LES CALCULS DE PRECISION POUR LES ENQUETES MENAGES DE L'INSEE TIREES DANS OCTOPUSSE

Emmanuel Gros[1], Sébastien Faivre[2] et Karim Moussallam[3]

[1] *Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE – emmanuel.gros@insee.fr*
[2] *Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE – sebastien.faivre@insee.fr*
[3] *Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE – karim.moussallam@insee.fr*

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 système d'échantillonnage plus efficace 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.

Nous présentons ici une formule de variance analytique « générique » valable pour toute enquête standard tirée dans Octopusse. Prolongeant les travaux exposés par Guillaume Chauvet dans [2], cette méthode repose d'une part sur l'utilisation d'estimateurs de variance de Yates-Grundy, avec estimation par réplication des probabilités d'inclusion double, pour estimer la variance relative aux deux premières phases du système – sélection de l'enquête annuelle de recensement, sélection des unités primaires du nouvel échantillon-maître – et d'autre part sur une application itérée de la formule de Rao pour prendre en compte les degrés de sondage ultérieurs relatifs aux sélections de logements – tirage des logements au sein de la fraction recensée des unités primaires de l'échantillon-maître, non-réponse.

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.

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

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse[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] 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[2] 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 de rotation. 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.

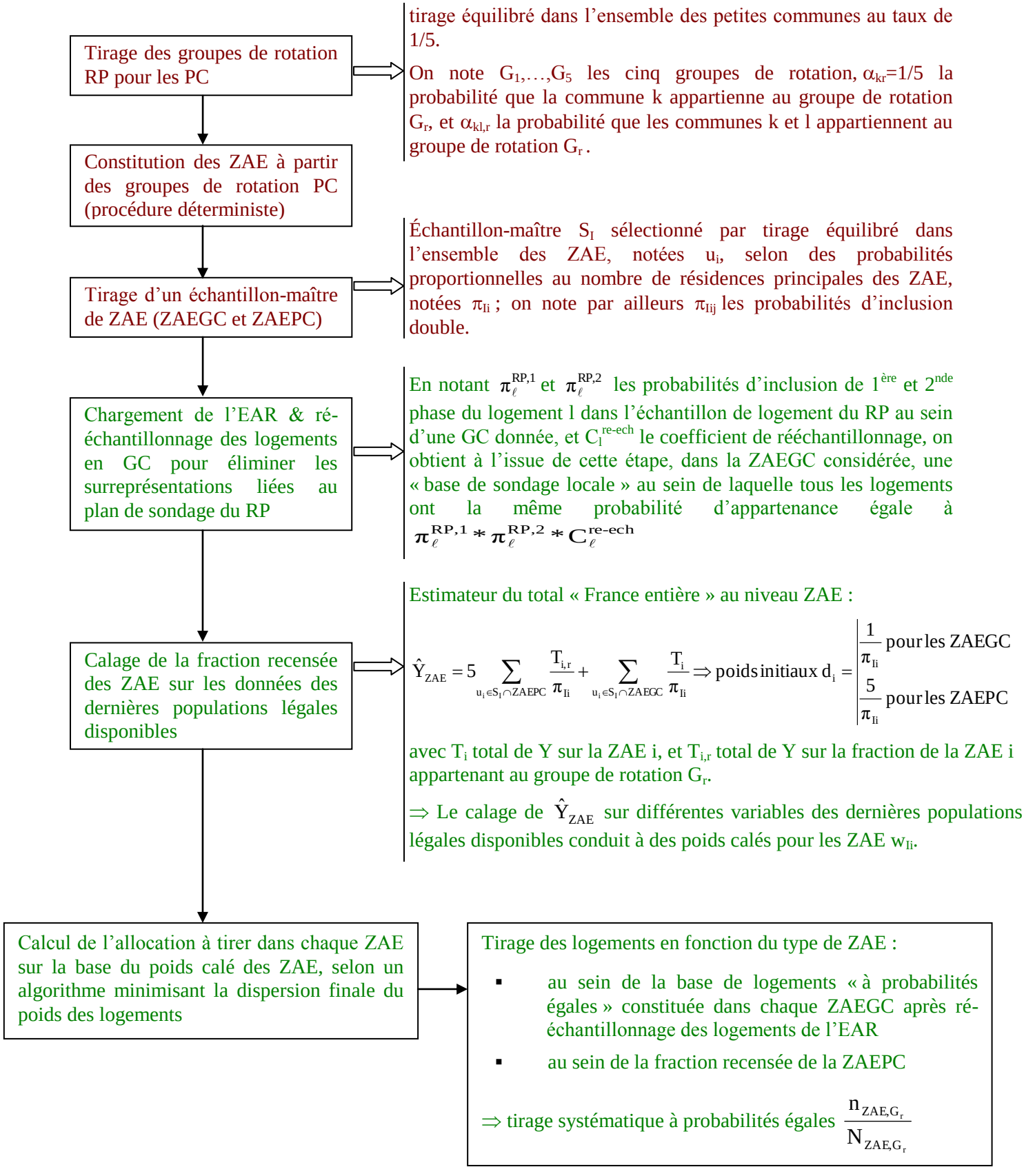
À l'initialisation du système, un échantillon-maître de 525 ZAE – 37 ZAEGC exhaustives, 202 ZAEGC non exhaustives et 286 ZAEPC – a été sélectionné pour la réalisation des enquêtes ménages nationales. Puis chaque année, la base de sondage annuelle d'Octopusse est constituée en chargeant les logements 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 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 au moment du chargement 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 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 encore disponibles (logements qui n'ont pas déjà été sélectionnés pour une enquête).

[2] Les grandes communes, au sens du recensement, sont les communes de 10 000 habitants ou plus.

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'initiation d'Octopusse, en vert italique, procédures réalisées tous les ans pour les tirages d'échantillon de la campagne en cours :



Dans ce contexte, l'estimateur utilisé pour estimer le total d'une variable Y est le suivant[3] :

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

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 des logements de l'échantillon 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[4].

Sous ces hypothèses, une formule de variance analytique « générique » – i.e. valable pour toute enquête standard tirée dans Octopusse – a pu être établie, en s'appuyant sur deux principes centraux :

[3] On ne prend pas en compte à ce stade la non-réponse observée lors de l'enquête, ni un éventuel calage final.

[4] Cette hypothèse revient à dire que l'on procède par « plug-in direct » dans l'estimateur de variance estimant la variance liée au tirage des groupes de rotation du RP, en remplaçant directement les totaux par communes inconnus par leurs estimateurs obtenus à partir de l'échantillon de logements au sein des communes.

- 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 la partie 2.2. ;
- 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 grâce à la formule de Rao. La partie 2.3. explicite cette formule et détaille la façon dont elle est appliquée dans le cadre de l'estimation de variance d'Octopusse.

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 ;
- $\alpha_{kr}=1/5$ la probabilité que la petite commune k appartienne au groupe de rotation G_r, et $\alpha_{kl,r}$ la probabilité que les petites 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 ;
- S_I l'échantillon-maître de ZAE sélectionné 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 ;
- S_r l'ensemble des 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 ;
- 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 $w_{Il}= w_{Ik}=w_{Ii}$ pour tout logement l ∈ commune k ∈ ZAE u_i ;
- enfin, S_l l'échantillon de logements final 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 $\pi_{l\,|\,Gr,\,ui}$.

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

On raisonne ici en supposant connus les totaux de la variable d'intérêt Y par commune – 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[5] suivant :

$$\hat{t}_{yr} = \sum_{k\in S_r} \frac{y_k}{\alpha_{kr}\pi_{Ik}}, \text{ avec } \pi_{Ik} = \pi_{Ii} \text{ pour toute commune k appartenant à une ZAE } u_i$$

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

En suivant l'approche de Guillaume Chauvet dans [3] et en l'adaptant au contexte d'Octopusse[6], la variance de cet estimateur, 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 :

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

avec $\tilde{Y}_{ir} = \sum_{k\in u_i} \frac{y_k \mathbb{1}_{k\in G_r}}{\alpha_{kr}}$ 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].

Des études par simulation ont permis d'une part d'évaluer la qualité de l'estimation des probabilités d'inclusion double[7] 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[8] 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 (*) permettant d'estimer correctement la variance liée aux deux premières phases du plan de sondage d'Octopusse – constitution des groupes de rotation du recensement puis sélection des ZAE de l'échantillon-maître – à partir de l'échantillon de commune 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.

---

[6] Ces adaptations concernent d'une part l'utilisation de l'estimateur de Yates-Grundy généralisé tenant compte du fait que les groupes de rotation du recensement ne sont pas de taille fixe et d'autre part la prise en compte des ZAEGC – 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 $\tilde{Y}_{ir} = y_i$.

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

[8] 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 le calage des ZAE effectué dans Octopusse, le total d'une variable Y s'estime par :

$$\hat{Y} = \sum_{l \in S_l} w_{ll} \frac{y_l}{\pi_{l|G_r,u_i}} = \sum_{u_i \in S_I} w_{Ii} \underbrace{\sum_{l \in S_l \cap G_r \cap u_i} \frac{y_l}{\pi_{l|G_r,u_i}}}_{\hat{y}_i}$$

Pour estimer la variance de cet estimateur, on va s'appuyer sur la formule d'estimation de variance de Rao – que l'on trouve également démontrée dans [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_I de m unités primaires sélectionnées selon des probabilités d'inclusion π_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 $Q(T_1,...,T_m) = \sum_{i \in S_I} q_i T_i^2 + \sum_{(i,j) \in S_I, i \neq j} q_{ij} T_i T_j$ permettant d'estimer sans biais sur l'échantillon S_I 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 la variance de l'estimateur d'Horvitz-Thompson $\hat{T}$ du total T par :

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

Le cadre théorique d'Octopusse ne correspond pas exactement à celui requis pour l'application de la formule de Rao : d'une part la phase de sélection des groupes de rotation du recensement n'est pas indépendante des autres tirages, et d'autre part un calage est effectué sur les ZAE dans Octopusse. Afin de pouvoir cependant procéder à l'estimation de variance en nous appuyant sur la formule de Rao et sur l'estimateur (*) de la variance liée aux deux premières phases du plan de sondage d'Octopusse, nous allons donc avoir recours aux hypothèses simplificatrices ❶ et ❸[9] qui nous ramènent au cadre théorique de la formule de Rao.

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'EM est assimilé à un tirage à probabilités inégales[10], et sa variance estimée à l'aide de la formule proposée par Deville dans [5] page 7 :

[9] Cette hypothèse revient à dire que l'on procède, faute de pouvoir faire mieux, par « plug-in direct » dans l'estimateur de variance de Yates-Grundy généralisé estimant la variance liée au tirage des groupes de rotation du RP, en remplaçant directement les y_k inconnus par les $\hat{y}_k$ estimés à partir de l'échantillon de logement au sein de la commune k.

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

$$\hat{V}(\hat{y}_i) = \frac{n_{ir}}{n_{ir}-1} \sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i}) \left(w_{Il} y_l - \frac{\sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i}) w_{Il} y_l}{\sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i})} \right)^2$$

La forme quadratique Q retenue pour estimer la variance liée aux deux premières phases du plan de sondage d'Octopusse étant (*), on obtient au final la formule de variance suivante :

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

$$+ \sum_{u_i \in S_I} \left(\frac{1}{\pi_{Ii}^2} - q_i \right) \frac{n_{ir}}{n_{ir}-1} \sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i}) \left(w_{Il} y_l - \frac{\sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i}) w_{Il} y_l}{\sum_{l \in S_l \cap G_r \cap u_i} (1-\pi_{l|G_r,u_i})} \right)^2$$

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

En pratique, pour prendre en compte la correction de la non-réponse par repondération et le calage sur marges usuellement mis en œuvre dans les enquêtes à l'Insee, on procède comme suit :

- la non-réponse est modélisée par un mécanisme poissonnien. Dans ce cadre, 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 une probabilité de tirage correspondant à la probabilité de réponse du logement, et tirages indépendants entre les logements – et donc appliquer à nouveau la formule de Rao pour la prendre en compte ;
- le calage de l'estimateur 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, sur les variables de calage.

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] Jay Breidt F. 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., Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE, *document de travail Insee M9806*, 1998, Insee.