L'Echantillon-Maître 99 (EM99) et Application au tirage des unités primaires par la macro « Cube »

Laurent WILMS (INSEE, Unité Méthodes Statistiques)

Introduction

Ce papier a pour objectif d'expliquer le système de tirage de la plupart des échantillons d'enquêtes nationales auprès des ménages menées par l'*INSEE*. Nous insisterons plus particulièrement sur la constitution de l'Echantillon-Maître (EM), étape préalable au tirage proprement dit d'une enquête et qui est réalisée après chaque recensement de la population.

Le nouvel Echantillon-Maître (EM99) devra être opérationnel fin octobre 2001 afin de pouvoir y tirer l'enquête Logement dont la collecte est prévue de décembre 2001 à février 2002.

Précisons d'ores et déjà que le ménage est identifié au logement qu'il occupe habituellement. In fine, la sélection des ménages s'effectue selon une **méthode d'échantillonnage probabiliste** des logements. Cette méthode doit à la fois tenir compte de contraintes pratiques (assurer une stabilité géographique du réseau d'enquêteurs INSEE et minimiser les coûts de déplacement) et garantir une précision statistique suffisante.

Ce qui suit peut également être vu comme une information sur une méthodologie générale dont on peut s'inspirer pour réaliser n'importe quelle enquête auprès des ménages à partir d'une base de sondage. D'ailleurs, des missions de coopération en ce sens sont en cours avec des pays tels que la Roumanie, l'Algérie et prochainement la République Tchèque.

Enfin, ceci ne concerne pas les extensions régionales d'enquêtes nationales ni les enquêtes locales qui ne sont traditionnellement pas tirées dans l'Echantillon-Maître et pour lesquelles des outils spécifiques devront être constitués. Nous ne détaillerons pas non plus ici le suivi des logements neufs complémentaire à la constitution de l'Echantillon-Maître.

1. Qu'est ce qu'un Echantillon-Maître?

1.1. Définition d'un Echantillon-Maître

C'est une réserve de logements destinée à alimenter la plupart des enquêtes-ménages nationales de l'INSEE[1] entre deux recensements de la population (RP). Cette réserve est donc constituée à l'issue de chaque recensement et contient initialement seulement des logements du dernier RP.

Afin que cette base de logements offre au cours du temps une bonne couverture des ménages, elle doit être continuellement enrichie par un échantillon des logements construits après le dernier RP (on parle alors de *logements neufs*). Cependant, cette précaution apportée par l'adjonction de logements neufs, ne permet pas de lutter contre le vieillissement de la base Échantillon-Maître des logements RP. Certains de ces logements sont détruits ou changent de statut, ce qui n'assure plus une couverture correcte des ménages[2]. Dans ces conditions, on comprend la nécessité d'une remise à jour complète de l'Echantillon-Maître après chaque RP.

1.2 Entre contraintes pratiques et désir de précision statistique

La construction de l'Echantillon-Maître prend d'abord en compte le mode d'organisation de la collecte des données auprès des ménages. Plus précisément, la collecte est organisée à partir d'un réseau d'enquêteurs ayant une certaine stabilité, renouvelé en partie après chaque recensement. Cette collecte se déroulant la plupart du temps en face à face, il devient alors nécessaire de concentrer géographiquement les zones d'enquêtes, si l'on souhaite obtenir la maîtrise des coûts de déplacement des enquêteurs.

En fonction de cette contrainte organisationnelle forte, il s'agit de sélectionner un Echantillon-Maître qui garantisse une bonne précision des enquêtes nationales. L'Echantillon-Maître ne possède malheureusement pas de bonnes vertus en termes de précision régionale. On pourra se reporter au 2.3.1. pour une explication.

[1] L'enquête Emploi constituant l'exception majeure. D'autres enquêtes nationales telles HID (Handicap-Incapacités et Dépendances) et les enquêtes de nature purement locale sont également tirées hors EM.

[2] Rappelons que la plupart des enquêtes visent les logements principaux.

1.3. Dimensionnement et délai de mise en service de l'Echantillon-Maître

La réserve Échantillon-Maître est traditionnellement calibrée ou dimensionnée en fonction du nombre d'enquêtes prévues entre deux recensements et du nombre moyen de ménages interrogés par enquête. Ainsi, l'Echantillon-Maître de 1990 tablait sur une période intercensitaire de 10 ans (en réalité 9) et prévoyait un nombre total de 100 enquêtes nationales (soit une moyenne de 10 par année). Enfin, la taille moyenne d'une enquête était évaluée à 10 000 ménages.

Pour l'Echantillon-Maître de 1999, son calibrage est effectué selon des prévisions analogues, le volume global des enquêtes nationales restant vraisemblablement stable. Mais l'évaluation de la durée de vie de cet Échantillon-Maître ne peut plus se faire en fonction de la date du prochain RP puisque ce dernier devrait être remplacé par le Recensement Rénové de la Population (RRP). D'ailleurs, le RRP alimentera un Échantillon-Maître « nouvelle génération ». De ce fait, la durée de vie de l'Echantillon-Maître de 1999 a été estimée à 6 ans (d'octobre 2001 à la fin de l'année 2007).

Enfin, notons que la mise en service de l'Echantillon-Maître est effective environ deux ans après le dernier RP[3]. Les raisons sont multiples :

- raisons techniques externes (mise à disposition des résultats du RP)
- raisons techniques internes (définition de la méthodologie, constitution par les DR des potentielles futures zones d'enquêtes, préparation de la chaîne de tirage des logements)
- raisons juridiques (contrôle CNIL, appel d'offre pour le matériel d'impression des fiches-adresses).

[3] Les RP, depuis la création de l'INSEE, ont eu lieu en 1946, 1954, 1962, 1968, 1975, 1982, 1990 et 1999.

Quelques chiffres

en 1990 :

-- 26 237 000 logements (principaux, secondaires, vacants et occasionnels) sont recensés.

-- l'Echantillon-Maître contient 1 859 000 logements, soit 7,1% des logements.

-- l'Echantillon-Maître a été disponible en 1992.

en 1999 :

-- 28 696 156 logements (principaux, secondaires, vacants et occasionnels) sont recensés.

-- l'Echantillon-Maître contient 5% des logements environ[4].

-- l'Echantillon-Maître sera disponible en 2001.

Rappel :

- on prévoit traditionnellement 10 enquêtes nationales par an en moyenne.

- une enquête standard touche environ 10 000 ménages

(soit un taux de sondage moyen de $\frac{1}{2000}$).

[4] Le chiffre exact sera connu vers la mi-janvier 2001 après tirage des districts.

2. Constitution de l'Echantillon-Maître

L'Echantillon-Maître est initialement tiré dans les logements du RP99.

Rappelons que cet Échantillon-Maître a pour objectifs prioritaires :

- de fournir une réserve suffisante de logements
- d'assurer une certaine concentration géographique des enquêtes, compatible avec une localisation du réseau d'enquêteurs (sur période intercensitaire ou jusqu'à la date de première disponibilité du RRP).
- d'assurer, pour les enquêtes nationales, une précision acceptable.

2.1. Démarche générale

Première étape

La localisation du réseau étant une contrainte incontournable, la première étape est donc de définir une partition de la France métropolitaine (les DOM n'entrant pas dans le champ des enquêtes de type Échantillon-Maître). Ces parties seront très exactement des unités urbaines (UU) ou des regroupements de communes rurales contiguës[5]. Ainsi, aucune zone du territoire national n'est exclue.

Ces parties ou unités seront classées dans quatre strates, dites de gestion (SG), correspondant au degré d'urbanisation des unités :

SG0 = unités rurales (regroupements de communes rurales)

SG1 = UU de moins de 20.000 habitants ou associations de telles UU

SG2 = UU de 20.000 à 100.000 habitants

SG3 = UU de plus de 100.000 habitants (hors UU de Paris)

SG4 = UU de Paris

Pour parler simplement, cette stratification est judicieuse si l'on admet que le degré d'urbanisation est un facteur explicatif important des comportements des ménages couverts par les enquêtes INSEE. On peut penser à des enquêtes telles que « Conditions de vie des ménages (PCV) », « Loisir » ou encore « Emploi du temps ».

[5] A chaque recensement, on affecte un code TU (taille de l'unité urbaine) aux communes. Si ce code vaut 0, la commune est dite rurale. Les autres codes, variant de 1 à 8, indiquent que la commune appartient à une UU dont la taille (en termes de tranche) est donnée par la valeur de ce code.

Les SG 3 et 4 sont les seules strates exhaustives. Autrement dit, une enquête nationale ayant pour champ l'ensemble des ménages de la métropole impactera forcément toutes les UU de plus de 100 000 habitants mais n'impactera qu'un échantillon des UU des SG1 et SG2 ainsi que des unités rurales.

Deuxième étape

Une solution naturelle de constitution de l'Echantillon-Maître serait de retenir l'ensemble des logements dans chacune des unités sélectionnées lors de la première étape ; on éviterait ainsi tout risque d'épuisement. Une enquête ménages serait alors une enquête à un degré dans la SG3 et la SG4 et à deux degrés dans les autres.

Pourtant, des raisons à la fois techniques et historiques[6] nous conduisent à ajouter un degré de tirage supplémentaire lors de la constitution de l'Echantillon-Maître. Ce degré porte précisément sur le tirage de groupes de districts dans les SG2, SG3 et SG4.

Ainsi le tirage de (groupes de) districts dans ces strates nous permet d'abord de réduire sensiblement le nombre de logements, donc de données à stocker dans la base Échantillon-Maître. Ensuite, la chaîne automatisée de tirage de logements pour une enquête, développée pour l'Echantillon-Maître de 1990 et réutilisée pour l'Echantillon-Maître de 1999, n'a pas été conçue pour un tirage direct dans les UU. Par ailleurs, on aurait tort de croire que l'ajout de ce niveau « district » nuit à la précision des enquêtes. En effet, l'effet de grappe lié aux districts est très fortement réduit, voire annulé, puisqu'il est prévu que la chaîne ne tire pas plus de deux logements par groupe de districts.

A l'issue de ces étapes, l'Echantillon-Maître initial est définitivement constitué. Il sera enrichi, au cours du temps, par l'adjonction de logements neufs.

En résumé, l'Echantillon-Maître de 1999 résulte d'un tirage stratifié à un ou deux degrés.

D'abord, nous allons décrire la constitution des unités de première étape des SG0 et SG1, seules strates où l'unité ne correspond pas exactement à une définition administrative. Ensuite, nous décrirons le mode de tirage des unités de première étape ou Unités Primaires (UP) des SG0, SG1 et SG2.

[6] Lors des précédents Échantillons-Maîtres, pour obtenir l'adresse des personnes à enquêter, il fallait la récupérer dans les bulletins-papier du RP qui étaient rangés par district ; elle était ensuite recopiée à la main sur la fiche de l'enquêté. Pour l'Échantillon-Maître de 1990, cette procédure a été entièrement automatisée après scannérisation des adresses des logements de la base Échantillon-Maître. Pour l'Échantillon-Maître99, les adresses seront stockées sous forme de base-images.

2.2. Constitution des Unités Primaires (UP)

2.2.1 Définition d'une UP en SG0

A l'issue de chaque RP, les communes sont classées selon un code TU traduisant leur degré d'urbanisation. Les communes de TU=0 sont des communes dites rurales. Elles ont souvent peu d'habitants, peuvent être parfois assez étendues géographiquement et comporter des zones difficiles d'accès.

L'UP qui sera constituée à partir de ces communes doit répondre à deux impératifs.

Premier impératif : un enquêteur et un seul est associé à une UP (localisation du réseau).

D'où :

• *règle 1* : l'UP doit nécessairement être un regroupement de communes rurales contiguës[7].

Deuxième impératif : sachant que,

- à une UP est affecté un enquêteur

- une charge maximale de 30 Fiches-Adresses (FA) par enquêteur est préconisée pour chaque enquête nationale

- le nombre maximum d'enquêtes à prévoir sur la période de vie de l'Echantillon-Maître est de 60 (10 par an sur 6 ans)

- un logement ne peut être enquêté plus d'une fois,

$\Rightarrow$ une UP contient, au moins, 60*30 = 1800 logements.

• *règle 2* : l'UP doit contenir un nombre suffisant de logements, soit au moins 1800 logements principaux.

[7] Regroupement réalisé en pratique sur la base des cantons.

Pour la définition des UP rurales, aux règles de contiguïté et de nombre minimal de logements, on ajoute les deux suivantes :

• *règle 3* : le nombre de logements par UP ne doit pas excéder 3600 et si possible se rapprocher de 1800.

• *règle 4* : l'UP doit être entièrement incluse dans une région.

La règle 3 a été élaborée à partir du mode de tirage des UP que nous décrivons en détail au paragraphe 2. 3. Il s'agit d'une volonté de limiter la dispersion des tailles des UP tout en assurant un nombre d'UP « important », ce qui est toujours conseillé dans un tirage à deux degrés avec premier degré proportionnel à la taille.

La règle 4 découle du mode de gestion régionale des UP.

2.2.2 Constitution pratique d'une UP en SG0

La France métropolitaine est constituée à la date du RP99 de 36.565 communes dont environ 30.000 rurales. Par le passé, les UP rurales étaient constituées à partir de cartes-papier au sein des 22 Directions Régionales en regroupant les communes sur la base des cantons. Pour l'EM99, un outil cartographique a été spécialement développé pour aider les DR dans cette tâche (cf. la contribution « Utilisation de la cartographie en vue de la constitution d'échantillons » Georges BOURDALLE, VII ^{èmes} JMS).

2.2.3 Constitution d'une UP en SG1

Les règles de constitution de ces UP sont similaires à celle des UP rurales.

• l'UP est une unité urbaine ou une association d'unités urbaines.

• l'UP doit comporter au moins 1800 logements principaux.

• les unités urbaines regroupées sont géographiquement proches et dans la même région.

Pour la description de la constitution pratique, on se réfèrera à la contribution aux VII ^{èmes} JMS, « Utilisation de la cartographie en vue de la constitution d'échantillons », précédemment citée.

2.2.4 Définition d'une UP en SG2

Nous rappelons simplement que les UP sont exactement les UU de 20.000 à 100.000 habitants.

2.3. Tirage des UP des SG0, SG1 et SG2

Les UP sont, avant tirage, non seulement classées selon leur SG, mais également selon leur région d'appartenance.

On obtient ainsi :

22 régions X 3 premiers degrés d'urbanisation (SG0, SG1,SG3) = 66 strates.

La stratification régionale permet essentiellement de répartir équitablement l'effort de collecte sur chaque région.

2.3.1 Nombre d'UP à tirer par strate

La réponse sur le nombre d'UP à tirer par strate découle de deux contraintes pratiques, à savoir :

- le taux de sondage moyen pour une enquête nationale standard, soit $\frac{1}{2000}$.
- le nombre moyen de fiches-adresse (FA) par enquêteur pour une enquête nationale standard, fixé à environ 23.

Donnons un exemple de calcul. Supposons qu'une région contienne 300.000 logements en SG0, alors on détermine *m,* nombre d'UP à tirer ou d'enquêteurs à recruter, par :

$$m = \frac{300\ 000}{2\ 000} \times \frac{1}{23} \approx 6$$

Pour l'Echantillon-Maître de 1999, le nombre d'UP (rurales) tirées par région est en moyenne de 6 parmi 100. On comprend alors aisément pourquoi l'Echantillon-Maître ne peut posséder de bonnes vertus régionales.

Le tableau suivant montre l'influence du nombre de FA par enquêteur (ou UP) sur le nombre d'UP à tirer, donc sur l'organisation du réseau.

Influence du nombre de FA/UP à enquêter sur le nombre d'UP à tirer

	NOMBRE DE FICHES ADRESSES PAR UP			
	30	**25**	**20**	**15**
Strate 0	92	110	137	184
Strate 1	66	79	98	132
Strate 2	55	66	82	110

Lecture : en strate de gestion 0 (toutes régions confondues), pour 30 FA/UP, il faut tirer 92 UP (Source RP99).

Finalement, l'Echantillon-Maître de 1999 comporte 128 UP en SG0, 75 en SG1 et 93 en SG2.

2.3.2 Tirage équilibré des UP[8]

Ayant déterminé l'allocation au sein de chacune des 66 strates définies, on effectue un tirage à probabilités inégales des UP tel que chaque UP ait une probabilité de sélection proportionnelle à son nombre de résidences principales, nombre évalué lors du RP99. Rappelons que ce type de tirage fournit un bon estimateur d'un total (ou d'une moyenne) dès lors que la variable d'intérêt agrégée au niveau UP est bien corrélée avec la taille de l'UP, ou que les moyennes par UP varient peu d'une UP à l'autre, ce qui est souvent le cas pour les enquêtes INSEE[9].

[8] Réalisé avec la macro SAS CUBE.

[9] Par exemple, supposons que l'on veuille estimer le nombre total de télévisions détenues par les ménages. Ce nombre évalué (ou agrégé) au niveau UP est très certainement fortement corrélé au nombre de résidences principales de l'UP.

De plus, le tirage bénéficie d'un équilibrage « super-régional »[10] sur quatre variables auxiliaires, au sein de chaque SG. Ces quatre variables sont :

- revenu net imposable (sources fiscales, année 1996)
- effectif de la tranche d'âge [0,19] (RP99)
- effectif de la tranche d'âge [20,59] (RP99)
- effectif de la tranche d'âge [60, +] (RP99)

A titre d'exemple :

Soit ***s*** l'échantillon d'UP rurales tirées dans l'ensemble des régions composant la Super-Région1 (notée ***SR1***). La propriété d'équilibrage est :

$$\sum_s \frac{X_i}{P_i} = \sum_{i \in SG0 \cap SR1} X_i$$

P_i désigne la probabilité de sélection de l'UP_i
X_i désigne, au niveau de l'UP_i, soit l'effectif de l'une des trois tranches d'âge, soit le revenu net imposable total.

Ainsi, l'estimateur d'Horvitz-Thomson ou CUBE $\sum_s \frac{X_i}{P_i}$ découlant du tirage possède une propriété de calage naturel. Contrairement à une procédure de calage réalisée après tirage (type CALMAR) qui génère un estimateur calé biaisé, l'estimateur d'Horvitz-Thomson équilibré reste sans biais.

Précisons que la propriété d'équilibrage est en réalité difficile à respecter rigoureusement. On peut montrer que plus le nombre de variables d'équilibrage (supposées non colinéaires) est grand et plus la taille de l'échantillon est petite, plus alors cette propriété sera difficile à obtenir (la variance de l'estimateur d'Horvitz-Thomson se détériorant corrélativement).

[10] **Composition des super-régions**

Super région 1 : (Champagne-Ardenne, Nord-Pas de Calais, Lorraine, Alsace)
Super région 2 : (Ile de France, Picardie, Haute-Normandie)
Super région 3 : (Centre, Bourgogne, Franche-Comté)
Super région 4 : (Auvergne, Languedoc-Roussillon, Limousin)
Super région 5 : (Basse-Normandie, Bretagne)
Super région 6 : (Rhône-Alpes, PACA, Corse)
Super région 7 : (Aquitaine, Midi-Pyrénées)
Super région 8 : (Pays-de-la-Loire, Poitou-Charentes)

On se réfèrera à la partie 4 pour l'analyse du comportement de la partie rurale de l'Echantillon-Maître de 1999 vis à vis de l'équilibrage.

Enfin, il n'était pas possible d'envisager un équilibrage à un niveau régional. La taille des échantillons par strate (région X SG) est trop petite (en moyenne 5 UP) pour, à la fois, garantir une propriété d'équilibrage satisfaisante et conférer à la méthode un caractère aléatoire suffisant.

2.4. Tirage équilibré des groupes de districts en SG2 et des districts en SG3 et SG4[11]

2.4.1 Tirage équilibré des groupes de districts en SG2

Stratification de l'UU

Avant le tirage des groupes de districts, on réalise une stratification de l'UU par regroupement de communes. Elle a pour objet d'assurer une dispersion géographique des groupes de districts tirés.

Cette procédure de constitution des strates est entièrement automatisée (il y a 93 UU tirées en SG2). On commence par trier les communes dans l'ordre croissant de leur nombre de résidences principales. On parcourt alors le fichier commune par commune en cumulant leur nombre de logements principaux. Dès que ce cumul dépasse le sixième des logements de l'UU, un groupe de communes ou strate est alors créé.

Cette procédure aboutit, en pratique, à la constitution de strates qui distinguent le centre de l'UU et sa périphérie urbaine.

[11] Réalisé avec la macro SAS CUBE

Constitution des groupes de districts

Les groupes de districts sont constitués de telle sorte que l'échantillon qui y sera tiré comporte une réserve suffisante de logements pour l'ensemble des futures enquêtes nationales. Sachant que, pour les enquêtes nationales (sans extensions régionales), le nombre maximum de FA à enquêter par UU ne dépasse pas 30[12], il faut donc prévoir une réserve de 30 X 60 enquêtes = 1800 logements principaux. Sachant de plus que l'on s'impose un tirage de 30 groupes de districts dans chaque UU, il faut donc constituer des groupes de districts comprenant au moins 60 logements principaux.

Tirage des groupes de districts

La répartition par strate des 30 groupes de districts à tirer s'effectue selon une allocation proportionnelle au nombre de logements principaux contenus dans les strates. Les groupes sont alors sélectionnés selon un tirage à probabilités égales et équilibrés sur l'UU selon les trois tranches d'âge [0,19], [20,59], [60,+][13].

2.4.2 Tirage équilibré des districts en SG3 et SG4

Stratification de l'UU

Avant tirage, on réalise une stratification de l'UU par regroupement de communes. Elle a pour objet d'assurer une dispersion géographique des districts tirés.

La procédure de constitution de ces strates consiste d'abord à trier les communes dans l'ordre croissant de leur nombre de résidences principales. Puis on parcourt le fichier commune par commune en cumulant leur nombre de districts. Dès que ce cumul dépasse 150 districts, un groupe de communes ou strate est alors créé.

Cette procédure aboutit, en pratique, à la constitution de strates qui distinguent le centre de l'UU et sa périphérie urbaine.

Pour l'UU de Paris, la même procédure est mise en oeuvre au niveau de chacun des départements.

[12] Rappelons que, pour une enquête standard de taux de sondage 1/2000ème, environ 23 FA sont tirées dans chaque UU de SG2. Ce nombre de FA/UU peut fluctuer fortement selon la taille de l'enquête : de 3 à 26 selon nos estimations.

[13] L'équilibrage au niveau UU implique également de définir H variables d'équilibrage contrôlant les tailles des échantillons dans chacune des H strates ou groupes de communes.

Tirage des districts

La réserve de logements principaux dans les UU des SG3 et SG4 doit être de 3%[14].

Le tirage des districts est un tirage stratifié, à probabilités égales et équilibré au niveau UU (sauf pour Paris où l'équilibrage est effectué au niveau du département). Les probabilités de sélection des districts sont toutes fixées à 3%. Les variables d'équilibrage sont les trois tranches d'âge [0,19], [20,59], [60,+][15] ainsi qu'une variable d'équilibrage pour contrôler la taille de l'échantillon en nombre de résidences principales[16].

Le tableau suivant donne, pour quelques UU de plus de 100.000 habitants, la taille minimale de la réserve de logements ainsi que le nombre d'enquêteurs à prévoir pour une enquête standard (taux de sondage au 1/2000ème).

Unité Urbaine	Nombre d'habitants	Nombre de logements principaux	Réserve logements principaux	Nombre de FA (Enquête Standard)	Nombre d'enquêteurs
Angoulême	100 000	46 000	1 380	23	1
Reims	215 000	94 000	2 820	47	2
Nantes	500 000	234 000	7 020	117	4
Lille	1 001 000	388 000	11 640	194	7
Paris	9 644 000	4 000 000	120 000	2000	67

Lecture : pour Angoulême, la réserve nécessaire dans l'Echantillon-Maître de 1999 doit être de 1 380 logements principaux, le nombre de FA à enquêter pour une enquête standard est de 23 et le nombre d'enquêteurs à mobiliser de 1.

[14] Ces strates sont exhaustives et comme la réserve Échantillon-Maître est calibrée sur la base de 60 enquêtes au taux de sondage moyen de 1/2000ème, on choisit donc de retenir dans chaque UU de SG3 et Paris 60/2000 = 3% des logements principaux.

[15] L'équilibrage au niveau UU implique également de définir H variables d'équilibrage contrôlant les tailles des échantillons dans chacune des H strates. Ainsi, on fixe à 3% le taux de sondage dans chaque groupe de communes.

[16] On montre que cette variable doit être définie par (0.03) X (nombre de résidences principales du district).

2.5. Constitution définitive de l'Echantillon-Maître

En strates de gestion 0 et 1 (strates des UP rurales et des UU de moins de 20.000 habitants), tous les logements de l'UP tirée sont retenus dans l'Echantillon-Maître. Pour les autres strates, seuls les logements composant les districts ou groupes de districts sélectionnés sont retenus.

C'est donc dans cette base de logements du RP99 que seront tirées les enquêtes ménages de type Echantillon-Maître.

Pour une description de la chaîne de tirage des logements dans l'Echantillon-Maître, on pourra lire pp. 180-192 « Les Techniques de Sondage » de P. ARDILLY.

3. Simulations de tirage Cube des UP[17]

3.1. Principe et raisons des simulations

3.1.1 Principe

Lorsque l'on s'intéresse à l'estimation d'une quantité (un total, une moyenne,...), la loi[18] ou distribution de l'estimateur mis en jeu est rarement connue de façon exacte. Rappelons que cette connaissance est nécessaire si l'on souhaite par exemple fournir des intervalles de confiance valides du paramètre mesuré.

Cependant, on peut assez souvent fournir une approximation théorique de la vraie loi de l'estimateur mais seulement dans un cadre asymptotique, c'est-à-dire lorsque l'échantillon est « suffisamment gros »[19]. Lorsque l'échantillon est petit, les simulations restent l'unique moyen pour approcher la vraie loi de l'estimateur.

Ainsi, si l'on souhaite étudier la loi d'un estimateur agissant sur des échantillons de taille n, on génère K échantillons de taille n selon la loi d'échantillonnage fixée. On dispose donc de K estimations dont la distribution empirique (l'histogramme)

[17] La réalisation informatique a été assurée par Jean-Noël PETIT (CNIP).

[18] Cette loi dépend à la fois de la loi de tirage des échantillons et de la fonction définissant l'inférence.

[19] Il n'est pas aisé d'établir un lien entre la taille de l'échantillon et la qualité de l'approximation de la vraie loi car ce lien dépend en général de la population que l'on étudie. Seules des études empiriques, telles que des simulations, permettront de préciser ce lien.

constitue une approximation acceptable de la loi de l'estimateur dès lors que K est assez grand.

3.1.2 Raisons

Premièrement, le tirage CUBE, fruit des recherches récentes de J-C. DEVILLE et Y. TILLE ne bénéficie d'aucune expérience pouvant se rattacher à celle de l'Echantillon-Maître. De plus, les échantillons d'UP dans les super-régions étant de petites tailles, cela nous mettait dans un cadre limite d'utilisation de CUBE préconisée par leur inventeurs[20]. Il fallait donc s'assurer, grâce aux simulations, du bon comportement de ce type de tirage appliqué à l'Echantillon-Maître de 1999.

Deuxièmement, l'Echantillon-Maître de 1999 devant concilier la contrainte d'une répartition régionale contrôlée[21] des UP et l'objectif d'un équilibrage super-régional de nature à améliorer la précision des estimateurs nationaux, il nous paraissait intéressant de comparer 3 stratégies de tirage définies selon un arbitrage entre cette contrainte et cet objectif.

3.2 Simulations appliquées au tirage des UP de l'Echantillon-Maître de 1999

3.2.1 Les stratégies de tirage

Les stratégies portent toutes sur le tirage d'UP **rurales** au sein des super-régions 4 et 6[22][23].

stratégie 1 : équilibrage simple

[20] On conseille d'utiliser un tirage CUBE sous la condition :

$$n > c+4$$

où n = taille de l'échantillon, c = nombre de variables d'équilibrage.
Hors de ce cadre, le tirage ne garantit plus une bonne propriété d'équilibrage et le caractère aléatoire (entropique) de la méthode n'est plus maximal.

[21] Cette contrainte permet à l'EM de « tomber » dans chaque région, donc de lui conférer un minimum de « représentativité » régionale, mais permet surtout d'assurer une répartition de l'effort de collecte équitable entre les régions pour chaque enquête.

[22] **Super région 4** : (Auvergne, Languedoc-Roussillon, Limousin)
Super région 6 (Rhône-Alpes, PACA). Corse non incluse lors des simulations, mais incluse lors du tirage effectif.

[23] Précisons que les premiers tests CUBE de tirage des UP rurales ont été réalisés en équilibrant sur les régions. Ils se sont avérés décevants en termes de précision des estimateurs de totaux (CV de l'ordre de 12% sur des variables d'intérêt égales aux variables d'équilibrage ou à une fonction de ces dernières), le nombre des UP tirées étant trop faible. Ces résultats nous ont donc confortés dans la nécessité d'une approche super-régionale.

- On tire les UP dans leur super-région d'appartenance, selon des probabilités proportionnelles à leur nombre de résidences principales.

- Le nombre d'UP tirées dans la super-région est égal à celui de l'Echantillon-Maître de 1990.

- On équilibre sur le revenu et les 3 tranches d'âge [0,19], [20,59] et [60,+].

Cette stratégie n'impose pas un nombre d'UP tirées par région. C'est la stratégie qui optimise donc la recherche d'un équilibrage super-régional (donc national) sans chercher à imposer une répartition régionale de l'effort de collecte pour les enquêtes.

stratégie 2 : équilibrage avec stratification régionale à allocation proportionnelle

- On stratifie la super-région en régions et on tire les UP proportionnellement à leur nombre de résidences principales dans chacune des strates régionales.

- Le nombre d'UP tirées dans chaque région est proportionnelle à la taille de la région en termes de nombre de résidences principales.

- De plus, le nombre total d'UP tirées dans la super-région est égal à celui de l'Echantillon-Maître de 1990 .

- On équilibre sur le revenu et les 3 tranches d'âge à un niveau super-régional.

Cette stratégie concilie la contrainte d'une répartition régionale contrôlée[24] des UP et l'objectif d'un équilibrage super-régional. Le nombre d'UP tirées par région est proportionnel à leur nombre de résidences principales.

stratégie 3 : équilibrage avec stratification régionale à allocation identique à celle de l'Echantillon-Maître de 1990

- On stratifie la super-région en régions et on tire les UP proportionnellement à leur nombre de résidences principales dans chacune des strates régionales.

- Le nombre d'UP tirées dans chaque région est égal à celui de l'Echantillon-Maître de 1990.

[24] Rappelons que cette contrainte permet à l'Échantillon-Maître de « tomber » dans chaque région, donc de lui conférer un minimum de « représentativité » régionale, mais permet surtout d'assurer une répartition de l'effort de collecte équitable entre les régions pour chaque enquête.

- On équilibre sur le revenu et les 3 tranches d'âge à un niveau super-régional.

Cette stratégie concilie la contrainte d'une répartition régionale contrôlée des UP et l'objectif d'un équilibrage super-régional. Le nombre d'UP tirées par région est celui de l'Echantillon-Maître de 1990. Ce nombre étant souvent égal ou proche à la répartition proportionnelle, les stratégies 2 et 3 sont quasiment identiques.

stratégie 4 : stratification régionale à allocation identique à celle de l'Echantillon-Maître de 1990 (sans équilibrage super-régional)

- On stratifie la super-région en régions et on tire les UP proportionnellement à leur nombre de résidences principales dans chacune des strates régionales.

- Le nombre d'UP tirées dans chaque région est égal à celui de l'Echantillon-Maître de 1990.

Cette stratégie, qui omet les variables d'équilibrage et respecte la contrainte de répartition de l'effort de collecte, permet de mesurer le gain de précision des autres stratégies qui utilisent l'information auxiliaire d'équilibrage.

3.2.2 Méthodologie

Les deux tableaux suivants fournissent, pour chaque super-région, le nombre d'UP rurales dans la base et la taille des échantillons tirés selon les stratégies.

Super région 4

	Nombre UP base	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4
Auvergne	105	indéterminé	5 ou 6[25]	5	5
Languedoc	93	indéterminé	5 ou 4	5	5
Limousin	58	indéterminé	3	3	3
TOTAL	**256**	**13**	**13**	**13**	**13**

[25] 5 ou 6 car la somme des probabilités d'inclusion d'ordre 1 au niveau de l'Auvergne est comprise entre 5 et 6 (même explication pour Languedoc-Roussillon).

Super région 6

	Nombre UP base	Stratégie 1	Stratégie 2	Stratégie 3	Stratégie 4
Rhône-Alpes	195	indéterminé	12 ou 13	12	12
PACA	64	indéterminé	4 ou 3	4	4
TOTAL	**259**	**16**	**16**	**16**	**16**

Pour chacune des stratégies de tirage, on effectue 5.000 tirages indépendants d'échantillons. A partir de ces 5.000 échantillons, on calcule les 5.000 estimations du total super-régional (partie rurale) des variables d'intérêt suivantes :

Variables d'informations auxiliaires utilisées pour le tirage

- Nombre de logements principaux (RP99)
- Revenu net imposable (Source DGI 1996) : ARNET96
- Age de 0 à 19 ans (RP99) : AAGE1
- Age de 20 à 59 ans (RP99) : AAGE2
- Age de 60 ans et +(RP99) : AAGE3

• *Variables réelles n'intervenant pas dans l'échantillonnage*

- Nombre de logements occasionnels (RP99)
- Nombre de logements secondaires (RP99)
- Nombre de logements vacants (RP99)

• *Variable artificielle corrélée aux variables auxiliaires*

- FONC = (AAGE1+AAGE2+AAGE3)/3 + 100*LOG(ARNET96)

• *Variables artificielles non corrélées aux variables auxiliaires*

Leur valeur qui est attribuée à chaque UP est générée à partir des réalisations de la loi normale centrée réduite et de la loi uniforme sur [0,1].

- ALEA1 = 250*normal(0,1) + 1000
- ALEA2 = 2000*uniforme([0,1])
- CONST = 1000

On construit pour chaque variable d'intérêt la loi empirique de l'estimateur CUBE. On peut alors répondre aux questions suivantes :

Comportement de Cube :

- Les estimateurs d'un total sont-ils sans biais ?
- Que gagne-t-on par rapport à un sondage aléatoire simple (calcul du design effect) ?

- La propriété d'équilibrage est-elle bien respectée ?

- Peut-on assimiler la loi de l'estimateur CUBE d'un total à une loi symétrique et, si oui, peut-on utiliser un calcul de l'intervalle de confiance sur le modèle d'une loi normale ?

Stratégies de tirages

- Quel niveau de précision statistique pour chaque stratégie ?

- Qu'apporte l'utilisation de variables d'équilibrage ?

- La contrainte de répartition régionale de collecte perturbe-t-elle fortement la précision super-régionale (i.e. : nationale) ?

Avant de présenter les résultats, précisons que le tirage de 5.000 échantillons de type CUBE, à partir de la macro SAS dont nous disposions, demande environ 12 heures sur un micro Pentium-260. Un ensemble de macros destiné à automatiser les simulations a d'ailleurs été développé par l'équipe de projet Échantillon-Maître.

Le choix des « 5.000 » offre un compromis acceptable entre la précision des simulations et leur temps d'exécution. En effet, on peut alors considérer que l'intervalle de confiance (à 95%) de la vraie variance de l'estimateur Cube (noté V) est de l'ordre de 2%[26].

3.2.3 Résultats

Les tests menés sur les super-régions 4 (Limousin + Auvergne + Languedoc-Roussillon) et 6 (Rhône-Alpes + Provence-Alpes-Côte-d'Azur) indiquent que le biais relatif de l'estimateur CUBE d'un total est bien nul pour chacune des variables d'intérêt[27].

[26] Avec 5000 tirages, on détermine

$IC_{95}(V) = \left[\hat{V}_{sim} - \delta, \hat{V}_{sim} + \delta\right]$ avec $\frac{\delta}{\hat{V}_{sim}} = 2\%$ et où $\hat{V}_{sim}$ désigne la dispersion des 5.000 estimateurs CUBE.

[27] Le biais relatif (empirique) est de l'ordre du millième. On le calcule selon la formule :
[(moyenne des 5000 estimateurs CUBE) - Vrai Total] / Vrai Total

Son coefficient de variation (*CV*) ainsi que son design-effect sont reportés dans les tableaux suivants. Précisons que le design-effect est calculé par le rapport de la variance CUBE à celle de l'estimateur d'Horvitz-Thompson issu d'un sondage aléatoire simple des UP.

On trouvera en annexe, pour la super-région 4, la distribution d'un estimateur CUBE du total des variables ayant servi pour les simulations.

Coefficient de variation (en %)

CV	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
Logement principaux	2,56	2,45	4,96	3,23	5,00	3,30	0	0
Revenu imposable 96	2,56	2,51	5,07	3,27	5,10	3,32	3,48	4,29
Age de 0 à 19 ans	2,96	2,50	5,25	3,16	5,25	3,22	5,00	4,26
Age de 20 à 59 ans	2,54	2,40	4,99	3,14	4,98	3,22	2,88	2,47
Age de 60 ans et +	3,06	2,77	5,28	3,63	5,35	3,88	4,09	4,78

Commentaire

Globalement, les *CV* se situent entre 0 et 5,35% pour la super-région 4 et entre 0 et 4,78 pour la super-région 6 qui obtient, quelle que soit la variable d'équilibrage, de meilleurs résultats que la super-région 4. La raison est probablement due à un taux de sondage plus élevé et à un nombre de régions plus faible.

Si l'on classe les stratégies selon la qualité de l'équilibrage, on peut proposer :

Super région 4 : stratégie 1 > stratégie 4 > stratégie 2 > stratégie 3

Super région 6 : stratégie 1 > stratégie 2 > stratégie 3 > stratégie 4

La stratégie 1 arrive logiquement en tête (stratégie optimisant l'utilisation de l'information auxiliaire).

Pour la super-région 6, c'est effectivement l'ordre attendu : le contrôle des tailles des échantillons par région est bien de nature à détériorer l'équilibrage super-régional. La stratégie 4, qui n'est pas équilibrée, arrive bien en dernière position.

Pour la super-région 4, la stratégie 4 se situe, de façon surprenante[28], devant la deuxième et troisième stratégie. L'explication tient sûrement, en partie, à l'existence d'une corrélation assez forte entre le nombre de résidences principales et les autres variables d'équilibrage. Une deuxième raison pourrait être l'utilisation d'un nombre de variables d'équilibrage trop élevé dans les stratégies 2 et 3. Une troisième raison serait la difficulté de l'actuelle macro-CUBE à respecter rigoureusement les allocations super-régionales[29] et régionales, ce qui pourrait provenir d'un mauvais relâchement des contraintes lors de la phase d'atterrissage.

CV	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
Logements occasionnels	14,32	53,34	15,03	52,90	15,18	52,39	14,88	59,44
Logements secondaires	32,09	33,67	32,11	33,32	32,79	33,27	32,60	39,17
Logements vacants	7,71	17,52	8,69	17,63	8,89	17,55	8,69	19,24

Les *CV* sont élevés, notamment pour la super-région 6. Mais rappelons que ces variables ne sont pas utilisées pour l'échantillonnage. Il sera ici plus intéressant de raisonner en termes de design-effect de façon à mieux mesurer l'apport de l'information auxiliaire lors du tirage.

Le classement des stratégies selon ces variables reste identique au précédent :

Super région 4 : stratégie 1 > stratégie 4 > stratégie 2 > stratégie 3

Super région 6 : stratégie 1 > stratégie 2 > stratégie 3 > stratégie 4

[28] 'Décevante' serait peut-être plus approprié ! En effet, l'introduction des variables d'équilibrage de revenu et d'âge dans les stratégies 1, 2 et 3 ne permet pas d'obtenir un gain de précision substantiel sur l'estimation de leur total par rapport à la stratégie 4.

[29] Pour les stratégies 1,2 et 3, on observe, sur les tailles d'échantillons tirés dans les super-régions, des fluctuations de $\pm$ 1.

CV	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
FONC	6,34	3,02	7,62	3,77	7,55	3,79	2,25	2,24

Les CV sont bons, quelle que soit la stratégie. Le classement pour cette variable corrélée aux variables d'équilibrage est :

Super région 4 : stratégie 4 > stratégie 1 > stratégie 3 > stratégie 2

Super région 6 : stratégie 4> stratégie 1 > stratégie 2 > stratégie 3

La stratégie 4 est la meilleure pour la variable FONC. La seule information auxiliaire du nombre de résidences principales semble donc suffire dans ce cas.

CV	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
ALEA 1	17,80	7,77	18,13	8,24	17,77	8,20	10,21	7,52
ALEA 2	24,36	14,87	24,40	15,14	24,50	15,05	21,05	14,72
CONST	15,31	4,98	15,71	5,58	15,42	5,57	5,69	4,41

Classement pour ces variables très faiblement corrélées avec les variables d'équilibrage :

Super région 4 : stratégie 4 > stratégie 1 > stratégie 2 > stratégie 3

Super région 6 : stratégie 4 > stratégie 1 > stratégie 2 > stratégie 3

La stratégie 4 l'emporte. Peut-être est-ce lié au fait que cette stratégie contrôle de façon parfaite la taille des échantillons tirés dans chacune des régions. Ce résultat pose toutefois la question d'une possible dégradation de la précision par l'introduction de nouvelles variables d'équilibrage (stratégies 1, 2 et 3)[30].

[30] La théorie indique qu'il ne peut y avoir une dégradation. Reste donc l'éventualité d'un mauvais relâchement des contraintes dans la macro CUBE.

Design-effect (en %)

Design-Effect	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
Logements principaux	0,64	28,87	2,42	50,14	2,46	52,22	0	0
Revenu imposable 96	0,61	16,28	2,38	27,69	2,41	28,58	1,12	47,43
Age de 0 à 19 ans	0,74	13,82	2,32	22,04	2,32	22,95	2,10	40,20
Age de 20 à 59 ans	0,58	19,73	2,24	33,99	2,24	35,68	0,75	20,91
Age de 60 ans et +	1,05	17,79	3,12	30,32	3,22	30,55	1,88	52,59

Le gain par rapport à un sondage aléatoire simple est substantiel.

Design-Effect	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
Logements occasionnels	21,00	71,40	22,99	70,60	23,77	69,01	22,83	88,76
Logements secondaires	60,76	72,18	60,91	69,67	64,63	70,11	63,60	96,00
Logements vacants	4,79	84,96	6,04	85,50	6,36	85,39	6,05	103,19

Le gain par rapport à un sondage aléatoire simple (SAS) est important sauf dans un cas, celui de la super-région 6 où la stratégie 4 fournit un estimateur du nombre de logements vacants moins bon que l'estimateur issu d'un SAS.

Design-Effect	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
FONC	10,13	126,83	14,65	196,73	14,42	199,27	1,28	69,53

On constate qu'un tirage des UP selon un SAS serait meilleur en super-région 6 pour cette variable, sauf pour la stratégie 4.

Design-Effect	Stratégie 1		Stratégie 2		Stratégie 3		Stratégie 4	
	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6	S-Reg 4	S-Reg 6
ALEA 1	712,48	177,89	744,27	199,07	714,55	198,04	234,90	166,53
ALEA 2	231,12	117,63	231,61	120,62	232,27	118,87	170,63	114,44
CONST	+∞	+∞	+∞	+∞	+∞	+∞	+∞	+∞

L'estimateur SAS est logiquement le meilleur pour la variable d'intérêt constante (CONST) ou fluctuant peu (ALEA1). En revanche, pour la variable ALEA2, ce résultat était-il prévisible ?

3.3 Validation du tirage CUBE

Suite aux simulations, le tirage de type CUBE est finalement retenu pour le tirage effectif de l'EM99. Les simulations ont révélé un bon comportement de la macro CUBE en termes de biais des estimateurs (biais nuls) et d'un respect acceptable de la propriété d'équilibrage (CV < 5%). De plus, le mode même de tirage (équilibré et à probabilités inégales) s'est trouvé légitimé par des design-effect appréciables sur la plupart des variables d'intérêt.

Cependant, un problème de réglage de la macro est apparu, se manifestant par une difficulté à respecter rigoureusement les tailles imposées d'échantillons au sein des régions, voire parfois au niveau des super-régions. Peut-être est-ce dû à la phase d'atterrissage CUBE lors du relâchement des contraintes d'équilibrage.

Nous attirons aussi une nouvelle fois l'attention sur la nécessité d'une utilisation prudente de CUBE dans le cadre de petits échantillons[31], une alternative intéressante étant une méthode de tirage de type simplexe.

[31] Rappelons que l'on conseille d'utiliser CUBE dans les cas où :

$$n > c+4$$

où n = taille de l'échantillon, c = nombre de variables d'équilibrage.
Hors de ce cadre, le tirage ne garantit plus une bonne propriété d'équilibrage et le caractère aléatoire (entropie) de la méthode n'est plus maximal.

4. Propriétés de l' Echantillon-Maître de 1999

Les simulations ayant permis d'accepter « l'outil » de tirage CUBE, nous l'utilisons donc pour le tirage effectif de l'Echantillon-Maître dans les strates de gestion 0,1 et 2. Nous avons opté pour la démarche suivante. En toute rigueur, nous ne devrions générer qu'un unique échantillon d'UP. La tentation est forte cependant de générer un grand nombre d'échantillons-maîtres et de retenir celui qui nous conviendrait le mieux ! Cette démarche est naturellement à proscrire car elle conduirait à un choix quasi-déterministe de l'Echantillon-Maître[32]. Cependant, au sein de chacune des huit super-régions, nous nous sommes autorisés à générer deux Échantillons-Maîtres et à opérer à un choix. Par cette procédure, nous estimons ne pas perturber le mécanisme aléatoire. Les critères de choix sont :

- retenir l'Echantillon-Maître qui aurait les meilleures propriétés d'équilibrage

- retenir l'Echantillon-Maître présentant une bonne dispersion géographique et, si possible, assez proche de l'ancien Echantillon-Maître[33].

Une fois l'Echantillon-Maître de 1999 retenu, nous avons cherché à déterminer certaines de ses propriétés. Rappelons, en effet, que ce premier degré de tirage influe fortement sur la qualité des estimateurs des futures enquêtes ménages issues du nouvel échantillon-maître. Afin de s'assurer que les échantillons d'UP présentent de « bonnes propriétés », nous avons réalisé un contrôle a posteriori sur deux niveaux.

La démarche et les résultats que nous présentons ne concernent que la partie rurale de l'Echantillon-Maître de 1999.

4.1. Définition des propriétés

Nous distinguons les propriétés de niveau 1 et de niveau 2.

- ***1er niveau*** : les échantillons sont-ils bien représentatifs d'une opposition entre deux composantes du monde rural : « développé » / « profond » ?

- ***2ème niveau :*** les échantillons présentent-ils de bonnes qualités inférentielles sur des variables d'intérêt jugées comme ayant un pouvoir explicatif assez fort des phénomènes étudiés dans les enquêtes ménages ?

[32] La méthode n'offrirait plus la garantie d'estimateurs sans biais et poserait le problème de la définition des poids d'inférence.

[33] L'outil cartographique est pour cela d'une aide précieuse.

Pour le 1er niveau, de nature purement descriptive, il ne s'agit pas de préciser les propriétés inférentielles de l'Echantillon-Maître mais de vérifier que sa composition, sur des variables jugées importantes et qui n'ont pas été explicitement intégrées lors de l'échantillonnage, ne présente pas de distorsions majeures par rapport à la base de sondage.

Nous avons tout d'abord contrôlé la composition des échantillons super-régionaux (partie rurale) en termes de répartition de leurs communes selon un indice d'évolution démographique. Plus précisément, nous avons calculé, pour toutes les communes rurales (y compris celles hors échantillon) d'une super-région, l'évolution démographique relative[34] de la commune entre le RP90 et le RP99. Il devient alors possible de classer les communes selon les tierciles de la distribution de cet indice démographique (commune en déclin, stable, en expansion). Enfin, on dénombre sur l'échantillon les communes en déclin, stables et en hausse. L'échantillon est considéré comme représentant parfaitement la super-région si sa répartition est de 1/3-1/3-1/3 entre ces trois catégories.

Ensuite, nous avons vérifié que les échantillons d'UP super-régionaux possèdent une bonne répartition entre les communes les plus pauvres et les plus riches. Pour cela, nous avons déterminé, sur l'ensemble des communes rurales d'une super-région donnée, les quartiles du revenu moyen. On classe alors les communes en « Pauvres », « Neutres » et « Riches » selon qu'elles se situent sous le 1er quartile, entre le 1er et 3ème quartile, et au-dessus du 3ème quartile. On peut alors comparer la répartition des communes « Pauvres », « Neutres » et « Riches » de l'échantillon super-régional considéré. L'échantillon est considéré comme représentant parfaitement la super-région si sa répartition est de 1/4-1/2-1/4 entre ces trois catégories.

Pour le 2ème niveau, nous avons contrôlé les qualités inférentielles de l'Echantillon-Maître. Ainsi, nous vérifions d'abord que l'estimation d'Horvitz-Thompson ou estimation CUBE du total de chacune des 4 variables d'équilibrage est proche du vrai total (total restreint aux communes rurales) de la super région considérée.

Ensuite, nous estimons l'effectif super-régional (pour la strate rurale) de 13 postes.

Les estimations sont ensuite comparées aux vrais effectifs par le calcul de l'écart relatif (les données utilisées sont celles du RP90 en raison de la non-disponibilité des données de la codification de la CS du RP99 au moment de cette étude).

[34] Indice calculé selon la formule :
(Population 99 de la commune-Population 90 de la commune) / Population 90

4.2 Résultats

Les résultats ont été communiqués à l'ensemble des DR lors de la livraison de leur échantillon après avoir effectué le choix entre les deux Echantillons-maîtres. Nous ne donnons ici, à titre d'exemple, que les résultats obtenus sur la partie rurale de la super-région 1 (Champagne-Ardenne, Nord-Pas de Calais, Lorraine, Alsace).

4.2.1 Résultats de niveau 1

Répartition des communes de l'échantillon selon les (vrais) tierciles de l'évolution démographique

Super-région 1 (Champagne-Ardenne, Nord-Pas de Calais, Lorraine, Alsace)

	Effectifs	%
Evolution démographique		
En déclin	130.00	32.26
Stable	129.00	32.01
En hausse	144.00	35.73

Le résultat montre que l'échantillon d'UP rurales de la super-région 1 traduit bien, au sens de l'indicateur démographique utilisé, le comportement réel de la super-région. On observe aussi de bons résultats pour les autres super-régions.

Répartition des communes de l'échantillon selon les (vrais) quartiles[35] du revenu net imposable moyen

Super-région 1 (Champagne-Ardenne, Nord-Pas de Calais, Lorraine, Alsace)

	Effectifs	%
STATUT		
Pauvre	76.00	18.86
Moyen	196.00	48.64
Riche	131.00	32.51

[35] Rappelons qu'une commune est dite
- « pauvre » si son revenu net imposable moyen est inférieur au 1er quartile
- « moyenne » si son revenu net imposable moyen est compris entre le 1er et le 3ème quartile
- « riche » si son revenu net imposable moyen est supérieur au 3ème quartile.

Les résultats suivants montrent que l'échantillon d'UP de la super-région 1 traduit relativement bien, au sens du revenu net imposable moyen, le comportement réel de la super-région. Les résultats sont similaires pour les autres super-régions, sauf, peut-être, pour la super-région 6 (Rhône-Alpes, PACA, Corse), dont l'échantillon a tendance à sur-représenter les communes définies comme « pauvres » au détriment de celles qualifiées de « moyennes ».

4.2.2 Résultats de niveau 2

Estimation du total des variables d'équilibrage et variables d'intérêt de type catégorie sociale (strate de gestion 0)

Super-région 1 (Champagne-Ardenne, Nord-Pas de Calais, Lorraine, Alsace)

		Vrai total	Estimateur du total	Ecart relatif (%)
Type de variable	Variables			
Variables d'équilibrage	Revenu net imp. en MF (1996)	87339	86672	-0.76
	0-19	596188	607588	1.91
	20-59	1078450	1083021	0.42
	60 et +	406886	410971	1.00
Autres variables	Pop. totale	2033099	2073516	1.99
	Agriculteurs	86476	102251	18.24
	Artisans, commerçants	53029	53581	1.04
	Cadres	45524	41402	-9.05
	Professions Intermédiaires	137560	134274	-2.39
	Employés	183856	175932	-4.31
	Ouvriers	328315	319593	-2.66
	Retraités	315013	319841	1.53
	Inactifs	883326	926641	4.90
	Etudiants	174388	180169	3.32
	Pop. active avec emploi	771714	759146	-1.63
	Chômeurs	70626	77851	10.23
	Pop. active totale	852932	847928	-0.59

L'estimateur d'Horvitz-Thompson (ou CUBE), utilisé pour les estimations fournies ci-dessus, possède une propriété de calage naturel sur les variables d'équilibrage, d'où un écart relatif faible. Les autres super-régions ont des résultats tout à fait comparables, avec des écarts relatifs de l'ordre de 1% à 3%. Les plus mauvais résultats sont obtenus pour la super-région 6 avec un écart relatif pouvant atteindre les 5%.

Notons que, pour les variables autres que celles d'équilibrage, les écarts relatifs sont plus importants, mais toutefois acceptables pour chacune des super-régions. Bien entendu, pour une enquête donnée, la procédure classique de redressement (calage **après** tirage) est toujours possible.

Conclusion

Parallèlement au projet de constitution de l'Echantillon-Maître de 1999 ou dans son prolongement, plusieurs projets sont en cours d'étude.

Tout d'abord, la possibilité de constituer un second Echantillon-Maître, dit EMEX, qui serait utilisé pour le tirage des extensions régionales des enquêtes nationales. Rappelons que, par le passé, de tels tirages étaient réalisés hors EM dans les bulletins-papier du RP. La création de l'EMEX permettrait d'automatiser complètement le tirage des logements, y compris l'impression de l'adresse. Cependant, l'EMEX, qui comme l'Echantillon-Maître, reposerait sur le principe de zones fixes d'enquêtes, ne permettrait pas de s'adapter, pour certaines extensions, à un besoin particulier de localisation fine des enquêtes, pour lequel d'autres outils doivent être mis en place.

Ensuite, les travaux effectués et les réflexions sur la manière de construire un Echantillon-Maître offrent des perspectives d'amélioration pour la constitution d'un futur Echantillon-Maître issu de RRP. Ce projet permettrait également, sans doute, d'apporter une aide toujours plus efficace dans le cadre des missions de coopération sur ce sujet.

- *Un premier axe de réflexion* porterait sur la définition des strates de gestion qui ne reposerait plus uniquement sur la taille des UU mais pourrait intégrer également d'autres critères, par exemple de type économique (dynamisme, spécificité en termes d'activité, prix moyen du logement au m²...).

- *Un deuxième axe* concernerait le développement d'un algorithme de constitution automatisée des UP rurales, s'affranchissant des limites cantonales, respectant les contraintes pratiques (contiguïté des communes, seuils de nombres de logements) et créant les UP selon un critère d'optimalité statistique (par exemple, la recherche d'UP les plus hétérogènes ou les plus homogènes en intra dans le cas d'une stratification...). [cf. « Un algorithme de regroupement d'unités statistiques selon certains critères de similitude », Marc CHRISTINE et Michel ISNARD, présenté aux VII ^{èmes} JMS]. Précisons qu'un algorithme a déjà été développé dans le cadre de l'échantillon Emploi, pour regrouper les petites communes rurales en respectant la contiguïté, mais sans tenir compte d'aucun critère d'optimalité.

- *Un troisième axe* s'appliquerait à définir une méthode utilisant au mieux l'information auxiliaire des variables d'équilibrage (création de variables synthétiques d'équilibrage).

- *Un dernier axe* porterait sur le gain éventuel de précision apporté par la suppression du niveau de tirage des districts dans les UU et, plus généralement, sur les études comparatives de précision des enquêtes en fonction de différents plans de sondage ou de différents critères de constitution des UP, rendues possibles notamment par la mise en oeuvre de l'algorithme évoqué ci-dessus.

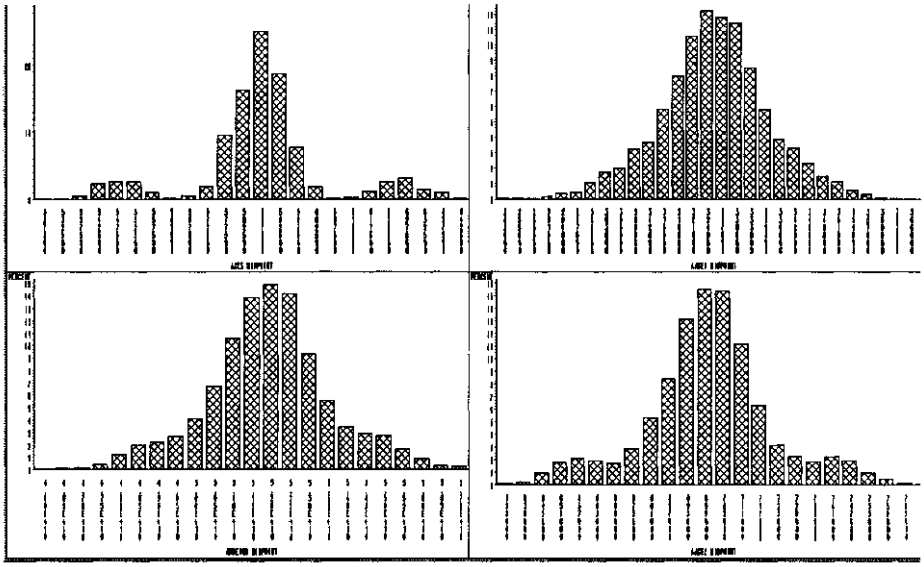
Enfin, les simulations entreprises afin de mieux connaître le comportement de l'algorithme de tirage CUBE appliqué à la sélection de l'Echantillon-Maître pourraient être approfondies en le comparant à d'autres types d'algorithmes.

ANNEXE

Ci-dessous, pour la super-région 4, sont représentées les 12 distributions des estimateurs CUBE du total des variables suivantes :

Nombre de résidences principales (ARES)

Effectif tranche d'âge [0,19] (AAG1)

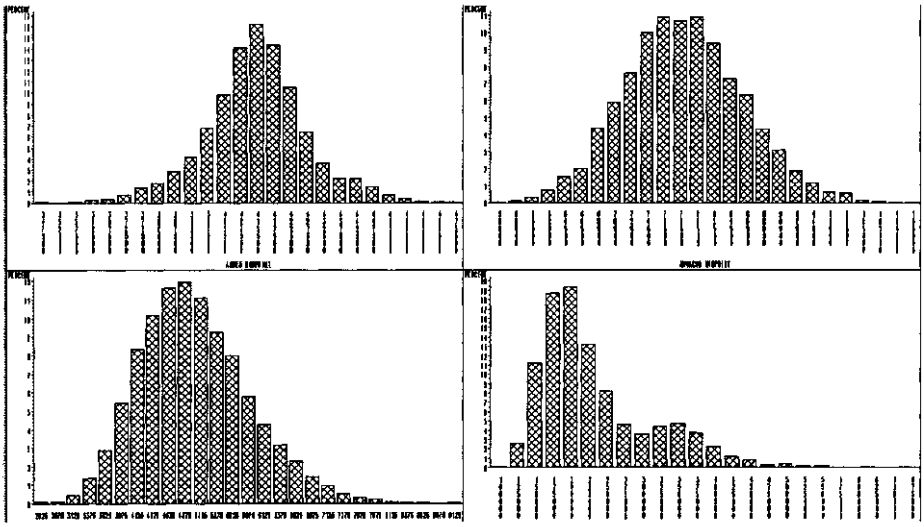


Revenu net imposable 1996 (ARNET96)

Effectif tranche d'âge [20,59] (AAG2)

Effectif tranche d'âge [60,+] (AAG3)

Nombre de résidences vacantes (ANVAC99)

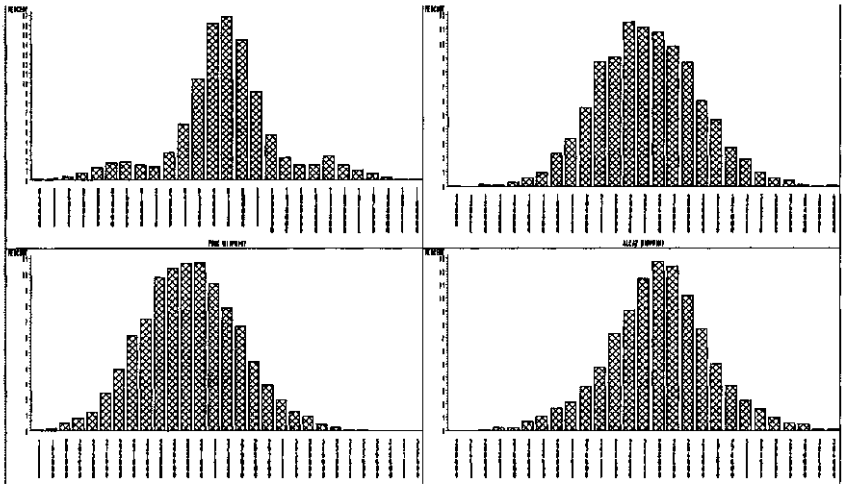


Nombre de résidences occasionnelles (ANOCC99)

Nombre de résidences secondaires (ANRSC99)

FONC

ALEA2



ALEA1

CONST