

CALCUL DE PRÉCISION POUR UNE RÉGION AYANT PROCÉDÉ À UNE EXTENSION RÉGIONALE ET LOCALE DE L'ENQUÊTE LOGEMENT 2006

Guillaume CHAUVET (*)

(*) Insee, Direction Régionale de Rennes - CREST/ENSAI

guillaume.chauvet@insee.fr - chauvet@ensai.fr

Introduction

L'Enquête Logement 2006 a pour but de connaître le parc de logements en France et de décrire les conditions de logement des ménages. Afin de disposer d'estimations à des niveaux régionaux et infra-régionaux, certaines régions (Bretagne, Corse, Midi-Pyrénées, Provence Alpes Côte d'Azur) ont procédé à une extension régionale et locale d'échantillon. L'objet de ce travail est de fournir un outil de calcul de précision pour les estimations associées à la région Bretagne. Dans le cas de variables financières affectées par de la non-réponse partielle, ce travail est utilisable mais ne permet pas de prendre en compte la variance d'imputation. Ce problème de non-réponse partielle n'apparaît pas pour les variables non financières, en raison du processus CAPI de saisie des questionnaires.

La note est organisée de la façon suivante. En section 1, le plan de sondage de l'Enquête Logement et les procédures de redressement de l'échantillon sont brièvement rappelés. Le calcul de précision est présenté dans ses grandes lignes en section 2. Le paramétrage de la macro SAS de calcul de précision est détaillé en section 3, avec des exemples d'utilisation. Cette macro est utilisée en section 4 pour proposer une formule approximative de variance pour l'estimation de la taille de domaines. Les détails sont donnés en annexe.

Pour une utilisation de la macro de calcul de précision, seules la section 3 et l'annexe 3 sont nécessaires. Pour une utilisation de la formule approximative de variance, seule la section 4 est nécessaire.

1. Opérations d'échantillonnage et de redressement dans les régions à extension pour l'Enquête Logement 2006

On présente ici les grandes lignes du plan de sondage correspondant à l'Enquête Logement, ainsi que les traitements de correction de la non-réponse et de calage sur l'échantillon de répondants. On trouvera une présentation plus détaillée du plan de sondage dans Fesseau et al. (2005) et Ouradou (2006). Pour les opérations de redressement, voir Le Guennec (2008a,b). Un schéma récapitulatif des opérations d'échantillonnage et de redressement est donné en Annexe 1.

L'Enquête Logement 2006 avait pour population cible les logements résidences principales au 31 juillet 2006. L'échantillon de base est celui de l'enquête nationale réalisée France entière dans les communes de l'Echantillon-Maître (EM), et sélectionné dans le RP99 et la Base de Sondage des Logements Neufs (BSLN). En cas d'extension régionale d'échantillon, un échantillon supplémentaire est sélectionné dans les communes de l'Echantillon-Maître pour les EXTensions régionales (EMEX) dans le RP99 et la Base de Logements Neufs (BLN¹) construite à partir du fichier Sitadel des permis

¹ La BLN donne la liste des logements prévus par les permis de construire. La BSLN donne une liste des logements neufs achevés dans les communes de l'échantillon-maître, et font l'objet d'un suivi.

de construire. Un complément d'échantillon est également sélectionné, d'une part au niveau des communes incluses dans une ZUS dans les logements du RP99, et d'autre part dans des fichiers externes d'allocataires (CNAF, DGI et ANPEEC). A la réunion des échantillons nationaux s'ajoute un échantillon sélectionné localement (cf Fesseau, 2006) :

- dans les aires urbaines de Brest, Lorient, Quimper, Rennes, Saint-Brieuc et Vannes pour la région Bretagne,
- dans le pôle urbain de Toulouse Sicoval pour la région Midi-Pyrénées,
- dans les pôles urbains d'Avignon, Marseille et Nice pour la région PACA.

Les zonages utilisés pour la sélection des échantillons locaux seront simplement appelés zones à extension, et désigneront selon le cas des aires urbaines ou des pôles urbains.

1.1. Echantillon national

L'extension régionale de l'échantillon dans l'EMEX a été coordonnée avec le tirage de l'échantillon national ; en pratique, la sélection a été réalisée de façon simultanée dans le fichier constitué de la réunion entre les bases EM et EMEX, on ne distinguera donc pas les deux échantillons correspondant. Les échantillons tirés dans les fichiers fiscaux de la DGI et de l'ANPEEC ont finalement été écartés du fichier de diffusion national en raison de la difficulté d'identification des bénéficiaires à partir du questionnaire. Les deux échantillons tirés dans le fichier de la CNAF ont été conservés dans le fichier national, mais écartés des fichiers régionaux en raison de la non disponibilité des résultats de la recherche de liens. Au niveau d'une région à extension, l'échantillon national est donc simplement constitué d'un échantillon tiré dans la base EM-EMEX (que l'on notera simplement l'échantillon EM) et d'un échantillon tiré dans la base ZUS (que l'on notera l'échantillon ZUS).

L'échantillon EM est sélectionné dans les communes de la base EM+EMEX, dans le RP99 et la BLN, par un tirage stratifié selon le groupe de régions (cf Ardilly, 2006, p. 226-229), la strate de gestion (au sens de l'échantillon-maître) et la catégorie de logements. Notons que les unités de l'EM et de l'EMEX sont elle-mêmes issues d'un échantillonnage qui doit être pris en compte dans le calcul de précision (cf Annexe 2.4). L'échantillon ZUS est sélectionné au niveau des communes incluses dans une ZUS dans les logements du RP99, selon un tirage stratifié par catégorie de logements.

1.2. Extension locale

L'extension locale de l'Enquête Logement a visé des zones centrées sur des grandes agglomérations, avec tirage stratifié au sein de chaque zone. Bien que le tirage ait été coordonné avec celui de l'échantillon national (les logements sélectionnés dans l'échantillon national ont été exclus de la base de sondage avant tirage de l'extension locale) on supposera, compte tenu de la taille modeste de ce dernier dans la zone à extension, que l'extension locale est approximativement indépendante des échantillons nationaux. La réunion entre échantillons nationaux et extension locale s'effectue à l'aide d'une méthode d'estimation composite.

1.3. Correction de la non-réponse globale sur l'échantillon fusionné

La non-réponse globale est modélisée par une seconde phase de tirage, conditionnelle à l'échantillon théoriquement sélectionné, au cours de laquelle chaque individu statistique (ici, un logement) possède une certaine probabilité de répondre à l'enquête. C'est le modèle des groupes homogènes de réponse qui a été retenu ici : on suppose que l'échantillon est stratifié en sous-populations au sein desquelles la probabilité de réponse est constante (voir Le Guennec, 2008b, p. 10, et Ardilly, 2006, p. 399). La recherche des facteurs corrélés à la non-réponse est réalisée au moyen d'une régression logistique. La correction de la non-réponse se traduit par une modification des poids, et par une augmentation de la variance due à la seconde phase de tirage.

1.4. Calage de l'échantillon fusionné

Les poids obtenus à l'issue de la correction de la non-réponse sont ensuite redressés sur une information externe. Ce redressement s'effectue en deux temps :

- Préalage sur les totaux des bases de sondage,
- Calage final des poids précalés sur des marges 2006 (estimations fournies par le kit Zones Mixtes, cf Baccaïni et Barret, 2007).

Le précalage étant partiellement détérioré par le calage final sur les marges 2006, seul ce dernier est pris en compte dans le calcul de précision.

2. Principe du calcul de précision

On s'intéresse à la population U de logements d'une région à extension. On note S l'échantillon sélectionné (obtenu en fusionnant les différents sous-échantillons), et R l'échantillon de répondants.

Soient y et z deux variables d'intérêt, de totaux respectifs sur la population $t_y = \sum_{k \in U} y_k$ et

$t_z = \sum_{k \in U} z_k$. On se restreint ici à l'estimation de paramètres de type total $\theta = t_y$ ou ratio

$\theta = t_y / t_z$ traités par le programme de calcul de précision.

On note $\hat{\theta}_w = \hat{t}_{y,w} = \sum_{k \in R} w_k y_k$ (respectivement $\hat{\theta}_w = \hat{t}_{y,w} / \hat{t}_{z,w}$) l'estimateur du paramètre calculé sur

l'échantillon R de répondants, où w_k désigne le poids de l'individu k après fusion des échantillons, correction de la non-réponse et calage. En utilisant la technique de linéarisation (Deville, 1999) on a

$$V(\hat{\theta}_w) \approx V(\hat{t}_{u,w}) \quad (1)$$

avec $u_k = y_k$ si $\theta = t_y$ (respectivement $u_k = (1/t_z)(y_k - \theta z_k)$ si $\theta = t_y / t_z$). Le même principe se généralise au cas d'un paramètre plus complexe que le ratio (coefficient de régression, coefficient de corrélation, fractile, ...), et Deville (1999) donne des règles de calcul de la variable linéarisée u_k .

La prise en compte du calage se fait en remplaçant la variable u_k par le résidu e_k de la régression (pondérée) de cette variable sur les variables utilisées dans le calage (Deville et Särndal, 1992), et

$$V(\hat{t}_{u,w}) \approx V(\hat{t}_{e,dp}) \quad (2)$$

avec $\hat{t}_{e,dp} = \sum_{k \in R} \frac{d_k}{p_k} e_k$, d_k le poids du logement k après fusion des sous-échantillons, et p_k la

probabilité pour l'unité k de S de faire partie des répondants. d_k / p_k correspond au poids avant calage. En utilisant la formule de décomposition de la variance, $V(\hat{t}_{e,dp})$ s'obtient en isolant la variance liée à la non-réponse et la variance liée à l'échantillonnage :

$$\begin{aligned}
V(\hat{t}_e) &= E_p V_m \hat{t}_{e,dp} + V_p E_m \hat{t}_{e,dp} \\
&= \underbrace{E_p V_m \hat{t}_{e,dp}}_{\text{variance liée à la non-réponse}} + V_p \underbrace{\left(\sum_{k \in S} d_k e_k \right)}_{\text{variance liée à l'échantillonnage}}
\end{aligned} \tag{3}$$

où E_p et V_p désignent l'espérance et la variance sous le plan de sondage, E_m et V_m l'espérance et la variance sous le mécanisme de réponse. Un estimateur du 1^{er} terme de variance, associé à la non-réponse, est donné en Annexe 2.1.

On note $\hat{t}_{e,d} = \sum_{k \in S} d_k e_k$. Alors

$$\begin{aligned}
V_p(\hat{t}_{e,d}) &= V_p \left(\sum_{k \in S_N} d_k^N \tilde{e}_k^N + \sum_{k \in S_L} d_k^L \tilde{e}_k^L + \sum_{k \in S_Z} d_k^Z \tilde{e}_k^Z \right) \\
&= V_p(\hat{t}_{\tilde{e}}^N + \hat{t}_{\tilde{e}}^L + \hat{t}_{\tilde{e}}^Z)
\end{aligned} \tag{4}$$

où S_N (respectivement S_L et S_Z) désigne le sous-échantillon associé à l'enquête nationale (respectivement, à l'extension locale et à l'échantillon ZUS), d_k^N (respectivement d_k^L et d_k^Z) les poids de sondage associés, et \tilde{e}_k^N (respectivement \tilde{e}_k^L et \tilde{e}_k^Z) sont des variables dépendant des résidus e_k et dont le calcul, lié à la forme de l'estimateur composite, est détaillé en Annexe 2.2.

Les échantillons S_N et S_Z sont indépendants par construction. En admettant que S_L est également approximativement indépendant de S_N et S_Z (cf le paragraphe 1.2), on obtient

$$V_p(\hat{t}_{e,d}) = V_p(\hat{t}_{\tilde{e}}^N) + V_p(\hat{t}_{\tilde{e}}^L) + V_p(\hat{t}_{\tilde{e}}^Z). \tag{5}$$

Le calcul d'un estimateur de $V_p(\hat{t}_{\tilde{e}}^L)$ et $V_p(\hat{t}_{\tilde{e}}^Z)$ est détaillé en Annexe 2.3. Le calcul d'un estimateur de $V_p(\hat{t}_{\tilde{e}}^N)$ est détaillé en annexe 2.4.

3. La macro SAS de calcul de précision

3.1. Les données en entrée

Dans l'enquête logement 2006, l'individu statistique était le logement. Les données doivent donc être disponibles au niveau de chaque logement, sous forme d'une table SAS contenant les informations nécessaires. En particulier, des variables d'intérêt disponibles au niveau individu ou ménage doivent être remontées au niveau du logement pour être exploitées par le programme.

La table en entrée doit contenir l'ensemble des logements **enquêtés** (y compris les logements hors champ) de la **région** (y compris si l'estimation ne porte que sur un domaine), et au moins :

- Les variables d'intérêt,
- La variable donnant les poids finaux.

C'est la table ADRESSE livrée par la division Logement et le Pôle Ingénierie Statistique Ménages qui correspond au bon niveau d'information. Une variable d'intérêt contenue dans une autre table (LOGEMENT, MENAGE ou INDIVIDU) doit être remontée au même niveau que la table ADRESSE.

Une table contenant des informations nécessaires au calcul de précision doit également être appariée avec l'échantillon de logements. Cette table est livrée avec le programme de calcul de précision. L'appariement est géré par la macro. La macro est disponible sur demande auprès de l'auteur.

3.2. Syntaxe de la macro

3.2.1. Paramètres relatifs aux données en entrée

❖ **LIB** = nom d'une librairie SAS

Nom de la librairie SAS contenant les données de l'échantillon et les données à appairer sous forme de deux tables SAS.
Ce paramètre est obligatoire.

❖ **TAB_ENQ** = nom d'une table SAS

Nom de la table SAS contenant les données de l'échantillon.
Ce paramètre est obligatoire.

❖ **AAPPARIER** = nom d'une table SAS

Nom de la table SAS contenant les données à appairer avec la table échantillon.
Ce paramètre est obligatoire. Par défaut, AAPPARIER = *__aappariier*.

❖ **VARTOT_QUAL** = variable(s)

Nom des variables qualitatives concernées par l'estimation de précision, séparées par des blancs. Ces variables peuvent être de type numérique ou caractère. L'estimation porte sur les effectifs estimés par catégorie.
Ce paramètre est facultatif.

Exemple : les variables y1 et y2 sont catégorielles. On s'intéresse à l'estimation des effectifs par modalité de ces deux variables. Paramétrage : VARTOT_QUAL = y1 y2

❖ **VARTOT_QUANT** = variable(s)

Nom des variables quantitatives concernées par l'estimation de précision, séparées par des blancs. Ces variables doivent être de type numérique. L'estimation porte sur les totaux de ces variables.
Ce paramètre est facultatif.

Exemple : les variables y3, y4, y5 et y6 sont quantitatives. On s'intéresse à l'estimation des totaux de ces quatre variables. Paramétrage : VARTOT_QUANT = y3 y4 y5 y6

❖ **VARRAT** = variable(s)

Nom des ratios concernés par l'estimation de précision, séparés par des blancs. Les variables concernées doivent être de type numérique. L'estimation porte sur le ratio des totaux des deux variables.
Ce paramètre est facultatif.

Exemple : On s'intéresse à l'estimation du ratio des totaux des variables y3 et y4, d'une part, et des totaux des variables y5 et y6, d'autre part. Paramétrage : VARRAT = y3/y4 y5/y6

3.2.2. Paramètres relatifs à l'estimation

❖ **PERCENT** = valeur

Ce paramètre est égal à 1 si les effectifs des variables catégorielles sont estimés sous forme de pourcentages, et à 0 sinon.

Ce paramètre est facultatif. Par défaut, PERCENT = 1.

❖ **DOMAINE** = variable

Variable indiquant l'appartenance du logement au domaine d'étude. La variable vaut 1 si le logement appartient au domaine, et 0 sinon.

Ce paramètre est obligatoire.

3.2.3. Paramètres relatifs au calcul de variance

Tous ces paramètres sont obligatoires, mais renseignés par défaut.

3.2.3.1. Paramètres relatifs aux pondérations

❖ **W** = variable

Variable de pondération finale.

Par défaut, W = *poidsreg*. La variable *poidsreg* est contenue dans la table TAB_ENQ.

❖ **POIDS1** = variable

Poids de sondage initial. Selon le logement, il correspond au poids de sondage dans l'échantillon EM, dans l'échantillon ZUS ou dans l'échantillon local.

Par défaut, POIDS1 = *poids1*. La variable *poids1* est contenue dans la table AAPPARIER.

❖ **POIDS2** = variable

Poids de tirage initial dans l'échantillon ZUS quand le logement a été tiré à la fois dans l'échantillon EM et dans l'échantillon ZUS (cf Le Guennec, 2008b, page 39).

Par défaut, POIDS2 = *poids2*. La variable *poids2* est contenue dans la table AAPPARIER.

❖ **D** = variable

Poids après la fusion des différents sous-échantillons selon la technique de l'estimation composite et après redressement de la non-réponse (cf Le Guennec, 2008b, page 6).

Par défaut, D = *ppopt2*. La variable *ppopt2* est contenue dans la table AAPPARIER.

❖ **P** = variable

Poids « de réponse » associé aux répondants à l'enquête (inverse de la probabilité de réponse, cf Le Guennec, 2008b, page 10).

Par défaut, P = *poidsrepfin*. La variable *poidsrepfin* est contenue dans la table AAPPARIER.

❖ **VARGHR** = variable

Variable identifiant les groupes homogènes de réponse.

Par défaut, VARGHR = *grhfin*. La variable *grhfin* est contenue dans la table AAPPARIER.

3.2.3.2. Paramètres relatifs aux plans de sondage initiaux

❖ **VARSTR** = variable(s)

Variable(s) donnant la stratification utilisée pour la sélection de l'extension locale.
Par défaut, VARSTR = *ident_synt*. La variable *ident_synt* est contenue dans la table AAPPARIER.

❖ **N_H** = variable

Variable donnant la taille d'échantillon attendue dans chaque strate pour l'extension locale.
Par défaut, N_H = *nechinit*. La variable *nechinit* est contenue dans la table AAPPARIER.

❖ **TAIL_ST** = variable

Variable donnant la taille de chaque strate utilisée pour l'extension locale.
Par défaut, TAIL_ST = *Nhpop*. La variable *Nhpop* est contenue dans la table AAPPARIER.

❖ **VARSTR_ZUS** = variable(s)

Variable(s) donnant la stratification utilisée pour la sélection de l'échantillon ZUS.
Par défaut, VARSTR_ZUS = *rg*. La variable *rg* est contenue dans la table AAPPARIER.

❖ **N_H_ZUS** = variable

Variable donnant la taille d'échantillon attendue dans chaque strate pour l'échantillon ZUS.
Par défaut, N_H_ZUS = *n_zus*. La variable *n_zus* est contenue dans la table AAPPARIER.

❖ **TAIL_ST_ZUS** = variable

Variable donnant la taille de chaque strate utilisée pour l'échantillon ZUS.
Par défaut, TAIL_ST_ZUS = *tail_zus*. La variable *tail_zus* est contenue dans la table AAPPARIER.

3.2.4. Paramètres relatifs aux sorties

❖ **LIBSOR** = nom d'une librairie SAS

Nom de la librairie SAS contenant les données en sortie.
Ce paramètre est obligatoire. Par défaut, LIBSOR = *work*.

❖ **NIV_SORTIE** = valeur

Spécifie le niveau de détail des sorties (cf paramètre SORTIE). Ce paramètre peut prendre deux valeurs :

- 1 : niveau de détail limité
- 2 : niveau de détail complet

Ce paramètre est obligatoire. Par défaut, NIV_SORTIE = 1.

❖ **SORTIE** = nom d'une table SAS

Table SAS donnant les sorties générées par la macro. Cette table contient autant de lignes que de paramètres à estimer, ainsi que les noms des paramètres à estimer (colonne 1), les estimations des paramètres (colonne 2), la variance du paramètre estimé (colonne 3), le coefficient de variation du paramètre estimé, exprimé en pourcentage (colonne 4), la borne inférieure de l'intervalle de confiance à 95 % pour le paramètre à estimer (colonne 5), la borne supérieure de l'intervalle de confiance à 95 % pour le paramètre à estimer (colonne 6), l'effet de plan, i.e. le rapport entre la variance associée à l'enquête et la variance associée à un

sondage aléatoire simple de même taille (colonne 7, non éditée si NIV_SORTIE = 1), l'effet de calage, i.e. le rapport entre la variance associée à l'enquête et la variance associée à la même enquête mais sans utiliser de calage (colonne 8, non éditée si NIV_SORTIE = 1), la part de la variance due à la non-réponse globale, exprimée en pourcentage (colonne 9, non éditée si NIV_SORTIE = 1).

3.3. Exemples d'utilisation

3.3.1. Exemple 1

On souhaite estimer, sur l'ensemble de la Bretagne, le nombre de logements par catégorie (résidences principales, résidences occasionnelles ou secondaires, logements vacants). La variable d'intérêt est CATL06, contenue dans la table ADRESSE. Tous les logements sont concernés par cette estimation, la variable DOMAINE est donc égale à 1 pour chaque logement. Les estimations sont données sous forme de pourcentage (PERCENT = 1). On utilise le niveau de sortie le plus faible (NIV_SORTIE = 1).

Le tableau 1 donne les résultats obtenus : le code SAS correspondant est donné en Annexe 3. Notons que la variance (estimée par le programme) est nulle pour l'estimateur du nombre de résidences principales, car l'échantillon est entre autres calé sur le nombre de résidences principales en 2006.

Tableau 1 : Estimation du nombre de logements par catégorie sur l'ensemble de la Bretagne

Paramètre à estimer	Estimation	Variance	Coefficient de variation (en %)	Borne Inf intervalle de confiance à 95 %	Borne Sup intervalle de confiance à 95 %
Catl06_1	0.8065	0	0	0.8065	0.8065
Catl06_2	0.117	0.00002563	4.3	0.1078	0.1277
Catl06_4	0.0758	0.00002563	6.68	0.0659	0.0857

Note de lecture : le pourcentage de résidences occasionnelles et secondaires (modalité 2 de la variable Catl06) est estimé en Bretagne à 11,7 %, avec un coefficient de variation de 4.32 %. Un intervalle de confiance à 95 % pour ce paramètre est donné par [0.1078 ; 0.1277].

3.3.2. Exemple 2

On souhaite estimer, sur l'ensemble des résidences principales de l'aire urbaine de Rennes :

- La structure des logements selon le nombre de chambres,
- Le nombre total de chambres,
- La surface moyenne par habitant.

Les variables nécessaires sont :

- Hcha : nombre de chambres du logement,
- Hsh1 : surface du logement (en dehors de la véranda et des pièces professionnelles),
- Nhab : nombre de personnes qui habitent dans le logement.

Les variables Hcha, Hsh1 et Nhab sont contenues dans la table LOGEMENT, et doivent être appariées au niveau de ADRESSE. En utilisant :

- VARTOT_QUAL = Hcha : on obtient la structure en pourcentage (PERCENT = 1) des logements selon le nombre de chambres,

- VARTOT_QUANT = Hcha : on obtient l'estimation du nombre total de chambres,
- VARRAT = Hsh1/nhab : on obtient l'estimation de la surface moyenne par habitant.

La variable DOMAINE est égale à 1 pour chaque résidence principale de l'aire urbaine de Rennes, et 0 sinon. On utilise le niveau de sortie le plus détaillé (NIV_SORTIE = 2). Le tableau 2 donne les résultats obtenus : le code SAS correspondant est donné en Annexe 3.

Le programme peut être utilisé en vue d'opérer des regroupements pour obtenir une précision suffisante sur les estimations. On peut par exemple se fixer un coefficient de variation limite de 10 %, ou 5 %. Si on retient le premier seuil, on voit que la précision est médiocre pour l'estimation du nombre de résidences principales ayant 5, 6 ou 7 chambres (on obtient des coefficients de variation respectivement égaux à 22.3 %, 46.3 % et 144 %) car les modalités correspondantes ne concernent que peu de logements. On crée la variable Hcha_bis en regroupant les logements ayant 4 logements ou plus au sein d'une même modalité. Si on refait tourner le programme pour cette nouvelle variable, on obtient les résultats donnés dans le tableau 3 : les coefficients de variation sont systématiquement inférieurs à 10 %, la précision est donc jugée suffisante.

Tableau 2 : Estimation du nombre total de chambres, de la surface moyenne par habitant et de la structure des logements selon le nombre de chambres sur le parc de résidences principales de l'aire urbaine de Rennes

Paramètre à estimer	Estimation	Variance	Coefficient de variation (en %)	Borne Inf intervalle de confiance à 95 %	Borne Sup intervalle de confiance à 95 %	Effet de plan	Effet de calage	Part de la variance due à la non-réponse (en %)
Hcha	578349,3	55453820,879	1,29	563753,7	592944,9	0,12	0,21	22,06
rat_hsh1nhab	38,2719	0,22739484	1,25	37,337	39,2065	0,48	0,42	21,78
Hcha_0	0,0849	0,00004344	7,76	0,072	0,0978	0,46	0,99	17,34
Hcha_1	0,1808	0,00010598	5,69	0,1607	0,201	0,59	0,9	21,77
Hcha_2	0,2566	0,00021231	5,68	0,2281	0,2852	0,91	0,95	19,12
Hcha_3	0,259	0,00029554	6,64	0,2254	0,2927	1,26	0,98	18,04
Hcha_4	0,1764	0,00017171	7,43	0,1507	0,2021	0,97	0,8	18,79
Hcha_5	0,0391	0,00008506	23,60	0,021	0,0572	1,86	0,97	19,94
Hcha_6	0,0027	0,00000169	48,50	0,0001	0,0052	0,52	1,01	17,36
Hcha_7	0,0004	0,00000033	151,65	-0,0008	0,0015	0,72	1,01	17,21

Notes de lecture :

Le nombre total de chambres (Paramètre = Hcha) est estimé à 578 349, avec un coefficient de variation de 1.29 %. Un intervalle de confiance à 95 % pour ce total est donné par [563 754 ; 592 945]. La variance représente 12 % de la variance que l'on obtiendrait avec un sondage aléatoire simple, et 21 % de la variance que l'on obtiendrait sans calage. La non-réponse représente 22.1 % de la variance.

La surface moyenne par habitant (Paramètre = rat_hsh1nhab) est estimée à 38.27 m², avec un coefficient de variation de 1.25 %. Un intervalle de confiance à 95 % pour ce ratio est donné par [37.34 ; 39.21]. La variance représente 48 % de la variance que l'on obtiendrait avec un sondage aléatoire simple, et 42 % de la variance que l'on obtiendrait sans calage. La non-réponse représente 21.8 % de la variance.

Le pourcentage de résidences principales avec 3 chambres (Paramètre = Hcha_3) est estimé à 25.9 %, avec un coefficient de variation de 6.64 %. Un intervalle de confiance à 95 % pour ce pourcentage est donné par [22.54 ; 29.27]. La variance représente 1.26 fois la variance que l'on obtiendrait avec un sondage aléatoire simple, et 98 % de la variance que l'on obtiendrait sans calage. La non-réponse représente 18.0 % de la variance.

Tableau 3 : Estimation du nombre total de chambres, de la surface moyenne par habitant et de la structure des logements selon le nombre de chambres (après regroupement) sur le parc de résidences principales de l'aire urbaine de Rennes

Paramètre à estimer	Estimation	Variance	Coefficient de variation (en %)	Borne Inf intervalle de confiance à 95 %	Borne Sup intervalle de confiance à 95 %	Effet de plan	Effet de calage	Part de la variance due à la non-réponse (en %)
Hcha	578349,3	55453820,879	1,29	563753,7	592944,9	0,12	0,21	22,06
rat_hsh1nhab	38,2719	0,22739484	1,25	37,337	39,2065	0,48	0,42	21,78
Hcha_bis_0	0,0849	0,00004344	7,76	0,072	0,0978	0,46	0,99	17,34
Hcha_bis_1	0,1808	0,00010598	5,69	0,1607	0,201	0,59	0,9	21,77
Hcha_bis_2	0,2566	0,00021231	5,68	0,2281	0,2852	0,91	0,95	19,12
Hcha_bis_3	0,259	0,00029554	6,64	0,2254	0,2927	1,26	0,98	18,04
Hcha_bis_4	0,2186	0,00020123	6,49	0,1908	0,2464	0,97	0,81	19,93

4. Formule approchée de variance pour l'estimation de la taille de domaines

Afin d'éviter le recours systématique à la macro de calcul de variance, on propose dans cette section une formule approchée de calcul de précision. Le programme et les sorties SAS associés sont donnés en Annexe 4.

Cette approximation est basée sur une liaison entre la variance de l'estimation de la taille d'un domaine et la taille du domaine elle-même (cf Ardilly, 2006, p. 577). Plus exactement, pour un domaine d de taille estimée \hat{N}_d , on a

$$\frac{\hat{V}(\hat{N}_d)}{\hat{N}_d^2} \approx \frac{\alpha}{\hat{N}_d} + \beta \quad (6)$$

où α et β sont des constantes à estimer, dépendant de la population et du plan de sondage. Le modèle proposé ci-dessous est ajusté à l'aide des données recueillies en Bretagne. Les domaines utilisés sont donnés par le croisement de la catégorie de logements (on ne retient ici que les résidences principales), de la zone géographique (Bretagne entière, AU de Brest, AU de Lorient, AU de Quimper, AU de Rennes, AU de Saint-Brieuc, AU de Vannes) et des variables suivantes :

- Nombre de chambres du logement (variable HCHA),
- Surface habitable du logement (variable HSHT),
- Nombre d'habitants du logement (variable NHAB),
- Nombre de pièces d'habitation du logement (variable HNPH1),
- Nombre de salles d'eau (variable KSE),
- Existence de WC (variable KWC),
- Existence d'une cave (variable KCA).

Par exemple, le croisement du champ « Résidences Principales », de la zone « AU de Brest » et de la modalité 1 de la variable HCHA correspond au domaine « Résidences principales de l'AU de Brest ne comptant qu'une seule chambre ». A l'aide de la macro de calcul de précision, on obtient ainsi 283 estimations pour les tailles de domaine, et les 283 estimations de variance correspondantes. Un extrait correspondant aux résultats obtenus pour la Bretagne est présenté dans le tableau 8, page 28. On utilise ces observations pour estimer les paramètres du modèle (6). On obtient pour la

modélisation avec constante un coefficient de détermination de 79.7 %, mais le paramètre β n'est que faiblement significatif. On choisit finalement de retenir le modèle sans constante, qui donne un coefficient de détermination de 81.5 %. On obtient la formule approchée :

$$\frac{\hat{V}(\hat{N}_d)}{\hat{N}_d^2} \approx \frac{110}{\hat{N}_d} \Rightarrow CV(\hat{N}_d) \approx \frac{10.5}{\sqrt{\hat{N}_d}} \quad (7)$$

On retient de la formule (7) que la précision se dégrade logiquement au fur et à mesure que la taille du domaine diminue. Si on se fixe un seuil à 10 % pour le coefficient de variation, on évitera donc de produire des estimations pour des domaines de taille (estimée) inférieure à $(10.5/0.10)^2 \approx 11000$ logements.

Afin d'évaluer la robustesse de cette formule approchée, on réalise à nouveau le même travail de modélisation dans le cadre suivant :

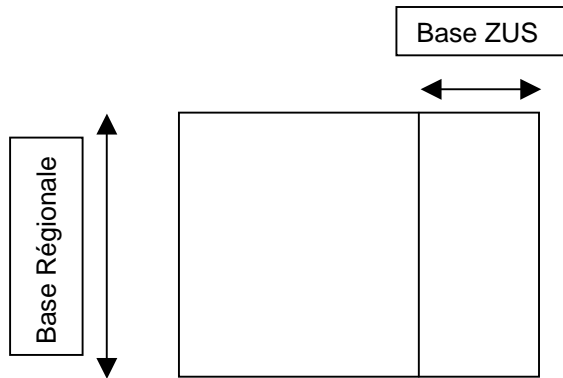
- On exclut les domaines associés à l'AU de Vannes,
- On exclut les domaines associés au nombre de chambres (variables HCHA).

Les résultats obtenus dans chacun des deux cas sont très proches de ceux obtenus avec l'ensemble des données. Ce résultat est rassurant puisqu'il va dans le sens d'une non-dépendance des coefficients α et β vis à vis du domaine, hypothèse sous-jacente à la modélisation utilisée.

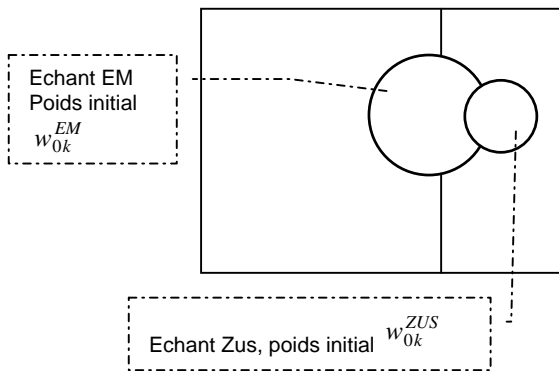
Bibliographie

- [1] Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.
- [2] Baccaïni, B., et Barret, C. (2007). *Zones mixtes : Guide Méthodologique*. Investissement AT31, PSAR Analyse Territoriale, Insee, France.
http://preprod-php.insee.fr/pacapsar/article.php?id_article=211
- [3] Bourdalle, G., et Christine, M., et Wilms, L. (2000). *Echantillons maître et emploi*. Article présenté aux VIIes journées de méthodologie statistique de l'Insee, Paris, France.
http://www.dg-dsds.insee.fr/F4_UMS/F410/Accueil/accueilV2_f410.htm
- [4] Christine, M., et Wilms, L. (2000). *Problèmes théoriques et pratiques de la construction de l'EMEX*. Article présenté aux VIIes journées de méthodologie statistique de l'Insee, Paris, France.
http://www.dg-dsds.insee.fr/F4_UMS/F410/Accueil/accueilV2_f410.htm
- [5] Deville, J-C. (1999). *Variance estimation for complex statistics and estimators : linearization and residual techniques*. Survey Methodology, 25, p. 193-204.
- [6] Deville, J-C., et Särndal, C-E. (1992). *Calibration estimators in Survey Sampling*. Journal of the American Statistical Association, 88, p. 376-382.
- [7] Fesseau, M., et Léon, O., et Chauvet, G. (2005). *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, Insee, Rennes.
http://10.64.230.2/menages/sommaireinvestissementsencours/Plan_loc.pdf
- [8] Fesseau, M. (2006). *Détail des stratifications des extensions locales de l'enquête logement 2006*. Note interne 78/SES/ISM, Insee, Rennes.
http://10.64.230.2/menages/sommaireinvestissementsencours/Note_78SESISM.pdf
- [9] Kish, L. (1980). *Design and estimation for domains*. The Statistician, 29, p. 209-222.
- [10] Le Guennec, J. (2008a). *Enquête logement 2006 - Mise à disposition des pondérations issues des redressements par calage des régions avec extensions locales*. Note interne 31-SES/PISM, Insee, Rennes.
http://10.64.230.2/menages/sommairetravauxterminees/e-mise_a_disposition_des_fichiers_avec_ponderations_locale%e2%80%a6.pdf
- [11] Le Guennec, J. (2008b). *Enquête logement 2006 - Repondération des échantillons des régions avec extensions locales*. Note interne non numérotée, Insee, Rennes.
- [12] Ouradou, F. (2006). *Note descriptive d'étape, relative au plan de sondage de l'enquête Logement 2006*. Note interne 1017/DG75-F410, Insee, Paris.
- [13] Särndal, C-E., et Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New-York.

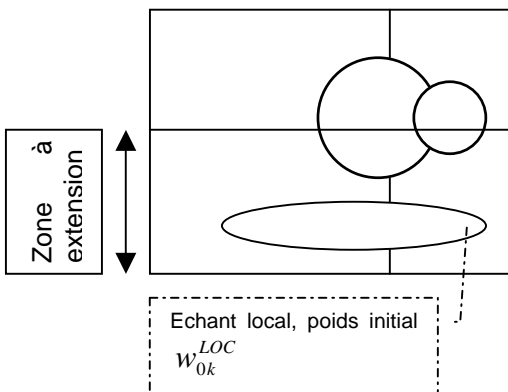
Annexe 1 : Echantillonnage et redressement de l'Enquête Logement, schéma récapitulatif



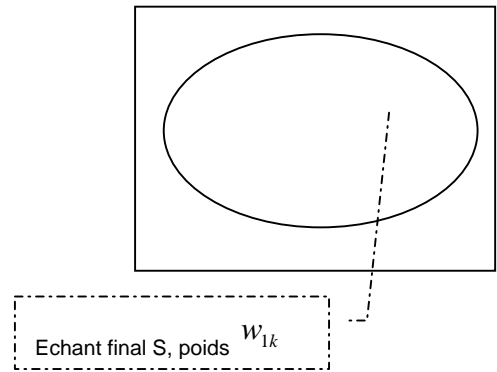
Etape 1 : sélection des échantillons EM et ZUS



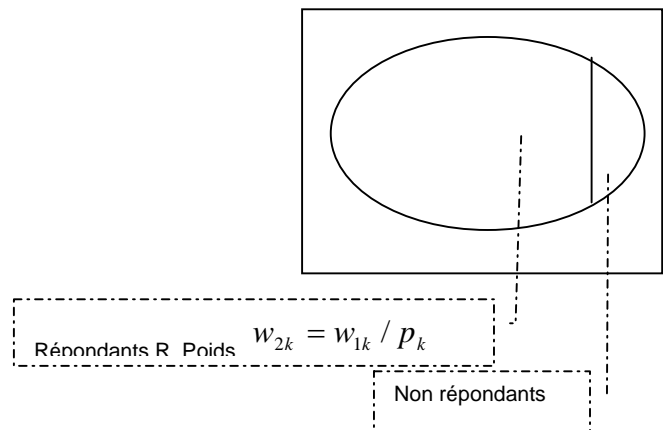
Etape 2 : sélection de l'extension locale



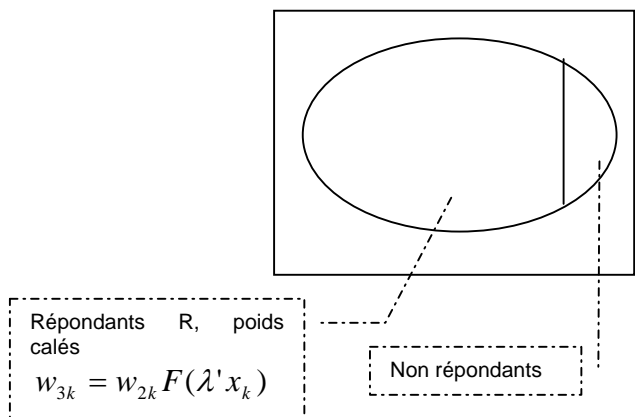
Etape 3 : estimation composite



Etape 4 : correction de la NR



Etape 5 : calage



Annexe 2 : Détails sur l'estimation de précision

A2.1. Estimation de la variance dûe à la non-réponse

On rappelle que la non-réponse globale est modélisée par une phase supplémentaire de tirage, en postulant un tirage de seconde phase dans l'échantillon S stratifié en groupes appelés groupes homogènes de réponse, avec des tirages bernoulliens et indépendants dans chacun des groupes. La probabilité de réponse de l'individu k est notée p_k . Elle est supposée constante au sein d'un groupe de réponse.

On notera I_1, \dots, I_I les I groupes homogènes de réponse, partitionnant la population U , $S_i = I_i \cap S$ l'échantillon de taille n_i appartenant au groupe I_i , $R_i = I_i \cap R$ l'ensemble de taille m_i des répondants appartenant au groupe I_i . On a donc $S = S_1 \cup \dots \cup S_I$, $n = n_1 + \dots + n_I$ et $m = m_1 + \dots + m_I$, et avec les notations précédentes, une probabilité de réponse (estimée) $p_k = m_i / n_i$ pour tout individu k appartenant à S_i . On peut donc encore réécrire :

$$\begin{aligned} \hat{t}_{e,dp} &= \sum_{k \in R} \frac{d_k}{p_k} e_k \\ &= \sum_{i=1}^I \frac{n_i}{m_i} \sum_{k \in R_i} d_k e_k \end{aligned}$$

On calcule alors la variance liée à la non-réponse en raisonnant conditionnellement aux nombres de répondants obtenus par groupe de réponse (ce qui revient à assimiler le tirage de 2^{nde} phase à un sondage aléatoire simple stratifié) :

$$E_p V_m \hat{t}_{e,dp} = E_p \left(\sum_{i=1}^I n_i^2 \left(1 - \frac{m_i}{n_i} \right) \frac{s_{ed,i}^2}{m_i} \right)$$

avec $s_{ed,i}^2 = \frac{1}{n_i - 1} \sum_{k \in S_i} (d_k e_k - \bar{e}_{d,i})^2$ et $\bar{e}_{d,i} = \frac{1}{n_i} \sum_{k \in S_i} d_k e_k$. On peut l'estimer par

$$\hat{V}_{NR} = \sum_{i=1}^I n_i^2 \left(1 - \frac{m_i}{n_i} \right) \frac{\tilde{s}_{ed,i}^2}{m_i}$$

avec $\tilde{s}_{ed,i}^2 = \frac{1}{m_i - 1} \sum_{k \in R_i} (d_k e_k - \tilde{e}_{di})^2$ et $\tilde{e}_{di} = \frac{1}{m_i} \sum_{k \in R_i} d_k e_k$.

A2.2. Calcul de l'estimateur composite

Un schéma récapitulatif de la décomposition des différents sous-échantillons selon le zonage est rappelé en figure 1. L'échantillon S_N , de taille $n = n_0 + n_1 + n_2 + n_3$, est l'échantillon tiré dans l'Echantillon Maître et l'Echantillon Maître pour les Extensions Régionales de la région, avec :

- n_0 le nombre de logements de S_N situé hors ZUS et hors zone à extension locale,
- n_1 le nombre de logements de S_N situé dans une ZUS et hors zone à extension locale,
- n_2 le nombre de logements de S_N situé hors ZUS et dans une zone à extension locale,
- n_3 le nombre de logements de S_N situé dans une ZUS et dans une zone à extension locale.

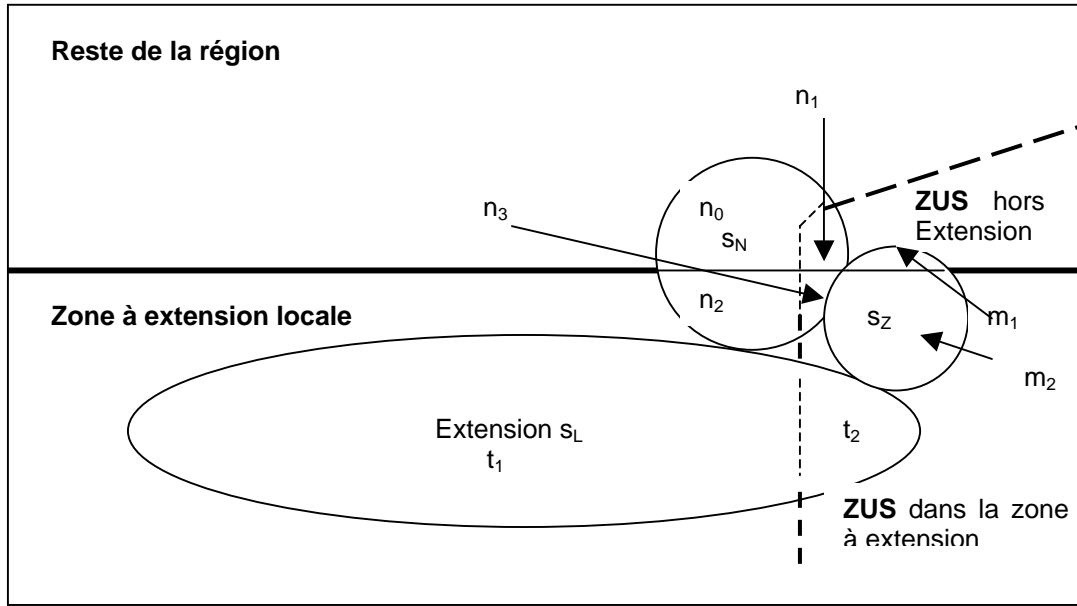


Fig. 1 : Fusion des échantillons dans une région avec extension (Le Guennec, 2008b, p. 7)

L'échantillon S_Z , de taille $m = m_1 + m_2$, est l'échantillon ZUS tiré dans la région, avec :

- m_1 le nombre de logements de S_Z situé hors zone à extension locale,
- m_2 le nombre de logements de S_Z situé dans une zone à extension locale.

L'échantillon S_L , de taille $t = t_1 + t_2$, est l'échantillon tiré dans la zone à extension locale, avec :

- t_1 le nombre de logements de S_L situé hors ZUS,
- t_2 le nombre de logements de S_L situé dans une ZUS.

Le tableau 4 donne la méthode de calcul des variables \tilde{e}_k utilisées pour le calcul de précision.

A2.3. Estimation de variance pour un SAS stratifié

Nous nous plaçons dans le cas du sous-échantillon S_L , sélectionné au titre de l'extension locale dans la base U_L . S_L est sélectionné à l'aide d'un tirage stratifié à un seul degré, avec sondage systématique ordonné à probabilités égales dans chaque strate. Un plan de sondage analogue est utilisé pour la sélection de l'échantillon S_Z dans la base ZUS U_Z . L'estimateur proposé est donc également utilisable pour l'estimation de la variance associée à l'échantillonnage de S_Z .

On note U_{L1}, \dots, U_{LH} les H strates correspondantes, de tailles respectives N_{L1}, \dots, N_{LH} . En assimilant le tirage systématique à un sondage aléatoire simple, on obtient :

$$V_p(\hat{t}_{\tilde{e}}^L) = \sum_{h=1}^H N_{Lh}^2 \left(1 - \frac{n_{Lh}}{N_{Lh}}\right) \frac{S_{\tilde{e}h}^2}{n_{Lh}}$$

Tableau 4 : Calcul des variables synthétiques en fonction du sous-échantillon, de la zone d'appartenance et de la nature du logement (neuf ou ancien)

Sous-échantillon	Logement situé dans une Zone à extension	Logement situé dans une ZUS	Variable synthétique
Echantillon national S_N	oui	oui	$\tilde{e}_k^N = \frac{n_3}{n_3 + m_2 + t_2} e_k$ si logement créé avant 1999, $\tilde{e}_k^N = \frac{n_3}{n_3 + t_2} e_k$ sinon
		non	$\tilde{e}_k^N = \frac{n_2}{n_2 + t_1} e_k$
	non	oui	$\tilde{e}_k^N = \frac{n_1}{n_1 + m_1} e_k$
		non	$\tilde{e}_k^N = e_k$
Echantillon local S_L	-	oui	$\tilde{e}_k^L = \frac{t_2}{n_3 + m_2 + t_2} e_k$ si logement créé avant 1999, $\tilde{e}_k^L = \frac{t_2}{n_3 + t_2} e_k$ sinon
		non	$\tilde{e}_k^L = \frac{t_1}{n_2 + t_1} e_k$
Echantillon ZUS S_Z	oui	-	$\tilde{e}_k^Z = \frac{m_2}{n_3 + m_2 + t_2} e_k$
	non		$\tilde{e}_k^Z = \frac{m_1}{n_1 + m_1} e_k$

avec $S_{\tilde{e}_h}^2 = \frac{1}{N_{L_h} - 1} \sum_{k \in U_{L_h}} (\tilde{e}_k - \mu_{\tilde{e}_h})^2$ et $\mu_{\tilde{e}_h} = \frac{1}{N_{L_h}} \sum_{k \in U_{L_h}} \tilde{e}_k$. Un estimateur sans biais de $S_{\tilde{e}_h}^2$ est

donné par $s_{\tilde{e}_h}^2 = \frac{1}{n_{L_h} - 1} \sum_{k \in S_{L_h}} (\tilde{e}_k - \bar{\tilde{e}}_h)^2$ (avec $\bar{\tilde{e}}_h = \frac{1}{n_{L_h}} \sum_{k \in S_{L_h}} y_k$, $S_{L_h} = U_h \cap S_L$ et $n_{L_h} = \text{Card}(S_{L_h})$),

mais cette quantité est inconnue en raison de la non-réponse. On choisit donc d'exprimer $S_{\tilde{e}_h}^2$ comme une fonction de totaux que l'on peut ensuite estimer par substitution. On a en effet

$$S_{\tilde{e}_h}^2 = \frac{1}{N_{L_h} - 1} \sum_{k \in U_{L_h}} \tilde{e}_k^2 - \frac{1}{N_{L_h} (N_{L_h} - 1)} \left(\sum_{k \in U_{L_h}} \tilde{e}_k \right)^2$$

que l'on peut estimer par

$$\hat{S}_{\tilde{e}_h}^2 = \frac{1}{N_{L_h} - 1} \sum_{k \in U_{L_h} \cap R} w_k \tilde{e}_k^2 - \frac{1}{N_{L_h} (N_{L_h} - 1)} \left(\sum_{k \in U_{L_h} \cap R} w_k \tilde{e}_k \right)^2.$$

On obtient donc l'estimateur

$$\hat{V}_p(\hat{t}_{\tilde{z}}^L) = \sum_{h=1}^H N_{Lh}^2 \left(1 - \frac{n_{Lh}}{N_{Lh}}\right) \frac{\hat{S}_{\tilde{z}h}^2}{n_{Lh}}$$

Une autre possibilité consiste à décomposer la dispersion inconnue $s_{\tilde{z}h}^2$ sur les groupes homogènes de réponse. En pratique, certaines des cellules correspondant à l'intersection des groupes de réponse et des strates U_{Lh} seront vraisemblablement de taille faible (voire nulle) ; une solution consiste à utiliser une technique de collapsing consistant à les regrouper pour avoir un nombre minimal de répondants par cellule. Le choix de ce nombre minimal est arbitraire, et son impact sur l'estimation de variance est difficile à évaluer.

L'échantillon ZUS est sélectionné nationalement par un tirage stratifié selon la catégorie de logement, mais pas selon la région. Le tableau 5 donne la répartition de l'échantillon initial selon la catégorie de logements

Tableau 5 : répartition de l'échantillon ZUS selon la région et la catégorie de logement

Région	Résidences Principales	Résidences occasionnelles et secondaires	Logements vacants
Bretagne	147	0	7
Midi-Pyrénées	90	0	9
PACA	519	4	64

Bien que l'estimateur calculé régionalement soit un estimateur sur domaine obtenu par tirage stratifié, on a considéré dans le calcul de variance que l'échantillon ZUS était issu d'un sondage aléatoire simple dans la région, les gains dûs à la stratification ayant tendance à disparaître pour des domaines de taille restreinte (Kish, 1980, p. 215). On applique donc la méthode précédente dans le cas particulier d'une seule strate (supposée) de tirage.

A2.4. Estimation de variance pour l'échantillon tiré dans l'échantillon-maître

Pour le tirage de l'échantillon S_N , les communes sont tout d'abord regroupées au sein de cinq groupes appelés **strates de gestion**, selon le degré d'urbanisation :

- SG0 = unités rurales (regroupements de communes rurales),
- SG1 = UU de moins de 20.000 habitants ou associations de telles UU,
- SG2 = UU de 20.000 à 100.000 habitants,
- SG3 = UU de plus de 100.000 habitants (hors UU de Paris),
- SG4 = UU de Paris.

Dans chaque strate de gestion, un échantillon noté $S_N^i, i = 0, \dots, 3$ est sélectionné indépendamment.

On note $\hat{t}_{\tilde{z}}^{N,i}$ l'estimateur correspondant. S_N est donné par la réunion des S_N^i , et $\hat{t}_{\tilde{z}}^N$ par la somme des $\hat{t}_{\tilde{z}}^{N,i}$. On a :

$$V_p(\hat{t}_{\tilde{z}}^N) = V_p(\hat{t}_{\tilde{z}}^{N,0}) + V_p(\hat{t}_{\tilde{z}}^{N,1}) + V_p(\hat{t}_{\tilde{z}}^{N,2}) + V_p(\hat{t}_{\tilde{z}}^{N,3}).$$

On rappelle ci-dessous le principe de tirage dans chaque strate de gestion (hors strate SG4), et un estimateur de variance correspondant à chacun des termes de la somme précédente.

A2.4.1. Strate de gestion SGO

A2.4.1.1. Sélection de l'échantillon-maître et de l'EMEX

Dans la strate des communes rurales, l'UP est un regroupement de communes rurales contiguës. Les UP sont stratifiées par super-régions (cf Ardilly, 2006, p. 226-229). L'échantillon maître est sélectionné avec des probabilités d'inclusion proportionnelles au nombre de résidences principales de l'UP, en équilibrant sur :

- Le revenu net imposable (sources fiscales, année 1996),
- L'effectif de la tranche d'âge [0,19] (RP99),
- L'effectif de la tranche d'âge [20,59] (RP99),
- L'effectif de la tranche d'âge [60, +] (RP99),
- La probabilité d'inclusion multipliée par l'indicatrice d'appartenance aux régions.

L'équilibrage est réalisé sur les super-régions. Les contraintes correspondant aux probabilités d'inclusion éclatées permettent d'obtenir une taille fixe dans chaque région. La sélection de l'EMEX conduit à tripler le nombre d'UP sélectionnées, toujours à probabilités proportionnelles au nombre de résidences principales, mais l'équilibrage super-régional ne peut plus être assuré (cf Ardilly, 2006, p. 240). Afin de compenser une éventuelle dispersion géographique des UP, ces UP sont éventuellement fractionnées a posteriori en deux fractions qui seront sollicitées tour à tour lors des tirages (on parlera de fraction active pour la fraction d'UP sollicitée). Bien que ce fractionnement corresponde à un 2nd degré de tirage, on considère dans le calcul de précision que l'on tire directement des UP (ou fractions d'UP).

A2.4.1.2. Sélection de l'échantillon final dans l'EM-EMEX

Un échantillon d'UP est sélectionné dans l'EM-EMEX par sondage aléatoire simple (afin de respecter au final la contrainte de probabilités proportionnelles au nombre de résidences principales). Chaque UP (ou fraction active d'UP) est ensuite découpée en groupes de communes comptant au moins 100 logements, et ces groupes de communes sont découpés en deux strates :

- La strate des groupes de communes dont le tirage est certain,
- La strate des groupes de communes ayant une probabilité de tirage inférieure strictement à 1.

Dans la strate non exhaustive, un échantillon de groupes de communes est sélectionné avec des probabilités proportionnelles à la taille. Dans les groupes de communes sélectionnés, les logements sont ensuite stratifiés par catégorie, et l'échantillon final est sélectionné par sondage systématique à probabilités égales (que l'on assimile à un sondage aléatoire simple) dans chaque strate. Comme on obtient pratiquement de très petites strates, le tirage stratifié est assimilé à un tirage systématique selon la catégorie de logements.

Au niveau de chaque région ayant fait l'objet d'une extension régionale, la sélection d'un échantillon de logements en strate SGO est réalisée selon les étapes suivantes :

- **1^{er} degré : sélection d'un échantillon d'UP (ou fractions d'UP) à probabilités proportionnelles à la taille,**
- **2^{ème} degré : découpage de l'UP (ou de la fraction d'UP) sélectionnée en groupes de communes et sélection d'un échantillon de groupes de communes à probabilités proportionnelles à la taille,**
- **3^{ème} degré : stratification de chaque groupe de communes sélectionné selon la catégorie de logements, et sélection d'un échantillon à probabilités égales dans chaque strate (le tirage stratifié étant assimilé à un tirage systématique dans chaque groupe de communes).**

A2.4.2. Strate de gestion SG1

A2.4.2.1. Sélection de l'échantillon-maître et de l'EMEX

Dans la strate des unités urbaines de moins de 20 000 habitants, l'UP est un regroupement d'Unités Urbaines (pas nécessairement contigües). La sélection de l'EM-EMEX s'effectue de la même façon qu'en strate SG0.

A2.4.2.2. Sélection de l'échantillon final dans l'EM-EMEX

Un échantillon d'UP est sélectionné dans l'EM-EMEX par sondage aléatoire simple (afin de respecter au final la contrainte de probabilités proportionnelles au nombre de résidences principales). Chaque UP (ou fraction active d'UP) est ensuite découpée en groupes de communes comptant au moins 100 logements. Dans chaque groupe de communes, les logements sont ensuite stratifiés par catégorie, et l'échantillon final est sélectionné par sondage systématique à probabilités égales (que l'on assimile à un sondage aléatoire simple) dans chaque strate. Il n'y a donc pas de sélection de groupes de communes comme en strate SG0.

Au niveau de chaque région ayant fait l'objet d'une extension régionale, la sélection d'un échantillon de logements en strate SG1 est réalisée selon les étapes suivantes :

- **1^{er} degré : sélection d'un échantillon d'UP à probabilités proportionnelles à la taille,**
- **2nd degré : sélection (éventuelle) d'une fraction d'UP dans les UP sélectionnées,**
- **3^{ème} degré :**
 - **découpage de l'UP (ou de la fraction d'UP) sélectionnée en groupes de communes**
 - **stratification de chaque groupe de communes selon la catégorie de logements**
 - **sélection d'un échantillon à probabilités égales dans chaque strate.**

A2.4.3. Strate de gestion SG2

A2.4.3.1. Sélection de l'échantillon-maître et de l'EMEX

En strate SG2, une UP est une unité urbaine. Les UP sont stratifiées par super-régions. L'échantillon maître est sélectionné avec des probabilités d'inclusion proportionnelles au nombre de résidences principales de l'UP, en équilibrant sur les mêmes variables que pour la strate SG0. La sélection de l'EMEX conduit à retenir l'intégralité des UP existantes. Pour les régions ayant fait l'objet d'une extension régionale d'enquête, il n'y a donc pas à proprement parler de tirage d'UP.

Dans chaque UP, les groupes de districts sont « pseudo-stratifiés » par regroupement de communes. Dans chaque UP, on sélectionne un échantillon de 30 groupes de districts avec des probabilités d'inclusion constantes dans chaque regroupement de communes (l'allocation étant proportionnelle au nombre de résidences principales), en équilibrant sur :

- L'effectif de la tranche d'âge [0,19] (RP99),
- L'effectif de la tranche d'âge [20,59] (RP99),
- L'effectif de la tranche d'âge [60, +] (RP99),
- La probabilité d'inclusion multipliée par l'indicatrice d'appartenance aux pseudo-strates (= regroupement de communes).

L'équilibrage porte sur l'UP. Les contraintes correspondant aux probabilités d'inclusion éclatées permettent d'obtenir une taille fixe dans chaque regroupement de communes.

A2.4.3.2. Sélection de l'échantillon final dans l'EM-EMEX

Dans chaque groupe de communes, les logements des districts sélectionnés sont stratifiés par catégorie, et l'échantillon final est sélectionné par sondage systématique à probabilités égales (que l'on assimile à un sondage aléatoire simple) dans chaque strate. L'étape de sélection des districts n'est pas prise en compte dans le calcul de précision ; en effet, l'échantillon final compte de 1 à 3 logements dans chaque district, ce qui limite l'effet de grappe (cf Ardilly, p. 120 et 232).

Au niveau de chaque région ayant fait l'objet d'une extension régionale, la sélection d'un échantillon de logements en strate SG2 est schématiquement réalisée selon les étapes suivantes :

- **découpage de chaque UP en groupes de communes,**
- **stratification de chaque groupe de communes selon la catégorie de logements,**
- **sélection d'un échantillon à probabilités égales dans chaque strate.**

A2.4.4. Strate de gestion SG3

A2.4.4.1. Sélection de l'échantillon-maître et de l'EMEX

En strate SG3, une UP est une unité urbaine. Toutes les UP sont sélectionnées. Dans chaque UP, les districts sont « pseudo-stratifiés » par regroupement de communes et on sélectionne un échantillon de districts avec des probabilités d'inclusion égales à 5 %, en équilibrant sur les mêmes variables que pour la strate SG2. La sélection de l'EMEX conduit à doubler le nombre de districts sélectionnés (toujours à probabilités égales).

A2.4.4.2. Sélection de l'échantillon final dans l'EM-EMEX

Dans chaque groupe de communes, les logements des districts sélectionnés sont stratifiés par catégorie, et l'échantillon final est sélectionné par sondage systématique à probabilités égales (que l'on assimile à un sondage aléatoire simple) dans chaque strate. Comme dans le cas de la strate SG2, l'étape de sélection des districts n'est pas prise en compte dans le calcul de précision.

Au niveau de chaque région ayant fait l'objet d'une extension régionale, la sélection d'un échantillon de logements en strate SG3 est schématiquement réalisée selon les étapes suivantes :

- **découpage de chaque UP en groupes de communes,**
- **stratification de chaque groupe de communes selon la catégorie de logements,**
- **sélection d'un échantillon à probabilités égales dans chaque strate.**

A2.4.5. Estimation de variance

A2.4.5.1 Strate de gestion SG0

Dans la strate de gestion 0, l'échantillon est sélectionné selon un plan de sondage à 3 degrés. La variance est estimée en se basant sur l'estimateur sans biais pour un sondage à trois degrés donné par exemple dans Särndal et al. (1992), section 4.4, page 148. On note U_I la population d'unités primaires (ici, une UP ou une fraction d'UP au sens de l'échantillon maître), U_{II} la population d'unités secondaires (ici, un regroupement de communes) et U_{III} la population d'unités tertiaires (ici, le logement). Soit s_I l'échantillon d'unités primaires sélectionné avec des probabilités d'inclusion $(\pi_i)_{i \in U_I}$, s_{IIi} l'échantillon d'unités secondaires sélectionné avec des probabilités d'inclusion $(\pi_{j|i})_{j \in i}$ dans l'unité primaire i retenue dans s_I , et s_{IIIij} l'échantillon d'unités tertiaires sélectionné avec des probabilités d'inclusion $(\pi_{k|i,j})_{k \in j}$ dans l'unité secondaire j retenue dans s_{IIi} . On utilise l'estimateur de variance :

$$\hat{V}_p(\hat{t}_{\tilde{z}}^{N,0}) = \frac{m}{m-1} \sum_{i \in s_I} (1 - \pi_i) \left(\frac{\tilde{t}_i}{\pi_i} - \tilde{R} \right)^2 + \sum_{i \in s_I} \frac{\hat{V}_i}{\pi_i}$$

avec

$$\tilde{t}_i = \sum_{j \in s_{III}} \frac{\tilde{t}_{ij}}{\pi_{j|i}}, \tilde{R}_i = \frac{\sum_{i \in s_I} \tilde{t}_i (1 - \pi_i)}{\sum_{i \in s_I} (1 - \pi_i)} \text{ et } \hat{V}_i = \frac{m_i}{m_i - 1} \sum_{j \in s_{III}} (1 - \pi_{j|i}) \left(\frac{\tilde{t}_{ij}}{\pi_{j|i}} - \tilde{R}_i \right)^2 + \sum_{j \in s_{III}} \frac{\hat{V}_{ij}}{\pi_{j|i}},$$

$$\tilde{t}_{ij} = \sum_{k \in R \cap s_{IIIij}} \frac{\tilde{e}_k}{\pi_{k|i,j} p_k}, \tilde{R}_i = \frac{\sum_{j \in s_{III}} \frac{\tilde{t}_{ij}}{\pi_{j|i}} (1 - \pi_{j|i})}{\sum_{j \in s_{III}} (1 - \pi_{j|i})} \text{ et } \hat{V}_{ij} = \frac{m_{ij}}{m_{ij} - 1} \sum_{k \in R \cap s_{IIIij}} \frac{(1 - \pi_{k|i,j})}{p_k} \left(\frac{\tilde{e}_k}{\pi_{k|i,j}} - \tilde{R}_{ij} \right)^2,$$

et

$$\tilde{R}_{ij} = \frac{\sum_{k \in R \cap s_{IIIij}} \frac{\tilde{e}_k}{\pi_{k|i,j}} \frac{1 - \pi_{k|i,j}}{p_k}}{\sum_{k \in R \cap s_{IIIij}} \frac{1 - \pi_{k|i,j}}{p_k}}$$

où m désigne le nombre d'unités primaires sélectionné, m_i désigne le nombre d'unités secondaires sélectionné dans l'unité primaire i , m_{ij} désigne le nombre d'unités tertiaires sélectionné dans l'unité secondaire j contenue dans l'unité primaire i , R l'ensemble des logements répondants et p_k la probabilité de réponse (estimée) du logement k .

Le calcul des termes \hat{V}_i pose problème car dans les unités primaires i , on ne sélectionne qu'un ou deux groupes de communes. Si un seul groupe de communes est échantillonné, l'estimation du premier terme

$$\frac{m_i}{m_i - 1} \sum_{j \in s_{III}} (1 - \pi_{j|i}) \left(\frac{\tilde{t}_{ij}}{\pi_{j|i}} - \tilde{R}_i \right)^2$$

de \hat{V}_i associé (notamment) au tirage des unités secondaires est impossible. Même quand deux groupes de communes sont échantillonnés, cette même estimation peut être impossible si un de ces groupes est sélectionné de façon certaine. On opère donc des regroupements d'unités primaires afin de disposer d'au moins deux regroupements de communes effectivement échantillonnés dans chacune (cf tableau 6). Notons que ce regroupement ne modifie que le calcul du 1^{er} terme des \hat{V}_i , les autres formules restant inchangées.

A2.4.5.2. Strate de gestion SG1

Dans la strate de gestion 1, l'échantillon est sélectionné selon un plan de sondage à 2 degrés. La variance est estimée en se basant sur l'estimateur sans biais pour un sondage à deux degrés. On note U_I la population d'unités primaires (ici, une UP ou une fraction d'UP au sens de l'échantillon maître) et U_{II} la population d'unités secondaires (ici, le logement). Soit s_I l'échantillon d'unités primaires sélectionné avec des probabilités d'inclusion $(\pi_i)_{i \in U_I}$. Chaque unité primaire $i \in s_I$ est partitionnée en H_i groupes de communes, et un échantillon s_{ih} de logements est sélectionné dans la

strate h de l'unité primaire i avec des probabilités d'inclusion $(\pi_{j|i})_{j \in i}$. On utilise l'estimateur de variance :

$$\hat{V}_p(\hat{t}_{\tilde{e}}^{N,anc,1}) = \frac{m}{m-1} \sum_{i \in s_I} (1 - \pi_i) \left(\frac{\tilde{t}_i}{\pi_i} - \tilde{R} \right)^2 + \sum_{i \in s_I} \hat{V}_i$$

avec

$$\tilde{t}_i = \sum_{h=1}^{H_i} \sum_{j \in s_{hi}} \frac{\tilde{e}_j}{\pi_{j|i} p_j}, \tilde{R} = \frac{\sum_{i \in s_I} \frac{\tilde{t}_i (1 - \pi_i)}{\pi_i}}{\sum_{i \in s_I} (1 - \pi_i)} \text{ et } \hat{V}_i = \sum_{h=1}^{H_i} \frac{m_{ih}}{m_{ih} - 1} \sum_{j \in R \cap s_{hi}} \frac{(1 - \pi_{j|i})}{p_j} \left(\frac{\tilde{e}_j}{\pi_{j|i}} - \tilde{R}_{hi} \right)^2$$

et $\tilde{R}_{hi} = \frac{\sum_{j \in R \cap s_{hi}} \frac{\tilde{e}_j (1 - \pi_{j|i})}{\pi_{j|i} p_j}}{\sum_{j \in R \cap s_{hi}} \frac{1 - \pi_{j|i}}{p_j}}$, m le nombre d'unités primaires sélectionné, m_{ih} le nombre d'unités

secondaires sélectionné dans la strate h de l'unité primaire i , R l'ensemble des logements répondants et p_j la probabilité de réponse (estimée) du logement j .

Comme pour la strate de gestion 0, le calcul des termes \hat{V}_i pose problème pour certains regroupements de communes dans lesquels un seul logement est observé. On réalise donc quelques regroupements de communes (cf tableau 7) afin de disposer d'au moins deux logements répondants dans chaque regroupement de communes.

Tableau 6 : Regroupement d'unités primaires pour l'estimation de la variance due au second degré de tirage dans la strate de gestion 0

Identifiant UP	Identifiant UP après regroupement	Identifiant groupe de communes	Identifiant groupe de communes après regroupement	Probabilité de sélection du groupe de communes dans l'UP
530025	530025	1	1	1
530025	530025	2	2	0,688
530030	530025	3	3	0,045
530053	530053	1	1	1
530055	530053	3	3	0,231
530072	530053	2	2	0,360
530073	530073	1	1	1
530073	530073	3	3	0,189
530077	530073	5	5	0,077
530077	530073	7	7	0,186
530097	530097	2	2	1
530097	530097	3	3	0,233
530100	530097	2	4	0,142
530121	530121	1	1	0,084
530127	530121	3	3	0,083
530132	530121	3	4	0,140
530144	530144	1	1	1
530144	530144	2	2	0,845
530145	530144	1	5	1
530145	530144	4	4	0,341

Tableau 7 : Regroupement de regroupements de communes pour l'estimation de la variance due au second degré de tirage dans la strate de gestion 1

Identifiant UP	Identifiant UP après regroupement	Identifiant groupe de communes	Identifiant groupe de communes après regroupement
531049	531049	1	1
531049	531049	2	1
531049	531049	5	5
531049	531049	6	5
531055	531055	1	1
531055	531055	2	1
531061	531061	1	1
531061	531061	2	1
531068	531068	1	1
531068	531068	3	1

A2.4.5.3. Strates de gestion SG2 et SG3

Dans ces deux strates, le tirage peut être assimilé à un sondage stratifié selon le groupe de communes. La variance est estimée en utilisant les résultats de l'annexe 2.3.

Annexe 3 : Exemples de mise en oeuvre de la macro

```

/*****
/* Calcul de précision Enquête Logement 2006 */
/*****
libname basetest "U:\SED\A1_TRAVAUX EN COURS\All_Projets\projets en
cours\Ext enquête logement\Qualité";

/* Nom du répertoire contenant la macro de calcul de précision */
%let repl=...;
/* Appel de la macro de calcul de précision */
%include "&repl.\calcul variance enquete logement 180608.sas";

/* Nom du répertoire contenant les tables supplémentaires nécessaires */
%let rep2=...;
/* Librairie contenant les tables supplémentaires nécessaires */
libname basefic "&rep2";

/*****
/* Exemple 1 */
/* Estimation, sur l'ensemble de la Bretagne, du nombre de */
/* logements par catégorie */
/*****
data adresse;
    set basefic.adresse;
    /* On s'intéresse à tous les logements --> domaine=1 pour chaque
logement */
    domaine=1;
run;
/* On récupère la table contenant les infos nécessaires */
data __aappariier; set basefic.__aappariier;
run;

```

```

%CALC_PREC_LOG(lib=work,tab_enq=adresse,aapparier=__aapparier,
               vartot_qual=catl06,
               percent=1,
               domaine=domaine1,
               libsor=work,sortie=sortie_bret);

/*****
/* Exemple 2
/* Sur l'ensemble des résidences principales de l'Aire Urbaine
/* de Rennes, estimation :
/* De la structure des logements selon le nombre de chambres,
/* Du nombre total de chambres,
/* De la surface moyenne par habitant.
*****/

/* On récupère les variables nécessaires */
proc sort data=basefic.adresse;by ident_log;run;
proc sort data=basefic.logement;by ident_log;run;
data adresse;
    merge basefic.adresse(in=ok)
          basefic.logement(keep=ident_log Hcha hsh1 nhab);
    by ident_log;if ok;
run;
/* On récupère la table contenant les infos nécessaires */
data __aapparier; set basetest.__aapparier; run;

data adresse;set adresse;
    /* On s'intéresse aux résidences principales de l'AU de Rennes */
    /* --> domaine=1 si CATL06="1" et aire="rennes" */
domaine2=(catl06="1")*(aire="rennes");
run;

%CALC_PREC_LOG(lib=work,tab_enq=adresse,
               vartot_qual=Hcha,
               vartot_quant=Hcha,
               varrat=hsh1/nhab,
               percent=1,
               domaine=domaine2,
               libsor=work,sortie=sortie_rennes_1);

/*****
/* Exemple 2 (suite)
/* Précision insuffisante pour les modalités 4 et + de la
/* variable donnant le nb de chambres
/* --> on regroupe
*****/
data adresse;
    set adresse;
    if Hcha=0 then Hcha_bis=0;
    else if Hcha=1 then Hcha_bis=1;
    else if Hcha=2 then Hcha_bis=2;
    else if Hcha=3 then Hcha_bis=3;
    else if Hcha>3 then Hcha_bis=4;
run;
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
               vartot_qual=Hcha_bis,
               vartot_quant=Hcha,
               varrat=hsh1/nhab,
               domaine=domaine2,
               libsor=work,sortie=sortie_rennes2);

```


Annexe 4 : Modélisation pour la formule approchée de variance

A4.1. Préparation du fichier

```
/* Nom du répertoire contenant la macro de calcul de précision */
%let rep1=. . .;
/* Appel de la macro de calcul de précision */
%include "&rep1.\calcul variance enquete logement 180608.sas";

/* Nom du répertoire contenant les tables supplémentaires nécessaires */
%let rep2=. . .;
/* Librairie contenant les tables supplémentaires nécessaires */
libname basetest "&rep2";

/* Nom du répertoire contenant les tables de l'échantillon */
%let rep3=. . .;
/* Librairie contenant les tables supplémentaires nécessaires */
libname basefic "&rep3";

/* On récupère les variables nécessaires */
proc sort data=basefic.adresse;by ident_log;run;
proc sort data=basefic.logement;by ident_log;run;
data adresse;
    merge basefic.adresse(in=ok)
          basefic.logement(keep=ident_log
                           Hcha /* Nb de chambres */
                           Hsht /* surface habitable (tranches) */
                           nhab /* Nb d'habitants */
                           Hnph1 /* Nb de pièces d'habitation */
                           Kse /* Nb de salles d'eau */
                           Kwc /* Existence de WC */
                           Kca /* Existence d'une cave */);
by ident_log;if ok;run;

/* On récupère la table contenant les infos pour le calcul de précision */
data __aapparier;set basetest.__aapparier;run;

data adresse;set adresse;
    domaine=(catl06="1");
    domaine_bre=(catl06="1")*(aire="brest");
    domaine_lor=(catl06="1")*(aire="lorient");
    domaine_qui=(catl06="1")*(aire="quimper");
    domaine_ren=(catl06="1")*(aire="rennes");
    domaine_stb=(catl06="1")*(aire="saint-brieuc");
    domaine_van=(catl06="1")*(aire="brest");
run;

/* Estimation pour les domaines associés à la Bretagne */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
               vartot_qual=Hcha Hsht nhab hnph1 kse kwc kca,
               percent=0,
               vartot_quant=,
               varrat=,
               domaine=domaine,
               libsor=work,sortie=sortie_bret);
```

```

/* Estimation pour les domaines associés à l'AU de Brest */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_bre,
    libsor=work,sortie=sortie_bre);

/* Estimation pour les domaines associés à l'AU de Lorient */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_lor,
    libsor=work,sortie=sortie_lor);

/* Estimation pour les domaines associés à l'AU de Quimper */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_qui,
    libsor=work,sortie=sortie_qui);

/* Estimation pour les domaines associés à l'AU de Rennes */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_ren,
    libsor=work,sortie=sortie_ren);

/* Estimation pour les domaines associés à l'AU de Saint-Brieuc */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_stb,
    libsor=work,sortie=sortie_stb);

/* Estimation pour les domaines associés à l'AU de Vannes */
%CALC_PREC_LOG(lib=work,tab_enq=adresse,
    vartot_qual=Hcha Hsht nhab hnphl kse kwc kca,
    percent=0,
    vartot_quant=,varrat=,domaine=domaine_van,
    libsor=work,sortie=sortie_van);

data sortie_bret;set sortie_bret;zone="bretagne";run;
data sortie_bre;set sortie_bre;zone="brest";run;
data sortie_lor;set sortie_lor;zone="lorient";run;
data sortie_qui;set sortie_qui;zone="quimper";run;
data sortie_ren;set sortie_ren;zone="rennes";run;
data sortie_stb;set sortie_stb;zone="saint-brieuc";run;
data sortie_van;set sortie_van;zone="vannes";run;

/* On supprime les estimations de variance associées à l'estimation du
nombre de résidences principales ayant un seul individu (le calage assure
que cette variance est théoriquement nulle) */
data sortie;set sortie_bret sortie_bre sortie_lor
    sortie_qui sortie_ren sortie_stb sortie_van;
    if _parametre_="nhab_1" then delete;
    y=_VARIANCE_/(_ESTIM_**2);x=1/_ESTIM_;
run;

```

A4.2. Ajustement des modèles de régression

A4.2.1. Modèle sur les données complètes (avec constante)

```
proc reg data=sortie;model y=x;quit;
```

```
          The REG Procedure
          Model: MODEL1
    Dependent Variable: y

    Analysis of Variance

    Source                DF          Sum of          Mean
    Model                  1          70.52388          70.52388
    Error                  281          17.93948           0.06384
    Corrected Total       282          88.46336

    Root MSE              0.25267    R-Square          0.7972
    Dependent Mean        0.18940    Adj R-Sq          0.7965
    Coeff Var             133.40794
```

```
          Parameter Estimates

    Variable    DF      Parameter      Standard
    Estimate    Error    t Value    Pr > |t|

    Intercept    1      0.03672      0.01571
    x             1     108.25773      3.25719      2.34      0.0201
                   33.24      <.0001
```

A4.2.2. Modèle sur les données complètes (sans constante)

```
proc reg data=sortie;model y=x / noint;quit;
```

```
          The REG Procedure
          Model: MODEL1
    Dependent Variable: y

    Analysis of Variance

    Source                DF          Sum of          Mean
    Model                  1          80.32627          80.32627
    Error                  282          18.28851           0.06485
    Corrected Total       283          98.61478

    Root MSE              0.25466    R-Square          0.8145
    Dependent Mean        0.18940    Adj R-Sq          0.8139
    Coeff Var             134.46043
```

```
          Parameter Estimates

    Variable    DF      Parameter      Standard
    Estimate    Error    t Value    Pr > |t|

    x             1     110.48508      3.13935      35.19      <.0001
```

A4.2.3. Modèle hors domaines associés à la variable Hcha

```
proc reg data=sortie(where=(substr(_PARAMETRE_,1,4) ne "Hcha"));
  model y=x / noint;quit;
```

```

The REG Procedure
  Model: MODEL1
  Dependent Variable: y

Analysis of Variance

Source                DF          Sum of          Mean
                    Squares          Square    F Value    Pr > F
Model                  1      60.45469      60.45469    889.81    <.0001
Error                 227      15.42257      0.06794
Corrected Total       228      75.87725

Root MSE              0.26065    R-Square      0.7967
Dependent Mean        0.18541    Adj R-Sq      0.7958
Coeff Var             140.58113

Parameter Estimates

Variable    DF      Parameter      Standard
          Estimate      Error    t Value    Pr > |t|
x              1      115.11015      3.85890    29.83    <.0001

```

A4.2.4. Modèle hors domaines associés à l'AU de Vannes

```
proc reg data=sortie(where=(zone ne "vannes"));
  model y=x / noint;quit;
```

```

The REG Procedure
  Model: MODEL1
  Dependent Variable: y

Analysis of Variance

Source                DF          Sum of          Mean
                    Squares          Square    F Value    Pr > F
Model                  1      75.88957      75.88957   1118.77    <.0001
Error                 243      16.48344      0.06783
Corrected Total       244      92.37301

Root MSE              0.26045    R-Square      0.8216
Dependent Mean        0.19388    Adj R-Sq      0.8208
Coeff Var             134.33740

Parameter Estimates

Variable    DF      Parameter      Standard
          Estimate      Error    t Value    Pr > |t|
x              1      108.50285      3.24393    33.45    <.0001

```

Tableau 8 : Estimation du nombre de résidences principales selon le nombre de chambres, la surface habitable, le nombre d'habitants, le nombre de pièces d'habitation, le nombre de salles d'eau, l'existence de WC, l'existence d'une cave, pour l'ensemble de la Bretagne

Paramètre à estimer	Estimation	Variance	CV (en %)	Borne Inf int. de confiance à 95 %	Borne Sup int. de confiance à 95 %
Hcha_0	60 366	6,7E+07	13,5	44 354	76 377
Hcha_1	186 301	1,1E+08	5,8	165 319	207 283
Hcha_2	338 267	1,9E+08	4,1	311 207	365 327
Hcha_3	382 244	2,3E+08	4,0	352 381	412 106
Hcha_4	303 617	1,9E+08	4,5	276 912	330 322
Hcha_5	66 495	5,4E+07	11,1	52 100	80 890
Hcha_6	6 771	5,3E+06	34,0	2 257	11 285
Hcha_7	632	6,7E+04	41,0	124	1 141
Hcha_8	148	1,4E+04	80,7	-86	383
Hsht_1	6 981	8,8E+05	13,4	5 147	8 814
Hsht_2	67 457	9,4E+07	14,4	48 487	86 427
Hsht_3	152 095	9,0E+07	6,2	133 494	170 697
Hsht_4	280 462	1,3E+08	4,1	258 035	302 889
Hsht_5	263 273	2,5E+08	6,0	232 536	294 009
Hsht_6	215 623	1,6E+08	5,8	191 079	240 167
Hsht_7	358 951	2,5E+08	4,4	328 200	389 702
nhab_2	446 658	6,7E+07	1,8	430 659	462 656
nhab_3	190 434	1,4E+08	6,2	167 421	213 446
nhab_4	162 832	8,8E+07	5,8	144 460	181 204
nhab_5	71 966	4,7E+07	9,5	58 507	85 425
nhab_6	7 715	1,2E+07	45,7	806	14 625
nhab_7	1 043	2,0E+05	42,5	174	1 911
nhab_8	333	5,8E+04	72,2	-138	803
nhab_9	48	1,1E+04	217,9	-156	251
hnph1_1	56 165	6,5E+07	14,4	40 325	72 006
hnph1_2	165 461	8,8E+07	5,7	147 036	183 885
hnph1_3	238 780	1,3E+08	4,8	216 395	261 165
hnph1_4	282 150	1,9E+08	4,8	255 374	308 926
hnph1_5	289 970	2,1E+08	5,0	261 536	318 403
hnph1_6	189 492	1,4E+08	6,3	166 178	212 805
hnph1_7	91 735	8,9E+07	10,3	73 217	110 253
hnph1_8	20 515	1,8E+07	20,8	12 169	28 861
hnph1_9	7 974	7,8E+06	34,9	2 515	13 432
hnph1_10	922	2,0E+05	48,4	47	1 796
hnph1_11	131	1,3E+04	87,6	-94	356
hnph1_12	1 547	1,3E+06	73,2	-674	3 768
kse_0	12 307	1,9E+07	35,2	3 822	20 792
kse_1	987 618	2,6E+08	1,6	955 945	1 019 292
kse_2	344 916	2,7E+08	4,7	312 850	376 981
kwc_1	1 179 650	1,2E+08	0,9	1 158 250	1 201 049
kwc_2	153 645	1,1E+08	6,9	132 764	174 526
kwc_3	4 387	8,9E+05	21,5	2 541	6 232
kwc_4	7 160	9,9E+06	43,9	1 006	13 314
kca_1	556 448	2,6E+08	2,9	524 650	588 246
kca_2	788 393	2,6E+08	2,1	756 595	820 191