\pi_{1,\text{Zae}}}(z_1) - \frac{\widehat{y_{\text{Zae}}}^s}{\pi_{1,\text{Zae}}}(z_2) \right)^2 \qquad \textbf{(eIV.2.25)} \\
& + \sum_{z \in \text{Zae}} (1 - q_{\text{Zae}}(z)) \frac{|s \cap \text{zae}|}{|s \cap \text{zae}| - 1} \sum_{s \cap \text{zae}} (1 - \pi_{s|\text{Gr,Zae}}) \left(wy_L - \frac{\sum_{s \cap \text{zae}} (1 - \pi_{s|\text{Gr,Zae}}) wy_L}{\sum_{s \cap \text{zae}} (1 - \pi_{s|\text{Gr,Zae}})} \right)^2
\end{aligned}
$$

– Le coefficient diagonal de la forme quadratique Q_L vaut (pour un logement l d'une petite commune, hors Corse) [38] :

$$
\begin{aligned}
q_l = & - \sum_{c_2 \in \text{gr} \cap \text{zae} \neq c} \frac{\widehat{\widehat{\Delta}}_{\text{gr}}(c, c_2)}{\widehat{\pi_{2,\text{Zae}}}(z, z_2)} \pi_{1,\text{Zae}}(z)^2 + \pi_{1,\text{Zae}}(z) \left(\frac{\widehat{\pi_{2,\text{gr}}}(c, +)}{\pi_{1,\text{gr}}(c)} - \pi_{1,\text{gr}}(+) \right) + q_{\text{Zae}}(z) \\
& + (1 - q_{\text{Zae}}(z)) \frac{|s \cap \text{zae}|}{|s \cap \text{zae}| - 1} (1 - \pi_{s|\text{Gr,Zae}}) \left[1 - (1 - \pi_{s|\text{Gr,Zae}}) / \left(\sum_{s \cap \text{zae}} (1 - \pi_{s|\text{Gr,Zae}}) \right) \right](l) \qquad \textbf{(eIV.2.26)}
\end{aligned}
$$

Notes :

– Le facteur $\pi_{1,\text{Zae}}(z)^2$ dans le premier terme et $\pi_{1,\text{Zae}}(z)$ dans le second proviennent de ce que $\frac{\widehat{y_c}^s}{\pi_{1,\text{gr}}} = \sum_{s \cap c} \frac{y_L}{\pi_{1,\text{gr}} \pi_{s|\text{Gr,Zae}}} = \sum_{s \cap c} w y_L \pi_{1,\text{Zae}}$.

– Pour les trois premiers termes de (**eIV.2.25**), les stratifications sont implicites. Par ailleurs, les petites communes et les ZAE sans échantillon sont à prendre en compte dans le calcul de variance avec des valeurs nulles (pour le résidu du calage).

39. Pour le dernier terme, le ième coefficient diagonal d'un écart quadratique $m\left[(y - m(y)/m(1))^2\right]$ vaut $m\left[(\delta_i - m(\delta_i)/m(1))^2\right] = m(1)\left(\frac{m(\delta_i^2)}{m(1)} - \left(\frac{m(\delta_i)}{m(1)}\right)^2\right) = m(\delta_i)(1 - m(\delta_i)/m(1))$ (vu que $\delta_i^2 = \delta_i$), par le théorème de König-Huygens (qui stipule que $\text{Var}(x) = E(x^2) - E(x)^2$).

– Les valeurs mesurées pour le coefficient q_l sont toutes positives lorsque la composante Corse-GR-PC est annulée (Tableau 37). Les valeurs négatives sont observées uniquement en Corse, de même que les valeurs supérieures à 2.

Tableau 37 – Eléments sur la distribution du coefficient q_l

	min	p1	median	mean	p99	max
avec annulation	0,7	0,9	1,0	1,0	1,5	2,0
sans annulation	-55,6	0,9	1,0	1,0	1,5	10,5

Note : La Corse est traitée ici spécifiquement pour la variance de l'échantillon des ZAE.

IV.3 prise en compte de la non-réponse

– L'exploitation de l'enquête traite la non-réponse par un modèle logit de niveau ménage, sans stratification [39].

– L'option retenue ici est de tenir compte de la non-réponse dans la variance par 'plug-in'. C'est-à-dire que dans l'expression de l'estimateur de la variance d'un total corrigé exactement de la non-réponse, la probabilité de réponse (conditionnelle) π_r est remplacée par $\widehat{\pi}_r$.

– Sous un modèle de réponse poissonnien, la variance du total corrigé exactement de la non-réponse (noté $\widehat{Y}^{r^*}$) se décompose selon (**eIV.3.27**), en conditionnant par l'échantillon.

$$\mathrm{Var}\left(\widehat{Y}^{r^*}\right) = \mathrm{Var}\left(\widehat{Y}^{s}\right) + E\left[\mathrm{Var}\left(\widehat{Y}^{r^*}\middle|s\right)\right] = \mathrm{Var}\left(\widehat{Y}^{s}\right) + E\left[\sum_{s_B}\left(\frac{wy_B}{\pi_r}\right)^2 \pi_r(1-\pi_r)\right] \quad \textbf{(eIV.3.27)}$$

• La variance du total corrigé de la non-réponse peut être estimée en fonction de l'échantillon s par :

$$\widehat{\mathrm{Var}\left(\widehat{Y}^{r^*}\right)}^{s} = \widehat{V}_r^{\,s}(y) = \widehat{V}_s^{\,s}(y) + \sum_{s_B}\left(\frac{wy_B}{\pi_r}\right)^2 \pi_r(1-\pi_r)$$

→ Cette statistique est estimée sur l'échantillon des ménages répondants sans biais conditionnel par :

$$\widehat{V}_r^{\,r^*}(y) = Q_L\left(w\widehat{y}_L^{\,r*}\right) - \sum_{s_B} q_L\left(w\frac{y_B r}{\pi_r}\right)^2 (1-\pi_r) + \sum_{s_B}\left(\frac{wy_B r}{\pi_r}\right)^2 (1-\pi_r)$$

(notation : pour $b \in s_B$ et $l \ni b$, $q_L(b) = q_l$; par ailleurs $\widehat{y}_L^{\,r*} = \dfrac{y_L r}{\pi_r}$)

→ La dernière étape approxime l'estimateur précédent en remplaçant la probabilité de réponse par son estimation. L'estimateur obtenu est calculable sur l'échantillon des répondants (**eIV.3.28**).

$$\begin{aligned}\widehat{V}_r^{\,r}(y) &= Q_L\left[w\widehat{y}_L(r)\right] - \sum_{s_B} q_L\left(w\frac{y_B r}{\widehat{\pi}_r}\right)^2 (1-\widehat{\pi}_r) + \sum_{s_B}\left(\frac{wy_B r}{\widehat{\pi}_r}\right)^2 (1-\widehat{\pi}_r) \\ &= Q_L\left[\sum_{r_L} w_r y_B r\right] + \sum_{s_B}(1-q_L)\left(w_r y_B r\right)^2 (1-\widehat{\pi}_r) \qquad \textbf{(eIV.3.28)} \\ &= Q_r\left[w_r y_B(r)\right]\end{aligned}$$

avec $w_r = \dfrac{w}{\widehat{\pi}_r}$ et $r_L : l \longmapsto r_l$ l'échantillon des (budgets) répondants du logement

40. En revanche, la région fait partie des variables explicatives.

– Le coefficient diagonal de la forme quadratique Q_r qui estime la variance de l'échantillon des ménages répondants est obtenu en remplaçant $w_r y_b r$ par δ_b (l'indicatrice de b). Il vaut, en $b \in r$ et $b \in l$:

$$q_r(b) = q_l - q_l(1-\widehat{\pi}_r) + 1 - \widehat{\pi}_r = q_l\widehat{\pi}_r + 1 - \widehat{\pi}_r$$

IV.4 estimation de la variance de l'échantillon des individus

– Les tirages d'individus sont indépendants conditionnellement à l'échantillon des ménages.

– Le tirage d'un individu par ménage-budget séparé (répondant) peut être considéré comme à probabilités égales sur les individus éligibles.

• Le conditionnement par l'échantillon de ménages répondants décompose la variance individuelle ainsi, où $\widehat{y_b}^{s_I} = \sum_{i \in s_I \cap b} \frac{y_i}{\pi_{i|b}}$:

$$\mathrm{Var}\left(\widehat{Y}^{s_I}\right) = \mathrm{Var}\left(\widehat{Y}^{r}\right) + E\left[\mathrm{Var}\left(\widehat{Y}^{s_I}\middle| r\right)\right] = \mathrm{Var}\left(\widehat{Y}^{r}\right) + E\left\{\sum_{b\in r}\mathrm{Var}\left(w_r\widehat{y_b}^{s_I}|r\right)\right\}$$

▷ Un estimateur de $\mathrm{Var}\left(\widehat{Y}^{s_I}\right)$ est donné en fonction de r par :

$$\widehat{V_{s_I}}^{r} = \widehat{V}_r^{\,r}(y_B) + \sum_{b\in r}\mathrm{Var}\left(w_r\widehat{y_b}^{s_I}|r\right)$$

▷ Cet estimateur est lui-même estimé sans biais conditionnel sur l'échantillon d'individus par (**eIV.4.29**).

$$\widehat{V_{s_I}}^{s_I} = \widehat{V}_r^{\,r}\left(\widehat{y_B}^{s_I}\right) + \sum_{b\in r}\widehat{(1-q_r)\mathrm{Var}\left(w_r\widehat{y_b}^{s_I}|r\right)}^{s_I} \qquad \textbf{(eIV.4.29)}$$

– Comme $|s_I \cap b| = 1$, sous l'équiprobabilité :

$$\mathrm{Var}\left(w_r\widehat{y_b}^{s_I}|r\right) = \mathrm{Var}\left(w_r\frac{y_{s_I\cap b}}{\pi_{s_I\cap b|b}}\middle| r\right) = \mathrm{Var}\left(w_I y_I|r\right)$$
$$= \mathrm{Var}_b\left(w_I y_I\right)$$
(où Var_A désigne la variance sur l'ensemble A)

$w_i = \frac{w_r}{\pi_{i|b}}$ est le poids en entrée du calage [40]. (Il peut dépendre de l'échantillon r.)

• $\widetilde{b}$ désigne génériquement les unités de collapse formées de deux ou trois ménages répondants. Pour un ensemble A de cardinal $|A|$, on note $\mathcal{S}^2_A(y) = \frac{1}{|A|-1}\sum_A\left(y - \frac{\sum_A y}{|A|}\right)^2$ [41]. L'estimateur (**eIV.4.30**) majore le deuxième terme de l'estimateur (**eIV.4.29**).

$$\widehat{V_{s_I|s}}^{s_I} = \sum_{\widetilde{b}}\left|\widetilde{b}\right|\mathcal{S}^2_{\widetilde{b}}\left\{\sqrt{(1-q_r)^+}w_I y_I\right\} \qquad \textbf{(eIV.4.30)}$$

(où $x^+ = \max(x, 0)$)

41. Ce poids intègre le facteur du calage des ZAE.

42. Le coefficient diagonal de $\mathcal{S}^2_A$ vaut $\frac{1}{|A|}(= \frac{1}{|A|-1}1(1-\frac{1}{|A|})$ en appliquant la note de bas de page nº *39*).

→ Pour récapituler, la variance de l'échantillon des individus est estimée sur celui-ci par (**eIV.4.31**).

$$\widehat{\mathrm{Var}\left(\widehat{Y}^{s_I}\right)} = Q_I\left[w_I y_I(s_I)\right] = Q_r\left(w_r \widehat{y_B}^{s_I}\right) + \sum_{\widetilde{b}} \left|\widetilde{b}\right| \mathcal{S}^2_{\widetilde{b}}\left\{\sqrt{(1-q_r)^+} w_I y_I\right\}$$

$$= \left(-\frac{1}{2}\right) \sum_{c_1,c_2 \in \mathrm{gr}\cap\mathrm{Zae}} \frac{\widehat{\widehat{\Delta}}_{\mathrm{gr}}(c_1,c_2)}{\widehat{\pi_{2,\mathrm{Zae}}}(z_1,z_2)} \left(\frac{\widehat{y_C}^{s_I}}{\pi_{1,\mathrm{gr}}}(c_1) - \frac{\widehat{y_C}^{s_I}}{\pi_{1,\mathrm{gr}}}(c_2)\right)^2 \tag{1}$$

$$+ \sum_{c\in \mathrm{gr}\cap\mathrm{Zae}} \left(\frac{\widehat{y_C}^{s_I}}{\pi_{1,\mathrm{gr}}}(c)\right)^2 \frac{1}{\pi_{1,\mathrm{Zae}}(z)} \left(\frac{\widehat{\pi_{2,\mathrm{gr}}}(c,+)}{\pi_{1,\mathrm{gr}}(c)} - \pi_{1,\mathrm{gr}}(+)\right) \tag{2}$$

$$+ \left(-\frac{1}{2}\right) \sum_{z_1,z_2 \in s^{ne}_{\mathrm{ZAE}}} \widehat{\widehat{\Delta}}_{\mathrm{Zae}}(z_1,z_2) \left(\frac{\widehat{y_{\mathrm{Zae}}}^{s_I}}{\pi_{1,\mathrm{Zae}}}(z_1) - \frac{\widehat{y_{\mathrm{Zae}}}^{s_I}}{\pi_{1,\mathrm{Zae}}}(z_2)\right)^2 \tag{3}$$

$$+ \sum_{\mathrm{zae}\in\mathrm{Zae}} (1-q_{\mathrm{Zae}}) \frac{|s\cap \mathrm{zae}|}{|s\cap \mathrm{zae}|-1} \sum_{s\cap\mathrm{zae}} (1-\pi_{s|\mathrm{Gr,Zae}}) \left(w_r \widehat{y_L}^{s_I} - \frac{\sum_{s\cap\mathrm{zae}} (1-\pi_{s|\mathrm{Gr,Zae}}) w_r \widehat{y_L}^{s_I}}{\sum_{s\cap\mathrm{zae}} (1-\pi_{s|\mathrm{Gr,Zae}})}\right)^2 \tag{4}$$

$$- \sum_{s_B} q_L \left(w_r \widehat{y_B}^{s_I} r\right)^2 (1-\widehat{\pi_r}) \tag{5}$$

$$+ \sum_{s_B} \left(w_r \widehat{y_B}^{s_I} r\right)^2 (1-\widehat{\pi_r}) \tag{6}$$

$$+ \sum_{\widetilde{b}} \left|\widetilde{b}\right| \mathcal{S}^2_{\widetilde{b}} \left\{\sqrt{(1-q_r)^+} w_I y_I\right\} \qquad \textbf{(eIV.4.31)} \tag{7}$$

Notes :

– Les trois derniers termes sont de niveau ménage-budget séparé, alors que le précédent est de niveau logement.

– Pour le calcul de variance, les logements et les ménages sans individus dans le champ de l'enquête peuvent être considérés comme répondants avec des variables nulles. Ceci intervient dans les 4 derniers termes de (**eIV.4.31**).

– La quatrième composante (4), qui correspond à la variance du tirage des logements, est parfois négative, comme il existe des coefficients $q_{\mathrm{Zae}} > 1$. Mais ceci est exceptionnel : un seul cas parmi la cinquantaine de variables étudiées.

– w_r=`poids_corr` est le poids du ménage après correction de la non-réponse. Il exclut ici le facteur de calage des ZAE.

– Pour tout ensemble A, $\widehat{y_A}^{s_I}$ est le total de y sur A estimé sur l'échantillon s_I. Dans le contexte de l'expression ci-dessus, pour une unité a tirée dans une phase antérieure à celle des individus, $\widehat{y_a}^{s_I}$ est un estimateur du total de y sur a conditionnellement à l'échantillon du niveau de a. En pratique, le calcul est effectué ainsi, avec w_i le poids de l'individu i avant calage et hors facteur de calage des ZAE :

▷ $\dfrac{\widehat{y_{\mathrm{zae}}}^{s_I}}{\pi_{1,\mathrm{zae}}} = \sum_{i\in s_I\cap\mathrm{zae}} w_i y$

▷ $\dfrac{\widehat{y_c}^{s_I}}{\pi_{1,\mathrm{gr}}} = \sum_{i\in s_I\cap c} w_i y \pi_{1,\mathrm{zae}}$

▷ $w_r \widehat{y_l}^{s_I} = \sum_{i\in l} w_i y$ (idem au niveau budget)

IV.5 prise en compte du calage AES

Le calage du total estimé est pris en compte dans le calcul de sa variance en remplaçant la variable par le résidu de sa régression par le vecteur de calage, sur l'échantillon des individus répondants pondéré par le poids final w_{rc} (**eIV.5.32**).

$$\widehat{\mathrm{Var}\left(\widehat{Y}^{rc}\right)} = Q_I\left[w_I\left(y_I - x_{\mathrm{cal}}^{I\prime}\widehat{\beta}\right)\right] \qquad \textbf{(eIV.5.32)}$$

avec $\widehat{\beta} = \widehat{m}^{rc}\left(x_{\mathrm{cal}}x_{\mathrm{cal}}{}'\right)^{-}\widehat{m}^{rc}\left(yx_{\mathrm{cal}}\right)$

– Le fait que le calage est réalisé par rapport à des totaux aléatoires, estimés par une moyenne annuelle de l'enquête Emploi, n'est pas pris en compte dans le calcul de variance. Cette omission sous-estime la variance.

IV.6 programmes et mise en oeuvre

– Les programmes sont ci-joints :

variance_aes.sas

- **variance_aes.sas** (macro principale)

variance_aes_test_2.sas

- **variance_aes_test_2.sas** (exemple de lancement)

– La variance d'un ratio peut être calculée directement, par linéarisation, en déclarant ce ratio sous la forme :

`ratio(numérateur,dénominateur,variable)`

où `variable` est le nom à attribuer au ratio résultant. Aucun blanc ne doit figurer entre les parenthèses.

– La table des données d'enquête n'est pas nécessairement localisée dans une base SAS permanente. Ceci permet de définir des variables provisoires sur lesquelles le calcul de variance est ensuite demandé. Il est nécessaire de définir une variable donnant la probabilité conditionnelle de tirage d'un individu `pi_i_cond_l`.

– Le nom du principal classeur excel créé par le programme est paramétré par `nom_sortie`. Un second classeur est suffixé par `_check`. Il fournit la décomposition de la variance estimée selon (**eIV.4.31**).

– Un calcul de variance standard est lancé par une instruction du type :

```
%variance_aes(liste_variables=%bquote(nfenum_sup_0
ratio(nfenum_sup_0,i1,part_nfenum_sup_0)
nfenum_s_1 ratio(nfenum_s_1,sexe_1,part_nfenum_s_1)
nfenum_s_2 ratio(nfenum_s_2,sexe_2,part_nfenum_s_2)
nfenum_age_25_34 ratio(nfenum_age_25_34,age_25_34,part_nfenum_age_25_34)
nfenum_age_35_49 ratio(nfenum_age_35_49,age_35_49,part_nfenum_age_35_49)
nfenum_age_50_64 ratio(nfenum_age_50_64,age_50_64,part_nfenum_age_50_64)
)
/* pas de blancs dans la définition des ratios */
, data_enquete=donnees_enquete
, poids_final=poidsfi
```

```
, poids_entree_calmar=poids_indiv
, poids_menage_corrige=poids_corr
, pi_i_cond_l=pi_i_cond_l
, base_echantillon=&racine\basesto
, ear=2010
, base_ZAE=&racine_application\octopusse\zae\basesto
, base_PC=&racine_application\octopusse\petites_communes\basesto

, liste_variables_calage=ZUS agq20s1 agq20s2 agq25s1 agq25s2 agq30s1 agq30s2
agq35s1 agq35s2 agq40s1 agq40s2 agq45s1 agq45s2
agq50s1 agq50s2 agq55s1 agq55s2 agq60s1 agq60s2
csp1_1 csp1_2 csp1_3 csp1_4 csp1_5 csp1_6
dip12 dip21 dip22 dip30 dip31 dip32 dip33 dip41 dip42 dip43 dip44
dip50 dip60 dip70 dip71
natio1 natio2
reg21 reg22 reg23 reg24 reg25 reg26 reg31 reg41 reg42 reg43 reg52
reg53 reg54 reg72 reg73 reg74 reg82 reg83 reg91 reg93 reg94

, liste_variables_non_reponse=CODE_REG_ADMIN  NAT_IMMIGRE OCC_STATUT_OCC
RP_COUPLE decile1_9_10 nbpi pos_prof tabard type_hab
, nom_sortie=variance_aes_test_c);
```

– Le paramètre `base_ZAE` donne le nom physique du répertoire contenant d'une part la table de référence des ZAE et l'échantillon maître, d'autre part l'estimation des probabilités d'inclusion doubles.

– Le paramètre `base_PC` donne la localisation de la table des probabilités d'inclusion double des petites communes, ainsi qu'une table de référence des petites communes.

– La base contenant le fichier échantillon `capi_echantillon_principal` est localisée dans `base_echantillon`.

IV.7 résultats

– Le calcul de variance d'AES 2012 a été effectué sur les 20 132 ménages (et logements hors champ). Les pondérations utilisées sont décrites dans la note N° 109/DG75-F230/ du 15 janvier 2013.

– Les résultats sont produits ici pour la variable nombre de formations non formelles `nfenum>0` croisée avec le sexe et avec trois tranches d'âge.

• Le Tableau 38 donne les estimations de variance (calée) sans annulation de la contribution de la Corse à l'estimation de la variance du groupe de rotation des petites communes.

Tableau 38 – Estimations de variance sans annuler la variance PC de la Corse

statistique totaux	nfenum sup 0 18 967 828	nfenum s 1 9 365 368	nfenum s 2 9 602 459	nfenum age 25 34 4 416 883	nfenum age 35 49 6 983 071	nfenum age 50 64 4 600 275
écart-type	224 605	164 213	150 089	98 735	118 760	119 149
écart-type approché	160 765	138 739	139 899	103 222	124 709	105 054
CV	1,18	1,75	1,56	2,24	1,70	2,59
Deff	2,37	2,41	1,94	1,97	1,75	1,99

Notes :

– *L'écart-type approché d'un total estimé* $\widehat{Y}\,(<\widehat{N})$ *est une approximation de la variance d'un sondage aléatoire simple (sans calage), par la formule* $\widehat{N}^2\left[\left(1-\frac{|r|}{\widehat{N}}\right)/|r|\right]\frac{\widehat{Y}}{\widehat{N}}\left(1-\frac{\widehat{Y}}{\widehat{N}}\right)$.

– *Le coefficient de variation (CV) est en pourcentage.*

– *Un Deff plus élevé pour le total général par rapport à celui par catégories peut se comprendre comme l'effet d'une taille de grappe plus grande.*

→ Il apparait des contributions négatives du tirage du groupe de rotation RP-PC à la variance estimée (Tableau 39). C'est gênant, même s'il y a une compensation possible par le terme 'correctif variance logement', via des coefficients q_l très négatifs en Corse.

Tableau 39 – Décomposition de la variance sans annulation de la composante PC de la Corse

statistique	nfenum sup 0	nfenum s 1	nfenum s 2	nfenum age 25 34	nfenum age 35 49	nfenum age 50 64
variance	50 447 523 357	26 965 968 642	22 526 734 558	9 748 595 723	14 103 869 423	14 196 493 194
variance PC	234 193 429	1 069 882 338	248 829 030	**-939 341 023**	**-658 641 304**	**-459 758 784**
variance ZAE	36 000 610 905	17 552 368 398	14 929 442 913	5 450 648 174	9 579 865 410	9 649 926 014
variance logement	7 313 266 515	3 647 653 428	4 271 730 632	1 603 280 412	2 455 699 384	2 703 317 341
correctif variance logement	-5 094 067 570	-2 354 131 329	-2 674 835 403	-447 227 016	-1 749 528 786	-1 816 011 578
variance non reponse	6 339 301 926	3 296 822 532	3 497 907 406	1 317 362 188	2 403 924 899	2 151 724 405
variance individu	5 654 218 150	3 753 373 275	2 253 659 980	2 763 872 989	2 072 549 821	1 967 295 796

Note : Cette décomposition est effectuée selon la formule (**eIV.4.31**)*. Il ne s'agit pas d'estimateurs sans biais des différentes composantes.*

• En annulant la contribution corse à la variance du GR-PC, les estimations de variance sont inférieures aux précédentes (Tableau 40). Les Deff restent supérieurs à 1, ce qui est rassurant.

Tableau 40 – Estimations de variance avec annulation de la composante Corse de la variance du GR-PC

statistique totaux	nfenum sup 0 18 967 828	nfenum s 1 9 365 368	nfenum s 2 9 602 459	nfenum age 25 34 4 416 883	nfenum age 35 49 6 983 071	nfenum age 50 64 4 600 275
écart-type	215 722	150 861	143 090	85 287	112 495	114 972
écart-type approché	160 765	138 739	139 899	103 222	124 709	105 054
CV	1,14	1,61	1,49	1,93	1,61	2,50
Deff	2,19	2,03	1,77	1,47	1,57	1,85

– Il n'apparait plus de contribution négative de la variance du groupe de rotation des petites communes, au niveau national [42] (Tableau 41). La variance estimée de la variable `nfenum_age_25_34` baisse, malgré une contribution accrue de la composante des petites communes. Ceci s'explique par un terme correctif moins négatif. En effet, la contribution de la Corse à cette composante est simultanément supprimée.

Tableau 41 – Décomposition de la variance estimée en annulant la contribution corse pour les petites communes

statistique	nfenum sup 0	nfenum s 1	nfenum s 2	nfenum age 25 34	nfenum age 35 49	nfenum age 50 64
variance	46 535 955 133	22 759 162 002	20 474 711 097	7 273 927 473	12 655 132 991	13 218 648 216
variance PC	3 187 130 940	1 501 493 477	1 259 428 635	210 425 955	621 113 777	834 418 532
variance ZAE	36 000 610 905	17 552 368 398	14 929 442 913	5 450 648 174	9 579 865 410	9 649 926 014
variance logement	7 313 266 515	3 647 653 428	4 271 730 632	1 603 280 412	2 455 699 384	2 703 317 341
correctif variance logement	-6 644 373 609	-3 438 234 617	-3 681 885 103	-1 371 895 113	-2 525 993 226	-2 265 825 479
variance non reponse	6 339 301 926	3 296 822 532	3 497 907 406	1 317 362 188	2 403 924 899	2 151 724 405
variance individu	340 018 455	199 058 785	198 086 613	64 105 857	120 522 747	145 087 403

– La composante 'individu' de la variance est la plus faible. Toutefois, sa mesure intègre le correctif de l'estimation plug-in de la variance des ménages répondants. Quoi qu'il en soit, l'enjeu du collapse effectué à ce niveau est sans doute mineur.

– L'élimination de la contribution des petites communes au coefficient q_l (via les deux premiers termes de (**eIV.2.26**)), dans le terme correctif de la variance logement, par le paramètre `pc_annule_q_l=oui`, augmente légèrement la variance estimée (Tableau 42). Ceci signifie qu'il joue bien à la baisse de la variance estimée. Il pourrait être prudent d'omettre ce terme, comme il suppose une stratification par la commune, et donc ne mesure pas précisément le correctif de la variance des petites communes. Mais l'effet marginal n'incite pas à retenir cette option.

Tableau 42 – Estimation de variance sans la contribution des petites communes au coefficient q_l

statistique totaux	nfenum sup 0 18 967 828	nfenum s 1 9 365 368	nfenum s 2 9 602 459	nfenum age 25 34 4 416 883	nfenum age 35 49 6 983 071	nfenum age 50 64 4 600 275
écart-type	215 893	150 852	143 242	85 328	112 686	115 009
écart-type approché	160 765	138 739	139 899	103 222	124 709	105 054
CV	1,14	1,61	1,49	1,93	1,61	2,50
Deff	2,19	2,03	1,77	1,47	1,58	1,85

– L'estimation 'plug-in' de la variance du GR-PC corse, décrite en note de bas de page n°*36*, augmente légèrement l'estimation, d'au plus 1% de l'écart-type. Ce raffinement parait superflu.

Références

[BC] Breidt FJ , Chauvet G, Improved variance estimation for balanced samples drawn via the Cube method, Journal of Statistical Planning and Inference, 2010

[GC] Chauvet G, On variance estimation for the French master sample, Journal of Official Statistics, vol.27 N° 4, 2011

[MCSF] Christine M., Faivre S., Octopusse : un système d'échantillon-maître pour le tirage des échantillons dans la dernière enquête annuelle de recensement, Insee, JMS 2009

43. En revanche, ça reste le cas pour une à trois régions, selon la variable. L'estimation de la composante 'petites communes' de la variance doit être considérée comme fragile. Elle est maintenue pour rendre compte de cette contribution.

Annexe A expression des moyennes des deux estimateurs de variance

Il s'agit ici d'estimer $E\left(\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}\right)$ [43] sur les simulations S, pour une variable connue sur l'univers.

La moyenne des estimateurs de variance d'Horvitz-Thompson et de Yates-Grundy sur l'ensemble des simulations peut s'exprimer en fonction de la matrice des probabilités d'inclusion double estimées sur cet ensemble. Les formules ci-dessous sont appliquées à une variable unidimensionnelle y. Mais elles peuvent s'étendre à une variable vectorielle. Elles sont également valables lorsque $E(|R)$ est remplacé par $\sum_S \Big/ |S|$.

- pour l'estimateur de variance d'Horvitz-Thompson :

$$E\left(\widehat{V}^{HT}\right) = E\left\{\sum_{i,j\in s}\frac{y}{\pi_1}(i)\widehat{\widehat{\Delta}}(i,j)\frac{y}{\pi_1}(j)\right\} = E\left\{\sum_{i,j}\frac{y}{\pi_1}(i)\widehat{\widehat{\Delta}}(i,j)s(i)s(j)\frac{y}{\pi_1}(j)\right\}$$
$$= \sum_{i,j}\frac{y}{\pi_1}(i)\widehat{\widehat{\Delta}}(i,j)\pi_2(i,j)\frac{y}{\pi_1}(j) \qquad \text{d'où :}$$

$$\widehat{E\left(\widehat{V}^{HT}\right)}^{S} = \left(\frac{y}{\pi_1}\right)'\left(\widehat{\widehat{\Delta}}\#\widehat{\pi}_2{}^{S}\right)\frac{y}{\pi_1} \qquad \textbf{(eA.33)}$$

- pour l'estimateur de Yates-Grundy (généralisé) : L'expression (**eI.2.13**) se reformule ainsi :

$$\widehat{V}^{YG} = \widehat{V}^{HT} - \left\{\left(\widehat{\widehat{\Delta}}\#ss'\right)\left(\frac{y}{\pi_1}\right)^2\right\}[+,] + \left\{\left(\frac{\widehat{\pi}_2{}^{R}(,+)}{\pi_1} - \pi_1(+)\right)\#s\#\left(\frac{y}{\pi_1}\right)^2\right\}[+,]$$

d'où l'estimateur sur les simulations :

$$\widehat{E\left(\widehat{V}^{YG}\right)}^{S} = \widehat{E\left(\widehat{V}^{HT}\right)}^{S} - \mathbb{1}'\left(\widehat{\widehat{\Delta}}\#\widehat{\pi}_2{}^{S}\right)\left(\frac{y}{\pi_1}\right)^2 + \left[\left(\widehat{\pi}_2{}^{R}(,+) - \pi_1(+)\,\pi_1\right)\#\left(\frac{y}{\pi_1}\right)^2\right][+,] \qquad \textbf{(eA.34)}$$

Notes :

– L'ensemble des simulations utilisées pour moyenner les estimations de variance par échantillon est identique à celui utilisé pour estimer par simulation la variance. Donc il n'y a pas d'indépendance entre les deux statistiques.

– Les termes de $\widehat{\pi}_2{}^{S} < 0$ sont pris en compte.

– Dans ces deux expressions, la correction de la diagonale de $\widehat{\pi}_2{}^{S}$ par π_1 peut sembler une bonne idée, pour réduire l'aléa sur l'estimation de π_2. Les résultats obtenus ont dissuadé de retenir cette option. En effet, l'incidence est très faible et ne joue pas clairement à la réduction du biais estimé.

– La moyenne sur les simulations du troisième terme de (**eI.2.13**) s'exprime littéralement par : $\left[\left(\frac{\widehat{\pi}_2{}^{R}(,+)}{\pi_1} - \pi_1(+)\right)\#\left(\frac{y}{\pi_1}\right)^2\#\bar{s}^{S}\right][+,]$. Cependant, comme $E\left(\bar{s}^{S}\right) = \pi_1$ est connu, il est sans doute préférable d'utiliser cette dernière valeur pour estimer $E\left(\widehat{V}\right)$.

– Les mêmes formules (**eA.33**, **eA.34**) s'appliquent pour calculer la variance estimée sur un échantillon s donné ($\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}$), en remplaçant $\widehat{\pi}_2{}^{S}$ par ss' et le dernier terme par : $\left[\left(\frac{\widehat{\pi}_2{}^{R}(,+)}{\pi_1} - \pi_1(+)\right)\#\left(\frac{y}{\pi_1}\right)^2\#s\right][+,]$.

44. Le conditionnement par les réplications est omis : $E\left(\widehat{V}\middle|R\right)$ est confondu avec $E\left(\widehat{V}\right)$.

– La précision de l'estimateur des probabilités d'inclusion par la méthode de Breidt-Chauvet semble meilleure que celle par la fréquence (Tableau 1). Donc il pourrait être préférable de calculer le biais en utilisant cette méthode pour $\widehat{\pi}_2^{BC,S}$. Cette option a été testée dans l'étude sur les petites communes, sans amélioration manifeste de la mesure du biais. En revanche, c'est le choix retenu pour les ZAE, comme il permet ensuite de réunir les réplications et les simulations pour l'estimation finale des probabilités d'inclusion.

Annexe B défauts de l'appariement avec les données sur les groupes de rotation des PC

Tableau B.43 – Défauts de l'appariement sur le code communal entre les données sur les groupes de rotation et le référentiel des petites communes utilisé

REGION	depcom	libgeo	p99 pop	type
21	10 054	Bourdenay	402	pas dans données GR
21	51 136	Châtillon-sur-Marne	843	pas dans données GR
21	52 359	Nully	301	pas dans données GR
22	80 369	Frohen-sur-Authie	174	pas dans données GR
22	80 370	Frohen-le-Petit	25	pas dans données GR
25	14 149	Cesny-aux-Vignes	498	pas dans données GR
25	50 066	Jullouville	2 414	pas dans données GR
25	50 216	Graignes-Mesnil-Angot	618	pas dans données GR
25	50 303	Mesnil-Angot	50	pas dans données GR
25	61 022	Bagnoles-de-l'Orne	895	pas dans données GR
25	61 136	Couvains	121	pas dans données GR
25	61 254	Marnefer	53	pas dans données GR
25	61 483	Bagnoles-de-l'Orne	1 279	pas dans données GR
26	71 211	Géanges	363	pas dans données GR
26	71 443	Saint-Loup-Géanges	764	pas dans données GR
42	67 242	Kirrwiller	684	pas dans données GR
43	39 367	Morbier	2 059	pas dans données GR
43	39 524	Tancua	167	pas dans données GR
52	49 206	Montfaucon-Montigné	518	pas dans données GR
52	49 210	Montigné-sur-Moine	1 204	pas dans données GR
52	85 036	Bretonnière-la-Claye	440	pas dans données GR
52	85 068	Claye	46	pas dans données GR
53	22 179	Fréhel	2 047	pas dans données GR
53	35 130	Hédé-Bazouges	1 833	pas dans données GR
54	79 013	Argenton-les-Vallées	1 038	pas dans données GR
54	79 017	Aubiers	2 877	pas dans données GR
54	79 037	Boësse	429	pas dans données GR
54	79 195	Nueil-les-Aubiers	2 115	pas dans données GR
54	79 305	Sanzay	243	pas dans données GR
72	47 010	Antagnac	333	pas dans données GR
72	64 269	Idron	5 151	pas dans données GR
73	81 107	Guitalens	337	pas dans données GR
73	81 132	Guitalens-L'Albarède	264	pas dans données GR
82	26 091	Chauvac-Laux-Montaux	37	pas dans données GR
82	26 158	Laux-Montaux	8	pas dans données GR
26	71 353	PLOTTES		pas dans pc_donnees_rp99
26	89 288	PAROY-EN-OTHE		pas dans pc_donnees_rp99
41	55 039	BEAUMONT-EN-VERDUNOIS		pas dans pc_donnees_rp99
41	55 050	BEZONVAUX		pas dans pc_donnees_rp99
41	55 068	BRABANT-EN-ARGONNE		pas dans pc_donnees_rp99
41	55 082	BROCOURT-EN-ARGONNE		pas dans pc_donnees_rp99
41	55 189	FLEURY-DEVANT-DOUAUMONT		pas dans pc_donnees_rp99
41	55 239	HAUMONT-PRES-SAMOGNEUX		pas dans pc_donnees_rp99
41	55 307	LOUVEMONT-COTE-DU-POIVRE		pas dans pc_donnees_rp99
72	47 163	MAUVEZIN-SUR-GUPIE		pas dans pc_donnees_rp99
82	74 106	DRAILLANT		pas dans pc_donnees_rp99

Annexe C défauts de l'appariement des petites communes avec les données du RP 2009

Tableau C.44 – Défauts de l'appariement sur le code communal des données du RP 2009 avec le référentiel des petites communes

REGION	depcom	libgeo	p99 pop	p09 pop	type
22	80 370	Frohen-le-Petit	25		pas dans RP 2009
25	50 303	Mesnil-Angot	50		pas dans RP 2009
25	61 022	Bagnoles-de-l'Orne	895		pas dans RP 2009
25	61 254	Marnefer	53		pas dans RP 2009
26	21 551	Saint-Germain-Source-Seine	29		pas dans RP 2009
26	71 211	Géanges	363		pas dans RP 2009
31	59 248	Fort-Mardyck	3 766		pas dans RP 2009
43	39 524	Tancua	167		pas dans RP 2009
52	49 210	Montigné-sur-Moine	1 204		pas dans RP 2009
52	85 068	Claye	46		pas dans RP 2009
54	79 017	Aubiers	2 877		pas dans RP 2009
54	79 037	Boësse	429		pas dans RP 2009
54	79 305	Sanzay	243		pas dans RP 2009
73	81 107	Guitalens	337		pas dans RP 2009
82	26 158	Laux-Montaux	8		pas dans RP 2009
21	10 038	Bercenay-le-Hayer		135	pas dans pc_donnees_rp99
21	10 383	Trancault		206	pas dans pc_donnees_rp99
21	51 201	Cuisles		138	pas dans pc_donnees_rp99
21	52 495	Trémilly		88	pas dans pc_donnees_rp99
25	14 482	Ouézy		256	pas dans pc_donnees_rp99
25	50 102	Carolles		765	pas dans pc_donnees_rp99
26	71 353	Plottes		590	pas dans pc_donnees_rp99
26	89 288	Paroy-en-Othe		207	pas dans pc_donnees_rp99
26	89 326	Rosoy		1 052	pas dans pc_donnees_rp99
31	62 847	Verquigneul		2 056	pas dans pc_donnees_rp99
41	55 039	Beaumont-en-Verdunois	0	0	pas dans pc_donnees_rp99
41	55 050	Bezonvaux	0	0	pas dans pc_donnees_rp99
41	55 068	Brabant-en-Argonne		105	pas dans pc_donnees_rp99
41	55 082	Brocourt-en-Argonne		47	pas dans pc_donnees_rp99
41	55 189	Fleury-devant-Douaumont	0	0	pas dans pc_donnees_rp99
41	55 239	Haumont-près-Samogneux	0	0	pas dans pc_donnees_rp99
41	55 307	Louvemont-Côte-du-Poivre	0	0	pas dans pc_donnees_rp99
42	67 057	Bosselshausen		178	pas dans pc_donnees_rp99
53	22 201	Plévenon		732	pas dans pc_donnees_rp99
53	35 317	Saint-Symphorien		527	pas dans pc_donnees_rp99
72	47 163	Mauvezin-sur-Gupie		477	pas dans pc_donnees_rp99
72	47 227	Ruffiac		176	pas dans pc_donnees_rp99
72	64 439	Ousse		1 472	pas dans pc_donnees_rp99
72	64 518	Sendets		846	pas dans pc_donnees_rp99
73	31 300	Lieoux		120	pas dans pc_donnees_rp99
82	74 106	Draillant		683	pas dans pc_donnees_rp99

Annexe D remarques sur la majoration de $\widehat{\Delta}^{BC} = \overline{\sum_t \underline{\lambda}_t \overline{\lambda_t} u_t u_t'}^{R}$

– pour mémoire, en notant $\pi_t \left(\in [0,1]^{\mathscr{P}} \right)$ la martingale de Cube à l'itération t :

$$\begin{aligned} \sum_t \underline{\lambda}_t \overline{\lambda_t} u_t u_t' &= \sum \mathrm{Var}\left[\pi_t - \pi_{t-1} | \mathscr{F}_{t-1}\right] \\ s &= \sum \pi_t - \pi_{t-1} + \pi_1 \\ \mathrm{Var}(s) &= \sum \mathrm{Var}\left(\pi_t - \pi_{t-1}\right) = E\left\{\sum \mathrm{Var}\left[\pi_t - \pi_{t-1} | \mathscr{F}_{t-1}\right]\right\} \\ &= E\left\{\sum \left(\pi_t - \pi_{t-1}\right)\left(\pi_t - \pi_{t-1}\right)'\right\} \end{aligned}$$

– La performance de la méthode de Breidt-Chauvet par rapport à l'estimation des probabilités d'inclusion par la fréquence reste inexpliquée rigoureusement. Même la raison pour laquelle l'estimateur se retrouve rapidement dans l'intervalle $[0,1]$ (du moins pour la diagonale) n'est pas évidente. Il y a donc un manque de maîtrise méthodologique de cet estimateur. Cette annexe esquisse quelques réflexions sur ce thème. Il ne s'agit ici que de la qualité de l'estimation des probabilités d'inclusion. La qualité de l'estimateur 'échantillon' de la variance, calculé comme une fonction non linéaire de ces probabilités estimées, est au delà du champ de ces investigations.

– L'étude de la majoration de $\widehat{\mathrm{Var}(s)}^{R}$ n'est qu'un élément secondaire dans la comparaison des estimateurs de variance, par rapport à la variance de ceux-ci. Néanmoins, en l'absence de cette dernière, la majoration pourrait être utile pour vérifier l'acceptabilité de l'estimateur.

– $\overline{\lambda_t} = \min\left\{\dfrac{\mathbb{1}_{u_t^i>0} - \pi_{t-1}^i}{u_t^i}, i \in \mathscr{P} \,/\, u_t^i \neq 0\right\}$, $\underline{\lambda}_t = \min\left\{\dfrac{\pi_{t-1}^i - \mathbb{1}_{u_t^i<0}}{u_t^i}, i \in \mathscr{P} \,/\, u_t^i \neq 0\right\}$

– Comme $\underline{\lambda}_t$ et $\overline{\lambda_t} > 0$, $\widehat{\pi_1(i)} = \sum_t \underline{\lambda}_t \overline{\lambda_t} (u_t^i)^2 + \pi_1(i)^2 \geqslant \pi_1(i)^2$. En revanche, il est possible que $\widehat{\pi_2}(i,j) < 0$ pour $i \neq j$, et c'est effectivement observé dans le présent travail.

– Empiriquement, sur $1\,000$ réplications du tirage de l'EM-ZAE réalisées sur 22 régions, le maximum de $\widehat{\Delta}$ par région et par réplication ne dépasse 1 que dans 3 cas. De plus, ce maximum n'excède pas 1.2 [44]. Ces résultats vont dans le sens d'une faible variance de l'estimateur de $\mathrm{Var}(s)$ par la méthode de Breidt et Chauvet, comme suggéré par ces auteurs.

– Du fait que $\pi_{t-1}^i + \overline{\lambda_t} u_t^i$ et $\pi_{t-1}^i - \underline{\lambda}_t u_t^i \in [0,1]$, il découle que [45] :

$$\underline{\lambda}_t \overline{\lambda_t} (u_t^i)^2 \leqslant \pi_{t-1}^i (1 - \pi_{t-1}^i)$$

Mais cette majoration n'est pas suffisante pour expliquer la borne supérieure remarquablement réduite de $\widehat{\Delta}$, puisqu'il s'avère empiriquement que $\sum_t \pi_{t-1}(1-\pi_{t-1})$ est souvent supérieur à 1, et atteint des valeurs largement plus élevées que ça.

– Sur 100 réplications, le maximum de $\sum_t \left(\pi_t - \pi_{t-1}\right)^2$ ne dépasse pas 1.8. Cette statistique peut être interprétée comme un indicateur de la rapidité de convergence de l'algorithme vers un sommet du cube.

---

45. Toutefois ce critère est encore plus favorable à l'estimateur $\overline{ss'} - \pi_1 \pi_1'$, dont les coefficients sont toujours dans $\left[-\pi_1^i \pi_1^j, 1 - -\pi_1^i \pi_1^j\right]$.

46. autre démonstration (l'indice i est omis ici) : $\mathrm{Var}\left[\pi_t - \pi_{t-1} | \mathscr{F}_{t-1}\right] = E\left\{\left(\pi_t - \pi_{t-1}\right)^2 \middle| \mathscr{F}_{t-1}\right\} = E\left\{\pi_t^2 | \mathscr{F}_{t-1}\right\} - \pi_{t-1}^2 \leqslant E\left\{\pi_t | \mathscr{F}_{t-1}\right\} - \pi_{t-1}^2 = \pi_{t-1}(1-\pi_{t-1})$ et $\mathrm{Var}\left\{\pi_t | \mathscr{F}_{t-1}\right\} = \underline{\lambda}_t \overline{\lambda_t} u_t^2$

– En notant $\overline{\pi}_t = \pi_{t-1} + \overline{\lambda_t} u_t$ et $\underline{\pi}_t = \pi_{t-1} - \underline{\lambda_t} u_t$, on peut écrire $\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2 = \sum_t \left(\overline{\pi}_t^i - \pi_{t-1}^i\right)\left(\pi_{t-1}^i - \underline{\pi}_t^i\right)$.

Comme $\pi_t^i \in \left\{\underline{\pi}_t^i, \overline{\pi}_t^i\right\}$ et que $\pi_{t-1}^i \in \left(\underline{\pi}_t^i, \overline{\pi}_t^i\right) \subset [0,1]$ [46], il s'ensuit que :

$$\begin{aligned}\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2 &= \sum_t |\Delta\pi_t|\left(|\overline{\pi}_t - \underline{\pi}_t| - |\Delta\pi_t|\right)\\ &\leqslant \sum_t |\Delta\pi_t|\left(1 - |\Delta\pi_t|\right)\end{aligned}$$

Ce majorant s'avère fort, puisque cette grandeur dépasse 6, sur 100 réplications. Mais il suggère un lien entre la faible variance de $\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2$ et la rapidité de convergence de l'algorithme du Cube.

– L'inégalité de Markov permet une majoration en probabilité :

$$\begin{aligned}proba\left\{\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2 > \eta\right\} &\leqslant E\left\{\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2\right\}/\eta = \mathrm{Var}\,(s_i)/\eta\\ &\leqslant \frac{\pi_i(1-\pi_i)}{\eta}\end{aligned}$$

Mais cette inégalité est aussi vérifiée par les autres estimateurs sans biais de $\mathrm{Var}\,(s)$, tels que $ss' - \pi_1\pi_1'$ et $\sum_t \Delta\pi_t\Delta\pi_t'$.

– La matrice $\widehat{\Delta}^{BC} = \overline{\sum_t \underline{\lambda_t}\overline{\lambda_t}u_t u_t'}^{R}$ est symétrique et positive, vu que $\underline{\lambda_t}\overline{\lambda_t} \geqslant 0$. Par conséquent $\widehat{\pi}_2 = \widehat{\Delta}^{BC} + \pi_1\pi_1'$ aussi. L'inégalité de Cauchy-Schwartz implique alors :

$$\begin{aligned}{\widehat{\pi}_2}^{BC}(i,j)^2 = \left(\mathbb{1}_i'\widehat{\pi}_2\mathbb{1}_j\right)^2 &\leqslant \mathbb{1}_i'\widehat{\pi}_2\mathbb{1}_i\mathbb{1}_j'\widehat{\pi}_2\mathbb{1}_j = \widehat{\pi}_1(i)\widehat{\pi}_1(j)\\ &\leqslant \max\left\{\widehat{\pi}_1(i)^2, \widehat{\pi}_1(j)^2\right\} \text{ (comme } {\widehat{\pi}_1}^{BC} \geqslant 0)\end{aligned}$$

$\Longrightarrow$ Il suffirait donc de majorer ${\widehat{\pi}_1}^{BC}$.

– Si $\pi_1(i) = 0$ alors $\mathbb{1}_i'\mathrm{Var}\,(s)\,\mathbb{1}_i = 0 = \sum \mathbb{1}_i'\mathrm{Var}\,(\Delta\pi_t)\,\mathbb{1}_i \Longrightarrow \forall t,\ \mathbb{1}_i'\mathrm{Var}\,(\Delta\pi_t)\,\mathbb{1}_i = 0 \Longrightarrow \forall t,\ \mathbb{1}_i'\mathrm{Var}\left[\Delta\pi_t|\mathscr{F}_{t-1}\right]\mathbb{1}_i = 0$ presque sûrement. Donc dans ce cas ${\widehat{\pi}_1}^{BC}(i) \overset{ps}{=} 0$. En revanche, si $\pi_2(i,j) = 0$, il ne semble pas assuré que ${\widehat{\pi}_2}^{BC}(i,j) \overset{ps}{=} 0$.

– $\mathbb{1}'\widehat{\Delta}^{BC}\mathbb{1} = 0 = \mathbb{1}'\mathrm{Var}\,(s)\,\mathbb{1}$. Ceci découle du fait qu'à chaque itération t de Cube, $u_t \in \mathrm{Ker}\left\{\left(\frac{x}{\pi}\right)^{\mathscr{P}}\right\}$ et par suite $\langle u_t, \mathbb{1}\rangle = 0$, si le sondage est de taille fixe (cas de l'EM-ZAE). Cette propriété est également vérifiée par les deux autres estimateurs de Δ.

– L'estimateur alternatif $\sum_t \Delta\pi_t\Delta\pi_t'$ donne des résultats proches de ceux de l'estimateur de Breidt et Chauvet (BC), pour la proximité aux probabilités d'inclusion simples. Néanmoins la distance $(\pi_1, \widehat{\pi}_1)$ est supérieure à celle de $\widehat{\Delta}^{BC}$ pour toutes les régions, sur $50\,000$ simulations. Une interprétation est que l'estimateur BC offre l'avantage de tenir compte exactement de la variance conditionnelle du tirage effectué par Cube à chacune de ses itérations, ce qui élimine une source de variabilité de l'estimateur de variance (Il est toujours vrai que $\mathrm{Var}\left[(\pi_t - \pi_{t-1})^2\right] \geqslant \mathrm{Var}\left[E\left\{(\pi_t - \pi_{t-1})^2\middle|\mathscr{F}_{t-1}\right\}\right]$.).

– Il parait difficile de majorer $\sum_t \underline{\lambda_t}\overline{\lambda_t}(u_t^i)^2$ plus finement, que ce soit en utilisant les propriétés géométriques ou statistiques de l'algorithme Cube.

47. notation : $(a,b) = [a,b]$ si $a \leqslant b$ et sinon $[b,a]$

Annexe E probabilités d'inclusion double des ZAE dans l'EMEX restreint

Cette annexe décrit l'estimation des probabilités d'inclusion double des ZAE dans l'échantillon maître restreint pour les extensions régionales [MCSF].

E.1 simulations du tirage de l'EMEX restreint

La taille par région de l'échantillon des ZAE de l'EMEX restreint est deux fois plus grande que celle de l'EM (Tableau E.45/Tableau 23) . Les simulations de 10 000 échantillons de ce tirage ont été stockées dans une table SAS par strate de tirage, nommée `zaeemexr_s_simules_&strate`. Ce nombre de simulations permet d'estimer directement (par la fréquence) les probabilités d'inclusion simple avec une marge d'erreur inférieure à 11%. Ce premier lot de simulations est utilisé pour calculer une dispersion des estimateurs de variance.

Tableau E.45 – Taille de l'échantillon de l'EMEX restreint par strate de tirage

région	nombre zae	$\lvert s\rvert$	$100min(\widehat{\pi_1}/\pi_1 - 1)$	$100max(\widehat{\pi_1}/\pi_1 - 1)$
11_pc	108	80	-2,6	5,2
11_gc	253	88	-5,8	7,7
21	115	25	-5,6	5,2
22	183	35	-7,6	7,2
23	141	30	-5,7	6,9
24	194	46	-8,7	5,8
25	148	27	-7,3	5,4
26	144	31	-5,5	6,5
31	235	71	-6,9	6,6
41	181	44	-6,8	7,2
42	123	28	-6,1	6,2
43	114	21	-6,5	6,2
52	198	53	-5,3	4,7
53	188	54	-5,2	5,9
54	138	33	-6,5	5,5
72	221	55	-7,7	10,5
73	213	45	-5,8	6,4
74	57	13	-4,4	6,1
82	363	98	-5,4	6,5
83	115	25	-7,0	4,9
91	140	39	-6,1	7,0
93	150	64	-4,5	4,3
94	19	6	-2,9	2,8
total	3 741	1 011	-8,7	10,5

Notes :

– *Les deux dernières colonnes donnent les taux d'écart minimal et maximal de* $\sum_{s \in S_d} s/\lvert S_d\rvert$ *par rapport à la probabilité d'inclusion simple dans l'EMEX restreint, en %.*

– *Les ZAE de Paris et de Boulogne-Billancourt sont exclues de ces comptages, ainsi que des tables stockées.*

– *Le programme de simulation de l'échantillonnage des ZAE de l'EMEX restreint est ci-joint* ***zaeemexr_s_simules***.

– Un deuxième lot de 1 000 000 réplications est utilisé pour une première estimation des probabilités d'inclusion par la méthode de Breidt-Chauvet.

– Un troisième lot d'un million de simulations sert au calcul du biais de la variance estimée, par l'estimateur Breidt-Chauvet. Il est pris en compte dans l'estimation des probabilités d'inclusion

définitives, qui sont donc calculées sur les deux millions de réplications réalisées.

E.2 qualité de l'estimation des probabilités d'inclusion dans l'EMEX-r

– La qualité de l'estimation des probabilités d'inclusion des ZAE dans l'EMEX restreint semble adéquate avec un million de réplications. Le taux d'erreur maximal sur la probabilité d'inclusion simple est inférieur à 0.4% pour toutes les strates de tirage (Tableau E.46), avec ce nombre de réplications. Les probabilités d'inclusion double paraissent strictement positives à l'exception éventuelle de la Corse [47]. Pour l'échantillon maître des ZAE, de taille inférieure de 50%, les deux plus petites régions ont une probabilité d'inclusion minimale quasiment nulle.

Tableau E.46 – Qualité de l'estimation des probabilités d'inclusion des ZAE dans l'EMEX restreint

	sur 2 000 000 réplications			taux d'erreur maximal (%)			
région	$\min(\widehat{\pi}_2)$	$\max(\widehat{\pi}_2)$	$\max[\lvert\widehat{\pi}_1 - \pi_1\rvert]$	2 000 000	1 000 000	100 000	50 000
11_pc	4,4E-02	1,00	0,0003	0,1	0,1	0,9	1,4
11_gc	7,3E-03	1,00	0,0003	0,3	0,3	0,2	0,5
21	4,8E-03	1,00	0,0003	0,2	0,3	0,8	1,2
22	5,4E-03	1,00	0,0004	0,3	0,3	0,9	1,5
23	5,7E-03	1,00	0,0004	0,2	0,3	1,0	1,3
24	6,6E-03	1,00	0,0003	0,2	0,3	1,1	1,2
25	4,1E-03	1,00	0,0004	0,3	0,3	1,0	1,4
26	4,5E-03	1,00	0,0003	0,2	0,3	0,9	2,1
31	5,6E-03	1,00	0,0003	0,2	0,4	1,3	1,5
41	2,9E-03	1,00	0,0003	0,3	0,3	0,9	1,6
42	7,4E-03	1,00	0,0003	0,2	0,3	0,7	1,3
43	2,8E-03	1,00	0,0002	0,3	0,3	1,0	1,5
52	7,9E-03	1,00	0,0003	0,3	0,3	0,9	1,1
53	1,0E-02	1,00	0,0003	0,2	0,2	0,8	0,9
54	6,6E-03	1,00	0,0004	0,2	0,3	1,1	1,2
72	6,4E-03	1,00	0,0004	0,2	0,4	0,8	1,2
73	5,6E-03	1,00	0,0003	0,3	0,4	0,9	1,5
74	3,3E-03	1,00	0,0003	0,2	0,3	0,8	1,1
82	6,2E-03	1,00	0,0003	0,2	0,3	1,4	1,6
83	2,9E-03	1,00	0,0003	0,2	0,3	1,1	1,2
91	3,1E-03	1,00	0,0005	0,2	0,2	0,8	1,1
93	3,7E-03	1,00	0,0003	0,1	0,1	0,8	0,9
94	1,0E-05	1,00	0,0002	0,1	0,2	0,3	0,7
total	1,0E-05	1,00	0,0005	0,3	0,4	1,4	2,1

Notes :

– *La durée d'exécution est de* $10^h 15$ *pour* 100 000 *réplications.*

– *Le programme d'estimation des probabilités d'inclusion double des ZAE dans l'EMEX restreint est ci-joint* ***pi2_hat_zaeemexr***.

• Le stockage des probabilités d'inclusion double de l'EMEX restreint est effectué dans une table SAS contenant la matrice carrée des probabilités d'inclusion double $\widehat{\pi}_2^{BC}$, par strate ZAE, non restreinte à l'échantillon effectif, sans correction de la diagonale, de nom générique :

`pi2_hat_zaeemexr_c&strate._2000000`

Cette table ne comporte pas d'identifiants. Les lignes et les colonnes sont dans l'ordre de la table de référence `zae_complet`.

48. Pour cette région, la probabilité d'inclusion minimale estimée sur un lot d'un million de réplications est négative.

E.3 qualité de l'estimation de variance des variables d'équilibrage - EMEX-r

– Comme pour l'échantillon maître des ZAE, le biais de la variance estimée sur l'échantillon pour les variables d'équilibrage (celles spécifiques à l'Ile-de-France sont exclues [48]) est calculé par rapport à la variance estimée sur un lot supplémentaire de 1 000 000 simulations. Les estimations 'échantillon' de variance sont moyennées sur ces mêmes simulations dans le calcul du biais.

→ L'estimateur de Yates-Grundy calculé avec les probabilités d'inclusion double estimées sur 1 000 000 réplications estime correctement la variance des variables d'équilibrage (Tableau E.47). En effet, le biais estimé représente moins de 0.04% de l'écart-type estimé sur les simulations. La qualité de l'estimation de variance ne semble plus s'améliorer par rapport à la version calculée avec les probabilités d'inclusion estimées sur 300 000 réplications. Cela suggère que le nombre de réplications réalisées est suffisant pour estimer la variance de l'EMEX restreint avec un biais quasiment nul.

Tableau E.47 – Biais relatif de l'estimation de la variance des totaux d'équilibrage estimés, en % de l'écart-type simulé

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-1	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	0,3	0,3	0,3	0,3	-100,0	-100,0	0,0	0,0	0,1	0,0	0,0	
11_gc	-0,2	0,1	0,0	0,1	0,1	0,0	-0,2	0,0	0,0	0,0	0,0	
21	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	
22	0,0	0,1	0,0	-0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,1	
23	-0,1	0,0	-0,1	0,1	-0,1	0,0	0,0	0,0	-0,1	0,1	0,0	
24	0,0	0,0	-0,1	0,0	-0,1	-0,1	0,0	-0,1	-0,1	0,0	0,0	
25	-0,1	0,0	0,1	0,0	0,0	-0,1	-0,1	0,0	0,1	-0,1	-0,2	
26	0,0	0,0	0,1	0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,0	
31	0,1	0,1	-0,1	-0,1	-0,1	-0,1	0,1	0,1	-0,1	0,1	-0,1	
41	-0,1	0,0	0,1	0,1	0,1	0,0	0,0	0,1	0,0	0,1	-0,1	
42	-0,1	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	
43	-0,1	0,1	-0,1	0,2	0,1	0,1	-0,1	0,1	0,0	0,2	0,1	
52	-0,2	0,0	0,0	-0,2	0,0	0,0	-0,2	0,0	-0,1	-0,2	0,0	
53	0,0	0,0	0,1	0,0	0,0	0,0	0,1	0,1	0,1	0,0	0,0	
54	0,0	0,0	0,1	0,0	-0,1	-0,1	-0,1	0,0	0,0	0,0	0,0	
72	0,0	-0,1	0,0	-0,1	-0,1	-0,1	0,0	-0,1	-0,1	-0,1	0,1	
73	0,0	0,0	0,0	-0,2	0,0	0,1	0,0	0,0	-0,1	-0,1	-0,1	
74	0,0	0,0	-0,1	0,0	-0,1	-0,1	0,1	-0,1	-0,1	-0,1	0,0	
82	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,1	0,1	0,0	
83	0,0	0,0	0,1	-0,2	0,1	0,0	0,0	0,0	0,1	-0,2	0,1	
91	0,0	0,0	0,1	0,2	0,0	0,1	0,0	0,0	0,1	0,1	0,0	
93	0,1	0,0	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,1	0,0	
94	-0,6	-0,4	-1,4	-1,1	-0,6	-0,8	-0,9	-0,5	-1,8	-1,2	-0,5	
1 000 000	-0,03	-0,01	0,00	-0,04	-0,02	-0,03	-0,03	-0,01	-0,03	-0,02	-0,01	
300 000	-0,03	-0,01	0,01	-0,04	0,02	0,00	-0,03	-0,02	-0,01	0,00	-0,02	
100 000	-0,04	0,00	0,00	-0,04	-0,03	-0,01	-0,05	0,04	-0,06	0,07	-0,03	
50 000	0,14	0,14	0,12	0,02	0,27	0,11	0,14	0,02	0,10	0,01	0,01	

Notes :

– *Il s'agit du taux d'écart entre la racine carrée de la moyenne des estimations de variance sur* 1 000 000 *simulations et celle de la variance calculée sur ces simulations.*

– *Les 4 dernières lignes donnent l'effet du nombre de réplications.*

– *Le programme est ci-joint* **`biais_relatif`** + **`biais_relatif_equilib_yg`** .

49. Les variables d'équilibrage de la petite couronne de l'Ile-de-France ne comportent pas la distinction par groupe de rotation du revenu fiscal et du nombre de résidences principales.

– L'estimateur d'Horvitz-Thompson semble moins précis, mais le biais relatif reste largement inférieur à 1% (Tableau E.48). La troncature des probabilités d'inclusion inférieures à 10^{-3} induit une surestimation, sensible pour deux des variables d'équilibrage.

Tableau E.48 – Comparaison des biais des estimateurs de variance de Yates-Grundy et Horvitz-Thompson, en % de l'écart-type simulé

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HT	0,3	0,0	0,1	0,4	0,2	0,2	0,4	0,0	0,1	0,3	0,2	
HTc	0,4	0,1	0,1	0,4	0,2	0,2	0,4	0,0	0,1	0,3	0,2	
HTt	0,9	0,3	0,3	0,6	0,3	0,6	1,0	0,2	0,1	0,4	0,4	

Note : HTc= Horvitz-Thompson après correction de la diagonale de $\widehat{\pi}_2$; HTt= Horvitz-Thompson avec troncature des $\widehat{\pi}_2 < 10^{-3}$ et correction de la diagonale

• L'avantage de l'estimateur de Yates-Grundy apparait encore plus nettement lorsque la comparaison porte sur l'estimation de variance réalisée uniquement sur le premier échantillon simulé (Tableau E.49). Ce résultat peut s'interpréter comme un signal de plus grande stabilité de cet estimateur de variance par rapport à la version Horvitz-Thompson.

Tableau E.49 – Comparaison des estimateurs de variance de Yates-Grundy et Horvitz-Thompson appliqués au premier échantillon simulé, en % de l'écart-type simulé

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	21,5	23,1	-24,2	14,8	6,1	29,0	18,3	22,7	-9,6	1,5	8,9	
HT	71,7	22,0	-100,0	-6,5	-53,7	49,7	73,1	4,6	-100,0	6,7	8,0	

Note : Une erreur de -100% correspond à une estimation de variance négative.

→ Pour les 138 variables issues des recensements (voir la partie sur la variance de l'EM), le biais relatif est compris -0.11% et +0.03%, et négatif dans 129 cas. Le taux d'erreur de l'écart-type estimé sur le premier échantillon simulé varie entre -23% et +21% sur ces variables.

– La dispersion de l'estimateur de variance de l'EMEX restreint est comprise entre celles de l'EM et de l'EMEX élargi (Tableau E.50), pour toutes les variables d'équilibrage.

Tableau E.50 – Dispersion relative de l'estimateur de variance Yates-Grundy, en % de l'écart-type simulé

$\lvert R\rvert$	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
1 000 000	79,7	78,6	79,3	75,2	67,7	70,1	77,0	75,0	74,4	70,5	72,0	
100 000	79,6	78,6	79,3	75,2	67,6	70,0	77,0	75,2	74,3	70,5	72,0	
50 000	79,7	78,6	79,4	75,2	68,0	70,6	77,0	74,8	74,7	70,5	72,0	
EM	81,8	82,6	81,5	88,2	89,5	81,8	78,9	89,3	76,4	74,7	77,0	
EMEX élargi	75,8	74,8	75,5	70,7	60,2	64,9	73,6	69,0	71,7	66,9	67,7	

– Pour la grande majorité des variables 'recensement', la dispersion de l'estimateur Breidt-Chauvet est inférieure à celle de Deville-Tillé.

E.4 erreur observée sur l'échantillon EMEX-r effectivement tiré

Sur l'EMEX restreint effectivement tiré, le taux d'erreur atteint des valeurs élevées (Tableau E.51). Pour la moitié des variables d'équilibrage, le taux d'erreur absolu est supérieur à celui observé sur les deux premiers échantillons simulés. Néanmoins, il ne semble pas nécessaire de réaliser une estimation spécifique pour la Corse, contrairement au cas de l'échantillon maître.

Tableau E.51 – Taux d'erreur de l'estimation sur l'EMEX restreint effectif

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	1,7	1,7	1,7	1,7	-100,0	-100,0		79,3	83,8	59,7	77,0	82,6
11_gc	90,9	112,7	101,1	98,2	-100,0	-100,0		43,5	31,6	97,3	31,3	45,9
21	8,4	35,4	-100,0	-100,0	-28,9	80,3		32,9	59,3	-100,0	-100,0	-29,6
22	23,4	-62,2	-100,0	50,9	0,2	-100,0		-100,0	16,1	-30,3	-100,0	-19,2
23	-100,0	-60,6	24,8	50,8	20,5	92,0		-100,0	-100,0	-100,0	56,0	100,5
24	-100,0	21,8	74,9	24,7	91,4	101,4		44,1	-10,2	-14,3	33,1	-100,0
25	30,4	-49,7	58,8	61,6	-15,6	47,4		43,4	-100,0	31,8	73,8	-22,7
26	-81,9	80,2	65,6	75,3	24,8	-100,0		-100,0	77,9	49,8	81,0	59,6
31	126,4	1,4	20,5	89,9	22,2	-4,8		158,8	44,8	12,4	11,7	7,0
41	-49,9	-100,0	18,8	110,3	25,8	97,0		45,8	-100,0	-100,0	106,1	22,9
42	-100,0	66,7	23,9	59,7	43,3	-100,0		-100,0	63,2	70,8	46,8	-100,0
43	-100,0	-100,0	126,6	-100,0	296,4	56,2		-44,3	-100,0	74,6	-23,5	70,2
52	181,3	54,0	1,5	98,8	-37,1	16,7		199,0	-100,0	-100,0	114,9	180,1
53	8,1	98,7	-100,0	61,1	-100,0	-100,0		-21,4	-93,0	-100,0	81,8	67,3
54	73,9	42,3	16,7	-90,6	-12,1	26,8		69,9	4,1	45,2	-100,0	129,9
72	59,7	-100,0	-100,0	-53,0	76,1	37,8		11,7	-100,0	-100,0	19,6	62,4
73	99,6	-100,0	49,9	104,1	83,2	69,1		62,4	-50,0	34,4	64,5	58,4
74	-100,0	67,6	3,6	-57,6	124,5	51,5		-100,0	88,2	118,3	-75,2	179,4
82	-100,0	-100,0	-100,0	52,6	17,7	-0,3		-100,0	-100,0	50,0	110,3	-100,0
83	-100,0	96,6	38,7	51,9	23,6	-100,0		-100,0	66,4	24,4	32,2	40,4
91	-100,0	58,4	-100,0	58,1	-100,0	-100,0		-100,0	62,0	-100,0	66,9	11,5
93	120,1	77,3	67,2	22,2	-32,9	-100,0		123,6	57,9	79,2	69,8	-100,0
94	152,9	218,9	57,0	-5,8	33,3	57,5		140,8	195,7	-11,5	-22,6	105,5
total	-78,8	19,0	-1,0	44,9	47,5	-13,8		-40,6	9,9	-3,9	43,6	38,2
HT	-100,0	19,9	76,5	2,9	71,0	-100,0		-100,0	-44,7	37,4	15,4	-100,0
s1	21,5	23,1	-24,2	14,8	6,1	29,0		18,3	22,7	-9,6	1,5	8,9
s2	2,0	24,9	32,3	5,2	21,1	18,1		-2,9	-12,4	19,5	19,4	-13,1

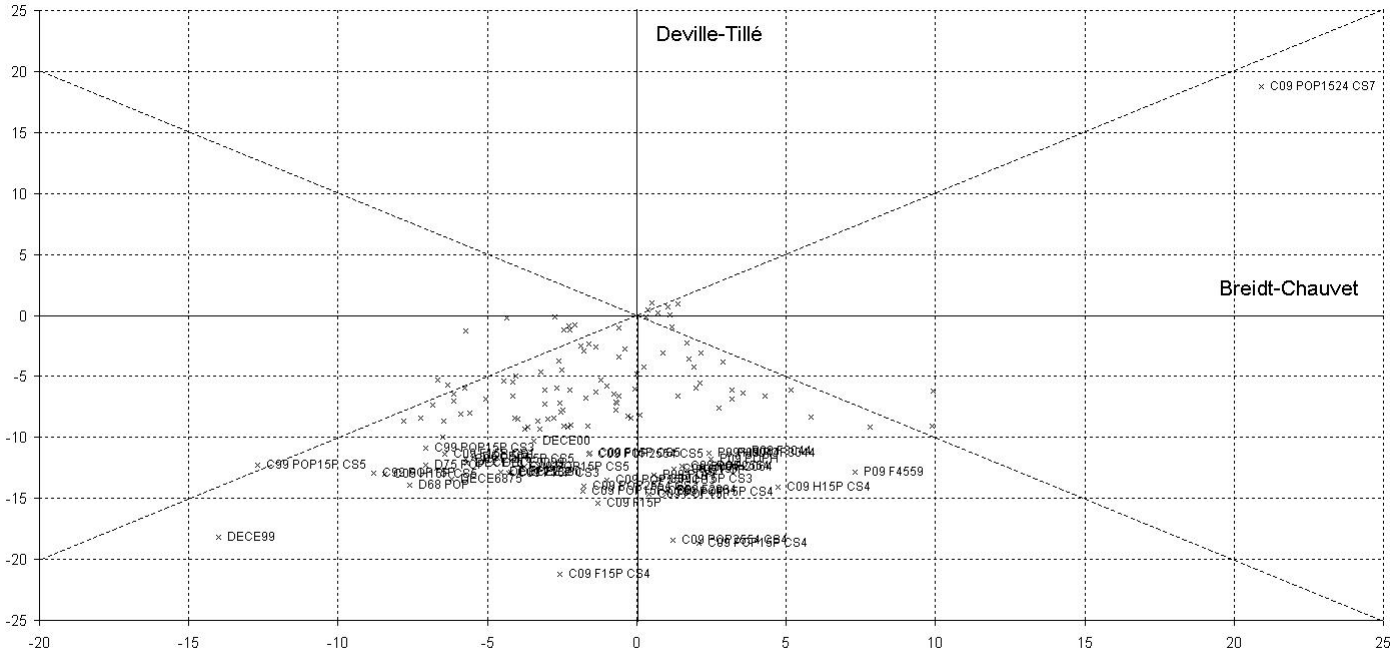
Notes :

– *Le taux d'erreur est mesuré par rapport à l'écart-type estimé sur les* 1 000 000 *simulations, en pourcentage.*

– *-100% correspond à une estimation de variance négative.*

– Pour les variables des recensements, le taux d'erreur de l'estimation sur l'échantillon effectif est compris entre -14 et +21%. L'approximation de Deville-Tillé [49] sous-estime la variance de la quasi-totalité de ces variables (Graphique 13). De plus, l'ampleur de l'erreur est dans la majorité des cas plus grande.

Graphique 13 – Erreur mesurée sur l'EMEX restreint effectif, en pourcentage de l'écart-type simulé



– L'approximation de Deville surestime la variance de toutes ces variables sur l'échantillon effectif, à une exception près. Pour presque toutes ces variables, l'ampleur de l'erreur est encore plus grande que celle de l'approximation de Deville-Tillé.

E.5 conclusions sur l'Emex restreint

- La méthode de Breidt-Chauvet appliquée à 1 000 000 réplications fournit des probabilités estimées d'inclusion double de l'EMEX restreint qui permettent d'estimer correctement la variance de cet échantillonnage.
- L'estimateur de variance de Yates-Grundy est préférable à celui d'Horvitz-Thompson pour l'échantillonnage de l'EMEX restreint, selon les indicateurs de qualité étudiés.
- Cet estimateur parait nettement plus précis que les approximations de Deville et de Deville-Tillé, pour les variables étudiées.

50. Le programme est ci-joint **biais_relatif_varrpdt_yg**.

Annexe F probabilités d'inclusion double des ZAE dans l'EMEX élargi

Cette annexe décrit l'estimation des probabilités d'inclusion double des ZAE dans l'échantillon maître élargi [MCSF].

F.1 simulations du tirage de l'EMEX élargi

La taille par région de l'échantillon des ZAE de l'EMEX élargi est accrue de 50% par rapport à l'EMEX restreint (Tableau F.52). Les simulations de 10 000 échantillons de ce tirage sont stockées dans une table SAS par strate de tirage, nommée `zaeemexe_s_simules_&strate`. Ce premier lot de simulations sert à estimer une dispersion des estimateurs de variance.

Tableau F.52 – Taille de l'échantillon de l'EMEX élargi par strate de tirage

région		nombre zae	$\lvert s\rvert$	$100min(\widehat{\pi}_1/\pi_1 - 1)$	$100max(\widehat{\pi}_1/\pi_1 - 1)$
11_pc	Île-de-France/petite couronne	108	108	0,0	0,0
11_gc	Île-de-France/grande couronne	253	132	-4,5	4,2
21	Champagne-Ardenne	115	37	-4,0	3,7
22	Picardie	183	52	-5,5	5,5
23	Haute-Normandie	141	44	-5,6	4,3
24	Centre	194	68	-5,7	4,8
25	Basse-Normandie	148	40	-5,2	4,5
26	Bourgogne	144	46	-8,5	5,0
31	Nord-Pas-de-Calais	235	106	-6,7	4,3
41	Lorraine	181	65	-4,0	5,6
42	Alsace	123	41	-4,9	3,7
43	Franche-Comté	114	31	-5,1	5,5
52	Pays de la Loire	198	78	-5,0	3,8
53	Bretagne	188	80	-3,3	5,0
54	Poitou-Charentes	138	49	-4,4	6,0
72	Aquitaine	221	82	-7,2	6,7
73	Midi-Pyrénées	213	67	-5,0	5,1
74	Limousin	57	19	-3,2	4,5
82	Rhône-Alpes	363	145	-4,9	5,2
83	Auvergne	115	37	-3,9	4,4
91	Languedoc-Roussillon	140	57	-5,2	4,7
93	Provence-Alpes-Côte d'Azur	150	94	-2,7	4,0
94	Corse	19	9	-3,1	1,9
	total	3 741	1 487	-8,5	6,7

Notes :

– *Les deux dernières colonnes contiennent les taux d'écart minimal et maximal de* $\bar{s}^{S_d} = \sum_{s \epsilon S_d} s/|S_d|$ *par rapport à la probabilité d'inclusion des ZAE dans l'EMEX élargi, en %.*

– *Les ZAE de Paris et de Boulogne-Billancourt sont exclues de ces comptages, ainsi que des tables stockées.*

– *Le programme de simulation de l'échantillonnage des ZAE de l'EMEX élargi est ci-joint* ***zaeemexe_s_simules***.

– 300 000 réplications donnent une première estimation des probabilités d'inclusion double par la méthode de Breidt-Chauvet.

– Un troisième lot de 400 000 simulations a été réalisé pour la validation (via les estimations de l'espérance de l'estimateur 'échantillon' de la variance $\widehat{\mathrm{Var}\left(\widehat{Y}\right)}^{s}$ et de la variance de référence $\mathrm{Var}\left(\widehat{Y}\right)$).

F.2 qualité de l'estimation des probabilités d'inclusion dans l'EMEX-e

– La qualité de l'estimation Breidt-Chauvet sur 300 000 réplications des probabilités d'inclusion des ZAE dans l'EMEX élargi semble satisfaisante. Le taux d'erreur maximal sur la probabilité d'inclusion simple est inférieur à 0.5% pour toutes les strates de tirage (Tableau F.53). Cette proximité est meilleure que celle atteinte par un million de réplications de l'échantillon maître. Ceci peut se comprendre par les probabilités d'inclusion plus élevées [50]. Les probabilités d'inclusion double paraissent strictement positives dans toutes les régions, contrairement aux deux autres échantillons maîtres. Leurs minimums sont supérieurs à ceux de l'EMEX restreint, ce qui est cohérent avec la taille plus élevée de l'échantillon.

Tableau F.53 – Qualité de l'estimation des probabilités d'inclusion des ZAE dans l'EMEX élargi

	sur 700 000 réplications			taux d'erreur maximal			
région	$\min(\widehat{\pi}_2)$	$\max(\widehat{\pi}_2)$	$\max[\lvert\widehat{\pi}_1 - \pi_1\rvert]$	700 000	300 000	100 000	50 000
11_pc	1,0E+00	1,00	0,0E+00	0,0	0,0	0,0	0,0
11_gc	1,9E-02	1,00	7,3E-04	0,3	0,5	0,6	0,8
21	1,7E-02	1,00	5,8E-04	0,3	0,4	0,8	0,7
22	1,5E-02	1,00	6,2E-04	0,3	0,4	0,7	0,9
23	1,4E-02	1,00	5,4E-04	0,2	0,5	0,5	1,1
24	1,7E-02	1,00	5,2E-04	0,3	0,4	0,8	0,8
25	1,2E-02	1,00	5,2E-04	0,3	0,4	0,9	1,4
26	1,5E-02	1,00	6,2E-04	0,3	0,5	0,9	0,9
31	1,7E-02	1,00	4,9E-04	0,2	0,5	0,8	0,8
41	1,2E-02	1,00	5,6E-04	0,3	0,4	0,9	0,9
42	1,9E-02	1,00	6,2E-04	0,4	0,4	0,5	1,2
43	1,2E-02	1,00	5,4E-04	0,3	0,5	0,9	0,9
52	2,0E-02	1,00	4,8E-04	0,2	0,4	0,6	1,0
53	2,6E-02	1,00	6,6E-04	0,2	0,4	0,7	0,7
54	1,7E-02	1,00	5,7E-04	0,3	0,4	0,6	0,9
72	1,6E-02	1,00	5,3E-04	0,2	0,5	0,8	1,0
73	1,5E-02	1,00	4,4E-04	0,3	0,4	0,6	1,1
74	1,4E-02	1,00	5,0E-04	0,2	0,3	0,6	0,8
82	1,5E-02	1,00	6,6E-04	0,3	0,5	1,0	1,0
83	1,3E-02	1,00	5,0E-04	0,2	0,3	0,6	0,8
91	1,2E-02	1,00	4,8E-04	0,3	0,4	0,8	0,8
93	1,5E-02	1,00	5,4E-04	0,2	0,2	0,5	0,9
94	8,1E-03	1,00	2,7E-04	0,1	0,3	0,4	0,6
total	8,1E-03	1,00	7,3E-04	0,4	0,5	1,0	1,4

Notes :

– *Toutes les ZAE de la petite couronne de l'Ile-de-France sont dans l'EMEX élargi.*

– *La durée d'exécution est de* 9^h10 *pour* 100 000 *réplications.*

– *Le programme d'estimation des probabilités d'inclusion double des ZAE de l'EMEX élargi est ci-joint* ***pi2_hat_zaeemexe***.

• Les probabilités d'inclusion double de l'EMEX élargi sont stockées dans une table SAS, contenant la matrice carrée des probabilités d'inclusion double $\widehat{\pi}_2^{BC}$, par strate ZAE, non restreinte à l'échantillon effectif, sans correction de la diagonale, sous le nom générique :

`pi2_hat_zaeemexe_c&strate._700000`

L'ordre des lignes et des colonnes correspond à celui de la table de référence des ZAE `zae_complet`.

51. Avec l'approximation de $n\widehat{p}$ par une loi binomiale $\mathcal{B}(n,p)$, où n est le nombre de simulations, le coefficient de variation est de l'ordre de $\sigma(\widehat{p}) \,/\, p = \sqrt{\left(\frac{1}{p} - 1\right) \Big/ n}$, donc décroissant en fonction de p.

F.3 qualité de l'estimation de variance des variables d'équilibrage - EMEX-e

– La qualité de l'estimation des variables d'équilibrage de l'EMEX élargi (à l'exclusion de celles spécifiques à l'Ile-de-France) est calculée par rapport à la variance estimée sur 400 000 simulations par la méthode de Breidt-Chauvet. L'estimation de variance sur l'échantillon est moyennée sur les mêmes simulations.

→ L'estimateur de Yates-Grundy estime convenablement la variance des variables d'équilibrage (Tableau F.54). Le biais relatif devient imperceptible ($\leqslant 0.04\%$) avec des probabilités d'inclusion estimées sur 300 000 réplications. Il est inférieur à 0.3% pour toutes les régions.

Tableau F.54 – Biais relatif de l'estimateur Yates-Grundy de la variance des totaux d'équilibrage estimés, en % de l'écart-type simulé de l'EMEX élargi

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0
11_gc	0,0	0,1	0,0	0,3	-0,1	-0,1	0,0	-0,1	0,1	0,3	0,0	
21	0,2	-0,1	0,0	0,1	0,0	0,0	0,1	-0,1	0,0	-0,2	0,1	
22	0,0	0,0	0,0	0,0	-0,1	-0,1	-0,1	0,0	-0,2	0,0	0,1	
23	0,1	-0,1	0,0	-0,1	-0,1	-0,2	0,0	-0,2	-0,1	-0,1	0,0	
24	0,0	-0,2	0,1	0,0	-0,1	0,0	0,1	-0,2	0,0	-0,1	-0,1	
25	0,0	0,1	-0,1	0,0	0,1	0,0	-0,1	0,2	-0,1	-0,1	0,0	
26	-0,3	0,0	-0,1	-0,2	-0,1	0,1	-0,3	0,0	-0,2	-0,1	-0,2	
31	-0,1	0,2	-0,1	0,1	0,0	0,1	0,0	0,1	0,0	-0,1	0,1	
41	0,1	0,1	-0,1	0,0	0,1	0,0	0,0	-0,1	0,0	0,0	0,3	
42	-0,2	0,2	0,1	0,2	-0,2	0,0	-0,1	0,2	0,0	0,2	0,2	
43	0,0	0,1	0,0	0,0	-0,1	0,2	0,0	0,2	0,0	0,0	0,3	
52	0,1	-0,1	0,1	0,1	0,0	-0,1	0,0	0,0	0,1	0,0	0,1	
53	0,0	0,0	0,2	0,0	-0,1	0,0	0,1	0,0	0,1	0,0	0,2	
54	-0,1	-0,1	-0,2	0,0	0,0	0,0	0,0	-0,1	-0,2	0,1	-0,1	
72	0,0	-0,2	0,2	-0,2	0,0	0,0	0,1	-0,1	0,2	-0,2	-0,1	
73	0,1	-0,1	0,1	-0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,2	
74	0,0	0,2	-0,2	-0,1	0,0	0,0	0,1	0,1	-0,1	0,0	0,0	
82	0,1	0,1	0,2	0,2	0,2	0,0	0,1	0,1	0,0	0,3	-0,1	
83	-0,1	-0,2	-0,2	0,1	0,0	-0,1	-0,1	-0,2	-0,1	0,0	0,0	
91	0,1	0,1	0,0	-0,1	0,1	0,1	-0,1	0,1	0,0	-0,1	0,1	
93	-0,1	0,1	0,0	-0,1	0,0	-0,2	0,0	0,1	0,0	0,1	0,0	
94	-0,1	-0,1	-0,1	0,0	-0,2	-0,2	-0,2	-0,1	0,0	-0,1	-0,1	
300 000	0,00	0,02	0,00	0,00	-0,02	-0,01	-0,01	-0,01	-0,01	0,01	0,04	
100 000	0,08	0,00	0,00	-0,01	-0,01	0,01	0,09	0,01	0,00	0,03	0,08	
50 000	0,07	-0,04	-0,01	0,06	-0,01	0,01	0,03	0,03	-0,02	0,04	0,09	

Note : Les programmes utilisés sont **`biais_relatif`** + **`biais_relatif_equilib_yg`**.

• Pour les 138 variables socio-démographiques, celles étudiées pour l'EM, le biais relatif est compris entre -0.02% et +0.02%. Il est négatif dans 101 cas.

• L'estimateur d'Horvitz-Thompson semble légèrement plus biaisé pour la plupart des variables d'équilibrage (Tableau F.55). La correction de la diagonale a un effet négligeable. Pour les variables des recensements, le biais relatif de cet estimateur est compris entre -0.27% et +0.10%.

Tableau F.55 – Comparaison des versions Yates-Grundy et Horvitz-Thompson de l'estimation de la variance des totaux d'équilibrage estimés, en % de l'écart-type simulé de l'EMEX élargi

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HT	0,1	0,0	-0,1	-0,3	-0,2	0,0	0,1	0,0	-0,1	-0,1	-0,1	
HTc	0,2	0,0	-0,1	-0,3	-0,2	0,0	0,2	0,1	0,0	-0,1	0,0	

– La version de Yates-Grundy apparait encore plus précise par rapport à celle d'Horvitz-Thompson lorsque l'estimation de variance n'est pas moyennée, mais calculée sur le premier échantillon simulé (Tableau F.56). Ce résultat illustre la plus grande stabilité de l'estimateur de variance de Yates-Grundy, observée également pour les autres échantillonnages de ce document.

Tableau F.56 – Comparaison des estimateurs de variance calculés sur le premier échantillon simulé

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
YG	4,4	-2,5	17,5	19,2	4,3	4,5	1,1	19,0	2,4	29,2	30,8	
HT	-100,0	18,2	-45,3	-26,4	13,6	-58,2	-100,0	9,3	-57,7	18,3	9,1	

Note : -100% correspond à une estimation de variance négative.

– La dispersion de l'estimateur de variance de l'EMEX élargi est sensiblement plus basse que celle de l'échantillon maître (Tableau F.57), ce qui était prévisible. Il ne semble pas y avoir d'effet du nombre de réplications sur cette dispersion. Ceci pourrait signifier que l'impact sur la variance estimée par Yates-Grundy de l'estimation des probabilités d'inclusion est très petit par rapport à l'effet de l'aléa d'échantillonnage [51]. C'est un élément favorable à l'estimateur de variance construit (robustesse par rapport à $\widehat{\pi}_2{}^R$).

Tableau F.57 – Dispersion relative de l'estimateur de Yates-Grundy

$\|R\|$	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
300 000	75,8	74,8	75,5	70,7	60,2	64,9	73,6	69,0	71,7	66,9	67,7	
50 000	75,9	74,8	75,5	70,7	60,2	64,9	73,6	69,0	71,7	67,0	67,7	
EM	81,8	82,6	81,5	88,2	89,5	81,8	78,9	89,3	76,4	74,7	77,0	

note : Au niveau régional, la dispersion de l'estimateur de variance des variables d'équilibrage reste supérieure à 100%, quasiment pour toutes les mesures.

52. $\widehat{\mathrm{Var}^{s}\left(\widehat{Y}\right)}$ dépend de s et de $\widehat{\pi}_2{}^{BC,R}$. Une autre interprétation est que le nombre de simulations de 10 000 est trop réduit pour permettre de détecter l'effet sur la dispersion du nombre de réplications.

– L'estimateur d'Horvitz-Thompson semble sensiblement plus dispersé pour les variables d'équilibrage (Tableau F.58).

Tableau F.58 – Dispersion relative de l'estimateur d'Horvitz-Thompson

estimateur	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
HT	117,0	115,6	116,2	108,8	79,4	90,2		116,4	107,0	110,3	106,9	104,3
YG	75,8	74,8	75,5	70,7	60,2	64,9		73,6	69,0	71,7	66,9	67,7

Note : La correction de la diagonale n'a quasiment aucune incidence sur la dispersion de l'estimateur d'Horvitz-Thompson.

– La dispersion de l'estimateur de variance des variables 'recensement' est presque toujours plus faible avec la méthode Breidt-Chauvet qu'avec Deville-Tillé (dans 135 cas sur 138).

F.4 erreur observée sur l'échantillon effectif de l'EMEX-e

L'erreur observée sur l'échantillon effectif est similaire à celle du premier échantillon simulé (Tableau F.59).

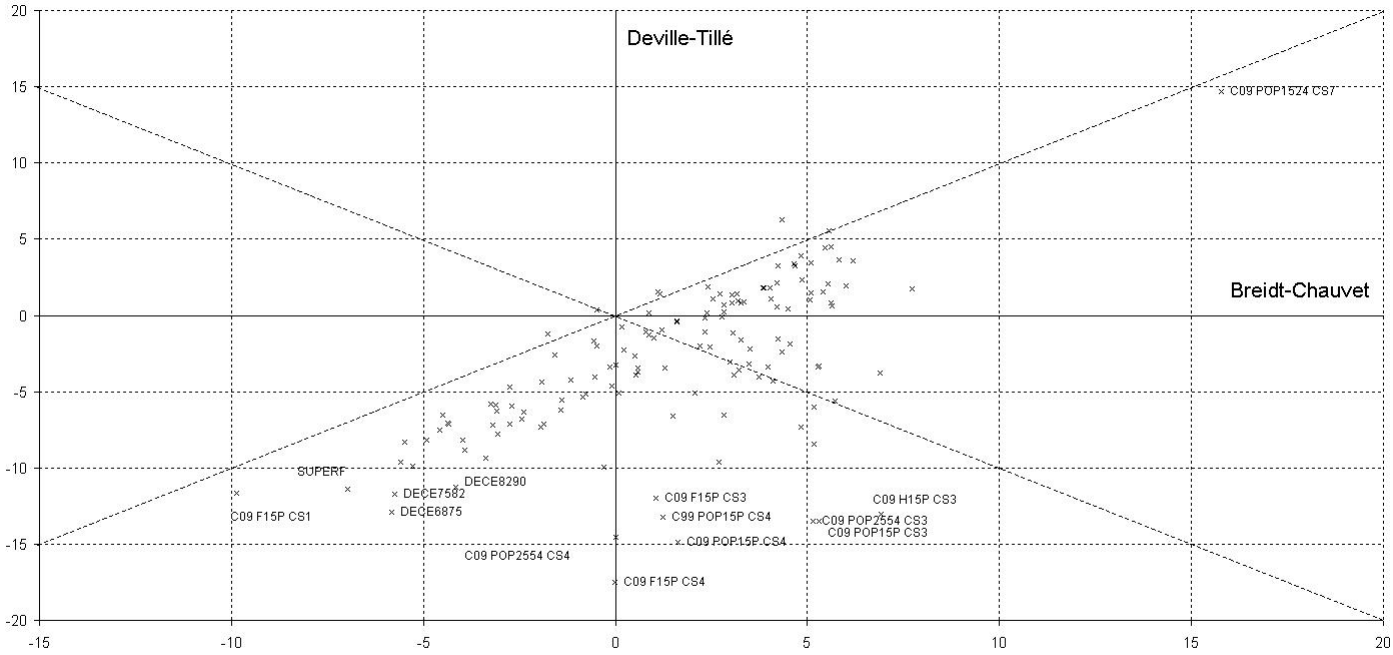
Tableau F.59 – Taux d'erreur mesuré sur l'EMEX élargi, en % de l'écart-type simulé

région	nresgr1	nresgr2	nresgr3	nresgr4	nres rural	nres bain	periur-	revenufisc04 1	revenufisc04 2	revenufisc04 3	revenufisc04 4	revenufisc04 5
11_pc	-100,0	-100,0	-100,0	-100,0	-100,0	-100,0		-100,0	-100,0	-100,0	-100,0	-100,0
11_gc	71,6	59,1	42,3	124,8	-100,0	-100,0		-100,0	10,3	66,2	93,3	31,8
21	-43,6	-100,0	-22,3	-100,0	-9,0	8,0		-15,0	9,3	-66,9	-100,0	-47,2
22	-70,4	38,2	-100,0	8,3	31,1	-21,4		-100,0	66,0	-35,7	-100,0	27,2
23	0,5	-60,1	26,8	60,1	7,0	36,7		16,2	-87,5	-41,3	50,3	89,8
24	-32,9	102,7	80,3	107,7	95,7	127,8		-16,4	44,7	-100,0	90,4	48,4
25	-100,0	23,6	55,4	-42,1	67,0	46,3		33,8	18,8	46,4	-0,6	47,1
26	-2,0	37,2	20,8	8,4	25,2	-100,0		-13,0	45,0	29,5	18,6	5,3
31	174,2	-100,0	115,4	49,8	51,2	-43,2		153,5	-100,0	69,5	-77,4	48,4
41	35,0	-82,7	15,4	84,5	-38,5	64,7		120,6	-100,0	57,1	58,5	22,3
42	-100,0	10,4	-58,2	86,5	65,0	-67,0		-57,6	29,9	91,9	70,9	32,5
43	-100,0	-12,0	61,1	-27,0	137,4	47,4		36,1	-11,1	5,4	26,3	-100,0
52	144,5	11,6	42,8	-75,5	-81,3	-23,3		155,4	-100,0	7,8	30,8	117,1
53	-19,3	127,9	-100,0	34,0	-72,8	-100,0		41,5	58,7	-100,0	102,3	16,1
54	53,8	-1,4	35,4	-100,0	29,0	-13,6		50,8	-9,4	57,3	-100,0	-100,0
72	102,3	50,9	-100,0	63,6	57,3	14,9		58,4	42,1	-100,0	82,2	-24,4
73	95,7	-35,6	-7,0	115,0	2,4	-100,0		85,6	8,3	-14,8	86,9	75,5
74	-100,0	27,2	-27,3	-100,0	43,4	-11,0		-100,0	24,2	24,8	-100,0	60,5
82	-100,0	-100,0	-100,0	29,3	18,5	33,0		-100,0	-100,0	34,3	32,4	-26,9
83	-42,9	3,7	6,8	-13,8	-100,0	-100,0		-100,0	-27,0	-63,9	-20,4	53,4
91	-42,9	119,6	-0,5	43,9	-100,0	-100,0		-16,7	46,3	-50,7	73,9	59,2
93	43,1	-3,1	97,5	-61,6	-100,0	-100,0		142,4	46,9	93,8	-100,0	-100,0
94	-100,0	40,4	-100,0	-20,1	-10,5	-24,7		-100,0	39,1	-42,5	-6,4	-100,0
total	-8,3	14,0	1,4	28,6	15,9	-33,2		5,2	0,8	1,3	33,5	29,3
HT	-100,0	84,4	28,9	50,8	34,6	-45,7		-21,5	54,4	17,9	53,8	16,8
s_1	4,4	-2,5	17,5	19,2	4,3	4,5		1,1	19,0	2,4	29,2	30,8

– Pour les variables des recensements, le taux d'erreur observé sur ce seul échantillon est compris entre -10% et +16% de l'écart-type simulé, ce qui parait relativement satisfaisant.

Pour la majorité des variables issues des recensements, l'ampleur de l'erreur de l'approximation de Deville-Tillé mesurée sur l'EMEX élargi effectivement tiré est supérieure à celle de l'estimation par la méthode de Breidt-Chauvet (Graphique 14). Néanmoins le nombre de ces variables pour lesquelles l'approximation est moins précise est plus réduit que pour l'échantillon EMEX restreint (respectivement 77 et 119).

Graphique 14 – Erreur mesurée sur l'EMEX élargi effectif, en pourcentage de l'écart-type simulé



F.5 conclusions sur la probabilité d'inclusion de l'Emex élargi

- Un nombre de réplications de 300 000 parait suffire pour estimer par la méthode de Breidt-Chauvet les probabilités d'inclusion de l'EMEX élargi avec une précision telle que le biais des estimations de variance soit pratiquement nul.
- L'estimateur de Yates-Grundy avec les probabilités estimées par Breidt-Chauvet parait clairement préférable à l'approximation de Deville-Tillé pour cet échantillonnage.
- L'augmentation du nombre de réplications ne parait pas susceptible de réduire la variabilité de l'estimateur de variance. Mais cette conclusion est fragile.
- Les probabilités d'inclusion définitives sont estimées sur les 700 000 réplications y compris celles utilisées pour la validation.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS, K. MOUSSALLAM