SIMULATIONS DE TIRAGES DE ZONES D'ACTION POUR LES ENQUETES DE L'INSEE

Fabien GUGGEMOS (*)

(*) Université de Neuchâtel (Suisse), Institut de Statistiques

Introduction

Le recensement exhaustif de la population est par définition le meilleur moyen d'établir des chiffres de population légale et de fournir - au-delà de simples données démographiques - des indicateurs sur des caractéristiques de natures diverses de la population, telles des caractéristiques économiques ou socio-démographiques. C'est ainsi que l'Institut National de la Statistique et des Etudes Economiques a régulièrement mis en œuvre depuis sa création en 1946 des recensements exhaustifs de la population française. Néanmoins, la réalisation de ces recensements est soumise à un certain nombre de contraintes pratiques et budgétaires importantes. Aussi a-t-on assisté ces dernières décennies à un allongement progressif de la période intercensitaire séparant deux recensements successifs, le dernier d'entre eux initialement prévu pour 1997 n'ayant d'ailleurs été réalisé qu'en 1999.

Ce phénomène d'allongement de la période intercensitaire pose naturellement la question de la bonne couverture temporelle des ménages offerte par les systèmes d'échantillonnage des enquêtes-ménages tels que l'échantillon-maître, réserve de logements constituée ad hoc après chaque recensement et dans laquelle sont tirés jusqu'au recensement suivant la plupart des échantillons des enquêtes ménages menées par l'INSEE. Face aux mutations socio-économiques rapides, la nécessité d'avoir régulièrement des informations fraîches a conduit l'INSEE à adopter une nouvelle méthode de recensement de la population française, mise en œuvre depuis 2004, passant d'un recensement exhaustif ponctuel à un recensement rotatif partiel mais continu.

Un tel changement représente un véritable bouleversement méthodologique et a eu notamment pour conséquence la redéfinition complète du système actuel d'échantillonnage des enquêtes ménages. La constitution des unités primaires, entités formant une partition du territoire national, a notamment dû être intégralement refondue en prenant en considération l'aspect rotatif du recensement ; elle s'est achevée fin janvier 2007. Dès lors, le choix des nouvelles unités primaires au sein desquelles seront tirés les logements pour les enquêtes de l'INSEE durant les dix prochaines années s'avère être une étape fondamentale : il est en effet nécessaire de tester diverses stratégies de tirages aléatoires de ces zones et de mesurer leur impact sur les précisions et représentativités des enquêtes à venir afin de pouvoir retenir la meilleure stratégie possible à utiliser pour le tirage final effectif.

L'objet de cette présentation est d'exposer les travaux effectués au printemps 2007 au sein de l'Unité Méthodes Statistiques (UMS) de l'INSEE, qui visaient précisément à mettre en œuvre ces tests des différentes stratégies de tirage aléatoire envisageables. Après avoir d'abord présenté plus en détail les enjeux de ces travaux dans le cadre du projet « Octopusse », on exposera la mise en œuvre théorique des simulations réalisées, puis l'on procédera dans un troisième temps à la comparaison empirique des diverses stratégies de tirage adoptées.

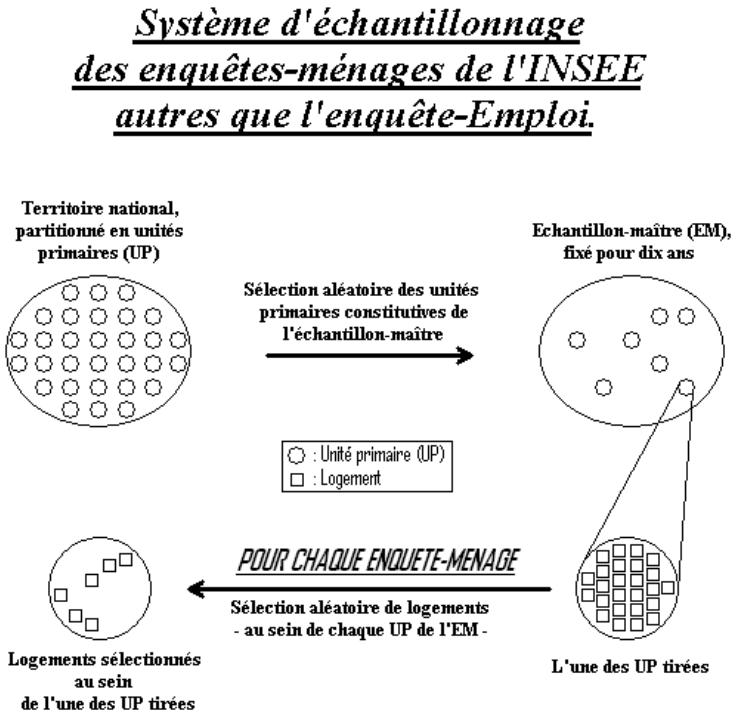
1. Objectifs et perspectives dans le cadre du projet « Nouveaux Systèmes d'Echantillonnage ».

1.1. Le développement du projet OCTOPUSSE.

1.1.1. Quelques mots sur le système d'échantillonnage actuel.

Jusqu'en 2009, les échantillons des enquêtes nationales auprès des ménages conduites par l'INSEE (à l'exception de ceux de l'enquête Emploi) sont tirés au sein de l'échantillon-maître (EM), réserve de logements qui est reconstituée à l'issue de chaque recensement national de la population et complétée durant la période intercensitaire par des sources annexes assurant la couverture des logements nouvellement construits. En d'autres termes, l'échantillon-maître constitue la base de sondage pour les enquêtes ménages de l'INSEE.

Plus précisément, le parc de logements à enquêter lors des enquêtes-ménages résulte d'un sondage aléatoire à deux degrés: Dans un premier temps sont tirées aléatoirement des unités primaires (communes ou regroupements de communes) pour constituer l'échantillon-maître ; au sein de ces dernières sont ensuite tirés aléatoirement, pour chaque enquête-ménage de la période intercensitaire, les logements à enquêter. Ceci permet d'éviter la dispersion des zones à enquêter et donc de respecter davantage les contraintes – sociales et financières – liées au déplacement des enquêteurs, tout particulièrement dans les zones reculées (rural « profond », montagne…).



1.1.2. Le nouveau recensement de la population.

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 partiel rotatif continu. Celui-ci a définitivement été mis en place en janvier 2004.

Désormais, les communes de moins de 10000 habitants sont recensées exhaustivement tous les cinq ans par roulement, tandis que les communes comprenant 10000 habitants ou plus font l'objet d'une enquête de recensement par sondage chaque année au taux moyen de 8%, les échantillons annuels étant disjoints au cours d'un même cycle de cinq ans.

On peut ainsi définir 5 groupes de rotation dans lesquels sont réparties les petites communes (PC, moins de 10000 habitants) selon l'année du cycle quinquennal où elles sont recensées. L'un des principaux intérêts de cette manière d'opérer réside dans la fraîcheur des données de la base de sondage issue du nouveau recensement puisque le temps séparant la date de collecte de l'enquête de la date à laquelle a été effectué le recensement est considérablement réduit. En effet, la sélection des échantillons pour les enquêtes s'opérera désormais dans la partie de la base qui aura été recensée durant l'année précédente.

1.1.3. Le projet OCTOPUSSE.

Le principe évoqué précédemment constitue le fondement du projet « Nouveaux systèmes d'échantillonnages », baptisé *OCTOPUSSE* (Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Echantillons), et dont la conception a été développée par l'Unité Méthodes Statistiques de l'INSEE (UMS). Lancé fin 2003, il vise à reconstruire les méthodes de sélection des échantillons pour les enquêtes en tenant compte du nouveau mode de recensement. Il est prévu qu'il soit opérationnel en 2009, date à laquelle le premier cycle quinquennal du nouveau recensement aura été achevé.

1.2. La nécessité de repenser la construction des unités primaires.

1.2.1. Des unités primaires 99 aux zones d'action enquêteurs.

Comme tout échantillon d'une enquête ayant lieu l'année *N* devra avoir été puisé dans la base des logements recensés durant l'année $N-1$, il faut s'assurer que les unités primaires susceptibles d'être sélectionnées possèdent des logements appartenant au groupe de rotation correspondant à l'année $N-1$. Puisque les grandes communes sont recensées chaque année – certes via une enquête par sondage –, le problème se pose avant tout pour les unités primaires constituées de petites communes. Rien en effet n'assure que les unités primaires de l'échantillon-maître 99 constituées de petites communes présentent des logements situés dans chacun des cinq groupes de rotation, puisque cette notion n'existait pas en 1999. D'où la nécessité de redéfinir de nouvelles unités primaires, renommées - à l'occasion du nouveau recensement - Zones d'Action Enquêteurs (ZAE). Ce travail a été mené à l'automne 2006 par la division Echantillonnage et Traitement Statistique des Données de l'UMS.

1.2.2. La constitution des ZAE.

A cet effet, la construction des ZAE a reposé sur les principes suivants :

- les ZAE doivent être des zones fixes pour une durée de 10 ans (pour pouvoir leur affecter un enquêteur stable dans le temps et localisé à proximité).
- toute grande commune constitue à elle seule une ZAE, appelée ZAE grande commune ou ZAEGC.
- pour assurer une réserve de logements chaque année, les ZAE petites communes (ZAEPC) doivent comporter des communes de chaque groupe de rotation.
- les ZAE doivent respecter les frontières régionales afin de pouvoir les rattacher sans ambiguïté à l'une des directions régionales (DR) de l'INSEE.
- en outre, chaque groupe de rotation d'une ZAEPC doit contenir un nombre minimal de logements principaux susceptibles d'être enquêtés, fixé à 300, et ce afin d'assurer d'une part une charge annuelle de travail suffisante pour l'enquêteur affecté à la zone en question, de satisfaire d'autre part le principe de disjonction qui impose d'effectuer plusieurs enquêtes la même année sans réinterroger les mêmes logements.

Chaque ZAEPC a été construite autour d'une commune pivot (le « centre » de la ZAE, en général la commune la plus peuplée de la ZAE), en se basant sur des critères de distance au pivot inférieure à un seuil fixé. Cela permet d'obtenir des zones compactes, d'étendue raisonnable, dans une optique de limitation des déplacements des enquêteurs au sein même de leur ZAE d'affectation.

1.2.3. Découpage final du territoire national.

Au final, les 36613 communes du code géographique officiel 2006 ont pu être affectées à une ZAE à la fin du mois de janvier 2007. Néanmoins, les contraintes de constitution mentionnées précédemment n'ont pas toujours pu être toutes scrupuleusement respectées. On retiendra en effet que le territoire national a pu être découpé en 3785 Zones d'Action Enquêteurs, dont 892 ZAEGC et 2893 ZAEPC. Parmi ces dernières, 4 ne possèdent des communes que dans seulement quatre des cinq groupes de rotation (2 ZAE en Pays de la Loire, 1 en Provence Alpes Côte d'Azur, 1 en Rhône-Alpes). Quelques autres ne satisfont pas en outre la contrainte plancher de 300 logements par groupe de rotation tout en en possédant néanmoins 1500 sur l'ensemble de la ZAE. Signalons par ailleurs le statut particulier des villes de Paris, Lyon et Marseille pour lesquelles chacun des arrondissements a été considéré comme constitutif d'une ZAE. Ces quelques particularités vont jouer un rôle dans le choix des plans de sondage adoptés ultérieurement ainsi que dans l'interprétation de certains résultats.

1.3. Les enjeux des travaux de simulation réalisés.

Du fait de la nouvelle partition du territoire national en ZAE, les stratégies de tirage d'unités primaires pour les enquêtes, employées sur l'échantillon-maître 99, s'avèrent caduques et doivent donc être intégralement remises en question. L'enjeu essentiel est de trouver un plan de sondage qui assure que, lorsque les tirages effectifs pour les enquêtes seront réalisés, les résultats de ces dernières seront suffisamment précis et représentatifs de l'ensemble de la population.

A ce stade, plusieurs grandes questions se posent :

- Tout d'abord, est-on assuré qu'un échantillon bien représentatif une année donnée du cycle quinquennal le soit encore une autre année ? Le problème qui est soulevé là est fondamental, il met en lumière un aspect contraignant de la façon dont ont été constituées les ZAE, à savoir qu'une ZAE ne présente pas systématiquement une forte homogénéité entre ses différents groupes de rotation.
- Le principe du respect du zonage ZAUER (découpage en aires urbaines) n'ayant finalement pas été retenu comme critère lors de la constitution des ZAE, comment assurer une bonne représentativité des différents types d'espaces - urbain, périurbain et rural - dans les enquêtes ?
- Le bon comportement d'un plan de sondage résiste-t-il correctement à l'épreuve du temps, en particulier à l'évolution des données contenues dans la base de sondage ? L'enjeu est de taille puisque le nouvel échantillon-maître est destiné à être utilisé de 2009 jusqu'en 2019.
- Face aux disparités (notamment démographiques) qui existent entre les différentes régions, peut-on envisager d'adopter des plans de sondage différant légèrement d'une région à l'autre, et cela permettrait-il d'améliorer les résultats des enquêtes ?

Les simulations de tirage aléatoire de ZAE qui ont été entreprises visent avant tout à apporter des réponses à ces enjeux, pour guider le choix du plan de sondage finalement retenu lors du tirage effectif du nouvel échantillon-maître à l'automne 2007. Elles font l'objet des deux prochaines parties de cette présentation.

2. Présentation théorique des calculs de précision effectués par simulations.

Dans cette partie est présentée la démarche théorique qui a été suivie puis mise en pratique pour apprécier la qualité des divers plans de sondage envisagés. Elle vise ainsi à exposer comment formellement on a cherché à apporter des réponses aux enjeux posés dans la partie précédente. Après quelques mots généraux sur la notion d'échantillons équilibrés – notion centrale dans les

travaux réalisés – on s'attachera à détailler sur quels critères sont effectués les tirages aléatoires des simulations puis quelles sont les estimations entreprises sur les échantillons obtenus pour évaluer la pertinence des plans de sondage testés.

2.1. La notion d'échantillonnage équilibré.

Considérons une population U, composée d'un ensemble de N individus indexés par l'indice k. Dans le cas présent, les individus seront en fait les ZAE qui forment une partition du territoire national. Un échantillon sur cette population n'est autre qu'une partie de cette dernière, on en dénombre par conséquent 2^N. Définir un plan de sondage sur la population U consiste alors à choisir une distribution de probabilités *p* sur l'ensemble *S* des 2^N échantillons, le tirage aléatoire d'un échantillon étant alors effectué selon cette distribution. Dans la pratique, le nombre d'individus de la population étant élevé, la définition du plan de sondage ne passe pas par la détermination de la loi *p* (Avec 3785 ZAE, il y aurait notamment $2^{3785} \approx 2.5 \cdot 10^{1139}$ échantillons potentiels !) mais par celle d'un certain nombre d'autres paramètres, au premier rang desquels figurent les probabilités d'inclusion d'ordre 1.

La probabilité d'inclusion (d'ordre 1) d'un individu k est définie comme étant la probabilité que cet individu soit présent dans l'échantillon tiré. Autrement dit, $\pi_k = \sum_{s \ni k} p(s)$ où la somme porte sur l'ensemble des échantillons *s* de *S* qui contiennent l'individu k. Remarquons qu'on peut définir de manière analogue les probabilités d'inclusion d'ordres supérieurs, $\pi_{i_1 \dots i_k} = \sum_{s \ni i_1 \dots i_k} p(s)$ étant la probabilité que les individus i_1, …, i_k appartiennent simultanément à l'échantillon tiré ; il est alors aisé de prouver que la donnée des probabilités d'inclusion d'ordre 1 à N est équivalente à la donnée de la distribution *p*.

La grandeur retenue afin d'apprécier la « représentativité » des échantillons tirés vis-à-vis d'une variable d'intérêt donnée est le total de cette variable sur l'ensemble de la population. Si X_1, …, X_N désignent les valeurs que prend la variable X auprès des individus, un estimateur naturel - calculé sur l'échantillon *s* tiré - du total de X sur l'ensemble de la population est l'estimateur d'Horvitz-Thompson (ou π-estimateur) :

$$\hat{T}_X = \sum_{k \in s} \frac{X_k}{\pi_k}.$$

Cet estimateur est en particulier non biaisé et sa variance est donnée par :

$$Var(\hat{T}_X) = \sum_{k \in U} \frac{X_k^2}{\pi_k^2} Var(I_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{X_k X_l}{\pi_k \pi_l} Cov(I_k, I_l) = \sum_{k \in U} \frac{X_k^2 (1 - \pi_k)}{\pi_k} + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{X_k X_l (\pi_{kl} - \pi_k \pi_l)}{\pi_k \pi_l}$$

dont un estimateur sans biais est:

$$\hat{Var}(\hat{T}_X) = \sum_{k \in S} \frac{X_k^2 (1 - \pi_k)}{\pi_k^2} + \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{X_k X_l (\pi_{kl} - \pi_k \pi_l)}{\pi_{kl} \pi_k \pi_l}.$$

En pratique, les probabilités d'inclusion d'ordre 2 ne sont pas connues et l'estimateur précédent de la variance ne peut donc pas être déterminé.

La notion d'échantillonnage équilibré sur la variable d'intérêt X répond à l'idée intuitive que l'on se fait d'un échantillon qualifié de « représentatif, pour la variable X, de l'ensemble de la population ». Concrètement, un échantillon *s* est dit équilibré si l'estimateur d'Horvitz-Thompson du total de X est exactement égal au vrai total de X sur l'ensemble de la population. Un plan de sondage est lui-même équilibré si tous les échantillons susceptibles d'être tirés suivant ce plan de sondage (c'est-à-dire les échantillons *s* tels que $p(s) > 0$) sont équilibrés, ce qui revient à dire que la variance de l'estimateur

d'Horvitz-Thompson pour le total de X doit être nulle. A titre d'exemple, un équilibrage sur les probabilités d'inclusion permet de s'assurer le tirage d'un échantillon de taille fixe $n = \sum_{k \in U} \pi_k$, car

$$card(s) = \sum_{k \in s} 1 = \sum_{k \in s} \frac{\pi_k}{\pi_k} = \sum_{k \in U} \pi_k = n .$$

Pour que les résultats des enquêtes soient « représentatifs » de l'ensemble de la population, il est donc essentiel de tirer des échantillons équilibrés sur un certain nombre de variables à définir. Un algorithme, appelé algorithme du CUBE, a été conçu à cet effet par Jean-Claude Deville et Yves Tillé à la fin des années 90.

2.2. L'algorithme du CUBE.

2.2.1. Présentation de l'algorithme du CUBE.

L'algorithme du CUBE a pour objet d'effectuer le tirage aléatoire d'un échantillon équilibré sur un certain nombre de variables pré-choisies, en respectant le jeu de probabilités d'inclusion d'ordre 1 des individus.

L'algorithme doit son nom au fait qu'il se base sur une représentation géométrique des différents échantillons possibles de la base de sondage de taille N dans l'hypercube unité de dimension N. Chaque sommet de l'hypercube représente l'un des 2^N échantillons possibles, ses coordonnées formant un vecteur de taille N, constitué exclusivement de 0 ou de 1, la k-ième coordonnée étant égale à l'indicatrice de présence de l'individu k dans l'échantillon en question.

Notons $\pi = (\pi_1, ..., \pi_N)'$ le vecteur des probabilités d'inclusion des individus et $X_i = (X_{1,i}, ..., X_{N,i})'$ le vecteur des valeurs prises par la variable d'équilibrage X_i auprès de chaque individu et définissons alors la matrice « des contraintes » à p lignes et N colonnes (où p est le nombre de variables d'équilibrage) :

$$A = \begin{pmatrix} \frac{X_{1,1}}{\pi_1} & \cdots & \frac{X_{k,1}}{\pi_k} & \cdots & \frac{X_{N,1}}{\pi_N} \\ \frac{X_{1,2}}{\pi_1} & \cdots & \frac{X_{k,2}}{\pi_k} & \cdots & \frac{X_{N,2}}{\pi_N} \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ \frac{X_{1,p}}{\pi_p} & \cdots & \frac{X_{k,p}}{\pi_k} & \cdots & \frac{X_{N,p}}{\pi_N} \end{pmatrix}$$

L'ensemble des échantillons équilibrés sur les p variables X_i - c'est-à-dire l'ensemble des échantillons s vérifiant $\forall i \in [1; p], \sum_{k \in s} \frac{X_{k,i}}{\pi_k} = \sum_{k \in U} X_{k,i}$ - est alors exactement l'ensemble des sommets $s = (s_1, ..., s_N)'$ de l'hypercube qui satisfont l'équation matricielle $A \cdot s = A \cdot \pi$. C'est encore l'intersection de l'ensemble des sommets de l'hypercube avec le sous-espace affine des contraintes $KerA + \pi$. Noter qu'une telle intersection n'est pas nécessairement non vide, autrement dit il n'existe pas nécessairement d'échantillon satisfaisant exactement toutes les contraintes d'équilibrage. Dans les situations rencontrées en pratique, c'est en général le cas et l'on cherche alors à tirer un échantillon respectant « au mieux » les contraintes d'équilibrage.

Le principe général est le suivant : en partant du vecteur π, l'algorithme du CUBE génère ainsi une marche aléatoire à l'intérieur de l'intersection de l'hypercube et de l'espace des contraintes

jusqu'à atteindre l'un des sommets de l'hypercube si toutes les contraintes sont vérifiées ou l'une de ses faces dans le cas contraire. Si cette dernière éventualité se produit, cette première phase, dite de « vol », est suivie d'une deuxième, la phase d' « atterrissage » qui permet d'atteindre l'un des sommets de l'hypercube à la fois « proche » de l'équilibre souhaité et du point déjà atteint. Plusieurs options existent pour préciser la façon d'opérer durant la phase d'atterrissage. Celle qui a été retenue pour les simulations consiste à relâcher progressivement les contraintes d'équilibrage suivant l'ordre croissant d'importance qui a été attribué à ces dernières.

L'algorithme du CUBE renvoie donc en sortie un échantillon tiré aléatoirement selon un jeu de probabilités d'inclusion donné et équilibré - ou approximativement équilibré - sur les variables d'équilibrage choisies.

2.2.2. La nouvelle version de l'algorithme du CUBE : Fast-CUBE.

La marche aléatoire générée par l'algorithme du CUBE suppose la recherche, à chaque étape, d'un vecteur du noyau de la matrice A, ce qui est en général assez coûteux en temps : La matrice A étant de dimension $p \times N$, son noyau est de dimension au moins N-p, nombre en général très élevé (le nombre d'individus de la base de sondage est souvent très important, tandis que le nombre p de contraintes d'équilibrage retenues n'excède pas quelques dizaines en général).

L'amélioration de l'algorithme du cube, nommée fast-cube, proposée par Guillaume Chauvet et Yves Tillé en 2004 consiste à rechercher à chaque étape un vecteur non nul dans le noyau d'une sous-matrice de A de taille $p \times (p+1)$ (un tel noyau est de dimension au moins égale à 1), réduisant considérablement de la sorte le temps de calcul.

2.3. Définition des plans de sondage adoptés pour les simulations.

Dans cette sous-partie est exposée la démarche adoptée pour déterminer les plans de sondage testés lors des simulations. Ceux-ci présentent un certain nombre de points communs qui sont développés ici, ce qui les distingue étant avant tout le choix des variables d'équilibrage.

2.3.1. Le principe de stratification régionale.

Le premier principe sous-jacent à l'élaboration des plans de sondage testés est celui de la stratification régionale : les travaux effectués, qu'il s'agisse de tirages aléatoires de ZAE ou de calculs d'estimateurs, sont systématiquement entrepris région par région. Les résultats au niveau national sont alors obtenus par agrégation des résultats de chacune des 22 régions. Ce principe est imposé par l'organisation structurelle et administrative de l'INSEE qui possède une antenne (direction régionale) dans chaque région française, sous la tutelle de laquelle se trouvent les enquêteurs de la région en question. Cela implique notamment que lorsque 1000 tirages de ZAE sont réalisés, comme c'est le cas dans la plupart des simulations qui suivent, ce sont en réalité 22000 tirages qui ont été effectués.

2.3.2. Calcul des probabilités d'inclusion et détermination des ZAE exhaustives.

Comme pour le plan de sondage adopté pour sélectionner l'échantillon-maître 99, la variable de référence est le nombre de résidences principales, notée *nres*. Il s'agit de sélectionner les ZAE avec des probabilités d'inclusion proportionnelles à leur nombre de résidences principales. On peut montrer que ce choix conduit à des estimations précises pour toute variable corrélée à *nres*.

A cet effet, notons τ le taux de sondage moyen (proportion des résidences principales qui doivent être enquêtées), e le nombre de fiches-adresses par enquêteur (égal au nombre de résidences principales enquêtées par un même enquêteur lors d'une enquête) et *nres_région* le total

du nombre de résidences principales de la région considérée. Le rapport $\frac{nres\_région \times \tau}{e}$ définit donc le nombre d'enquêteurs à mobiliser pour chaque enquête dans la région. Conformément aux usages antérieurs, le taux de sondage a été fixé à 1/2000 et le nombre de fiches-adresses par enquêteur à 20. Le rapport e/τ vaut donc 40000 et correspond à la capacité de couverture d'un enquêteur, en termes de nombre de résidences principales.

2.3.2.1. Combien d'enquêteurs allouer à une ZAE ?

Dans un premier temps, l'allocation « 1 enquêteur pour une ZAE » a été envisagée et mise en œuvre. Néanmoins, il est très vite apparu que celle-ci ne s'avérait pas nécessairement très pertinente. En effet, certaines grandes villes de province ainsi que certains arrondissements de Paris, Lyon et Marseille possèdent un nombre de résidences principales bien supérieur à 40000, nombre théorique maximal de résidences principales que peut couvrir un enquêteur. Aussi ce point de vue a-t-il vite été abandonné.

L'allocation choisie a finalement été la suivante :

- les ZAE pour lesquelles *nres* est inférieur à $e/\tau = 40000$ se voient affecter, si elles sont tirées, un seul enquêteur.

- les ZAE pour lesquelles *nres* est supérieur à $e/\tau = 40000$ se voient affecter un nombre d'enquêteurs égal à $\frac{nres \times \tau}{e} = \frac{nres}{40000}$, arrondi à l'entier le plus proche. (Ces ZAE sont tirées avec probabilité 1, cf section suivante).

2.3.2.2. Détermination des probabilités d'inclusion.

Partant de ces considérations ainsi que du principe selon lequel les probabilités d'inclusion doivent être proportionnelles au nombre de résidences principales des ZAE (dans la mesure où elles demeurent inférieures à 1), le mode de calcul des probabilités d'inclusion qui a été suivi est le suivant :

1. Initialisation : $nbenq = \frac{nres\_région \times \tau}{e}$ et $reste = nres\_région$.
2. Tri des ZAE de la région considérée par nombre décroissant de résidences principales.
3. Pour chaque ZAE, traitée dans l'ordre établi à l'étape précédente,
   - Si $nbenq \times \frac{nres_{ZAE}}{reste} > 1$, alors faire $\begin{cases} \pi_{ZAE} = 1 \\ nbenq = nbenq - arrondi\left(\frac{nres_{ZAE} \times \tau}{e}\right) \\ reste = reste - nres_{ZAE} \end{cases}$
   - Sinon faire $\pi_{ZAE} = nbenq \times \frac{nres_{ZAE}}{reste}$.

Ce procédé fait ressortir une liste de ZAE qui sont systématiquement tirées ($\pi_{ZAE} = 1$), qualifiées par la suite d'exhaustives. Les probabilités d'inclusion des ZAE non exhaustives sont quant à elles parfaitement proportionnelles à leur nombre de résidences principales. Il est aisé de prouver que le seuil d'exhaustivité, c'est-à-dire le nombre de résidences principales au-delà duquel une ZAE est exhaustive, est simplement égal au rapport e/τ. Il est en particulier intéressant de constater que ce seuil est indépendant de la région considérée.

2.3.2.3. Traitement des arrondissements de Paris, Lyon et Marseille

Le statut des villes de Paris, Lyon et Marseille est quelque peu particulier puisque chacun de leurs arrondissements constitue une ZAE. Or, en appliquant l'algorithme de calcul des probabilités d'inclusion précédent, les petits arrondissements ne vont pas être sélectionnés exhaustivement, alors que dans la pratique, ils seront systématiquement enquêtés, en raison notamment de la concentration des enquêteurs dans ces trois villes. Pour pallier cette contrainte pratique, il a été décidé - pour les simulations réalisées - d'agréger les arrondissements de Paris, Lyon et Marseille et de considérer ainsi que chacune de ces trois villes ne forme qu'une seule ZAE. Etant donné l'importance démographique de ces ZAE, elles seront évidemment retenues comme exhaustives par l'algorithme précédent, ce qui assure implicitement la sélection systématique de l'ensemble de leurs arrondissements (et pas seulement les plus importants d'entre eux.)

2.3.2.4. Quelques statistiques sur les ZAE.

Les arrondissements (20 à Paris, 9 à Lyon, 16 à Marseille) ayant été remplacés par 3 ZAE seulement, le nombre de ZAE sur l'ensemble de la France a ainsi été ramené à 3743. On trouvera en annexe 1 la liste des 37 ZAE exhaustives. Le tableau 1 présente un certain nombre de statistiques régionales et nationales concernant les ZAE, obtenues après avoir mis en œuvre l'algorithme de calcul des probabilités d'inclusion. Les deux dernières colonnes, seuil+ et seuil- indiquent respectivement le nombre de résidences principales de la plus petite ZAE exhaustive et de la plus grande ZAE non-exhaustive.

On notera que, conformément à une précédente remarque, le rapport $e/\tau = 40000$ semble bien définir le seuil d'exhaustivité indépendamment de la région considérée, puisque la colonne seuil+ ne fait figurer que des nombres supérieurs à 40000 et la colonne seuil- des nombres inférieurs à 40000. On remarquera également que dans certaines régions le nombre de ZAE à tirer est extrêmement faible, ce qui va limiter significativement les marges de manœuvre, notamment quant au choix des variables d'équilibrage. Les cas de la Corse et du Limousin sont à souligner.

2.3.3. Le choix des variables d'équilibrage.

Le choix des variables d'équilibrage est un problème épineux et c'est essentiellement sur lui que repose la latitude dont on dispose pour faire varier le plan de sondage utilisé. Néanmoins, six variables d'équilibrage sont systématiquement retenues : le nombre de résidences principales *nres* des ZAE ainsi que le nombre de résidences principales *nresgr1, nresgr2, nresgr3, nresgr4, nresgr5* des 5 groupes de rotation des ZAE, permettant d'assurer chaque année du cycle quinquennal une « représentativité » de l'ensemble des logements, qui sont les entités de « base » enquêtées lors des enquêtes-ménages. On peut toutefois remarquer que la première de ces variables est égale à la somme des cinq autres, si bien que, si l'équilibrage est assuré sur 5 de ces 6 variables, il le sera alors automatiquement sur la sixième. Aussi abandonnera-t-on la variable d'équilibrage superflue *nresgr5.*

Par ailleurs, comme les probabilités d'inclusion sont proportionnelles à la variable *nres*, l'équilibrage sur cette variable revient à celui sur les probabilités d'inclusion et permet donc (cf partie 2.1) de tirer dans chaque région des échantillons de ZAE de taille fixe, cette taille figurant dans le tableau 1.

Les variables d'équilibrage supplémentaires seront précisées lors de l'analyse de chaque plan de sondage étudié. Avec le jeu des probabilités d'inclusion obtenues précédemment et l'ensemble des variables retenues, on peut alors tirer aléatoirement des échantillons équilibrés à l'aide de fast-cube.

Région	nombre de ZAE	nombre de ZAE exhaustives	nombre d'enquêteurs à mobiliser pour les ZAE exhaustives	nombre de ZAE non exhaustives à tirer	nombre total d'enquê-teurs à mobiliser	seuil+	seuil-
Ile-de-France	363	2	29	84	113	52333	38214
Champagne-Ardenne	115	1	2	12	14	83262	28128
Picardie	183	1	1	17	18	57593	24661
Haute-Normandie	141	2	3	14	17	54133	21654
Centre	194	2	3	22	25	50689	32311
Basse-Normandie	148	1	1	13	14	54358	18681
Bourgogne	144	1	2	15	17	71334	23321
Nord Pas-De-Calais	235	1	2	35	37	99846	34143
Lorraine	181	2	2	21	23	52981	17209
Alsace	123	2	4	13	17	45926	27831
Franche-Comté	114	1	1	10	11	55159	22671
Pays de la Loire	198	3	7	25	32	66487	29109
Bretagne	188	2	4	26	30	70552	29128
Poitou-Charentes	138	1	1	16	17	42337	37761
Aquitaine	221	1	3	27	30	114133	38911
Midi-Pyrénées	213	1	5	22	27	199430	23175
Limousin	57	1	2	6	8	66271	22922
Rhône-Alpes	363	4	10	47	57	55136	28965
Auvergne	115	1	2	12	14	67612	20396
Languedoc-Roussillon	140	3	6	18	24	49902	31560
PACA	150	4	17	30	47	60880	38319
Corse	19	0	0	3	3	-	22333
Total France	**3743**	**37**	**107**	**488**	**595**	**42337**	**38911**

Tableau 1 : Statistiques descriptives concernant les ZAE et le seuil d'exhaustivité

2.4. Principe des estimations et des calculs empiriques de précision.

On souhaite que, lors du tirage des ZAE du nouvel échantillon-maître, celui-ci soit le plus « représentatif » possible du territoire national. Il faut donc que l'estimation du total d'une variable soit la plus proche possible du vrai total de cette variable sur l'ensemble de la population, et ce pour un maximum de variables d'intérêt pertinentes.

Par ailleurs, la sélection finale du nouvel échantillon-maître demeurant un processus fondamentalement aléatoire, il faut s'assurer que, même si sous le plan de sondage considéré on obtient une bonne « représentativité » en moyenne, le risque de s'éloigner sensiblement de cette dernière lors du tirage final soit minimisé.

Les simulations consistent donc à effectuer, pour un plan de sondage donné, un grand nombre de tirages aléatoires indépendants, à réaliser un certain nombre d'estimations sur chacun de ces tirages puis à analyser la distribution empirique de ces estimations, en termes d'espérance et de variance.

2.4.1. Estimations par groupes de rotations.

2.4.1.1. L'estimateur retenu.

Les premières estimations menées l'ont été en prenant en compte l'année de rotation du cycle quinquennal du nouveau mode de recensement, car le plan de sondage retenu au final se doit d'être bon chaque année - et pas seulement en moyenne sur le cycle quinquennal. Si *X* désigne la variable d'intérêt étudiée, le simple estimateur d'Horvitz-Thompson du total de *X*, à savoir $\sum_{\substack{ZAE \\ tirées}} \left(\frac{X_{ZAE}}{\pi_{ZAE}} \right)$, n'est pas satisfaisant puisque, pour chaque ZAE, il ne prend en compte que la valeur X_{ZAE} agrégée au niveau de la ZAE, et non la valeur prise par *X* lors de l'année de rotation considérée.

Dans l'expression précédente, si l'on s'intéresse à l'année de rotation *i*, le terme X_{ZAE} a donc été remplacé par $\hat{X}_{ZAE} = X_{ZAE,i} \cdot \frac{nres_{ZAE}}{nres_{ZAE,i}}$ où $X_{ZAE,i}$ et $nres_{ZAE,i}$ désignent respectivement la valeur de la variable *X* et le nombre de résidences principales pour le groupe de rotation *i* de la ZAE *ZAE*. En d'autres termes, la vraie valeur de *X* au niveau ZAE étant inconnue une année donnée (puisque l'affectation en groupes de rotation pour le recensement est aléatoire), on la remplace par un estimateur par le ratio basé sur la valeur prise par *X* dans le groupe de rotation *i* de la ZAE. En réalité, $X_{ZAE,i}$ sera lui-même estimé car seul un échantillon de logements tiré aléatoirement au sein de chaque ZAE sélectionnée sera enquêté lors des enquêtes-ménages.

Les estimateurs retenus par année de rotation sont finalement les suivants, pour *i* variant de 1 à 5 :

$$\hat{T}_{X_i} = \sum_{\substack{ZAE \\ tirées}} \left(\frac{X_{ZAE,i} \dfrac{nres_{ZAE}}{nres_{ZAE,i}}}{\pi_{ZAE}} \right)$$

où : - i représente l'année de rotation du cycle quinquennal.

- $X_{ZAE,i}$: modalité de la variable X dans le groupe de rotation i de la ZAE *ZAE*.
- $nres_{ZAE,i}$: nombre de résidences principales du groupe de rotation i de la ZAE *ZAE*.
- $nres_{ZAE}$: nombre de résidences principales de la ZAE *ZAE*.
- π_{ZAE} : probabilité d'inclusion de la ZAE *ZAE*.

Rappelons que les ZAE grandes communes ne possèdent pas de groupe de rotation puisqu'elles font l'objet d'un recensement partiel chaque année. Pour ces ZAEGC, on a donc considéré des groupes de rotation « fictifs », au sein desquels les variables d'intérêt sont égales au cinquième de leur valeur dans l'ensemble de la ZAE. Cela a été fait notamment pour la variable *nres*. Comme $\left(X_{ZAE,i} = X_{ZAE} / 5 \text{ et } nres_{ZAE,i} = nres_{ZAE} / 5 \right)$ implique que $X_{ZAE,i} \frac{nres_{ZAE}}{nres_{ZAE,i}} = X_{ZAE}$, cela ne modifie en rien le principe du calcul des estimations mais permet d'une part de simplifier leur mise en œuvre informatique et d'autre part de définir les variables *nresgr1, nresgr2, nresgr3, nresgr4, nresgr5* pour les ZAEGC, nécessaires à l'équilibrage (cf partie 2.3.). L'estimateur retenu ci-dessus peut donc aussi se réécrire de la sorte :

$$\hat{T}_{X_i} = \sum_{\substack{ZAEGC \\ tirées}} \left(\frac{X_{ZAE}}{\pi_{ZAE}} \right) + \sum_{\substack{ZAEPC \\ tirées}} \left(\frac{X_{ZAE,i} \dfrac{nres_{ZAE}}{nres_{ZAE,i}}}{\pi_{ZAE}} \right) .$$

2.4.1.2. Existence d'un biais.

Si le simple estimateur d'Horvitz-Thompson est par construction sans biais, il n'en est a priori pas de même pour celui qui a été retenu puisque l'on a remplacé la vraie valeur prise par la variable *X* au niveau *ZAE* par une estimation par le ratio. Comme le montre la reformulation de l'estimateur ci-dessus, le biais est en réalité exclusivement engendré par les petites communes.

Néanmoins, si l'on a choisi un tel estimateur, c'est parce qu'on peut raisonnablement penser que les variables d'intérêt étudiées (population, population par tranche d'âge, revenu, etc…) sont fortement corrélées à la variable de repondération *nres*. Plus cette corrélation est importante, plus le biais engendré doit être faible. Le cas extrême est celui où il y aurait parfaite proportionnalité entre *X* et *nres*, ce qui conduirait à annuler le biais. C'est d'ailleurs ce qui produit si pour *X*, on choisit la variable *nres* elle-même.

En définitive, la prise en compte de l'aspect rotatif du recensement conduit au choix d'un estimateur qui présente un biais, que l'on espère le plus faible possible compte tenu de la supposée forte corrélation entre les variables d'intérêt et la variable de repondération. Cette question du biais sera réexaminée dans la partie 3 à la lueur des résultats obtenus.

2.4.1.3. Analyse de la distribution des estimations.

Après avoir effectué un grand nombre N de tirages d'échantillons indépendants (1000 par région en général, sauf indication contraire explicite), on procède donc de la manière suivante : Pour chaque région, année de rotation et variable d'intérêt étudiée, on calcule l'estimateur du total de cette variable à partir de chaque échantillon tiré. On obtient donc N estimations ; sont alors déterminés la moyenne empirique de ces N estimations, leur coefficient de variation (rapport de l'écart-type empirique à la moyenne empirique) ainsi que le biais empirique relatif (écart relatif entre la moyenne empirique des estimations et la vraie valeur du total de la variable).

Dans un deuxième temps, par agrégation des résultats régionaux, on obtient des estimations de même nature, mais au niveau national, sur lesquelles on effectue les mêmes analyses en moyenne et variance.

2.4.2. Estimations sans distinction des groupes de rotation.

La dispersion des estimations s'avère relativement faible dès le premier plan de sondage testé (cf partie 3) et le biais empirique évolue peu d'une année de rotation à l'autre. Ceci incite dans un deuxième temps à faire abstraction de l'aspect rotatif du recensement, en moyennant avec des poids identiques et à chaque tirage les estimations des totaux dans chaque groupe de rotation :

$$\hat{T}_X = \frac{1}{5} \sum_{i=1}^{5} \hat{T}_{X_i} .$$

On détermine à nouveau la moyenne empirique des N estimations obtenues, le coefficient de variation de leur distribution ainsi que le biais empirique relatif, tout ceci dans chaque région puis au niveau national.

2.4.3. Evolutions temporelles des estimations.

La question du comportement du plan de sondage face aux évolutions temporelles de la base de sondage est essentielle. Une façon d'y répondre est de tester sa qualité sur des données des recensements nationaux antérieurs à 1999. Aussi pour certaines variables, essentiellement démographiques, a-t-on pu récupérer des données des recensements de 1999, 1990, 1982, 1975, 1968 et 1962. Un travail identique à celui mené dans les deux précédents paragraphes a été entrepris sur ces données.

2.4.4. Estimations par type d'espace.

Un autre aspect majeur de la représentativité du territoire national réside dans la répartition des unités tirées par type d'espace. L'abandon du respect du zonage ZAUER (découpage en aires urbaines) pour constituer les ZAE peut faire craindre un comportement erratique des estimateurs et implique la nécessité d'introduire des variables caractérisant le type d'espace parmi les variables d'équilibrage.

Un indicateur d'urbanisation issu du recensement de 1999, intitulé POL99, a été retenu pour chaque commune. Les 36613 communes de France ont alors été réparties en trois types d'espace, urbain, périurbain et rural selon le principe suivant :

- Espace urbain : communes pour lesquelles POL99 vaut 1 (« communes appartenant à un pôle urbain »).
- Espace périurbain : communes pour lesquelles POL99 vaut 2 (« communes monopolarisées, appartenant à une couronne périurbaine ») ou vaut 3 (« communes multipolarisées »).
- Espace rural : communes pour lesquelles POL99 vaut 4 (« Espace à dominante rurale »).

On notera que ce ne sont pas les ZAE qui sont ainsi affectées à un type d'espace mais les communes elles-mêmes, si bien qu'une ZAE peut se retrouver « à cheval » sur deux espaces distincts (c'est le cas pour 1733 ZAE) voire sur les trois espaces (307 ZAE).

Pour chaque tirage réalisé, on a estimé le total des variables d'intérêt par type d'espace, en retenant cette fois-ci l'estimateur de Horvitz-Thompson « simple » qui, dès le départ, ne prend pas en compte l'aspect rotatif du recensement :

$$\hat{T}_{X\,urbain} = \sum_{\substack{ZAE \\ tirées}} \frac{X_{ZAE,urbain}}{\pi_{ZAE}}$$

où $X_{ZAE,urbain}$ est la somme des modalités prises par la variable *X* dans les communes de la ZAE *ZAE* qui sont affectées à l'espace urbain (Expressions similaires pour le périurbain et le rural).

Ce choix se justifie par le besoin de vérifier si les craintes relatives à un possible comportement erratique des estimateurs sont fondées ou non. Il est avant tout nécessaire de s'assurer que l'on peut obtenir des comportements acceptables de la part des estimateurs « simples ». Comme on le verra dans la troisième partie, il est déjà assez complexe de trouver un plan de sondage qui produise de bons résultats par type d'espace, ce qui valide a posteriori le choix de l'estimateur qui a été fait ici. Celui-ci permet en outre un fort gain en temps de calcul.

3. Comparaison empirique des divers plans de sondage testés lors des simulations.

A présent, les principaux résultats obtenus sont présentés et analysés sous les divers angles proposés dans la partie 2 de cette présentation. Le propos sera essentiellement axé sur la comparaison entre les plans de sondage qui ont été testés et on tâchera ainsi de mettre en exergue les raisons invoquées pour les choix successifs de nouvelles variables d'équilibrage.

3.1. Les variables d'intérêt.

3.1.1. Quelques mots sur la base de sondage.

Pour l'étude, les variables utilisées sont celles du dernier recensement national exhaustif disponible (celui de 1999) - à l'exception notable des données fiscales. Au final, treize variables ont été retenues pour leur pertinence vis-à-vis des travaux effectués, fournissant des données générales démographiques, fiscales, d'état civil… Ces variables, sur lesquelles sont effectuées les estimations, sont les suivantes :

- Population sans double compte 1999,
- Nombre de résidences principales 1999,
- Population des résidences principales 1999,
- Nombre de résidences principales grandes communes 1999,
- Nombre de décès entre les recensements de 1990 et 1999,
- Nombre de naissances entre les recensements de 1990 et 1999,
- Nombre d'individus de [0 ;19] ans 1999,
- Nombre d'individus de [20 ;59] ans 1999,
- Nombre d'individus de + 60 ans 1999,
- Revenu net imposable 1996,
- Revenu fiscal 2004,
- Nombre de ménages fiscaux 2004,
- Nombre de personnes dans les ménages fiscaux 2004.

Les données ayant été initialement fournies par communes, les valeurs prises par ces variables dans les groupes de rotation des ZAE ont été simplement calculées par agrégation des données de leurs communes constitutives.

Par ailleurs, la base de sondage recensait les communes de France selon le code officiel géographique en vigueur en 1999, ce qui a posé quelques difficultés puisque la constitution des ZAE s'est opérée sur la base du code officiel géographique en vigueur au 1er janvier 2006. Quelques modifications ont ainsi dû lui être apportées au cas par cas afin de la rendre compatible avec ce dernier : le cas des communes ayant fusionné entre 1999 et 2006 a été traité par simple agrégation des données datées de 1999. En revanche, le cas des communes s'étant scindées depuis 1999 est plus délicat : la répartition du nombre des résidences principales de l'ancienne commune sur les nouvelles communes étant connue, les valeurs prises par les autres variables quantitatives ont été réparties de la même façon sur les nouvelles communes. Il en résulte que les données dont on dispose donc sur ces dernières sont des estimations et non des vraies valeurs. Néanmoins, il faut souligner que les communes concernées sont en général de petites (voire très petites) communes et qu'elles sont en nombre très restreint : Une trentaine sur l'ensemble de la France (qui - rappelons-le - possède plus de 36000 communes). L'impact de ces approximations sur les résultats des travaux est donc parfaitement négligeable.

3.1.2. Les cinq plans de sondage testés.

La spécification des ZAE exhaustives, du nombre d'enquêteurs à leur affecter et le calcul des probabilités d'inclusion ont été décrits dans la partie 2. Le degré de liberté quant à la détermination du plan de sondage à tester se situe essentiellement au niveau du choix des variables d'équilibrage. On a eu l'occasion de souligner le caractère primordial dans l'équilibrage du nombre de résidences principales par ZAE et par groupe de rotation. L'ajout de contraintes supplémentaires ou la modification de contraintes déjà présentes répond majoritairement au besoin d'améliorer les estimations par type d'espace.

Le premier plan de sondage testé est le plus simple possible, il se contente d'équilibrer, par priorité décroissante, sur le nombre de résidences principales des ZAE et sur le nombre de résidences principales dans chacun des quatre premiers groupes de rotation. Le deuxième plan de sondage consiste à rajouter le nombre de résidences principales grandes communes comme sixième variable d'équilibrage (variable égale au nombre de résidences principales dans les ZAEGC et à 0 dans les ZAEPC). Dans le troisième plan, on ajoute également le revenu fiscal 2004 décliné par groupe de

rotation. Enfin, dans le quatrième, le nombre de résidences principales grandes communes est remplacé par deux autres variables, nombre de résidences des ZAE en zone rurale d'une part, en zone périurbaine d'autre part. Le cinquième et dernier plan de sondage est une variante du précédent, les deux dernières variables citées étant dépriorisées dans les trois régions les plus urbanisées, à savoir l'Ile-de-France, le Nord-Pas-de-Calais et Provence-Alpes-Côte-d'-Azur. Le tableau 2 qui suit résume les variables d'équilibrage des cinq plans de sondage, hiérarchisées selon l'ordre de priorité décroissant. Cet ordre est celui suivi par l'algorithme fast-cube pour relâcher les contraintes d'équilibrage lors de la phase d'atterrissage.

Variables d'équilibrage	**Numéro du Plan de sondage**					
	N° 1	**N° 2**	**N° 3**	**N° 4**	**N° 5**	
Nombre de résidences principales 99	1	1	1	1	1	
Nombre de résidences principales dans le groupe de rotation 1	2	2	2	2	2	
Nombre de résidences principales dans le groupe de rotation 2	3	3	3	3	3	
Nombre de résidences principales dans le groupe de rotation 3	4	4	4	4	4	
Nombre de résidences principales dans le groupe de rotation 4	5	5	5	5	5	
Nombre de résidences grandes communes	-	6	6	-	-	
Nombre de résidences en zone rurale	-	-	-	6	6*	11**
Nombre de résidences en zone périurbaine	-	-	-	7	7*	12**
Revenu fiscal 2004 dans le groupe de rotation 1	-	-	7	8	8*	6**
Revenu fiscal 2004 dans le groupe de rotation 2	-	-	8	9	9*	7**
Revenu fiscal 2004 dans le groupe de rotation 3	-	-	9	10	10*	8**
Revenu fiscal 2004 dans le groupe de rotation 4	-	-	10	11	11*	9**
Revenu fiscal 2004 dans le groupe de rotation 5	-	-	11	12	12*	10**
Total du nombre de variables d'équilibrage	**5**	**6**	**11**	**12**	**12**	

* toutes régions sauf Ile-de-France, Nord-Pas-de-Calais et PACA
** régions Ile-de-France, Nord-Pas-de-Calais et PACA

Tableau 2 : Listes des variables d'équilibrage pour les plans de sondage testés, ordonnées par priorité décroissante.

Précisons enfin que, comme la part de la variance issue des données des ZAE exhaustives est nulle – par définition même du caractère exhaustif – les ZAE exhaustives ne sont jamais prises en compte lors de l'équilibrage et dans les calculs d'estimations ou de vraie valeur d'une quelconque variable d'intérêt. On peut donc garder à l'esprit que les résultats en termes de dispersion relative des estimations (coefficients de variation) seraient légèrement améliorés par la prise en compte des ZAE exhaustives. On a choisi délibérément d'omettre ces dernières dans le calcul des estimations afin de mieux faire ressortir les qualités et défauts des plans de sondage testés.

On trouvera en annexe 2 la valeur des totaux des variables étudiées sur l'ensemble des ZAE exhaustives et la part de ces totaux dans les totaux calculés sur l'ensemble des ZAE du territoire national. On pourra retenir que pour ces variables, les ZAE exhaustives représentent environ 15% de l'ensemble des ZAE, avec néanmoins un « pic » prévisible à 35% pour la variable Nombre de résidences grandes communes.

3.2. Analyse des résultats avec et sans distinction des groupes de rotation.

Pour chaque plan de sondage adopté, 1000 simulations de tirages aléatoires de ZAE par région ont été réalisées. Les résultats de l'analyse en moyenne et variance des estimations nationales avec et sans distinction des groupes de rotation, dont les principes théoriques ont été exposés aux paragraphes 2.4.1. et 2.4.2., figurent dans le tableau 3. **Dans les colonnes intitulées « biais » et « cv » apparaissent respectivement les biais empiriques relatifs et les coefficients de variation empiriques, exprimés en pourcentages.** Sur fond clair sont distinguées les variables qui ont servi à l'équilibrage.

Biais (relatif) et cv en %

Groupe de rotation	plan de sondage N°1		plan de sondage N°2		plan de sondage N°3		plan de sondage N°4		plan de sondage N°5	
	biais	cv	biais	cv	biais	cv	biais	cv	biais	cv
Population sans double compte au RP99										
1	0,689	**0,363**	0,689	**0,335**	0,696	**0,297**	0,683	**0,311**	0,702	**0,324**
2	0,565	**0,372**	0,576	**0,349**	0,557	**0,313**	0,566	**0,320**	0,577	**0,317**
3	0,530	**0,381**	0,548	**0,354**	0,524	**0,322**	0,538	**0,321**	0,539	**0,336**
4	0,690	**0,362**	0,692	**0,346**	0,696	**0,312**	0,700	**0,324**	0,702	**0,307**
5	0,832	**0,391**	0,845	**0,352**	0,826	**0,329**	0,850	**0,340**	0,846	**0,349**
SANS	*0,661*	***0,327***	*0,670*	***0,295***	*0,660*	***0,260***	*0,667*	***0,270***	*0,673*	***0,275***
Nombre de résidences principales 99										
1	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**
2	-0,027	**0,072**	-0,024	**0,068**	-0,027	**0,072**	-0,023	**0,067**	-0,030	**0,075**
3	-0,031	**0,077**	-0,030	**0,075**	-0,030	**0,076**	-0,029	**0,075**	-0,030	**0,077**
4	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**
5	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**	0,000	**0,000**
SANS	*-0,011*	***0,021***	*-0,011*	***0,021***	*-0,011*	***0,022***	*-0,010*	***0,021***	*-0,012*	***0,022***
Population des résidences principales au RP99										
1	0,878	**0,370**	0,880	**0,337**	0,885	**0,294**	0,877	**0,313**	0,891	**0,320**
2	0,800	**0,379**	0,811	**0,353**	0,793	**0,317**	0,802	**0,321**	0,815	**0,319**
3	0,721	**0,385**	0,733	**0,351**	0,710	**0,317**	0,727	**0,320**	0,729	**0,332**
4	0,824	**0,370**	0,832	**0,341**	0,841	**0,313**	0,841	**0,316**	0,849	**0,307**
5	0,940	**0,387**	0,949	**0,334**	0,932	**0,316**	0,950	**0,324**	0,948	**0,329**
SANS	*0,832*	***0,338***	*0,841*	***0,296***	*0,832*	***0,264***	*0,840*	***0,272***	*0,846*	***0,276***
Nombre de résidences grandes communes										
1	-	-	0,047	**0,925**	-0,017	**0,924**	0,056	**2,637**	-0,032	**2,616**
2	-	-	0,047	**0,925**	-0,017	**0,924**	0,056	**2,637**	-0,032	**2,616**
3	-	-	0,047	**0,925**	-0,017	**0,924**	0,056	**2,637**	-0,032	**2,616**
4	-	-	0,047	**0,925**	-0,017	**0,924**	0,056	**2,637**	-0,032	**2,616**
5	-	-	0,047	**0,925**	-0,017	**0,924**	0,056	**2,637**	-0,032	**2,616**
SANS	-	**-**	*0,047*	***0,925***	*-0,017*	***0,924***	*0,056*	***2,637***	*-0,032*	***2,616***
Nombre de décès entre le RP90 et le RP99										
1	-1,974	**1,274**	-1,968	**1,262**	-2,015	**1,143**	-2,006	**1,114**	-1,958	**1,106**
2	-1,359	**1,405**	-1,354	**1,401**	-1,392	**1,292**	-1,418	**1,257**	-1,384	**1,255**
3	-1,173	**1,322**	-1,173	**1,339**	-1,161	**1,192**	-1,246	**1,164**	-1,167	**1,146**
4	-1,024	**1,305**	-1,094	**1,258**	-1,078	**1,177**	-1,114	**1,153**	-1,094	**1,155**
5	-0,712	**1,389**	-0,763	**1,369**	-0,665	**1,242**	-0,702	**1,268**	-0,695	**1,202**
SANS	*-1,248*	***0,968***	*-1,270*	***0,943***	*-1,262*	***0,783***	*-1,297*	***0,770***	*-1,260*	***0,742***
Nombre de naissances entre le RP90 et le RP99										
1	-0,608	**0,807**	-0,620	**0,756**	-0,615	**0,725**	-0,606	**0,731**	-0,599	**0,739**
2	-0,764	**0,818**	-0,808	**0,758**	-0,814	**0,726**	-0,792	**0,713**	-0,785	**0,718**
3	-0,645	**0,836**	-0,694	**0,779**	-0,707	**0,727**	-0,655	**0,726**	-0,627	**0,735**
4	-0,802	**0,846**	-0,831	**0,752**	-0,837	**0,723**	-0,822	**0,742**	-0,809	**0,727**
5	-0,399	**0,846**	-0,420	**0,778**	-0,445	**0,748**	-0,404	**0,753**	-0,394	**0,750**
SANS	*-0,644*	***0,717***	*-0,675*	***0,642***	*-0,684*	***0,600***	*-0,656*	***0,606***	*-0,643*	***0,604***

Biais (relatif) et cv en %

Groupe de rotation	plan de sondage N°1		plan de sondage N°2		plan de sondage N°3		plan de sondage N°4		plan de sondage N°5	
	biais	cv	biais	cv	biais	cv	biais	cv	biais	cv
Nombre d'individus de f0,19[ans au RP99										
1	1,326	**0,783**	1,324	**0,786**	1,334	**0,716**	1,318	**0,728**	1,349	**0,743**
2	1,046	**0,811**	1,052	**0,806**	1,020	**0,734**	1,037	**0,720**	1,067	**0,710**
3	0,963	**0,828**	0,975	**0,831**	0,927	**0,741**	0,974	**0,722**	0,982	**0,739**
4	1,035	**0,806**	1,046	**0,806**	1,058	**0,740**	1,062	**0,743**	1,074	**0,714**
5	1,352	**0,824**	1,367	**0,782**	1,322	**0,740**	1,359	**0,728**	1,366	**0,738**
SANS	*1,145*	***0,698***	*1,153*	***0,685***	*1,132*	***0,613***	*1,150*	***0,605***	*1,168*	***0,605***
Nombre d'individus de [20,59[ans au RP99										
1	0,963	**0,494**	0,959	**0,485**	0,969	**0,397**	0,950	**0,416**	0,971	**0,417**
2	0,768	**0,518**	0,780	**0,505**	0,749	**0,438**	0,768	**0,434**	0,775	**0,424**
3	0,711	**0,514**	0,743	**0,515**	0,700	**0,439**	0,728	**0,427**	0,716	**0,439**
4	0,937	**0,498**	0,950	**0,496**	0,944	**0,422**	0,954	**0,435**	0,954	**0,415**
5	1,071	**0,537**	1,098	**0,509**	1,056	**0,458**	1,088	**0,458**	1,084	**0,481**
SANS	*0,890*	***0,444***	*0,906*	***0,426***	*0,883*	***0,350***	*0,897*	***0,352***	*0,900*	***0,356***
Nombre d'individus de + 60 ans au RP99										
1	-0,715	**0,818**	-0,703	**0,818**	-0,703	**0,712**	-0,702	**0,716**	-0,700	**0,669**
2	-0,496	**0,871**	-0,479	**0,848**	-0,455	**0,762**	-0,472	**0,729**	-0,483	**0,710**
3	-0,412	**0,837**	-0,423	**0,836**	-0,366	**0,744**	-0,430	**0,711**	-0,404	**0,687**
4	-0,353	**0,850**	-0,380	**0,820**	-0,361	**0,735**	-0,373	**0,738**	-0,375	**0,713**
5	-0,384	**0,872**	-0,408	**0,834**	-0,336	**0,765**	-0,352	**0,716**	-0,365	**0,721**
SANS	*-0,472*	***0,718***	*-0,479*	***0,687***	*-0,444*	***0,588***	*-0,466*	***0,563***	*-0,466*	***0,534***
Revenu net imposable de l'année 1996										
1	0,304	**0,881**	0,322	**0,848**	0,302	**0,573**	0,335	**0,562**	0,332	**0,546**
2	0,158	**0,898**	0,201	**0,857**	0,200	**0,555**	0,248	**0,542**	0,209	**0,545**
3	0,236	**0,853**	0,323	**0,835**	0,291	**0,567**	0,296	**0,523**	0,287	**0,556**
4	0,525	**0,914**	0,570	**0,874**	0,576	**0,601**	0,584	**0,579**	0,590	**0,581**
5	0,315	**0,846**	0,409	**0,822**	0,346	**0,550**	0,389	**0,535**	0,393	**0,546**
SANS	*0,308*	***0,771***	*0,365*	***0,738***	*0,343*	***0,421***	*0,370*	***0,397***	*0,362*	***0,399***
revenu fiscal 2004										
1	1,374	**0,911**	1,392	**0,820**	1,391	**0,422**	1,407	**0,438**	1,397	**0,432**
2	1,452	**0,929**	1,499	**0,880**	1,473	**0,455**	1,538	**0,467**	1,496	**0,459**
3	1,388	**0,879**	1,475	**0,857**	1,437	**0,454**	1,459	**0,449**	1,435	**0,441**
4	1,401	**0,915**	1,465	**0,840**	1,464	**0,456**	1,488	**0,457**	1,481	**0,459**
5	1,380	**0,892**	1,464	**0,832**	1,411	**0,452**	1,447	**0,441**	1,441	**0,444**
SANS	*1,399*	***0,814***	*1,459*	***0,746***	*1,435*	***0,307***	*1,468*	***0,311***	*1,450*	***0,306***
Nombre de ménages fiscaux en 2004										
1	0,361	**0,250**	0,357	**0,238**	0,360	**0,206**	0,356	**0,210**	0,362	**0,215**
2	0,363	**0,268**	0,361	**0,241**	0,359	**0,213**	0,371	**0,222**	0,359	**0,224**
3	0,331	**0,258**	0,337	**0,245**	0,339	**0,221**	0,343	**0,226**	0,341	**0,225**
4	0,243	**0,243**	0,251	**0,228**	0,250	**0,210**	0,250	**0,205**	0,257	**0,204**
5	0,322	**0,267**	0,314	**0,256**	0,322	**0,217**	0,318	**0,225**	0,329	**0,222**
SANS	*0,324*	***0,199***	*0,324*	***0,178***	*0,326*	***0,148***	*0,328*	***0,151***	*0,330*	***0,148***
Nombre de personnes dans les ménages fiscaux en 2004										
1	1,376	**0,527**	1,372	**0,466**	1,387	**0,393**	1,376	**0,402**	1,385	**0,416**
2	1,345	**0,537**	1,351	**0,481**	1,335	**0,401**	1,362	**0,424**	1,356	**0,416**
3	1,234	**0,525**	1,249	**0,472**	1,237	**0,399**	1,265	**0,405**	1,250	**0,415**
4	1,133	**0,516**	1,149	**0,470**	1,162	**0,403**	1,160	**0,415**	1,177	**0,397**
5	1,297	**0,534**	1,289	**0,470**	1,288	**0,404**	1,296	**0,404**	1,298	**0,421**
SANS	*1,277*	***0,459***	*1,282*	***0,392***	*1,282*	***0,319***	*1,292*	***0,328***	*1,293*	***0,332***

Tableau 3: Analyse en moyenne et variance de la distribution empirique des estimations nationales avec et sans distinction des groupes de rotation (calculées sur la base de 1000 tirages indépendants dans chaque région)

3.2.1. Quelques remarques générales.

La première observation peut être conduite sur la variable fondamentale *nres*, nombre de résidences principales, pour laquelle on observe systématiquement un biais et une dispersion parfaitement nuls dans trois des cinq groupes de rotation, non nuls mais très faibles dans les deux autres. Cette variable est proportionnelle aux probabilités d'inclusion, si bien que, si l'on note *nbenq* le nombre d'enquêteurs à mobiliser pour les ZAE non-exhaustives, c'est-à-dire le nombre de ZAE non-exhaustives à tirer, on a $\pi_{ZAE} = nbenq \times \frac{nres_{ZAE}}{\sum_{\substack{toutes \\ ZAE}} nres_{ZAE}}$. L'équilibrage sur *nres* assurant le tirage d'échantillons de taille fixe égale à *nbenq*, il résulte que l'estimateur retenu doit a priori coïncider avec la vraie valeur puisque l'on obtient, pour *i=1,2,3,4,5* :

$$\hat{T}_{nres_i} = \sum_{\substack{ZAE \\ tirées}} \left(\frac{nres_{ZAE,i} \dfrac{nres_{ZAE}}{nres_{ZAE,i}}}{\pi_{ZAE}} \right) = \left(\sum_{\substack{toutes \\ ZAE}} nres_{ZAE} \right) \left(\frac{1}{nbenq} \sum_{\substack{ZAE \\ tirées}} 1 \right) = \sum_{\substack{toutes \\ ZAE}} nres_{ZAE}$$

Ceci permet d'expliquer la nullité des valeurs observées dans les groupes de rotation 1, 4 et 5. Pour expliquer les valeurs a priori inattendues des groupes de rotation 2 et 3, rappelons l'existence de quatre ZAE pour lesquelles il manque un groupe de rotation. Lorsqu'une telle ZAE est tirée, sa contribution à l'estimation lors de l'année de rotation où elle n'a pas de groupe est nulle au lieu d'être égale à son nombre de résidences principales et l'estimation qui en résulte est en conséquence légèrement inférieure à la vraie valeur. D'où l'existence d'un faible biais empirique négatif. Il a d'ailleurs été vérifié scrupuleusement que les estimations ne coïncidant pas parfaitement avec la vraie valeur du total sont exactement celles correspondant aux échantillons contenant au moins l'une de ces quatre ZAE particulières.

Concernant le nombre *nresgc* de résidences principales en grandes communes, cette variable a été introduite dans l'équilibrage afin d'améliorer les résultats par type d'espace observés sur le premier plan de sondage, comme on le verra par la suite. La forte disparité des valeurs qu'elle prend (on rappelle que cette variable est nulle dans les ZAEPC et égale à *nres* dans les ZAEGC) est sans doute à l'origine d'une dispersion relativement importante des estimations, avec des coefficients de variation proches encore du pourcent lorsque cette variable sert à l'équilibrage, et de l'ordre de 2,6% lorsque ce n'est plus le cas. La réécriture de l'estimateur retenu, mentionnée au paragraphe 2.4.1., permet de constater que :

$$\hat{T}_{nresgc_i} = \sum_{\substack{ZAEGC \\ tirées}} \left(\frac{nresgc_{ZAE}}{\pi_{ZAE}} \right) = \sum_{\substack{ZAE \\ tirées}} \left(\frac{nresgc_{ZAE}}{\pi_{ZAE}} \right)$$

Celui-ci doit donc être sans biais (Horvitz-Thompson) et indépendant du groupe de rotation. C'est bien ce que l'on semble observer dans le tableau 3.

3.2.2. Le rôle des variables fiscales dans l'équilibrage.

Concernant le rôle des variables fiscales, le revenu fiscal 2004 constitue un indicateur social de la population et a donc été introduit dans l'équilibrage à partir du troisième plan de sondage. On peut malheureusement constater que son influence sur la précision des estimations est relativement faible. Certes les coefficients de variation de l'ensemble des variables d'intérêt subissent une légère diminution, mais celle-ci est trop faible (guère plus de 0,1%) pour être vraiment intéressante. Tout juste observe-t-on qu'ils sont diminués de moitié pour la variable revenu fiscal 2004 elle-même et du tiers pour le revenu net imposable 1996, dont on pouvait s'attendre à ce qu'il soit fortement corrélé à cette dernière.

Il est nécessaire de garder à l'esprit que le revenu fiscal 2004 n'intervient qu'au mieux comme septième variable d'équilibrage et que les tirages ont lieu par région. Dans les petites régions (tout particulièrement la Corse) où le nombre de ZAE n'est que de quelques dizaines, l'équilibrage obtenu ne peut pas être très bon en général et risque même d'être médiocre sur les variables à faible priorité. On peut se demander si cela n'est pas en partie à l'origine de la relative stabilité des résultats obtenus d'un plan de sondage à l'autre.

Dans l'ensemble, les résultats sont malgré tout assez satisfaisants en termes de dispersion, les coefficients de variation étant de l'ordre de quelques dixièmes de pourcent pour les variables étudiées, avec toutefois un bémol à apporter sur le nombre de décès entre 1990 et 1999. On notera en outre que pour une variable donnée et pour un plan de sondage donné, biais et dispersion varient peu d'un groupe de rotation à l'autre, tendant à valider une bonne homogénéité au niveau national entre les cinq groupes de rotation.

3.2.3. Le caractère biaisé des estimations

Lors de la description de l'estimateur retenu, on a mentionné que celui-ci présente a priori l'inconvénient d'être légèrement biaisé et que cela devrait être compensé par la corrélation plus ou moins importante existant entre les variables d'intérêt et la variable de repondération *nres*. Dans le tableau 3, on observe effectivement des biais empiriques de l'ordre de quelques dixièmes de pourcent, voire du pourcent, qui persistent au gré des différents plans de sondage. Afin d'étudier plus précisément cette question, on a choisi une variable présentant un fort biais, en l'occurrence le nombre d'individus de [0,19] ans, sur laquelle ont à nouveau été effectuées des estimations à partir de 1000 nouveaux échantillons, puis 10000 nouveaux échantillons (tirés sous le deuxième plan de sondage). Les résultats sont consignés dans le tableau 4 ci-dessous.

année de rotation	nombre de tirages	moyenne des totaux estimés	total réel	Biais relatif (%)	Cv (%)	student
nombre d'individus de [0, 19] ans au RP 99						
1	1 000	12646905,785	12484503	**1,301**	0,842	0,482
	10 000	12649745,768	12484503	**1,324**	0,820	1,592
2	1 000	12614547,826	12484503	**1,042**	0,811	0,402
	10 000	12614474,430	12484503	**1,041**	0,822	1,254
3	1 000	12598144,217	12484503	**0,910**	0,851	0,335
	10 000	12602030,018	12484503	**0,941**	0,833	1,119
4	1 000	12614468,050	12484503	**1,041**	0,846	0,385
	10 000	12615759,748	12484503	**1,051**	0,830	1,254
5	1 000	12651868,678	12484503	**1,341**	0,824	0,507
	10 000	12652981,952	12484503	**1,350**	0,822	1,621

Tableau 4 : Analyse du biais empirique par augmentation du nombre de simulations

La dernière colonne fait apparaître la valeur de la statistique de Student calculée pour tester la nullité du biais. On remarque qu'à 95%, l'hypothèse nulle n'est jamais rejetée, que l'on fasse 1000 ou 10000 simulations. Néanmoins, en ayant multiplié par 10 le nombre de tirages réalisés, le biais empirique n'évolue pas, du moins ne subit-il aucune diminution sensible. Cela tend à prouver son existence, déjà justifiée théoriquement. En supposant que le biais empirique observé ne dépend pas du nombre de simulations effectuées comme le tableau 4 le laisse penser, et en considérant le fait que la dispersion des estimations n'a aucune raison d'évoluer en augmentant le nombre de tirages (c'est d'ailleurs ce que l'on observe sur les coefficients de variation présents dans le tableau 4), on peut s'attendre à ce que l'hypothèse nulle finisse par être systématiquement rejetée, à 95%, si l'on élève encore davantage le nombre de simulations (avec 40000 simulations par exemple, on devrait obtenir des valeurs des statistiques de Student environ deux fois plus élevées que celles obtenues

avec 10000 simulations). Ceci n'a pu être entrepris, le coût en temps de calcul devenant trop important (avec 40000 simulations, ce sont déjà près d'un million de tirages qui doivent être effectués à cause du principe de stratification régionale). On retiendra donc que le caractère biaisé des estimations apparaît dans l'analyse des résultats, qu'il demeure relativement faible mais néanmoins problématique. On trouvera en annexe 3 une étude supplémentaire sur la question du choix d'un estimateur présentant un léger biais.

3.3. Analyse des résultats sur les données des recensements antérieurs.

On s'intéresse ici à l'évolution de la dispersion des estimateurs du total d'une même variable lorsqu'on considère les données des recensements antérieurs à 1999. Sur la base de 1000 tirages indépendants réalisés dans chaque région sous le premier plan de sondage, les estimations avec distinction des groupes de rotation, puis sans distinction, ont été calculées pour chaque recensement depuis celui de 1962. Dans le tableau 5 ci-dessous sont exposés les résultats, à l'échelle nationale, des estimations sans distinction des groupes de rotation, pour les cinq variables d'intérêt pour lesquelles on dispose des données sur les recensements antérieurs. (**Biais relatif et cv sont toujours exprimés en pourcentages**)

RP	Population sans double compte		Nombre de résidences principales		Population des résidences principales	
	biais	cv	biais	cv	biais	cv
1962	1,559	**1,349**	1,353	**1,419**	1,746	**1,370**
1968	0,395	**1,104**	0,331	**1,186**	0,568	**1,122**
1975	-0,522	**0,776**	-0,556	**0,851**	-0,359	**0,787**
1982	-0,185	**0,481**	-0,450	**0,524**	-0,028	**0,486**
1990	0,380	**0,334**	-0,188	**0,248**	0,529	**0,339**
1999	0,661	**0,327**	-0,011	**0,021**	0,832	**0,338**

RP	Nombre de décès entre RP cité et RP précédent*		Nombre de naissances entre RP cité et RP précédent*	
	biais	cv	biais	cv
1962	3,343	**1,835**	2,119	**1,470**
1968	2,247	**1,643**	-0,127	**1,207**
1975	1,441	**1,484**	-2,129	**1,165**
1982	0,487	**1,322**	-2,696	**1,071**
1990	-0,242	**1,160**	-1,368	**0,839**
1999	-1,248	**0,968**	-0,644	**0,717**

* : le recensement précédant celui de 1962 a eu lieu en 1954.

Tableau 5 : Evolutions temporelles des estimations, analyse en moyenne et variance (calculs sur la base de 1000 tirages indépendants par région, sous le 1er plan de sondage)

Le plan de sondage semble se dégrader en considérant des données de plus en plus anciennes, ce qui se traduit par une augmentation des coefficients de variation. Toutefois, ces derniers restent relativement faibles, ne dépassant guère plus de 0,5% pour les variables démographiques et 1,3% pour les variables d'état civil sur les trois derniers recensements, ces seuils ne s'élevant qu'à 1,4% et 1,8% si l'on remonte jusqu'en 1962.

Ces observations suggèrent la persistance d'un bon comportement des plans de sondage dans le temps et sont d'autant plus satisfaisants qu'ils ont été obtenus avec le plan de sondage le plus simple qui soit. Il a donc été décidé de ne pas réitérer ce genre d'étude sur les autres plans de sondage.

3.4. Analyse des résultats par type d'espace.

La question de la représentativité des différents types d'espace a été mentionnée parmi les enjeux majeurs qui sous-tendent au choix d'un bon plan de sondage. A nouveau, pour chaque plan, 1000 tirages par région ont été entrepris sur lesquels ont été effectuées les estimations par type d'espace, rural périurbain et urbain. Le procédé théorique a été décrit dans le paragraphe 2.4.4. Les résultats de l'analyse en moyenne et variance sont consignés dans le tableau 6. (**Biais relatif et cv en %,** variables ayant servi à l'équilibrage sur fond clair.)

La forme des estimateurs qui ont été retenus, ne faisant plus intervenir la repondération via la variable *nres*, doit conduire théoriquement à l'absence de biais. Les observations faites sur le tableau 6 tendent à valider ce constat, les biais empiriques s'avérant extrêmement faibles dans l'ensemble.

3.4.1. Remarque sur les grandes communes.

Les communes étant considérées comme « grandes communes » dès lors qu'elles possèdent plus de 10000 habitants, la très grande majorité d'entre elles figurent donc logiquement dans l'espace urbain. Cependant, il existe quelques grandes communes appartenant à l'espace périurbain et quelques-unes également appartenant à l'espace rural. Les estimations du nombre de résidences grandes communes dans ces deux espaces varient donc très fortement selon que l'une de ces communes fait partie d'une des ZAE tirées ou non. Cela explique les très forts coefficients de variation observés pour cette variable en rural et en périurbain ; ces résultats sont donc justifiés mais inexploitables.

3.4.2. L'ajout des variables d'équilibrage pour réduire la dispersion des estimations.

Manifestement, les craintes évoquées dans la deuxième partie de cette présentation suite à l'abandon du zonage ZAUER dans la constitution des ZAE étaient fondées, puisque la qualité des estimations s'avère relativement médiocre pour le premier plan de sondage. Les coefficients de variation sont de l'ordre de 3% dans l'espace urbain, de 6% dans les deux autres quand ils étaient de quelques dixièmes de pourcent seulement sans distinction du type d'espace (cf paragraphe 3.2.). Aussi les divers choix de variables d'équilibrage se sont-ils essentiellement faits avec pour objectif d'améliorer ce point précis.

C'est pour cette raison notamment qu'a été introduite la variable nombre de résidences grandes communes qui présente l'avantage de ne pas varier selon l'année de rotation (deuxième plan de sondage). Le fait que les ZAE soient tirées avec des probabilités proportionnelles à leur taille implique que les ZAE à dominante urbaine sont plus souvent sélectionnées que les autres. Cela apporte un élément d'explication au fait que la dispersion, même médiocre, soit plus faible dans l'urbain que dans le périurbain et le rural. Malgré tout, on peut observer que le deuxième plan de sondage n'est guère meilleur que le premier, les coefficients ayant légèrement diminué, mais demeurant insatisfaisants.

L'ajout des variables fiscales dans l'équilibrage a également été décidé dans la perspective de réduire cet épineux problème. On peut en effet légitimement supposer que les gros revenus - par exemple - soient plutôt concentrés dans les grandes villes et par conséquent dans l'espace urbain. Les résultats obtenus à ce titre pour le troisième plan de sondage montrent à nouveau une légère amélioration, davantage sensible dans le rural et le périurbain que dans l'urbain, mais encore une fois largement insuffisante. On conserve toujours des coefficients de l'ordre de 2%, 4,5% et 5% dans les trois espaces considérés.

Face à ces observations et à la faible utilité constatée de l'équilibrage sur le nombre de résidences grandes communes, variable pourtant supposée intrinsèquement liée à la notion de type d'espace, il a été décidé de remplacer cette dernière par deux autres variables, le nombre de résidences principales en zone rurale et le nombre de résidences principales en zone périurbaine. C'est l'objet du quatrième plan de sondage. Par une remarque similaire à celle faite au paragraphe 2.3.3., l'équilibrage sur ces deux variables ainsi que sur la variable *nres* assure automatiquement celui

type d'espace	plan de sondage N°1		plan de sondage N°2		plan de sondage N°3		plan de sondage N°4		plan de sondage N°5	
	biais	cv	biais	cv	biais	cv	biais	cv	biais	cv
Population sans double compte au RP99										
urbain	0,030	**3,050**	0,111	**2,181**	0,028	**2,099**	0,039	**0,899**	0,001	**0,904**
périurbain	-0,080	**6,017**	-0,348	**5,147**	-0,171	**4,503**	0,015	**1,824**	0,001	**1,865**
rural	0,011	**6,155**	0,141	**5,854**	0,093	**5,008**	-0,103	**2,127**	0,037	**2,060**
Nombre de résidences principales 99										
urbain	0,040	**3,051**	0,095	**2,059**	0,027	**2,009**	0,038	**0,793**	-0,011	**0,791**
périurbain	-0,084	**5,989**	-0,336	**5,117**	-0,165	**4,524**	0,011	**1,789**	0,003	**1,804**
rural	-0,010	**6,162**	0,116	**5,819**	0,106	**4,956**	-0,107	**2,093**	0,024	**2,022**
Population des résidences principales au RP99										
urbain	0,033	**3,044**	0,108	**2,188**	0,028	**2,096**	0,036	**0,902**	0,002	**0,909**
périurbain	-0,083	**6,004**	-0,342	**5,139**	-0,169	**4,494**	0,016	**1,824**	0,000	**1,859**
rural	0,006	**6,147**	0,135	**5,851**	0,092	**5,003**	-0,098	**2,116**	0,034	**2,053**
Nombre de résidences grandes communes										
urbain	-	-	0,053	**1,450**	0,019	**1,432**	0,065	**2,485**	0,003	**2,514**
périurbain	-	-	-1,353	**40,151**	-1,954	**43,788**	-1,309	**40,585**	-1,365	**39,921**
rural	-	-	3,341	**74,372**	0,928	**73,581**	3,025	**74,045**	-0,772	**77,476**
Nombre de décès entre le RP90 et le RP99										
urbain	0,036	**3,318**	0,119	**2,343**	0,079	**2,381**	0,069	**1,316**	0,002	**1,322**
périurbain	-0,057	**6,420**	-0,290	**5,456**	-0,192	**4,992**	0,022	**2,405**	0,070	**2,478**
rural	0,054	**6,361**	0,094	**5,960**	0,103	**4,930**	-0,177	**2,521**	0,030	**2,404**
Nombre de naissances entre le RP90 et le RP99										
urbain	0,053	**3,141**	0,064	**2,102**	0,019	**2,048**	0,012	**1,071**	-0,004	**1,085**
périurbain	-0,077	**6,221**	-0,401	**5,404**	-0,224	**4,804**	-0,001	**2,083**	-0,005	**2,098**
rural	0,004	**6,425**	0,207	**6,162**	0,034	**5,419**	-0,037	**2,558**	0,090	**2,534**

Biais (relatif) et cv en %

type d'espace	plan de sondage N°1		plan de sondage N°2		plan de sondage N°3		plan de sondage N°4		plan de sondage N°5	
	biais	cv	biais	cv	biais	cv	biais	cv	biais	cv
Nombre d'individus de [0,19] ans au RP99										
urbain	0,036	**3,138**	0,114	**2,369**	0,022	**2,257**	0,020	**1,199**	0,017	**1,206**
périurbain	-0,095	**6,094**	-0,368	**5,255**	-0,179	**4,566**	0,009	**1,993**	-0,012	**2,024**
rural	0,036	**6,278**	0,194	**6,041**	0,056	**5,245**	-0,061	**2,406**	0,065	**2,343**
Nombre d'individus de [20,59] ans au RP99										
urbain	0,031	**3,041**	0,116	**2,191**	0,018	**2,090**	0,032	**0,935**	-0,003	**0,929**
périurbain	-0,083	**6,043**	-0,358	**5,172**	-0,171	**4,514**	0,008	**1,889**	-0,014	**1,921**
rural	0,011	**6,197**	0,165	**5,923**	0,084	**5,123**	-0,093	**2,217**	0,050	**2,176**
Nombre d'individus de + 60 ans au RP99										
urbain	0,020	**3,239**	0,090	**2,308**	0,065	**2,309**	0,083	**1,165**	-0,007	**1,176**
périurbain	-0,055	**6,093**	-0,292	**5,212**	-0,158	**4,651**	0,042	**2,028**	0,061	**2,097**
rural	-0,008	**6,239**	0,059	**5,822**	0,137	**4,874**	-0,152	**2,248**	-0,007	**2,170**
Revenu net imposable de l'année 1996										
urbain	-0,046	**3,099**	0,125	**2,344**	0,022	**2,046**	0,042	**0,996**	0,019	**0,998**
périurbain	-0,122	**6,251**	-0,370	**5,165**	-0,165	**4,500**	0,021	**2,032**	-0,020	**2,088**
rural	-0,018	**6,205**	0,123	**5,923**	0,073	**5,206**	-0,106	**2,286**	-0,019	**2,208**
revenu fiscal 2004										
urbain	-0,045	**3,109**	0,149	**2,431**	0,035	**2,084**	0,048	**0,979**	0,010	**0,976**
périurbain	-0,086	**6,161**	-0,369	**5,146**	-0,169	**4,447**	0,037	**2,025**	-0,029	**2,050**
rural	-0,024	**6,130**	0,128	**5,848**	0,069	**5,201**	-0,076	**2,229**	0,018	**2,175**
Nombre de ménages fiscaux en 2004										
urbain	0,037	**3,043**	0,104	**2,113**	0,042	**2,036**	0,040	**0,830**	-0,006	**0,819**
périurbain	-0,073	**5,992**	-0,342	**5,148**	-0,171	**4,514**	0,018	**1,818**	0,010	**1,835**
rural	-0,007	**6,134**	0,121	**5,793**	0,093	**5,001**	-0,094	**2,070**	0,029	**2,020**
Nombre de personnes dans les ménages fiscaux en 2004										
urbain	0,026	**3,024**	0,109	**2,222**	0,049	**2,116**	0,040	**0,938**	-0,003	**0,928**
périurbain	-0,070	**6,012**	-0,344	**5,165**	-0,183	**4,487**	0,031	**1,863**	0,008	**1,901**
rural	0,007	**6,117**	0,135	**5,821**	0,086	**5,048**	-0,089	**2,114**	0,038	**2,063**

Tableau 6: Analyse en moyenne et variance de la distribution empirique des estimations nationales par type d'espace (calculées sur la base de 1000 tirages indépendants dans chaque région)

Biais (relatif) et cv en %

sur le nombre de résidences principales en zone urbaine. Cette fois-ci, on note une très nette amélioration de la précision des estimations pour l'ensemble des variables d'intérêt étudiées, avec des coefficients de variation inférieurs au pourcent dans l'espace urbain et avoisinant les 2% ailleurs. Les variables d'état civil, en particulier le nombre de décès, continuent de présenter un degré de précision légèrement moins bon que celui observé pour les autres variables.

Le cinquième plan de sondage est en réalité une variante du précédent, qui envisage une modification de l'ordre de priorité dans le relâchement des contraintes d'équilibrage. Dans les régions les plus urbanisées et par conséquent les plus « grosses », on peut raisonnablement espérer que le nombre de ZAE à sélectionner soit suffisamment important pour que des contraintes d'équilibrage classées onzièmes ou douzièmes dans l'ordre de priorité jouent un véritable rôle sans être systématiquement relâchées. C'est la raison pour laquelle on a choisi de déprioriser les variables d'équilibrage nombre de résidences principales en espace rural et en espace périurbain, dans les trois régions où l'espace urbain représente plus de 75% des logements (Ile-de-France, Nord-Pas-de-Calais, PACA). Le tableau 10 figurant en annexe 4 donne pour chaque région la part occupée par les trois types d'espaces en termes de nombre de logements.

Les résultats obtenus n'évoluent guère par rapport au plan de sondage précédent, que ce soit pour les trois régions en question ou pour les estimations nationales. On observe en effet pour les coefficients de variation des valeurs semblables à celles résultant du quatrième plan de sondage.

On retiendra donc essentiellement la nécessité d'équilibrer sur le nombre de résidences principales en zone rurale et en zone périurbaine pour obtenir des estimations qui ne présentent pas de dispersion insatisfaisante. La bonne représentativité des différents types d'espace semble de toute façon passer par là. Ajoutons également que la qualité des estimations pourra encore être davantage accrue en utilisant des procédures de calage a posteriori.

Conclusion

Moyennant un certain nombre d'ajustements concernant en particulier la détermination des ZAE exhaustives et les cas spécifiques des arrondissements de grandes villes, diverses stratégies de tirage aléatoire des ZAE ont été mises en œuvre et comparées en termes de représentativité et de précision, en jouant sur le choix des variables servant à l'équilibrage des échantillons.

Il ressort des simulations que l'estimateur retenu pour le total des variables d'intérêt présente certes un très léger biais mais aussi une variance empirique particulièrement faible dès le premier plan de sondage testé. Par ailleurs, ce bon comportement semble résister aux évolutions de la base de sondage dans un laps de temps raisonnable. Quant à la spécification du type d'espace des zones enquêtées, la qualité de la précision s'est avérée beaucoup plus complexe à obtenir ; retenons essentiellement que la construction des ZAE n'ayant pas pris en compte le critère du zonage ZAUER, il semble incontournable d'introduire des variables caractéristiques des types d'espace dans l'équilibrage pour obtenir des résultats satisfaisants. Un calage sur ce type de variables permettrait aussi sans doute d'améliorer la précision des estimations a posteriori. Quant à l'ajout d'un équilibrage sur des variables fiscales, il semble également apporter un gain significatif en précision.

Les résultats obtenus ont ainsi pu être mis à profit pour choisir de manière définitive le plan de sondage final du nouvel échantillon-maître. Cette étude a par la suite été poursuivie par Emmanuel Gros pour réaliser, toujours dans un double objectif de représentativité et de précision, le tirage des logements à enquêter au sein des ZAE sélectionnées et pour résoudre le problème de la sélection conjointe des ZAE destinées à l'échantillon-maître et de celles destinées aux extensions régionales.

Annexe 1

Variable	Total sur les ZAE exhaustives	Total national sur l'ensemble des ZAE	Part du total des ZAE exhaustives dans le total national (en %)
population sans double compte au RP99	8 980 723	58 518 395	15,35
Nombre de résidences principales 99	4 329 856	23 809 194	18,19
population des résidences principales au RP99	8 695 333	57 217 451	15,20
Nombre de résidences grandes communes	4 329 856	12 365 887	35,01
Nombre de décès entre le RP90 et le RP99	714 729	4 743 282	15,07
Nombre de naissances entre le RP90 et le RP99	1 128 463	6 606 463	17,08
Nombre d'individus de [0, 19] ans au RP 99	1 896 937	14 381 440	13,19
Nombre d'individus de [20, 59] ans au RP 99	5 286 207	31 661 121	16,70
nombre d'individus de + 60 ans au RP 99	1 797 428	12 478 127	14,40
revenu net imposable de l'année 1996	501 747 570	2 730 738 611	18,37
revenu fiscal 2004	124 631 621 580	737 515 347 721	16,90
nombre de ménages fiscaux en 2004	4 049 531	24 222 850	16,72
nombre de personnes dans les ménages fiscaux en 2004	8 388 039	58 208 535	14,41

Tableau 7 : Total des variables d'intérêt sur les ZAE exhaustives et part de ces totaux dans les totaux nationaux (calculés sur l'ensemble de toutes les ZAE).

Annexe 2

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

Tableau 8 : Liste des 37 ZAE exhaustives

Annexe 3

Dans cette annexe, on s'intéresse à la méthode de repondération utilisée dans l'estimateur retenu pour le total des variables d'intérêt. A priori, la probabilité que l'ensemble des communes d'un groupe de rotation d'une ZAEPC soit enquêté une année donnée, conditionnellement au fait que la ZAEPC en question soit elle-même sélectionnée, est exactement d'un cinquième. Il en résulte donc qu'un estimateur sans biais (estimateur d'Horvitz-Thompson) du total des variables étudiées pourrait être :

$$\hat{T}_{X_i} = \sum_{\substack{ZAE \\ tirées}} \left(\frac{5 \cdot X_{ZAE,i}}{\pi_{ZAE}} \right) = \sum_{\substack{ZAEGC \\ tirées}} \left(\frac{X_{ZAE}}{\pi_{ZAE}} \right) + \sum_{\substack{ZAEPC \\ tirées}} \left(\frac{5 \cdot X_{ZAE,i}}{\pi_{ZAE}} \right) .$$

(i variant de 1 à 5)

Cependant un tel estimateur risque de présenter une dispersion plus importante, puisqu'il ne tient pas compte de la différence des poids réels des groupes de rotation au sein des ZAE, que l'on a choisi de mesurer en termes de nombre de logements principaux.

On a choisi de retenir trois variables pour effectuer quelques analyses supplémentaires, dont les résultats sont fournis dans le tableau 9 ci-dessous. Sur un jeu de 1000 échantillons par région tirés sous le quatrième plan de sondage, on a estimé le total de ces trois variables à l'aide de l'estimateur biaisé, détaillé au paragraphe 2.4.1. (résultats dans les colonnes « estimateur 1 » du tableau 9).

Le même travail a alors été effectué en utilisant l'estimateur non biaisé précisé ci-dessus (résultats dans les colonnes « estimateur 2 » du tableau 9).

Enfin, on a calculé le rapport *r* de l'erreur quadratique de l' « estimateur 1 » sur l'erreur quadratique de l' « estimateur 2 » : $r = \frac{B_1^2 + \sigma_1^2}{B_2^2 + \sigma_2^2}$ où B_k et σ_k sont respectivement le biais empirique absolu et l'écart-type empirique de l'estimateur *k*.

année de rotation	estimateur 1 (biaisé)		Estimateur 2 (non biaisé)		rapport de l'erreur quadratique de l'estimateur 1 à celle de l'estimateur 2
	biais (%)	cv (%)	biais (%)	cv (%)	
Population sans double compte au RP99					
1	0,683	**0,311**	0,078	**0,729**	1,046
2	0,566	**0,320**	-0,057	**0,802**	0,658
3	0,538	**0,321**	-0,191	**0,746**	0,666
4	0,700	**0,324**	0,015	**0,836**	0,852
5	0,850	**0,340**	0,044	**0,798**	1,315
Nombre de naissances entre le RP90 et le RP99					
1	-0,606	**0,731**	0,127	**1,006**	0,869
2	-0,792	**0,713**	-0,064	**1,040**	1,041
3	-0,655	**0,726**	-0,196	**0,978**	0,957
4	-0,822	**0,742**	-0,126	**1,108**	0,982
5	-0,404	**0,753**	0,071	**1,072**	0,628
Revenu net imposable 1996					
1	0,335	**0,562**	-0,048	**0,853**	0,589
2	0,248	**0,542**	0,066	**0,884**	0,453
3	0,296	**0,523**	-0,088	**0,848**	0,501
4	0,584	**0,579**	0,181	**0,917**	0,777
5	0,389	**0,535**	-0,044	**0,842**	0,618

Tableau 9 : Comparaison de l'estimateur retenu avec l'estimateur sans biais (résultats nationaux)

L'estimateur n°2 présente des biais empiriques relatifs bien plus faibles mais des coefficients de variation bien plus élevés que l'estimateur n°1. En observant la dernière colonne du tableau précédent, il apparaitrait toutefois que, pour les trois variables étudiées ici, l'erreur quadratique soit dans l'ensemble plus faible dans le premier cas, même si ce n'est pas toujours très « prononcé ». Le choix, pour les simulations, de l'estimateur utilisant une repondération par la variable nombre de résidences principales semble donc pertinent.

Annexe 4

région	nombre de résidences principales	proportion de résidences principales en aire rurale (%)	proportion de résidences principales en aire périurbaine (%)	proportion de résidences principales en aire urbaine (%)
Ile-de-France	4510369	0,07	9,89	90,04
Champagne-Ardenne	540024	28,50	20,35	51,15
Picardie	700971	22,85	38,02	39,12
Haute-Normandie	698563	10,44	29,26	60,30
Centre	999962	28,19	23,25	48,55
Basse-Normandie	572019	35,11	24,09	40,80
Bourgogne	670956	33,07	23,47	43,46
Nord Pas-De-Calais	1491693	5,03	17,50	77,47
Lorraine	908678	16,84	25,37	57,79
Alsace	678837	6,67	34,21	59,12
Franche-Comté	452124	25,12	28,90	45,98
Pays de la Loire	1292740	29,18	20,13	50,70
Bretagne	1209664	28,56	25,33	46,11
Poitou-Charentes	686209	37,92	21,39	40,69
Aquitaine	1212578	29,29	13,61	57,10
Midi-Pyrénées	1070770	32,78	15,44	51,79
Limousin	311495	38,74	17,94	43,33
Rhône-Alpes	2274059	13,88	20,00	66,12
Auvergne	556291	34,77	21,66	43,57
Languedoc-Roussillon	968654	29,13	22,71	48,16
Provence Alpes Côte d'Azur	1896302	8,34	9,82	81,84
Corse	106236	41,40	16,74	41,86
Total France	**23809194**	**18,00**	**18,93**	**63,07**

Tableau 10 : Proportions des logements dans les différents types d'espace.

Bibliographie

[1] Brossier, G., Dussaix, A.-M., « Enquêtes et sondages, Méthodes, modèles, applications, nouvelles approches », *Dunod*, Liège, août 1999.

[2] Chauvet, G., Tillé, Y., « A fast algorithm of balanced sampling », Computational Statistics, vol 21, pp 53-61, 2006.

[3] Clairin, R., Brion, P., « Manuel de sondages, applications aux pays en développement », *CEPED*, Paris, Documents et manuels du CEPED n°3, novembre 1997.

[4] Deville, J.-C., Tillé, Y., « Efficient balanced sampling : the cube method », *Biometrika*, vol 91, pp 893-912, 2004.

[5] Rousseau, S., Tardieu, F., « La macro SAS CUBE d'échantillonage équilibré », *INSEE*, *méthodologie statistique* n°0402, documentation de l'utilisateur.

[6] Tillé, Y., « Théorie des sondages, échantillonnage et estimation en populations finies », *Dunod*, Paris, avril 2001.

[7] Wilms, L., « Présentation de l'échantillon-maître 1999 et application au tirage des unités primaires par la macro CUBE », INSEE, Journées de méthodologie statistique.

[8] « Actes des journées de méthodologie statistique des 4 et 5 décembre 2000 », Tome 1, *INSEE METHODES* n°100, Paris, novembre 2002.