OCTOPUSSE[1] : un système d'Echantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement

Marc CHRISTINE (*), Sébastien FAIVRE (*)

(*) Insee, Unité Méthodes Statistiques

Remarque : le projet OCTOPUSSE, qui a permis la mise au point d'un Echantillon-Maître adapté au contexte original du nouveau recensement, a constitué un projet innovant et complexe, nécessitant de nombreuses études méthodologiques. L'objectif principal de cet article est de présenter les principales caractéristiques méthodologiques d'OCTOPUSSE pour donner une vue d'ensemble de son fonctionnement. Certains aspects méthodologiques de l'Echantillon-Maître OCTOPUSSE (plan de sondage des ZAE, méthode de tirage équilibrés emboîtés pour le tirage des ZAE pour les Extensions régionales), qui font l'objet de présentations complémentaires détaillées[2], ne sont abordés ici que de manière succincte, avec un renvoi pour plus de détail aux présentations complémentaires sur le sujet.

1. L'impact du Recensement Rénové sur le tirage des enquêtes ménages[3]

1.1 Le système antérieur d'Echantillon-Maître 1999

Depuis la décennie 60, les échantillons des enquêtes nationales auprès des ménages réalisées par l'Insee sont sélectionnés dans des listes de logements ad hoc constituées à partir de chaque recensement de la population. Ces listes sont en général complétées par des sources annexes (fichiers des permis de construire) permettant la couverture des logements construits postérieurement au dernier recensement disponible, dits « neufs », avec la constitution d'une base de sondage additionnelle (Base de Sondage des Logements Neufs, BSLN).

Pour assurer le tirage des échantillons des principales enquêtes, des systèmes d'échantillonnage standardisés ont été conçus et mis en œuvre depuis de longues années. La philosophie d'ensemble de ces systèmes qui se sont succédés s'est peu modifiée. **Elle résulte de compromis entre, d'une part, des objectifs statistiques de précision et d'optimisation des plans de sondage à coût donné et, d'autre part, les contraintes induites par le choix de réaliser la plupart de ces enquêtes en face à face, en recourant à un réseau d'enquêteurs localisé à proximité des logements enquêtés et relativement stable dans le temps.**

Le système actuel des échantillons des enquêtes ménages est donc organisé autour d'un **échantillon-maître** (EM) qui constitue la base de sondage principale dans laquelle seront sélectionnés la plupart des échantillons des enquêtes ménages nationales, à l'exception de **l'enquête Emploi qui utilise un autre système (aréolaire).**

Dans l'échantillon-maître, les unités finales sont des logements, mais ceux-ci sont concentrés dans des unités primaires afin de ne pas disperser les lieux d'enquêtes et de limiter les frais de déplacement des enquêteurs (surtout dans la partie rurale).

Ce système d'échantillonnage et la base de sondage sur laquelle il s'appuie étaient renouvelés après chaque recensement de la population et restaient fixes pendant la période inter-censitaire. Le système actuel (qui doit s'achever en 2009) est basé sur les données du recensement de 1999.

---

[1] Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Echantillons.

[2] Voir l'article de F. GUGGEMOS sur le plan de sondage des ZAE, et l'article de M. CHRISTINE et E. GROS sur les tirages équilibrés emboîtés des ZAE EM/EMEX.

[3] Le paragraphe 1 s'appuie sur l'expression des besoins du projet OCTOPUSSE, rédigée par Nathalie Caron.

1.2 Le contexte du nouveau recensement

Dans le courant de la précédente décennie, l'INSEE a décidé de passer du principe de recensement exhaustif de la population française effectué à intervalles de temps quasi-réguliers (7 à 9 ans) à un nouveau mode de recensement rotatif continu. Celui-ci a définitivement été mis en place en janvier 2004.

Désormais, les communes de moins de 10000 habitants (au dénombrement du RP 1999), ou « petites communes », sont recensées exhaustivement tous les cinq ans par roulement : pour cela, cinq groupes de rotation ont été définis aléatoirement, dans lesquels ont été réparties ces petites communes.

Pour ce qui est des communes comprenant 10000 habitants ou plus, ou « grandes communes », elles font l'objet d'une enquête de recensement plus complexe, par sondage chaque année au taux moyen de 8%. Les adresses de ces communes sont réparties aléatoirement entre cinq groupes de rotation disjoints : chaque année, des logements appartenant à un échantillon d'adresse puisé dans le groupe de rotation « actif » sont recensés. Au total, sur un cycle de 5 ans, environ 40% des logements de la commune sont recensés. Plus précisément, le traitement des logements diffère selon que le logement appartient à une « grande adresse[4] », une « adresse neuve » ou une « autre adresse » :

- lors de la première phase du RP de construction des groupes de rotation, les grandes adresses et les adresses neuves ont été affectées, pour la majorité des grandes communes, de manière semi-déterministe à un des cinq groupes tandis que les autres adresses étaient réparties de manière aléatoire entre ces groupes ;

- une fois les groupes de rotation constitués, la seconde phase du tirage sélectionne, pour un groupe de rotation donné, les adresses qui seront enquêtées. Les grandes adresses et les adresses neuves sont enquêtées exhaustivement, puis les autres adresses sont échantillonnées de telle sorte que l'échantillon total (y compris les grandes adresses et les adresses neuves enquêtées exhaustivement) représente 40% des logements du groupe de rotation enquêté.

Le plan de sondage du Recensement en grandes communes est détaillé de manière précise à l'annexe A.

Ce changement de méthodologie, s'il induit certes la perte du caractère exhaustif du recensement, offre en contrepartie de nombreux avantages, dont le principal réside dans la fraîcheur des données recueillies : avec cette méthode de collecte, il y aura en effet chaque année un recensement exhaustif dans environ 7000 petites communes et une enquête de recensement par sondage dans environ 900 grandes communes.

2. L'opportunité de disposer de bases de sondage annuelles fraîches

Les constantes liées à l'organisation de la ***collecte en face à face***, et la nécessité de ne pas trop disperser les lieux d'enquêtes et de limiter les déplacements des enquêteurs, subsistent, ce qui est un facteur de maîtrise de la qualité et des coûts. Elles rendent toujours utile un système de type « Echantillon-maître ».

Cependant, le changement de contexte lié à la ***mise en place d'une nouvelle méthodologie de recensement*** depuis janvier 2004 entraîne une refonte globale du système actuel d'échantillonnage des enquêtes ménages. En effet, le « nouveau recensement » conduit à une modification radicale des méthodes employées pour la construction des échantillons des enquêtes, parce que la base fournissant les listes d'unités échantillonnables sera renouvelée chaque année en permettant l'apport

[4] Est considérée comme grande adresse toute adresse dont le nombre de logements est au moins égal à 60 et qui est telle que l'ensemble des grandes adresses ne représentent pas plus de 10% des logements de la commune.

d'« information fraîche » mais, en contrepartie, ne sera plus exhaustive sur l'ensemble du territoire. De surcroît, les concepts manipulés dans le cadre du nouveau recensement ne correspondent plus à ceux habituellement utilisés dans les échantillons (unité urbaine par exemple), tandis que d'autres apparaissent (distinction entre petites et grandes communes...).

Du point de vue tant des enquêteurs que des utilisateurs des enquêtes, les perspectives offertes par le nouveau recensement portent surtout sur la **fraîcheur de la base de sondage** obtenue en réduisant au maximum le décalage temporel qui existe entre la date de collecte de l'enquête et la date de recensement.

Par conséquent, **le principe fondamental retenu consiste à sélectionner les échantillons des enquêtes dans la partie de la base recensée l'année précédente**[5]. Les avantages de ce principe de fraîcheur sont nombreux :

- Il devrait permettre de **minimiser le nombre de logements détruits ou en cours de destruction dans les échantillons** ainsi que le nombre de « transformations » d'une résidence principale en résidence secondaire (hors champ), qui sont souvent des points évoqués très négativement par les enquêteurs sur leurs conditions de travail, ainsi que des causes de surcoût pour les enquêtes.

- Par ailleurs, la **qualité du ciblage dans les enquêtes de certaines catégories de population** (au moyen de la surreprésentation de ces populations dans les échantillons tirés) **sera fortement améliorée** par la fraîcheur de l'information disponible sur les logements de la base de sondage. Dans le cas de certaines enquêtes, la fraîcheur de la base de sondage constitue un impératif absolu pour tirer un échantillon répondant aux spécifications des concepteurs d'enquêtes[6]. Par ailleurs, les enquêteurs se plaignent actuellement de se déplacer régulièrement pour des enquêtes complexes chez des ménages qui s'avèrent hors champ.

- Il permettra également de pouvoir **s'affranchir d'un système complémentaire de type BSLN pour l'échantillonnage des logements « neufs »** et, en particulier, d'en faire disparaître le coût.

- Enfin, il rendra possible **une disjonction maximale des échantillons** en garantissant qu'une même feuille de logement ne puisse être sélectionnée qu'une seule fois par OCTOPUSSE[7] pour des enquêtes ménages au cours d'une période de cinq ans.

3. Une conséquence : la constitution d'unités primaires particulières (ZAE, Zones d'Action Enquêteurs)

3.1. Principes de constitution des ZAE

Une conséquence directe de ce principe de fraîcheur est la nécessité de repenser la construction des unités primaires (ZAE, Zones d'Action Enquêteur) au sein desquelles seront tirés les échantillons de logements. En effet, ces zones doivent être construites selon les principes suivants :

a) comme par le passé, les ZAE doivent être des zones fixes pour pouvoir leur associer un enquêteur stable dans le temps et localisé à proximité
b) mais - ce qui est la nouveauté - elles doivent comporter des communes des cinq groupes de rotation pour pouvoir réaliser des enquêtes chaque année sur un échantillon tiré parmi les logements recensés l'année précédente

[5] A l'exception de certaines enquêtes comme l'enquête Victimation, dont le mode de tirage particulier envisagé est détaillée au paragraphe 7.3.2.

[6] C'est le cas par exemple pour l'enquête Modes de garde, ciblant notamment les logements dans lesquels vit un enfant de moins de trois ans.

[7] Ainsi, un intervalle de cinq ans minimum entre deux interrogations successives d'un même logement pourra être assuré pour l'essentiel, sauf exception, cf. paragraphe 7.3.2.

c) elles doivent comporter un nombre minimal de logements « échantillonnables » par groupe de rotation du RP pour que l'enquêteur affecté sur cette zone ait chaque année une charge de travail suffisante et que les principes de disjonction entre échantillons permettent de tirer plusieurs échantillons distincts d'enquête la même année sans devoir réinterroger les mêmes logements.

Afin de réduire au maximum l'étendue des ZAE, les règles de constitution des ZAE ont été établies comme suit :

- **constitution des ZAE en respectant les frontières régionales**[8]
- **séparation ZAE grandes communes (ZAEGC) et ZAE petites communes (ZAEPC)**
- **une grande commune constitue à elle seule une ZAEGC**[9]
- **au moins 300 résidences principales par groupe de rotation dans les ZAEPC**
- **avec l'objectif de minimiser leur étendue.**

Les ZAE ont été construites et tirées en 2007 sur la base des données du RP 1999, compte-tenu du fait qu'on ne disposait pas encore à cette date des premières populations légales (publiées fin décembre 2008).

3.2. Algorithme de constitution automatisée des ZAE et choix d'un scénario de constitution

Si la constitution des ZAEGC est immédiate (compte-tenu du principe une grande commune égal une ZAEGC), celle des ZAEPC apparaît en revanche beaucoup plus délicate : il s'agit en effet d'obtenir des zones d'étendue la plus faible possible sous les contraintes de nombre minimal de 300 résidences principales par groupe de rotation. Vu le grand nombre de petites communes à affecter à une ZAEPC (35 721 au 1er janvier 2006), et compte-tenu de l'impact de l'étendue des ZAE sur les conditions de travail des enquêteurs, une application de constitution automatisée des ZAE a été développée. L'algorithme de constitution des ZAEPC mis en œuvre dans chaque région est le suivant :

Algorithme de construction des ZAEPC

La construction des ZAEPC (c'est-à-dire la réalisation automatique d'une partition « optimale » du territoire métropolitain dans le respect des limites régionales et sous des contraintes de nombre minimal de logements dans chacun des cinq groupes de rotation et d' « étendue »[10] maximale des ZAEPC) a constitué un problème nouveau sur le plan algorithmique.

Plusieurs algorithmes de construction ont été proposés par Vincent Loonis et Marc Christine. Ces algorithmes ont tous pour point commun de construire une ZAEPC autour d'une commune-pivot constituant le « centre » de la ZAE. Les communes disponibles (non encore affectées à une étape donnée de la procédure) situées à une distance **à vol d'oiseau**[11] de la commune-pivot inférieure à une distance maximale fixée sont alors incorporées successivement à la ZAEPC, jusqu'à obtenir 300 logements par groupe de rotation. Dès que cette condition est remplie, la ZAEPC est constituée et les communes qui la composent sont définitivement affectées. Si les communes

[8] Dans le cas de l'Ile-de-France, les ZAE constituées respectent aussi la séparation grande couronne/petite couronne.

[9] Le choix de former une ZAEGC avec une seule grande commune vient de la nécessité de prendre en compte le plan de sondage RP qui utilise une base de sondage au niveau communal (RIL, Répertoire d'Immeubles Localisés), et calcule des poids au niveau communal. Par construction, une grande commune contient les cinq groupes de rotation. De plus, même dans le cas des grandes communes juste au-dessus du seuil de 10 000 habitants (cas extrême de la commune de Behren-les-Forbach, avec 10 073 habitants et 3312 résidences principales), le Recensement fournit chaque année un échantillon de plus de 250 logements, ce qui a été jugé suffisant pour assurer le tirage des enquêtes de l'année parmi les logements appartenant à la dernière campagne de Recensement.

[10] L'étendue d'une ZAE est mesurée ici comme le maximum sur les cinq années du cycle de recensement du déplacement moyen effectué par un enquêteur au sein de la ZAE une année donnée.

[11] Distance calculée sur la base de fonds de carte SAS au niveau communal en géographie au 1er janvier 1999.

disponibles situées à une distance inférieure au seuil limite ne permettent pas de remplir les critères de constitution (notamment si un groupe de rotation n'atteint pas la taille requise en nombre de logements), alors la tentative de constitution est un échec et les communes examinées restent non affectées.

En pratique, les algorithmes proposés diffèrent par le mode d'agrégation des communes (critère de taille minimale parmi les communes contiguës ou de distance minimale à la commune-pivot). Dans sa version définitive, l'algorithme proposé par Vincent Loonis prévoit une phase de « rejet » dans laquelle seules les communes les plus proches de la commune-pivot et strictement nécessaires à la constitution de la ZAEPC (pour le respect de la contrainte de taille) sont conservées dans la ZAEPC, les autres étant remises dans les communes « non encore affectées ». Cela permet ainsi d'augmenter le nombre de ZAEPC constituées et de diminuer leur nombre de communes (donc de diminuer le nombre de communes différentes dans lesquelles un enquêteur devra se rendre pour réaliser des enquêtes si la ZAEPC est tirée[12]).

Une phase « de vol » se déroule alors dans une région de la manière suivante :

- l'algorithme démarre en essayant de constituer une ZAEPC ayant pour pivot la commune la plus grande (en termes de nombre de résidences principales).

- à une étape quelconque de la phase de vol, l'algorithme essaie de constituer une ZAEPC en prenant comme commune-pivot la plus grande (en termes de nombre de résidences principales) des communes non encore affectées à ce stade.

- la phase de vol se termine quand la plus petite des communes non encore affectées a été examinée comme commune-pivot.

A l'issue de la phase de vol, une partie des communes n'ont pas été affectées. On passe alors à la phase « d'atterrissage » consistant à affecter les communes non affectées lors de la phase de vol à la ZAEPC la plus « proche » (c'est-à-dire à la ZAEPC ayant la commune-pivot la plus proche de la commune à affecter), sous réserve que la commune-pivot de la ZAE la plus proche soit située à une distance inférieure au seuil de distance maximale fixé[13].

Un paramètre essentiel est alors le « rayon » maximal de la ZAE, c'est-à-dire la distance maximale (à vol d'oiseau) à la commune-pivot d'une commune candidate pour appartenir à la ZAE. En effet, un « rayon » trop large conduit à constituer des ZAE très étendues, entraînant des déplacements importants pour les enquêteurs. Mais, à l'inverse, un « rayon » trop étroit conduit à constituer un nombre faible de ZAE, avec un grand nombre de communes non affectées à l'issue de la phase d'atterrissage. Ces communes doivent alors être affectées aux ZAE existantes bien qu'elles soient situées à une distance de la commune pivot supérieure au « rayon maximal » autorisé, ce qui entraîne également la création de ZAE de grande taille avec un grand nombre de communes.

De ce fait, un grand nombre de simulations de constitution de ZAE ont été lancées grâce à un applicatif développé par Bruno Berlemont, Chef de Projet Informatique du projet OCTOPUSSE, en faisant varier la « rayon » maximal de la ZAE.

La stratégie de choix de scénario s'est basée sur le critère « nombre de communes non affectées à l'issue de la phase d'atterrissage » : l'objectif était alors de voir jusqu'à quel seuil l'augmentation du « rayon » conduisait à diminuer de manière significative le nombre de communes non affectées en phase d'atterrissage.

Pour juger de la qualité des ZAE construites en termes de distances à parcourir pour les enquêteurs, ***on étudié l'étendue moyenne des ZAE constituées à l'issue de la phase d'atterrissage***, l'étendue d'une ZAE étant définie comme le maximum sur les cinq années du cycle RP du trajet moyen effectué

[12] L'objectif est ainsi de limiter le nombre d'endroits où l'enquêteur aura à trouver une place de stationnement, à éventuellement se signaler à la mairie…

[13] Voir dans la Présentation Power Point associée une illustration réalisée par Vincent Loonis du processus de constitution des ZAE dans le cas de la ZAE de Sainte-Gauburge.

par un enquêteur résidant dans la commune pivot de la ZAE pour atteindre un logement recensé dans la ZAE au cours de l'année (et donc susceptible d'être sélectionné pour une enquête).
En pratique, l'étendue d'une ZAEPC correspond au maximum sur cinq ans des « étendues annuelles moyennes» de la ZAE calculées une année donnée comme la distance moyenne des communes de la fraction recensée de la ZAEPC à la commune-pivot de la ZAEPC pondérées par le nombre de logements de chaque commune.

Sur la base des différents scénarios de constitution des ZAE, on a obtenu les résultats suivants :

Rayon maximal de la ZAE	Nombre de ZAEPC constituées	Nombre de petites communes non affectées à l'issue de la phase d'atterrissage	Etendue moyenne à l'issue d'atterrissage des ZAEPC constituées
10	1 788	**10 996**	7,8
15	2 565	**1 746**	10
18	2 779	**645**	10,9
19	2 848	**465**	11,2
20	2 886	**363**	11,4
21	2 944	**247**	11,7
22	2 969	**175**	11,9
23	3 005	**130**	12,1
24	3 037	**107**	12,3
25	3 056	**83**	12,5
26	3 093	**68**	12,7
27	3 115	**32**	12,9
28	3 144	**15**	13,2

On constate au vu des résultats que, pour des valeurs allant jusqu'à 19 km, le nombre de communes non affectées à l'issue de la phase de vol diminue beaucoup lorsqu'on augmente le « rayon maximal », tandis qu'au delà d'un rayon de 21 km, ce nombre diminue peu, alors que l'étendue des ZAEPC constituée augmente (on affecte ainsi 180 communes supplémentaires quand on augmente la rayon de 18 à 19km, alors qu'on n'affecte que 102 communes en plus quand on augmente la rayon de 19km à 20 km, 118 quand on l'augmente de 20 à 21 km et seulement 72 quand on passe de 21 à 22km). **Sur la base de ces résultats, le choix s'est porté sur un rayon maximal de 20 km.**

D'autre part, les simulations menées ont permis de vérifier l'impact de certains choix sur la « qualité » des ZAE constituées.

En effet, si les options prises pour fixer le scénario de référence (et notamment les contraintes comme le choix du seuil minimum de 250 logements ou 300 logements par groupes de rotation, la stratification régionale, la séparation grande commune/petites communes,…) l'ont été pour des raisons de fond (réservoir minimum de logements, simplification de l'application, cohérence avec l'EMEX[14],…) généralement incontournables, il est néanmoins intéressant d'avoir une estimation du coût lié à chacune de ces contraintes.

Une estimation du coût lié à une contrainte (ou du gain lié au fait de ne pas imposer une contrainte) peut être faite en comparant le nombre de communes affectées en présence et en l'absence de la contrainte, toutes choses égales par ailleurs[15]. On obtient alors les résultats suivants :

[14] Echantillon-Maître pour les Extensions Régionales, cf. article de M. CHRISTINE et E. GROS.
[15] Avec toujours une distance maximale de 20 km à vol d'oiseau.

Contrainte	Nombre de communes non affectées à l'issue de la phase d'atterrissage sans la contrainte	Nombre de communes non affectées à l'issue de la phase d'atterrissage avec la contrainte	« Coût » de la contrainte (en termes de nombre de communes non affectées)
Séparation ZAEGC/ZAEPC[16]	320	363	43
Stratification régionale[17]	168	363	195
Passage du seuil de 250 à 300 résidences principales par groupe de rotation	231	363	134

On constate ainsi que les contraintes introduites ont relativement peu d'impact sur le nombre de communes non affectées, la plus coûteuse étant la stratification régionale avec une hausse de 195 du nombre de communes non affectées

Au final, outre les 850 ZAEGC correspondant aux 850 grandes communes, le choix d'un scénario avec un rayon maximal de 20 km conduit à la constitution de 2886 ZAEPC. Suite à certaines modifications effectuées à la marge (rééclatement de quelques ZAE trop étendues en termes de distances réelles et constitution de quelques ZAE formées de quatre groupes uniquement), on obtient au final 2893 ZAEPC[18] regroupant 35 721 petites communes, dont 12 ZAEPC dont l'un des cinq groupes est manquant ou déficitaire (moins de 250 résidences principales dans le groupe).

Pour juger de l'optimalité en termes d'étendue, on a vérifié que les distances intra-ZAE étaient comparables avec celles des unités primaires constituées pour l'Echantillon-Maître actuel et demeuraient acceptables. Sur la base des travaux menés à cet effet, la moyenne des distances par la route entre la commune-pivot et les logements à enquêter sur les 2893 ZAEPC constituées est estimée à 10 km, à comparer à celle des 3202 UP 1999 constituées dans les strates « rural » et « petit urbain »[19], soit 8 km. La distance annuelle maximale moyenne pour les ZAEPC est de 18 km. Ces distances ont été considérées comme raisonnables, ce qui a permis de valider les ZAE constituées.

A noter en revanche que, contrairement aux Echantillons-Maîtres précédents, il n'a pas été possible de réaliser une stratification des unités primaires par types d'espace (pour l'Echantillon-Maître 1999, le critère de stratification utilisé était la taille d'unité unité urbaine répartie en cinq tranches).

[16] Lever cette contrainte signifie qu'il est possible d'affecter une petite commune à une ZAEGC en phase d'atterrissage

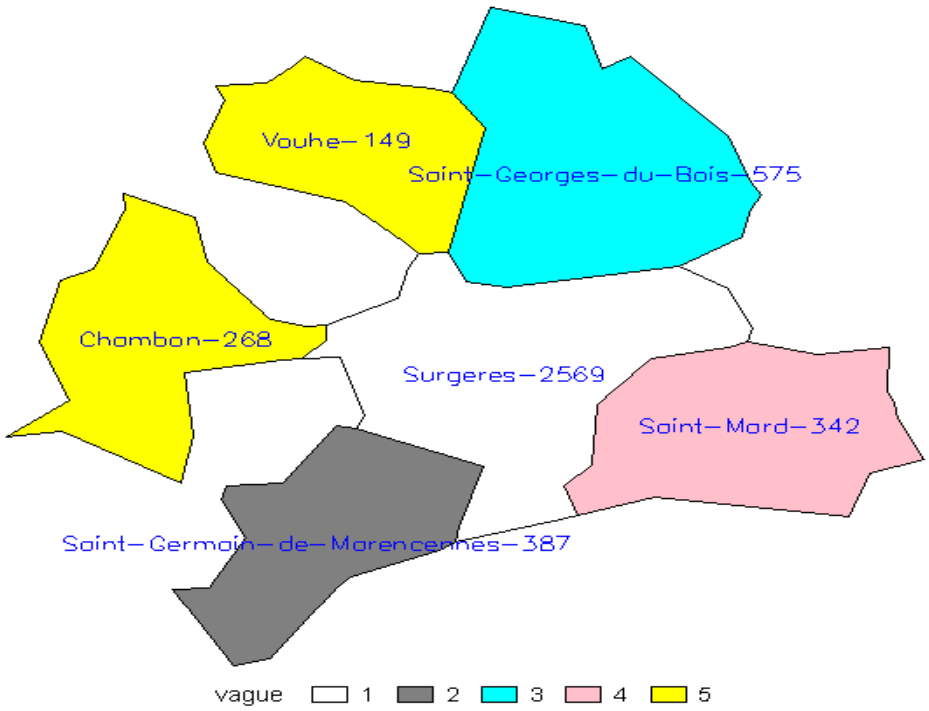
[17] Lever cette contrainte signifie qu'il est possible de constituer des ZAE à cheval sur plusieurs régions.

[18] Les communes restantes non encore affectées sont alors affectées à la ZAE dont la commune-pivot est la plus proche.

[19] Par convention, la « commune-pivot » d'une UP 1999 est la plus grande commune de l'UP en termes de nombre de résidences principales.

3.3. Exemple de tirage des échantillons dans le cas d'une ZAE constituée

Sélection des logements dans une ZAE formée de Petites Communes : exemple de la ZAEPC Z17434 (Surgères)



En pratique, dans le cas de la ZAE de Surgères, le tirage des logements à enquêter au sein de la ZAE est effectué de la manière suivante :

- En novembre 2009, la campagne de recensement de l'année 2009 (recensement du groupe 1) est chargée dans OCTOPUSSE. Dans la ZAE Z17434, c'est la commune de Surgères qui est recensée l'année du groupe 1. Pour les échantillons tirés entre novembre 2009 et novembre 2010, les logements tirés dans la ZAE sont tirés dans la commune de Surgères.

- En novembre 2010, c'est la campagne de Recensement de l'année 2010 (groupe 2 recensé) qui est chargée. Les logements tirés dans la ZAE entre novembre 2010 et novembre 2011 sont alors tirés dans la commune de Saint-Germain-de-Marencennes, qui appartient au groupe 2.

- De même, les logements tirés entre novembre 2011 et novembre 2012 dans la ZAE seront tirés à Saint-Georges du Bois (groupe 3), ceux tirés entre novembre 2012 et novembre 2013 à Saint-Mard (groupe 4) et enfin ceux tirés entre novembre 2013 et novembre 2014 à Chambon et Vouhé (groupe 5).

3.4. La ZAEPC : une unité primaire aléatoire

Une difficulté vient du fait que les ZAEPC sont des objets aléatoires, construits conditionnellement à l'affectation aléatoire des petites communes en groupes de rotation, effectuée par le Recensement. Ainsi, dans toute la suite, les probabilités manipulées correspondent en réalité à des probabilités conditionnelles au tirage des groupes de rotation des petites communes.

On donne en annexe B un exemple (cas de la commune de Vitry-en-Charollais) montrant que la probabilité de tirage d'une commune dans l'Echantillon-Maître OCTOPUSSE, conditionnellement au tirage des groupes de Rotation du Recensement, peut s'avérer très différente de la probabilité de tirage non conditionnelle (qui reste inconnue dans le cas général, faute de pouvoir simuler la constitution des ZAE sur un grand nombre de tirages de groupes de rotation du recensement)

Cet aspect devra ainsi être pris en compte dans les travaux de calcul de la précision des échantillons tirés dans OCTOPUSSE qui seront menés à partir de septembre 2009.

4. Allocations et tirage des ZAE de l'Echantillon-Maître national

4.1. Calcul du nombre de ZAE à tirer

Les ZAE sont tirées proportionnellement à leur taille (nombre de logements principaux), certaines étant retenues d'office (« exhaustives »). Le tirage a lieu de manière indépendante dans chaque région, avec dans le cas particulier de la région Ile-de-France une stratification grande couronne/petite couronne.

Le nombre de ZAE-EM à tirer a été fixé en prenant l'hypothèse conventionnelle suivante (analogue à celle prise pour l'EM 99) : **pour une enquête nationale au taux moyen de 1/2000** (un peu moins de 12.000 logements principaux), on **affecte 20 Fiches-adresses par enquêteur**[20].

On montre que le seuil d'exhaustivité vaut : $S = \frac{e}{\tau}$, pour un échantillon d'enquête au taux moyen τ, chaque enquêteur ayant une charge moyenne de e fiches-adresses. **Ce seuil ne dépend pas de la région considérée. On obtient aussi le nombre de ZAE à tirer dans la sous-strate non exhaustive :**

$$k = \frac{\tau(N - N^{exh})}{e} \text{ , où : } N^{exh} = \sum_{i / N_i \geq \frac{e}{\tau}} N_i ,$$

avec : N_i = la taille de la ZAE i et N=la taille de la région (nombre de logements principaux).

Principe du calcul du seuil d'exhaustivité

Considérons une région de taille N (en nombre de logements principaux). On veut tirer un échantillon d'enquête au taux moyen τ. La taille de l'échantillon de logements est donc : $n = \tau\ N$.

On suppose que chaque enquêteur a une charge moyenne de e fiches-adresses. Il en résulte un nombre d'enquêteurs nécessaire dans la strate de : $\tau\ N / e$.

On va tirer des ZAE à l'intérieur de la strate. Il y a deux types de ZAE :

- celles retenues exhaustivement : ce sont celles dont la taille est supérieure à un certain seuil S.

- celles tirées aléatoirement, avec une probabilité proportionnelle à leur taille.

[20] Le principe général étant qu'une ZAE est affectée à un seul enquêteur, hormis celles formées de grandes communes exhaustives.

On note N_i la taille de la ZAE i, N^{exh} la taille de la sous-strate exhaustive : $N^{exh} = \sum_{i/N_i \geq S} N_i$, k le nombre de ZAE à tirer dans la sous-strate non exhaustive.

- Dans la sous-strate exhaustive, la taille de l'échantillon de logements tirés sera : $n^{exh} = \tau\ N^{exh}$. Cet échantillon sera attribué à : $\tau\ N^{exh} / e$ enquêteurs.

- Dans la sous-strate non exhaustive, la probabilité de tirage d'une ZAE i est donc : $\boxed{\pi_i = k\ \frac{N_i}{N - N^{exh}}}$.

Chaque ZAE sera affectée à 1 enquêteur avec une charge de e fiches-adresses. Les ZAE tirées génèreront donc un échantillon de logements de : k e et on a la relation :

$\boxed{n = \tau\ N = \tau\ N^{exh} + k\ e}$, d'où la valeur de k : $\boxed{k = \frac{\tau(N - N^{exh})}{e}}$.

Il en résulte que : $\boxed{\pi_i = \frac{\tau N_i}{e}}$.

Ceci va permettre de définir le seuil d'exhaustivité : en effet, les probabilités de tirage π_i doivent être inférieures à 1, d'où la relation : $N_i < \frac{e}{\tau}$. Le seuil d'exhaustivité peut donc être défini par : $\boxed{S = \frac{e}{\tau}}$.

On notera que le seuil d'exhaustivité ne dépend pas de la strate considérée.

Résultats :

- le seuil d'exhaustivité résultant est à 40.000 logements principaux
- ***37*** grandes communes exhaustives (qui seront affectées à plusieurs enquêteurs), dont la liste est donnée en annexe B.
- ***488*** ZAE non exhaustives tirées dont :
  - 286 ZAE-PC
  - 202 ZAE-GC non exhaustives.

Par ailleurs, des ZAE complémentaires ont été tirées pour la constitution de l'EMEX[21] utilisé en cas de réalisation d'une extension régionale (échantillon complémentaire à l'enquête nationale tiré dans la région et financé par des organismes régionaux). En fonction de la taille de l'extension régionale à tirer, on mobilisera :

- soit les ZAE de l'EMEX restreint (ZAE de l'Echantillon-Maître national plus ZAE complémentaires tirées pour l'EMEX restreint), permettant de « doubler » le nombre de ZAE impactées dans la région par rapport à l'Echantillon-Maître national[22].

[21] Echantillon-Maître pour les Extensions régionales.

[22] Le nombre de ZAE tirées n'est pas exactement doublé compte-tenu du fait que certaines ZAE non exhaustives pour l'EM deviennent exhaustives pour l'EMEX restreint. Au final, 525 ZAE ont été tirées pour l'EM et 1013 pour l'EMEX restreint (et 1489 pour l'EMEX élargi).

- soit les ZAE de l'EMEX élargi (ZAE de l'Echantillon-Maître national plus ZAE complémentaires tirées pour l'EMEX restreint plus ZAE complémentaires tirées pour l'EMEX élargi), permettant de « tripler » le nombre de ZAE impactées dans la région par rapport à l'Echantillon-Maître national.

Le mode de tirage de l'EMEX est décrit dans l'article CHRISTINE/GROS.

4.2. Tirage des ZAE

Le tirage est stratifié *par région* (Rappel : dans le cas particulier de l'Ile-de-France, on sépare la « petite » et la « grande » couronne). Il est également équilibré sur des *totaux régionaux*:. Il est nécessaire d'équilibrer non seulement au niveau de l'ensemble de la ZAE *mais aussi de chacun des 5 groupes de rotation*, de manière à avoir chaque année une base de sondage « représentative ». Cela multiplie le nombre de contraintes d'équilibrage (cinq contraintes pour une variable) et réduit d'autant le nombre de variables indépendantes à introduire. On notera que la base de sondage annuelle est équilibrée sur le total des communes du groupe de rotation impacté mais pas sur le total France métropolitaine et que les groupes de rotation ne sont pas rigoureusement équivalents.

Interprétation de l'équilibrage lors du tirage des ZAE.

Lors du tirage des ZAE, on a introduit des conditions d'équilibrage **par groupe de rotation** de la forme :

$$\forall u \in \{1, \ldots, 5\} : \sum_{k \in s_{ZAE}} \frac{T_{k,u}(Z)}{\pi_k} = T_u(Z),$$

où Z est une variable d'équilibrage, $T_u(Z)$ est le **total** de cette variable **sur les communes du groupe de rotation u** et $T_{k,u}(Z)$ le total de Z sur l**es communes du groupe de rotation u au sein de la ZAE k.**

Il faut bien comprendre **que cet équilibrage doit être entendu conditionnellement à la répartition aléatoire en groupes de rotation.** En effet, le total sur lequel on équilibre, soit $T_u(Z)$, peut s'écrire sous la forme : $T_u(Z) = \sum_{i=1}^{C} T_i^C(Z) G_{i,u}$, où : $T_i^C(Z)$ représente le total de la variable Z dans la commune C_i. Le facteur $G_{i,u}$ intervient par le fait que l'on ne somme que sur les i tels que : $G_{i,u}$ = 1, c'est-à-dire les communes affectées au groupe de rotation u. **Ce total $T_u(Z)$ est donc aléatoire.**

Naturellement, par sommation sur u, on a : $\sum_{u=1}^{5} \sum_{k \in s_{ZAE}} \frac{T_{k,u}(Z)}{\pi_k} = \sum_{u=1}^{5} T_u(Z)$, soit :

$$\sum_{k \in s_{ZAE}} \frac{T_{k,\bullet}(Z)}{\pi_k} = T(Z).$$

Le total sur lequel on équilibre n'est plus aléatoire mais le terme de gauche, qui incorpore la probabilité de tirage des ZAE et, implicitement, leur mode de construction, dépend encore de l'aléa d'affectation des communes en groupes de rotation.

Si l'on prend l'espérance par rapport à cet aléa (et par rapport au tirage conditionnel des ZAE), on obtiendra : $\forall u \in \{1, \ldots, 5\} : E\left[\sum_{k \in s_{ZAE}} \frac{T_{k,u}(Z)}{\pi_k}\right] = E[T_u(Z)] = \sum_{v=1}^{5} \frac{1}{5} T_v(Z) = \frac{1}{5} \sum_{v=1}^{5} T_v(Z) = \frac{1}{5} T(Z)$,

soit encore :

$$E\left[5 \sum_{k \in s_{ZAE}} \frac{T_{k,u}(Z)}{\pi_k}\right] = T(Z).$$

Il s'agit très précisément d'un estimateur de type « en expansion », dans lequel on supposerait connus les vrais totaux $T_{k,u}(Z)$ de la variable Z sur les communes du groupe de rotation u de toute ZAE k, totaux que l'on remplace ensuite par des estimations à partir du second degré de tirage par les estimateurs $N_{k,u}\ \bar{y}_{k,u}$.

Le choix des variables d'équilibrage est issu de nombreuses simulations visant à déterminer les variables d'équilibrages « optimales » vis-à-vis de la qualité de l'équilibrage. On a finalement retenu : ***le nombre de résidences principales des ZAE par groupe de rotation ; le revenu fiscal 2004 ventilé par groupe de rotation ; enfin, le nombre de résidences principales dans les différents types d'espace (rural, périurbain et urbain)***.

Les simulations de tirage ayant conduit au choix des variables d'équilibrage pour le tirage des ZAE sont détaillées dans l'article de Fabien Guggemos présenté lors des JMS 2009.

Cas particulier de la région Ile-de-France :

En Ile-de-France, le tirage a été stratifié en « grande couronne » et « petite couronne » afin d'assurer une bonne représentation de ces deux zones géographiques aux caractéristiques différentes. Par ailleurs, il a été possible d'introduire des variables d'équilibrage supplémentaires compte-tenu du nombre plus important de ZAE tirées et des particularités de la région (quasi-absence de communes en zone rurale au sens de la classification ZAUER dans la région, ensemble des communes situées en zone urbaine et très faible nombre de petites communes dans la petite couronne…), qui rendent inutiles certaines conditions d'équilibrage standard. Ces variables d'équilibrage supplémentaires ont ainsi été déterminées en collaboration avec la DR d'Ile-de-France, pour rendre compte le mieux possible des spécificités de la région. Au final, les conditions d'équilibrage introduites sont les suivantes :

- Pour la grande couronne, l'équilibrage a été réalisé sur les variables suivantes : nombre de résidences principales par groupe de rotation, nombre de résidences total en zone périurbaine, revenu fiscal 2004 ventilé par groupe de rotation, nombre de personnes de moins de 20 ans, nombre de personnes entre 20 et 59 ans, nombre de personnes de 60 ans et plus, nombre d'étrangers, nombre de familles monoparentales, nombre de familles de grande taille (quatre enfants ou plus), nombre de propriétaires de leur logement, nombre de logements HLM et nombre de personnes habitant en logement collectif.

- En petite couronne, vu qu'il n'y a que 2 ZAEPC sur les 108 ZAE, l'équilibrage a été réalisé sur les variables suivantes, **non ventilées par groupe de rotation** : nombre de résidences principales, revenu fiscal 2004, nombre de personnes de moins de 20 ans, nombre de personnes entre 20 et 59 ans, nombre de personnes de 60 ans et plus, nombre d'étrangers, nombre de familles monoparentales, nombre de familles de grande taille (quatre enfants ou plus), nombre de propriétaires de leur logement, nombre de logements HLM et nombre de personnes habitant en logement collectif.

En ce qui concerne le tirage simultané de l'EM et de l'EMEX, le mode de tirage emboîté des ZAE appartenant aux différents domaines de tirage (tirage des ZAE formant l'EMEX élargi, puis tirage au sein de ces ZAE de celles formant l'EMEX restreint, puis tirage au sein de ces dernières des ZAE formant l'EM), tout en conservant à chaque étape un échantillon équilibré sur des totaux nationaux, est détaillé dans l'article de Marc Christine et Emmanuel Gros également présenté lors des JMS 2009.

Les unités primaires (ZAE) tirées en 2007 seront fixées et sont prévues au tirage des échantillons des enquêtes ménages pendant 10 ans entre 2009 et 2019, date à laquelle il sera sans doute nécessaire de procéder à un tirage de nouvelles unités primaires pour limiter les réinterrogations successives des logements des unités primaires tirées et pour prendre en compte les modifications démographiques survenues pendant les dix années écoulées**. A cette date, il sera cependant encore possible de continuer à utiliser la chaîne OCTOPUSSE en chargeant ces nouvelles unités primaires tirées.** Plus généralement, il est possible au niveau de l'application de changer les ZAE quand on le souhaite, de préférence sur des périodes de 5 ou 10 ans correspondant à un ou deux cycles complets de recensement.

5. Le tirage des logements dans les ZAE : calcul des allocations et des pondérations

5.1. Constitution des bases de sondage annuelles OCTOPUSSE

Une année donnée, la base de sondage annuelle OCTOPUSSE est constituée à partir des listes de logements suivantes issues de la dernière Enquête Annuelle de Recensement disponible :

- Dans les ZAECG tirées (ou exhaustives), liste des logements de la commune recensés lors de la dernière Enquête Annuelle de Recensement (environ 8% des logements de la commune)
- Dans les ZAEPC tirées, liste des logements du recensement exhaustif des communes appartenant à la fraction recensée de la ZAEPC (commune de la ZAEPC appartenant au groupe de rotation impacté par la dernière Enquête Annuelle de Recensement disponible)

Toutefois, dans les grandes communes, une phase intermédiaire est nécessaire. En effet :

– la 1ère phase RP (affectation des adresses aux groupes de rotation) conduit à une affectation inégale et « semi-déterministe » (et non à probabilités égales 1/5) des adresses dans les différents groupes de rotation (notamment les grandes adresses) : il est donc nécessaire de reconstituer une « pseudo-probabilité » d'affectation.

– lors de la 2ème phase RP (tirage de logements à recenser dans le groupe de rotation annuel en grande commune), il y a une surreprésentation des « adresses neuves » et des « grandes adresses » (recensées d'office dans chaque groupe de rotation d'adresses).

Ainsi, la probabilité qu'un logement soit recensé varie suivant le type d'adresse (grande, neuve ou autre) : cela nécessite donc une opération statistique préalable (***rééchantillonnage*** des logements situés en grandes adresses et en adresses neuves) pour disposer d'une base effective de logements à poids identiques. En pratique, l'objectif du rééchantillonnage est de ne conserver dans la base de sondage qu'une partie des logements recensés en grandes adresses et en adresses neuves de manière à éliminer la surreprésentation de ces strates par le Recensement.

Cadre conceptuel général de la procédure de rééchantillonnage.

Population U de taille N.

Echantillon S_1, tiré avec des probabilités d'inclusion d'ordre 1 : π_1^{2/S_1}.

Au sein de S_1 et *conditionnellement à ce dernier*, on veut tirer un échantillon S_2, avec des probabilités d'inclusion conditionnelles π_i^{2/S_1}, de telle sorte que l'estimateur en expansion d'un total,

s'appuyant sur les unités de l'échantillon final S_2, soit : $\sum_{i \in S_2} \frac{Y_i}{\pi_1^1 \pi_1^{2/S_1}}$, **affecte le même poids à toutes les unités de S_2**. On est donc dans le cadre d'un ***sondage en deux phases***.

On notera que π_1^{2/S_1} est une variable aléatoire dépendant de i et de S_1. Par construction, π_1^{2/S_1} doit être nul sur toutes les unités n'appartenant pas à S_1.

On doit donc imposer la condition : $\exists \omega, \ \forall i \in S_1 : \frac{1}{\pi_1^1 \pi_1^{2/S_1}} = \omega$.

Le poids ω ne doit pas dépendre de i mais peut (et doit) dépendre de S_1 : on le notera $\omega(S_1)$.

On peut s'astreindre à ce que, conditionnellement à S_1, l'échantillon final S_2 soit de taille fixe k_2 (*dépendant de S_1*), qu'on notera : $k_2(S_1)$. On doit donc vérifier la condition :

$\forall S_1 : \sum_{i \in S_1} \pi_i^{2/S_1} = k_2(S_1)$,

soit : $\forall S_1 : \sum_{i \in S_1} \frac{1}{\omega(S_1) \pi_i^1} = k_2(S_1)$.

On en déduit :

$\forall S_1 : \omega(S_1) = \frac{1}{k_2(S_1)} \sum_{i \in S_1} \frac{1}{\pi_i^1}$,

d'où, finalement :

$$\boxed{\forall S_1, \ \forall i \in S_1 : \pi_1^{2/S_1} = k_2(S_1) \frac{1/\pi_i^1}{\sum_{j \in S_1} (1/\pi_j^1)}}.$$

Le tirage conditionnel de S_2 sachant S_1 est donc un tirage à probabilités inégales, proportionnelles à l'inverse de la probabilité d'inclusion d'ordre 1 de l'unité i lors du tirage de S_1.

On peut chercher à faire en sorte de « perdre » le moins possible d'unités lors du rééchantilonnage, ce qui conduit à prendre pour taille de S_2 la valeur maximale possible de $k_2(S_1)$.

Comme on doit avoir : $\forall S_1, \ \forall i \in S_1 : \pi_1^{2/S_1} \leq 1$, ceci entraîne :

$k_2(S_1) \leq \frac{\sum_{j \in S_1} (1/\pi_j^1)}{1/\pi_i^1} = \pi_i^1 \sum_{j \in S_1} (1/\pi_j^1)$.

Comme cette inégalité doit être valable pour tout i de S_1, on en déduit :

$k_2(S_1) \leq (\underset{i \in S_1}{Min}\, \pi_i^1) \sum_{j \in S_1} (1/\pi_j^1)$.

On prendra donc comme valeur de la taille de S_2 :

$$k_2^*(S_1) = \text{Int}\,[(\underset{i \in S_1}{Min}\, \pi_i^1) \sum_{j \in S_1} (1/\pi_j^1)].$$

On obtient au final :

$$\forall S_1 : \omega(S_1) = \frac{1}{k_2^*(S_1)} \sum_{j \in S_1} \frac{1}{\pi_j^1} \# \frac{1}{\underset{i \in S_1}{Min}\, \pi_i^1}.$$

$$\forall S_1, \forall i \in S_1 : \pi_1^{2/S_1} = k_2^*(S_1) \frac{1/\pi_i^1}{\sum_{j \in S_1} 1/\pi_j^1} \# \frac{\underset{j \in S_1}{Min}\, \pi_j^1}{\pi_1^1}.$$

Un estimateur du total d'une variable Y, issu de cette procédure de rééchantillonnage, aura alors pour expression :

$$\hat{T}(Y) = \sum_{i \in S_2} \omega(S_1) Y_i = \omega(S_1) \sum_{i \in S_2} Y_i.$$

▲ Il peut y avoir des difficultés, en toute rigueur, à étudier ses propriétés et calculer correctement son biais et sa variance, dans la mesure où l'on ne peut ignorer sa dépendance par rapport au 1er échantillon S_1.

Ainsi, la principale innovation méthodologique dans la constitution des bases de sondage annuelles en grandes communes est l'application du processus de rééchantillonnage, qui vise à constituer à l'intérieur d'une ZAE une « base de sondage locale » dans à laquelle tous les logements de la ZAE ont la même probabilité d'appartenance. L'objectif est alors de pouvoir effectuer des tirages de logements à probabilités égales dans chacune des ZAE tirées, et éviter ainsi que la disjonction altère la représentativité de la base après chaque tirage.

Cette opération ne s'applique qu'au grandes communes, compte-tenu du recensement exhaustif des petites communes qui fait que tous les logements d'une petite commune ont la même probabilité d'être recensés, égale à 1/5.

Cependant, une difficulté vient du fait que **le RP ne gère pas explicitement de notion de probabilité de première phase en grandes communes**. Par ailleurs, les déséquilibres entre groupes de rotation sont relativement importants au niveau de la strate des grandes adresses, notamment dans les communes ayant moins de cinquante grandes adresses, du fait du mode de répartition initiale des grandes adresses dans ces communes[23]. (voir tableau annexe D). De plus, les mouvements importants de régularisation effectués entre 2004 et 2006 sur la strate des grandes adresses ne permettent pas de modéliser la présence des logements en grandes adresses dans les différents groupes de rotation en se basant sur le mode de répartition initial des grandes adresses.

[23] Dans ces communes, les grandes adresses ont été triées par nombres de logements décroissants puis réparties une par une dans les groupes de rotation successifs, la plus grande des grandes adresses étant placée dans un groupe sélectionné aléatoirement parmi les trois premiers groupes. A noter que ce déséquilibre, parfois très fort au niveau communal, s'estompe au niveau régional du fait d'une répartition initiale des grandes adresses plus équilibrée entre les groupes de rotation dans les grandes communes ayant plus de 50 grandes adresses.

De ce fait, il a été décidé de développer une méthodologie ad hoc de reconstitution des probabilités de première phase basée sur les données du RIL. **Ainsi, pour la strate des grandes adresses, la probabilité de première phase associée au groupe i (probabilité qu'un logement en grandes adresses de la commune se trouve dans le groupe i) est approchée par la part des logements en grandes adresses dans le groupe i (obtenue en divisant le nombre de logements en grandes adresses du groupe i par le nombre total de logements en grandes adresses de la commune)**. Pour ce faire, on importera chaque année dans OCTOPUSSE des données du RIL donnant pour chaque commune le nombre de logements par groupe de rotation et par strate (grandes adresses, adresses neuves et autres adresses). De même, la probabilité de première phase pour la strate des petites adresses (adresses neuves et autres adresses) est approchée par la part des logements en petites adresses dans le groupe i [24].

En pratique, on conservera donc dans OCTOPUSSE tous les logements recensés en « autres adresses » et une fraction des logements recensés en « grandes adresses » et en « adresses neuves » de telle sorte que tous les logements aient la même probabilité de présence finale dans OCTOPUSSE (la sélection des logements conservés en « grandes adresses » et en « adresses neuves » étant effectuée au moyen d'un tirage systématique à probabilités égales).

Un exemple de mise en œuvre du rééchantillonnage dans le cas de la commune de Grenoble est détaillé ci-dessous.

Exemple de rééchantillonnage des logements en grande commune

On prend ici l'exemple du chargement des logements du groupe 1 (en novembre 2009 par exemple), pour la ville de Grenoble, qui est une commune exhaustive pour OCTOPUSSE. D'après le RIL au 1er juillet 2006, on a sur les logements de Grenoble les données suivantes :

Nombre de logements	Grandes adresses	Adresses neuves	Autres adresses	Total
Groupe 1	1 396	137	15 743	17 276
Total commune	6 369	1 702	77 649	85 720

On suppose d'autre part un taux uniforme de 83% de résidences principales dans les logements recensés[25], quel que soit le groupe de rotation et la strate d'adresse RP.

Calcul des probabilités d'inclusion de première phase :

Pour les grandes adresses : $\pi \log R_{GA}^{1P} = \frac{1396}{6369} = 21{,}9\%$

Pour les petites adresses (adresses neuves et autres adresses) : $\pi \log R_{PA}^{1P} = \frac{137+15743}{1702+77649} = 20{,}0\%$

La probabilité de première phase (d'inclusion dans le groupe 1) d'un logement en adresse neuve ou en autre adresse est donc de $20{,}0\%$.

[24] Il n'a pas été jugé pertinent de calculer une probabilité spécifique pour la strate des adresses neuves, du fait notamment de l'hétérogénéité de la notion d'adresses neuves. En effet, une adresse est neuve tant que le groupe de rotation dans lequel elle se trouve n'a pas été recensé. Ainsi, dans le RIL au 1er juillet 2006, les adresses neuves du groupe 3 (recensé en 2006) ne contiennent que des adresses apparues depuis le 1er juillet 2005, les adresses neuves du groupe 2 (recensé en 2005) contiennent des adresses apparues entre juillet 2004 et juillet 2006, celles du groupe 1 des adresses apparues entre juillet 2003 et juillet 2006 et celles des groupes 4 et 5 (non encore recensés) des adresses apparues depuis le RP 1999.

[25] Taux de résidences principales constaté au niveau national au RP 1999. Il s'agit ici d'un exemple théorique : en pratique, on utilise le nombre de résidences principales effectivement recensé dans la commune dans chaque strate par le Recensement.

Probabilités de seconde phase (fournies par le RP) :

On tire un échantillon d'adresses comprenant au total 40% des logements du groupe de rotation. Le groupe de rotation 1 comprenant au total 17 276 adresses, on tire donc un échantillon total de $17276 * 0,4 \approx 6910$ logements. Dans cet échantillon, les logements du groupe 1 en grandes adresses et en adresses neuves sont inclus automatiquement (probabilité de seconde phase de 1). Il reste donc un échantillon de logements en autres adresses de taille $6910 - 137 - 1396 = 5377$ logements à tirer. La probabilité de seconde phase (probabilité de tirage au sein du groupe 1) d'un logement en « autres adresses » est donc : $\pi \log R_{AA}^{2P} = \frac{5377}{15743} = 34,2\%$.

Calcul des probabilités de chargement dans OCTOPUSSE

Cette probabilité est calculée comme le produit de la probabilité de première phase et de la probabilité de seconde phase . On obtient alors par type de logement les résultats suivants :

	Grandes adresses	Adresses neuves	Autres adresses
Proba 1ère phase	21,9%	20,0%	20,0%
Proba 2nde phase	1	1	34,2%
Proba chargement	21,9%	20,0%	6,84%

On conserve l'ensemble des résidences principales recensées en autres adresses (qui représentent par hypothèse 83% des logements recensés en autres adresses), soit $5377 \times 0,83 \approx 4463$ résidences principales en autres adresses, avec une probabilité de chargement de $6,84\%$ pour une résidence principale en autres adresses.

Rééchantillonnage des logements en adresses neuves :

-on conserve une fraction $\frac{6,84}{20,0} = 34,2\%$ des résidences principales en adresses neuves. Comme on a recensé 137 logements en adresses neuves, on a chargé dans OCTOPUSSE $137 \times 0,83 \approx 114$ résidences principales : on tirera au moment du rééchantillonnage un échantillon de $114 \times 0,342 \approx 39$ logements en adresses neuves (soit un coefficient de rééchantillonnage de $\frac{39}{114}$). La probabilité d'inclusion finale dans OCTOPUSSE d'un logement en adresse neuve sera donc :

$$\pi \log FIN_{AN} = \pi \log R_{PA}^{1P} \times \frac{39}{114} \approx 6,84\%$$

Rééchantillonnage des grandes adresses :

On conserve alors une fraction $\frac{6,84}{21,9} \approx 31,2\%$ des résidences principales en grandes adresses : comme on a recensé 1396 logements en grandes adresses, on a chargé au final $1396 \times 0,83 \approx 1159$ résidences principales dans OCTOPUSSE, parmi lesquelles on garde au final $1159 \times 0,312 \approx 362$ résidences principales dans OCTOPUSSE (soit un coefficient de rééchantillonnage de $\frac{362}{1159}$). On tirera donc lors du rééchantillonnage 362 résidences principales en grandes adresses sur les 1159 chargées. La probabilité d'inclusion finale dans OCTOPUSSE d'un logement en grande adresse vaut donc :

$$\pi \log FIN_{GA} = \pi \log R_{GA}^{1P} \times \frac{362}{1159} \approx 6,84\%$$

On a donc constitué au final une base de sondage dans la commune de Grenoble de $4463 + 39 + 362 = 4864$ logements principaux dans laquelle on peut réaliser des tirages successifs à probabilités égales.

Au final, la base de sondage annuelle OCTOPUSSE est constituée à partir des listes de logements de l'Enquête Annuelle de Recensement par :

- **les logements de l'EAR conservés à l'issue du rééchantillonnage dans les ZAEGC tirées**
- **les logements appartenant à la fraction recensée (logements des petites communes appartenant au groupe de rotation impacté) dans les ZAEPC tirées.**

5.2. Calcul de l'allocation de logements à tirer dans chaque ZAE

Les calculs d'allocation cherchent à ***minimiser la dispersion des pondérations finales logements***, sous des contraintes :

- de taille totale d'échantillon fixée
- de charge minimale et maximale par enquêteur (c'est-à-dire par ZAE non exhaustive) : par exemple, fourchette 20-40 pour une enquête de 20 000 logements.

En pratique, l'algorithme de calcul des allocations de logements fonctionne de la manière suivante :

- une allocation régionale est calculée pour chaque région en ventilant l'allocation totale à tirer entre les différentes régions proportionnellement au nombre de résidences principales de chaque région (sur la base des estimations fournies par le Recensement)
- dans chaque région, l'allocation à tirer est ventilée entre chacune des ZAE exhaustives et « l'ensemble des ZAE non exhaustives » proportionnellement à la taille de chaque ZAE exhaustive[26] et de la zone « ensemble des communes non exhaustives de la région»
- l'allocation à tirer dans la zone « ensemble des communes non exhaustives de la région » est ventilée entre les ZAE non exhaustives tirées dans la région, de façon à minimiser la dispersion du poids final des logements sous des contraintes de nombre minimum et maximum de logements dans chaque ZAE non exhaustive. La minimisation peut se faire au niveau de chaque région, au bien au niveau national sur la zone « ensemble des communes non exhaustives de « France Métropolitaine ». L'algorithme de minimisation mis en œuvre, proposé par Vincent Loonis, est détaillé en annexe D.

En l'absence de contraintes sur le nombre de fiches-adresses à tirer par ZAE non exhaustive, on obtiendrait ainsi, en ajustant les allocations, l'équipondération des logements. Mais au prix d'allocations très différentes d'une ZAE à l'autre et d'une année à l'autre.

Cependant, compte-tenu des bornes minimum et maximum imposées aux allocations dans les ZAE non exhaustives, il n'est pas possible d'obtenir l'égalité des poids des logements tirés : plus les poids des ZAE sont dispersés et plus les allocations assurant l'équipondération des logements tombent en dehors de la plage autorisée et doivent alors être ramenées à la borne inférieure ou à la borne supérieure.

5.3. Procédure de tirage des logements dans les ZAE

Le tirage des logements dans les ZAE est effectué de la manière suivante :

Pour un tirage dans la dernière campagne, la « base de sondage utile » de la ZAE est constituée de l'ensemble des logements de la ZAE appartenant à la dernière campagne et échantillonnables (n'ayant pas été déjà tirés dans un échantillon antérieur et n'ayant pas été marqués dans le cadre de la disjonction avec l'enquête Emploi, cf. infra.). Ces logements sont triés selon des critères de tri (5

[26] Ce qui assure le taux de sondage moyen général dans chaque ZAE exhaustive.

variables au maximum) choisis par l'expert-sondage[27]. Les logements sont ensuite tirés par tirage systématique à probabilités égales. Au cas où les logements chargés pour la dernière campagne sont épuisés, on tire alors parmi les logements échantillonnables des campagnes antérieures (campagne N-1, puis N-2, puis N-3, puis N-4). Dans ce cas, on affecte aux logements tirés dans les campagnes antérieures un poids correspondant au poids d'un logement « standard » (logement quelconque en petites communes et logements de la strate des « autres adresses » en grandes communes) tiré dans la campagne active dans la même ZAE.

Pour un tirage dans cinq campagnes en ZAEPC (voir cas exceptionnel de l'enquête Victimation au paragraphe 7.3.2), on calcule à partir de l'allocation globale à tirer dans la ZAEPC cinq sous-allocations (à tirer dans chacune des cinq bases annuelles de la ZAE) proportionnellement au nombre de résidences principales situées dans chacun des cinq groupes de rotation de la ZAE. On effectue ensuite de manière indépendante le tirage systématique de l'allocation à tirer dans chacune des cinq bases de sondage annuelles. Si l'une des bases de sondage annuelle de la ZAE est épuisée, on se contente de tirer les logements disponibles restants mais on ne compense pas en tirant plus de logements dans les autres bases de sondage annuelles. On obtient alors au final un échantillon incomplet.

Remarque : pour certaines enquêtes sur des domaines présentant des effets de saisonnalité (Emploi du Temps, Budget de Famille…), il est nécessaire de répartir l'échantillon total en plusieurs vagues de collecte. Le processus de répartition par vagues d'un échantillon est donné en annexe F.

5.4. Calcul du poids de sondage final des logements

La « probabilité » de tirage d'un logement dans l'échantillon est calculé comme ***le produit de la probabilité de « présence finale » du logement dans la base de sondage annuelle OCTOPUSSE impactée par le tirage*** par ***la probabilité de tirage dans l'échantillon sachant qu'il est présent dans la base de sondage annuelle OCTOPUSSE (égale au rapport entre le nombre de logements tirés dans la ZAE et le nombre de logements chargés dans la ZAE pour la campagne active)***.

Dans le cas d'un logement appartenant à une ZAEPC, la « probabilité de présence finale » vaut :

$$\pi LogFIN = \pi ZAE \times \frac{1}{5}$$

où πZAE est la probabilité de tirage de la ZAE.

D'autre part, la probabilité de tirage du logement dans l'échantillon sachant qu'il appartient à la base de sondage OCTOPUSSE vaut $\frac{n_{ZAE}}{N_{ZAE,GRi}}$ où n_{ZAE} est l'allocation tirée dans la ZAE et $N_{ZAE,GRi}$ le nombre de logements chargés dans OCTOPUSSE pour le groupe de rotation i correspondant à la campagne active.

Dans le cas d'une ZAEPC, le poids de sondage final d'un logement tiré dans une ZAEPC, égal à l'inverse de la « probabilité de tirage dans l'échantillon » vaut donc :

$$wLog = \frac{1}{\pi LogFIN} x \frac{1}{\frac{n_{ZAE}}{N_{ZAE,GRi}}}$$

[27] Le tri de la variable de sondage avant le tirage systématique des logements a pour but d'assurer le tirage d'un échantillon de logements « représentatif » de la base de sondage au sens de la variable de tri. On choisira donc comme critère de tri les variables issues du recensement qui semblent les plus corrélées avec le phénomène étudié (par exemple, dans le cas de l'enquête IVQ sur l'illettrisme, on pourra choisir parmi les critères de tri le niveau de diplôme du chef de ménage).

$$wLog = \frac{5}{\pi ZAE} x \frac{N_{ZAE,GRi}}{n_{ZAE}}$$

Dans le cas d'un logement appartenant à une ZAEGC, la probabilité de présence finale dans base annuelle OCTOPUSSE dépend de la strate d'adresse à laquelle le logement appartient et fait intervenir, outre le tirage des ZAE, le plan de sondage du recensement (probabilité de tirage de première phase et de seconde phase), ainsi que le rééchantillonnage des logements pour les strates dans lequel il est effectué.

De manière générale, la probabilité de présence finale d'un logement dans la base de sondage annuelle OCTOPUSSE (qui dépend de la ZAEGC et de la strate d'adresse à laquelle le logement appartient) vaut :

$$\pi LogFIN = \pi ZAE \times \pi LogR^{1P} \times \pi LogR^{2P} \times Coeff\ \mathrm{Re}ech$$

où πZAE est la probabilité de tirage de la ZAE, $\pi LogR^{1P}$ la probabilité de première phase du recensement modélisée (probabilité d'affectation du logement à un groupe de rotation, dont la valeur est fonction de la strate d'adresses à laquelle le logement appartient), $\pi LogR^{2P}$ la probabilité de seconde phase du recensement (probabilité de tirage du logement pour l'Enquête Annuelle de Recensement sachant qu'il appartient au groupe de rotation impacté, valant 1 pour les logements des strates des grandes adresses et des adresses neuves) et $Coeff\ \mathrm{Re}ech$ la part des logements de la strate d'adresse à laquelle le logement appartient chargés dans OCTOPUSSE conservés dans la base de sondage annuelle OCTOPUSSE à l'issue du rééchantillonnage (ce coefficient valant 1 pour les logements en « autres adresses »).

On a alors :

$$wLog = \frac{1}{\pi ZAE_{ZAE} \times \pi LogR^{1P}_{ZAE,SA} \times \pi LogR^{2P}_{ZAE,SA} \times Coeff\ \mathrm{Re}ech_{ZAE,SA}} x \frac{N_{ZAE,GRi}}{n_{ZAE}}$$

5.5. Gestion de la disjonction avec l'échantillon Emploi

La disjonction entre l'enquête Emploi et les autres enquêtes ménages tirées dans OCTOPUSSE constitue un impératif de qualité essentiel pour limiter la charge d'enquête pesant sur les ménages. Elle correspond à un engagement important de l'INSEE vis-à-vis de ses partenaires extérieurs (CNIS, Eurostat…).

Compte-tenu du caractère prioritaire de l'enquête Emploi, l'objectif est alors de marquer dans OCTOPUSSE les logements impactés par l'Echantillon Emploi afin qu'ils ne puissent pas être sélectionnés dans l'échantillon d'une autre enquête ménages.

Pour les échantillons actuels (construits sur la base du RP 1999), la disjonction a pu être gérée de manière fine au niveau logement sur la base de l'identifiant du logement au RP 1999, commun aux deux échantillons : les logements appartenant aux aires Emploi tirées ont alors été directement marqués dans l'Echantillon-Maître 1999.

Ce processus de disjonction apparaît plus difficile à mettre en œuvre pour les nouveaux échantillons entrant en vigueur en 2009 (nouveaux échantillons Emploi et OCTOPUSSE[28]) compte tenu du fait que les bases de sondage employées sont différentes : fichiers de la Taxe d'Habitation pour le nouvel

[28] Il n'est pas prévu de gérer une disjonction entre les nouveaux échantillons et ceux issus du RP 1999, au moins en petites communes. Il est donc possible de tirer via OCTOPUSSE un logement appartenant aux aires Emploi de l'échantillon Emploi actuel construit sur la base du RP 1999.

Echantillon Emploi et données des Enquêtes Annuelles de Recensement pour les échantillons tirés dans OCTOPUSSE.

En particulier, en l'absence d'identifiant logement commun entre les fichiers TH et RP, la disjonction ne pourra pas être effectuée au niveau logement comme cela avait été le cas pour les échantillons issus du RP 1999.

De ce fait, il est nécessaire d'envisager une disjonction à un niveau moins fin que le logement.

- En grandes communes, il est possible d'effectuer une disjonction au niveau adresse (marquage dans OCTOPUSSE de tous les logements d'une adresse impactée par l'enquête Emploi) grâce à l'appariement RP/TH au niveau adresse. En pratique, il est nécessaire de récupérer la liste des identifiants CICN2 des adresses impactées par l'enquête Emploi. Cette opération a été effectuée par la division ETSD grâce à un accès à la table de correspondance du RP entre adresse TH (code Rivoli+numéro dans la voie) et identifiant CICN2 de l'adresse : la liste des codes CICN2 des adresses impactées par l'enquête Emploi en grandes communes a été produite en juillet[29]. L'échantillon Emploi 2009-2019 compte 86012 adresses en grandes communes (hors adresses neuves rajoutées fin 2008), dont 38199 dans les ZAEGC EM sur lesquelles on constate un taux d'appariement de 90% avec le RIL (34 397 adresses appariées). Il n'est pas possible de gérer de disjonction pour les adresses non appariées, mais la probabilité d'interroger un logement d'une adresse Emploi non appariée dans une autre enquête ménage devient très faible.

- En petites communes, la disjonction est effectuée au niveau du district de collecte RP[30] (marquage dans OCTOPUSSE des logements des districts RP impactés par un ou plusieurs secteurs Emploi) pour les petites communes qui ont plusieurs districts de collecte, ou, à défaut, au niveau de la commune entière pour celles qui ne sont pas découpées en districts. Pour les communes concernées ayant plusieurs districts, l'identification des districts impactés est effectuée par les services des Directions Régionales de l'INSEE. Au total, sur les 215 petites communes concernées, 56 n'avaient qu'un seul district de collecte et ont été entièrement marquées. Le repérage des districts de collecte en petites communes a donc concerné 159 petites communes. Au total, 528 districts de collecte RP ont été identifiés comme impactés par l'enquête Emploi sur les 2505 districts de collecte constitués dans ces 159 communes.

6. Amélioration de la qualité des bases de sondage annuelles : le principe du calage des ZAE

6.1. Importance de la « représentativité » des unités primaires tirées pour l'Echantillon-Maître

La représentativité des unités primaires tirées pour l'Echantillon-Maître est un paramètre essentiel de la qualité des échantillons tirés.

En effet, les Echantillon-Maîtres sont fondés sur le plan méthodologique sur un système de tirage des logements à deux degrés :

- tirage au premier degré d'un échantillon d'unités primaires fixées une fois pour toutes (ZAE)
- tirage au deuxième degré dans chaque unité primaire d'un échantillon de logements « représentatifs » de l'unité primaire[31]

[29] Elle devra cependant être mise à jour avec les codes CICN2 des adresses des logements neufs rajoutés aux grappes Emploi fin 2008.

[30] Les districts de collecte RP constituent le seul découpage infra-communal disponible pour les logements des EAR en petites communes.

[31] Dans cette note la « représentativité » doit être comprise au sens de fidélité de l'échantillon des ZAE tirées par rapport à l'univers ; cette fidélité est mesurée par l'intermédiaire des erreurs relatives observées de l'estimation de HORVITZ-THOMSON (à partir des ZAE tirées ou de la fraction annuelle impactée) du total d'un certain nombre de variables auxiliaires jugées pertinentes, par rapport aux vraies valeurs connues dans l'univers. Une bonne représentativité est d'autant plus importante que la variance de premier degré est généralement prépondérante dans la variance totale.

Ainsi, la représentativité des ZAE tirées conditionne la représentativité des échantillons de logements tirés par OCTOPUSSE.

6.2. Mesure dans un cadre général de la représentativité d'un échantillon d'unités primaires tirées

La représentativité d'un échantillon d'Unités Primaires tirées ne peut être mesurée par définition que pour des variables auxiliaires connues sur l'ensemble de la population (Variables issues du RP 1999 puis variables issues des premières populations légales à partir de mars 2009[32]) pour lesquelles on vérifie que les Unités Primaires tirées sont bien équilibrées, non seulement sur les variables d'équilibrage introduites lors du tirage mais également sur d'autres variables indépendantes.

En pratique, on étudie la « représentativité » des Unités Primaires tirées pour une variable auxiliaire donnée en comparant l'estimateur du total « France entière » de cette variable, obtenu à partir des totaux observés sur les ZAE tirées, avec le vrai total France entière : par exemple, pour étudier la « représentativité » des unités primaires tirées pour la variable « nombre de personnes de moins de 20 ans » (au RP 1999), on estime le nombre total de personnes de moins 20 ans en France à partir du nombre total de personnes de moins de 20 ans dans chacune des ZAE tirées et on compare l'estimation obtenue avec le vrai nombre total de personnes de moins de 20 ans.

Dans le cadre d'un tirage à deux degrés standard, pour une variable d'intérêt Y donnée, on calcule l'erreur relative entre la vraie valeur $T(Y)$ du total « France entière » de la variable Y avec l'estimateur de ce total sur l'échantillon d'Unités Primaires tirées :

$$\hat{T}(Y) = \sum_{UPtirees} d_k T_k$$

où T_k est le total de la variable Y sur l'ensemble des logements de l'unité primaire et $d_k = \frac{1}{\pi_k}$ le poids de sondage associé à l'Unité Primaire k (égal à l'inverse de la probabilité d'inclusion de l'unité primaire). L'erreur relative pour la variable d'intérêt Y de l'échantillon d'unités primaires tiré est alors calculée comme :

$$ERR_Y = \frac{\hat{T}(Y) - T(Y)}{T(Y)}$$

Dans le cadre d'OCTOPUSSE, le problème est un peu plus compliqué compte-tenu du caractère rotatif du recensement. Ainsi, une année donnée, les unités géographiques dans lesquelles les logements seront tirées sont :

- les ZAEGC tirées (dont le territoire est entièrement couvert par le plan de sondage du Recensement compte-tenu de l'équilibrage au niveau IRIS des échantillons tirés)
- la fraction recensée des ZAEPC tirées

On a alors l'estimation suivante du total « France entière » d'une variable

$$\hat{T}(Y) = 5 \sum_{k \in s \cap ZAEPC} \frac{T_{k,t}}{\pi_k} + \sum_{k \in s \cap ZAEGC} \frac{T_k}{\pi_k}.$$

Où T_k représente le total de la variable Y sur l'unité primaire k

[32] Données non encore disponibles au moment de la rédaction de cet article, qui utilise donc pour l'étude les données du RP 1999.

Et $T_{k,t}$ le total de la variable Y sur la fraction de l'unité primaire k recensée l'année t

On a donc les poids d'extrapolations initiaux suivants : $d_k = \frac{1}{\pi_k}$ pour les ZAEGC et $d_k = \frac{5}{\pi_k}$ pour les fractions recensées des ZAEPC.

Il convient cependant de rappeler que ceci est conditionnel à l'affectation des communes en groupes de rotation (la constitution des ZAE et les π_{ZAE} sont en fait aléatoires).

6.3. Résultats observés sur la représentativité des bases de sondage annuelles

Les résultats en termes d'équilibrage des ZAE tirées (sur la base du tirage effectif des ZAE) ont montré des fluctuations très importantes d'une année sur l'autre, et tout particulièrement pour les variables de segmentation ZAUER « urbain/periurbain/rural ».

Exemple : erreur relative observée sur la variable « nombre de résidences en zone rurale au RP 1999» en fonction des groupes de rotation

Groupe de rotation	Erreur relative sur la variable « nombre de résidences principales en zone rurale »
GR1	+3,4%
GR2	-3,3%
GR3	-7,9%
GR4	-8,1%
GR5	-9,4%

Ainsi, un tirage dans le groupe 1 conduit à retenir un échantillon de communes présentant une erreur relative de +3,4% en termes de nombre de résidences principales en zone rurale, alors qu'un tirage dans le groupe 5 conduit à une erreur relative de -9,4% en termes de résidences principales en zone rurale.

On observe également des variations importantes au niveau des variables « nombre d'emploi par secteur », avec, pour la variable « nombre de personnes employées dans l'industrie au RP 1999», une erreur relative de -3,9% dans le groupe 3 et de +5,0% dans le groupe 4.

6.4. Une solution de calage des ZAE

Le moyen retenu pour améliorer la représentativité des bases de sondage annuelles a été de caler les ZAE tirées sur un certain nombre de variables d'intérêt (segmentation urbain/peri-urbain/rural, emploi par secteur, tranches d'unités urbaines) ce qui conduit à repondérer les ZAE pour que l'estimation des variables auxiliaires considérées à partir de l'échantillon de ZAE tirées coïncide avec le total « France entière » de ces variables d'intérêt.

Procédure de calage des ZAE

Le calage des ZAE a pour but de calculer un nouveau système de pondération des ZAE $wcale_k$ permettant pour certaines variables auxiliaires $x^1,...,x^N$ disponibles pour toutes les ZAE d'obtenir une estimation exacte du total (noté X^j pour la variable auxiliaire x^j) à partir de l'échantillon de ZAE tiré tout en conservant des pondérations les plus proches possibles des pondérations initiales d_k (au sens d'une certaine distance D fixée par l'utilisateur). La pondération finale des ZAE est donc obtenue en résolvant le programme de minimisation sous contraintes suivant :

$$Min\left[D(w_k, d_k)\right]$$
$$sc : \forall j = 1..n, \sum_{ZAEtirées}\left(w_k \times x_k^j\right) = X^j$$

Ainsi, plus le nombre de variables de calage augmente et plus l'espace dans lequel on recherche la pondération finale se réduit, ce qui limite les possibilités de minimisation et conduit en règle générale à une augmentation de la distance entre la pondération initiale et la pondération finale (c'est-à-dire à une plus grande distorsion des poids liée au calage).

On peut alors considérer qu'avec le calage de ZAE, on arrivera à tirer un échantillon de logements représentatif dans la mesure où :

- on aura au premier degré une « base de sondage annuelle » (échantillon de communes formé par les fractions recensées lors de la dernière campagne RP des ZAE tirées) représentative de l'ensemble du territoire.

- on tirera au sein de chacune des « fractions recensées » des ZAE tirées un échantillon de logements représentatif de la fraction recensée.

D'autre part, compte-tenu du mode de calcul des allocations (par minimisation de la dispersion du poids final des logements sous des contraintes de nombre minimum et maximum de fiches-adresses tirées par ZAE non exhaustives, cf. supra paragraphe 5.2.), le calage des ZAE permet en théorie d'améliorer la représentativité des échantillons de logements tirés en augmentant les allocations dans les types de zones sous-représentées dans l'échantillon de ZAE[33]. Cependant, les conséquences du calage sur le poids des logements doivent être étudiées afin de vérifier que cela n'entraîne pas une trop grande dispersion des poids, source de variance des estimations.

En pratique, le calage des ZAE est effectué séparément pour chacune des cinq bases de sondage annuelles sur les 488 ZAE non exhaustives tirées pour l'Echantillon-Maître national : de ce fait, la pondération calée des ZAE non exhaustives varie suivant le groupe de rotation chargé dans OCTOPUSSE.

6.5. Variables de calage retenues

Après comparaison de plusieurs scenarii de calage[34], les variables de calage retenues ont été :

- les variables d'équilibrage lors du tirage des ZAE (nombre de résidences principales, revenu fiscal total, nombre de résidences principales dans les espaces urbain/périurbain/rural)

- l'âge en trois tranches (nombre de personnes de moins de 20 ans, nombre personnes de 20 à 59 ans et nombre de personnes de plus de 60 ans)

[33] Ce point est détaillé en annexe F.

[34] Travail effectué au début 2008 sur la base des données du RP 1999.

- les variables d'emploi par secteur (agriculture, industrie, construction, tertiaire)

- la répartition des logements principaux par tranches de taille d'unité urbaine.

En pratique, **à partir de la campagne 2008**, le calage sera effectué sur la base des données détaillées du dernier cycle de recensement : le calcul s'appuiera donc sur les données diffusées des populations légales sous forme de fichier regroupant les listes de logements et d'individus recensés lors des cinq dernières campagnes de recensement, avec pour chaque logement recensé un poids population légale. **Pour la campagne de l'année N, le calage des ZAE sera donc effectué sur des variables au 1er janvier N-2.**

6.6. Impact du calage sur la représentativité des bases de sondage annuelles

6.6.1. Variables d'intérêt étudiées

Les variables étudiées sur lesquelles on juge de la qualité des différents scenarii proviennent de différentes sources :
- variables sociodémographiques du RP 1999 (répartition par âge, type de ménage, emploi par secteur, position professionnelle, niveau de diplôme, date d'achèvement du logement…).
- variables fiscales des données de la TH (revenu global du foyer, nombre total de parts fiscales, nombre de personnes de référence mariées…)
- variables ANPE (DEFM)

Les variables issues des sources TH et ANPE (disponibles au niveau communal), qui n'ont pas servi dans la mise au point de la liste des variables de calage, permettent notamment de vérifier que le calage ne dégrade pas la « représentativité » des bases de sondage annuelles pour des variables « indépendantes » de la source RP 1999.

6.6.2. Résultats de l'étude

Les travaux menés montrent que le calage permet d'obtenir une erreur relative égale à zéro sur les variables de calage, sans la voir augmenter pour les autres variables d'intérêt, qu'elles soient issus du recensement ou de sources administratives indépendantes[35]. A titre d'exemple, on observe les erreurs relatives suivantes pour la variable « DEFM au 31 décembre 2007 » (total égal à 2 100 415)

Groupe de rotation	Erreur relative « avant calage »	Erreur relative « après calage »
GR1	+ 0,2	-0,5
GR2	- 0,7	-0,2
GR3	+0,0	+0,8
GR4	-1,2	-0,6
GR5	+0,5	+0,0

On constate ainsi que l'erreur relative maximale après calage reste toujours inférieure à 1% en valeur absolue, et n'augmente pas de manière systématique par rapport à celle observée avant calage.

6.7. Impact du calage des ZAE sur le poids final des logements

6.7.1. Calcul du poids de sondage final des logements avec calage des ZAE

Le poids de sondage final des logements est calculé selon les formules données au paragraphe 5.4, dans lesquelles on remplace les poids de sondage des ZAEGC et les poids de sondage des fractions recensées des ZAEPC par les poids calés.

[35] Les résultats de l'étude sur différentes variables sont présentés en annexe G.

- Pour une ZAEGC, on remplace son poids de sondage $d_k = \frac{1}{\pi ZAE}$ par le poids calé de la ZAEGC.
- Pour une ZAEPC, on remplace le poids de sondage de la fraction recensée $d_k = \frac{5}{\pi ZAE}$ par le poids calé de la fraction recensée de la ZAEPC.

6.7.2. Effet du calage sur la dispersion des poids des logements d'un échantillon

Le calage des ZAE a pour effet de disperser leur poids, la dispersion étant d'autant plus grande que le nombre de variables de calage est important. Le paramétrage choisi pour l'algorithme de calage (fonction de lien « linéaire tronquée ») permet de borner le rapport entre le poids final et le poids initial de la ZAE. Dans les études menées ici, on a imposé que ce rapport soit compris entre 0.1 et 10, ce qui signifie que le poids final de la ZAE est au moins 10 fois inférieur et au plus 10 fois supérieur à son poids initial, ce qui rend possible tout de même une dispersion finale importante.

Il apparaît donc nécessaire d'étudier en détail l'impact du calage des ZAE sur le poids des logements. Dans toute la suite, on se place dans la situation suivante : tirage d'un échantillon de 20 000 logements avec un nombre minimum de 20 fiches-adresses et un nombre maximum de 40 fiches-adresses tirées dans chaque ZAE non exhaustive.

On a alors simulé le calcul des allocations à tirer dans les ZAE de l'EM dans le cas où les ZAE ne sont pas calées et dans le cas où elles sont calées.

Tout d'abord, l'analyse des allocations montre que le calage des ZAE augmente la proportion (en moyenne sur les cinq groupes) de ZAE pour lesquelles le calcul des allocations bute sur une des bornes minimum ou maximum. Ainsi, on bute sur une borne dans 57,7% des ZAE non exhaustives avec le calage contre 53,4% des ZAE non exhaustives en l'absence de calage

Proportion moyenne sur cinq campagnes	ZAE non calées	ZAE calées
ZAE ayant une allocation égale à 20 logements (borne minimum)	15,4%	13,8%
ZAE ayant une allocation strictement entre 20 et 40 logements	47,6%	42,3%
ZAE ayant une allocation égale à 40 logements (borne maximum)	37,0%	43,9%

On remarque avec le calage ainsi une hausse globale de la proportion de ZAE pour lesquelles on atteint les bornes (57,7% après calage contre 52,4% avant calage) liée à une dispersion du poids des ZAE plus grande après calage[36].

Ensuite, il convient alors de vérifier que cela n'entraîne pas une dispersion trop grande des poids.

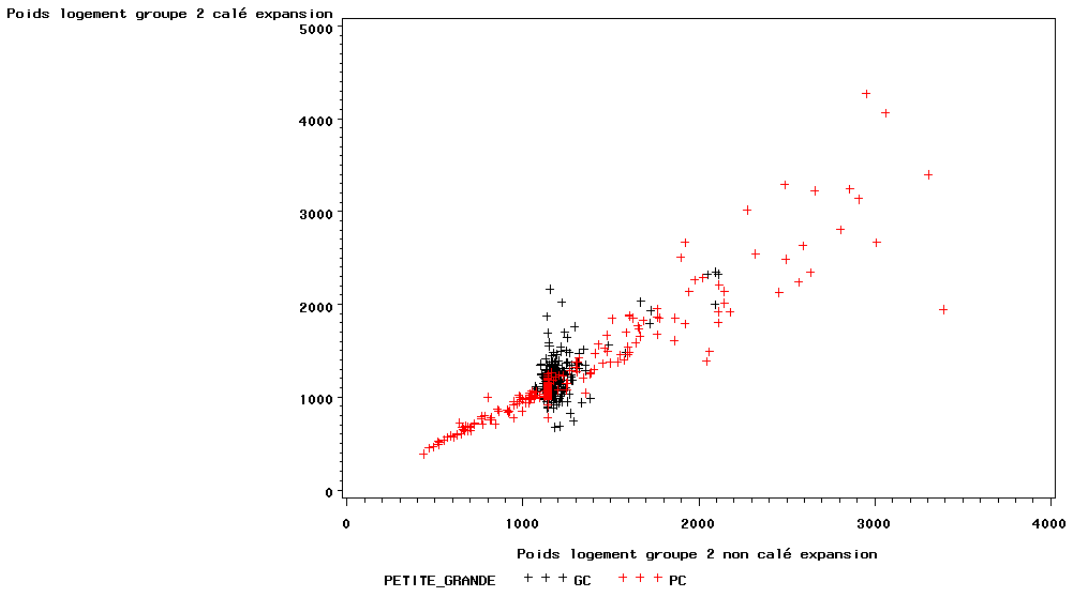
Pour chaque groupe de rotation, on a simulé le tirage d'un échantillon de 20 000 logements sans calage des ZAE puis avec calage des ZAE, et calculé la fonction de répartition du poids des 20 000 logements tirés dans les deux scénarios. On peut calculer les rapports inter-fractiles des fonctions de répartition pour comparer la dispersion des poids dans les deux scénarios. On obtient alors les résultats suivants :

[36] La hausse de la dispersion du poids des ZAE liée au calage reste cependant limitée, puisque l'année la plus défavorable, la variance du poids calé des ZAE non exhaustives (avec les variables de calage RP 1999) est supérieure de 15% à la variance du poids non calé des ZAE non exhaustives.

Pondération	Groupe	D90/D10	D95/D5	D97,5/D2,5
Scénario « ZAE non calées »	1	1,6	2,3	3,4
	2	1,4	2,1	3,5
	3	1,5	2,1	3,1
	4	1,5	2,4	3,7
	5	1,5	2,3	3,4
Scénario « ZAE calées »	1	1,7	2,3	3,2
	2	1,7	2,3	3,5
	3	1,5	2,0	3,1
	4	1,6	2,7	4,0
	5	1,6	2,4	3,5

On constate que les rapports interfractiles varient peu d'un scénario à l'autre, ce qui indique que le calage des ZAE n'augmente pas significativement la dispersion des poids des logements.

Par ailleurs, la stabilité de la fonction de répartition des poids des logements avant et après calage vient d'une bonne stabilité des poids des logements au niveau de chaque ZAE[37] (à l'exception de quelques ZAEGC pour lesquelles le calage entraîne une hausse significative du poids des logements). Le graphique ci-dessous, qui représente le poids d'un logement avant et après calage pour chacune des ZAE EM, donne un nuage de points proche de la bissectrice y=x (signe que le poids reste très proche avant et après calage).



[37] On rappelle qu'au sein d'une ZAE tous les logements ont le même poids

Conclusion :

On a donc mis en place une méthodologie innovante de calage des unités primaires. Une validation empirique de cette méthode a été opérée sur la base de l'échantillon de ZAE tiré, en vérifiant que le calage des ZAE permet de réduire à zéro l'erreur relative sur les variables de calage tout en maintenant ou en améliorant la représentativité des bases de sondage annuelles pour des variables indépendante des variables de calage et sans augmenter non plus de manière significative la dispersion du poids de sondage des logements tirés.

En outre, la méthode permettra également d'incorporer une ***information récente*** en calant sur les données du nouveau RP, disponibles en mars 2009 (alors que la construction et le tirage des ZAE ont utilisé des données du RP 1999).

Dans le fonctionnement courant d'OCTOPUSSE, les nouveaux poids calés des ZAE (recalculés chaque année pour s'ajuster à la campagne annuelle dans laquelle seront puisés les échantillons***) sont réinjectés dans le calcul des allocations de logements par ZAE, lesquelles*** dépendent donc de ces nouveaux poids aléatoires ***et non plus directement des poids initiaux des ZAE.***

7. Retour sur un choix fondamental : tirer dans la dernière campagne ou dans un cycle complet de cinq campagnes

Un des avantages du nouveau RP mis en avant étant de pouvoir disposer chaque année d'une base de sondage annuelle « fraîche » pour les enquêtes ménages, le projet OCTOPUSSE a retenu dès l'origine le principe d'un tirage des échantillons d'enquêtes dans la dernière campagne RP (cf. partie 2).

Cependant, il est indispensable d'obtenir une base de sondage annuelle « représentative » du territoire métropolitain lors de chacune des cinq années du cycle RP pour assurer la qualité des échantillons tirés dans OCTOPUSSE.

A titre dérogatoire au principe de tirage dans la dernière campagne, il est néanmoins envisagé certaines enquêtes, comme par exemple l'enquête Victimation, d'impacter cinq campagnes de Recensement en petites communes : il s'agit en effet d'enquêtes réalisées à un rythme annuel et visant avant tout à mesurer des variations annuelles ou infra-annuelles. On étudie ainsi notamment l'hypothèse de tirer les échantillons des enquêtes ménages dans cinq campagnes en ZAEPC afin d'éviter qu'un « biais de rotation » lié au changement de communes composant la base de sondage annuelle en ZAEPC ne vienne perturber l'estimation des évolutions annuelles ou infra-annuelles mesurées.

Plus généralement, les variations observées sur les résultats des premières Enquêtes Annuelles de Recensement (EAR) du RP, notamment au niveau des estimations de population communale en grandes communes, ont conduit à étudier l'importance des fluctuations des EAR en grandes communes et de leur impact sur la stabilité de la représentativité des bases de sondage annuelles, ainsi que sur l'opportunité éventuelle d'un tirage dans cinq campagnes, aussi bien en petites communes qu'en grandes communes, en réponse à ces fluctuations.

On présente ici le résultat des travaux menés sur le sujet.

- La première sous-partie étudie l'importance des fluctuations des EAR, importantes au niveau communal mais plus limitées lorsqu'on se place au niveau de l'ensemble des grandes communes ou de l'ensemble des communes, ce qui conduit à relativiser l'impact des fluctuations des EAR sur la représentativité des bases de sondage annuelles.

- La seconde sous-partie aborde plus en détail le tirage dans cinq campagnes en grandes communes et les difficultés importantes qu'il entraîne, qui conduisent à ne pas retenir cette alternative.

- La troisième partie revient, elle, sur les conséquences du tirage dans cinq campagnes en ZAEPC (envisagé par exemple pour l'enquête Victimation), notamment en termes de couverture des logements neufs et de disjonction, qui amènent à envisager de limiter au maximum son usage.

7.1. Fluctuations observées sur les EAR

7.1.1. Des fluctuations relativement importantes localement en grandes communes

L'analyse des résultats des premières Enquêtes Annuelles de Recensement a montré que les estimations de populations communales effectuées à partir de la dernière campagne de recensement[38] pouvaient connaître des fluctuations relativement importantes dans certaines grandes communes.

Compte-tenu du plan de sondage RP en grandes communes (détaillé en annexe A), on peut envisager plusieurs causes possibles au déséquilibre des groupes de rotation :

- La répartition initiale des grandes adresses, qui conduit à privilégier les trois premiers groupes de rotation pour l'affectation des grandes adresses (le groupe contenant le plus grand nombre de logements en grandes adresses, en moyenne sur l'ensemble des grandes communes, étant le groupe 3). Plusieurs opérations de mise à jour des grandes adresses ayant eu lieu depuis 2003, la répartition actuelle des grandes adresses ne correspond pas à la répartition initiale. On retrouve néanmoins ce déséquilibre, notamment au niveau des communes ayant un nombre réduit de grandes adresses, comme le montre le tableau en annexe D.

- Un équilibrage à l'IRIS pas toujours opérant (compte tenu du nombre relativement important de variables d'équilibrage)

- Absence d'équilibrage sur certaines variables importantes en termes de tirage d'échantillons (logements ZUS par exemple).

Ainsi, ces facteurs peuvent expliquer (en partie) des fluctuations importantes au niveau local des EAR.

Il reste cependant à mesurer l'impact global de ces fluctuations sur les estimations de différentes variables socio-démographiques au niveau de l'ensemble des grandes communes.

7.2.2. Analyse des fluctuations observées sur l'ensemble des grandes communes et au niveau France entière

L'importance des variations observées au niveau communal doit être nuancée par le fait qu'OCTOPUSSE ne vise pas à mesurer des variables sur une commune en particulier mais sur l'ensemble du territoire.

L'analyse a porté sur les évolutions annuelles de grandeurs socio-démographiques sur les quatre enquêtes annuelles de recensement, mesurées comme la somme des valeurs observées sur les logements recensés dans l'EAR, pondérées par le poids EAR[39] du logement calculé par le RP, d'une part sur le champ des grandes communes et d'autre part sur l'ensemble des communes.

[38] Il est rappelé que les populations légales des grandes communes diffusées par l'INSEE s'appuient non pas sur une seule Enquête Annuelle de Recensement, mais, de manière beaucoup plus robuste, sur un cycle complet de cinq campagnes annuelles consécutives de recensement.

[39] Poids d'extrapolation du logement pour les résultats de l'Enquête Annuelle de Recensement.

Sur l'ensemble des grandes communes, les fluctuations apparaissent beaucoup moins importantes, comparables à celles observées sur l'ensemble des communes (incluant les petites communes) d'une EAR sur l'autre. A titre d'exemple, ci-joint les évolutions annuelles des résultats de l'EAR observées sur la variable « population par tranches d'âge », sur l'ensemble des communes (grandes et petites), d'une part, et sur l'ensemble des grandes communes, d'autre part.

	Variation EAR 05/04 France métropolitaine	Variation EAR 05/04 ensemble des GC	Variation EAR 06/05 France métropolitaine	Variation EAR 06/05 ensemble des GC	Variation EAR 07/06 France métropolitaine	Variation EAR 07/06 ensemble des GC
De 0 à 19 ans	0,9%	1,1%	-0,3%	-0,2%	-0,2%	0,0%
De 20 à 59 ans	0,9%	1,1%	0,2%	0,5%	-0,2%	0,0%
60 ans ou plus	1,3%	1,2%	0,6%	0,9%	2,8%	2,8%

La variabilité des résultats en grandes communes entre les différentes EAR n'apparaît pas significativement plus forte que la variabilité des résultats sur l'ensemble des communes, malgré un taux de sondage plus réduit (8% des logements en grandes communes contre 20% des logements en petites communes, soit environ 14% des logements sur l'ensemble des communes).

Cela a conduit à ne pas retenir l'idée d'une variabilité « spécifique » qui serait observée sur les grandes communes compte-tenu du taux de sondage plus faible qu'en petites communes.

7.2. Peut-on prendre en compte la variabilité des EAR en grandes communes par un tirage dans cinq campagnes en grandes communes?

Une possibilité évoquée pour traiter le problème du déséquilibre des grandes adresses est le tirage simultané dans cinq campagnes en grandes communes. En effet, on dispose à première vue à chaque instant d'une base « exhaustive » des grandes adresses, ce qui permet d'éviter le problème de déficit de logements en grandes adresses dans les groupes 4 et 5 dans les petites « grandes communes ».

Cependant, le tirage simultané dans cinq campagnes pose les questions méthodologiques suivantes :

7.2.1. Repondération des logements sur la base d'un RIL moyen pour tenir compte de l'évolution du bâti pendant les cinq années du cycle.

Exemple : nombre total de logements de la Base de Sondage Annuelle dans la commune d'Ambérieu -en-Bugey (01004)

Campagne RP	2004	2005	2006	2007	2008
Nombre logements RIL	5451	5768	5768	5942	5950

On s'aperçoit ainsi que, si on souhaite ramener les logements à un poids comparable (poids estimé pour le point moyen du cycle, c'est-à-dire au 1er janvier 2006), il est nécessaire de repondérer les logements. Dans le cas de la commune d'Ambérieu-en-Bugey, il faut alors augmenter en moyenne de 5% le poids des logements tirés dans la campagne 2004, tandis qu'il faut réduire en moyenne de 3% celui des logements tirés dans les campagnes 2007 et 2008 (en revanche, celui des logements tirés dans la campagne 2005 reste stable en moyenne compte tenu de la stabilité du RIL).

On serait alors conduit à adopter une démarche de pondération complexe au niveau de chacune des trois strates d'adresses (grandes adresses, adresses neuves et autres adresses), analogue à celle employée pour le calcul des populations légales (gestion et mise à jour d'un RIL médian…).

Dans la pratique, la repondération peut cependant s'avérer parfois difficile dans les cas de disparition d'adresses, notamment dans la strate des grandes adresses (dégroupement de l'adresse suite à un retour de collecte ou à une mise à jour administrative, recalcul à la hausse du seuil des grandes adresses si elles regroupent plus de 10% des logements de la commune, destruction des logements…).

Ainsi, pour la commune d'Ambérieu-en-Bugey, on observe l'évolution suivante de la composition du groupe 1 :

Nombre de logements du groupe 1	RIL campagne 2004	RIL campagne 2006
Grandes Adresses	86	0
Adresses Neuves	60	87
Autres Adresses	944	1080
Total	**1090**	**1167**

On voit ici que tous les logements en grandes adresses recensés dans le groupe 1 en 2004 (86 logements) ne font plus partie de la strate des grandes adresses en 2006 (puisque la strate des grandes adresses est vide dans le groupe 1 pour l'année 2006). Dans ce cas, il n'est pas possible de calculer une pondération des logements en grandes adresses recensés en 2004 selon les règles générales mais il faut adopter une règle ad hoc (regroupement de strates par exemple) pour leur attribuer un poids « calé 2006 ».

7.2.2. Hétérogénéité de la définition des strates

La question du tirage simultané dans cinq campagnes pose également la question de la variabilité de la définition des strates au cours du temps, notamment sur les campagnes du premier cycle RP. En particulier, la strate des grandes adresses (outre des mises à jour automatiques du seuil pour respecter la limite de 10% des logements en grandes adresses) a fait l'objet d'une mise à jour importante au deuxième trimestre 2006 qui a commencé à être effective à partir de la collecte 2007[40] pour prendre en compte le fait que la strate des grandes adresses s'était vidée dans plusieurs grandes communes : cette mise à jour s'est traduite en pratique par une modification du seuil des grandes adresses (nombre de logements à partir duquel l'adresse est affectée à la strate des grandes adresses).

Le tableau ci-dessous (extrait de la note d'Aude Mulliez 1491/F520 du 10 mai 2006) récapitule les mises à jour effectuées dans certaines communes

[40] Ainsi, en 2004, le nombre total d'adresses de grande taille était de 9 844, représentant 1 031 475 logements, et, entre 2004 et 2006, 957 adresses représentant 93 240 logements sont apparues, tandis que 1 839 adresses représentant 202 831 logements ont disparu.

DR de collecte	DEPCOM	NOM	nb log	nb log GA	seuil GA actuel	seuil GA recalculé	nb adr GA ajoutées	nb log GA ajoutés	part des nouvelles GA dans la commune
DR34	66037	CANET-EN-ROUSSILLON	14 751	0	187	**66**	**16**	**1 387**	**9,4%**
DR20	2A004	AJACCIO	30 611	163	121	**63**	**35**	**2 838**	**9,3%**
DR51	77284	MEAUX	20 823	539	177	**139**	**9**	**1 503**	**7,2%**
DR69	69081	ECULLY	7 638	226	171	**83**	**4**	**534**	**7,0%**
DR45	91521	RIS-ORANGIS	11 122	256	128	**127**	**6**	**762**	**6,9%**
DR34	34172	MONTPELLIER	139 202	4 688	113	**77**	**101**	**9 167**	**6,6%**
DR34	34003	AGDE	42 616	1 677	180	**124**	**17**	**2 477**	**5,8%**
DR20	2B033	BASTIA	19 015	795	129	**81**	**10**	**1 045**	**5,5%**
DR13	13215	MARSEILLE 15E	29 327	1 320	101	**62**	**19**	**1 601**	**5,5%**
DR34	34108	FRONTIGNAN	12 245	137	95	**60**	**9**	**665**	**5,4%**
DR33	64024	ANGLET	21 473	979	90	**60**	**16**	**1 162**	**5,4%**
DR33	33069	BOUSCAT	11 932	505	147	**94**	**5**	**645**	**5,4%**
DR78	93014	CLICHY-SOUS-BOIS	9 427	365	85	**83**	**6**	**503**	**5,3%**
DR20	2A247	PORTO-VECCHIO	8 069	263	251	**199**	**2**	**425**	**5,3%**

On voit ainsi par exemple que, pour la commune de Canet-en-Roussillon, le seuil des grandes adresses a été descendu de 187 à 66 logements entre les campagnes de recensement 2006 et 2007.

Ainsi, pour cette commune, la « base exhaustive » des grandes adresses constituée à partir des cinq premières campagnes du cycle (2004-2008) pour les premiers tirages à partir de mai 2009 aurait en fait contenu les logements suivants :

- campagnes 2004 à 2006 : ensemble des logements des adresses de plus de 187 logements appartenant au groupe de rotation recensé
- campagnes 2007 et 2008 : ensemble des logements des adresses de plus de 66 logements appartenant au groupe de rotation recensé.

Il semble donc difficile, dans ces conditions, de raisonner sur une strate des grandes adresses homogène, au moins tant que l'effet lié à la mise à jour de la strate des grandes adresses persiste (c'est-à-dire jusqu'au chargement de la campagne 2011 en novembre 2011).

Une couverture « disparate » des grandes adresses pose alors des problèmes importants en termes de représentativité des échantillons tirés compte-tenu du caractère spécifique des ménages des grandes adresses en termes de niveau de vie, d'emploi…notamment dans les adresses de très grande taille.

7.2.3. Nécessité de fractionner l'allocation en cinq sous-allocations

Un tirage global de l'allocation (dont le mode de calcul serait à définir) dans la « base complète » formée des cinq campagnes ne semble pas possible en raison du taux de ponction différent des campagnes (du fait du renouvellement par cinquième de la base de sondage chaque année) selon la date à laquelle on se place.

Par exemple, si la campagne 2009 vient d'être chargée, aucun logement n'a encore été tiré dans la campagne 2009 alors que 200 logements ont déjà été tirés dans la campagne 2005[41] (soit les deux tiers des logements chargés pour une petite « grande commune » de 5000 résidences principales dans laquelle la base de sondage OCTOPUSSE contient environ 300 logements) : il n'est donc pas possible d'effectuer directement un tirage systématique à probabilités égales dans « la base

[41] Sur la base d'un tirage de 250 logements par ZAE chaque année conduisant à tirer chaque année 50 logements dans chacune des cinq campagnes actives. Ainsi, après le chargement de la campagne de l'année N, aucun logement n'a encore été tiré dans la campagne N alors que 50x4=200 logements ont déjà été tirés dans la campagne N-4 qui a déjà été active pendant quatre ans.

complète » formée des cinq campagnes. Comme c'est le cas pour les tirages dans cinq campagnes dans les ZAEPC, il est alors nécessaire de fractionner l'allocation en cinq sous-allocations tirées dans chacun des cinq campagnes.

Cela pose alors le problème de la représentativité des échantillons tirés dans chacune des cinq bases annuelles de logements de la commune, qui rend plus difficile une bonne couverture géographique de tous les IRIS de la commune.

Par exemple, dans le cas d'une commune de 20 000 résidences principales (comptant 10 IRIS) dans laquelle on tire un échantillon de 40 logements, on peut garantir, dans le cas d'un tirage dans la dernière campagne, sans surreprésentation (on trie alors les logements par identifiant RP, donc par IRIS, comme premier critère de tri), qu'on aura 4 logements dans chacun des dix IRIS. Dans l'hypothèse d'un tirage dans cinq campagnes en grande commune, on tire huit logements dans chacune des cinq bases annuelles, ce qui présente le risque d'avoir au final des IRIS pas (ou peu) représentés dans l'échantillon : on ne maîtrise plus en effet la stratification implicite par IRIS.

D'autre part, compte-tenu du processus de rééchantillonnage, il n'est pas garanti qu'il reste encore des logements en grandes adresses pour les tirages dans les groupes déficitaires.

Par exemple, dans le cas de la commune de Pamiers qui compte 140 logements en grandes adresses tous situés dans le groupe 1 (sur un total de 7216 logements), on ne retient à l'issue d'un rééchantillonnage communal que 8 logements en grandes adresses, ce qui est clairement insuffisant pour avoir des logements en grandes adresses dans tous les échantillons tirés.

Ainsi, le tirage dans cinq campagnes en grandes communes pose des problèmes méthodologiques importants en termes de pondération des logements. D'autre part, compte-tenu de l'hétérogénéité de la définition des strates d'adresses et de la nécessité d'effectuer cinq tirages de logements dans chacune des cinq campagnes, la représentativité des échantillons de logements tirés ne semble pas assurée. Cette solution n'a donc pas été retenue pour OCTOPUSSE.

7.3. Quel impact du tirage dans cinq campagnes en ZAEPC sur les échantillons tirés ?

La possibilité d'un tirage dans cinq campagnes en ZAEPC a été ouverte pour assurer une meilleure représentativité des bases de sondage. Cependant cette méthode entraîne des inconvénients importants en termes de défaut de couverture de la construction neuve et de disjonction, qui conduisent à restreindre fortement son usage, compte-tenu par ailleurs de la possibilité d'assurer une représentativité stable des bases de sondage annuelles OCTOPUSSE grâce au calage des ZAE (cf. partie 6).

7.3.1. Un défaut de couverture important des logements neufs qui conduit à préférer en règle générale un tirage dans la dernière campagne

Une des difficultés méthodologiques liées au tirage dans cinq campagnes en petites communes est le défaut de couverture des logements neufs dans celles-ci.

En effet, en novembre de l'année N, la base de sondage en ZAE Petites Communes est, dans ce cas, constituée de l'agrégation de cinq campagnes de recensement des années N, N-1, N-2 , N-3 et N-4. Environ 20% de l'allocation totale tirée en petites communes est donc tirée dans chacune des cinq campagnes[42], ce qui entraîne un défaut de couverture de 80% des logements construits l'année N-1.

[42] Compte-tenu du fait que les ZAE ont été tirées avec un équilibrage sur le nombre de résidences principales au niveau groupe de rotation et que l'allocation totale à tirer dans la ZAEPC est ventilée entre les cinq fractions proportionnellement à la taille de la fraction en termes de nombre de résidences principales (avec généralement une part importante de l'allocation de la ZAE tirée dans la commune pivot compte tenu de sa taille).

Ainsi, 20% des échantillons sont tirés dans une base ne contenant pas les logements construits à partir de janvier N-4, 20% dans une base ne comportant pas les logements construits après janvier N-3, 20% dans une base sans logements après janvier N-2, 20% dans une base après janvier N-1 et 20% dans une base après janvier N.

Ainsi, en moyenne, le défaut de couverture moyen est de 80% pour les logements neufs construits l'année N-1, de 60% pour ceux de l'année N-2, de 40% pour ceux de l'année N-3 et de 20% pour ceux de l'année N-4.

Cette situation est donc susceptible d'entraîner un biais d'échantillonnage qui peut se révéler problématique, notamment pour les enquêtes pour lesquelles les logements neufs correspondent à des comportements spécifiques (Patrimoine, Emploi du Temps…).

Par ailleurs, le principe du calage des ZAE, qui permet d'assurer une représentativité stable des bases de sondage annuelle, atténue le gain en termes de « représentativité » lié à un tirage dans cinq campagnes en ZAEPC[43].

Il est donc proposé de limiter au maximum le tirage dans cinq campagnes en ZAEPC, en ne l'envisageant que pour des enquêtes de périodicité annuelle ou infra-annuelle pour lesquelles le biais de rotation lié au changement de base annuelle est susceptible de créer des fluctuations supplémentaires dans l'observation des évolutions.

Ainsi, il a été décidé que l'ensemble des enquêtes apériodiques ou effectuées tous les cinq ou six ans (Emploi du Temps, Patrimoine, IVQ, Logement, FQP…) seront tirées dans la dernière campagne en ZAEPC. Il en est de même pour les nouveaux entrants des échantillons de panels.

Remarque : dans le cas des panels avec interrogations sur plusieurs années, les cinq groupes de rotation des ZAEPC finissent par être impactés compte-tenu de l'arrivée de nouveaux entrants chacune des cinq années du cycle.

7.3.2. Cas particulier de l'enquête Victimation

La seule enquête pour laquelle un tirage dans cinq campagnes en ZAEPC n'est pas exclu est l'enquête Victimation. En effet, il s'agit d'une enquête effectuée chaque année et sans réinterrogation des logements (pas de dimension de panel), ce qui fait que le « biais de rotation » lié au changement de base de sondage annuelle OCTOPUSSE serait maximal dans le cas d'un tirage dans la dernière campagne.

En pratique, compte tenu du caractère local des phénomènes mesurés, l'impact de la rotation des fractions recensées des ZAEPC pourrait être important, dans le cas par exemple d'une ZAE où la fraction correspondant au groupe 1 contient une ZUS et celle correspondant au groupe 2 n'en contient pas. Le tirage de l'échantillon dans cinq campagnes pourrait alors permettre d'éviter ces fluctuations des bases de sondage si le calage des ZAE (effectué sur les variables de segmentation urbain/rural en tranches d'unités urbaines et en aires urbaines, de structure par âge, d'emploi par secteur et de revenu fiscal total) ne permettait pas d'assurer une stabilité suffisante de la base de sondage pour les phénomènes étudiés (criminalité…).

Le gain lié à la stabilité des petites communes impactées par l'enquête devra être mis en balance avec le défaut de couverture des logements neufs évoqué au paragraphe précédent.

Par ailleurs, le tirage dans cinq campagnes pose un problème spécifique **en matière de disjonction**. En effet, compte tenu de l'absence d'identifiant pérenne entre deux campagnes de recensement successives, il n'est pas possible de repérer, parmi les logements tirés en novembre de l'année N dans la campagne de l'année N, ceux qui ont été tirés l'année précédente (N-1) au titre de la campagne N-5.

[43] Le calage des ZAE s'appliquera aussi au tirage dans cinq campagnes : dans ce cas, on recalcule un poids appliqué à l'ensemble de la ZAE.

Par exemple, en novembre 2014, la campagne annuelle du groupe 1 sera chargée. Or des tirages dans cinq campagnes impactant les campagnes 2009 à 2013 auront été faits au cours de l'année 2013, sans qu'on puisse savoir, pour les logements chargés au titre de la campagne de recensement 2014, quels sont ceux qui ont été tirés l'année précédente dans la campagne 2009.

Il apparaît donc nécessaire de quantifier plus en détail l'impact sur la disjonction d'un tirage dans cinq campagnes de l'enquête Victimation.

Un premier essai de quantification en termes de disjonction peut être effectué sur la base des hypothèses suivantes (détaillées en annexe B)[44] : 180 FA tirées chaque année dans chaque ZAEPC, réparties entre 35 FA tirées dans cinq campagnes pour l'enquête Victimation (soit 7 FA dans chacune des cinq campagnes) et 145 FA tirées dans la dernière campagne pour les autres enquêtes.

On suppose pour simplifier que tous les tirages d'enquêtes de l'année à venir ont lieu en novembre après le chargement de la campagne RP. En novembre de l'année N, on tirera alors :

- 145+7=152 logements dans la campagne de l'année N
- 7 logements dans chacune des quatre campagnes antérieures N-4 à N-1

Année de tirage Campagne RP	2010	2011	2012	2013	2014
2006	7				
2007	7	7			
2008	7	7	7		
2009	7	7	7	7	
2010	152	7	7	7	7
2011		152	7	7	7
2012			152	7	7
2013				152	7
2014					152

Méthode de calcul du nombre de ménages déjà interrogés il y a moins de quatre ans

En 2014, on tire donc 152 logements dans le groupe 1 (campagne 2014) et 7 logements dans chacun des quatre autres groupes. Dans le cas le plus défavorable d'une ZAE composée de 300 logements dans chacun des cinq groupes de rotation, on a alors l'impact suivant sur la disjonction :

- tirage dans le groupe 1 : 152 logements sont tirés dans la campagne 2014. En moyenne, on a dans la campagne 2014 2,3% (7/300) des logements tirés en 2013 (dans la campagne 2009) et de même 2,3% des logements tirés en 2012, 2011 et 2010. Donc parmi les 152 logements tirés dans la campagne 2014, 3,5 en moyenne (152x2,3%) ont déjà été tirés il y a un an, 3,5 il y a deux ans et de même 3,5 il y a trois et quatre ans.

- tirage dans les autres groupes : dans le groupe 5 : 7 logements sont tirés dans la campagne 2013. Ils sont disjoints avec les 152 logements tirés en 2013 dans la campagne 2013. Ils ne sont pas disjoints en revanche des 7 logements tirés en 2012, 2011, 2010 et 2009 dans la campagne 2008. Il s'agit cependant d'un recouvrement très faible, de l'ordre de 0,16 logements pour chacune des trois campagnes. On observe de même un recouvrement de 0,16 logements pour deux campagnes pour les tirages dans le groupe 4 et pour une campagne pour le groupe 3.

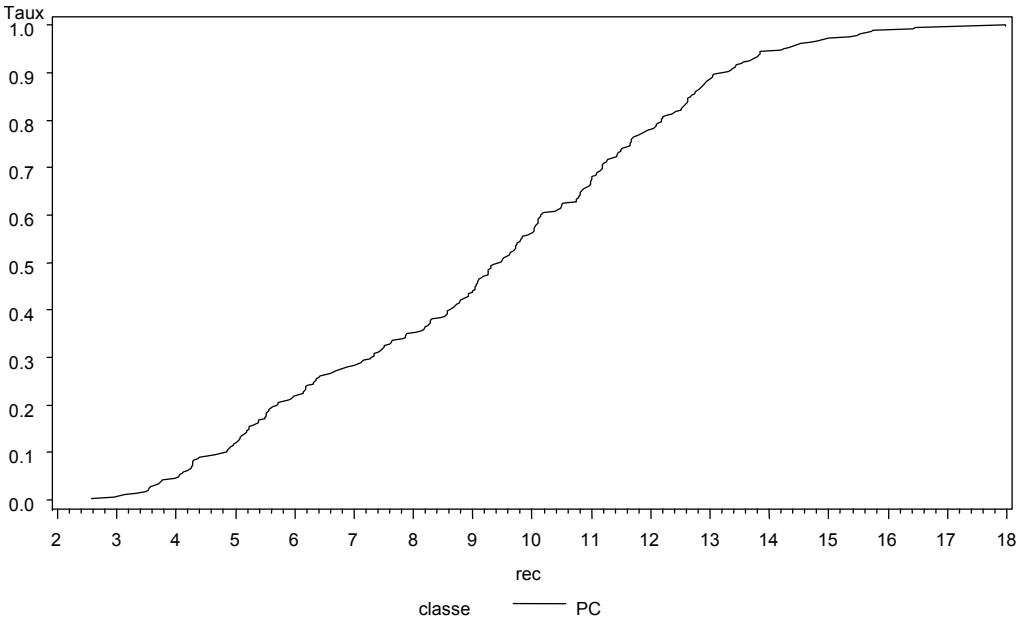
[44] Il s'agit ici uniquement d'hypothèses de travail élaborées sur la base des enquêtes passées, qui ne préjugent pas du volume d'enquêtes pour les années à venir.

Au total, pour une ZAE de taille minimale, 15 logements parmi les 180 logements tirés une année donnée ont déjà été interrogés il y a quatre ans ou moins (dont 3,5 il y a moins un an ou moins, 7,2 il y a moins de deux ans ou moins et 11,1 il y a moins de trois ans ou moins).

Il s'agit de la situation la plus défavorable. En effet, le nombre de ménages réinterrogés à moins de cinq ans d'intervalle diminue avec la taille de la ZAE (et notamment de la fraction recensée de la ZAE).

En moyenne, sur les 286 ZAEPC tirées pour l'EM, on observe sur les cinq groupes de rotation une réinterrogation de 9 logements à moins de quatre ans d'intervalle[45] , soit 5% des logements interrogés dans l'année.

La fonction de répartition du nombre de logements réinterrogés à quatre ans d'intervalle ou moins dans les 286 ZAEPC de l'EM donne les résultats suivants:



Il faudrait ainsi retirer en moyenne chaque année 9 logements des échantillons tirés pour ne pas réinterroger des ménages déjà interrogés dans l'enquête Victimation il y a quatre ans ou moins, ce nombre pouvant aller au maximum jusqu'à 18 comme le montre le graphique (cas particuliers de certaines ZAE ayant moins de 300 logements dans un groupe de rotation).

Ainsi, le tirage dans cinq campagnes en ZAEPC, même s'il apparaît répondre à la question de la stabilité des bases de sondage d'une année sur l'autre, pose des problèmes en termes de couverture des logements neufs et de disjonction.

8. Le cas particulier des échantillons ZUS

8.1. Le tirage d'échantillons ZUS dans le contexte du RP 1999

Au début des années 2000, le besoin d'information « spécifique » sur les populations des Zones Urbaines Sensibles s'est fortement accru, notamment pour des enquêtes visant à mesurer les difficultés « socio-économiques » des Français (logement, illettrisme, victimation). Pour répondre à cette demande nouvelle d'échantillons (apparue après la constitution de l'EM 99), l'UMS a constitué en 2004 une base ZUS spécifique regroupant l'ensemble des logements ZUS au RP 1999, soit 1 842

[45] On se limite ici au recouvrement dû aux 152 logements tirés dans la dernière campagne en négligeant celui lié aux 7 logements tirés dans les campagnes antérieures, inférieur à un logement et proche de zéro dans la très grande majorité des cas.

744 logements (dont 1 672 520 résidences principales), dont 98,2% sont susceptibles d'être tirés au sein d'une enquête ménage « standard »[46]. Actuellement cette base est utilisée de façon régulière pour le tirage d'échantillons spécifiques en ZUS, dans le cadre d'enquêtes à périodicité annuelle (enquêtes Victimation) ou pluriannuelles (enquête Logement, enquête IVQ).

Le zonage ZUS

Les ZUS (Zones Urbaines Sensibles) sont des territoires infra-communaux définis par les pouvoirs publics (loi du 14 novembre 1996) pour être la cible prioritaire de la politique de la ville, en fonction des difficultés que connaissent les habitants de ces territoires. Ces territoires sont de taille très variable suivant les communes, leur taille au RP 1999 allant de 109 logements (commune de Valmont, 57690) à 29 335 logements (commune de Roubaix, 59612).

Données sur les ZUS au RP 1999[47] (les chiffres en italique correspondent à des estimations)

Données	Grandes communes	Petites communes	Total métropole
Nombre de communes ayant une ZUS	378	111	489
Nombre de logements en ZUS	1 557 357	285 387	1 842 744
Nombre total de logements	14 144 523	14 551 476	28 695 999
Part des logements en ZUS	11,4%	2,0%	6,4%
Nombre de résidences principales en ZUS	*1 413 488*[48]	*259 032*	1 672 520
Nombre total de résidences principales	12 368 718	11 446 313	23 815 031
Part des résidences principales en ZUS	*11,0%*	*2,3%*	7,0%

Ainsi, 85% des logements en ZUS sont situés en grandes communes. En proportion, les ZUS représentent environ 7% des résidences principales sur l'ensemble de la métropole, et 11% des résidences principales des grandes communes.

Contrairement aux tirages en population générale où seuls les logements situés dans les unités primaires tirées peuvent être impactés par le tirage d'un échantillon, l'ensemble des logements recensés en ZUS peuvent être impactés par un tirage ZUS (un seul degré de tirage). Seules quelques ZUS spécifiques « non atteignables » sont exclues de la base de sondage. Cependant, une des difficultés du système est qu'il est difficile de gérer une disjonction entre l'Echantillon-Maître et la base ZUS.

Il a été décidé dans le cadre du projet OCTOPUSSE de reconduire cette fonctionnalité de tirage d'échantillons ZUS. Cependant, compte-tenu du nouveau système de recensement, il a été nécessaire de repenser la méthode de tirage des échantillons ZUS.

8.2. L'adaptation du tirage des échantillons ZUS au contexte du nouveau Recensement

Au moment du chargement de la collecte de l'année N, une base ZUS « année N » est créée en chargeant tous les logements des ZUS présents dans l'enquête annuelle RP de l'année N

[46] Soit une base utilisable de 1 809 570 logements, les 1,8% restants ayant été considérés comme « non atteignables » par les Directions Régionales de l'INSEE car trop éloignées du réseau enquêteurs ou trop dangereuses.
[47] Données calculées à partir de la table com_zus04 donnant par commune le nombre de logements en ZUS
[48] Estimation calculée en appliquant la proportion de résidences principales en ZUS (90,7%) au nombre de logements en ZUS en grandes communes

(éventuellement, tous les logements ZUS des grandes communes si on ne sait pas identifier les logements ZUS des petites communes), ***que ces logements soient situés ou non dans une ZAE tirée***.

Cependant, afin d'éviter des problèmes de disjonction avec les échantillons tirés en population générale dans la dernière campagne active, les bases de sondage annuelles ZUS entrent en service avec un an de décalage par rapport à la campagne active. Ainsi, l'activation de la campagne de Recensement de l'année N conduit à l'activation de la base annuelle ZUS de l'année N-1 (formée de tous les logements recensés en ZUS pendant la campagne de Recensement N-1). Tous les logements déjà tirés en ZUS l'année précédente dans le cadre des tirages en population générale sont marqués et ne peuvent donc plus être sélectionnés dans un échantillon ZUS. Les logements tirés dans les échantillons ZUS sont également marqués et ne peuvent donc plus être sélectionnés par la suite, que ce soit pour un autre échantillon ZUS ou pour un tirage en population générale impactant une campagne antérieure compte tenu d'un épuisement local des logements chargés pour la campagne active (cf. paragraphe 5.3).

Une premier scénario a ainsi été de tirer les échantillons ZUS dans la base annuelle ZUS de l'année N-1.

Cependant, la constitution d'une base ZUS pose également des problèmes sur le plan méthodologique. En effet, le plan de sondage du RP n'est pas équilibré sur le nombre de logements en ZUS (au niveau infra-communal, le domaine sur lequel a lieu l'équilibrage est l'IRIS, qui ne recouvre pas la notion de ZUS). Ainsi, rien de garantit que la proportion de logements en ZUS soit stable d'une collecte sur l'autre. De plus, dans les petites « grandes communes », on peut penser que les logements ZUS correspondent dans de nombreux cas aux logements en grandes adresses, qui ont été répartis prioritairement dans les trois premiers groupes.

Ainsi, il paraît difficile d'extrapoler à partir des logements ZUS de l'échantillon de l'enquête annuelle le nombre de logements ZUS présents en métropole, pour obtenir un vrai calcul des poids de sondage.

Pour éviter ce problème d' « hétérogénéité » des groupes de rotation en termes de couverture des logements ZUS, la solution retenue est celle d'une base ZUS « exhaustive sur cinq ans» formée de cinq bases ZUS « exhaustives » annuelles.

On effectue ainsi le tirage des échantillons ZUS dans les cinq bases annuelles ZUS actives, c'est-à-dire les bases annuelles de N-1 à N-5 (si la campagne active est celle de l'année N), compte-tenu du fait que les bases annuelles ZUS rentrent en service avec un an de décalage par rapport à la base de sondage principale.

Concrètement, on tire cinq sous-échantillons ZUS dans chacune des cinq bases ZUS annuelles impactées, chaque sous-échantillon étant représentatif d'un des cinq groupes de rotation : l'utilisateur rentre alors le nombre de logements à tirer dans chacune des cinq bases (la séparation en cinq bases annuelles permet de gérer le fait que les collectes anciennes ont été ponctionnées plus fortement que les collectes récentes au moment du tirage de l'échantillon).

Ainsi, chaque sous-échantillon ZUS est représentatif non pas de l'ensemble des logements ZUS mais uniquement de l'ensemble des logements ZUS du groupe de rotation dans lequel il est tiré. Son poids de sondage final dépend uniquement de la probabilité de seconde phase du Recensement (probabilité de tirage du logement pour l'Enquête Annuelle de Recensement au sein du groupe de rotation, gérée de façon explicite par le Recensement), du nombre total de logements initialement présents dans la base annuelle ZUS et de l'allocation tirée dans la base annuelle ZUS. En particulier, il n'est pas nécessaire de modéliser les comportements d'affectation des logements aux différents groupes de rotation (probabilité de première phase).

Remarque : la disjonction n'est pas gérée parfaitement, puisque le même logement ZUS recensé deux fois à cinq ans d'intervalle peut être tiré deux fois à des dates rapprochées, via un tirage dans la campagne active de l'année N et via un tirage ZUS dans la campagne N-5. La probabilité que ce cas survienne est cependant des plus faibles et a été considérée comme négligeable.

9. Le ciblage de certaines sous-populations au moyen d'un tirage en deux phases

L'objectif de la fonctionnalité « groupes de seconde phase » d'OCTOPUSSE est de permettre le tirage d'échantillons avec une surreprésentation de certaines catégories de logements (définies à partir des variables RP chargées dans OCTOPUSSE sur les logements et les individus qui y sont rattachés) : par exemple, surreprésentation des ménages dont le chef de famille est peu diplômé pour l'enquête IVQ 2004.

Un groupe de seconde phase est défini par un ensemble de conditions logiques sur des variables RP disponible dans OCTOPUSSE (par exemple « âge du chef de ménage >= 50). Les groupes de seconde phase doivent être construits de sorte à ce que chaque logement de la base de sondage soit rattaché à un unique groupe de seconde phase. Les groupes de seconde phase doivent donc former une partition de la base de sondage[49].

Ce système de « groupe de seconde phase » est déjà mis en œuvre au niveau de la chaîne EM 99. Un échantillon de première phase est tiré à probabilité égale dans l'EM 99 (tous les logements de l'échantillon de première phase sont marqués afin de ne pas déformer la base de sondage). Au sein de cet échantillon de première phase, chaque logement est rattaché à un groupe de surreprésentation : en seconde phase, un échantillon de logements est tiré au sein de chaque groupe pour constituer l'échantillon de seconde phase, le nombre de logements à conserver étant déterminer par la taille totale de l'échantillon de première phase, la proportion de logements du groupe présents dans l'échantillon de première phase et le coefficient de surreprésentation associé au groupe.

Un point important est la taille de l'échantillon de première phase à tirer (calcul du « N de première phase »): l'objectif est alors de déterminer la taille minimale de l'échantillon de première phase à tirer pour avoir dans chaque groupe au niveau de l'échantillon de première phase un nombre de logements supérieur ou égal au nombre de logements à tirer dans le groupe pour l'échantillon de seconde phase.

Pour ce faire, il est nécessaire de connaître la proportion de chacun des groupes dans la population totale, alors qu'on ne dispose dans les Echantillons-Maîtres que de la partie des logements recensés constituant la base de sondage de l'Echantillon-Maître : c'est sur la base des seuls logements chargés dans l'Echantillon-Maître que se fait le calcul de la proportion dans la population des différents groupes de seconde phase.

Dans le cas de l'EM 99, cela ne pose de problème dans la mesure où les logements chargés sont issus d'un recensement exhaustif effectué de manière homogène sur tout le territoire.

Dans le cas d'OCTOPUSSE, les logements chargés sont issus d'un sondage pour les grandes communes et d'un recensement exhaustif pour les petites communes. Comme on charge 8% des logements des grandes communes tirées et l'ensemble des logements des petites communes du groupe de rotation des ZAE tirées, il semble difficile d'effectuer une analyse globale sur l'ensemble de la base de logements.

Il est donc proposé de distinguer pour chaque groupe de seconde phase la proportion en grandes communes $\widetilde{p}_{Gi,GC}$ et la proportion en petites communes $\widetilde{p}_{Gi,PC}$, la proportion sur la population $\widetilde{p}_{Gi}$ étant alors estimée comme la moyenne des proportions en grandes communes et en petites communes pondérée par l'effectif des grandes communes et des petites communes (en fonction des dernières données des populations légales).

On calcule alors le nombre de logements estimés dans l'échantillon final pour chaque groupe de seconde phase avec la formule :

[49] On appelle partition de seconde phase un ensemble de groupe de seconde phase formant une partition de la base de sondage : il est nécessaire de définir une partition de seconde phase pour pouvoir lancer un tirage en deux phases. Le cas de groupes non représentés dans l'échantillon final est géré en affectant au groupe un coefficient de sur-représentation égal à 0.

$$\hat{N}_{Gi} = N_{TOT} \times \frac{\tilde{p}_{Gi} \times \alpha_{Gi}}{\sum_{j=1}^{J} (\tilde{p}_{Gj} \times \alpha_{Gj})}$$

où N_{TOT} est la taille totale de l'échantillon à tirer, $\tilde{p}_{Gi}$ la proportion estimée du groupe i et α_{Gi} le coefficient de sur-représentation du groupe i.

Pour chaque groupe de seconde phase, on réalise alors une estimation de la taille de l'échantillon de première phase à tirer (taille minimum pour avoir un nombre de logement dans le groupe supérieur ou égal au nombre de logements attendus dans l'échantillon final) pour obtenir en moyenne dans l'échantillon de première phase un nombre de logement appartenant au groupe de seconde phase suffisant (au moins égal au nombre de logements du groupe de seconde phase de l'échantillon final).

- Une estimation de la taille minimale basée sur la proportion moyenne dans la base :

$$N_{PMOY,Gi} = \frac{\hat{N}_{Gi}}{\tilde{p}_{Gi}}$$

- Une estimation basée sur la proportion minimum du groupe de seconde phase i entre la proportion observée en grande commune et celle observée en petites communes :

En notant $p\min_{Gi} = \min(\tilde{p}_{Gi,GC}, \tilde{p}_{Gi,PC})$ on obtient alors :

$$N_{PMIN,Gi} = \frac{\hat{N}_{Gi}}{p\min_{Gi}}$$

La taille de l'échantillon de première phase à tirer est alors calculée comme le maximum des tailles d'échantillon de première phase à tirer pour obtenir le nombre de logements nécessaire dans un groupe de seconde phase donné.

On obtient alors :

- Une estimation de la taille minimale de l'échantillon de première phase basée sur les proportions moyennes dans les groupes de seconde phase :

$$N_{PMOY} = \max_i (N_{PMOY,Gi})$$

- Une estimation de la taille minimale de l'échantillon de première phase basée sur les proportions « minimum » dans les groupes de seconde phase :

$$N_{PMIN} = \max_i (N_{PMIN,Gi})$$

Sauf problème d'épuisement de la campagne en cours, il sera ainsi plus sûr de tirer un échantillon de première phase de taille N_{PMIN} afin de prendre une marge de sécurité par rapport à la taille minimale plancher N_{PMOY} nécessaire pour garantir qu'on aura en moyenne dans l'échantillon de première phase un nombre suffisant de logements dans chacun des groupes de seconde phase.

On tire donc un échantillon de première phase, de taille N_{1P}. Notons $\hat{p}_{Gi}^{1P} = \frac{\sum_{j \in S1 \cap Gi} w_j^{1P}}{\sum_{j \in S1} w_j^{1P}}$ la proportion estimée du groupe i dans la population à partir de l'échantillon de première phase.

Comme on souhaite obtenir au final un échantillon de taille N_{TOT}, on tire dans chaque groupe de seconde phase un échantillon de taille :

$$N_{2P,GRi} = N_{TOT} \times \frac{\hat{p}_{Gi}^{1P} \times \alpha_{Gi}}{\sum_{j=1}^{J} (\hat{p}_{Gj}^{1P} \times \alpha_{Gj})}$$

Remarque : il sera également possible de mobiliser des « vraies proportions » issues par exemple de l'analyse « hors système » des données de l'ensemble d'une campagne RP ou d'un cycle de cinq campagnes. Dans ce cas, l'expert sondage pourra introduire directement le nombre de logements à conserver dans chaque groupe de seconde phase. OCTOPUSSE fournira alors une estimation de la taille de l'échantillon de première phase à tirer (basée sur la proportion moyenne ou « minimale » de chacun des groupes de rotation), puis effectuera le tirage de l'échantillon de première phase et le tirage de l'échantillon de seconde phase en gardant pour l'échantillon de seconde phase le nombre de logements spécifié dans chaque groupe de seconde phase.

Poids d'un logement tiré dans un échantillon de seconde phase :

Le poids d'un logement tiré dans un échantillon de seconde phase $w\log_{2P}$ appartenant au groupe de seconde phase i est calculé à partir du poids du logement dans l'échantillon de première phase $w\log_{1P}$ selon la formule :

$$w\log_{2P} = w\log_{1P} \times \frac{N_{1P,Gi}}{n_{2P,Gi}}$$

où $N_{1P,Gi}$ est le nombre de logements de l'échantillon de première phase présents dans le groupe de seconde phase i et $n_{2P,Gi}$ le nombre de logements tirés pour l'échantillon de seconde phase parmi les logements de l'échantillon de première phase appartenant au groupe i.

Annexe A : plan de sondage du RP en grandes communes

Plan de sondage RP en grandes communes

1. Constitution des cinq groupes d'adresses

a) L'initialisation de la base de sondages (à partir du RIL de juin 2003)

Le traitement des adresses est différent selon qu'il s'agit d'une adresse qui existait déjà lors du recensement de 1999 ou bien d'une adresse nouvelle (créée après le recensement de 1999), et aussi selon qu'il s'agit d'une petite ou d'une grande adresse (est considérée comme grande adresse toute adresse dont le nombre de logements est au moins égal à 60 et qui est telle que l'ensemble des grandes adresses ne représentent pas plus de 10% des logements de la commune).

La première étape consiste à répartir dans les cinq groupes les grandes adresses (y compris les grandes adresses neuves). Pour la majorité des communes, la répartition est faite de manière déterministe : on place la plus importante des grandes adresses par tirage aléatoire dans un groupe compris entre 1 et 3[50], puis la seconde plus grande adresse dans le groupe suivant, etc.…La répartition des grandes adresses n'est aléatoire que pour les communes qui en comptent plus de 50 (soit dans le RIL 2003 seulement 34 communes sur les 892 qui ont des grandes adresses). La répartition se fait alors par un tirage équilibré sur le nombre de logements uniquement.

On répartit ensuite dans les cinq groupes les adresses « classiques» (adresses qui ne sont ni des grandes adresses ni des adresses nouvelles), qui constituent la grande majorité des adresses de la commune. La répartition se fait de manière équilibrée au niveau de la commune sur des variables du recensement de 1999 : nombre de logements et nombre de logements collectifs, répartition de la population par sexe, tranches d'âge (moins de 20 ans, 20-39 ans, 40-59 ans, 60-74 ans, 75 ans et plus).

Enfin, les adresses neuves sont réparties entre les cinq groupes. Si leur nombre est inférieur à 50, elles sont réparties de manière déterministe de façon à équilibrer le nombre de logements entre chacun des groupes. Si leur nombre est supérieur à 50, elles sont alors réparties aléatoirement de manière équilibrée sur le nombre de logements dans les cinq groupes de rotation (le nombre de logements indiqué dans le RIL est la seule variable auxiliaire dont on dispose pour ces adresses).

b) La mise à jour annuelle de la base de sondage

Chaque année, la base de sondage est mise à jour pour tenir compte des évolutions du RIL. Les mises à jour du RIL l'année N correspondent notamment à :
- l'ajout dans le RIL des adresses nouvelles déclarées créées entre N-1 et N et la suppression des adresses déclarées disparues ;
- la suppression des adresses hors champ tirées dans le RIL de l'année N-1 et enquêtées en janvier N.
- la mise à jour du nombre de logements sur les adresses tirées dans le RIL de l'année N-1 et enquêtées en janvier N.

Les adresses nouvelles de l'année N-1 qui ont été enquêtées en janvier N deviennent des adresses connues, tandis que les adresses nouvelles de l'année N sont réparties entre les cinq groupes de rotation. Cette répartition se fait de la manière suivante:
- les grandes adresses neuves sont affectées à un groupe de rotation de manière déterministe afin de répartir les grandes adresses le plus équitablement possible entre les cinq groupes (en termes de nombre de logements).
- S'il y en a plus de 50, les « petites » adresses neuves sont réparties aléatoirement en cinq paquets de manière équilibrée sur le nombre de logements, chaque paquet étant ensuite affecté à un groupe de rotation. S'il y a moins de 50 « petites » adresses neuves, elles sont réparties de manière déterministe pour équilibrer le nombre de logements entre les cinq groupes de rotation.

[50] Afin que cette adresse soit recensée rapidement dans le premier cycle RP pour des premières estimations de population.

2. Tirage de l'échantillon de l'EAR au sein du groupe de rotation impacté

Une fois le groupe de rotation tiré (1ère phase du tirage RP), la 2nde phase du tirage sélectionne les adresses du groupe de rotation qui seront enquêtées.

Les grandes adresses et les « petites » adresses neuves sont recensées exhaustivement (c'est-à-dire que la probabilité d'inclusion en 2ème phase de ces adresses dans l'échantillon est de 1). Les adresses « classiques » sont ensuite échantillonnées de telle sorte que l'échantillon total (y compris les grandes adresses et les adresses neuves enquêtées exhaustivement) représente 40% des logements du groupe de rotation enquêté. Le nombre de logements enquêtés au sein d'adresses « classiques » est alors :

$$\frac{40}{100} NLOG^t - (NLOG^t_{GC} + NLOG^t_{GN} + NLOG^t_{PN})$$

(où $NLOG^t$ représente le nombre de logements des adresses du groupe t, $NLOG^t_{GC}$ le nombre de logements des grandes adresses connues, $NLOG^t_{GN}$ le nombre de logements des grandes adresses neuves et $NLOG^t_{PN}$ le nombre de logements des petites adresses neuves).

La sélection des adresses « classiques » se fait par un tirage équilibré sur les variables ayant servi à la constitution des groupes de rotation ainsi que sur le nombre de logements de chaque IRIS de la commune. Cependant, les variables « nombre de logements de l'IRIS » étant les dernières dans l'ordre des variables d'équilibrage, l'équilibrage suivant ces variables n'est pas toujours opérant.

Annexe B : Liste des 37 grandes communes exhaustives dans OCTOPUSSE

région	identifiant des ZAE exhaustives		nombre d'enquêteurs à mobiliser	nombre de résidences principales
Ile-de-France	Z75056	PARIS	28	1110912
	Z92012	BOULOGNE-BILLANCOURT	1	52333
Champagne-Ardenne	Z51454	REIMS	2	83262
Picardie	Z80021	AMIENS	1	57593
Haute-Normandie	Z76351	HAVRE	2	79863
	Z76540	ROUEN	1	54133
Centre	Z37261	TOURS	2	66627
	Z45234	ORLEANS	1	50689
Basse-Normandie	Z14118	CAEN	1	54358
Bourgogne	Z21231	DIJON	2	71334
Nord Pas-De-Calais	Z59350	LILLE	2	99846
Lorraine	Z57463	METZ	1	53048
	Z54395	NANCY	1	52981
Alsace	Z67482	STRASBOURG	3	116767
	Z68224	MULHOUSE	1	45926
Franche-Comté	Z25056	BESANCON	1	55159
Pays de la Loire	Z44109	NANTES	3	130582
	Z49007	ANGERS	2	70810
	Z72181	MANS	2	66487
Bretagne	Z35238	RENNES	2	99462
	Z29019	BREST	2	70552
Poitou-Charentes	Z86194	POITIERS	1	42337
Aquitaine	Z33063	BORDEAUX	3	114133
Midi-Pyrénées	Z31555	TOULOUSE	5	199430
Limousin	Z87085	LIMOGES	2	66271
Rhône-Alpes	Z69123	LYON	5	216157
	Z42218	SAINT-ETIENNE	2	82269
	Z38185	GRENOBLE	2	75227
	Z69266	VILLEURBANNE	1	55136
Auvergne	Z63113	CLERMONT-FERRAND	2	67612
Languedoc-Roussillon	Z34172	MONTPELLIER	3	112008
	Z30189	NIMES	2	60191
	Z66136	PERPIGNAN	1	49902
Provence Alpes Côte d'Azur	Z13055	MARSEILLE	9	346820
	Z06088	NICE	4	164910
	Z83137	TOULON	2	73849
	Z13001	AIX-EN-PROVENCE	2	60880

Annexe C : Impact du tirage des groupes de communes RP sur la constitution des ZAE

En petites communes, un logement est chargé dans OCTOPUSSE une année N [51] (et fera donc partie des logements échantillonnables) si la commune à laquelle il appartient réunit les deux conditions suivantes :

- la ZAE à laquelle la commune appartient est tirée
- la commune est recensée l'année N .

Cependant, du fait du mode de constitution des ZAE[52], les deux évènements ne sont pas indépendants.

En particulier, la probabilité de tirage d'une commune dans OCTOPUSSE (via le tirage de la ZAE à laquelle elle appartient) n'a de sens que conditionnellement au tirage des groupes de rotation : en effet, la composition de la ZAE dans laquelle elle a été placée (et donc sa probabilité de tirage) dépend de l'association des communes au sein des groupes de rotation.

Par exemple, une petite commune placée à proximité d'une « grande » petite commune fera partie (avec une grande probabilité) de la ZAE de grande taille formée autour de la « grande » petite commune si elle appartient à un groupe de rotation distinct de la « grande » petite commune, et sera donc tirée pour OCTOPUSSE avec une grande probabilité. En revanche, si elle appartient au même groupe de rotation que la « grande » petite commune, elle ne pourra pas faire partie de la même ZAE que la « grande » petite commune (puisque la grande petite commune sature déjà la contrainte de 300 logements par groupe de rotation)[53] et fera donc vraisemblablement partie d'une ZAE plus petite s'il n'y a pas d'autres petites communes de « grande taille » à proximité (ce qui est souvent le cas) : elle aura alors une probabilité plus faible d'être tirée pour OCTOPUSSE.

Ce phénomène est illustré par l'exemple suivant (ZAE de Paray-le-Monial et commune de Vitry-en-Charollais).

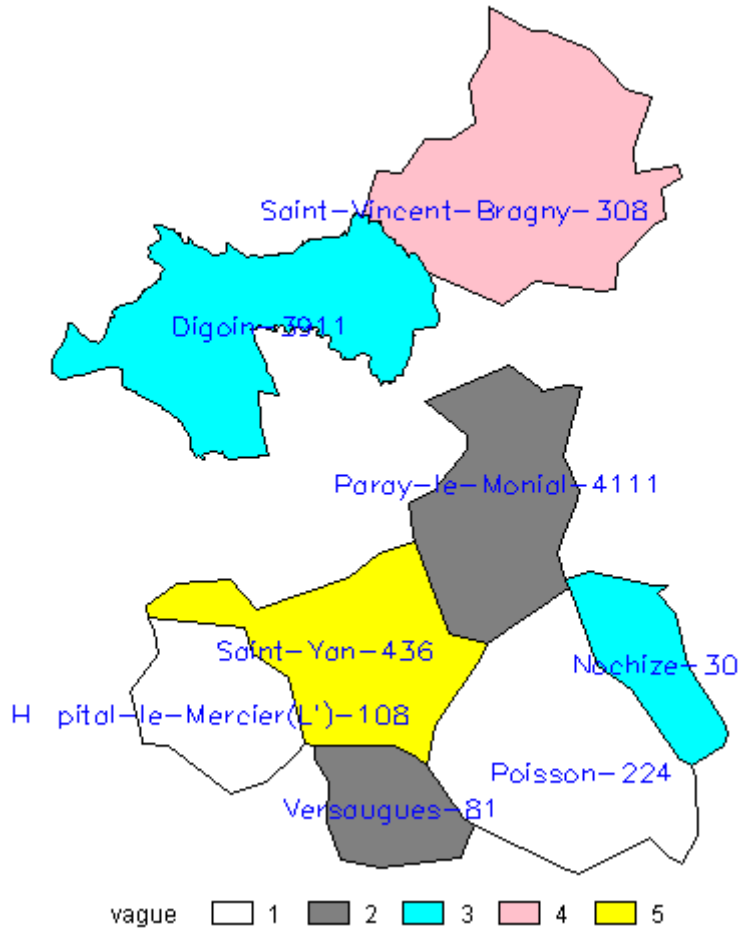
[51] Chargement en novembre de l'année N.

[52] Une ZAE est constituée en prenant les communes les proches de la commune-pivot strictement nécessaires pour satisfaire la contrainte de 300 résidences principales par groupe de rotation (sous réserve de la présence des communes nécessaires dans un rayon de 20 km à vol d'oiseau de la commune-pivot). En particulier, si la commune-pivot compte plus de 300 résidences principales, aucune une autre commune du même groupe ne sera incorporée dans la ZAE (sauf éventuellement en phase d'atterrissage) puisque la commune-pivot sature déjà la contrainte de 300 résidences principales dans son groupe de rotation.

[53] On se place ici dans un cadre simplifié où toutes les communes sont affectées en phase de vol.

ZAE de Paray-le-Monial (Z 71342)

9209 résidences principales au RP 1999
commune-pivot : Paray-le-Monial (groupe 2)



La ZAE Z 71342, représentée sur la carte ci-dessus, a été constituée avec Paray-le-Monial (4111 résidences principales) comme commune-pivot. La commune de Paray-le-Monial, qui appartient au groupe de rotation 2, est la plus grande des petites communes de la région Bourgogne : elle est donc testée en première en tant que commune-pivot. De plus, une ZAE formée autour de Paray-le-Monial compte au minimum 5311 résidences principales (4111 résidences à Paray-le-Monial, groupe 2, et au moins 300 résidences principales dans chacun des quatre autres groupes de rotation).

On étudie ici le cas de la commune de Vitry-en-Charollais (349 résidences principales), qui appartient également au groupe 2. La commune de Vitry-en-Charollais est la commune de plus de 300 résidences principales la plus proche de Paray-le-Monial. Si cette commune avait été dans un groupe de rotation différent de Paray-le-Monial, elle aurait été (sauf cas exceptionnel de répartition en groupes de rotation ne permettant pas de constituer une ZAE autour de Paray-le-Monial) obligatoirement incluse dans la ZAE de Paray-le-Monial, donc tirée pour OCTOPUSSE avec une probabilité supérieure ou égale à $k \times \frac{5311}{Nbourg}$ (où k est le nombre de ZAE non exhaustives tirées en Bourgogne et $Nbourg$ est le nombre de résidences principales total des communes non exhaustives de Bourgogne).

Schématiquement, on peut dire qu'il y avait (avant le tirage des groupes de rotation RP) 4 chances sur 5 que la commune de Vitry-en-Charollais soit dans un groupe de rotation différent de celui de Paray-le-Monial, donc qu'elle soit tirée pour OCTOPUSSE avec une probabilité supérieure à $k \times \frac{5311}{Nbourg}$.

Cependant, les communes de Paray-le-Monial et Vitry-en-Charollais sont toutes les deux dans le groupe 2. De ce fait, la commune de Vitry-en-Charollais n'est pas incluse dans la ZAE de Paray-le-Monial, mais a servi de pivot pour constituer une autre ZAE de taille beaucoup plus réduite avec au total 1770 résidences principales. Ainsi, conditionnellement au tirage des groupes RP, sa probabilité de tirage n'est que de $k \times \frac{1770}{Nbourg}$ (soit trois fois moins que si elle était tombée dans un autre groupe que Paray-le-Monial).

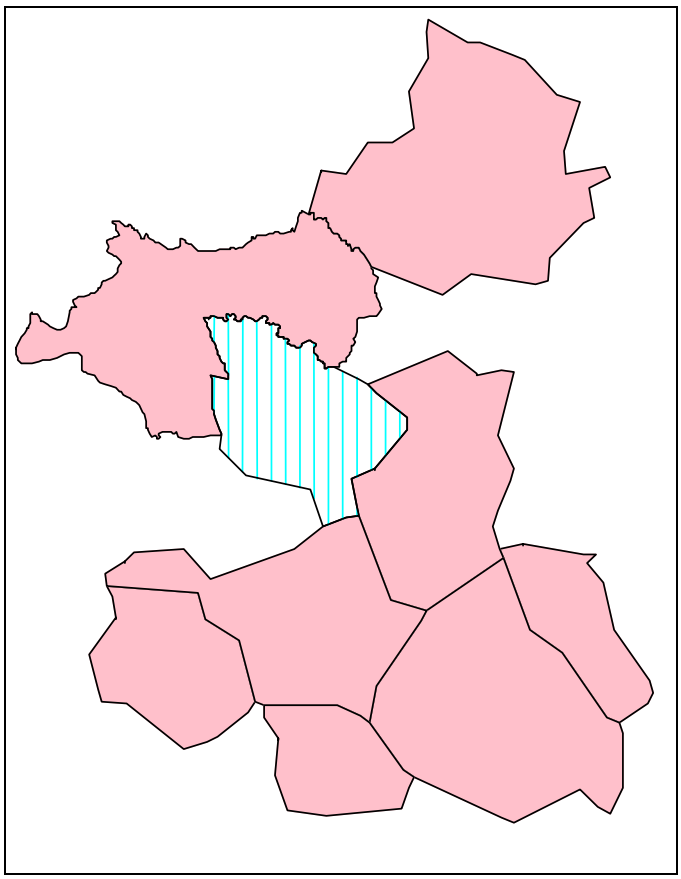
Ainsi, dans le cas de Vitry-en-Charollais, on constate que:

- la probabilité de tirage conditionnelle aux groupes RP est de $k \times \frac{1770}{Nbourg}$
- en « première approche »[54], la probabilité que deux communes appartiennent à un groupe de rotation donné est $\frac{1}{5} \times \frac{1}{5} = \frac{1}{25}$, ce qui donne une probabilité de $5 \times \frac{1}{25} = \frac{1}{5}$ que deux communes appartiennent au même groupe de rotation (soit donc une probabilité de $1 - \frac{1}{5} = \frac{4}{5}$ qu'elles appartiennent à deux groupes différents). Dans ce cadre simplifié, on constate que la probabilité de tirage non conditionnelle aux groupes RP de la commune de Vitry-en-Charollais est supérieure à : $\frac{4}{5} \times k \times \frac{5311}{Nbourg} + \frac{1}{5} \times k \times \frac{1770}{Nbourg} \approx k \times \frac{4548}{Nbourg}$.

De ce fait, les probabilités de tirages des ZAE et des communes ne peuvent être comprises que conditionnellement au tirage des groupes de rotation. Ainsi, on raisonnera toujours par la suite à groupes de rotation fixés (c'est-à-dire une fois le tirage des groupes effectué).

[54] C'est-à-dire en se plaçant dans le cadre d'un tirage poissonnien sans tenir compte de l'effet des conditions d'équilibrage sur les probabilités d'inclusion double des communes dans les groupes de rotation.

ZAE de Paray-le-Monial et commune de Vitry-en-Charollais



Annexe D : Données sur les grandes adresses (RIL 2006) pour les 775 grandes communes ayant des grandes adresses

	Ensemble des communes	Communes ayant au maximum 49 GA	Communes ayant au maximum 9 GA	Communes ayant au maximum 4 GA	Communes ayant 1 GA
Nb communes	775	671	375	227	89
Nb log GA GR 1	**203 595**	**115 072**	**26 530**	**9 747**	**2 925**
Nb log GA GR 2	195 835	110 640	26 070	9 287	1 625
Nb log GA GR 3	195 245	110 356	27 073	11 052	1 914
Nb log GA GR 4	183 420	102 500	20 237	5 776	437
Nb log GA GR 5	**184 694**	**98 821**	**16 982**	**3 239**	**68**
Total log GA	962 789	537 389	116 892	39 101	6 969

Annexe E : Algorithme de calcul des allocations de logements à tirer pour les ZAE non exhaustives proposé par Vincent Loonis

L'objectif du calcul d'allocations dans les ZAE non exhaustives est de minimiser la dispersion des poids des logements tirés dans l'échantillon sous des contraintes de nombre total de logements à tirer et de nombre minimal et maximal de logements à tirer par ZAE non exhaustive[55]. Pour les ZAE non exhaustives - ZAE notées NENA - la dispersion des poids liée à un jeu d'allocations est quantifiée au moyen de la fonction objectif : l'objectif du calcul d'allocations est alors de déterminer les allocations permettant de minimiser la valeur de la fonction objectif sous les contraintes de taille totale de l'échantillon dans le domaine et de nombre minimum et maximum de fiches-adresses à tirer par ZAE non exhaustive.

Pour le tirage dans la dernière campagne et pour le tirage dans cinq campagnes, de même que pour les différents estimateurs possibles utilisant les poids initiaux des ZAE ou les poids calés des ZAE au niveau national ou régional, la fonction objectif est de la forme :

$$F_{OBJ}(n_1,...,n_K) = \sum_{k=1}^{K} n_k \left(\frac{a_k}{n_k} - \frac{\sum_{j=1}^{K} a_j}{n}\right)^2$$

où a_k est un paramètre qui varie en fonction du nombre de campagnes impactées, de l'estimateur choisi et du type de la ZAE (ZAEGC ou ZAEPC), de sorte que $\frac{a_k}{n_k}$ corresponde au poids de sondage final d'un logement « standard » (logement quelconque en ZAEPC et logement de la strate « autres adresses » en ZAEGC[56]) tiré dans la ZAE k.

Ainsi, pour un tirage dans la dernière campagne avec calage des ZAE, le poids de sondage final d'un logement tiré dans une ZAEPC s'écrit (cf. paragraphes 5.4 et 6.7.1 pour les formules et les notations) :

$$wLog = w_{ZAE,GRi} x \frac{N_{ZAE,GRi}}{n_{ZAE}}$$

On a alors : $a_k = w_{ZAE,GRi} x N_{ZAE,GRi}$

De même, pour un logement « standard » tiré dans une ZAEGC, le poids de sondage s'écrit :

$$wLog = w_{ZAE} x \frac{1}{\pi LogR^{1P}_{ZAE,PA} \times \pi LogR^{2P}_{ZAE,AA}} x \frac{N_{ZAE,GRi}}{n_{ZAE}}$$

On a alors : $a_k = w_{ZAE} x \frac{1}{\pi LogR^{1P}_{ZAE,PA} \times \pi LogR^{2P}_{ZAE,AA}} x N_{ZAE,GRi}$

Algorithme de calcul des allocations non arrondies

[55] Dans les cas des ZAE exhaustives (37 grandes communes ayant plus de 40 000 résidences principales au RP 1999), l'allocation calculée est proportionnelle à la taille de la commune en nombre de résidences principales. On distinguera aussi à partir de 2011 les ZAE en Attente de Collecte Exhaustive (grandes communes devenues petites et n'ayant pas encore été recensées en tant que petite communes), pour lesquelles une allocation forfaitaire est calculée.

[56] On rappelle que, suite à l'étape de rééchantillonnage, le poids de sondage final des logements dans une même grande commune est rendu « presque » identique (aux variations liées aux arrondis près) quelque soit la strate d'adresses à laquelle il appartient.

Il s'agit d'un algorithme basé sur une recherche de la valeur minimum de la fonction objectif en calculant systématiquement la fonction objectif pour toutes les valeurs possibles des n_{ZAE} au sein de certaines limites fixées par l'algorithme.

En notant n le nombre total de logements à tirer dans la zone d'optimisation, on calcule pour chaque ZAE NENA de la zone d'optimisation une allocation brute n_{ZAE}^{0} selon la formule :

$$n_{ZAE}^{0} = n \times \frac{a_{ZAE}}{\sum_{\substack{ZAE \in Zone \\ ZAE = NENA}} a_k}$$

On note $n_{\min}$ et $n_{\max}$ les allocations minimales et maximales par ZAE NENA spécifiées par l'utilisateur et N_{NENA} le nombre de ZAE NENA de la zone d'optimisation.

Pour i variant de 0 à $N_{NENA} - 1$ et pour j variant de 0 à $N_{NENA} - i - 1$, on considère les allocations décimales suivantes (notée $n_{ZAE}^{i,j}$) :

- Pour les i ZAE ayant l'allocation brute n_{ZAE}^{0} la plus faible, on prend une allocation $n_{ZAE}^{i,j}$ égale à $n_{\min}$

- Pour les j ZAE ayant l'allocation brute n_{ZAE}^{0} la plus élevée, on prend une allocation $n_{ZAE}^{i,j}$ égale à $n_{\max}$

- Pour les $k_{NENA} - i - j$ ZAE restantes, on calcule une allocation décimale égale à :

$$n_{ZAE}^{i,j} = \frac{\left[n - i \times n_{\min} - j \times n_{\max}\right] \times a_{ZAE}}{\sum_{k \in ZAE restantes} a_k}$$

  - si toutes les allocations $n_{ZAE}^{i,j}$ sont comprises (bornes incluses) entre $n_{\min}$ et $n_{\max}$, le jeu d'allocation $n^{i,j}$ associé à (i,j) est admissible et on calcule la valeur de la fonction objectif associée
  - sinon, le jeu d'allocation $n^{i,j}$ n'est pas admissible.

On retient au final le jeu d'allocations admissibles correspondant à la plus petite valeur de la fonction objectif.

Les allocations sont ensuite arrondies par la méthode du cumul d'arrondis.

Annexe F : répartition par vague des logements dans OCTOPUSSE

1. Objectifs de la répartition par vagues

L'objectif de la répartition par vagues est d'aboutir pour chaque ZAE non exhaustive à une allocation globale de fiches-adresses entre les vagues sous forme de paquets « équilibrés » de fiches-adresses avec autant de paquets qu'il y a de vagues, pour autant que le nombre de fiches-adresses tirées soit suffisant pour constituer des vagues assez remplies (dans le cas contraire, on alimente un nombre réduit de vagues).

Sous réserve d'un nombre minimal de fiches-adresses par vague, on cherche donc à équilibrer au mieux la charge des enquêteurs en constituant des paquets de taille égale à une fiche-adresse près (différence liée aux arrondis).

Il a été décidé par ailleurs que, pour éviter de devoir solliciter des enquêteurs à des intervalles de temps espacés sur une même enquête à vagues, l'utilisateur pourra choisir d'affecter les paquets de fiches-adresses à des vagues contiguës lorsque le nombre de paquets constitués est inférieur à un certain seuil.

2. Calcul du nombre de fiches-adresses à affecter à chaque paquet

On définit en entrée de l'algorithme un nombre minimum $n_{\min,vague}$ de fiches-adresses par ZAE non exhaustive pour constituer une vague (identique pour toutes les ZAE). Soit n_{ZAE} le nombre de fiches-adresses tirées dans la ZAE, et v le nombre de vagues maximal à constituer (paramètre national).

On compare le nombre de vagues qu'on souhaite constituer (nombre v rentré en paramètre) et le nombre maximal de vagues qu'on peut constituer (au vu du nombre minimal de fiches-adresses par vagues en entrée). Ce nombre maximal de vagues possibles est égal à $Ent(\frac{n_{ZAE}}{n_{\min,vague}})$ où Ent est la partie entière[57]. Par exemple, si on a 55 fiches-adresses tirées et un minimum de 20 fiches-adresses par paquet, on ne peut pas constituer plus de $Ent(\frac{55}{20}) = Ent(2,75) = 2$ paquets.

Si $Ent(\frac{n_{ZAE}}{n_{\min,vague}}) \geq v$, alors il y a suffisamment de fiches-adresses pour constituer v paquets.

Si $Ent(\frac{n_{ZAE}}{n_{\min,vague}}) < v$, alors il n'y a pas assez de fiches-adresses pour constituer tous les paquets voulus et on en constitue seulement $Ent(\frac{n_{ZAE}}{n_{\min,vague}})$. Par exemple, si, pour une enquête à 4 vagues avec au moins 20 fiches-adresses par vague, on tire au niveau de l'échantillon global 55 fiches-adresses, on constituera au final un nombre de vagues égal à $Ent(\frac{55}{20}) = Ent(2,75) = 2$. A noter quand même qu'on constitue toujours au moins une vague, même si $Ent(\frac{n_{ZAE}}{n_{\min,vague}}) = 0$: dans ce

[57] Une option permettra d'arrondir à l'entier supérieur le nombre de vagues à constituer : on constituera alors $Ent(\frac{n_{ZAE}}{n_{\min,vague}}) + 1$ vagues dans la ZAE.

cas néanmoins, un message signale à l'expert sondage qu'il existe dans la ZAE une vague unique déficitaire.

Dans tous les cas, on répartit ensuite « équitablement » les fiches-adresses entre les différents paquets (différence maximale d'une fiche-adresse entre deux paquets suite aux arrondis). Dans l'exemple précédent, on constituera donc une vague de 28 logements et une autre de 27 logements.

3. Affectation des logements aux différentes vagues

L'affectation des logements aux différents paquets se fait en tirant pour chaque vague un sous-échantillon « représentatif » par tirage systématique à probabilités égales après avoir trié les logements suivant des critères de tri adaptés : ces critères de tri des logements (cinq maximum, à choisir parmi la même liste de critères de tri que ceux proposés pour le tirage d'échantillons dans OCTOPUSSE) seront spécifiés en entrée par l'utilisateur[58]. Dans l'exemple du paragraphe 22, on tirera donc parmi les 55 logements tirés dans la ZAE un sous-échantillon de 28 logements à probabilités égales, puis on constituera le second paquet avec les 27 logements restants[59].

4. Affectation des paquets de fiches-adresses aux vagues

On trie les ZAE non exhaustives par nombre croissant de fiches-adresses tirées.

Soit N le nombre total de ZAE non exhaustives dans lesquelles il est nécessaire de constituer des vagues et N_1 le nombre de ces ZAE dans lesquelles il n'y a pas assez de fiches-adresses tirées pour constituer v vagues et pour lesquelles on n'alimentera alors que certaines vagues. Comme on trie les ZAE par nombre croissant de fiches-adresses tirées, on examine d'abord les N_1 ZAE où on ne pourra pas constituer v vagues, puis ensuite les $N - N_1$ ZAE où il sera possible de les constituer. D'autre part, au sein des N_1 ZAE où on ne pourra pas constituer v vagues, on examine d'abord les N_2 ZAE pour lesquelles le nombre de vagues est supérieur à 2 et inférieur à un paramètre ρ fixé par l'utilisateur et pour lesquelles les paquets de fiches-adresses constitués doivent être affectés à des vagues contiguës (la première vague et la dernière vague n'étant pas considérées comme contiguës).

Les cas d'égalité sont résolus par ordre d'identifiant croissant de ZAE.

Tri aléatoire des vagues : un tri aléatoire des vagues est effectué pour déterminé un rang de priorité entre vagues ayant le même nombre de logements. Ce rang de priorité est notamment utilisé pour répartir de manière aléatoire les paquets de la première ZAE traitée, alors qu'aucune vague n'a encore de logements.

Répartition des paquets de fiches-adresses dans chacune des vagues :

On affecte chaque paquet à une vague de la manière suivante :

S'il s'agit d'une ZAE « affectation contiguë des paquets » :

- on trie les vagues par nombre croissant de fiches-adresses déjà affectées aux ZAE tirées antérieurement et les paquets par nombre décroissant de fiches-adresses.
- on affecte le paquet ayant le plus de fiches-adresses à la vague qui en a le moins
- on affecte le second paquet ayant le plus de fiches-adresses à celle des deux vagues contiguës à la vague d'affectation du premier paquet qui a le moins de fiches-adresses
- on affecte le troisième paquet ayant le plus de fiches-adresses à celle des vagues contiguës aux deux vagues d'affectation des deux premiers paquets qui a le moins de fiches-adresses

[58] Éventuellement : permettre de reprendre automatiquement les mêmes critères de tri que pour le tirage de l'échantillon global.

[59] Si cela simplifie l'algorithme sur le plan informatique, on peut aussi considérer qu'on tire un échantillon de 27 logements parmi les 27 logements restants.

Les cas d'égalité entre vagues (en termes de nombre de logements) sont résolus en affectant les paquets aux vagues selon l'ordre de priorité respectif (déterminé aléatoirement au début de l'algorithme) des vagues concernées par le cas d'égalité.
Les cas d'égalité entre paquets sont résolus par numéro croissant de paquet.

S'il ne s'agit pas d'une ZAE à « affectation contiguë des paquets »

- on trie les vagues par nombre croissant de fiches-adresses déjà affectées aux ZAE tirées antérieurement et les paquets par nombre décroissant de fiches-adresses.

- on affecte le plus grand paquet à la vague qui a le plus faible nombre de fiches-adresses, le second plus grand paquet à la vague qui a le second plus faible nombre de fiches adresses…

En cas d'égalité sur le nombre de fiches-adresses déjà affectées, on classe les vagues entre elles par ordre de priorité aléatoire respectif. Les cas d'égalité entre paquets sont résolus par numéro croissant des paquets.

5. Exemple de fonctionnement de l'algorithme de répartition par vague dans le cas où il n'y a pas de ZAE « à affectation contiguë des paquets »

Dans cet exemple, on a une enquête sur trois vagues, et on souhaite avoir au moins 20 fiches-adresses par vagues. On suppose que le système ne comporte que cinq ZAE, pour lesquelles on a tiré au niveau de l'échantillon global de 293 FA le nombre de FA suivant par ZAE (en ayant trié les ZAE par nombre de FA croissants) :

ZAE	Z1	Z2	Z3	Z4	Z5
Nombre de FA échantillon total	38	45	59	66	85

On suppose que l'ordre de tri aléatoire des vagues a donné comme ordre de priorité 2,1,3.

Constitution des vagues dans la ZAE Z1

38 logements seulement ayant été tirés en ZAE Z1, il n'est pas possible de constituer deux vagues. Toutes les vagues ayant le même nombre de logements (0), le paquet constitué par les 38 FA de la ZAE Z1 est affecté à la vague 2 qui est celle des trois vagues qui a le rang de priorité le plus élevé.

Constitution des vagues dans la ZAE Z2

45 logements ont été tirés en ZAE Z2, donc on constitue 2 paquets : un paquet A avec 23 logements (constitué par tirage systématique de 23 logements parmi les 45 logements présents) et un paquet B avec les 22 logements restants.
Le tri des vagues par ordre croissant de logements déjà affectés (et traitement des égalités avec le rang de priorité aléatoire) donne le classement suivant : vague 1 (0 logement), vague 3 (0 logement) et vague 2 (38 logements). On affecte donc le paquet A à la vague 1 et le paquet B à la vague 3.

On a donc alors 23 logements en vague 1, 38 logements en vague 2 et 22 logements en vague 3.

Affectation aux vagues des logements de la ZAE Z3

Comme il y a 59 logements dans la ZAE Z3, on ne peut constituer que deux paquets : un paquet A de 30 logements et un paquet B de 29 logements.
On affecte donc le paquet A à la vague la plus petite (vague 3) et le paquet B à la seconde vague la plus petite (vague 1).

On a donc alors 52 logements en vague 1, 38 logements en vague 2 et 52 logements en vague 3

Affectation aux vagues des logements de la ZAE Z4

Comme il y a 66 logements dans la ZAE Z3, on peut constituer trois paquets A, B et C de 22 logements chacun.
On affecte le paquet A à la vague la plus petite (vague 2), le paquet B à la vague 1 (qui a le même nombre de logements que la vague 3 et qui est prioritaire sur la vague 3 compte de l'ordre de priorité) et le paquet C à la vague 3.

On a donc alors 74 logements en vague 1, 60 logements en vague 2 et 74 logements en vague 3

Affectation aux vagues des logements de la ZAE Z5

Comme il y a 85 logements dans la ZAE Z3, on peut constituer trois paquets : un paquet A de 29 logements et deux paquets B et C de 28 logements.
On affecte le paquet A à la vague la plus petite (vague 2), le paquet B à la vague 1 (prioritaire sur la vague 3) et le paquet C à la vague 3.

On a donc alors 112 logements en vague 1, 89 logements en vague 2 et 112 logements en vague 3.

Synthèse : affectation aux vagues des logements tirés dans les ZAE

Répartition des logements tirés	Vague 1	Vague 2	Vague 3	**Total**
ZAE Z1	0	38	0	**38**
ZAE Z2	23	0	22	**45**
ZAE Z3	29	0	30	**59**
ZAE Z4	22	22	22	**66**
ZAE Z5	28	29	28	**85**
TOTAL	**102**	**89**	**102**	**293**

6. Exemple de fonctionnement de l'algorithme dans le cas où il y a des ZAE « à affectation contiguë des paquets »

On reprend ici le même cas de répartition par vagues que dans l'exemple précédent : on a une enquête sur trois vagues, et on souhaite avoir au moins 20 fiches-adresses par vagues. On suppose que le système ne comporte que cinq ZAE, pour lesquelles on a tiré au niveau de l'échantillon global de 293 FA le nombre de FA suivant par ZAE (en ayant trié les ZAE par nombre de FA croissants). On suppose également que lorsqu'il y a seulement deux paquets, ils doivent être affectés à des vagues contiguës : c'est-à-dire soit aux vagues 1 et 2, soit aux vagues 2 et 3, puisque les vagues 1 et 3 ne sont pas considérées comme contiguës.

ZAE	Z1	Z2	Z3	Z4	Z5
Nombre de FA échantillon total	38	45	59	66	85

Compte-tenu du nombre minimum de 20 fiches-adresses par paquet, on constitue par ZAE le nombre de paquets suivant :

ZAE	Z1	Z2	Z3	Z4	Z5
Nombre de paquets constitués	1	2	2	3	3

Les ZAE Z2 et Z3 ont un nombre de paquets supérieur ou égal à 2 et inférieur ou égal au seuil fixé par l'utilisateur (2 paquets), donc elles sont des ZAE « à affectation contiguë des paquets ». On commence donc par l'affectation aux vagues des paquets constitués dans les ZAE Z2 et Z3.

On suppose également qu'on a comme ordre aléatoire de priorité des vagues 2,1,3.

Constitution des vagues dans la ZAE Z2 « à affectation contiguë des paquets »

45 logements ont été tirés en ZAE Z2, donc on constitue 2 paquets : un paquet A avec 23 logements (constitué par tirage systématique de 23 logements parmi les 45 logements présents) et un paquet B avec les 22 logements restants.
Toutes les vagues ont le même nombre de logements (0). Compte tenu de l'ordre de priorité, on affecte le paquet qui a le plus de logements (paquet A) à la vague 2. On affecte ensuite le second paquet (B) à celle des vagues contiguës à la vague 2 qui a le plus de logements : comme elles n'ont toutes les deux aucun logement, on affecte le paquet à la vague 1 qui a l'ordre de priorité le plus important.

On a donc 22 logements en vague 1, 23 logements en vague 2 et 0 logements en vague 3.

Affectation aux vagues des logements de la ZAE Z3 « à affectation contiguë des paquets »

Comme il y a 59 logements dans la ZAE Z3, on ne peut constituer que deux paquets : un paquet A de 30 logements et un paquet B de 29 logements.
On affecte donc le paquet A à la vague ayant le moins de logements, c'est-à-dire la vague 3. Comme la vague 2 est la seule vague contiguë à la vague 3, on affecte le paquet B à la vague 2.

On a donc alors 22 logements en vague 1, 52 logements en vague 2 et 30 logements en vague 3

Constitution des vagues dans la ZAE Z1

38 logements seulement ayant été tirés en ZAE Z1, il n'est pas possible de constituer deux vagues. On affecte le paquet à la vague ayant le moins de logements, c'est-à-dire la vague 1.

On a donc alors 60 logements en vague 1, 52 logements en vague 2 et 30 logements en vague 3

Affectation aux vagues des logements de la ZAE Z4

Comme il y a 66 logements dans la ZAE Z3, on peut constituer trois paquets A, B et C de 22 logements chacun.
On affecte le paquet A à la vague la plus petite (vague 3), le paquet B à la vague intermédiaire (vague 2) et le paquet C à la vague la plus grande (vague 1).

On a donc alors 82 logements en vague 1, 74 logements en vague 2 et 52 logements en vague 3

Affectation aux vagues des logements de la ZAE Z5

Comme il y a 85 logements dans la ZAE Z3, on peut constituer trois paquets : un paquet A de 29 logements et deux paquets B et C de 28 logements.
On affecte le paquet A à la vague la plus petite (vague 3), le paquet B à la vague intermédiaire (vague 2) et le paquet C à la vague la plus grande (vague 1).

On a donc alors 110 logements en vague 1, 102 logements en vague 2 et 81 logements en vague 3.

Synthèse : affectation aux vagues des logements tirés dans les ZAE

Répartition des logements tirés	Vague 1	Vague 2	Vague 3	**Total**
ZAE Z1	38	0	0	**38**
ZAE Z2	22	23	0	**45**
ZAE Z3	0	29	30	**59**
ZAE Z4	22	22	22	**66**
ZAE Z5	28	28	29	**85**
TOTAL	**110**	**102**	**81**	**293**

Annexe G : Erreurs relatives mesurées sur les ZAE tirées, avant et après calage des ZAE, sur des variables issues du RP 1999, des DEFM et de la Taxe d'Habitation

variable	Groupe	Erreur relative EM 99	Erreur relative ZAE non calées	Erreur relative ZAE calées	Erreur liée au groupe
nb_res_princ	1		0,6	0,0	0,1
	2		0,3	-0,1	0,1
	3		-0,4	0,0	-0,1
	4		-0,9	0,0	0,0
	5		0,5	0,0	-0,1
	Total	0,4	0,0		
	Valeur réelle				
revenufisc04	1		0,5	0,0	0,1
	2		0,7	0,0	0,2
	3		-0,6	0,0	-0,2
	4		-1,3	0,0	0,0
	5		-0,1	0,0	-0,1
	total	1,1	-0,2		
	Valeur réelle				
age1	1		1,0	0,0	0,1
	2		0,4	0,0	0,0
	3		-1,5	0,0	-0,2
	4		-1,3	0,0	0,0
	5		0,7	0,0	0,1
	total	0,8	-0,2		
	Valeur réelle				
age2	1		1,1	0,0	0,1
	2		0,8	0,0	0,0
	3		-0,6	0,0	-0,2
	4		-0,8	0,0	0,1
	5		0,8	0,0	0,0
	total	0,6	0,3		
	Valeur réelle				
age3	1		-0,6	0,0	0,0
	2		-0,3	0,0	-0,1
	3		-0,4	0,0	0,0
	4		-0,9	0,0	0,0
	5		-0,3	0,0	0,0
	total	0,3	-0,5		
	Valeur réelle				
Emploi (C1A)	1		0,5	0,0	0,8
	2		-0,4	0,0	-0,6
	3		-0,2	0,0	0,3
	4		0,1	0,0	-0,7
	5		0,6	0,0	0,2
	total	0,3	0,1		
	Valeur réelle				
Emploi agricole (C11)	1		0,1	0,0	-1,8
	2		7,0	0,0	0,3
	3		-4,9	0,0	-0,5
	4		-0,2	0,0	1,2
	5		2,1	0,0	0,8
	total	6,0	0,8		
	Valeur réelle				
Emploi industriel (C21)	1		0,4	0,0	2,5
	2		-2,9	0,0	-2,0
	3		-4,4	0,0	-1,5
	4		-3,9	0,0	-1,3

	5		5,0	0,0	2,3
	total	0,9	-1,1		
	Valeur réelle				
Emploi construction (C31)	1		-0,4	0,0	0,3
	2		0,6	0,0	-1,0
	3		-1,4	0,0	0,5
	4		3,3	0,0	0,7
	5		2,1	0,0	-0,4
	total	2,9	0,9		
	Valeur réelle				
Emploi services (C41A)	1		0,7	0,0	0,6
	2		-0,3	0,0	-0,3
	3		1,3	0,0	0,8
	4		0,9	0,0	-0,8
	5		-0,7	0,0	-0,3
	total	-0,3	0,4		
	Valeur réelle				
zauer "urbain"	1		-0,8	0,0	-0,4
	2		-1,0	0,0	0,0
	3		2,3	0,0	0,2
	4		2,0	0,0	-0,6
	5		0,4	0,0	0,7
	total	-1,6	0,6		
	Valeur réelle				
zauer "périurbain"	1		2,3	0,0	-0,7
	2		7,7	0,0	2,3
	3		-2,3	0,0	-2,8
	4		-3,6	-0,1	2,6
	5		10,3	0,0	-1,5
	total	7,1	2,9		
	Valeur réelle				
zauer "rural"	1		3,4	0,0	2,4
	2		-3,3	-0,7	-2,1
	3		-7,9	-0,1	1,5
	4		-8,1	0,0	-0,6
	5		-9,4	0,0	-1,2
	total	0,5	-5,1		
	Valeur réelle				
diplôme "aucun"	1		0,6	0,1	0,0
	2		-1,0	-0,8	-0,1
	3		-0,7	0,5	0,1
	4		0,6	1,8	0,1
	5		1,2	0,3	0,0
	total	0,3	0,1		
	Valeur réelle				
diplôme "Certif"	1		-0,5	-0,4	-0,2
	2		0,5	0,4	-0,1
	3		-0,9	0,2	-0,1
	4		-1,2	-0,1	0,1
	5		1,0	0,5	0,2
	total	-0,3	-0,2		
	Valeur réelle				
diplôme "CAP"	1		0,8	0,0	0,3
	2		0,9	0,2	-0,2
	3		-1,1	-0,2	0,0
	4		-1,1	-0,1	0,1
	5		0,8	-0,5	-0,1
	total	0,7	0,1		
	Valeur réelle				
diplôme "BEP"	1		0,6	-0,2	0,0

	2		0,6	-0,1	-0,2
	3		-1,5	-0,4	-0,1
	4		-1,2	-0,3	0,1
	5		1,1	0,1	0,2
	total	1,3	-0,1		
	Valeur réelle				
diplôme "BEPC"	1		1,4	0,5	0,2
	2		0,9	0,3	-0,2
	3		-0,2	0,1	-0,1
	4		-1,1	-0,4	0,1
	5		0,0	-0,4	0,0
	total	0,7			
	Valeur réelle				
diplôme "Bac pro/techno"	1		2,1	1,0	0,2
	2		1,8	0,6	-0,1
	3		0,2	0,6	-0,2
	4		-1,0	-0,4	0,1
	5		0,3	-0,7	0,0
	total	1,6	0,7		
	Valeur réelle				
diplôme "Bac général"	1		1,3	0,3	0,3
	2		1,6	0,8	0,2
	3		0,0	-0,5	-0,4
	4		-1,3	-1,4	0,1
	5		-0,2	0,1	-0,2
	total	0,7	0,3		
	Valeur réelle				
diplôme "1er cycle univ, "	1		1,5	0,1	0,2
	2		1,4	0,2	0,3
	3		0,1	0,0	-0,4
	4		-1,1	-0,7	0,0
	5		0,5	0,2	0,0
	total	0,8	0,5		
	Valeur réelle				
diplôme "2 et 3ième cycle univ, "	1		0,7	0,0	-0,1
	2		0,7	0,2	0,5
	3		-0,4	-0,9	-0,5
	4		-1,5	-0,5	0,1
	5		-1,1	0,4	-0,1
	total	0,7	-0,3		
	Valeur réelle				
DEFM au 30/09/05	1		0,7	-0,2	0,0
	2		-0,3	0,2	0,4
	3		-0,1	0,7	0,0
	4		-1,0	-0,1	-0,2
	5		0,4	-0,2	-0,2
	total	0,3	0,0		
	Valeur réelle				
DEFM au 30/09/06	1		0,7	-0,1	0,0
	2		-0,2	0,3	0,3
	3		0,1	0,9	0,1
	4		-0,9	-0,2	-0,1
	5		0,8	0,3	-0,3
	total	0,7	0,1		
	Valeur réelle				
DEFM au 30/09/07	1		0,2	-0,5	0,0
	2		-0,7	-0,2	0,2
	3		0,0	0,8	0,0
	4		-1,2	-0,6	-0,1

	5		0,5	0,0	-0,1
	total	0,4	-0,2		
	Valeur réelle				
Nombre de référents célibataires (celib) Source : TH 2006	1		1,1	0,1	0,1
	2		0,3	-0,3	0,3
	3		0,5	0,6	-0,1
	4		-0,5	0,3	-0,2
	5		0,5	-0,3	-0,1
	total	0,0	0,4		
	Valeur réelle				
revenu global du foyer (foy_rev) Source TH 2006	1		0,3	0,3	-0,2
	2		0,5	0,3	0,3
	3		-0,2	0,3	-0,1
	4		-0,7	1,3	0,0
	5		-0,3	1,0	0,0
	total	1,1	-0,1		
	Valeur réelle				
Nombre de rés princ HLM (HLM)	1		2,0	0,6	0,1
	2		-0,6	0,2	0,4
	3		-0,2	0,3	-0,5
	4		2,7	4,1	0,0
	5		1,1	-0,5	0,1
	total	-0,9	1,0		
	Valeur réelle				
Nombre de locataires (locat)	1		1,5	0,3	0,3
	2		-0,5	-0,4	0,2
	3		0,7	0,6	-0,2
	4		0,4	1,2	0,0
	5		0,4	-0,4	-0,2
	total	-0,4	0,5		
	Valeur réelle				
famille monoparentale	1		1,5	0,5	0,2
	2		-0,1	0,0	0,2
	3		0,4	0,4	-0,2
	4		0,4	1,2	0,0
	5		1,2	0,8	-0,1
	total	0,4	0,7		
	Valeur réelle				
couple sans enfant	1		0,8	-0,2	0,1
	2		1,1	0,3	-0,1
	3		-1,5	-0,3	-0,1
	4		-1,4	-0,3	0,1
	5		0,9	0,0	0,0
	total	0,9	0,0		
	Valeur réelle				
couple avec enfants	1		0,2	0,0	0,0
	2		0,2	-0,2	0,0
	3		-0,2	0,2	0,0
	4		-0,8	0,1	-0,1
	5		0,6	0,4	0,0
	total	0,5	0,0		
	Valeur réelle				