LES EXTENSIONS RÉGIONALES ET LOCALES DE L'ENQUÊTE LOGEMENT 2006 ÉCHANTILLONNAGE ET REPONDÉRATION

J. Le Guennec

INSEE, pôle d'ingéniérie statistique ménages

Problématique

L'INSEE réalise tous les quatre ans une enquête nationale sur le logement. La dernière s'est déroulée entre janvier et décembre 2006 en six vagues de collecte sur tout le territoire métropolitain. Outre la description détaillée des conditions de logement des habitants, le questionnaire comprenait une analyse des consommations d'énergie et des dépenses liées au logement, et un historique de la mobilité résidentielle.

Les implications locales de cette problématique expliquent l'intérêt des acteurs régionaux pour les résultats de cette enquête dans les zones dont ils ont la charge. Le plan de sondage national ne permet cependant pas son exploitation à un niveau régional, hors peut-être l'Ile de France, et à coup sûr pas sur des périmètres infra-régionaux. La technique d'échantillonnage à plusieurs degrés, avec un premier degré aréolaire non stratifié par région, comprenant peu d'unités primaires dans chacune, et la taille de l'échantillon sélectionné au second degré par région, n'assurent pas une précision suffisante à ce niveau géographique.

C'est pourquoi six régions : Nord-Pas-de-Calais, Bretagne, Midi-Pyrénées, Provence-Alpes-Côte d'Azur, Corse, Ile de France, ont financé des extensions d'échantillon pour une diffusion de résultats à des niveaux régionaux ou locaux. La région Nord-Pas-de-Calais, à forte armature urbaine, recherchait une représentativité régionale globale. L'Ile de France souhaitait distinguer ses départements. Les quatre autres régions se sont intéressées à la situation du logement dans leurs grandes agglomérations. Les conditions de logement des populations immigrées, le logement locatif, l'endettement des ménages en accession à la propriété, la mobilité résidentielle étaient quelques-uns des thèmes que ces régions souhaitaient éclairer.

Selon le niveau géographique de diffusion recherché, l'extension d'échantillon peut être obtenue :

- par le simple accroissement de la taille d'échantillon dans la zone, dans le cadre du plan de sondage national ;
- ou donner lieu au tirage d'un échantillon complémentaire totalement distinct de l'échantillon national, avec un plan de sondage original.

Compte tenu des méthodes d'échantillonnage des enquêtes auprès des ménages à l'INSEE, la première solution est possible lorqu'on ne vise qu'une représentativité régionale globale. Une diffusion sur un périmètre plus restreint nous contraint à la seconde solution. Il faut alors fusionner les échantillons issus des différents plans de sondage pour exploiter l'enquête avec une pondération unique, ce qu'on réalise en appliquant la méthode du partage des poids. Lorsqu'on dispose d'informations auxiliaires propres au domaine visé par l'extension, on peut ensuite améliorer la précision des résultats par un calage spécifique.

L'on décrit ici les problèmes posés par la repondération d'une enquête à l'intérieur d'un domaine sondé au moyen de plusieurs échantillons issus de plans de sondage distincts. Cette configuration a concerné quatre régions où se sont faites des extensions locales d'échantillon : Bretagne, Midi-Pyrénées, Corse et Provence-Alpes-Côte d'Azur.

1. Les plans de sondage

1.1. Le plan de sondage national

Afin de répondre à des attentes multiples (situation des ménages habitant des quartiers urbains soutenus pas des politiques spécifiques, situation des bénéficiaires de dispositifs d'aides au logement), l'échantillon national résulte lui-même d'un plan complexe faisant appel à diverses bases de sondage. Il est la réunion de plusieurs échantillons tirés de façon indépendante.

1.1.1. L'échantillon principal

L'échantillon principal de l'enquête Logement a été tiré dans l'échantillon-maître de 1999. Comme dans toutes les enquêtes ménages de l'INSEE, il s'agit d'un sondage à deux degrés autopondéré, le premier degré étant constitué de l'échantillon-maître (EM). Rappelons que celui-ci est un échantillon de communes, ou de groupes de communes contigües, sélectionnées aléatoirement après le dernier recensement général de population, avec des probabilités proportionnelles à leur taille. Les communes de plus de 100 000 habitants figurent toutes dans l'EM, mais avec un échantillon d'ilôts. Au second degré, les logements ont été sélectionnés dans le fichier du recensement de 1999 et dans les bases complémentaires des logements construits depuis mars 1999 : la « base de sondage des logements neufs » (BSLN), à jour au 31 décembre 2005.

1.1.2. L'échantillon ZUS

Un échantillon complémentaire a été tiré dans les « zones urbaines sensibles », afin de renforcer la précision des résultats dans ce champ. L'intersection entre les ZUS et l'échantillon maître était trop restreinte pour se contenter d'un accroissement du nombre de logements au second degré de l'échantillon principal. Le tirage a donc été effectué directement dans le fichier du recensement de 1999 dans toutes les communes et ilôts composant les ZUS, de façon indépendante de l'échantillon national principal. Un logement situé en ZUS pouvait donc être inclus dans les deux échantillons. La base de sondage des ZUS est emboîtée dans celle de l'échantillon principal.

1.1.3. L'échantillon CNAF

Deux échantillons d'allocataires ont été sélectionnés dans les fichiers de la Caisse nationale d'allocations familiales : l'un parmi les bénéficiaires d'aide à l'accession à la propriété, l'autre parmi les locataires en situation d'impayé de loyer.

Alors que par nature, RP99 et BSLN constituent des strates disjointes, toutes les autres bases de sondage avaient des intersections non vides. Un logement situé en ZUS appartient par définition à la base RP. Un allocataire appartient soit à la base RP soit à la base des logements neufs.

1.2. Les extensions régionales stricto-sensu

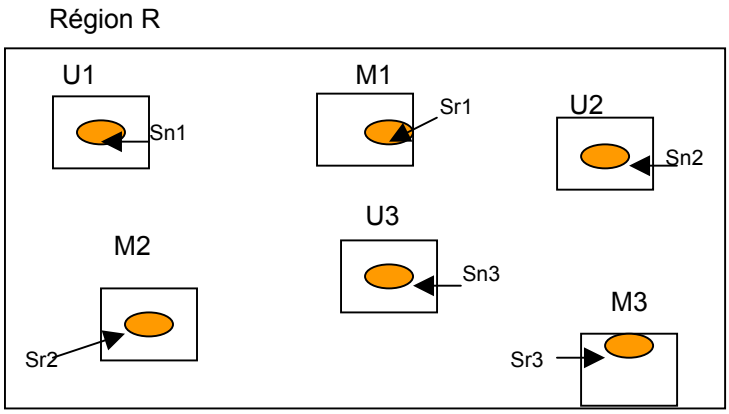
Il s'agit des échantillons complémentaires visant une représentativité régionale globale. Ces extensions sont obtenues par accroissement de la taille d'échantillon dans la région concernée, en respectant le schéma de tirage de l'échantillon principal. A l'INSEE, la technique consiste à mobiliser, au premier degré, un échantillon-maître complémentaire, l'EMEX, pour pouvoir augmenter la taille de l'échantillon de logements au second degré sans créer d'effets de grappe préjudiciables.

L'EMEX est un échantillon complémentaire de communes (ou groupes de communes, ou ilôts dans les plus grandes villes) sélectionné selon un plan de sondage similaire à l'EM, mais stratifié par région. L'EMEX ayant été tiré conditionnellement à l'échantillon-maître national, ses unités sont affectées d'une pondération initiale de tirage. La réunion EM et EMEX à l'intérieur d'une région donne lieu à une repondération des unités primaires par partage des poids initiaux des deux échantillons.

L'échantillon régional est alors sélectionné à deux degrés, en prenant au premier degré l'ensemble des unités de l'EMEX et celles de l'intersection entre l'échantillon-maître national et la région. L'effectif de logements sélectionné est égal à la somme de la taille de l'échantillon résultant de l'allocation par

région du tirage national, et de l'extension financée par les partenaires locaux. Le tirage national et celui de l'extension sont simultanés et intégrés dans un plan de sondage unique.
Sauf la Corse, les trois autres régions citées plus haut ont réalisé une extension régionale par tirage dans l'EMEX.

Figure 1. Schéma de tirage dans l'échantillon-maître et l'EMEX

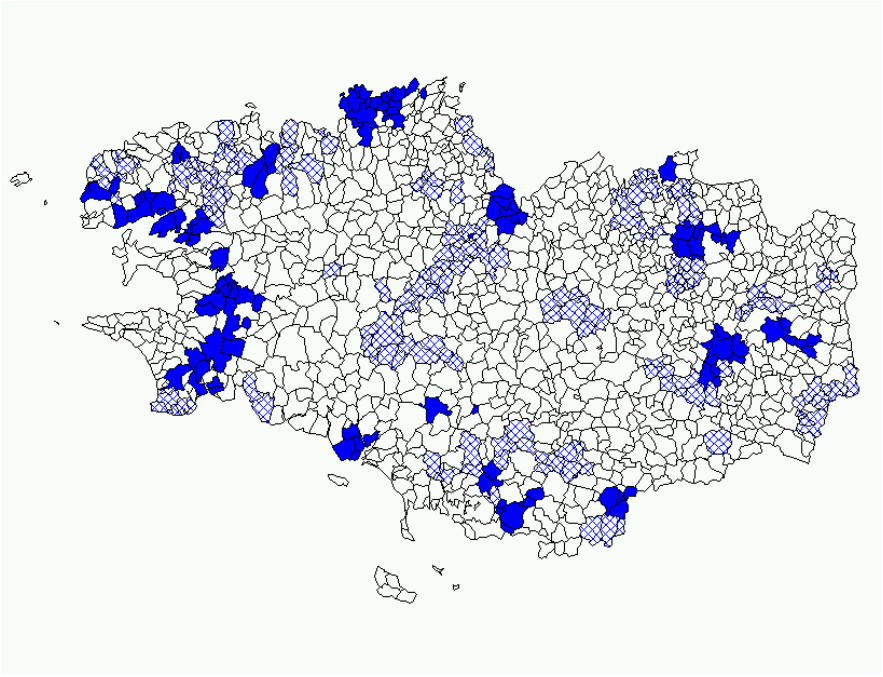


{U1, U2, U3} = Echantillon-maître national ∩ région R

{M1, M2, M3} = EMEX de la région R

{sn1, sn2, sn3, sr1, sr2, sr3} = échantillon complet de logements de la région R

Figure 2. L'échantillon-maître et l'EMEX 1999 en Bretagne



communes de l'échantillon-maître
communes de l'EMEX

1.3. Les extensions locales

Des échantillons complémentaires ont été tirés dans 12 zones centrées sur les plus grandes agglomérations des quatre régions déjà citées : 6 en Bretagne, 3 en PACA, 2 en Corse et une en Midi-Pyrénées. Chaque zone constituait un premier niveau de stratification géographique. Ces 12 échantillons sont donc indépendants.

Dans chaque zone, les logements ont été sélectionnés par sondage direct à un degré, dans deux bases de sondage disjointes : le recensement de 1999 et la base de logements « neufs » (BLN) construits après le dernier RP. La BLN est elle-même constituée à partir des permis de construire. Elle inclut les logements déclarés achevés entre mars 1999 et fin mars 2006 et enregistrés dans la base de données Sitadel du Ministère de l'Equipement.

Les échantillons tirés dans le recensement de 1999 ont été stratifiés selon des critères géographiques et socio-économiques. Les plus grandes agglomérations concernées ont été découpées en 3 à 6 groupes de quartiers en fonction du revenu médian des quartiers IRIS2000. Cette première typologie a été croisée avec un critère variable selon les régions : taille de logement en Bretagne, statut d'occupation du logement en Corse, origine immigrée ou non du ménage en PACA, localisation à l'intérieur de l'agglomération toulousaine (pôle urbain/périphérie) en Midi-Pyrénées. En Bretagne, les zones avec extension étant définies par un périmètre plus large que celui des pôles urbains, on a également distingué ville-centre et périphérie.

Dans chaque strate, les logements ont été sélectionnés par tirage systématique équiprobable, dans un fichier ordonné sur des critères corrélés aux variables d'intérêt : appartenance au secteur subventionné (HLM), nombre de pièces du logement, année d'achèvement de l'immeuble.

Les logements neufs ont été stratifiés sur le type d'habitat : individuel ou collectif. Dans chaque strate, l'échantillon a été sélectionné par tirage systématique dans un fichier ordonné par commune, nombre de logements de l'immeuble et année de dépôt du permis de construire.

Les extensions locales ont été tirées conditionnellement à l'échantillon national. Elles avaient donc une intersection vide avec l'échantillon national principal et avec l'échantillon ZUS. Les bases de sondage de logements utilisées étaient en revanche les mêmes : RP 99 et BLN.

2. Repondération post-collecte

Le calage des résultats sur des informations propres à chaque zone avec extension locale impliquait une recherche spécifique. Pour des raisons de calendrier, le redressement aux fins d'exploitations locales a été dissocié du redressement national et a donné lieu à un deuxième jeu de pondérations dans les quatre régions concernées.

Le redressement comporte trois étapes :

- la fusion des échantillons par partage des poids
- la correction de la non-réponse totale
- le calage sur des informations auxiliaires

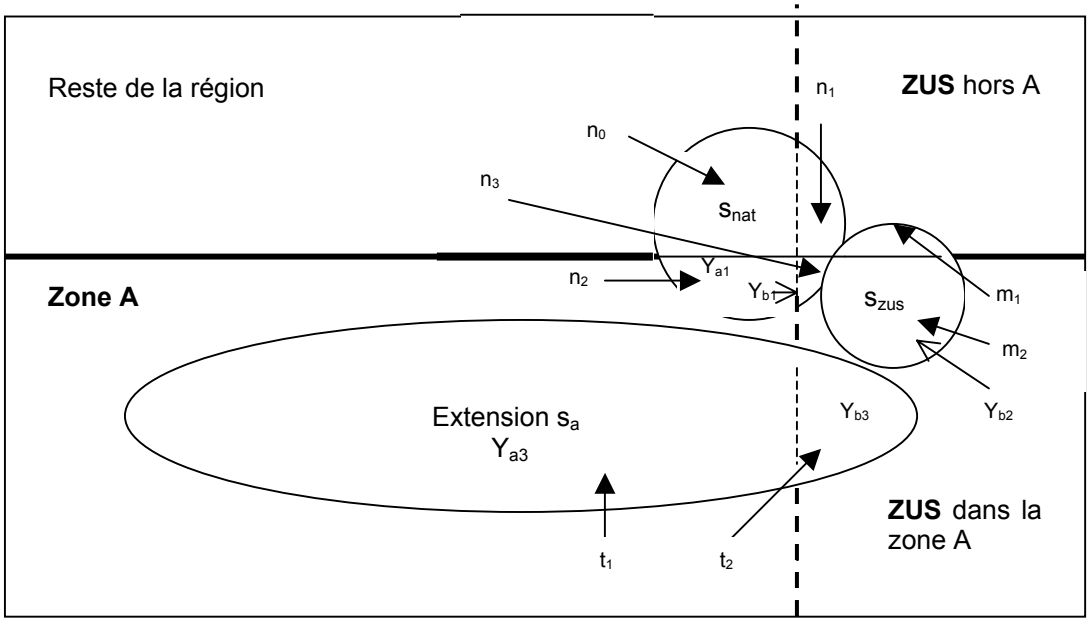
2.1. Fusion des échantillons pour une pondération unique

Dans une zone locale avec extension, l'échantillon initial complet comprenait :

1. des logements recensés en 1999, provenant du tirage national principal
2. des logements construits depuis 1999, provenant du tirage national principal (tirage BSLN)
3. des logements recensés en 1999 provenant du sous-échantillon ZUS
4. les logements recensés en 1999 de l'extension locale (tirage à un degré dans le RP99)
5. les logements construits depuis 1999, de l'extension locale (tirage à un degré dans la BLN)
6. des logements des sous-échantillons CNAF

L'absence de données de calage sur les allocataires, la très faible taille des sous-échantillons CNAF dans chaque zone concernée, l'absence d'enjeu de diffusion sur ce champ au niveau local ont conduit à abandonner l'échantillon CNAF dans les quatre régions avec extensions locales. Le plan de sondage dans chacune de ces régions peut alors être résumé par le schéma ci-dessous.

Figure 3. Schéma d'échantillonnage dans la région R



La zone A, incluse dans la région R, a fait l'objet d'une extension locale d'échantillon. Certains quartiers sont intégrés dans des « zones urbaines sensibles ». On trouve également des ZUS dans le reste de la région.

$\mathbf{s}_{nat}$, de taille $n = n_0 + n_1 + n_2 + n_3$, est l'intersection, avec la région R, de l'échantillon national tiré dans l'échantillon maître et l'Emex . Ses unités sont affectées des pondérations d_{1k}.

$\mathbf{s}_{zus}$, de taille $m = m_1 + m_2$, est l'intersection, avec la région R, de l'échantillon tiré dans les ZUS . Ses unités ont les pondérations d_{2k}.

$\mathbf{s}_a$, de taille $t = t_1 + t_2$, est l'extension tirée dans la zone A incluse dans la région R. Ses unités ont les pondérations d_{3k}.

Les échantillons $\mathbf{s}_{nat}$ et $\mathbf{s}_{zus}$ peuvent avoir une intersection non vide, alors que l'extension $\mathbf{s}_a$ a été sélectionnée conditionnellement aux précédents et leur est donc disjointe.

L 'intersection entre la zone A, objet d'une extension, et les ZUS, définit deux domaines : a (zone A hors ZUS) et b (les ZUS de la zone A). L'intersection du domaine avec chacun des sous-échantillons définit 5 sous-échantillons, fournissant chacun un estimateur sans biais du total Y_a ou Y_b d'une variable $\mathcal{Y}$ dans le domaine.

2.1.1. Application de la méthode de partage des poids au cas de bases de sondage multiples

B_1, … B_j … B_J sont J bases de sondage, non disjointes, dont la réunion recouvre totalement la population de référence U : $U = \bigcup_j B_j \qquad \bigcap_j B_j \neq \varnothing$

On a tiré un échantillon $\mathbf{s}_j$ dans chacune des bases B_j avec un plan de sondage $\mathbf{p}_j$. Les tirages sont indépendants et l'enquête a été réalisée dans la réunion des échantillons : $\mathbf{s} = \bigcup_{\mathbf{j}} \mathbf{s}_j$

On veut estimer le total *Y* d'une variable $\mathcal{Y}$ dans la population, à l'aide des observations de l'échantillon complet **s**, en attribuant un poids unique à chaque unité enquêtée.

Le lien entre une unité *k* de l'univers et une base de sondage B_j est codifié par la variable indicatrice $l_{j,k}$:

$$\begin{cases} l_{j,k} = 1 \Leftrightarrow k \in B_j \\ l_{j,k} = 0 \text{ sinon} \end{cases} \quad (1)$$

$$L_k = \sum_{j=1}^{J} l_{j,k} \quad (2)$$

L_k est le nombre total de liens entre une unité *k* et la base *B* ; c'est donc le nombre de bases de sondage auxquelles *k* est susceptible d'appartenir.

On opère le changement de variable suivant :

$$\forall k \in U \quad Z_{jk} = \frac{Y_k}{L_k} \times l_{j,k} \quad (3)$$

On montre que le total *Y* dans la région R peut s'écrire comme le total des nouvelles variables z_{jk} :

$$Z = \sum_{k \in U} \sum_{j=1}^{J} Z_{jk} = \sum_{k \in U} \sum_{j=1}^{J} \frac{Y_k}{L_k} l_{j,k} = \sum_{k \in U} \frac{Y_k}{L_k} \sum_{j=1}^{J} l_{j,k} = \sum_{k \in U} Y_k = Y$$

On substitue alors à l'estimation directe du total *Y* dans la population celle du total *Z* dans la réunion des bases de sondage, par la formule habituelle d'Horvitz-Thomson, et l'on obtient l'estimateur du partage des poids :

$$\hat{Y} = \hat{Z} = \sum_{k \in s} \sum_{j=1}^{J} \frac{z_{jk}}{\pi_{jk}} = \sum_{k \in s} \frac{y_k}{L_k} \sum_{j=1}^{J} \frac{l_{jk}}{\pi_{jk}} \quad (4)$$

où $\pi_{j,k} = \text{Prob}\{k \in \mathbf{s_j}\}$

Le poids de l'unité *k* après partage des poids est donc :

$$\boxed{d_k = \frac{1}{L_k} \sum_{j=1}^{J} \frac{1}{\pi_{jk}} l_{jk}} \quad (5)$$

Ici, le nombre de liens L_k selon la situation du logement dans la région R était le suivant :

- hors ZUS et hors zone avec extension : 1
- en ZUS et hors zone avec extension : 2
- hors ZUS et dans une zone avec extension : 2
- en ZUS et dans une zone avec extension : 3

Avec les notations de la figure 3, le partage des poids conduit aux poids suivants :

Zone	Sous-échantillons	Poids après partage des poids
$R \cap \overline{A} \cap \overline{Z} \overline{US}$	Échantillon national principal	$d_k = d_{1k}$ = poids initial
$R \cap \overline{A} \cap ZUS$	Échantillon national principal Echantillon ZUS	$d_k = \frac{1}{2} d_{1k} l_{1k} + \frac{1}{2} d_{2k} l_{2k}$
$R \cap A \cap \overline{Z} \overline{US}$	Échantillon national principal Extension locale	$d_k = \frac{1}{2} d_{1k} l_{1k} + \frac{1}{2} d_{3k} l_{3k}$
$R \cap A \cap ZUS$	Échantillon national principal Echantillon ZUS Extension locale	$d_k = \frac{1}{3} d_{1k} l_{1k} + \frac{1}{3} d_{2k} l_{2k} + \frac{1}{3} d_{3k} l_{3k}$

Un total Y dans la zone A est estimé par :

$$\hat{Y}_A = \frac{1}{2}\hat{Y}_{a1} + \frac{1}{3}\hat{Y}_{b1} + \frac{1}{2}\hat{Y}_{a3} + \frac{1}{3}\hat{Y}_{b2} + \frac{1}{3}\hat{Y}_{b3} \qquad (6)$$

où $\hat{Y}_{Dj}$ désigne l'estimateur Horvitz-Thomson dans le domaine D à partir de l'échantillon s_j avec les poids du plan de sondage p_j.

L'équation (4) peut aussi s'écrire :

$$\hat{Y} = \sum_{j=1}^{J} \sum_{k \in \mathbf{s_j}} \frac{y_k}{L_k \pi_{jk}} = \sum_{j=1}^{J} \hat{Z}_j \qquad (7)$$

Sous l'hypothèse d'indépendance des échantillons s_j, vérifiée ici entre l'échantillon national principal et l'échantillon ZUS, la variance de l'estimateur par partage des poids devient :

$$V\left(\hat{Y}\right) = V\left(\hat{Z}\right) = \sum_{j=1}^{J} V\left(\hat{Z}_j\right) \qquad (8)$$

2.1.2. Un partage des poids minimisant la variance : l' « estimateur composite »

Les solutions ci-dessus peuvent être appliquées dans tous les cas où la population cible est sondée au moyen d'un échantillonnage multiple. Cependant, si les plans de sondage p_j conduisent à des variances très inégales, ce qui se produit lorsque les tailles des échantillons s_j sont très différentes, l'estimateur (4) n'est pas optimal du point de vue de la précision.

Une autre approche consiste à rechercher des coefficients de pondération de chaque sous-échantillon qui minimisent la variance de l'estimateur.

Chaque échantillon s_j fournit un estimateur sans biais Horvitz-Thomson du total *Y* d'une variable d'intérêt $\mathcal{Y}$:

$$\hat{Y}_j = \sum_{k \in \mathbf{s}_j} \frac{y_k}{\pi_{jk}}$$

$$\pi_{jk} = \mathrm{Prob}\{k \in \mathbf{s}_j\}$$

$$E\left(\hat{Y}_j\right) = Y$$

L'estimateur composite est formé par une combinaison linéaire des $\hat{Y}_j$ issus de chaque échantillon qui minimise la variance de l'estimateur final $\widetilde{Y}$.

On recherche donc les coefficients λ_j vérifiant :

$$\widetilde{Y} = \sum_j \lambda_j \hat{Y}_j$$

$$E\left(\widetilde{Y}\right) = Y$$

$V\left(\widetilde{Y}\right)$ minimum

Espérance de $\widetilde{Y}$:

$$E\left(\widetilde{Y}\right) = \sum_j \lambda_j E\left(\hat{Y}_j\right) = \sum_j \lambda_j Y = Y \sum_j \lambda_j$$

Pour conserver un estimateur sans biais de *Y*, il suffit de prendre des coefficients λ_j vérifiant :

$$\sum_j \lambda_j = 1 \qquad (9)$$

Lorsque les échantillons sont indépendants, la variance de $\widetilde{Y}$ respecte exactement l'identité suivante :

$$V\left(\widetilde{Y}\right)=\sum_j \lambda_j^2 V\left(\hat{Y}_j\right) \qquad (10)$$

Dans le cas contraire, si les tailles d'échantillon sont très faibles devant celle de la population, on peut négliger les covariances et s'en tenir à l'expression (10) pour la suite des calculs.

Dans (10), $V\left(\widetilde{Y}\right)$ est une expression en λ_j. Elle atteint son minimum pour les valeurs de λ_j qui annulent ses dérivées partielles.

Par récurrence, on trouve que :

$$\lambda_j = \frac{\prod_{i\neq j} V\left(\hat{Y}_i\right)}{\sum_i \left(\prod_{l\neq i} V\left(\hat{Y}_l\right)\right)} \qquad (11)$$

Lorsque tous les échantillons ont été tirés avec un plan de sondage aléatoire simple, la résolution de (11) conduit aux coefficients :

$$\lambda_j = \frac{n_j}{n} \qquad (12)$$
$$\text{avec : } n=\sum_j n_j$$

Avec d'autres plans de sondage, l'optimum de précision n'est pas atteint avec des valeurs de coefficients proportionnelles aux tailles des échantillons fusionnés, mais celles-ci en constituent une approximation acceptable. C'est cette solution qui a été adoptée.

Avec les notations de la figure 3, l'estimateur composite dans une région avec une extension locale conduit aux poids suivants :

Zone	Sous-échantillons	Poids composite
$R \cap \overline{A} \cap \overline{Z} \overline{US}$	Échantillon national principal	$d_k = d_{1k}$ = poids initial
$R \cap \overline{A} \cap ZUS$	Échantillon national principal Echantillon ZUS	$d_k = \frac{n_1}{n_1+m_1} d_{1k} l_{1k} + \frac{m_1}{n_1+m_1} d_{2k} l_{2k}$
$R \cap A \cap \overline{Z} \overline{US}$	Échantillon national principal Extension locale	$d_k = \frac{n_2}{n_2+t_1} d_{1k} l_{1k} + \frac{t_1}{n_2+t_1} d_{3k} l_{3k}$
$R \cap A \cap ZUS$	Échantillon national principal Echantillon ZUS Extension locale	$d_k = \frac{n_3}{n_3+m_2+t_2} d_{1k} l_{1k} + \frac{m_2}{n_3+m_2+t_2} d_{2k} l_{2k} + \frac{t_2}{n_3+m_2+t_2} d_{3k} l_{3k}$

Si *d* est l'indice d'un domaine ayant une intersection non vide avec chacun des échantillons sélectionnés dans *J* bases de sondage, le partage des poids conduit à une combinaison linéaire des estimateurs $\hat{Y}_j^d$ formée avec des coefficients λ_j égaux à $\frac{1}{J}$.

2.1.3. Comparaison des résultats avec les deux estimateurs

Tableau 1. Effectifs total logements estimés avant et après partage des poids

aire	Estimateurs par échantillon				Estimateurs après partage des poids			
					Valeurs		Biais d'estimation[1] (%)	
	Base de sondage rp99	échantillon national	extension locale	échantillon ZUS[2]	mgpp	estimateur composite	mgpp	estimateur composite
Brest	142 304	157 718	142 304	*3 973*	149 979	143 902	5,4	1,1
Lorient	89 439	56 723	89 439	*4 584*	73 297	88 116	-18,0	-1,5
Quimper	62 696	67 791	62 696	*3 056*	65 285	63 065	4,1	0,6
Rennes	231 978	231 724	231 978	*20 781*	232 177	232 160	0,1	0,1
Saint-Brieuc	56 063	53 956	56 063	*6 112*	54 603	55 919	-2,6	-0,3
Vannes	54 791	80 242	54 791	*4 584*	67 503	57 553	23,2	5,0
reste de la région	855 346	729 677		*3 973*	728 897	728 463	-14,8	-14,8
Bretagne	1 492 617	1 377 832		*47 063*	1 371 741	1 369 179	-8,1	-8,3
Toulouse	375 926	369 709	375 926	*20 170*	373 827	375 391	-0,6	-0,1
reste de la région	943 668	858 608		*10 085*	859 697	860 370	-8,9	-8,8
Midi-Pyrénées	1 319 594	1 228 318		*30 254*	1 233 523	1 235 762	-6,5	-6,4
Avignon	134 484	147 813	134 484	*9 168*	141 481	137 038	5,2	1,9
Marseille	829 219	846 690	829 219	*118 878*	837 999	846 521	1,1	2,1
Nice	643 652	682 308	643 652	*25 670*	662 464	669 978	2,9	4,1
reste de la région	929 676	961 746		*26 893*	960 788	960 428	3,3	3,3
PACA[3]	2 521 655	2 630 567		*180 610*	2 591 147	2 600 043	2,8	3,1
Ajaccio	41 863	31 987	41 863	*9 168*	36 905	41 747	-11,8	-0,3
Bastia	30 566	23 990	30 566	*4 584*	26 963	30 473	-11,8	-0,3
reste de la région	104 937	115 918			115 918	115 918	10,5	10,5
Corse	177 366	171 896		*13 752*	179 786	188 139	1,4	6,1

[1] Ecart relatif entre l'estimateur après partage des poids et l'effectif dans la population de référence (colonne 1).
[2] Il s'agit, par construction, de l'effectif estimé sur le seul champ des ZUS. Il n'est donc pas directement comparable aux autres estimateurs, qui représentent l'ensemble de la zone.
[3] Le total régional diffère de la somme des quatre zones indiquées, car celle d'Avignon inclut des communes situées dans le département du Gard, en région Languedoc-Roussillon.

Tableau 2. Effectifs de résidences principales estimés avant et après partage des poids

aire	Base de sondage rp99	Estimateurs par échantillon			Estimateurs après partage des poids			
		échantillon national	extension	échantillon ZUS[5]	Valeurs		Biais d'estimation[4] (%)	
					mgpp	estimateur composite	mgpp	estimateur composite
Brest	126 372	134 199	125 899	*3 973*	130 017	126 754	2,9	0,3
Lorient	78 704	49 806	78 704	*4 584*	64 503	77 580	-18,0	-1,4
Quimper	51 319	48 422	51 319	*2 750*	49 878	51 121	-2,8	-0,4
Rennes	213 653	215 824	213 653	*20 170*	214 906	214 005	0,6	0,2
Saint-Brieuc	50 413	48 422	50 413	*5 806*	49 289	50 349	-2,2	-0,1
Vannes	48 148	74 708	48 148	*4 278*	61 380	51 030	27,5	6,0
reste de la région	641 059	574 148		*3 362*	573 062	572 459	-10,6	-10,7
Bretagne	1 209 668	1 145 530		*44 923*	1 143 034	1 143 298	-5,5	-5,5
Toulouse	341 783	336 099	337 826	*18 336*	338 318	339 147	-1,0	-0,8
reste de la région	728 989	640 277		*9 168*	641 401	642 097	-12,0	-11,9
Midi-Pyrénées	1 070 772	976 376		*27 504*	979 719	981 244	-8,5	-8,4
Avignon	118 634	133 553	120 026	*8 251*	127 174	122 571	7,2	3,3
Marseille	734 808	747 671	739 500	*106 349*	744 108	750 787	1,3	2,2
Nice	450 654	478 281	465 852	*22 003*	471 942	474 467	4,7	5,3
reste de la région	606 154	623 720		*22 003*	623 400	623 280	2,8	2,8
PACA[6]	1 896 302	1 975 805		*158 606*	1 956 026	1 958 439	3,1	3,3
Ajaccio	30 052	20 792	29 402	*7 334*	25 327	29 311	-15,7	-2,5
Bastia	25 818	23 990	25 497	*4 584*	24 457	25 490	-5,3	-1,3
reste de la région	50 366	59 177		*0*	59 177	59 177	17,5	17,5
Corse	106 236	103 959		*11 918*	108 960	113 977	2,6	7,3

[4] Ecart relatif entre l'estimateur après partage des poids et l'effectif dans la population de référence (colonne 1).
[5] Il s'agit, par construction, de l'effectif estimé sur le seul champ des ZUS. Il n'est donc pas directement comparable aux autres estimateurs, qui représentent l'ensemble de la zone.
[6] Le total régional diffère de la somme des quatre zones indiquées, car celle d'Avignon inclut des communes situées dans le département du Gard, en région Languedoc-Roussillon.

Tableau 3. Effectifs de logements HLM estimés avant et après partage des poids

aire	Base de sondage rp99	Estimateurs par échantillon			Estimateurs après partage des poids			
		échantillon national	extension locale	échantillon ZUS[8]	Valeurs		Biais d'estimation[7] (%)	
					mgpp	estimateur composite	mgpp	estimateur composite
Brest	16 938	16 602	16 661	*3 667*	16 535	16 613	-2,4	-1,9
Lorient	13 514	12 451	13 399	*3 362*	13 340	13 458	-1,3	-0,4
Quimper	5 423	8 301	5 564	*1 834*	6 996	5 756	29,0	6,1
Rennes	31404	24 903	31 447	*11 918*	27 605	30 457	-12,1	-3,0
Saint-Brieuc	6 490	9 684	6 502	*2 750*	8 030	6 606	23,7	1,8
Vannes	6 183	11 068	6 478	*2 445*	8 646	6 891	39,8	11,5
reste de la région	51 502	44 272		*2750*	43 572	43 183	-15,4	-16,2
Bretagne	131 454	127 281		*28 726*	124 724	122 964	-5,1	-6,5
Toulouse (pu+si)	44 402	41 518	41 022	*11 613*	42 362	42 746	-4,6	-3,7
reste de la région	47 033	46 461		*6 418*	47 198	47 655	0,4	1,3
Midi-Pyrénées	91 435	87 979		*18 030*	89 560	90 400	-2,1	-1,1
Avignon	17 929	22 259	17 901	*6 112*	20 466	18 974	14,2	5,8
Marseille	116 083	112 436	113 265	*48 896*	113 172	115 461	-2,5	-0,5
Nice	37 161	30 249	41 039	*9 474*	36 055	34 009	-3,0	-8,5
reste de la région	57 904	68 376		*10 390*	68 302	68 274		
PACA[9]	228 262	232 749		*74 872*	237 167	235 724	3,9	3,3
Ajaccio	3 059	1 599	3 037	*1 222*	2 202	2 980	-28,0	-2,6
Bastia	4 516	6 397	4 514	*3 667*	5 084	4 588	12,6	1,6
Total	*7 575*	*7 997*	*7 551*	*4 890*	*7 285*	*7 568*	-3,8	-0,1

[7] Ecart relatif entre l'estimateur après partage des poids et l'effectif dans la population de référence (colonne 1).

[8] Il s'agit, par construction, de l'effectif estimé sur le seul champ des ZUS. Il n'est donc pas directement comparable aux autres estimateurs, qui représentent l'ensemble de la zone.

[9] Le total régional diffère de la somme des quatre zones indiquées, car celle d'Avignon inclut des communes situées dans le département du Gard, en région Languedoc-Roussillon.

2.2. Calage de l'échantillon local sur une information auxiliaire

Une fois repondéré par partage des poids, chaque échantillon local a été calé sur les totaux, dans la zone, de variables auxiliaires.

2.2.1. Pourquoi un calage spécifique par zone avec extension d'échantillon

Le calage d'un échantillon sur des totaux connus dans la population a pour but de rendre cohérentes les estimations obtenues d'une même variable entre différentes sources statistiques tout en améliorant la précision de la variabe d'intérêt estimée par l'enquête.

Rappelons les propriétés de l'estimateur d'un total après calage.

Soient : U la population de référence
Y le total dans U de la variable d'intérêt $\mathcal{Y}$
$\hat{Y}$ un estimateur de Y
$\mathbf{X}$ le vecteur des totaux, dans U, de variables auxiliaires
π_k la probabilité d'inclusion de l'unité *k* dans l'échantillon s

L'estimateur après calage est asymptotiquement égal à l'estimateur redressé par régression sur les variables de calage X :

$$\hat{Y}_{cal} \approx \hat{Y}_{reg} = \hat{Y}_{HT} + \hat{\mathbf{b}}'\left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right) = Y + \hat{U}_{HT}$$

avec : $\hat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{Y}$$

$$\mathbf{D} = Diag\left(\frac{1}{\pi_k}\right)$$

$$\hat{U}_{HT} = \sum_{k \in s} \frac{y_k - \hat{\mathbf{b}}'\mathbf{x_k}}{\pi_k} \qquad (1)$$

Sa variance est celle des résidus de la régression de $\mathbf{Y}$ sur $\mathbf{X}$:

$$V\left(\hat{Y}_{cal}\right) \approx V\left(\hat{U}_{HT}\right) \qquad (2)$$

Que devient l'estimateur du total Y_A de la variable d'intérêt dans la sous-population U_A résidant dans la zone A avec extension d'échantillon, après calage de l'échantillon national sur des totaux France entière ? On est ici dans l'estimation à l'intérieur d'un domaine, c'est-à-dire d'une sous-population dont les contours ne sont pas définis par les paramètres du plan de sondage. L'expression en a été établie par P. Ardilly.

Le total Y_A se décompose comme le total France entière de la variable Z_k, produit de Y_k par l'indicatrice d'appartenance à la zone A.

$$Z_k = Y_k \times \varepsilon_{A,k}$$
$$\varepsilon_{A,k} = 1 \Leftrightarrow k \in U_A$$
$$\varepsilon_{A,k} = 0 \Leftrightarrow k \notin U_A$$
$$Y_A = \sum_{k \in U_A} Y_k = \sum_{k \in U} Z_k$$

Sans calage à un niveau local, l'estimateur d'un total dans la zone A devient, après calage national :

$$\hat{Y}_A = \hat{Z}_{cal} = \hat{Z}_{HT} + \hat{\mathbf{b}}'\left(\mathbf{X} - \hat{\mathbf{X}}_{\mathbf{HT}}\right) = Z + \hat{U}_{HT} = Y_A + \hat{U}_{HT}$$

$$\hat{U}_{HT} = \sum_{k \in s} \frac{z_k - \hat{\mathbf{b}}'\mathbf{x}_\mathbf{k}}{\pi_k} = \sum_{k \in s \cap U_A} \frac{y_k - \hat{\mathbf{b}}'\mathbf{x}_\mathbf{k}}{\pi_k} + \sum_{k \in s \cap U_{\bar{A}}} \frac{-\hat{\mathbf{b}}'\mathbf{x}_\mathbf{k}}{\pi_k} \qquad (3)$$

où U_A désigne la sous-population de la zone A et $U_{\bar{A}}$ la partie de la population résidant en-dehors de la zone A.

Il s'ensuit que la somme des résidus de la régression des Z_k sur les variables de calage incorpore une somme de résidus calculés sur les unités situées hors de la zone A, comme l'indique l'expression (3). Après calage sur des totaux nationaux, la variance de $\hat{Y}_A$ pourrait donc être supérieure à celle de l'estimateur direct Horvitz-Thomson dans la zone, sans prise en compte d'une information auxiliaire spécifique au domaine concerné par une extension.

La seule solution pour l'éviter est de caler les échantillons dans les zones avec extension sur une information locale propre à ces zones.

2.2.2. Les sources d'information : bases de sondage ou sources externes

Un calage sur les totaux des bases de sondage permet de corriger l'aléa d'échantillonnage sur les variables présentes dans ces bases.

Le recensement de 1999 contient une information riche sur le logement. Les caractéristiques des logements étant par nature stables dans le temps, les distributions de 1999 et de 2006 restent nécessairement corrélées, pour les logements existant aux deux dates. Un calage sur les effectifs de logements en 1999 par modalité conserve par conséquent toute sa pertinence.

Un tel redressement ne pouvait s'appliquer qu'à une fraction (77 %) de l'échantillon : les logements sélectionnés dans le recensement de 1999, et résidences principales à cette date. Les autres logements (constructions neuves, logements vacants ou résidences secondaires en 1999) ne peuvent être calés que sur leur effectif total dans la base de sondage. On ne contrôle pas, dans cette repondération, l'égalité stricte du parc total de logements estimé avec son effectif en 2006, connu par ailleurs. Enfin, les bases de sondage utilisées ne contenant aucune information sur les habitants des logements en 2006, elles ne permettent pas un calage de l'échantillon d'individus sur la population au moment de l'enquête.

D'autres sources statistiques donnent une estimation annuelle du parc de logements, à divers niveaux géographiques.

Le parc total de logements est estimé chaque année au niveau national dans le compte satellite du logement. Les premières enquêtes annuelles de recensement (EAR), de 2004 à 2006, fournissent une information sur la répartition territoriale du parc de logements. La moyenne des résultats annuels 2004 à 2006, complétée par une modélisation des petites communes non encore recensées, a fourni une estimation localisée en milieu de période, soit en février 2005. Celle-ci a été projetée au 1er janvier 2006 en s'appuyant sur les évolutions observées entre 2005 et 2006 dans les déclarations de fiscalité locale. Ces trois sources : compte satellite du logement, enquêtes annuelles de recensement et déclarations de taxe d'habitation, ont été mises à profit pour estimer l'effectif total de logements dans chaque zone concernée par une extension d'échantillon.

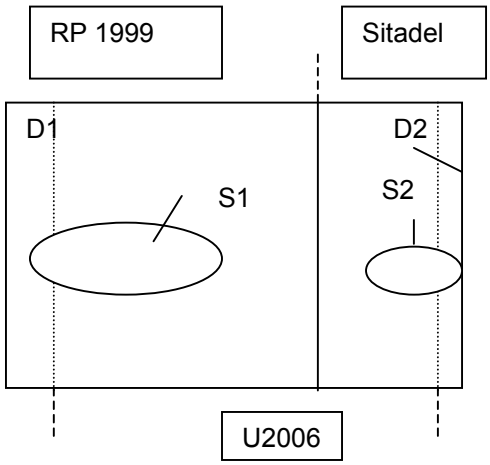
Le calage de l'échantillon d'individus sur un effectif de population au 1er janvier 2006 nécessitait le recours à une source externe. Les données de cadrage ont été prises dans les estimations annuelles de population par région. Les populations par zone à l'intérieur de chaque région ont été projetées sur 2006 à partir des structures territoriales de 2005 observées dans les enquêtes annuelles de recensement.

Les sources externes à l'enquête logement fournissent un chiffre de population totale, nombre de logements et nombre de personnes, en 2006, mais peu d'informations structurelles. Le recensement annuel a l'inconvénient de reposer sur des définitions non nécessairement identiques aux concepts du questionnaire de l'ENL. Le protocole de collecte diffère entre les des deux enquêtes. Le recensement recueille la déclaration spontanée de l'habitant à une question formulée de la façon la plus générale. Le même concept (nombre de pièces du logement par exemple) est cerné au travers d'une progression de questions posées par un enqêteur dans l'enquête logement. On ne peut donc pas caler les caractéristiques du logement observées dans l'enquête logement sur les effectifs correspondants dans le recensement par risque de non homogénéité des deux termes de l'équation.

2.2.3. La solution retenue

Le parc de logements 2006 peut être schématisé comme ci-dessous. L'échantillon s1 est tiré dans l'ensemble des logements recensés en 1999, l'échantillon s2 dans celui des logements construits depuis 1999. Certains logements disparaissent depuis la mise à jour de la base de sondage (D1 et D2) : immeubles détruits, locaux perdant leur usage d'habitation. Le parc de logements en 2006 (U2006) est constitué des logements recensés en 1999 et existant toujours en 2006 et des constructions neuves en usage au moment de l'enquête.

Figure 4. Des bases de sondage
à la population 2006



La solution idéale aurait consisté à caler simultanément sur les informations en provenance des bases de sondage et sur les effectifs en 2006 fournis par les sources externes.

Les équations de calage auraient été de la forme :

$$\sum_{s1\cap\overline{D}1} w_k x_{1k} + \sum_{s1\cap D1} w_l x_{1l} = X_{1,RP}$$
$$\sum_{s2\cap\overline{D}2} w_i x_{2i} + \sum_{s2\cap D2} w_j x_{2j} = X_{2,Sit}$$
$$\sum_{s1\cap\overline{D}1} w_k x_{3k} + \sum_{s2\cap\overline{D}2} w_i x_{3i} = X_{3,2006}$$

où w_k est le poids de calage de l'unité *k*, X_1 est une variable présente dans le RP de 1999, X_2 une variable présente dans la base de sondage Sitadel, X_3 une variable dont le total est connu dans la population de 2006.

Pour que ces équations soient compatibles, les effectifs N_j des bases de sondage doivent respecter la condition évidente :

$$N_{RP} + N_{Sit} = N_{2006} + N_{disparus} > N_{2006}$$

La diffusion progressive des premiers résultats du recensement rénové a mis en lumière le caractère incomplet de la base de logements neufs utilisée en complément du dernier recensement général pour l'échantillonnage. La somme des effectifs totaux des deux bases était en effet inférieure, ou à peine supérieure, au parc de logements estimé en janvier 2005 par les trois premières enquêtes annuelles de recensement, dans les régions concernées par des extensions. Les retards de déclaration d'achèvement de travaux à l'administration par les propriétaires explique la sous-estimation du nombre de logements neufs dans la base de sondage utilisée.

La solution retenue a consisté à effectuer successivement un double calage :

- l'échantillon complet de logements répondants, incluant les logements disparus, a été calé sur les effectifs des bases de sondage, pour redresser l'aléa d'échantillonnage
- l'échantillon repondéré, sans les logements hors champ, a ensuite été calé sur le parc de logements et la population de 2006.

2.2.3.1. Pré-calage sur les totaux des bases de sondage

Le calage sur les effectifs des bases de sondage distingue trois sous-ensembles, redressés sur des marges distinctes :

- les logements résidences principales en 1999, quel que soit leur usage en 2006
- les logements vacants, résidences secondaires ou occasionnelles en 1999
- les logements construits depuis 1999

Chacun de ces sous-ensembles est calé sur l'effectif correspondant dans la base de sondage, RP de 1999 ou base des permis de construire.

Schéma de pré-calage sur les totaux des bases de sondage

Situation du logement en 2006	Catégorie du logement en 1999			Logement construit depuis 1999
	Résidence principale	Résidence secondaire ou occasionnelle	Logement vacant	
Résidence principale				
Résidence secondaire ou occasionnelle				
Logement vacant				
Logement disparu				
Total de la base de sondage	↓ N1	↓ N2	↓ N3	↓ N4

Tableau 4. Les variables de calage sur les totaux des bases de sondage

Catégorie de logement	Variables de calage
Résidences principales en 1999	type d'habitat (individuel / collectif) type de propriété : hlm / autre taille du logement : nombre de pièces ou surface statut d'occupation : propriétaire / locataire nombre de personnes habitant le logement en 1999 mode principal de chauffage strate de l'extension locale
Résidences secondaires, occasionnelles, vacantes en 1999	catégorie de logement en 1999 type d'habitat
Logements neufs	strate de l'extension locale (croisement : quartier, type d'habitat)

L'exploitation de l'enquête exclut ensuite de l'échantillon repondéré les unités hors champ : logements disparus, vacants, résidences secondaires en 2006. A l'issue de ce calage, on obtient une première estimation de l'effectif de résidences principales en 2006.

2.2.3.2. Calage sur les effectifs 2006

Schéma du calage final sur l'effectif en 2006

Situation du logement en 2006	Catégorie du logement en 1999			Logement construit depuis 1999	Total en 2006
	Résidence principale	Résidence secondaire ou occasionnelle	Logement vacant		
Résidence principale					→ M1
Résidence secondaire, occasionnelle, logement vacant					→ M2

L'échantillon de logements et celui des individus ont été repondérés par un calage simultané sur les effectifs en 2006 fournis par les sources externes indiquées ci-dessus. Outre le parc total de logements par mode d'usage (résidences principales, secondaires, logements vacants), on a également utilisé sa répartition par type d'habitat (appartement dans un immeuble collectif ou autre type de logement) et taille du ménage habitant le logement (ménages d'une personne, ménages de plusieurs personnes). Ces deux caractéristiques, très corrélées au comportement de réponse, sont homogènes entre le recensement annuel et l'enquête logement.

L'utilisation de la version 2 de CALMAR a permis de caler simultanément sur les totaux de la région et de chacune des zones concernées par une extension locale, en utillisant l'option des matrices inverses généralisées du logiciel pour s'affranchir des colinéarités induites.

2.3. Correction de la non-réponse totale

2.3.1. La recherche d'un modèle de non-réponse

L'information auxiliaire utile à l'analyse des facteurs influençant le comportement de réponse n'est disponible que dans la base du recensement de population. Seule la partie de l'échantillon sélectionnée dans le RGP de 1999 pouvait donc être redressée autrement que de façon uniforme. Par ailleurs, les logements non habités de façon permanente en 1999 (résidences secondaires, logements vacants) échappent également à l'analyse. Celle-ci ne porte donc que sur les logements recensés en 1999, et résidences principales à cette date, soit 82% de l'échantillon des résidences principales en 2006.

La recherche des facteurs corrélés au comportement de réponse a été menée séparément dans chacune des quatre régions, mais a conduit à des modèles de réponse proches d'une région à l'autre. Le type de logement (individuel ou collectif), la taille du logement, l'âge ou le type d'activité (actif occupé, retraité, étudiant) de la personne de référence occupant le logement en 1999 sont les principaux critères de non-réponse. On répond d'autant moins qu'on habite un petit appartement dans un immeuble collectif, et que l'on est très jeune ou très âgé, étudiant ou retraité. Ces deux critères apparaissent malgré l'ancienneté de l'information, en raison d'une permanence du type d'habitat occupé respectivement par les étudiants ou les personnes âgées. Le type de logement traduit l'accessibilité du logement, les digicodes ne concernant que l'habitat collectif. La taille du logement est corrélée au nombre de personnes qui l'habitent. Elle exprime à la fois le degré d'accessibilité du logement, et la plus grande difficulté à joindre les ménages d'une personne.

La seule information utilisable pour corriger la non-réponse des logements construits depuis 1999 était le type de construction : maison individuelle ou immeuble collectif.

Il en était de même des logements recensés en 1999 mais résidences secondaires ou vacantes à l'époque. On disposait également, pour ces logements, du type de propriétaire : HLM ou non.

2.3.2. La technique de redressement

Deux possibilités étaient envisageables : redresser la non-réponse avant le pré-calage sur les totaux des bases de sondage, ou repondérer pour non-réponse après le pré-calage et avant le calage final sur les effectifs de 2006. C'est la première solution qui a été choisie, afin de minimiser l'écart entre l'estimateur intermédiaire obtenu avec les poids de pré-calage et l'effectif ciblé, et le risque de déformation des structures sur les variables du pré-calage par le redressement final.

Trois techniques ont été testées :

- repondération pour non-réponse avant pré-calage, par groupe homogène de réponse (GHR)
- redressement de la non-réponse directement par calage classique
- redressement de la non-réponse par calage généralisé

2.3.2.1. Repondération par groupe homogène de réponse

Dans le processus d'identification des non-répondants, on fait l'hypothèse que le mode d'usage du logement au moment de l'enquête : résidence principale, secondaire, occasionnelle ou logement vacant, est déterminé sans erreur. Il ne peut donc y avoir de non-répondants que parmi les logements résidences principales en 2006.

Les groupes de réponse sont définis par croisement des modalités des facteurs cités ci-dessus, à l'intérieur de chaque zone avec extension d'échantillon. Les logements répondants sont repondérés de façon uniforme par l'inverse du taux de réponse dans chaque groupe :

$$d_k^* = d_k \times \frac{n_{g(k)}}{r_{g(k)}}$$

où d_k est le poids de l'unité k après partage des poids, *g(k)* le groupe de réponse auquel appartient l'unité k, $n_{g(k)}$ l'effectif initial de l'échantillon et $r_{g(k)}$ l'effectif de répondants dans le groupe de réponse *g(k)*.

Le pré-calage sur les totaux des bases de sondage est ensuite réalisé en spécifiant d_k^* en poids initial en entrée de CALMAR, ce qui conduit aux équations de calage :

$$\sum_{k\in sr} d_k^* F(\lambda' x_k) x_k = \sum_{k\in sr} \frac{d_k}{\hat{p}_k} F(\lambda' x_k) x_k = \sum_{k\in U} X_k = X$$

avec :

$$\hat{p}_k = \text{Prob}\{k \in \text{sr}/k \in \text{s}\}$$

s étant l'échantillon initial et sr celui des répondants.

2.3.2.2. Redressement de la non-réponse par calage direct

Aucune repondération pour non-réponse n'est effectuée avant pré-calage. Les variables identifiées ci-dessus comme facteurs de non-réponse sont introduites en variables de calage et celui-ci est effectué avec, en poids initial, le poids d_k issu du partage des poids.
On résout les équations de calage de la forme :

$$\sum_{k\in sr} d_k F(\mu' x_k) x_k = \sum_{k\in U} X_k = X$$

2.3.2.3. Redressement de la non-réponse par calage généralisé

Comme dans le cas précédent, l'échantillon n'est pas repondéré pour non-réponse avant pré-calage. Les mêmes variables corrélées au comportement de réponse sont spécifiées en variables de calage. On introduit dans le calage un deuxième vecteur de paramètres $\mathbf{z_k}$, de même dimension que le vecteur $\mathbf{x_k}$ des variables de calage, contenant des variables explicatives de la non-réponse, mais observées sur les seuls logements répondants. Ces variables instrumentales sont choisies de façon à corriger l'ancienneté de la base de sondage, source de l'information auxiliaire utilisée pour expliquer la non-réponse. On prend les variables de même définition que celles identifiées comme facteurs de non-réponse dans la base de sondage, mais avec leurs valeurs dans le questionnaire d'enquête, c'est-à-dire en 2006.

Les équations de calage sont alors de la forme :

$$\sum_{k\in sr} d_k F(\lambda' \mathbf{z_k}) \mathbf{x_k} = \sum_{k\in U} X_k = X$$

où : d_k est la pondération après partage des poids.

Le nombre de personnes habitant le logement était le facteur de réponse le plus déterminant et le plus susceptible de modification entre 1999 et 2006. C'est donc la variable prioritairement introduite en variable instrumentale.

La catégorie du logement (résidence principale ou non) peut avoir changé entre la constitution de la base de sondage et l'enquête : il faut en tenir compte pour ne pas repondérer abusivement pour non-réponse des logements inoccupés en 2006. Son introduction en variable instrumentale conduit à corriger le poids initial en entrée de Calmar par une probabilité du logement à être une résidence principale ou un logement vacant en 2006.

Le vecteur Z de variables instrumentales comprenait donc les indicatrices d'appartenance aux modalités suivantes :

- Z_{11} =résidence principale en 2006
- Z_{12} =logement vacant, secondaire ou occasionnel en 2006
- Z_{13} =logement disparu depuis 1999
- Z_{21} =ménage d'une personne en 2006
- Z_{22} =ménage de deux personnes en 2006
- Z_{23} =ménage de trois personnes en 2006
- Z_{24} =ménage de quatre personnes ou plus en 2006

Pour satisfaire la contrainte d'égalité de dimension des vecteurs $\mathbf{x_k}$ et $\mathbf{z_k}$, le vecteur $\mathbf{z_k}$ est complété par les autres variables de calage figurant dans $\mathbf{x_k}$, avec les mêmes valeurs.

Les logements neufs ne peuvent être calés que sur leur effectif dans la base de sondage par type de construction : maisons individuelles ou immeubles collectifs, seule information disponible. Pour ce sous-échantillon, on pouvait utiliser en variable instrumentale la taille du ménage en deux modalités : ménage d'une personne et autre ménage.

A titre d'exemple, dans la région Midi-Pyrénées, où une extension d'échantillon avait été réalisée dans le pôle urbain de Toulouse, les vecteurs de calage étaient les suivants, où $1_{\in P}$ est l'indicatrice d'appartenance à la sous-population P :

$$X = \begin{bmatrix}
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ rés.principale en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ rés. secondaire en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logement vacant en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage d'une personne en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 2 personnes en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 3 personnes en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 4 personnes ou plus en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage propriétaire en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ appartement dans un immeuble collectif} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ HLM en 1999} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logement d'une pièce} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 2 pièces} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 3 pièces} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 4 pièces} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 5 pièces ou plus} \\
1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ strate de tirage}
\end{bmatrix}$$

$$Z = \begin{bmatrix} 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ rés.principale en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ rés. secondaire en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logement vacant en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage d'une personne en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 2 personnes en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 3 personnes en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage de 4 personnes ou plus en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ ménage propriétaire en 2006} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ appartement dans un immeuble collectif} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ HLM en 1999} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logement d'une pièce} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 2 pièces} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 3 pièces} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 4 pièces} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ logt de 5 pièces ou plus} \\ 1_{\in \text{ pôle urbain}} \times 1_{\in} \text{ strate de tirage} \end{bmatrix}$$

2.3.2.4. Résultats comparés

Le pré-calage sur les totaux des bases de sondage conduit à un effectif total de logements inférieur à celui attendu en raison d'un défaut de couverture du champ par la base de sondage des logements neufs.

Avec une repondération de la non-réponse par groupe homogène avant calage, les taux de résidence principale obtenus sont proches des taux estimés à partir des enquêtes annuelles de recensement : légèrement sous-estimés en Bretagne et en Midi-Pyrénées, légèrement surestimés en PACA. La taille moyenne des ménages est en général surestimée, particulièrement sur la région Bretagne, où l'on surestime également la population totale.

Le redressement de la non-réponse directement par pré-calage simple a pour effet de sous-estimer le nombre de logements résidences principales en 2006, et donc le taux de résidences principales, par rapport à un calage après repondération pour non-réponse. Dans le calage direct, tout se passe comme si l'on repondérait également pour non-réponse les logements devenus résidences secondaires ou vacants en 2006.

L'introduction de la catégorie du logement en 2006 en variable instrumentale, peut avoir pour effet de corriger la sous estimation du nombre de résidences principales par rapport à un calage direct simple. Cette méthode était plus efficace en Midi-Pyrénées et en région PACA hors zones de Marseille et Nice, qu'en Bretagne, où la corrélation entre les vecteurs $\mathbf{x_k}$ et $\mathbf{z_k}$ était moins assurée.

L'introduction de la taille du ménage en 2006 en variable instrumentale améliore l'estimation du nombre moyen de personnes par ménage dans quelques zones où le calage classique la surestimait. C'est le cas en Midi-Pyrénées et dans quatre des zones bretonnes avec extension (Brest, Quimper, Lorient, Saint-Brieuc). Dans ces zones, on obtient aussi une meilleure estimation de la part des ménages d'une personne, parmi lesquels on a le plus fort taux de non-réponse.

Le calage généralisé apparaît moins performant dans les zones très urbaines de Marseille, Nice et Rennes, ce qu'on peut expliquer par une moindre corrélation entre le vecteur de calage et celui des variables instrumentales.

On a opté pour un redressement de la non-réponse par groupe homogène de réponse avant pré-calage sur les totaux des bases de sondage.

Estimateurs obtenus avec les poids de « pré-calage » sur les bases de sondage, selon trois modes de redressement

Tableau 5. Nombre de résidences principales en 2006

Région	Estimation 2006 RRP	repondération pour non-réponse par GHR avant pré-calage	Non-réponse redressée par calage simple	Non-réponse redressée par calage généralisé
Midi-Pyrénées	1 212 695	1 152 407	1 110 580	1 175 243
Pôle urbain de Toulouse	390 796	375 763	361 106	379 835
Provence Alpes Côte d'Azur	2 077 591	2 020 033	1 909 365	2 041 750
Marseille	801 761	775 905	748 687	754 358
Nice	483 532	477 332	439 241	430 583
Avignon	132 681	127 468	121 040	130 246
Bretagne	1 344 841	1 309 265	1 268 517	1 267 002
Rennes	241 875	243 390	240 068	238 518
Brest	136 410	134 263	130 853	139 184
Vannes	57 847	57 125	54 970	59 890
Quimper	57 904	54 714	52 130	54360
Lorient	87 858	85 685	84 470	85 378
Saint-Brieuc	56 675	53 991	52 547	52 453

Tableau 6. Taux de résidences principales (%) en 2006

Région	Estimation 2006 RRP	repondération pour non-réponse par GHR avant pré-calage	Non-réponse redressée par calage simple	Non-réponse redressée par calage généralisé
Midi-Pyrénées	82,6	80,5	78,1	82,4
Pôle urbain de Toulouse	94,1	93,6	91,3	94,9
Provence Alpes Côte d'Azur	76,5	77,3	70,1	74,9
Marseille	90,2	90,8	88,5	88,6
Nice	71,5	73,5	68,4	66,2
Avignon	88,5	89,9	87,0	92,8
Bretagne	80,6	80,0	77,7	77,6
Rennes	93,6	93,7	92,6	92,1
Brest	88,5	87,3	85,5	90,5
Vannes	88,1	88,8	86,0	92,1
Quimper	79,4	78,8	75,3	78,2
Lorient	89,2	88,0	86,9	87,6
Saint-Brieuc	91,6	89,0	87,0	86,9

Tableau 7. Population en 2006

Région	Estimation 2006 RRP	repondération pour non-réponse par GHR avant pré-calage	Non-réponse redressée par calage simple	Non-réponse redressée par calage généralisé
Midi-Pyrénées	2 681 090	2 639 385	2 552 838	2 627 352
Pôle urbain de Toulouse	820 560	824 649	797 940	799 503
Provence Alpes Côte d'Azur	4 660 040	4 584 341	4 353 895	4 498 129
Marseille	1 838 146	1 831 882	1 772 579	1 769 825
Nice	1 038 314	1 32 027	956 473	886 611
Avignon	307 734	302 675	288 794	310 545
Bretagne	2 996 270	3 108 230	3 021 363	2 966 267
Rennes	549 234	584 690	579 083	586 483
Brest	301 456	316 495	310 172	302 830
Vannes	127 818	136 456	132 020	142 775
Quimper	125 259	125 485	120 917	121 362
Lorient	187 367	196 963	195 544	183 915
Saint-Brieuc	124 095	124 112	121 314	116 373

Tableau 8. Nombre moyen de personnes par ménage en 2006

Région	Estimation 2006 RRP	repondération pour non-réponse par GHR avant pré-calage	Non-réponse redressée par calage simple	Non-réponse redressée par calage généralisé
Midi-Pyrénées	2,23	2,29	2,30	2,23
Pôle urbain de Toulouse	2,10	2,19	2,21	2,10
Provence Alpes Côte d'Azur	2,24	2,27	2,28	2,20
Marseille	2,29	2,36	2,37	2,35
Nice	2,15	2,16	2,18	2,06
Avignon	2,32	2,37	2,38	2,38
Bretagne	2,23	2,37	2,38	2,34
Rennes	2,27	2,40	2,41	2,46
Brest	2,21	2,36	2,37	2,18
Vannes	2,21	2,39	2,40	2,39
Quimper	2,16	2,29	2,32	2,23
Lorient	2,13	2,30	2,31	2,15
Saint-Brieuc	2,19	2,30	2,31	2,22

Tableau 9. Part des ménages d'une personne (%) en 2006

Région	Estimation 2006 RRP	repondération pour non-réponse par GHR avant pré-calage	Non-réponse redressée par calage simple	Non-réponse redressée par calage généralisé
Midi-Pyrénées	33,4	30,7	30,3	32,2
Pôle urbain de Toulouse	39,7	36,0	35,4	36,0
Provence Alpes Côte d'Azur	34,0	32,4	31,9	35,5
Marseille	33,2	31,4	30,9	32,3
Nice	36,9	34,0	33,4	39,7
Avignon	32,2	29,5	28,9	27,0
Bretagne	34,5	30,8	30,5	29,9
Rennes	35,5	34,0	33,6	33,5
Brest	37,6	31,3	30,8	34,1
Vannes	33,0	31,2	30,6	33,2
Quimper	37,4	29,4	28,4	29,7
Lorient	37,0	36,8	36,4	40,3
Saint-Brieuc	36,2	32,4	31,9	34,6

Bibliographie

[1] P. Ardilly. Les techniques de sondage. Editions Technip, 2006.

[2] J.C.Deville, P. Lavallée. Sondage indirect : les fondements de la méthode généralisée du partage des poids. Techniques d'enquête, vol. 32 n° 2, décembre 2006.

[3] Pierre Lavallée. Le sondage indirect ou la méthode généralisée du partage des poids. Statistique et mathématiques appliquées, Paris, Ellipses, 2002.

[4] M. Fesseau, O. Léon, G. Chauvet. Plan de sondage préconisé pour les extensions locales de l'enquête logement 2006 et travaux réalisés dans les DR pour sa mise en oeuvre. Note interne 102/SES/ISM du 9/05/2005, Insee, Rennes.

[5] M. Fesseau. Détail des stratifications des extensions locales de l'enquête logement 2006. Note interne 78/SES/ISM du 19/04/2006, Insee, Rennes.

[6] J. Le Guennec. Enquête logement 2006 - Repondération des échantillons des régions avec extensions locales. Note interne du 19/03/2008, Insee, Rennes.

[7] F. Ouradou. Note descriptive d'étape, relative au plan de sondage de l'enquête Logement 2006. Note interne 1017/DG75-F410 du 21/04/2006, Insee, Paris.