Les calculs de précision dans Octopusse

Théorie et application à l'enquête Logement 2013





Martin Chevalier, Emmanuel Gros
Insee – DMCSI – Division Sondages

Karim Moussallam
Dares – Sous-direction Salaires, travail et relations professionnelles

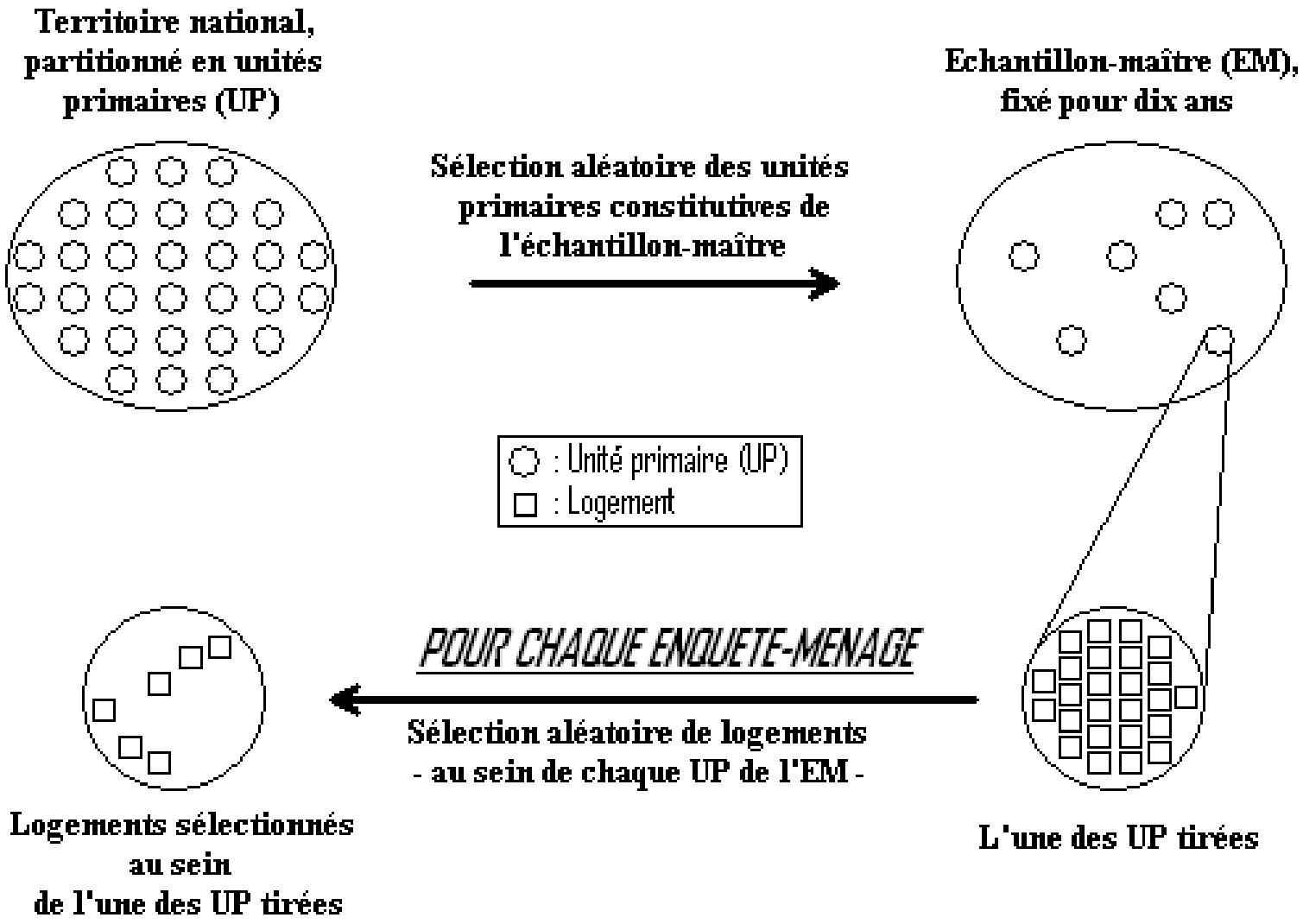
Insee
Mesurer pour comprendre

Plan de l'exposé

- Contexte et cadre théorique d'Octopusse
- Estimation de la variance liée aux deux premiers niveaux d'aléa d'Octopusse
- Prise en compte du degré de tirage des logements et estimation de la variance sur l'échantillon de logements final
- Application à l'enquête Logement 2013 en métropole
- Résultats de l'estimation de variance

Contexte et cadre théorique d’Octopusse

Le concept d'Échantillon-Maître (EM)



Le recensement rénové de la population

- Passage d'un recensement exhaustif tous les 7 à 9 ans à un recensement annuel par sondage.
- Pour les petites communes (PC) de moins de 10 000 habitants :
  - ✓ stratification par région ;
  - ✓ communes réparties en 5 groupes de rotation par tirage équilibré ;
  - ✓ une année donnée, enquête auprès de l'ensemble des logements des petites communes d'un groupe de rotation.
- Pour les grandes communes (GC) :
  - ✓ stratification par type d'adresse ;
  - ✓ répartition des adresses en 5 groupes de rotation ;
  - ✓ une année donnée, enquête auprès de 40 % environ des logements d'un groupe de rotation.

↳ **Perte de l'exhaustivité mais apport d'informations récentes**

Impact sur le système d'échantillonnage

- Double objectif :
  - ✓ conserver le principe de tirage à deux degrés des EM classiques ;
  - ✓ tirer les échantillons dans la partie la plus récente de la base de sondage ⇒ pour une enquête donnée, sélection de logements recensés l'année précédente.

↳ Nécessité d'une refonte du système d'échantillonnage :

- ✓ conception de nouvelles Unités Primaires intégrant l'aspect rotatif du nouveau recensement ⇒ Zone d'Action Enquêteur ;
- ✓ constitution d'un nouvel échantillon-maître assurant une bonne représentativité du territoire national ;
- ✓ opérations statistiques supplémentaires : calage des ZAE, rééchantillonnage des logements en GC.

Octopusse, un système d'échantillonnage complexe (1)

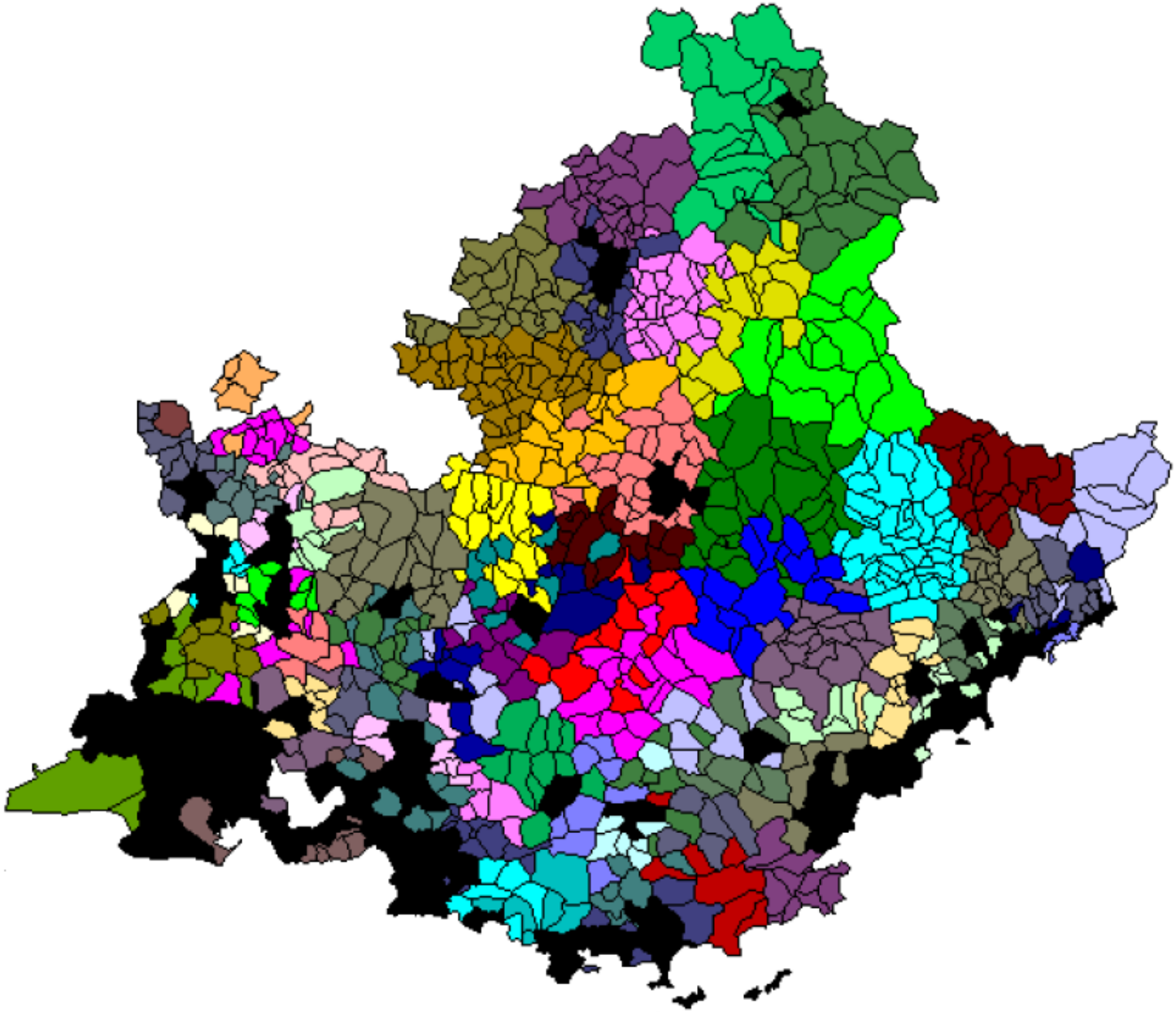
❶ Affectation des petites communes / adresses en grande commune aux groupes de rotation du RP par tirage équilibré.

❷ Conditionnellement aux groupes de rotation du RP, constitution des ZAE :

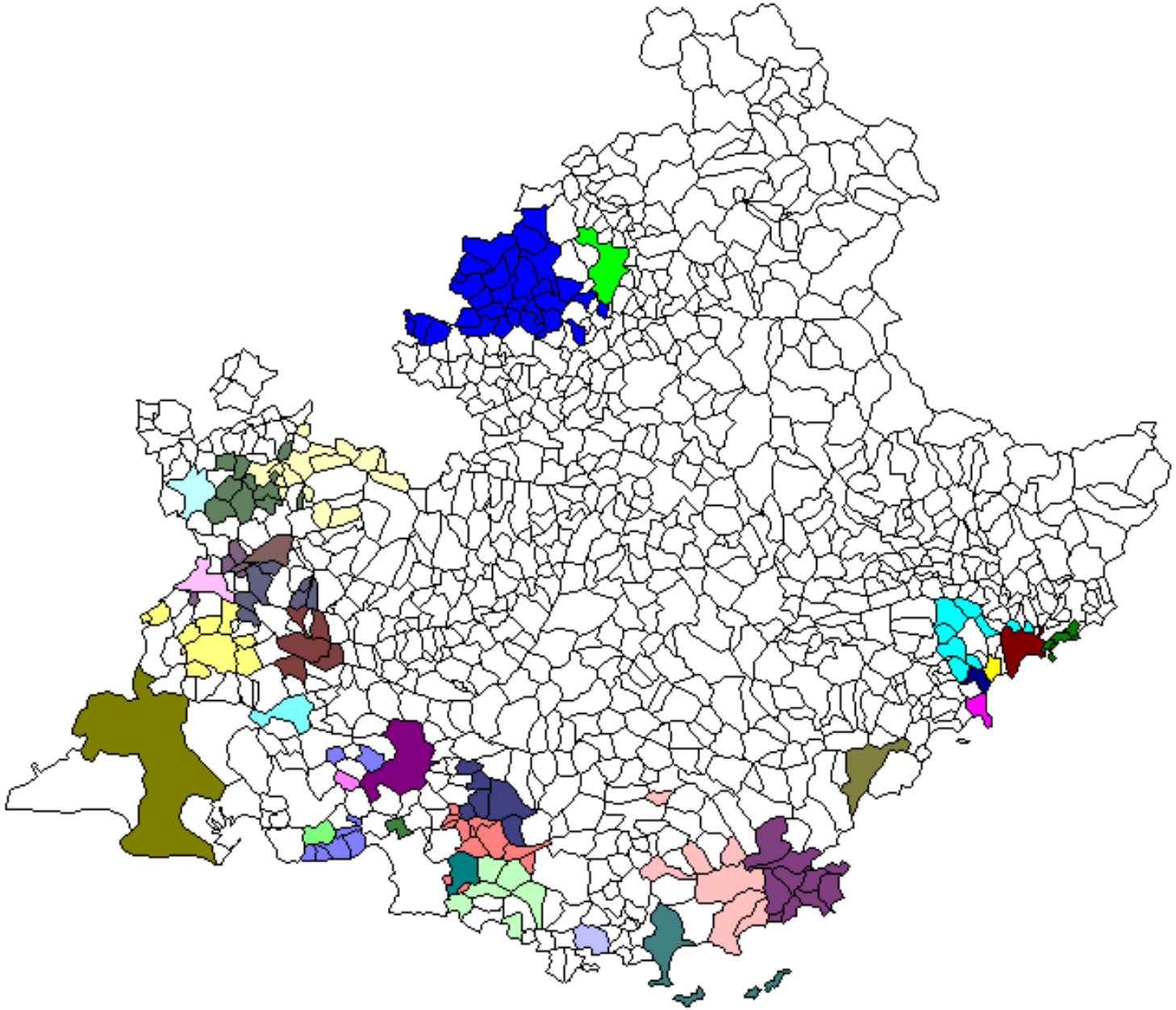
- ✓ chaque grande commune constitue une ZAEGC ;
- ✓ au niveau des petites communes, les ZAEPC sont constituées en regroupant des PC proches de telle sorte que chaque ZAEPC contienne au moins 300 résidences principales de chaque groupe de rotation.

❸ Sélection d'un Échantillon-Maître de ZAE, par tirage stratifié par région et équilibré.

Exemple : les ZAE en PACA



Exemple : l'EM en PACA



Octopusse, un système d'échantillonnage complexe (2)

❹ Opérations statistiques spécifiques : rééchantillonnage des logements en GC et calage des ZAE.

❺ Pour une enquête d'une année N donnée, sélection des logements au sein de la fraction des ZAE de l'EM recensée l'année N–1 :

- ✓ dans les ZAEGC, sélection des logements au sein de l'échantillon de logement recensé en N–1 ;
- ✓ dans les ZAEPC, sélection des logements dans les communes recensées en N–1.

➔ 3 niveaux d'aléas : tirage des GR du RP, tirage des ZAE de l'EM, tirage des logements au sein de la fraction recensée des ZAE de l'EM.

Estimation de la variance liée aux tirages des groupes de rotation du recensement et des ZAE de l'échantillon-maître

Estimateur en expansion

On suppose dans un 1er temps connus les totaux par commune

Soit U la population des communes. Pour une année N donnée, on va sélectionner les logements dans les communes appartenant aux ZAE u_i de l'EM et au groupe de rotation G_r recensé en N–1 :

$$S_r = \{ k \in U ; k \in u_i \in EM \text{ et } k \in G_r \}$$

➔ Le total d'une variable Y peut être estimé sans biais grâce à l'estimateur en expansion suivant :

$$\hat{t}_{yr} = \sum_{k \in S_r} \frac{y_k}{\alpha_{kr} \pi_{Ii}}$$

avec α_{kr} = probabilité de sélection de la commune k dans G_r

π_{Ii} = probabilité de sélection de la ZAE u_i dans l'EM

Variance de l'estimateur en expansion (1)

Dans une présentation aux JMS de 2012, Guillaume Chauvet a proposé un estimateur sans biais de cette variance :

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

avec $\tilde{Y}_{ir} = \sum_{k\in u_i} \frac{y_k 1I_{k\in G_r}}{\alpha_{kr}}$ et où les probabilités d'inclusion doubles

$\hat{\alpha}_{kl,r}$ et $\hat{\pi}_{Iij}$ sont estimées par réplication en s'appuyant sur les propriétés de martingale de l'algorithme de tirage équilibré Cube.

➔ Des études sur données simulées prouvaient l'efficacité de cette procédure d'estimation.

Variance de l'estimateur en expansion (2)

- Calcul des $\hat{\alpha}_{kl,r}$ et $\hat{\pi}_{Iij}$ sur les données d'Octopusse :
  - ✓ 430 000 réplications pour $\hat{\alpha}_{kl,r}$ et 2,3 millions pour $\hat{\pi}_{Iij}$;
  - ✓ bonne qualité d'estimation des probabilités : $\in$]0,1], taux d'erreur très faible sur les probabilités d'inclusion simple et décroissant avec le nombre de réplication, etc.
- Étude de la qualité des estimateurs de Yates-Grundy :
  - ✓ estimateurs non biaisés, contrairement à l'approximation de Deville-Tillé (DT) ;
  - ✓ pas de problème d'instabilité des estimateurs de variance : leur dispersion reste mesurée, et inférieure à celle de l'estimateur DT
  - ✓ exception : Corse pour $\hat{V}_{EM}$ → estimateur de Deville préférable.

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

Estimateur d'Octopusse et estimateur en expansion

➢ Estimateur d'Octopusse :

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

➢ Hypothèses simplificatrices :

✓ Calage des ZAE non pris en compte

✓ dans les ZAEGC, on assimile l'enchaînement « sélection des adresses du RP → rééchantillonnage → tirage final des logements » à un unique tirage à probabilités $\pi_{l \mid Gr,\, ui}$

➔ Estimateur en expansion : $\hat{Y}_\pi = \sum_{\ell \in S_l} \dfrac{y_\ell}{\alpha_{kr} \pi_{Ii} \pi_{\ell \mid G_r, u_i}}$

La formule de Rao

Contexte : plan de sondage à 2 degrés avec tirages de 2nd degré indépendants entre UP ; π_i probabilités de tirage de 1er degré, T_i=total de Y sur UP i, estimé sans biais par $\hat{T}_i$

Formule de Rao : Si $Q(T_1,...,T_m) = \sum_{i \in S_I} q_i T_i^2 + \sum_{(i,j) \in S_I,\, i \neq j} q_{ij} T_i T_j$ permet

d'estimer sans biais la variance de 1er degré sur l'échantillon S_I, et si $\hat{V}(\hat{T}_i)$ estime sans biais les variances de second degré, on peut estimer sans biais à partir de l'échantillon final S la variance de l'estimateur d'Horvitz-Thompson du total T par :

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

Vers une formule de variance générique pour Octopusse

➢ La formule de Rao permet :

- ✓ de prendre en compte le degré de tirage des logements au sein de la fraction des ZAE de l'EM recensée l'année précédente ;
- ✓ de prendre en compte le traitement de la non-réponse par repondération : modélisation de la NR par un mécanisme poissonnien → la phase de non-réponse peut être vue comme un degré de sondage supplémentaire et on peut appliquer à nouveau la formule de Rao.

➢ Prise en compte du calage sur marges via les résidus de la régression de la variable d'intérêt sur les variables de calage.

Application à l’enquête Logement 2013 en métropole

Les particularités de l'enquête Logement 2013

➢ La méthodologie du calcul de variance dans les enquêtes tirées avec Octopusse a déjà été mise en œuvre en pratique (*Adult education survey* 2012).

➢L'enquête Logement 2013 présente plusieurs spécificités qui appellent des adaptations de ce cadre général :

- ✓ Extensions régionales en Nord-Pas-de-Calais et en Île-de-France.
- ✓ Sous-échantillons spécifiques logements neufs et Zones urbaines sensibles (ZUS).

Plan de sondage de l'enquête Logement 2013

➢ Echantillon principal : 35 600 logements tirés dans l'Enquête annuelle de recensement 2011

- ✓ Régions métropolitaines hors extension : tirage dans les ZAE de l'échantillon-maître (EM).
- ✓ Région Nord-Pas-de-Calais : tirage dans les ZAE de l'échantillon-maître pour les extensions régionales élargi (EMEX-E) ➔ environ 3 fois plus de ZAE que l'EM.
- ✓ Région Île-de-France : tirage direct dans l'EAR 2011 sans passer par un système d'échantillon-maître ➔ objectif de précision infra-régional.

Plan de sondage de l'enquête Logement 2013

➢ Echantillon de logements neufs : 850 logements tirés dans la base DGI-Sitadel

- ✓ Pour être tiré avec Octopusse, un logement doit avoir été enquêté par la dernière enquête annuelle de recensement disponible (2011 pour l'enquête Logement 2013).

- ✓ Les logements achevés après mars 2011 ne peuvent donc pas être échantillonnés avec Octopusse : une base de sondage spécifique est établie en rapprochant les données du SOeS sur les permis de construire et de la DGI sur les locaux achevés.

- ✓ Un degré de sondage supplémentaire par rapport à l'échantillon principal : tirage d'un permis puis d'un logement par permis.

Plan de sondage de l'enquête Logement 2013

➢ Echantillon spécifique Zones urbaines sensibles : 6 000 logements tirés dans les EAR 2006-2010

- ✓ Tirage direct de logements dans les Zones urbaines sensibles (sans passer par les ZAE de l'échantillon-maître).

- ✓ Partage des poids avec l'échantillon principal (logements de l'échantillon principal en ZUS, logements de l'échantillon ZUS dans les ZAE de l'échantillon-maître).

➢Echantillon de réserve : 5 000 logements tirés comme l'échantillon principal, sauf en région Île-de-France (tirage dans l'EMEX-E et non directement dans l'EAR).

Plan de sondage de l'enquête Logement 2013

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

<table>
<tr><th>Echantillons</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>

GR : Groupe de rotation du recensement

Adaptation de la méthodologie de calcul de précision

- Echantillon principal : pas de difficulté particulière
  - ✓ Régions métropolitaines hors extension : application directe.
  - ✓ Région Nord-Pas-de-Calais : remplacement des probabilités d'inclusion simple et double de l'EM par celles de l'EMEX-E (estimées par réplication).
  - ✓ Région Île-de-France : tirage en un seul degré donc pas de terme de variance spécifique à l'échantillonnage dans un système d'échantillon-maître.

Adaptation de la méthodologie de calcul de précision

➢ Echantillon de logements neufs : variance d'un sondage à probabilités inégales stratifié

- ✓ Dans 90 % des cas, au plus un permis est échantillonné par ZAE dans les régions hors extensions et Nord-Pas-de-Calais : le premier degré de tirage n'est pas pris en compte.
- ✓ Dans tous les cas, un seul logement est tiré par permis, si bien que le terme de variance intra-permis n'est pas calculable : le second degré de tirage n'est pas pris en compte.
- ✓ Le tirage des logements neufs est assimilé à un tirage à un degré stratifié par région (et par taille de commune en Île-de-France).

Adaptation de la méthodologie de calcul de précision

➢ Echantillon spécifique Zones urbaines sensibles : pas de difficulté particulière.

➢Echantillon de réserve : variance d'un sondage à probabilités inégales stratifié

- ✓ La réserve n'a été déclenchée qu'en Île-de-France, et seulement partiellement (144 fiches-adresses). Dans ce contexte il n'est pas possible d'estimer la variance associée au tirage des ZAE dans l'EMEX-E.

- ✓ Le tirage de la réserve est assimilé à un tirage à un degré stratifié par taille de commune.

Plan de sondage de l'enquête Logement 2013

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

<table>
<tr><th>Echantillons</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>

GR : Groupe de rotation du recensement

Les étapes du calcul de variance

1. Calcul des résidus de la régression sur les variables de calage.

2. Calcul du terme de variance associé aux groupes de rotation du RP dans l'échantillon principal et l'échantillon de réserve.

3. Calcul du terme de variance associé au tirage des ZAE dans l'échantillon principal hors Île-de-France

   - ✓ Dans les régions métropolitaines hors extension, utilisation des probabilités d'inclusion de l'EM.

   - ✓ En région Nord-Pas-de-Calais, utilisation des probabilités d'inclusion de l'EMEX-E.

Les grandes étapes du calcul de variance

4. Calcul du terme de variance associé au tirage des logements

- ✓ Utilisation de la formule de la variance d'un sondage à probabilité inégales :

$$\hat{V}(\hat{T}) = \frac{n}{n-1}\sum_{\ell \in s}(1-\pi_\ell)\left(y_\ell - \frac{\sum_{\ell \in s}(1-\pi_\ell)y_\ell}{\sum_{\ell \in s}(1-\pi_\ell)}\right)^2$$

- ✓ Dans l'échantillon principal hors Île-de-France, application de la formule de Rao (sondage à 2 degrés).

Les grandes étapes du calcul de variance

5. Calcul du terme de variance associé à la non-réponse.

$$\hat{V}_{NR}(\hat{T}^R) = \sum_{\ell \in s} (w_\ell^2 - q_\ell^L)(1-\hat{p}_\ell)\left(\frac{y_\ell \mathbb{1}_{\ell \in R}}{\hat{p}_\ell}\right)^2$$

Remarque : empiriquement q_ℓ^L est négligeable devant w_ℓ^2

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

Moyenne	$1{,}43 \times 10^{-5}$
Maximum	$5{,}93 \times 10^{-3}$
D9	$4{,}87 \times 10^{-5}$
Q3	$2{,}27 \times 10^{-5}$
Médiane	$6{,}76 \times 10^{-6}$
Q1	$1{,}76 \times 10^{-6}$
D1	$5{,}19 \times 10^{-7}$
Minimum	$-3{,}76 \times 10^{-3}$

Résultats de l’estimation de variance

Résultats de l'estimation de variance

➢ Construction de la macro SAS de calcul de variance ***%precisionEnl13*** avec deux objectifs :

- ✓ Modularité : les termes de variance caractéristiques de l'échantillonnage avec Octopusse (groupes de rotation, ZAE) sont calculés par des modules facilement redéployables.

- ✓ Ergonomie : syntaxe analogue à celles des PROC MEANS et PROC FREQ (instructions VAR, TABLES, BY, FORMAT), plusieurs options d'affichage et d'exportation des résultats.

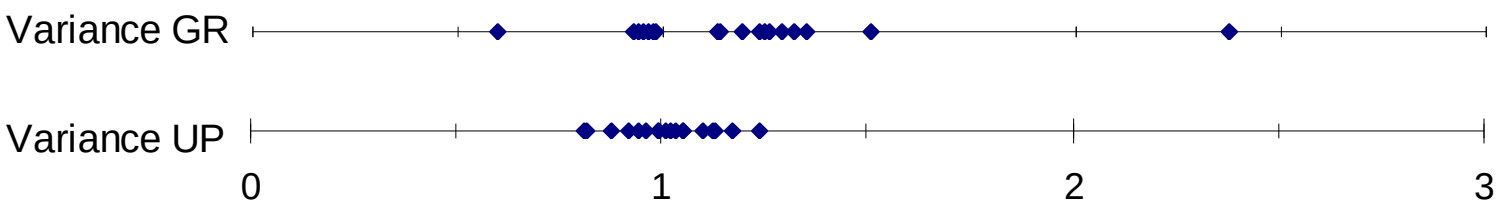
➢Mise en œuvre du calcul de précision sur les données provisoires de l'enquête Logement 2013, sur les variables utilisées dans l'Insee Première sur l'enquête Logement 2006.

Limitation des sur- et sous-estimations

➢ En l'absence de probabilités d'inclusion double, l'estimation de la variance de premier degré n'aurait pu être effectuée avec la formule de Yates-Grundy.

➢ Plusieurs formules auraient permis d'approcher cette variance, mais sans prendre totalement en compte le tirage équilibré des groupes de rotation ou des unités primaires de l'échantillon-maître.

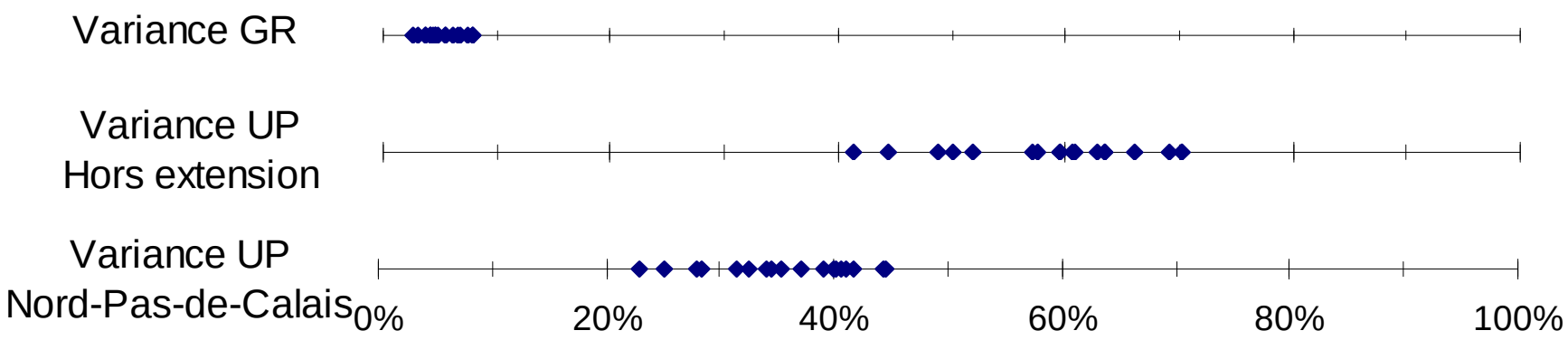
Figure 1 : Rapports entre les estimations de Deville et de Yates-Grundy



Impact limité des groupes de rotation du RP

➢ La particularité de l'échantillonnage avec Octopusse tient au tirage dans la dernière Enquête annuelle recensement disponible : la base de sondage est plus « fraîche », au prix d'une variance supplémentaire associée aux groupes de rotation du recensement.

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



Estimations précises au niveau national…

Tableau 4 : Coefficients 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
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
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

…et dans les régions à extension

Tableau 5 : Comparaison 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 %
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 %
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 %

31 : Nord-Pas-de-Calais ; 52 : Pays de la Loire
11 : Île-de-France ; 82 : Rhône-Alpes

…et dans les régions à extension

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 %
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 %
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	4,10 %	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 %

Merci de votre attention !

Contacts :

Martin Chevalier
Tél. : 01 41 17 53 75
Courriel : martin.chevalier@insee.fr

Emmanuel Gros
Tél. : 01 41 17 64 91
Courriel : emmanuel.gros@insee.fr

Karim Moussallam
Tél. : 01 44 38 24 86
Courriel : karim.moussallam@travail.gouv.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr

Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00