Problèmes théoriques et pratiques de la mise en œuvre d'une sur-représentation des ZUS dans les échantillons d'enquête : le cas de l'enquête IVQ.

Laurent DAVEZIES (*),Cédric LANDRE (**),
Fabrice MURAT (***) et Sylvie ROUSSEAU (**)

(*) Ministère de l'Education Nationale, DEP, Sous-Direction des Etude Statistiques
(**) INSEE, DSDS, Unité Méthodes Statistiques
(***) INSEE, DSDS, Département de l'Emploi et des Revenus d'Activité

1. Présentation de l'enquête IVQ2004.

L'enquête Information et Vie quotidienne (IVQ) porte sur un domaine encore assez mal connu : l'évaluation des compétences d'adultes en matière de littéracie et numératie. Les enquêtes statistiques menées par l'Insee sur l'illettrisme avaient jusqu'à récemment une base déclarative : la personne enquêtée indiquait si elle éprouvait des difficultés à lire des journaux, à remplir un chèque, etc. Une telle approche est très subjective et l'usage de tests pour vérifier ces déclarations est vite apparu nécessaire. Des exercices et un protocole d'enquête spécifiques ont dû être construits, en tenant compte des contraintes d'une enquête ménage conduite auprès d'une population d'adultes.

L'objectif de l'enquête est aussi bien de connaître les conséquences de l'illettrisme (les handicaps dans la vie personnelle et professionnelle) que ses causes potentielles (parcours familial et scolaire). Les données permettront de construire des indicateurs pour guider la politique de lutte contre l'illettrisme, d'évaluer le système éducatif et d'affiner l'analyse du marché du travail (en mesurant les compétences individuelles autrement que par le diplôme).

L'architecture de l'enquête prévoit une adaptation des épreuves au niveau de la personne interrogée. L'interview commence par un exercice d'orientation sur un support de la vie quotidienne (une page de programme TV). Si la personne n'obtient pas de résultats suffisants, alors elle passe les exercices simples du module « ANLCI » (du nom de l'Agence Nationale de Lutte Contre l'Illettrisme, à l'origine de ce module). Si au contraire, elle obtient de bons résultats, alors elle est interrogée sur les exercices du module « Haut ». Une personne dans une situation intermédiaire à l'issue de l'exercice d'orientation passe un module « intermédiaire » qui permet d'affiner le diagnostic avant d'orienter le questionnement définitif vers l'un des modules « ANLCI » ou « Haut ». Dans tous les cas, quels que soient les résultats à l'exercice d'orientation, la personne passe un test de compréhension orale, avant les modules ANLCI, « Haut » ou « Intermédiaire ». Viennent après ces différents modules, des questions sur la « numératie » : ces items portent sur la maîtrise des compétences de base en calcul et en raisonnement logique (là encore, le niveau de difficulté des exercices est adapté en fonction des réponses à quelques questions simples insérées dans le module d'orientation). Ensuite, des informations sur le parcours familial, scolaire et professionnel de la personne interrogée sont recueillies.

Une opération analogue (IVQ2002) s'était déroulée du 4 novembre au 14 décembre 2002, dans dix régions auprès d'un échantillon d'environ 4000 logements. Du fait des non-réponses (ménages hors-champ et refus), le nombre de répondants était de 2 083. Les premiers résultats (voir [3]) ont montré en particulier qu'environ 12 % de la population apparaissait éprouver des difficultés à lire, écrire ou compter.

L'objectif principal de la reprise de l'opération en 2004 (IVQ2004) était de permettre des études plus fines sur cette population des personnes lisant, écrivant ou comptant avec difficulté. En effet, l'échantillon de 2002 compte à peine plus de 200 répondants dans cette catégorie, ce qui limite les analyses. Il a donc été prévu de reconduire l'enquête auprès d'un échantillon plus grand et de le sélectionner selon un plan de sondage qui permette d'obtenir suffisamment de personnes ayant des difficultés à lire, écrire ou compter. Ainsi, les ménages où l'on avait le plus de chance de trouver une personne ayant de telles difficultés ont été sur-représentés dans l'échantillon (les critères précis sont présentés plus loin).

D'autre part, des contacts avec des partenaires régionaux ont été pris lors de la préparation de l'enquête. Ainsi, des extensions d'échantillon ont été financées en Aquitaine, dans le Nord-Pas-de-Calais et dans les Pays de la Loire. Dans d'autres régions, des partenaires locaux avaient aussi manifesté un intérêt très marqué pour l'opération mais des contraintes de temps et de financement n'ont pas permis à tous ces projets d'aboutir.

Enfin, la Délégation Interministérielle à la Ville (DIV), un des financeurs de l'enquête, souhaitait pouvoir produire des études dans les Zones Urbaines Sensibles (ZUS). Pour satisfaire cette demande, un échantillon spécifique de logements localisés en ZUS a été sélectionné ; son objectif est de garantir un nombre minimal de répondants dans ces zones.

2. Echantillonnage.

2.1 Les enjeux de l'échantillonnage

L'enquête IVQ2004 possède deux objectifs d'exploitation spécifiques qui la distinguent de la plupart des enquêtes ménages nationales métropolitaines et qui doivent être appréhendés par l'échantillonnage :

- Trois régions (Nord-Pas-de-Calais, Aquitaine et Pays de la Loire) souhaitaient disposer d'estimations à l'échelle régionale et ont financé dans ce but une extension d'échantillon.
- Les différents partenaires de l'opération souhaitaient cibler des populations rares :
  - La première population ciblée est constituée des personnes ayant des difficultés avec la lecture.
  - La deuxième population ciblée est constituée par les personnes vivant dans une ZUS.

Le budget global, tous financements réunis, a été établi pour 17 400 logements.

Tableau 1 : Les tailles d'échantillon prévues selon les sources de financement (en nombre de logements)

Source du financement	***Effectif à échantillonner***
Financement national	12 000
Financement national extension ZUS	1 000
Financement régional	
Nord-Pas-de-Calais	*1 500*
Pays de Loire	*1 500*
Aquitaine	*1 400*
TOTAL	***17 400***

2.2 Les bases de sondage

L'unité statistique interrogée dans l'enquête est l'individu. Cependant, en l'absence de base de sondage à ce niveau, c'est par l'intermédiaire des logements et des ménages que les individus sont approchés. Plus précisément, le champ de l'enquête est défini, au moment du passage de l'enquêteur, par l'ensemble des personnes âgées de 18 à 65 ans qui résident dans des résidences principales. Dans chaque ménage d'une résidence principale, l'enquêteur liste les individus dans le champ, classe leurs prénoms par ordre alphabétique et interroge la première personne présente de cette liste.

Pour la plupart des enquêtes ménages nationales métropolitaines[1], les bases de sondage sont l'Echantillon-Maître (EM), qui est une *"réserve"* localisée de logements construite à partir du dernier recensement, et la base de sondage des logements neufs (BSLN) qui complète l'EM pour couvrir les logements construits après mars 1999. Ces deux bases ont été complétées pour répondre aux deux objectifs particuliers de l'enquête :

- **Pour établir des résultats régionaux avec une précision acceptable**, il est nécessaire d'augmenter la taille des échantillons dans les régions concernées car l'EM, par construction, assure la pertinence des résultats nationaux mais non pas celles des statistiques régionales calculées sur la partie régionale de l'échantillon national. Pour l'enquête IVQ2004, les financements locaux ont permis de tripler la taille de l'échantillon dans les trois régions concernées par rapport à celle qui aurait été obtenue en absence d'extension. A ce surcroît d'échantillon, répond une base de sondage supplémentaire : l'échantillon-maître pour les extensions régionales (EMEX[2]). Conçu en 2001 pour homogénéiser le tirage et les traitements des extensions régionales associées à des enquêtes nationales, l'EMEX est une réserve localisée de logements recensés qui permet d'éviter les problèmes de réserve insuffisante par ponction excessive dans l'EM. Il permet de surcroît au niveau national de tirer parti des extensions en exploitant simultanément les données nationales et régionales. Ces dernières alimentent, quant à elles, dans les régions à extension, les exploitations régionales et les comparaisons inter-régionales. En pratique, dans les régions à extension, l'échantillon des logements recensés est sélectionné dans la base de sondage formée de la réunion « EM + EMEX » qui est « représentative»[3] de chaque région au regard de critères d'âge et de revenu. De manière analogue, pour les logements construits après le recensement, la base de sondage des logements neufs est enrichie dans les régions à extension par des logements issus de permis de construire déclarés achevés par le Ministère de l'Equipement[4].
- **L'autre enjeu de l'échantillonnage d'IVQ2004 avait trait au ciblage de deux populations rares :**
    - **La population des personnes ayant des difficultés avec la lecture** pouvait se prêter à un tirage dans l'EM à condition de sur-représenter les ménages de cette population. Leurs caractéristiques ont été mises en évidence à partir de l'enquête IVQ2002 et sont exposées au paragraphe 2.4. Afin de préserver l'équilibrage de la base de sondage, un tel tirage s'effectue en deux temps : d'abord, un échantillon dit « de 1ère phase » est sélectionné à probabilités égales, puis, une 2ème phase de tirage permet d'introduire les probabilités inégales dans l'échantillon obtenu précédemment. Enfin, dans la base de sondage, une marque « logement non enquêtable à l'avenir » est attribuée à tous les logements de l'échantillon de 1ère phase.

[1] Hors l'enquête Emploi, l'enquête Camme et quelques enquêtes ad-hoc comme Construction des identités, Seniors et immigrés, etc.
[2] Depuis 2001, l'enquête « IVQ » a été la 2ème enquête faisant l'objet d'extensions régionales.
[3] Au sens du tirage équilibré sur des critères d'âge et de revenu. Pour en savoir plus, voir [4] et [5].
[4] Source : SITADEL (Système d'Information et de Traitement Automatisé des Données Elémentaires sur le Logement et les locaux).

- **Le ciblage des personnes vivant en ZUS** posait problème car la construction de l'EM a ignoré ce zonage qui ne couvre qu'environ 6% des logements recensés. Dans ces conditions, un tirage dans l'EM sur-représentant les logements en ZUS aurait conduit à ponctionner très lourdement cette base de sondage, menaçant ainsi la capacité de la réserve pour les enquêtes ultérieures. Il a donc fallu développer une stratégie d'échantillonnage ad-hoc pour satisfaire la demande de la DIV, et en particulier, constituer une base de sondage spécifique pour les logements en ZUS. Cette base de sondage répond aussi à la demande croissante[5] de statistiques sur les ZUS, notamment pour analyser les disparités de comportements des ménages vivant en ZUS ou hors ZUS. Construite en juin 2004, elle liste tous les logements ordinaires recensés en mars 1999 et localisés dans des ZUS[6]. Comme l'EM, cette base ne concerne donc pas la construction neuve. Notons qu'il est actuellement pratiquement impossible d'échantillonner correctement des logements neufs en ZUS car la BSLN ne couvre qu'une partie des ZUS (en l'occurrence, les ZUS qui se trouvent dans les communes de l'EM) et le fichier SITADEL des permis de construire déclarés achevés n'est pas îloté à ce jour ce qui ne permet pas de repérer les logements en ZUS.

2.3 L'échantillonnage.

Deux tirages indépendants ont été réalisés :

- un tirage classique dans l'EM, complété par l'EMEX dans les régions à extension, et la BSLN, enrichie de logements issus des permis de construire déclarés achevés par le Ministère de l'Equipement dans les régions à extension.
- un tirage dans la base ZUS, sans extension régionale.

L'échantillon final réunit les deux parties prélevées dans les différentes bases (cf. tableau 3 et figure 1). Il liste 17 407 logements.

Chacun de ces deux tirages obéit à un plan en deux phases, où la deuxième étape permet de cerner la population du champ à la date du dernier recensement et de sur-représenter le cas échéant les catégories d'intérêt. Plus précisément :

- le tirage de 1ère phase s'effectue de sorte que tous les ménages aient la même chance d'être interrogé quelle que soit leur localisation sur le territoire. Ici, un distinguo s'établit naturellement dans les régions à extension où cette probabilité s'accroît avec la taille de l'échantillon, à hauteur des financements locaux. Partout ailleurs, le taux de sondage de tous les logements principaux est identique. Ce taux s'applique aussi aux logements neufs que l'on considère destinés à l'habitat principal. Par ailleurs, pour tenir compte des changements de catégorie de logement intervenus depuis le recensement de mars 1999, l'échantillon comprend également des logements déclarés occasionnels, secondaires ou vacants au moment du RP (cf. tableau 3). Leurs taux de sondage sont cependant moindres que celui des résidences principales compte-tenu du champ d'interrogation.

[5] Utilisée la première fois à l'occasion du tirage de l'enquête IVQ2004, la base ZUS a depuis été utilisée pour deux autres enquêtes : EPCV-Victimisation de janvier 2005 et Violence et Santé de novembre 2005. L'enquête Logement de 2006 prévoit également une extension ZUS.

[6] Il s'agit plus précisément des logements d'une liste d'îlots définie par l'INSEE en 2000. Cette liste a été constituée pour approcher au mieux le périmètre des 717 ZUS de France métropolitaine (Corse comprise). La base ZUS contient 1 842 744 logements ordinaires, dont 98.2% ont été considérés par les directions régionales comme « facilement enquêtables » par le réseau des enquêteurs « ménage » de l'INSEE. En effet, échantillonner dans les ZUS pose des problèmes de collecte car ce réseau ne couvre que les communes de l'échantillon maître et les aires emploi. Avant d'engager un échantillonnage spécifique dans les ZUS, il convenait de s'assurer que l'ensemble des logements en ZUS pouvait être facilement enquêtable par le réseau des enquêteurs traditionnels. Après consultation des directions régionales, la « représentativité » des logements en ZUS susceptibles d'être enquêtés en mobilisant les enquêteurs permanents a été considérée comme correcte.

- Le tirage de 2ème phase opère sur les logements principaux obtenus dans l'échantillon de 1ère phase et introduit des coefficients de sur-représentation qui permettent à la fois d'éliminer les ménages qui n'appartenaient pas au champ de l'enquête en 1999[7] et de viser les populations d'intérêt.
    - Pour la partie échantillonnée dans l'EM+EMEX, la population d'intérêt est composée des logements déclarés principaux au RP99 et où habitaient des ménages particulièrement susceptibles d'avoir des difficultés avec la lecture. La détermination de ces critères à partir des résultats de l'enquête IVQ2002 est développée dans le paragraphe suivant.
    - Pour la partie échantillonnée dans la base ZUS, contrairement à la partie précédente, la 2ème phase permet simplement de cerner les logements du champ d'âge au RP99 mais ne sur-représente aucune population particulière, les caractéristiques des personnes vivant en ZUS étant assez proches de celles ayant des difficultés avec la lecture d'après les résultats de l'enquête IVQ2002.

2.4 La détermination des critères de sur-représentation lors du tirage dans l'EM et l'EMEX.

Il s'agit ici d'exploiter l'enquête IVQ2002 afin de déterminer :

- les caractéristiques des personnes ayant des difficultés avec la lecture et l'écriture
- puis les coefficients de sur-représentation à appliquer aux ménages possédant ces caractéristiques

A la date de cette étude, le nombre total de logements à échantillonner pour IVQ2004 n'était pas encore fixé, notamment parce que les projets d'extensions régionales n'avaient pas tous abouti. Les résultats présentés par la suite portent sur une hypothèse de 12 000 logements échantillonnés.

- Dans un premier temps, nous avons recherché les caractéristiques de la population visée parmi les variables présentes dans les bases de sondage[8]. Nous avons ainsi modélisé le fait d'éprouver des difficultés à lire ou écrire au moyen d'une régression logistique effectuée sur les répondants à l'enquête IVQ2002 et avec les variables suivantes:
    - Le genre de la personne de la personne de référence du ménage,
    - Le nombre de voitures possédées par le ménage (aucune ou une contre deux ou plus),
    - La nationalité de la personne de référence du ménage (française ou étrangère),
    - L'âge de la personne de référence du ménage (regroupé en 6 classes : moins de 25 ans, de 25 à 35 ans, de 35 à 45 ans, de 45 à 55 ans, de 55 à 65 ans, 65 ans ou plus),
    - La typologie Tabard [2] en 27 postes (les modalités de la typologie sont présentées en annexe, différents regroupements de ces 27 modalités ont été testés),
    - Le diplôme de la personne de référence du ménage,
    - Le fait que le logement soit ou non un HLM.

[7] Le champ de l'enquête correspond aux ménages qui comptent au moment de l'enquête au moins une personne de 18 à 65 ans inclus (ou ce qui revient au même née entre 1939 et 1986 inclus). Le tirage de 2ème phase exclut les ménages repérés hors champ d'après les informations datant du RP99.

[8] L'échantillon maître et la BSLN (l'EM contient des variables collectées au recensement de 1999).

Cette étude permet de conclure que la probabilité d'éprouver des difficultés à lire, compter ou écrire augmente significativement dans les situations suivantes :

- le ménage vit dans des quartiers plutôt « défavorisés » (selon la typologie Tabard, caractérisés par des taux élevés de chômeurs ou d'ouvriers peu qualifiés[9]),
- le ménage possède moins de deux voitures,
- la personne de référence du ménage est de nationalité étrangère,
- a un niveau d'études primaire ou collège,
- possède comme diplôme le plus élevé le brevet, un bep ou un cap.

En revanche, le genre de l'individu, son âge et l'habitat en HLM ne semble pas avoir d'effet « toutes choses égales par ailleurs » sur le fait d'éprouver des difficultés avec la lecture.

- Dans un second temps, nous avons étudié l'influence de ces variables sur la non-réponse, en exploitant toujours les données de l'enquête IVQ2002. Nous avons ajouté dans notre analyse les variables supplémentaires suivantes :
  - le nombre de pièces du logement,
  - le nombre d'occupants du logement,
  - et la présence d'un digicode à l'entrée de l'immeuble.

  Nous concluons que le comportement de non réponse est significativement renforcé par :
  - la présence d'un digicode à l'entrée du logement,
  - le fait que la personne de référence du ménage soit âgé de plus de 65 ans,
  - et le fait que son niveau d'étude soit au plus celui du collège.

  Par contre, habiter certains quartiers plutôt « défavorisés »[10] au regard de la typologie Tabard, semble augmenter la probabilité de réponse.

Après cette étude préalable, nous avons envisagé différentes stratégies. Celle que nous avons retenue sur-représente avec un facteur 4 la population ainsi définie :

- habiter un quartier à fort taux de chômeurs ou de travailleurs peu qualifiés[11],
- ou être sans diplôme
- ou être de nationalité étrangère.

La population ainsi définie représente 38 % des ménages recensés au RP99. Dans l'enquête IVQ 2002, 11 % des personnes appartenant à ce groupe avaient répondu à l'enquête et montraient des difficulté à répondre correctement aux exercices.

Nous avons calculé les effectifs attendus dans différents groupes sous l'hypothèse d'un échantillon de 12 000 logements (les résultats concernent les logements principaux qui composent 90 % de l'échantillon). Il convient d'ajouter à ces effectifs environ 1400 logements occasionnels, secondaires, vacants ou neufs, ce qui permet d'espérer environ 70 répondants supplémentaires dans la population cible. Cette stratégie permettait a priori d'espérer plus de 1000 répondants en difficulté avec les exercices de l'enquête sans compter les extensions régionales ou l'extension ZUS.

---

[9] Les catégories Tabard significatives sont les suivantes : « INDOUV » (regroupement des modalités INDOUV1, INDOUV3, INDOUV4, INDOUV5), « AGRI » (regroupement des modalités AGRI12, AGRI13, AGRI21,AGRI22), « CHOMA » (regroupement des modalités CHOMA1, CHOMA2, CHOMA3, CHOMA4), « INDQ5 »(modalité INDQ5) et « INDQ24 » (regroupement des modalités INDQ2, INDQ3, INDQ4).

[10] Les catégories Tabard significatives sont ici : « INDOUV » (regroupement des modalités INDOUV1, INDOUV3, INDOUV4, INDOUV5), « ADPUB » (regroupement des modalités ADPUB1, ADPUB3), « AGRI » (regroupement des modalités AGRI12, AGRI13, AGRI21,AGRI22), et « INDQ24 » (regroupement des modalités INDQ2, INDQ3, INDQ4).

[11] Quartiers associés aux types « CHOMA » (regroupement des modalités CHOMA1, CHOMA2, CHOMA3, CHOMA4) ou « INDQ5 » (modalité INDQ5) de la typologie Tabard.

Tableau 2 : prévision du nombre de répondants en difficulté sur les exercices d'IVQ en fonction du coefficient de sur-représentation adopté
(partie échantillonnage national dans l'EM sur la base de 12000 unités échantillonnées)

Coefficient de sur-représentation[12] / Effectif attendu	1	2	4
Nombre de personnes échantillonnées du groupe sur-représenté	4 788	6 641	**8 234**
Nombre de répondants attendus en difficulté sur la lecture et l'écriture	755	911	**1 044**
Nombre de répondants attendus dans les quartiers sur-représentés	692	960	**1 190**
Nombre de répondants attendus sans diplôme	1 743	2 418	**2 998**
Nombre de répondants attendus de nationalité étrangère	521	722	**895**

Ces effectifs prévus à la date de l'étude étaient à prendre avec prudence pour de nombreuses raisons : l'enquête IVQ2002 n'a eu lieu que dans dix régions et pouvait être soumise à des effets régionaux ; ses résultats sont soumis à un aléa de sondage relativement fort car l'échantillon d'IVQ2002 était d'assez petite taille ; enfin, les résultats du recensement étant de plus en plus datés, les liaisons entre les variables observées en 2004 et celles collectées au RP99 sont sans doute moins fortes que les liaisons entre les variables observées en 2002 et celles du RP99. Enfin, au moment du tirage, le critère « nationalité étrangère » n'a finalement pas pu être retenu pour des raisons pratiques.

2.5 Bilan de l'échantillonnage

Tableau 3 : Nombre de logements échantillonnés par catégorie de logement et base de sondage

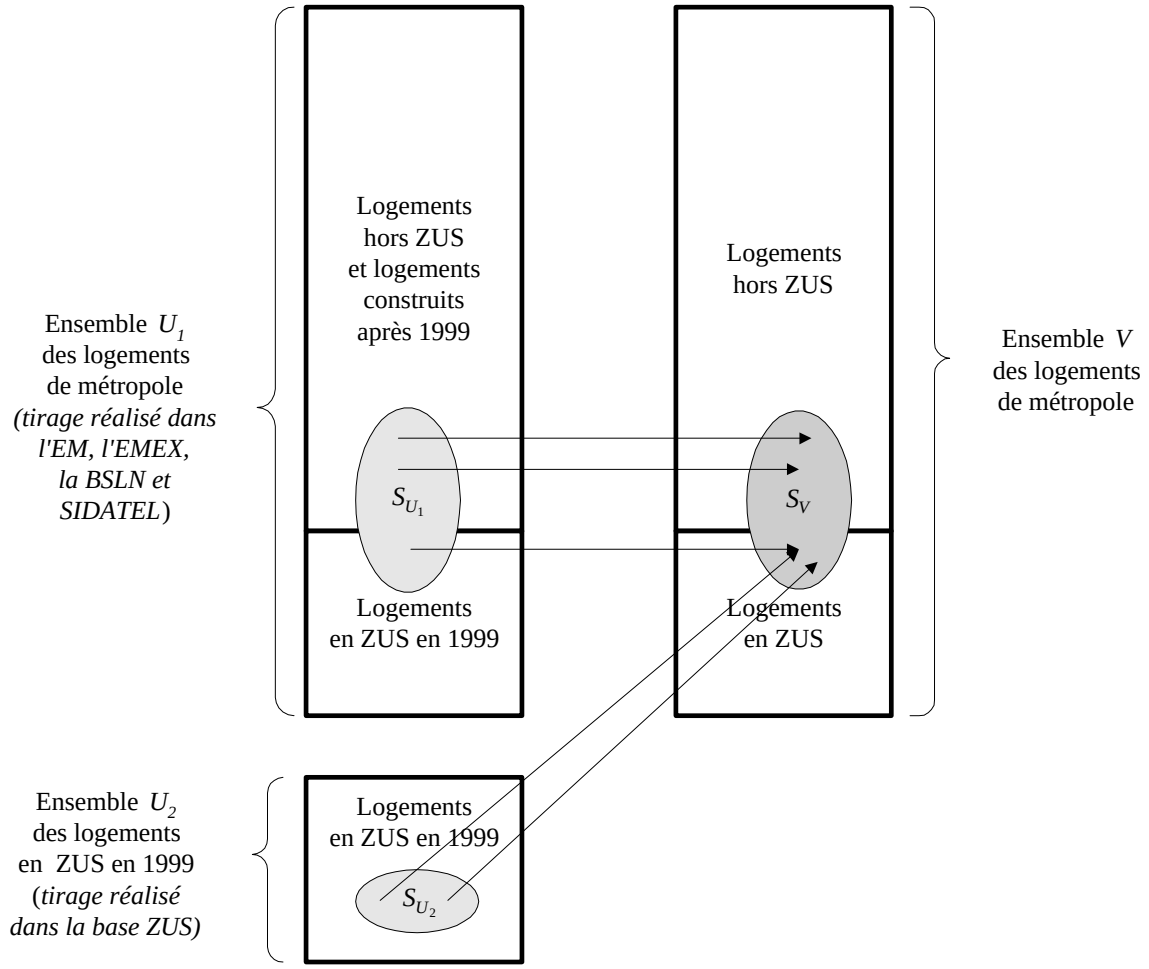
Base de sondage	*Catégorie de logement*	*Nord-Pas-de-Calais*	*Pays de la Loire*	*Aquitaine*	*Autres régions*	*Total*
Echantillonnage grâce à l'EM, l'EMEX, la BSLN et Sitadel	Résidences principales au RP99	1 733	1 414	1 253	6 780	11 180
	Dont groupe sur-représenté	*1 301*	*933*	*774*	*4 540*	*7 548*
	Dont groupe non sur-représenté	*432*	*481*	*479*	*2 240*	*3 632*
	Résidences secondaires ou occasionnelles au RP99	32	112	132	619	895
	Résidences vacantes au RP99	233	195	244	1 408	2 080
	Logements achevés après le RP99	215	383	360	1 290	2 248
	TOTAL	***2 213***	***2 104***	***1 989***	***10 097***	***16 403***
Echantillonnage grâce à la base ZUS	Résidences principales au RP99	88	40	34	739	901
	Résidences secondaires ou occasionnelles au RP99	0	0	0	3	3
	Résidences vacantes au RP99	9	3	2	86	100
	Logements achevés après le RP99	0	0	0	0	0
	TOTAL	***97***	***43***	***36***	***828***	***1 004***
TOTAL		**2 310**	**2 147**	**2 025**	**10 925**	**17 407**

Une fois déterminé l'échantillon des logements à enquêter, restait à traiter le problème de leur pondération que nous allons maintenant exposer.

[12] Les coefficients de sur-représentation utilisés pour échantillonner les enquêtes ménages ne dépassent généralement pas quatre ; en effet sur-représenter un groupe par 4, revient à éliminer quatre fois plus de logements de l'échantillon maître que ce qui devrait l'être pour le même nombre de personnes interrogées pour la même enquête sans sur-représentation.

3. Pondération de l'échantillon

Figure 1: L'échantillonnage d'IVQ2004



3.1 Systèmes de pondération envisagés.

3.1.1 Pondération d'Horvitz-Thompson

Nous notons V notre population d'intérêt (il s'agit dans notre cas de l'ensemble des ménages de métropole). Pour n'importe quel échantillon aléatoire s_V de la population V, nous cherchons à obtenir un système de pondération W telle que l'on ait : $E\left(\sum_{i\in s_V} w_i x_i\right) = X$ où X désigne le total sur la population d'intérêt d'une variable quelconque $\{x_i\}_{i\in V}$. Dans le cas où ce système de pondération est déterministe (c'est-à-dire que les poids ne sont pas aléatoires), alors nécessairement pour tout $i \in V$, $w_i = \frac{1}{\pi_i}$ où $\pi_i = P(i \in s_V)$. On obtient ainsi l'estimateur d'Horvitz-Thompson, que nous notons $\hat{X}^{HT}$.

Dans le cas qui nous intéresse, on procède à deux tirages indépendants :

- un premier tirage dans l'ensemble U_1 des logements de métropole (le premier degré de tirage étant le choix de l'échantillon maître, de l'échantillon maître pour les extensions

régionales et de la base de sondage des logements neufs éventuellement enrichie de SIDATEL)

- et un deuxième tirage U_2 dans l'ensemble des logements des ZUS.

Si on note π_i^1 et π_i^2 les probabilités de tirage de l'individu i lors du premier tirage et lors du deuxième tirage, on a alors $P(i \in s_V) = \pi_i^1 + \pi_i^2 - \pi_i^1\pi_i^2$ et la pondération d'Horvitz-Thompson associée est alors $w_i^{HT} = \dfrac{1}{\pi_i^1 + \pi_i^2 - \pi_i^1\pi_i^2}$. Si le logement i n'appartient pas à une ZUS, l'expression précédente se simplifie : $w_i^{HT} = \dfrac{1}{\pi_i^1}$.

3.1.2.Pondérations du partage des poids

Dans le cas qui nous intéresse, d'autres systèmes de pondération permettant d'obtenir des estimations sans biais pour tout total sur la population d'intérêt sont envisageables : en effet, le total X d'une variable $\{x_i\}_{i\in V}$ sur la population V peut s'écrire comme la somme du total sur la population U_1 et du total sur la population U_2.

Ainsi, pour $\alpha \in [0,1]$, on a :

$$\begin{aligned} X &= \sum_{i\in V} x_i = \sum_{i\notin ZUS} x_i + \sum_{i\in ZUS} [\alpha + (1-\alpha)]\cdot x_i \\ &= \sum_{i\in U_1} \left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot x_i + \sum_{i\in U_2} (1-\alpha)\cdot I_{i\in ZUS}\cdot x_i \end{aligned}$$

En considérant les estimateurs d'Horvitz-Thompson sur chacune des sous-populations U_1 et U_2, on obtient de nouveaux estimateurs de totaux sur la population V :

$$\begin{aligned} \hat{X}^{PP,\alpha} &= \sum_{i\in s_{U_1}} \left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot \frac{1}{\pi_i^1}\cdot x_i + \sum_{i\in s_{U_2}} (1-\alpha)\cdot I_{i\in ZUS}\cdot \frac{1}{\pi_i^2}\cdot x_i \\ &= \sum_{i\in s_V} \left[\left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot I_{i\in s_{U_1}}\cdot \frac{1}{\pi_i^1} + (1-\alpha)\cdot I_{i\in ZUS}\cdot I_{i\in s_{U_2}}\cdot \frac{1}{\pi_i^2} \right]\cdot x_i \\ &= \sum_{i\in s_V} w_i^\alpha \cdot x_i \qquad où \quad w_i^\alpha = \left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot I_{i\in s_{U_1}}\cdot \frac{1}{\pi_i^1} + (1-\alpha)\cdot I_{i\in ZUS}\cdot I_{i\in s_{U_2}}\cdot \frac{1}{\pi_i^2} \end{aligned}$$

Ces estimateurs sont sans biais car :

$$\begin{aligned} E\left(\hat{X}^{PP,\alpha}\right) &= \sum_{i\in V} E\left(w_i^\alpha \cdot I_{i\in s_V}\right)\cdot x_i \\ &= \sum_{i\in V} \left[\left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot E\left(I_{i\in s_{U_1}}\cdot I_{i\in s_V}\right)\cdot \frac{1}{\pi_i^1} + (1-\alpha)\cdot I_{i\in ZUS}\cdot E\left(I_{i\in s_{U_2}}\cdot I_{i\in s_V}\right)\cdot \frac{1}{\pi_i^2} \right]\cdot x_i \\ &= \sum_{i\in V} \left[\left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right)\cdot \pi_i^1\cdot \frac{1}{\pi_i^1} + (1-\alpha)\cdot I_{i\in ZUS}\cdot \pi_i^2\cdot \frac{1}{\pi_i^2} \right]\cdot x_i \\ &= \sum_{i\in V} \left[\left(I_{i\notin ZUS} + \alpha \cdot I_{i\in ZUS}\right) + (1-\alpha)\cdot I_{i\in ZUS} \right]\cdot x_i = \sum_{i\in V} x_i = X \end{aligned}$$

3.1.3. Autres pondérations envisageables

A ce stade les pondérations envisageables sont nombreuses : on a déjà cité la pondération d'Horvitz-Thompson w^{HT} et les pondérations w^{α} construites en utilisant le principe de la méthode du partage des poids ; mais on peut également envisager n'importe quelle combinaison convexe des pondérations précédemment présentées. Ainsi $w = (1-\mu)\cdot w^{HT} + \mu \cdot w^{\alpha}$ avec $0 \le \mu \le 1$ est une pondération qui fournit des estimations sans biais de totaux[13]. La pondération optimale est alors celle qui minimise la variance des estimations.

3.2 Calcul de précision et système de pondération retenu

Nous allons étudier la précision des estimateurs envisagés à la section précédente. Notons ces estimateurs sous la forme :

$$\hat{X} = (1-\mu)\cdot \hat{X}^{HT} + \mu \cdot \hat{X}^{PP,\alpha} \quad (\mu \in [0,1],\ \alpha \in [0,1]).$$

La variance de $\hat{X}$ s'exprime ainsi :

$$V(\hat{X}) = (1-\mu)^2 \cdot V(\hat{X}^{HT}) + \mu^2 \cdot V(\hat{X}^{PP,\alpha}) + 2\cdot(1-\mu)\cdot\mu\cdot Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})$$

L'expression théorique de chacun de ces trois termes est calculée en annexe

La complexité des différents termes (cf. annexe) rend inatteignable la résolution analytique du programme d'optimisation. Pour choisir le système de pondération, il nous a donc fallu faire appel à des critères plus simples que celui de minimisation de la variance.

On peut néanmoins remarquer que pour la variable $X_i = I_{i\notin ZUS}$, l'estimateur du partage des poids et l'estimateur d'Horvitz-Thompson coïncident, le choix de α et de μ n'a aucune influence sur l'estimation du nombre de logements hors ZUS (ou sur l'estimation de tout total défini sur les logements hors ZUS). Seule l'estimation sur le domaine des ZUS a un impact, ce domaine étant restreint (7 % des logements au RP99), on peut penser qu'en pratique les estimations au niveau national (ainsi que les variances associées à ces estimations) doivent être relativement peu sensibles au choix de la pondération.

---

[13] Dans le cas où la condition $0 \le \mu \le 1$ n'est pas vérifiée, la pondération obtenue fournie toujours des estimations sans biais des totaux mais cela peut conduire à obtenir des poids négatifs ce que le responsable d'enquête devrait vraisemblablement trouver peu satisfaisant. Remarquons également qu'une combinaison convexe de plus de deux termes se ramène au cas précédent car si $w' = \left(1-\sum_i \lambda_i\right) w^{HT} + \sum_i \lambda_i \cdot w^{\alpha_i}$ avec $\lambda_i \ge 0$ et $\sum_i \lambda_i \le 1$ alors w'$= (1-\mu)\cdot w^{HT} + \mu \cdot w^{\alpha}$ avec $\mu = \sum_i \lambda_i$ et $\alpha = \dfrac{\sum_i \lambda_i \alpha_i}{\sum_i \lambda_i}$.

Choix de α :

Le choix que nous avons fait pour α a été guidé par la volonté d'obtenir les poids les moins dispersés possibles pour la pondération du partage des poids, par principe d'équité entre les ménages[14]. Pour cela, nous avons été amené à considérer les différents cas de figure pour sélectionner un logement en ZUS dans l'échantillon. Les différentes situations sont résumées dans le tableau 4 ci-dessous :

Tableau 4 : Les 3 cas de figure pour sélectionner un logement en ZUS

Cas de figure	***Probabilité conditionnée par la présence dans l'échantillon*** s_V	***Poids de l'observation***
$i \in s_{U_1}$ et $i \notin s_{U_2}$	$\frac{\pi_i^1 \cdot (1-\pi_i^2)}{\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2}$	$\frac{\alpha}{\pi_i^1}$
$i \notin s_{U_1}$ et $i \in s_{U_2}$	$\frac{\pi_i^2 \cdot (1-\pi_i^1)}{\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2}$	$\frac{1-\alpha}{\pi_i^2}$
$i \in s_{U_1}$ et $i \in s_{U_2}$	$\frac{\pi_i^1 \cdot \pi_i^2}{\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2}$	$\frac{\alpha}{\pi_i^1} + \frac{1-\alpha}{\pi_i^2}$

La dispersion des poids est appréhendée par la variance des poids dans l'échantillon final :

$$V(w_i^{PP,\alpha} \mid i \in s_V) = E((w_i^{PP,\alpha})^2 \mid i \in s_V) - \left[E(w_i^{PP,\alpha} \mid i \in s_V)\right]^2$$

Or $E(w_i^{PP,\alpha} \mid i \in s_V) = \frac{1}{P(i \in s_V)}$ et ne dépend donc pas de α.

$$E((w_i^{PP,\alpha})^2 \mid i \in s_V) = \frac{1}{\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2} \cdot \left[\frac{\alpha^2 \cdot (1-\pi_i^2)}{\pi_i^1} + \frac{(1-\alpha)^2 \cdot (1-\pi_i^1)}{\pi_i^2} + \pi_i^1 \cdot \pi_i^2 \cdot \left(\frac{\alpha}{\pi_i^1} + \frac{1-\alpha}{\pi_i^2} \right)^2 \right]$$

$$\propto \frac{\alpha^2}{\pi_i^1} + \frac{(1-\alpha)^2}{\pi_i^2} + 2 \cdot \alpha \cdot (1-\alpha)$$

Cette expression est donc minimale pour $\alpha^* = \frac{\pi_i^1 - \pi_i^1 \cdot \pi_i^2}{\pi_i^1 + \pi_i^2 - 2 \cdot \pi_i^1 \cdot \pi_i^2} \approx \frac{\pi_i^1}{\pi_i^1 + \pi_i^2}$.

Les taux de sondage dans la base ZUS et dans l'échantillon-maître (hors extension régionale) sont tels que $\pi_i^1 = \frac{3}{2} \cdot \pi_i^2$, ce qui implique que $\boxed{\alpha^* \approx 0{,}6}$. C'est donc le choix que nous avons fait pour définir α.

Remarquons que ce résultat s'interprète intuitivement de la manière suivante : comme la variance des estimations est inversement proportionnelle à la taille d'échantillon, le coefficient α optimum introduit dans le partage des poids peut être approximé par le ratio suivant : nombre de logements recensés en ZUS échantillonnés via l' « EM+EMEX » (1539 unités) divisé par le nombre de logements recensés en ZUS échantillonnés via la base ZUS (1004 unités), c'est-à-dire : $\alpha^* \cong \frac{1539}{1004+1539} \cong 0{,}6$

[14] Ce système de pondération est naturel dans l'approche « ménage » par opposition aux statistiques d'entreprises où des probabilités inégales s'imposent du fait de l'hétérogénéité de la taille des unités.

Choix de μ :

Trois cas de figure sont envisageables :

1. Si $Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha}) \leq V(\hat{X}^{HT})$ et $Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha}) \leq V(\hat{X}^{PP,\alpha})$ alors l'estimateur optimal est :

$$\hat{X}^{*} = \frac{V(\hat{X}^{PP,\alpha}) - Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})}{V(\hat{X}^{HT}) + V(\hat{X}^{PP,\alpha}) - 2Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})} \hat{X}^{HT} + \frac{V(\hat{X}^{HT}) - Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})}{V(\hat{X}^{HT}) + V(\hat{X}^{PP,\alpha}) - 2Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})} \hat{X}^{PP,\alpha}$$

et sa variance est :

$$V(\hat{X}^{*}) = \frac{V(\hat{X}^{HT})V(\hat{X}^{PP,\alpha}) - \left(Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})\right)^2}{V(\hat{X}^{HT}) + V(\hat{X}^{PP,\alpha}) - 2Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha})}$$

2. Si $V(\hat{X}^{HT}) \leq Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha}) \leq V(\hat{X}^{PP,\alpha})$
alors l'estimateur optimal est : $\hat{X}^{*} = \hat{X}^{HT}$

3. Si $V(\hat{X}^{PP,\alpha}) \leq Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha}) \leq V(\hat{X}^{HT})$
alors l'estimateur optimal est : $\hat{X}^{*} = \hat{X}^{PP,\alpha}$

Nous n'avons pas trouvé de raison évidente pour penser que nous étions dans un de ces trois cas pour les variables d'intérêt de l'enquête, les expressions théoriques des variances et du terme de covariance sont en effet trop compliquées pour pouvoir être ordonnées de manière simple. Ce qui nous a guidé dans le choix de μ a relevé de raisons plus pratiques. Nous avons choisi la valeur $\mu = 1$, c'est à dire l'estimateur du partage des poids pour deux raisons :

- la simplicité de sa mise en œuvre,

- la facilité d'adaptation au« moule standard » des enquêtes ménages de l'INSEE pour le calcul de précision. En effet, en calculant les variables transformées $z_{1i} = \left[\left(I_{i \notin ZUS} + \alpha \cdot I_{i \in ZUS}\right) \cdot x_i\right]$ et $z_{2i} = \left[\left(1-\alpha\right) \cdot I_{i \in ZUS} \cdot x_i\right]$ ainsi que les estimateurs d'Horvitz-Thompson correspondant $\hat{Z}_1(\alpha) = \sum_{i \in s_{U_1}} \frac{1}{\pi_i^1} \cdot z_{1i}$ et $\hat{Z}_2(\alpha) = \sum_{i \in s_{U_2}} \frac{1}{\pi_i^2} \cdot z_{2i}$, l'estimateur $\hat{X}^{PP,\alpha}$ prend la forme suivante :

$$\hat{X}^{PP,\alpha} = \hat{Z}_1(\alpha) + \hat{Z}_2(\alpha)$$

La variance de l'estimateur du partage des poids s'écrit donc :

$$V(\hat{X}^{PP,\alpha}) = V(\hat{Z}_1(\alpha)) + V(\hat{Z}_2(\alpha)) + 2 \cdot Cov(\hat{Z}_1(\alpha), \hat{Z}_2(\alpha))$$

Comme le tirage dans la base des logements en ZUS est indépendant du tirage réalisé grâce à l'échantillon-maître et à la BSLN, $Cov(\hat{Z}_1(\alpha), \hat{Z}_2(\alpha)) = 0$. L'expression de la variance de $\hat{X}^{PP,\alpha}$ se simplifie de la manière suivante :

$$V(\hat{X}^{PP,\alpha}) = V(\hat{Z}_1(\alpha)) + V(\hat{Z}_2(\alpha))$$

Le premier terme peut s'estimer comme pour les autres enquêtes ménage au moyen du logiciel POULPE en utilisant les fichiers de l'échantillon-maître, et le second terme peut s'estimer à partir du plan de sondage utilisé sur la base des ZUS (ce plan de sondage étant relativement simple). On est donc ramené à un cas classique d'estimation de variance d'un estimateur d'Horvitz-Thompson sur le total d'une variable transformée.

Le système de pondération que nous avons adopté est donc le suivant :

Tableau 5 : Pondération du partage des poids

Type de logement	**Base de sondage**	**Pondération finale** w_i
Logement hors ZUS	Echantillonné grâce à l'EM, l'EMEX, la BSLN ou SITADEL	$\frac{1}{\pi_i^1}$
Logement en ZUS	Echantillonné grâce à l'EM, l'EMEX, la BSLN ou SITADEL	$\frac{0,6}{\pi_i^1}$
	Echantillonné grâce à la base ZUS	$\frac{0,4}{\pi_i^2}$
	Echantillonné deux fois ; une fois grace à l'EM ou l'EMEX ; une fois grâce à la base ZUS[15]	$\frac{0,6}{\pi_i^1}+\frac{0,4}{\pi_i^2}$

L'estimateur retenu est donc : $\hat{X}=\sum_{i\notin ZUS}\frac{x_i}{\pi_i^1}+\sum_{i\in ZUS}x_i\cdot\left(0,6\cdot\frac{I_{i\in s_{U_1}}}{\pi_i^1}+0,4\cdot\frac{I_{i\in s_{U_2}}}{\pi_i^2}\right)$

3.3 Bilan de la pondération

3.3.1. Pondérations d'échantillonnage dans les bases EM + EMEX + BSLN + SITADEL :

Les pondérations d'échantillonnage dans les bases EM, EMEX, BSLN et SITADEL sont présentées dans le tableau 6 ci-dessous. La pondération de chaque logement s'obtient en multipliant le « poids de base » W_1 par un coefficient c de représentation dépendant de la région et du type du logement considéré.

Tableau 6 : les pondérations de l'échantillon issu de l'EM, l'EMEX, la BSLN et SITADEL

Région / ***Type de logement***	***Nord-Pas-de-Calais***	***Pays de la Loire***	***Aquitaine***	***Autres régions***
Principaux sur-représentés	$c=1$	$c=1$	$c=1$	$c=1$
Principaux non sur-représentés	$c=4$	$c=4$	$c=4$	$c=4$
Occasionnels ou secondaires	$c=4$	$c=4$	$c=4$	$c=4$
Vacants dans le rural	$c=2$	$c=2$	$c=2$	$c=2$
Vacants dans l'urbain	$c=1$	$c=1$	$c=1$	$c=1$
Neufs	$c\approx 1$	$c\approx 1$	$c\approx 1$	$c\approx 1$
Poids de base	$W_1=376$	$W_1=327$	$W_1=332$	$W_1=1096$

[15] Ce cas de figure théoriquement possible ne s'est pas produit lors de la réalisation des tirages.

3.3.2 Pondérations d'échantillonnage dans la base « ZUS » :

Le deuxième jeu de pondérations issu d'échantillonnage renvoie au tirage dans la « base ZUS » est présenté dans le tableau 7 ci-dessous. La pondération de chaque logement s'obtient en multipliant le « poids de base » W_2 par un coefficient c de représentation dépendant du type du logement considéré.

Tableau 7 : pondérations d'échantillonnage dans la base ZUS

Type de logement	***Coefficient***
Principaux	$c = 1$
Occasionnels[16]	$c = 3{,}48$
Secondaires[17]	$c = 5{,}12$
Vacants dans le rural	$c = 2$
Vacants dans l'urbain	$c = 1$
Poids de base	$W_2 = 1497$

3.3.3. Pondérations finales après partage des poids :

Les pondérations issues du partage des poids s'obtiennent à partir des pondérations d'échantillonnage (tableaux 6 et 7) et du système retenu dans le paragraphe 3.2 (tableau 5).

3.3.4. Pondérations finales :

Pour améliorer la précision des estimations obtenues, un calage a été effectué sur deux marges : la catégorie de logement ventilée en trois modalités[18] (logement principal, logement occasionnel ou secondaire et logement vacant) et un critère géographique à quatre modalités (Nord-Pas de Calais, Pays de la Loire, Aquitaine, autres régions).

[16] Le coefficient prévu était 4, ce qui conduisait à une allocation théorique de 1,74 logements. En pratique deux logements occasionnels ont été échantillonnés donc $c = 4 \times \frac{1{,}74}{2} = 3{,}48$

[17] Le coefficient prévu était 4, ce qui conduisait à une allocation théorique de 1,28 logements. En pratique un seul logement secondaire a été échantillonnés donc $c = 4 \times \frac{1{,}28}{1} = 5{,}12$

[18] Le poids des logements neufs est calé au moment de l'échantillonnage dans la BSLN et Sitadel.

Les pondérations finales issues du calage se distribuent de la manière suivante :

Tableau 8 : pondérations finales

Région	Type de logement	Logements Hors ZUS	Logements en ZUS échantillonnés dans l'EM, l'EMEX, la BSLN et Sitadel	Logements en ZUS échantillonnés dans la base « ZUS »
Nord-Pas- de-Calais	Principaux du groupe sur-représenté	371,42	222,85	600,17
	Principaux du groupe non sur-représenté	1485,67	891,40	
	Occasionnels ou Secondaires	1 791,88		
	Vacants dans le rural	756,28		
	Vacants dans l'urbain	378,14	226,88	611, 06
	Neufs	376,75		
Pays de la Loire	Principaux du groupe sur-représenté	332,30	199,38	617,39
	Principaux du groupe non sur-représenté	1 329,21	797,53	
	Occasionnels ou Secondaires	1 762,76		
	Vacants dans le rural	663,89		
	Vacants dans l'urbain	331,95	199,17	616,76
	Neufs	338,38		
Aquitaine	Principaux du groupe sur-représenté	325,92	195,55	596,63
	Principaux du groupe non sur-représenté	1 303,67	782,20	
	Occasionnels ou Secondaires	1 338,48		
	Vacants dans le rural	706,02		
	Vacants dans l'urbain	353,01	211,81	646,26
	Neufs	340		
Autres régions	Principaux du groupe sur-représenté	1 095,67	657,40	607,15
	Principaux du groupe non sur-représenté	4 382,70	2629,62	
	Occasionnels ou Secondaires	3 980,60		Occ : 1 917,13 Sec : 2 819,15
	Vacants dans le rural	2 169,49		
	Vacants dans l'urbain	1 084,74	650,85	601,13
	Neufs	1 137,21		

4. Conclusion

L'enquête s'est déroulée du 11 octobre 2004 au 28 janvier 2005 (la date de fin était au départ prévue pour le 15 janvier mais la collecte a été prolongée en Aquitaine suite aux difficultés rencontrées). L'entretien a pu être réalisé dans à peu près les deux tiers des ménages (la moitié du tiers restant est constituée de ménages hors-champ, en particulier, du fait de la condition d'âge ; l'autre moitié regroupe les refus et les ménages « impossibles à joindre »). Pour les DR ayant participé à l'opération de 2002, le taux de réalisation a augmenté de 7 points : l'expérience de la première opération a pu être capitalisée mais c'est le caractère obligatoire de l'enquête de 2004 qui est sans doute la cause principale.

Compte tenu de ces taux de réponse assez favorables, les objectifs de l'enquête en terme d'échantillonnage ont pour la plupart été remplis : le nombre total de répondants est de 10 400 ; on en compte 1 500 en ZUS, 1 700 dans le Nord-Pas-de-Calais et 1 400 dans les Pays de la Loire, effectifs qui dépassent les prévisions. En revanche, les conditions de collecte en Aquitaine n'ont pas été aussi favorables, si bien qu'avec 950 répondants, la taille de l'échantillon est un peu inférieure à ce qui avait été fixé.

Les premiers traitements menés sur les données permettent aussi de tester l'efficacité de la sur-représentation. Divers classements provisoires de la population selon le niveau de compétences en lecture ont été construits : selon les cas, la proportion de personnes en difficulté en utilisant la pondération présentée ici est de 4 à 5 points inférieures à ce que donnent les chiffres bruts. Ceci montre que la sur-représentation a permis d'augmenter sensiblement l'effectif des plus en difficulté. Cependant, pour la diffusion des résultats définitifs, des analyses supplémentaires devront être menées pour construire un modèle de non-réponse et des calages sur données externes, permettant d'assurer la qualité des estimations obtenues, tant au niveau national que régional, pour les extensions.

Annexes :

Modalités de la typologie Tabard en 27 postes :

Typo 27	Libellé
ADPUB1	Classes moyennes de fonction publique / littoral
ADPUB3	Administration, cafés restaurants /PACA, littoral
AGRI12	IAA, bois, meubles, matériaux de construction / cantons des régions ouest
AGRI13	Agriculture, textiles et industries diverses / cantons des Pays de la Loire
AGRI21	Agriculture, bâtiment / littoral, bassin méditerranéen
AGRI22	Commerce de véhicules, commerce de gros (alimentation, bois, matériaux) / petits bourgs de la moitié ouest
AGRI31	Agriculture / rural isolé, quart sud-ouest
CHOMA1	Manutention, tri / Haute-Normandie, ZUS
CHOMA2	Tertiaire administratif et commercial peu qualifié, chômage / littoral, ZUS
CHOMA3	Petits métiers urbains, chômage / Ile-de-France, ZUS
CHOMA4	Chômage, service des villes / ZUS, grands centres, PACA, Nord-Pas-de-Calais
DIR1	Activités artistiques ou à clientèle aisée /Paris est
DIR3	Patronat, établissements financiers, services aux entreprises / Quartiers très aisés de Paris et de l'ouest de l'Ile-de-France
DIR4	Cadres de la santé et de l'environnement / quartiers des grands centres de province
DIR5	Haute technologie / ouest francilien
INDOUV1	Industries textile-cuirs, papier-carton, matériaux de construction / espace rural industriel
INDOUV3	Métallurgie, industrie qualifiée / périphérie industrielle du quart nord-est
INDOUV4	Mécanique, chimie, plastiques, faible qualification industrielle / moitié est
INDOUV5	Chômage industriel / banlieues des grandes Unités Urbaines de province, Nord-Pas-de-Calais
INDQ2	Transports ferroviaires
INDQ3	Salariés qualifiés de l'industrie / communes et cantons péri-urbains
INDQ4	Encadrement de la production
INDQ5	Métiers divers peu qualifiés / petits centres provinciaux
SEMAG2	Hôtellerie restauration / littoral, bassin méditerranéen
SEMAG3	Activités semi-agricoles / communes petits pôles
TEC2	Aéronautique, ordinateurs
TEC3	Catégories moyennes administratives d'entreprise publiques ou privées / banlieue parisienne

Calcul des termes intervenant dans la formule de variance d'un estimateur composite de total (combinaison linéaire d'un estimateur d'Horvitz-Thompson et d'un estimateur du partage des poids).

Variance de l'estimateur d'Horvitz-Thompson :

La variance de l'estimateur d'Horvitz-Thompson dépend des probabilités d'inclusion double $\pi_{i,k} = P(i \in s_V \cap k \in s_V)$:

$$V(\hat{X}^{HT}) = \sum_{k \in V} \sum_{i \in V} \frac{X_i \cdot X_k}{\pi_i \cdot \pi_k} \cdot (\pi_{ik} - \pi_i \pi_k)$$

Comme précédemment pour les probabilités d'inclusion simple, nous notons π_{ik}^1 et π_{ik}^2 les probabilités d'inclusion doubles pour le premier tirage (effectué dans la population U_1 des logements de métropole) et pour le deuxième tirage (effectué dans la population U_2 des logements en ZUS). Naturellement, si un des deux logements i ou k n'appartient pas à une ZUS alors $\pi_{ik}^2 = 0$.

$$\begin{aligned}
\pi_{ik} &= P(i \in s_V \cap k \in s_V) \\
&= P(i \in s_{U_1} \cap k \in s_{U_1}) + P(i \in s_{U_1} \cap k \notin s_{U_1} \cap k \in s_{U_2}) + P(i \notin s_{U_1} \cap k \in s_{U_1} \cap i \in s_{U_2}) \\
&\quad + P(i \notin s_{U_1} \cap k \notin s_{U_1} \cap i \in s_{U_2} \cap k \in s_{U_2}) \\
&= P(i \in s_{U_1} \cap k \in s_{U_1}) + P(i \in s_{U_1} \cap k \notin s_{U_1}) \cdot P(k \in s_{U_2}) + P(i \notin s_{U_1} \cap k \in s_{U_1}) \cdot P(i \in s_{U_2}) \\
&\quad + P(i \notin s_{U_1} \cap k \notin s_{U_1}) \cdot P(i \in s_{U_2} \cap k \in s_{U_2}) \quad \textit{car les tirages sont indépendants} \\
&= \pi_{ik}^1 + \pi_k^2 \cdot (\pi_i^1 - \pi_{ik}^1) + \pi_i^2 \cdot (\pi_k^1 - \pi_{ik}^1) + \pi_{ik}^2 \cdot (1 - \pi_i^1 - \pi_k^1 + \pi_{ik}^1) \\
&= \pi_{ik}^1 \cdot \pi_{ik}^2 + \pi_{ik}^1 \cdot (1 - \pi_i^2 - \pi_k^2) + \pi_{ik}^2 \cdot (1 - \pi_i^1 - \pi_k^1) + \pi_i^1 \cdot \pi_k^2 + \pi_k^1 \cdot \pi_i^2
\end{aligned}$$

Finalement on obtient comme expression de la variance :

$$V(\hat{X}^{HT}) = \sum_{i,k \in V^2} X_i \cdot X_k \frac{\pi_{ik}^1 \cdot \pi_{ik}^2 + \pi_{ik}^1 \cdot (1 - \pi_i^2 - \pi_k^2) + \pi_{ik}^2 \cdot (1 - \pi_i^1 - \pi_k^1) + \pi_i^1 \cdot \pi_k^2 + \pi_k^1 \cdot \pi_i^2 - (\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2)(\pi_k^1 + \pi_k^2 - \pi_k^1 \cdot \pi_k^2)}{(\pi_i^1 + \pi_i^2 - \pi_i^1 \cdot \pi_i^2)(\pi_k^1 + \pi_k^2 - \pi_k^1 \cdot \pi_k^2)}$$

Variance de l'estimateur du partage des poids :

Pour calculer la variance de l'estimateur $\hat{X}^{PP,\alpha}$, nous utilisons la propriété de dualité décrite par Pierre Lavallée [1] pour écrire $\hat{X}^{PP,\alpha}$ comme somme de deux estimateurs d'Horvitz-Thompson de totaux sur les populations U_1 et U_2. En posant $\hat{Z}_1(\alpha) = \sum_{i \in s_{U_1}} \frac{1}{\pi_i^1} \left[\left(I_{i \notin ZUS} + \alpha \cdot I_{i \in ZUS} \right) \cdot x_i \right]$ et $\hat{Z}_2(\alpha) = \sum_{i \in s_{U_2}} \frac{1}{\pi_i^2} \cdot \left[(1 - \alpha) \cdot I_{i \in ZUS} \cdot x_i \right]$, l'estimateur $\hat{X}^{PP,\alpha}$ prend la forme suivante :

$$\hat{X}^{PP,\alpha} = \hat{Z}_1(\alpha) + \hat{Z}_2(\alpha)$$

La variance de l'estimateur du partage des poids s'écrit donc :

$$V(\hat{X}^{PP,\alpha}) = V(\hat{Z}_1(\alpha)) + V(\hat{Z}_2(\alpha)) + 2 \cdot Cov(\hat{Z}_1(\alpha), \hat{Z}_2(\alpha))$$

Comme le tirage dans la base des logements en ZUS est indépendant du tirage réalisé grâce à l'échantillon-maître et à la BSLN, $Cov(\hat{Z}_1(\alpha),\hat{Z}_2(\alpha))=0$. L'expression de la variance de $\hat{X}^{PP,\alpha}$ se simplifie de la manière suivante :

$V(\hat{X}^{PP,\alpha})=V(\hat{Z}_1(\alpha))+V(\hat{Z}_2(\alpha))$

En reprenant les notations introduites dans le paragraphe précédent, on a de plus :

$$V(\hat{Z}_1(\alpha))=\sum_{i\in U_1}\sum_{k\in U_1}\left(I_{i\notin ZUS}+\alpha\cdot I_{i\in ZUS}\right)\left(I_{k\notin ZUS}+\alpha\cdot I_{k\in ZUS}\right)\frac{x_i\cdot x_k}{\pi_i^1\cdot\pi_k^1}\left(\pi_{ik}^1-\pi_i^1\cdot\pi_k^1\right)$$

$$V(\hat{Z}_2(\alpha))=\sum_{i\in U_2}\sum_{k\in U_2}(1-\alpha)^2\cdot I_{i\in ZUS}\cdot I_{k\in ZUS}\cdot\frac{x_i\cdot x_k}{\pi_i^2\cdot\pi_k^2}\cdot\left(\pi_{ik}^2-\pi_i^2\cdot\pi_k^2\right)$$

Et donc :

$$V(\hat{X}^{PP,\alpha})=\sum_{i\in U_1}\sum_{k\in U_1}\left(I_{i\notin ZUS}+\alpha\cdot I_{i\in ZUS}\right)\left(I_{k\notin ZUS}+\alpha\cdot I_{k\in ZUS}\right)\frac{x_i\cdot x_k}{\pi_i^1\cdot\pi_k^1}\left(\pi_{ik}^1-\pi_i^1\cdot\pi_k^1\right)+$$
$$\sum_{i\in U_2}\sum_{k\in U_2}(1-\alpha)^2\cdot I_{i\in ZUS}\cdot I_{k\in ZUS}\cdot\frac{x_i\cdot x_k}{\pi_i^2\cdot\pi_k^2}\cdot\left(\pi_{ik}^2-\pi_i^2\cdot\pi_k^2\right)$$

$$\boxed{V(\hat{X}^{PP,\alpha})=\sum_{i\in V}\sum_{k\in V}\left(I_{i\notin ZUS}+\alpha\cdot I_{i\in ZUS}\right)\left(I_{k\notin ZUS}+\alpha\cdot I_{k\in ZUS}\right)\frac{x_i\cdot x_k}{\pi_i^1\cdot\pi_k^1}\left(\pi_{ik}^1-\pi_i^1\cdot\pi_k^1\right)+ (1-\alpha)^2\cdot I_{i\in ZUS}\cdot I_{k\in ZUS}\cdot\frac{x_i\cdot x_k}{\pi_i^2\cdot\pi_k^2}\cdot\left(\pi_{ik}^2-\pi_i^2\cdot\pi_k^2\right)}$$

<u>Covariance de l'estimateur d'Horvitz-Thompson et de l'estimateur du partage des poids :</u>

Comme l'estimateur d'Horvitz-Thompson et l'estimateur du partage des poids sont tout les deux sans biais, $Cov(\hat{X}^{HT},\hat{X}^{PP,\alpha})=E\left(\hat{X}^{HT}\cdot\hat{X}^{PP,\alpha}\right)-X^2$. Il reste donc à calculer la quantité $E\left(\hat{X}^{HT}\cdot\hat{X}^{PP,\alpha}\right)$.

$$E\left(\hat{X}^{HT}\cdot\hat{X}^{PP,\alpha}\right)=\sum_{i\in V}\sum_{k\in V}x_i\cdot x_k\cdot w_k^{HT}\cdot E\left(w_i^{PP,\alpha}\cdot I_{i\in s_V}\cdot I_{k\in s_V}\right)$$
$$=\sum_{i\in V}\sum_{k\in V}x_i\cdot x_k\cdot\frac{1}{\pi_k^1+\pi_k^2+\pi_k^1\cdot\pi_k^2}\cdot\left[\frac{I_{i\notin ZUS}+\alpha\cdot I_{i\in ZUS}}{\pi_i^1}\cdot E\left(I_{i\in s_{U_1}}\cdot I_{k\in s_V}\right)+\frac{(1-\alpha)\cdot I_{i\in ZUS}}{\pi_i^2}\cdot E\left(I_{i\in s_{U_2}}\cdot I_{k\in s_V}\right)\right]$$

Or $I_{k\in s_V}=I_{k\in s_{U_1}}+I_{k\in s_{U_2}}-I_{k\in s_{U_1}}\cdot I_{k\in s_{U_2}}$, donc :

$$E(I_{i\in s_{U_1}}\cdot I_{k\in s_V})=E(I_{i\in s_{U_1}}\cdot I_{k\in s_{U_1}})+E(I_{i\in s_{U_1}}\cdot I_{k\in s_{U_2}})-E(I_{i\in s_{U_1}}\cdot I_{k\in s_{U_1}}\cdot I_{k\in s_{U_2}})$$
$$=\pi_{ik}^1+\pi_i^1\cdot\pi_k^2-\pi_{ik}^1\cdot\pi_k^2$$

De même : $E(I_{i\in s_{U_2}}\cdot I_{k\in s_V})=\pi_{ik}^2+\pi_i^2\cdot\pi_k^1-\pi_{ik}^2\cdot\pi_k^1$.

Finalement :

$$E\left(\hat{X}^{HT} \cdot \hat{X}^{PP,\alpha}\right) =$$
$$\sum_{i\in V}\sum_{k\in V} x_i x_k \frac{1}{\pi_k^1 + \pi_k^2 - \pi_k^1\pi_k^2}\left[\frac{I_{i\notin ZUS} + \alpha I_{i\in ZUS}}{\pi_i^1}\left(\pi_{ik}^1 + \pi_i^1\pi_k^2 - \pi_{ik}^1\pi_k^2\right) + \frac{(1-\alpha)I_{i\in ZUS}}{\pi_i^2}\left(\pi_{ik}^2 + \pi_i^2\pi_k^1 - \pi_{ik}^2\pi_k^1\right)\right]$$

Et :

$$Cov(\hat{X}^{HT}, \hat{X}^{PP,\alpha}) =$$
$$\sum_{i\in V}\sum_{k\in V} x_i x_k \left[\frac{\left(I_{i\notin ZUS} + \alpha I_{i\in ZUS}\right)\left(\pi_{ik}^1 + \pi_i^1 \cdot \pi_k^2 - \pi_{ik}^1 \cdot \pi_k^2\right)}{\left(\pi_k^1 + \pi_k^2 - \pi_k^1 \cdot \pi_k^2\right)\pi_i^1} + \frac{\left((1-\alpha)I_{i\in ZUS}\right)\left(\pi_{ik}^2 + \pi_i^2 \cdot \pi_k^1 - \pi_{ik}^2 \cdot \pi_k^1\right)}{\left(\pi_k^1 + \pi_k^2 - \pi_k^1 \cdot \pi_k^2\right)\pi_i^2} - 1\right]$$

Lexique des abréviations :

ANLCI : Agence National de Lutte Contre l'Illettrisme
BSLN : Base de Sondage des Logements Neufs
DIV : Délégation Interministérielle à la Ville
EM : Echantillon-Maître
EMEX : Echantillon-Maître pour les Extensions Régionales
IVQ : Information et Vie Quotidienne
RP : Recensement de la Population
SITADEL : Système de traitement Automatisé des Données Elémentaires sur le Logement et les locaux

Bibliographie :

[1] Lavallée, P (2002), *Le sondage indirect ou la méthode généralisée du partage des poids*, Statistique et Mathématiques Appliquées.

[2] Martin-Houssart, G et Tabard, N (2002), *Représentation socio-économique du territoire*, Document de travail de l'INSEE n°F0208.

[3] Murat, F (2004), *Les difficultés des adultes face à l'écrit*, Insee Première n°959.

[4] Rousseau, S et Tardieu, F (2004), *La macro SAS CUBE de tirage équilibré, documentation de l'utilisateur*, Document de travail de l'INSEE n°0402 de la série méthodologie statistique.

[5] Wilms, L (2000), *L'échantillon-maître 1999*, Insee Méthodes n°100, tome 1.