CALCULS DE PRÉCISION DANS OCTOPUSSE : THÉORIE ET APPLICATION À L'ENQUÊTE LOGEMENT 2013

Emmanuel GROS[1] (*), Martin CHEVALIER[2] (*), Karim MOUSSALLAM[3] (**)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**) Dares, Sous-direction Salaires, travail et relations professionnelles

Résumé

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse et présenté dans [1], 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.

En 2011, Guillaume Chauvet s'est intéressé dans [2] au problème des calculs de précision dans le contexte d'Octopusse, et a proposé, dans un cadre légèrement simplifié[4], une méthode d'estimation de variance fondée 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. En s'appuyant sur les programmes fournis par Guillaume Chauvet, l'Insee a prolongé ces travaux de façon à élaborer une formule de variance analytique « générique » valable dans le contexte réel d'Octopusse. Ces travaux complémentaires ont consisté :

- d'une part à calculer 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 – du système Octopusse, à analyser leur qualité ainsi que celle des estimateurs de Yates-Grundy les utilisant.
- 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.

Ce cadre méthodologique a été appliqué à l'enquête Logement 2013, et a été adapté aux spécificités de son plan de sondage : tirage dans l'EMEX élargi pour la région Nord-Pas-de-Calais et directement dans l'Enquête annuelle de recensement 2011 pour la région Île-de-France, constitution d'un échantillon de logements neufs (construits entre mars 2011 et mars 2013) par tirage dans la base de permis de construire DGI-SITADEL, constitution d'un échantillon de logements situés en Zone urbaine sensible (ZUS). La mise en œuvre de cette méthodologie a permis de mesurer l'impact des différentes composantes du plan de sondage sur la précision des estimations, en particulier le tirage des logements dans la dernière Enquête annuelle de recensement et les extensions régionales.

Du point de vue pratique, cet investissement méthodologique s'est traduit par la mise à disposition des partenaires en charge de l'exploitation des données de l'enquête d'un programme de calcul de précision spécifique à l'enquête Logement 2013. Ce programme est structuré de façon à ce que les modules

[1] emmanuel.gros@insee.fr
[2] martin.chevalier@insee.fr
[3] karim.moussallam@travail.gouv.fr
[4] 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.

correspondant aux différentes composantes du calcul de variance soient aisément réutilisables dans le cadre d'autres enquêtes échantillonnées avec le système Octopusse.

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 provides a recent sampling frame. The interaction between these two concepts leads to an efficient but singularly complex sampling procedure, which makes variance estimation particularly difficult.
This paper draws the methodological framework of the sampling system Octopusse and presents the different variance formulas and their justification as well as the results obtained for a first application of this variance estimation method to the Housing survey in 2013.

Mots-clés

Estimation de variance, échantillon-maître Octopusse, sondages à plusieurs phases/degrés, estimation de variance de Yates-Grundy, enquête Logement 2013.

Introduction

Le système actuel d'échantillonnage des enquêtes ménages à l'Insee, baptisé Octopusse et présenté dans [1], 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.

En 2011, Guillaume Chauvet s'est intéressé dans [2] au problème des calculs de précision dans le contexte d'Octopusse, et a proposé, dans un cadre légèrement simplifié[5], une méthode d'estimation de variance fondée 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. En s'appuyant sur les programmes fournis par Guillaume Chauvet, l'Insee a prolongé ces travaux de façon à élaborer une formule de variance analytique « générique » valable dans le contexte réel d'Octopusse. Ces travaux complémentaires ont consisté :

- d'une part à calculer 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 – du système Octopusse, à analyser leur qualité ainsi que celle des estimateurs de Yates-Grundy les utilisant.
- 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.

Ce cadre méthodologique a été appliqué à l'enquête Logement 2013, et a été adapté aux spécificités de son plan de sondage : tirage dans l'EMEX élargi pour la région Nord-Pas-de-Calais et directement dans

[5] 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.

l'Enquête annuelle de recensement 2011 pour la région Île-de-France, constitution d'un échantillon de logements neufs (construits entre mars 2011 et mars 2013) par tirage dans la base de permis de construire DGI-SITADEL, constitution d'un échantillon de logements situés en Zone urbaine sensible (ZUS). La mise en œuvre de cette méthodologie a permis de mesurer l'impact des différentes composantes du plan de sondage sur la précision des estimations, en particulier le tirage des logements dans la dernière Enquête annuelle de recensement et les extensions régionales.

Du point de vue pratique, cet investissement méthodologique s'est traduit par la mise à disposition des partenaires en charge de l'exploitation des données de l'enquête d'un programme de calcul de précision spécifique à l'enquête Logement 2013. Ce programme est structuré de façon à ce que les modules correspondant aux différentes composantes du calcul de variance soient aisément réutilisables dans le cadre d'autres enquêtes échantillonnées avec le système Octopusse.

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[6] 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 ;
- 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[7] 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 :

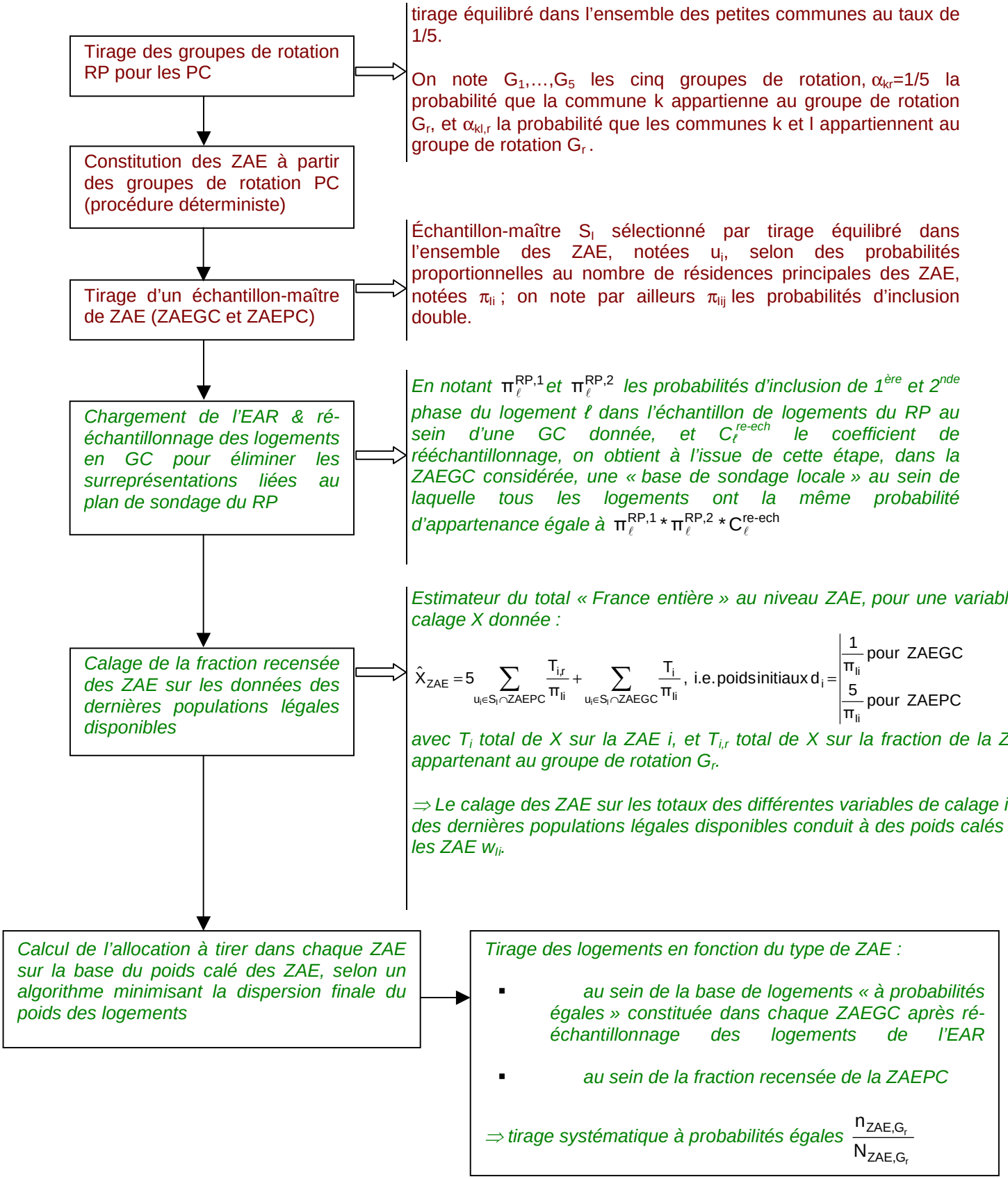
- d'une part, l'Enquête Annuelle de Recensement surreprésente certaines strates de logements en grandes communes : logements des grandes adresses et des adresses neuves. Afin d'éliminer ces surreprésentations, une procédure de rééchantillonnage – qui consiste à ne conserver, par tirage aléatoire, qu'une 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 ;

[6] Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Échantillons.
[7] Les grandes communes, au sens du recensement, sont les communes de 10 000 habitants ou plus.

- 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 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 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'initiation 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[8] :

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

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 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[9].

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 :

- 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

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

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[10], 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_l 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_l 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_{li} ; on note par ailleurs π_{lij} la probabilité (toujours conditionnelle) d'inclusion double des unités u_i et u_j au sein de S_l ; par convention, on notera $\pi_{lk}=\pi_{li}$ pour toute commune k appartenant à la ZAE u_i, et $\pi_{l\ell} = \pi_{lk} = \pi_{li}$ 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_l 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_{li} le poids calé de la ZAE u_i restreinte à la dernière EAR, $w_{lk}=w_{li}$ pour toute commune k appartenant à la ZAE u_i et $w_{l\ell} = w_{lk} = w_{li}$ 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 : $w_\ell = w_{l\ell} / \pi_{\ell|G_r,u_i}$;
- 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.

[10] 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.

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

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 :

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

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

Des études par simulation ont permis d'une part d'évaluer la qualité de l'estimation des probabilités d'inclusion double[12] 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[13] 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 $Q_{GR}(y_1,...,y_{n_c}) = \hat{V}_{GR}^{S_r}$; d'autre part la variance liée à la sélection des ZAE de l'échantillon-maître estimée par la forme quadratique $Q_{EM}(y_1,...,y_{n_c}) = \hat{V}_{EM}^{S_r}$. 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.

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 :

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

[13] 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.

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

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

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

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

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

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

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[14], 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 :

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

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

- 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 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 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 $S_{UP}=S_I$, $S_{US}=S_\ell$, $Q=Q_{EM}$, $T_i = \frac{\tilde{Y}_{ir}}{\pi_{Ii}} = \sum_{k \in u_i} \frac{y_k \mathbb{1}_{k \in G_r}}{\alpha_{kr}\pi_{Ii}}$ et

$\hat{T}_i = \sum_{k \in u_i} \frac{\sum_{\ell \in S_\ell \cap k} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}{\alpha_{kr}\pi_{Ii}} = \frac{1}{\alpha_{ir}\pi_{Ii}} \underbrace{\sum_{\ell \in S_\ell \cap u_i} \frac{y_\ell}{\pi_{\ell|G_r,u_i}}}_{\hat{y}_{u_i \cap G_r}} = \frac{\hat{y}_{u_i \cap G_r}}{\alpha_{ir}\pi_{Ii}}$. En notant q_i^{EM} le coefficient diagonal[15] 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 :

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

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

- 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_I, 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 :

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

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 :

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

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

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

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, que l'on trouve 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_{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 $\hat{Y}^{R}_{calé}$ s'estime donc à partir de l'échantillon de répondants R par :

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

3. Application à l'enquête Logement 2013

3.1. Plan de sondage et post-traitements de l'enquête Logement 2013 en métropole

Le plan de sondage de l'enquête Logement distingue quatre sous-échantillons en métropole[16] : l'échantillon principal, l'échantillon de réserve, l'échantillon de logements neufs et l'échantillon tiré spécifiquement dans les zones urbaines sensibles (ZUS). Le plan de sondage et le calcul des poids de tirage de l'enquête Logement 2013 en métropole sont décrits en détails dans [7].

Échantillon principal. La base de sondage de l'échantillon principal de l'enquête Logement 2013 est l'enquête annuelle recensement (EAR) 2011. Il est donc affecté par la constitution des groupes de rotation du recensement. Les unités primaires dans lesquelles ces logements sont tirés varient selon les régions :

- Dans les régions métropolitaines hors extensions, les logements sont tirés dans les unités primaires de l'Échantillon-maître (EM). Au sein de chaque unité primaire, les logements à enquêter sont sélectionnés par un tirage systématique sur fichier trié par statut d'occupation du logement, période d'achèvement et type d'habitation.
- En région Nord-Pas-de-Calais, les logements sont tirés dans les unités primaires de l'Échantillon-maître pour les extensions régionales élargi (EMEX-E), qui sont trois fois plus nombreuses que celles de l'EM. Le tirage des logements dans les unités primaires est effectué comme dans les régions métropolitaines hors extensions.
- En région Île-de-France, les logements sont tirés directement dans l'EAR 2011 sans passer par un système d'échantillon-maître : il s'agit d'un tirage à un seul degré stratifié par département, type d'habitation (individuel *versus* collectif), statut d'occupation (propriétaire *versus* non-propriétaire), taille de commune (grandes communes – plus de 10 000 habitants – *versus* petites communes). Associé à une taille d'échantillon beaucoup plus importante, ce tirage à un degré stratifié a pour objectif de garantir la représentativité infrarégionale de l'enquête Logement 2013 en Île-de-France.

Échantillon de réserve. Afin de pallier un éventuel épuisement du nombre de logements échantillonnés en cours de collecte, à l'échantillon principal est associé un échantillon de réserve. Son déclenchement n'est pas automatique mais dépend de l'avancement et du taux de réussite de l'enquête sur le terrain.

- Dans les régions métropolitaines hors Île-de-France, l'échantillon de réserve est tiré selon les mêmes modalités que l'échantillon principal : tirage dans l'EAR 2011 et dans les unités primaires de l'EM pour les régions hors extensions et de l'EMEX-E en région Nord-Pas-de-Calais.
- En région Île-de-France, la réserve est tirée dans l'EMEX-E : il s'agit donc d'un tirage à deux degrés.

Dans tous les cas les poids de tirage tiennent compte des différents scénarios de déclenchement de la réserve et du partage des poids à effectuer avec le sous-échantillon ZUS (*cf. infra*).

[16] Le tirage et les post-traitements de l'enquête Logement 2013 dans les départements d'outre-mer est assuré par le Centre de ressources interrégional des enquêtes ménages (CRIEM).

Échantillon de logements neufs. L'EAR 2011 comporte uniquement les logements achevés avant mars 2011. Les logements construits entre mars 2011 et l'arrivée de l'enquête Logement 2013 sur le terrain en juin 2013 ne peuvent donc pas être échantillonnés en utilisant la base de sondage « classique » des enquêtes ménages de l'INSEE.
Une base de sondage spécifique est utilisée pour sélectionner un échantillon de logements dits « neufs » : il s'agit d'un appariement entre la base Sitadel de permis de construire du Service de l'observation et des statistiques (SOeS) et la liste des locaux achevés transmise mensuellement par la Direction générale des finances publiques (anciennement Direction générale des impôts, DGI), sur la période mars 2011-mars 2013[17].
La base de sondage DGI-Sitadel permet d'atteindre les logements dans le cadre d'un sondage à deux degrés : tirage d'un permis de construire puis, au sein de ce permis, d'un logement à enquêter. La localisation des permis et donc des logements échantillonnés varie selon les régions :

- Dans les régions métropolitaines hors Île-de-France, l'échantillon de permis est tiré dans les mêmes unités primaires que l'échantillon principal : celles de l'EM pour les régions hors extensions et celles de l'EMEX-E pour la région Nord-Pas-de-Calais. On a donc affaire à un tirage à trois degrés (unité primaire, permis, logement).
- Dans la région Île-de-France, les permis sont directement tirés dans la base de sondage après stratification par taille de communes (grandes communes *versus* petites communes). On a donc affaire à un tirage à deux degrés (permis, logement).

À noter que dans tous les cas, ce tirage n'est pas affecté par les groupes de rotation du recensement.

Échantillon spécifique ZUS. Un dernier sous-échantillon de logements est tiré spécifiquement dans les zones urbaines sensibles. La base de sondage est constituée par les cinq campagnes de recensement qui précèdent l'EAR 2011 (EAR 2006 à 2010). Il s'agit d'un tirage de logements à un degré, stratifié par année de recensement, taille de commune (grandes communes *versus* petites communes) et, dans les grandes communes, type d'adresse (adresses neuves[18], grandes adresses ou autres adresses). Pour tenir compte du fait que certains logements échantillonnés dans ce sous-échantillon auraient également pu être atteints par le sous-échantillon principal ou la réserve, une méthode de partage des poids est mise en œuvre au niveau national.

Le Tableau 1 synthétise l'ensemble des caractéristiques du plan de sondage de l'enquête, en indiquant pour chaque configuration le nombre de logements échantillonnés.

Tableau 1 : Représentation synthétique du plan de sondage de l'enquête Logement 2013

<table>
<tr><th>Sous-échantillons</th><th>Métropole hors extensions</th><th>Nord-Pas-de-Calais</th><th>Île-de-France</th></tr>
<tr><td>Principal</td><td rowspan="2">2 degrés, GR
20 129 + 2 827</td><td rowspan="2">2 degrés, GR
4 676 + 657</td><td>1 degré, stratifié, GR
10 795</td></tr>
<tr><td>Réserve</td><td>2 degrés, GR
1 516</td></tr>
<tr><td>Neufs</td><td>3 degrés, pas de GR
486</td><td>3 degrés, pas de GR
109</td><td>2 degrés, pas de GR
255</td></tr>
<tr><td>ZUS</td><td colspan="3">1 degré, stratifié, pas de GR
6 000</td></tr>
</table>

Note : GR : Groupe de rotation
Lecture : Pour les régions métropolitaines hors extension, le tirage de l'échantillon principal suit un plan de sondage à deux degrés dans les groupes de rotation du recensement de la population. 20 129 logements sont échantillonnés au titre de l'échantillon principal et 2 827 au titre de la réserve.

[17] En raison des délais de transmission des données et de préparation de l'enquête, les informations sur la période avril 2013-mai 2013 n'ont pas pu être exploitées.
[18] Il s'agit d'adresses neuves au sens du recensement, dont les logements sont achevés au moment de l'EAR. Ces « adresses neuves » sont donc totalement distinctes des « logements neufs » du sous-échantillon tiré dans la base DGI-Sitadel.

Collecte et post-traitements. Sur les 47 450 logements échantillonnés, 42 594 ont été effectivement intégrés à la collecte : la réserve n'a en effet été déclenchée qu'en Île-de-France, et très partiellement (144 logements). À ces 42 594 fiches-adresses mises en collecte correspondent 27 158 questionnaires validés.
Pour les sous-échantillons tirés dans une ou plusieurs EAR (principal, réserve et ZUS), une correction de la non-réponse totale par groupes de réponse homogène est mise en œuvre séparément dans les trois zones géographiques définies par les extensions régionales (France métropolitaine hors extensions, Nord-Pas-de-Calais, Île-de-France). Les principales variables qui interviennent dans cette correction sont la taille du logement, le statut d'occupation, le fait que le logement soit HLM, la localisation (tranche d'unité urbaine, ZUS), et la date d'achèvement.
Pour le sous-échantillon de logements neufs et faute de variables pertinentes dans la base de sondage, la correction de la non-réponse totale prend la forme d'un calage préliminaire sur le nombre total de logements à différents niveaux géographiques (ZEAT pour les régions métropolitaines hors extensions, département pour le Nord-Pas-de-Calais et regroupement de départements pour l'Île-de-France).
En dernier lieu, un calage sur marges est réalisé, en respectant là encore les zones géographiques délimitées par les extensions régionales (France métropolitaine hors extensions, Nord-Pas-de-Calais, Île-de-France). Les variables de calage sont la surface et le nombre de pièces, le type de logement et son appartenance au secteur social, le statut d'occupation et le nombre de personnes dans le logement, la date d'achèvement de la construction et le mode de chauffage, la tranche d'unité urbaine. En région Île-de-France, le type de logement, le statut d'occupation, la date d'achèvement et la tranche d'unité urbaine sont introduites au niveau départemental quand cela est techniquement possible. Dans les régions métropolitaines hors extensions, le nombre total de résidences principales par région est introduit comme marge supplémentaire[19].

3.2. Adaptation de la méthodologie de calcul de précision à l'enquête Logement 2013

Échantillon principal. Le cadre général décrit dans la partie 2 de cet article est directement applicable au calcul de précision dans l'échantillon principal de l'enquête Logement 2013 :

- Dans les régions métropolitaines hors extensions, on se trouve dans la situation « classique » d'un tirage dans la dernière enquête annuelle de recensement disponible et dans les unités primaires de l'EM.
- En région Nord-Pas-de-Calais, la seule différence avec la situation décrite précédemment est que les probabilités d'inclusion simple et double des unités primaires à utiliser dans le calcul de variance sont celles de l'EMEX-E et non celles de l'EM. Comme pour les probabilités d'inclusion double de l'EM, les probabilités d'inclusion double de l'EMEX-E ont été calculées par réplication en exploitant les propriétés de martingale de l'algorithme du Cube (comme proposé dans [4]). À noter que les probabilités d'inclusion simple de l'EMEX-E interviennent à la fois dans le calcul du terme de variance relatif à la sélection des unités primaires et dans celui qui correspond au tirage des groupes de rotation du recensement. De même, le coefficient q_i associé à l'utilisation de la formule de Rao doit pour la région Nord-Pas-de-Calais faire intervenir les probabilités d'inclusion simple et double de l'EMEX-E, et non de l'EM.
- En région Île-de-France, le tirage direct dans l'enquête annuelle de recensement sans passer par un système d'échantillon-maître simplifie le calcul de variance : aucun terme relatif à la sélection des unités primaires n'a à être pris en compte ; le terme de variance relatif au tirage des groupes de rotation du recensement ne fait pas intervenir les probabilités d'inclusion des unités primaires ; en l'absence d'échantillonnage d'unités primaires le coefficient q_i est nul. L'échantillonnage étant stratifié, toutes les quantités sont calculées au sein de chaque strate de tirage puis additionnées.

Échantillon de réserve. La première simplification par rapport au plan de sondage de l'enquête concerne l'échantillon de réserve en Île-de-France (la réserve n'a pas été déclenchée dans les autres régions). Cette réserve étant tirée dans l'EMEX-E, le calcul de précision associé doit théoriquement être analogue à celui mis en œuvre dans la région Nord-Pas-Calais. Ici cependant, la réserve n'a été déclenchée que partiellement et dans quelques unités primaires seulement (20 ZAE concernées sur 167). Le calcul du

[19] Les post-traitements mis en œuvre sont détaillés dans la documentation de l'enquête Logement 2013.

terme de variance relatif au tirage dans l'EMEX-E n'est donc pas envisageable. Dans ce cadre, les 144 logements de l'échantillon de réserve (donc 92 répondants) sont pris en compte comme les autres logements de la région Île-de-France, c'est-à-dire comme s'ils avaient été sélectionnés par sondage à un seul degré stratifié[20] dans l'EAR 2011.

Échantillon de logements neufs. Le faible nombre de logements échantillonnés (850) et de questionnaires validés (467) pour l'échantillon de logements neufs induit plusieurs difficultés.
D'une part, pour les régions métropolitaines hors Île-de-France, le tirage des permis est effectué dans les unités primaires de l'EM pour les régions hors extensions et de l'EMEX-E pour la région Nord-Pas-de-Calais. Cependant, la taille particulièrement faible de l'échantillon conduit dans la plupart des cas (92,6 %) à au plus un seul logement répondant par ZAE ; plus encore, 37,3 % des ZAE ne présentent aucun répondant. Dans ce contexte, faute de pouvoir calculer correctement les termes de variance intra-ZAE (moins de deux logements répondants par ZAE) et inter-ZAE (de nombreuses ZAE sans aucun répondant), on assimile le tirage des permis dans les régions métropolitaines hors Île-de-France à un sondage à un degré stratifié par région[21]. L'ampleur de la sous-estimation de la variance associée à cette simplification devrait être limitée dans la mesure où les allocations par unité primaire particulièrement faibles atténuent l'impact de la corrélation spatiale intra-ZAE.
D'autre part, dans toutes les régions métropolitaines y compris l'Île-de-France, on rencontre une difficulté analogue pour la mesure de la variance associée à l'échantillonnage des logements neufs *via* leur permis de construire. Un seul logement ayant été tiré par permis échantillonné, il n'est pas possible de calculer une variance intra-permis (moins de deux répondants par permis) ; par ailleurs 45,1 % des permis ne présentent aucun logement répondant. Dans ces conditions, on assimile là encore ce tirage à deux degrés à un tirage à un seul degré.
Ces éléments conduisent ainsi à modéliser la variance associée au tirage des logements neufs comme la variance d'un tirage à un degré stratifié par région (et selon la taille des communes en Île-de-France).

Échantillon spécifique ZUS. Le plan de sondage de l'échantillon spécifique ZUS ne présente pas de difficulté majeure pour le calcul de précision : le tirage étant à un degré stratifié dans cinq campagnes de l'EAR, on applique la formule proposée par Deville [3] au sein de chaque strate sans prendre en compte les groupes de rotation du recensement. L'intégration du partage des poids associé au tirage de ce sous-échantillon dans le calcul de variance est transparente : dans la mesure où les échantillons après partage des poids peuvent être considérés comme indépendants, la variance de l'échantillon complet peut être estimée en agrégeant les termes de variance relatifs aux différents échantillons pour autant qu'on introduise le coefficient de partage des poids dans les formules présentées précédemment.

Le Tableau 2 synthétise le plan de sondage de l'enquête Logement 2013 tel qu'il est adapté à la problématique du calcul de variance. Les simplifications pour la réserve en Île-de-France et pour le sous-échantillon de logements neufs n'affectent qu'une faible fraction des logements répondants (559 sur 27 158), ce qui en limite fortement l'impact sur les estimations. On aboutit ainsi à quatre configurations distinctes, qui constituent le cadre général du calcul de variance dans l'enquête Logement 2013.

[20] Les informations nécessaires au classement des logements de la réserve dans les strates de l'échantillon principal en Île-de-France n'étant pas disponibles, dans le calcul de précision ce sous-échantillon est stratifié uniquement par taille de commune (grandes communes *versus* petites communes).
[21] Cette stratification est adoptée car le tirage des unités primaires de l'EM et de l'EMEX-E est stratifié par région. La région Corse ne présentant qu'un seul questionnaire validé dans ce sous-échantillon (sur 3 logements échantillonnés), elle est rapprochée de la région Provence-Alpes-Côte-D'azur pour cette stratification.

Tableau 2 : Adaptation du plan de sondage de l'enquête Logement 2013 pour le calcul de variance

<table>
<tr><th></th><th>Métropole hors extensions</th><th>Nord-Pas-de-Calais</th><th>Île-de-France</th></tr>
<tr><td>Principal</td><td>2 degrés, GR
$V_{GR} + V_{UP}^{EM} + V_{LOG}$</td><td>2 degrés, GR
$V_{GR} + V_{UP}^{EMEX-E} + V_{LOG}$</td><td rowspan="2">1 degré, stratifié, GR
$V_{GR} + V_{LOG}$</td></tr>
<tr><td>Réserve</td><td colspan="2">Réserve non-déclenchée</td></tr>
<tr><td>Neufs</td><td colspan="3" rowspan="2">1 degré, stratifié, pas de GR
V_{LOG}</td></tr>
<tr><td>ZUS</td></tr>
</table>

Note : GR : Groupe de rotation. Dans la configuration correspondant aux sous-échantillons de logements neufs ou spécifique ZUS, la stratification diffère d'un sous-échantillon à l'autre.
Lecture : Pour les régions métropolitaines hors extension, le tirage de l'échantillon principal est pris en compte dans le calcul de variance comme un tirage à deux degrés dans les groupes de rotation du recensement de la population. Les termes de variance associés à ce plan de sondage sont donc respectivement la variance due au tirage des groupes de rotation du recensement, la variance due au tirage des unités primaires de l'EM et la variance due au tirage des logements au sein de l'intersection entre groupe de rotation et unité primaire de l'EM.

3.3. Le calcul de précision dans l'enquête Logement 2013 étape par étape

Cette partie reprend l'ensemble des éléments présentés précédemment pour décrire, étape par étape, la méthode de calcul de précision mise en œuvre pour l'enquête Logement 2013. Aux notations introduites à la partie 2.1 s'ajoutent :

- c_ℓ le coefficient individuel de partage des poids du logement ℓ : en suivant [7] p. 12, c_ℓ vaut 1 pour le sous-échantillon de logements neufs ainsi que pour les logements de l'échantillon principal qui ne sont pas situés en ZUS, 0,281 pour les logements de l'échantillon principal situés en ZUS et 0,719 pour les logements du sous-échantillon ZUS ;
- $\hat{p}_\ell$ la probabilité de réponse issue de la procédure de correction de la non-réponse pour le logement ℓ (introduite à la partie 2.4) ;
- ε_ℓ la valeur pour le logement ℓ du résidu estimé par la régression de la variable d'intérêt y sur les variables de calage (introduit à la partie 2.4).

Prise en compte du calage. Le calage sur marges de l'enquête Logement 2013 est pris en compte comme proposé à la partie 2.4 : dans toutes les formules, la variable d'intérêt y est remplacée par son résidu ε estimé par la régression sur les variables de calage. Le calage ayant été mené séparément dans les trois zones délimitées par les extensions régionales (régions métropolitaines hors extensions, région Nord-Pas-de-Calais, région Île-de-France), un modèle différent est estimé pour chacune de ces trois zones.

Variance associée au tirage des groupes de rotation du recensement. Trois des quatre configurations distinguées pour le calcul de précision font intervenir les groupes de rotation du recensement. Dans les trois cas, le terme de variance estimé sur les logements répondants à l'enquête est de la forme :

$$\hat{V}_{GR}(\hat{Y}^{R}_{Calé}) = -\frac{1}{2}\sum_{\substack{k,l\in S_r \\ k\neq l}} \frac{\hat{\alpha}_{kl,r} - \alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}\ \underbrace{\hat{\pi}_{lij}}_{(i,j)\,/\,\substack{k\in u_i \\ l\in u_j}}} \left(\sum_{\ell\in S_\ell\cap k} \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\alpha_{kr}\pi_{\ell|G_r,u_i}\hat{p}_\ell} - \sum_{\ell\in S_\ell\cap l} \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\alpha_{lr}\pi_{\ell|G_l,u_i}\hat{p}_\ell} \right)^2$$

$$+ \sum_{k\in S_r} \frac{\sum_{l\in U}\hat{\alpha}_{kl,r} - \alpha_{kr}\sum_{l\in U}\alpha_{l,r}}{\alpha_{kr}\ \underbrace{\pi_{li}}_{i\,/\,k\in u_i}} \left(\sum_{\ell\in S_\ell\cap k} \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\alpha_{kr}\pi_{\ell|G_r,u_i}\hat{p}_\ell} \right)^2$$

Ce terme est nul pour toutes les grandes communes, dans la mesure où pour une grande commune k, $\alpha_{kr} = 1$ et $\hat{\alpha}_{kl,r} = \alpha_{lr}$. En pratique il est donc estimé uniquement pour les petites communes. La seule différence entre les trois configurations est la valeur des probabilités d'inclusion simple π_{li} et double $\hat{\pi}_{lij}$ des unités primaires :

- Pour les régions métropolitaines hors extensions, ce sont les probabilités de l'EM :

$$\pi_{li} = \pi^{EM}_{li} \text{ et } \hat{\pi}_{lij} = \hat{\pi}^{EM}_{lij}$$

- Pour la région Nord-Pas-de-Calais, ce sont les probabilités de l'EMEX-E :

$$\pi_{li} = \pi^{EMEX\text{-}E}_{li} \text{ et } \hat{\pi}_{lij} = \hat{\pi}^{EMEX\text{-}E}_{lij}$$

- Pour la région Île-de-France, dans la mesure où le tirage a eu lieu directement dans l'EAR, ces probabilités d'inclusion sont constantes et égales à 1 :

$$\pi_{li} = 1 \text{ et } \hat{\pi}_{lij} = \pi_{lij} = 1$$

Variance associée au tirage des unités primaires de l'échantillon-maître. Deux des quatre configurations distinguées pour le calcul de précision font intervenir les unités primaires de l'EM ou de l'EMEX-E. Dans les deux cas, le terme de variance estimé sur les logements répondants est de la forme :

$$\hat{V}_{UP}(\hat{Y}^{R}_{Calé}) = -\frac{1}{2}\sum_{\substack{u_i,u_j\in S_l \\ u_i\neq u_j}} \frac{\hat{\pi}_{lij} - \pi_{li}\pi_{lj}}{\hat{\pi}_{lij}} \left(\sum_{\ell\in S_\ell\cap u_i} \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\alpha_{ir}\pi_{li}\pi_{\ell|G_r,u_i}\hat{p}_\ell} - \sum_{\ell\in S_\ell\cap u_j} \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\alpha_{jr}\pi_{lj}\pi_{\ell|G_r,u_j}\hat{p}_\ell} \right)^2$$

Ce terme est nul pour toutes les ZAE exhaustives, dans la mesure où pour une ZAE exhaustive u_i, $\pi_{li} = 1$ et $\hat{\pi}_{lij} = \pi_{lj}$. En pratique il est donc estimé uniquement pour les ZAE non-exhaustives. Là encore, la seule différence entre les régions hors extensions d'une part et la région Nord-Pas-de-Calais d'autre part est la valeur des probabilités d'inclusion simple et double des unités primaires.

Variance associée au tirage des logements. Le terme de variance associé au tirage des logements varie sensiblement selon que le plan de sondage fait intervenir deux degrés (régions métropolitaines hors extensions, région Nord-Pas-de-Calais) ou un degré (région Île-de-France, sous-échantillons de logements neufs ou spécifique ZUS).
Dans les régions métropolitaines hors extensions et la région Nord-Pas-de-Calais, le terme de variance estimé sur les logements répondants est de la forme :

$$\hat{V}_{LOG}(\hat{Y}^{R}_{Calé}) = \sum_{u_i\in S_l} \frac{1-q_i}{(\alpha_{ir}\pi_{li})^2} \frac{n_{ir}}{n_{ir}-1} \sum_{\ell\in S_\ell\cap u_i} (1-\pi_{\ell|G_r,u_i}) \left(\frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\pi_{\ell|G_r,u_i}\hat{p}_\ell} - \frac{\sum_{\ell\in S_\ell\cap u_i}(1-\pi_{\ell|G_r,u_i})\dfrac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell\in R}}{\pi_{\ell|G_r,u_i}\hat{p}_\ell}}{\sum_{\ell\in S_\ell\cap u_i}(1-\pi_{\ell|G_r,u_i})} \right)^2$$

La seule différence entre les régions métropolitaines hors extensions et la région Nord-Pas-de-Calais provient une fois encore des probabilités d'inclusion associées au tirage respectivement dans l'EM et dans l'EMEX-E, qui affectent à la fois π_{li} et $q_i = -\sum_{\substack{u_j \in S_l \\ u_j \neq u_i}} \frac{\hat{\pi}_{lij} - \pi_{li}\pi_{lj}}{\hat{\pi}_{lij}}$.

Dans la région Île-de-France et pour les sous-échantillons de logements neufs ou spécifique aux ZUS, ce terme est de la forme :

$$\hat{V}_{LOG}(\hat{Y}^R_{Calé}) = \sum_{h=1}^{H} \frac{1}{\alpha_h^2} \frac{n_h}{n_h - 1} \sum_{\ell \in S_h} \left(1 - \pi_{\ell|G_r}\right) \left(\frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell \in R}}{\pi_{\ell|G_r} \hat{p}_\ell} - \frac{\sum_{\ell \in S_h} (1 - \pi_{\ell|G_r}) \frac{c_\ell \varepsilon_\ell \mathbb{1}_{\ell \in R}}{\pi_{\ell|G_r} \hat{p}_\ell}}{\sum_{\ell \in S_h} (1 - \pi_{\ell|G_r})} \right)^2$$

où $S_1,...,S_H$ désignent les strates de tirage (spécifiques à chaque sous-échantillon, *cf. supra*). Le conditionnement par le groupe de rotation du recensement et le coefficient α_h ne font sens que pour l'échantillon principal ou la réserve en Île-de-France (tirage dans l'EAR). Dans ce cas, les strates de tirage respectent la répartition entre petites et grandes communes si bien que α_h peut être défini pour tous les logements d'une même strate. Pour les sous-échantillons de logements neufs ou spécifique ZUS, $\pi_{\ell|G_r} = \pi_\ell$ et $\alpha_h = 1$.

Variance associée à la non-réponse. La variance associée à la non-réponse est prise en compte comme proposé à la partie 2.4, c'est-à-dire par un terme de la forme :

$$\hat{V}_{NR}(\hat{Y}^R_{Calé}) = \sum_{\ell \in S} (w_\ell^2 - q_\ell^L)(1 - \hat{p}_\ell) \left(\frac{\varepsilon_\ell \mathbb{1}_{\ell \in R}}{\hat{p}_\ell} \right)^2$$

où w_ℓ et q_ℓ^L désignent respectivement le poids de tirage et le terme diagonal de la forme quadratique estimant la variance d'échantillonnage associés au logement ℓ. Les valeurs de w_ℓ et q_ℓ^L varient selon les configurations :

- Pour les régions métropolitaines hors extensions, $w_\ell = \frac{c_\ell}{\alpha_{ir} \pi_{li}^{EM} \pi_{\ell|G_r, u_i}}$ et

$$q_\ell^L = -\sum_{\substack{l \in S_r \\ l \neq k}} \frac{\hat{\alpha}_{kl,r} - \alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r} \underbrace{\hat{\pi}_{lij}^{EM}}_{j/l \in u_j}} \left(\underbrace{\pi_{lj}^{EM}}_{j/l \in u_j} \right)^2 + \frac{\sum_{k \in U} \hat{\alpha}_{kl,r} - \alpha_{lr} \sum_{k \in U} \alpha_{k,r}}{\alpha_{lr}} \pi_{li}^{EM} + q_i^{EM}$$

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

- En région Nord-Pas-de-Calais, $w_\ell = \frac{c_\ell}{\alpha_{ir} \pi_{li}^{EMEX\text{-}E} \pi_{\ell|G_r, u_i}}$ et

$$q_\ell^L = -\sum_{\substack{l \in S_r \\ l \neq k}} \frac{\hat{\alpha}_{kl,r} - \alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r} \underbrace{\hat{\pi}_{lij}^{EMEX\text{-}E}}_{j/l \in u_j}} \left(\underbrace{\pi_{lj}^{EMEX\text{-}E}}_{j/l \in u_j} \right)^2 + \frac{\sum_{k \in U} \hat{\alpha}_{kl,r} - \alpha_{lr} \sum_{k \in U} \alpha_{k,r}}{\alpha_{lr}} \pi_{li}^{EMEX\text{-}E} + q_i^{EMEX\text{-}E}$$

$$+ \frac{1 - q_i^{EMEX\text{-}E}}{(\alpha_{ir} \pi_{li}^{EMEX\text{-}E})^2} \frac{n_{ir}}{n_{ir} - 1} (1 - \pi_{\ell|G_r, u_i}) \left(1 - \frac{1 - \pi_{\ell|G_r, u_i}}{\sum_{\ell \in S_\ell \cap u_i} (1 - \pi_{\ell|G_r, u_i})} \right)$$

- En région Île-de-France, $w_\ell = \dfrac{c_\ell}{\alpha_{ir}\pi_{\ell|G_r}}$ et

$$q_\ell^L = -\sum_{\substack{l\in S_r \\ l\neq k}} \frac{\hat{\alpha}_{kl,r} - \alpha_{kr}\alpha_{lr}}{\hat{\alpha}_{kl,r}} + \frac{\sum_{k\in U}\hat{\alpha}_{kl,r} - \alpha_{lr}\sum_{k\in U}\alpha_{k,r}}{\alpha_{lr}} + \frac{1}{\alpha_h^2}\frac{n_h}{n_h-1}(1-\pi_{\ell|G_r})\left(1-\frac{1-\pi_{\ell|G_r}}{\sum_{\ell\in S_h}(1-\pi_{\ell|G_r})}\right)$$

- Pour les sous-échantillons de logements neufs ou spécifique ZUS, $w_\ell = \dfrac{c_\ell}{\pi_\ell}$ et

$$q_\ell^L = \frac{n_h}{n_h-1}(1-\pi_\ell)\left(1-\frac{1-\pi_\ell}{\sum_{\ell\in S_h}(1-\pi_\ell)}\right)$$

Empiriquement, la valeur de q_ℓ^L est petite devant w_ℓ^2 (Tableau 3) : ne pas intégrer ce paramètre relativement complexe à calculer ne modifie ainsi que très marginalement les estimations de variance. Si dans le cas de l'enquête Logement 2013 ce terme est calculé précisément et introduit dans les formules, dans d'autres enquêtes il pourrait être négligé sans que cela n'affecte la qualité de l'estimation de variance.

Tableau 3 : Distribution de q_ℓ^L / w_ℓ^2

Moyenne	$1{,}43\times10^{-5}$
Maximum	$5{,}93\times10^{-3}$
P95	$7{,}54\times10^{-5}$
D9	$4{,}87\times10^{-5}$
Q3	$2{,}27\times10^{-5}$
Médiane	$6{,}76\times10^{-6}$
Q1	$1{,}76\times10^{-6}$
D1	$5{,}19\times10^{-7}$
P5	$-1{,}49\times10^{-5}$
Minimum	$-3{,}76\times10^{-3}$

Lecture : En moyenne, le rapport q_ℓ^L / w_ℓ^2 *vaut* $1{,}43\times10^{-5}$

Variance totale. L'échantillonnage étant indépendant dans chacune des quatre configurations (après prise en compte du partage des poids), la variance totale de la variable d'intérêt y est la somme des termes de variance associés à chacune de ces configurations et de la variance additionnelle associée à la non-réponse :

$$\hat{V}(\hat{Y}^R_{\text{Calé}}) = \hat{V}^{\text{Hors extension}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}^{\text{NPdC}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}^{\text{IdF}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}^{\text{Neufs+ZUS}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}_{\text{NR}}(\hat{Y}^R_{\text{Calé}})$$

En réorganisant les termes de variance, on obtient alors la formule effectivement estimée :

$$\hat{V}(\hat{Y}^R_{\text{Calé}}) = \hat{V}^{\text{Hors extension+NPdC+IdF}}_{\text{GR}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}^{\text{Hors extension+NPdC}}_{\text{UP}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}_{\text{LOG}}(\hat{Y}^R_{\text{Calé}}) + \hat{V}_{\text{NR}}(\hat{Y}^R_{\text{Calé}})$$

4. Estimations de précision dans l'enquête Logement 2013

Cette dernière partie propose une application de la méthodologie du calcul de précision dans l'enquête Logement 2013 sur un nombre restreint de variables[22]. Elles ont été choisies sur la base des premières exploitations réalisées à partir de l'enquête Logement 2006 [8]. Les estimations proposées ici reposent sur des données provisoires ; à ce titre, seules les statistiques relatives au calcul de précision (coefficient de variation, *design effect*, *cf. infra*) sont présentées dans cet article.

L'ensemble de ces calculs sont réalisés en utilisant la version 1.0 de la macro SAS ***%precisionEnl13*** de calcul de précision dans l'enquête Logement 2013. Ce programme a été spécifiquement développé pour mener à bien le calcul de la précision dans l'enquête Logement 2013. Il repose sur deux principes :

- Modularité : les différentes étapes du programme de calcul sont structurées sous la forme de modules spécifiques. Ceux-ci sont réutilisables d'une enquête à l'autre ce qui facilite le développement de programme *ad hoc* de calcul de précision tout en augmentant leur fiabilité. Des modules sont en particulier consacrés à la préparation des données techniques et au calcul des termes de variance relatifs au tirage des groupes de rotation du recensement ($\hat{V}_{GR}$) et au tirage des unités primaires de l'échantillon-maître ($\hat{V}_{UP}$).
- Ergonomie : une partie importante du programme est consacrée à le rendre simple d'utilisation. Il intègre des fonctionnalités directement orientées vers sa mise en œuvre pratique : prise en charge des variables qualitatives, des formats SAS, de l'estimation sur les sous-domaines définis par les modalités d'une variable. Sa syntaxe reprend les instructions des procédures SAS classiques : DATA, VAR, TABLES, FORMAT, WHERE et BY. De nombreux contrôles et messages d'information ou d'avertissement sont programmés. L'instruction OUTPUT permet de choisir les statistiques à afficher dans la fenêtre de résultats ou de rediriger les résultats détaillés de l'estimation de précision dans des tables pour une réutilisation dans SAS.

4.1. La méthode d'estimation retenue limite les sur- ou sous-estimations trop importantes de la précision

La spécificité de la méthodologie de calcul de précision mise en œuvre ici, présentée dans la partie 2 de cet article et développée dans [9], tient à l'utilisation de probabilités d'inclusion double estimées par simulation en exploitant les propriétés de martingale de l'algorithme du Cube [4]. Faute de disposer de ces probabilités d'inclusion double, l'estimation de la variance des termes associés au tirage des groupes de rotation du recensement et à l'échantillonnage des unités primaires de l'échantillon-maître avec la formule de Yates-Grundy serait impossible. Dans le cadre d'un sondage à probabilités inégales, une alternative serait d'utiliser, comme c'est le cas pour le second degré de tirage, la formule proposée par Deville [3].

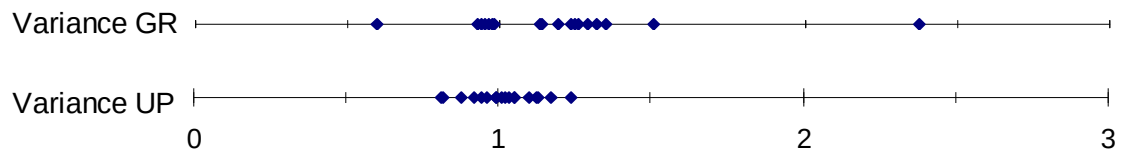
On compare ici les estimations de variance produites par les deux méthodologies, Yates-Grundy d'une part et Deville d'autre part. L'objectif n'est ni de valider la méthode d'estimation des probabilités d'inclusion double (effectué par simulation dans [9]) ni de discuter des propriétés statistiques des deux estimateurs. La portée de cet exercice est en effet fortement limitée par l'imprécision associée à l'estimation de la variance avec les deux méthodologies. L'objectif est de mesurer l'écart avec les estimations qui auraient été disponibles en l'absence d'investissement méthodologique dans le calcul des probabilités d'inclusion double associées au tirage avec Octopusse.

Pour une variable et un terme de variance donné, l'indicateur utilisé est le rapport entre la valeur estimée avec la formule de Deville et la valeur estimée avec la formule de Yates-Grundy. Plus la valeur de l'indicateur s'écarte de 1, plus l'écart entre les deux estimations est important.

[22] Surface du logement (variable HST), nombre de pièces d'habitation hors cuisine (HPH), année d'achèvement de la construction (IAAT), statut d'occupation (SEC), signes d'humidité sur certains murs du logement (GHUMI2), infiltrations ou inondations provenant d'une fuite d'eau dans la plomberie (GINOIB_1, GINOIB_2), bonne isolation phonique du logement (KBS0), installation électrique déficiente (GELEC2B, GELEC3), indicateur de surpeuplement du logement (KIP3), manque de confort sanitaire de base (KA0R, KWC1, KDLKB), déménagement envisagé (ODL), demande ou renouvellement de demande de HLM au cours des 12 derniers mois (OIH).

Les résultats fournis par les deux estimateurs sont rarement très proches l'un de l'autre (Figure 1) : dans plus de 80 % des cas l'écart entre les deux estimations est supérieur à 5 %. Pour l'estimation de la variance associée au tirage des groupes de rotation du recensement de la population, les écarts peuvent être importants : dans la plupart des cas, l'estimation obtenue avec la méthode de Deville conduit à une valeur supérieure de 10 % ou plus à celle obtenue avec la méthode de Yates-Grundy. Les écarts constatés dans l'estimation de la variance associée à l'échantillonnage des unités primaires de l'échantillon-maître sont plus faibles : ils n'excèdent jamais 20 %. Dans une majorité de cas les deux estimations diffèrent cependant de plus de 10 %. Ces éléments confirment ainsi l'intérêt pratique de l'investissement méthodologique dans l'estimation des probabilités d'inclusion double associées à l'échantillonnage avec Octopusse.

Figure 1 : Rapports entre les composantes de variance estimées avec les formules de Yates-Grundy et de Deville



Source : Échantillon principal de l'enquête Logement 2013, données provisoires
Champ : France métropolitaine
Note : GR : Groupe de rotation ; UP : Unité primaire.
Lecture : Pour les variables de l'enquête analysées, le rapport maximal entre le terme de variance associé à la sélection des groupes de rotation du recensement estimé avec la formule proposée par Deville [3] et l'estimateur de Yates-Grundy est d'environ 2,4. La valeur maximale de ce même rapport pour la variance associée à l'échantillonnage des unités primaires est de l'ordre de 1,2.

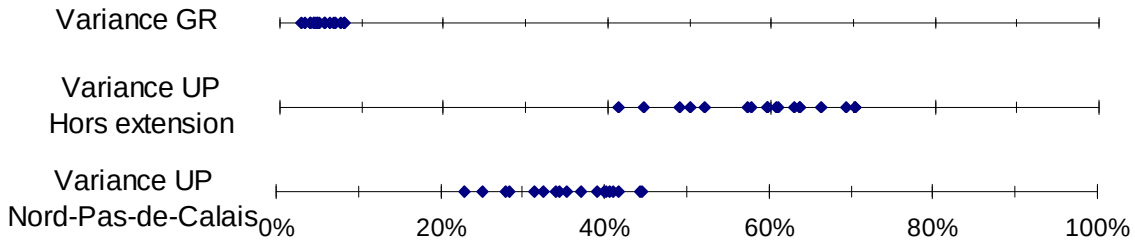
4.2. La perte de précision associée aux groupes de rotation du recensement est limitée

Comme système d'échantillon-maître, la principale particularité d'Octopusse est de reposer sur un échantillon d'unités primaires rendues en partie aléatoires par l'alternance des groupes de rotation du recensement et le tirage dans la dernière enquête annuelle de recensement disponible. Si elle permet de tirer des échantillons dans une base de sondage plus « fraîche », cette situation induit un aléa supplémentaire et une perte spécifique en termes de précision.

C'est cette perte de précision que mesure le terme de variance associé au tirage des groupes de rotation du recensement $\hat{V}_{GR}$. Pour chaque variable de l'enquête analysée, l'indicateur ici retenu est la part de la variance totale que représente ce terme de variance (Figure 2). Il apparaît très clairement que la part de la variance spécifiquement due au tirage des groupes de rotation du recensement est systématiquement faible, toujours inférieure à 8 %. Si ces estimations sont elles-mêmes sujettes à une certaine variance, ces éléments convergents semblent indiquer que dans une grande majorité de cas, la perte de précision associée aux groupes de rotation du recensement est limitée.

D'autre part, la part de variance associée à l'échantillonnage des unités primaires varie sensiblement selon que le tirage a eu lieu dans l'EM (échantillon principal en France métropolitaine hors extensions) ou dans l'EMEX élargi (échantillon principal en région Nord-Pas-de-Calais) : comme attendu, la part de la variance associée au tirage des unités primaires est plus importante quand le nombre d'unités primaires échantillonné est plus faible (EM).

Figure 2 : Décomposition de la variance dans l'échantillon principal



Source : Échantillon principal de l'enquête Logement 2013, données provisoires
Champ : France métropolitaine
Note : GR : Groupe de rotation ; UP : Unité primaire. En région Île-de-France, l'échantillonnage ayant eu lieu directement dans l'EAR 2011, il n'y a pas de variance associée à la sélection d'unités primaires.
Lecture : Pour les variables de l'enquête analysées, la part de la variance associée au tirage des groupes de rotation du recensement n'excède jamais 8 % de la variance totale. Dans les régions métropolitaines hors extensions, la variance associée à l'échantillonnage des unités primaires de l'EM oscille entre 45 % et 80 % de la variance totale. Dans la région Nord-Pas-de-Calais, la variance associée à l'échantillonnage des unités primaires de l'EMEX élargi oscille entre 20 % et 45 % de la variance totale.

4.3. Les extensions régionales augmentent sensiblement la précision des estimations

Dans le contexte d'une enquête par sondage, la finalité du calcul de précision est de nourrir la réflexion sur la qualité des informations recueillies pour en enrichir le commentaire. Deux indicateurs sont utilisés ici dans cette application à l'enquête Logement 2013 :

- Pour une variable donnée, le coefficient de variation (CV) est le rapport entre l'écart-type lié à l'estimation par sondage et la valeur de l'estimateur ponctuel :

$$CV(\hat{Y}^{R}_{Calé}) = \frac{\sqrt{\hat{V}(\hat{Y}^{R}_{Calé})}}{\hat{Y}^{R}_{Calé}}$$

  Plus le coefficient de variation est grand, plus l'imprécision associée à l'estimation par sondage est grande devant la valeur de la variable estimée par l'enquête.
- Pour une variable donnée, le *design effect* (*Deff*) est le rapport entre la variance estimée sous le plan de sondage et la variance associée à un sondage aléatoire simple de même taille :

$$Deff(\hat{Y}^{R}_{Calé}) = \frac{\hat{V}(\hat{Y}^{R}_{Calé})}{\hat{V}_{SAS}(\hat{Y}^{R}_{Calé})}$$

  Dans le cadre de la méthodologie présentée ici, le calcul de la variance du sondage aléatoire simple s'inspire de celle mise en œuvre dans le logiciel Poulpe [5, p. 20-21]. Pour un plan de sondage à plusieurs degrés, le *design effect* est en général strictement supérieur à 1 : plus il est proche de 1, plus le plan de sondage parvient à limiter l'effet de grappe induit par la corrélation spatiale au sein des unités primaires d'un tirage à plusieurs degrés [6].

Ces deux indicateurs répondent donc à deux questions distinctes : le coefficient de variation rend compte de la fiabilité de l'information recueillie quand le *design effect* renseigne sur l'efficacité de l'échantillonnage à plusieurs degrés.

Au niveau national, les indicateurs calculés sur les quelques variables mobilisées dans cette étude confirment la qualité générale des estimations obtenues (Tableau 4). Toutes les estimations sauf une présentent un coefficient de variation inférieur à 5 %. Le cas du manque de confort sanitaire de base est

particulier dans la mesure où l'estimation ponctuelle est très basse (350 000 logements en 2006 sur 26,3 M [8]) : cette proportion étant très marginale, le nombre d'individus sur lesquels porte effectivement l'estimation est particulièrement faible, d'où une précision moindre. D'autre part, les *design effect* compris entre 1,5 et 2,2 semblent indiquer que la perte liée à l'échantillonnage à plusieurs degrés en termes de précision est maîtrisée.

Une des particularités de l'enquête Logement 2013 est d'avoir donné lieu à deux extensions régionales, une en région Nord-Pas-de-Calais et une en région Île-de-France. Afin de comparer l'impact de ces extensions régionales sur la précision des estimations, les résultats au sein des deux régions à extensions sont comparés à ceux de deux régions aussi proches que possibles en termes de taille de population :

- Les estimations en région Nord-Pas-de-Calais (4 042 015 habitants en 2011) sont comparées à celles de la région Pays de la Loire (3 601 113 habitants en 2011).
- Faute de disposer d'une région de taille équivalente, les estimations en région Île-de-France (11 852 851 habitants en 2011) sont comparées à celles de la seconde région la plus peuplée, Rhône-Alpes (6 283 541 habitants en 2011).

Limité, cet exercice fournit quelques points de repères des apports des extensions régionales en termes de précision (Tableau 5). Comme attendu, la précision des estimations est supérieure dans les régions à extensions, tout particulièrement en région Île-de-France. En termes de précision, les extensions permettent ainsi une exploitation fiable au niveau régional, y compris quand les variables d'intérêt ne sont pas directement liées aux variables de calage (indicateur de surpeuplement du logement par exemple) ; cela n'est pas le cas dans les régions sans extension avec les estimateurs habituels[23].

Dans le cas de la région Île-de-France, l'objectif du tirage direct dans l'EAR (et non pas par le biais des unités primaires d'un échantillon-maître) était de garantir la représentativité infrarégionale des estimations. Les coefficients de variation des estimations sont ainsi calculés par sous-domaine régionaux en Île-de-France (Tableau 6), en distinguant Paris, les départements de la « petite couronne » (Hauts-de-Seine, Seine-Saint-Denis et Val-de-Marne) et les départements de la « grande couronne » (Seine-et-Marne, Yvelines, Essonne et Val-d'Oise). Dans l'ensemble, des niveaux satisfaisants de précision semblent être atteints pour la plupart des variables (notamment isolation phonique, surpeuplement, déménagement). Les coefficients de variation les plus élevés sont atteints pour les proportions les plus faibles, en particulier l'indicateur de manque de confort de base ainsi que la part des logements construits après 1999 pour le département de Paris.

Dans leur ensemble ces estimations de précision confirment la qualité des données recueillies par l'enquête Logement 2013, au niveau national et au niveau régional dans les régions à extension (Nord-Pas-de-Calais et Île-de-France). Elles appellent également à la prudence quand les proportions estimées sont faibles en niveau ou les domaines de diffusion restreints. Le programme ***%precisionEnl13*** est en mesure de fournir des estimations de précision dans des situations complexes (variables qualitatives recodées, sous-domaines spécifiques) et peut ainsi être utilisé pour juger en situation de la fiabilité des statistiques calculées.

[23] La méthode de calcul de précision utilisée ici porte sur des estimateurs « directs » des variables d'intérêt sur les domaines régionaux. Des méthodologies d'estimation alternatives, en particulier les méthodes d'estimation sur petits domaines, peuvent permettre des exploitations d'enquêtes nationales sur des domaines régionaux (voir par exemple [10]).

Tableau 4 : Coefficient de variation et *design effect* au niveau national

	Coefficient de variation	*Design effect*
Surface du logement	0,20 %	1,50
Nombre de pièces d'habitation hors cuisine	0,17 %	1,90
Année d'achèvement de la construction		
Avant 1948	0,76 %	1,77
De 1949 à 1974	1,04 %	1,80
De 1975 à 1998	1,18 %	2,00
En 1999 et après	1,22 %	1,91
Statut d'occupation		
Propriétaire non-accédant	0,76 %	2,17
Accédant à la propriété	1,49 %	2,19
Locataire secteur libre	0,82 %	1,65
Locataire HLM et autre logement social	0,52 %	1,54
Signes d'humidité sur certains murs du logement	1,46 %	1,65
Infiltrations ou inondations	4,06 %	1,41
Bonne isolation phonique du logement	0,61 %	2,03
Installation électrique déficiente	3,06 %	1,52
Indicateur de surpeuplement du logement	1,83 %	1,50
Manque de confort sanitaire de base	8,82 %	1,61
Déménagement envisagé	1,39 %	1,81
Demande ou renouvellement de demande de HLM au cours des 12 derniers mois	3,15 %	1,58

Source : Enquête Logement 2013, données provisoires
Champ : France métropolitaine
Lecture : Le coefficient de variation associé à l'estimation de la surface moyenne du logement est de 0,20 %, tandis que le design effect *est de 1,50.*

Tableau 5 : Comparaison des coefficients de variation entre les régions Nord-Pas-de-Calais et Pays de la Loire, Île-de-France et Rhône-Alpes

	CV 31	CV 52	CV 11	CV 82
Surface du logement	0,54 %	1,49 %	0,44 %	1,59 %
Nombre de pièces d'habitation hors cuisine	0,41 %	1,12 %	0,32 %	1,18 %
Année d'achèvement de la construction				
Avant 1948	1,81 %	6,59 %	1,30 %	5,38 %
De 1949 à 1974	2,44 %	5,59 %	1,58 %	4,60 %
De 1975 à 1998	2,82 %	3,98 %	1,93 %	4,68 %
En 1999 et après	3,16 %	5,40 %	2,99 %	7,88 %
Statut d'occupation				
Propriétaire non-accédant	1,79 %	3,12 %	1,60 %	3,63 %
Accédant à la propriété	3,75 %	4,59 %	2,67 %	5,64 %
Locataire secteur libre	2,12 %	5,32 %	2,13 %	4,73 %
Locataire HLM et autre logement social	0,86 %	7,74 %	0,88 %	8,41 %
Signes d'humidité sur certains murs du logement	3,14 %	7,89 %	2,41 %	6,61 %
Infiltrations ou inondations	9,46 %	13,23 %	7,54 %	14,98 %
Bonne isolation phonique du logement	1,41 %	2,26 %	1,29 %	2,92 %
Installation électrique déficiente	7,63 %	19,36 %	5,36 %	11,38 %
Indicateur de surpeuplement du logement	5,21 %	11,26 %	2,33 %	8,77 %
Manque de confort sanitaire de base	16,63 %	55,54 %	16,04 %	33,71 %
Déménagement envisagé	3,41 %	6,77 %	1,98 %	5,38 %
Demande ou renouvellement de demande de HLM au cours des 12 derniers mois	7,13 %	17,80 %	4,73 %	12,83 %

Source : Enquête Logement 2013, données provisoires
Champ : Régions Nord-Pas-de-Calais, Pays de la Loire, Île-de-France, Rhône-Alpes.
Note : CV : Coefficient de variation ; 31 : Région Nord-Pas-de-Calais ; 52 : Région Pays de la Loire ; 11 : Région Île-de-France ; 82 : Région Rhône-Alpes.
Lecture : Le coefficient de variation associé à l'estimation de la surface moyenne du logement est de 0,54 % en région Nord-Pas-de-Calais contre 1,49 % en région Pays de la Loire. Il est de 0,44 % en région Île-de-France contre 1,59 % en région Rhône-Alpes.

Tableau 6 : Comparaison des coefficients de variation au niveau infra-régional en Île-de-France

	CV Paris	CV Petite couronne	CV Grande couronne
Surface du logement	1,49 %	0,81 %	0,87 %
Nombre de pièces d'habitation hors cuisine	1,22 %	0,74 %	0,67 %
Année d'achèvement de la construction			
Avant 1948	1,22 %	2,78 %	3,59 %
De 1949 à 1974	3,50 %	2,29 %	2,75 %
De 1975 à 1998	4,79 %	3,38 %	2,74 %
En 1999 et après	19,17 %	4,25 %	4,24 %
Statut d'occupation			
Propriétaire non-accédant	3,46 %	2,74 %	2,51 %
Accédant à la propriété	8,39 %	4,78 %	3,58 %
Locataire secteur libre	3,32 %	3,85 %	3,98 %
Locataire HLM et autre logement social	2,30 %	1,39 %	1,17 %
Signes d'humidité sur certains murs du logement	4,41 %	3,93 %	4,30 %
Infiltrations ou inondations	13,17 %	12,51 %	13,55 %
Bonne isolation phonique du logement	3,64 %	2,32 %	1,67 %
Installation électrique déficiente	8,52 %	8,91 %	10,67 %
Indicateur de surpeuplement du logement	4,40 %	4,01 %	5,26 %
Manque de confort sanitaire de base	19,03 %	34,59 %	58,21 %
Déménagement envisagé	3,92 %	3,10 %	3,58 %
Demande ou renouvellement de demande de HLM au cours des 12 derniers mois	9,60 %	7,02 %	9,01 %

Source : Enquête Logement 2013, données provisoires
Champ : Régions Île-de-France et Rhône-Alpes
Note : CV : Coefficient de variation ; Petite couronne : départements 92, 93 et 94 ; Grande couronne : départements 77, 78, 91 et 95.
Lecture : Le coefficient de variation associé à l'estimation de la surface moyenne du logement est de 1,49 % à Paris, de 0,81 % dans la petite couronne et de 0,87 % dans la grande couronne.

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*.

[7] Insee, Note de spécification du tirage des échantillons pour l'enquête Logement 2013, N°714/DG75-L110/PA-P

[8] Castéran B. et Ricroch L. (2008), Les logements en 2006. Le confort s'améliore, mais pas pour tous, Insee Première, N°1202

[9] Gros E., Moussallam K. (2015), Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse, *Document de travail de l'Insee M 2015/01*

[10] Ardilly P. (2014), Estimation régionale de taux de pauvreté utilisant une technique de calage, Actes du 8ème colloque francophone sur les sondages