JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN DUMAIS
PHILIPPE BERTRAND
BERTRAND KAUFFMANN

Sondage, estimation et précision dans la rénovation du recensement de la population

Journal de la société française de statistique, tome 140, n° 4 (1999), p. 11-35

<http://www.numdam.org/item?id=JSFS_1999__140_4_11_0>

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

SONDAGE, ESTIMATION ET PRÉCISION DANS LA RÉNOVATION DU RECENSEMENT DE LA POPULATION

Jean DUMAIS, Philippe BERTRAND, Bertrand KAUFFMANN *

RÉSUMÉ

Dans certains pays, les agences statistiques nationales ont de plus en plus de difficultés à justifier et à obtenir les ressources nécessaires au dénombrement exhaustif de la population. La France a choisi de repenser complètement sa façon de recenser sa population. Cependant, contrairement à certains pays nordiques, la constitution en France d'un fichier de personnes ou de ménages est juridiquement impossible. La méthode proposée pour recenser par tranches annuelles et la façon de recomposer les estimations censitaires sont décrites en détail dans cet article. La question de la précision liée au sondage et au modèle est aussi traitée. Il faut noter que ce document présente l'état des options au moment d'aller sous presse et que l'avancement des recherches et les résultats des tests sur le terrain pourront venir le modifier; pour terminer, il donne donc quelques indications sur les travaux en cours et les perspectives d'évolution.

1. INTRODUCTION

Au cours de la dernière décennie, nombre de pays occidentaux ont réduit leurs dépenses publiques. Dans certains de ces pays, les agences statistiques nationales ont de plus en plus de difficultés à justifier et à obtenir les ressources nécessaires au dénombrement exhaustif de la population. De plus, dans plusieurs pays européens, les citoyens comprennent de moins en moins la nécessité d'un dénombrement décennal, étant donné la quantité et la richesse des informations administratives détenues par les agences gouvernementales. En France, les derniers recensements ont eu lieu en 1968, 1975, 1982, 1990 et 1999.

Au vu de ces changements environnementaux, la France a choisi de repenser complètement sa façon de recenser sa population. Dans de nombreux pays, nordiques en particulier, l'existence de fichiers de population avec mention obligatoire d'une adresse, couplée avec la possibilité d'interconnecter divers fichiers administratifs, permet une alternative intéressante à la pratique

* Troisième Programme de Rénovation du Recensement de la Population, INSEE 18, Boulevard Adolphe Pinard 75675 Paris, Cédex 14; courriel : jean.dumais@insee.fr

de recensements périodiques [Laihonen, 2000; Borchsenius, 2000]. Mais, en France, la constitution d'un fichier de personnes ou de ménages de cette nature est juridiquement impossible, celle d'un fichier de logements très difficile. Ainsi, pour fins de recensement, le dénombrement des personnes et des logements par des visites sur le terrain reste donc une nécessité.

La rénovation proposée repose sur le concept de « recensement continu » dont l'idée remonte à Kish [1990] et qu'on retrouve dans les études de la National Academy of Science (1994) en préparation du recensement américain de 2000. Une première approche des procédures envisageables en France peut être trouvée dans [Jacod et Deville, 1996]. Le présent article fait le point des développements méthodologiques depuis que la Direction de l'INSEE a décidé de mettre en route un projet de « Recensement Rénové » appliquant ce type d'approche.

L'idée de base repose sur l'observation, chaque année, d'une fraction de la population, ainsi qu'à l'utilisation des méthodes de sondage pour procéder à (une partie de) cette observation. Le recours aux données administratives et aux techniques d'estimation pour petits domaines permettra l'allègement de la pression statistique sur les ménages et un étalement des coûts sur plusieurs années ; jumelé à l'annualisation de la collecte, il permettra la publication d'un portrait statistique mis à jour annuellement plutôt que tous les 7 à 9 ans.

Un des objectifs de la rénovation est de continuer à pouvoir publier les résultats du recensement sur toute portion du territoire (même infracommunale), bien sûr dans les limites imposées par le respect du secret statistique, comme on sait le faire avec un recensement traditionnel. L'apport des données administratives et la modélisation permettront d'atteindre cet objectif tous les ans.

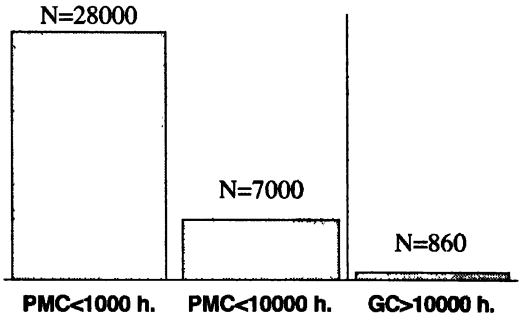
2. STRATÉGIE D'ÉCHANTILLONNAGE

Il est convenu qu'une modification des méthodes employées pour le recensement de la population ne peut s'accompagner d'une augmentation des coûts. La rénovation du recensement devra donc se faire à raison d'environ 8,4 millions de bulletins individuels par année, soit 60 millions de bulletins sur 7 ans, ce qui correspond à l'effort de collecte d'un recensement général chaque sept ans. Il est aussi convenu que la longueur du cycle de collecte du recensement rénové ne peut pas coïncider avec le cycle des élections municipales (6 ans). Un cycle de 10 ans existe ailleurs (par exemple, Argentine, États-Unis, Inde, Mexique, Suisse et un grand nombre de pays de l'Union européenne), mais il offrirait moins de fraîcheur que le cycle actuel. Le cycle de 5 ans paraît convenable, et conforme à plusieurs recommandations (ONU) ou pratiques internationales (Australie, Canada, Japon), en plus d'offrir une information mise à jour plus rapidement.

L'INSEE dispose d'une cartographie numérisée pour toutes les grandes communes (GC), celles de plus de 10 000 habitants. Cet ensemble de quelque 860 communes représentait 49 % de la population de 1990 et représente encore

la même proportion au recensement de 1999. Cette cartographie numérisée est complétée par le répertoire d'immeubles localisés (RIL). Le RIL permet le sondage direct d'immeubles d'habitation (au sens des bordereaux de district M6 du RP99, c'est-à-dire l'ensemble des logements desservis par la même cage d'escalier ou le même ascenseur) sans le passage coûteux (en termes de précision) par des aires.

Dans les petites et moyennes communes (PMC), celles de moins de 10 000 habitants, l'organisation d'un sondage pouvant donner des résultats d'une précision acceptable n'apparaît pas comme un opération économique. En effet, près de 80 % d'entre-elles (quelques 28 000 communes) comptent au plus 1 000 habitants, et on devrait y organiser des sondages à des taux dépassant les 75 %, en supposant un sondage aléatoire simple des logements. Pour les 7 000 communes dont la population s'établit entre 1 000 et 10 000 habitants, les taux de sondages devraient osciller entre 50 % et 20 % (respectivement), toujours sous l'hypothèse d'un sondage aléatoire simple de logements. Or, pour des enquêtes auprès de ménages utilisant un plan par grappes, il est courant de voir des effets de grappe de l'ordre de 1,5 ou 2, ce qui correspond à une correction des tailles d'échantillons données ci-dessus par un facteur du même ordre; on en arrive vite à la nécessité du dénombrement exhaustif.



Tous les éléments sont dès lors réunis pour la mise en place du plan de sondage proposé pour le recensement rénové.

En résumé, recenser sur un cycle de 5 ans, à raison de 8,4 millions de bulletins par année, soit par le RIL pour une moitié de la population, soit par listage exhaustif du territoire pour l'autre moitié donne :

$$\underbrace{\frac{1}{5} \times 29\ 900\ 000}_{PMC} + \underbrace{p \times 28\ 800\ 000}_{GC} = 8\ 400\ 000.$$

Ceci suppose que $p = 8$ %, c'est-à-dire 40 % de la population d'un cinquième des immeubles. On construit ainsi un cycle de 5 ans en grande commune.

Le plan de sondage proposé est donc le suivant : les PMC seront sondées au taux (moyen) d'un cinquième par an et tous leurs logements seront visités; toutes les GC seront visitées chaque année, mais seulement une fraction (environ 8 %) de leurs logements sera enquêtée.

Comme la population se répartit à peu près également entre « petites et moyennes » et « grandes » communes, on obtient un taux moyen de sondage annuel de l'ordre de 1/7 des logements. Sur la période quinquennale, on doit plutôt compter sur 41 420 000 bulletins, soit tout près de 70 % de la population.

2.1. Les petites et moyennes communes (PMC)

Considérons d'abord le domaine des PMC. Dans chaque région, les PMC seront éventuellement séparées en deux strates, « PMC rurales » ou « PMC urbaines ». Dans chacune des strates, 5 groupes de rotation de communes seront formés. Ces groupes de rotation seront créés à partir des renseignements du recensement de la population (RP99) par tirage d'échantillons équilibrés [Deville Tillé, 1999] sur les effectifs de la distribution âge-sexe des communes ; cette approche devrait permettre de minimiser les variations inter-annuelles dues au seul sondage.

Les figures 1 à 4 illustrent comment les 5 groupes de rotation sont équilibrés. Pour cette illustration, les variables d'équilibre sont les effectifs au Recensement général de la population de 1990 (RGP90) des classes d'âge-sexe ; les tranches d'âge retenues sont 0 à 19 ans, 20 à 39 ans, 40 à 59 ans, 60 à 74 ans et 75 ans et plus. Ces quatre figures donnent les « diagrammes à moustaches » de quelques variables mesurées sur les 2815 petites communes de Rhône-Alpes au RGP90. Pour chaque groupe de rotation, on voit et les quartiles (les bornes inférieure et supérieure du rectangle) et l'étendue de la distribution ; il est intéressant de noter la superposition des diagrammes d'une même variable. Des variables montrées dans ces quatre figures, seule la variable « Nombre de femmes âgées de 20 à 39 ans » a été utilisée pour la composition des groupes ; le nombre total d'hommes n'intervient pas directement dans l'établissement de l'équilibre ni aucune des variables associées au ménage ou au logement.

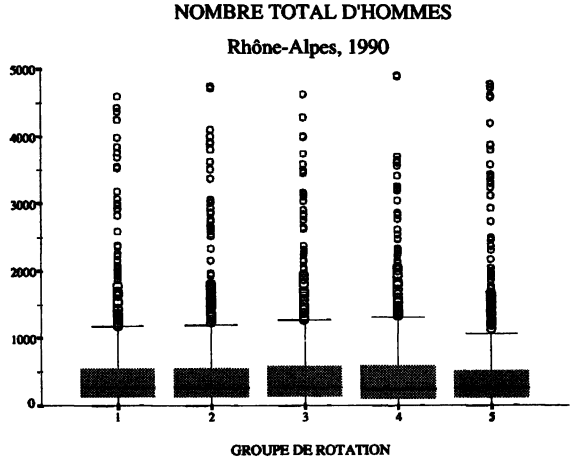


FIG 1. —

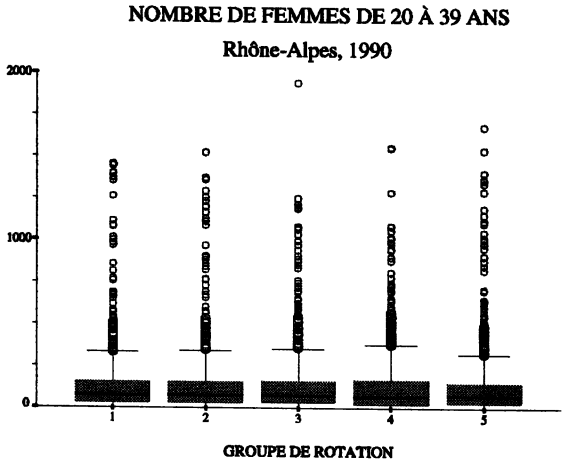


FIG 2. —

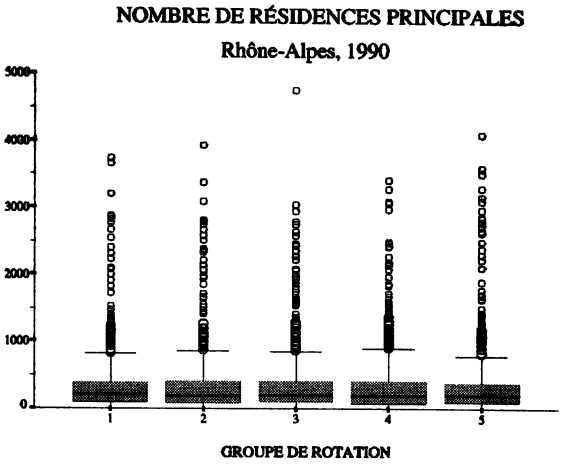


FIG 3. —

Chaque année, on fera le recensement (c'est-à-dire le dénombrement et la collecte exhaustifs) de la population et des logements de toutes les communes d'un des groupes de rotation. Ainsi, chaque PMC sera recensée une fois tous les 5 ans, et toutes les PMC à raison d'un cinquième par année.

Des tests de stabilité de l'équilibre (voir Tableau 1) ont été menés sur les petites et moyennes communes de la région Rhône-Alpes ($N = 2815$, $E(n) = N/5 = 563$). Pour ce faire, des groupes de rotation ont été formés par tirage équilibré en utilisant les données du Recensement de 1982. On a calculé la précision relative des estimations globales annuelles, c'est-à-dire obtenues à partir des seules données de l'enquête de l'année en cours. La précision relative est le rapport de l'erreur-type d'une estimation à celle-

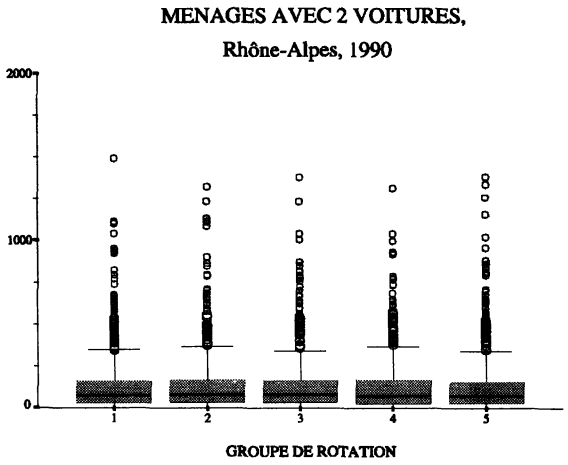


FIG 4. —

ci, exprimé en pourcentage; on l'appelle aussi « coefficient de variation » ou « cv ». Dans le tableau, il s'agit de la précision qu'aurait donné, en 1982, le plan proposé élaboré sur les données de 1982; on voit naturellement que la précision est absolue pour les variables d'équilibre pour la région (cv=0 %); pour avoir une idée de la précision sur des domaines plus petits que la région, la précision a aussi été calculée pour le domaine « département du Rhône » et pour le domaine « aire urbaine de Grenoble ». Ces résultats sont donnés dans les colonnes intitulées RP82. Pour avoir une idée de la stabilité des groupes de rotation, les mêmes statistiques ont été calculées en utilisant les données du recensement de 1990 pour les mêmes communes, dans les groupes de rotation de 1982, comme si ceux-ci avaient « vieilli »; les résultats sont présentés dans les colonnes intitulées RP90.

Le Tableau 1 montre que la perte de précision est nulle pour les domaines « département du Rhône » et « aire urbaine de Grenoble »; notons que la publication des résultats directs des enquêtes annuelles n'est pas prévue pour ce niveau de détail géographique. Pour l'ensemble de la strate « Rhône-Alpes », le Tableau 1 montre aussi qu'il n'y a pas de perte de précision évidente pour les quelques variables moins corrélées à l'âge ou au sexe, et qu'elle est de l'ordre de 1 % pour les variables d'équilibre. Le Tableau 1 illustre aussi comment l'équilibre, atteint au niveau de la région lors du tirage de l'échantillon, n'est pas préservé à des niveaux géographiques inférieurs (département, aire urbaine, pays, etc.). En fait, la précision sur des domaines infrarégionaux est équivalente à celle qu'on aurait obtenue d'un tirage aléatoire simple. La taille des structures infrarégionales et l'hétérogénéité des tailles des PMC qui les composent, rendent difficile le tirage équilibré des PMC à ce niveau.

TABLEAU 1. — Stabilité du tirage équilibré dans le temps

Nombre de...	Précision relative (cv en %) des estimations annuelles					
	Rhône-Alpes, N=2815, E(n)=563		Dépt Rhône, N=274, E(n)=55		AU Grenoble, N=25, E(n)=5	
	RP82	RP90	RP82	RP90	RP82	RP90
00-19 ans	0	1.3	16.0	15.9	45.9	45.1
20-39 ans	0	1.1	16.4	16.0	46.3	45.6
40-59 ans	0	0.9	15.7	15.7	43.8	44.4
60-74 ans	0	0.5	15.3	15.2	45.2	44.0
75-95 ans	0	0.8	16.1	16.2	50.9	47.3
Hommes	0	0.9	15.6	15.4	44.8	44.5
Femmes	0	0.8	15.9	15.7	45.1	44.5
Total	0	0.9	15.7	15.6	45.0	44.5
Etrangers	6.2	6.0	26.2	24.8	55.0	51.2
Migrants	1.5	2.3	17.2	16.7	45.9	45.0
Chômeurs	2.2	2.2	18.7	18.4	47.5	47.9
Employés	0.5	1.1	15.7	15.5	45.1	44.9

2.2. Les grandes communes (GC)

Le sondage en GC utilisera le « répertoire d'immeubles localisés » (RIL). Ce répertoire est une liste d'édifices adressés (édifices ésidentiels, institutionnels ou commerciaux) repérés géographiquement de façon à créer une cartographie numérisée. Le RIL sera d'abord alimenté par les résultats du RP99 permettant ainsi de décrire statistiquement chaque immeuble résidentiel du RP99. Le RIL sera mis à jour en continu à partir de permis de construire, de permis de démolir, de fichiers d'abonnés (eau, gaz, électricité,...), de renseignements fournis par les administrations locales et par l'observation directe sur le terrain. Ainsi, le RIL peut servir à la constitution d'une base de sondage « immeubles » en GC.

Dans chaque IRIS2000 [1] de chaque GC, on créera si nécessaire jusqu'à 3 strates d'immeubles – « petits », « moyens » et « trop grands » – selon la distribution de la taille des immeubles de l'IRIS2000. Au sein de chaque strate, on créera ensuite 5 groupes de rotation d'immeubles sur le modèle du sondage en PMC.

Trois strates supplémentaires seront prévues dans chaque IRIS2000 : une pour les immeubles d'activité (usines, entrepôts,...), une seconde pour les logements collectifs (établissements, collectivités, communautés, internats,...),

1. IRIS2000 = « *îlots regroupés selon des indicateurs statistiques* », zone homogène d'environ 2 000 habitants, soit environ 800 logements, ou de 200 à 300 immeubles en moyenne.

et une dernière pour les immeubles neufs. On visitera chaque année un cinquième des immeubles d'activité pour s'assurer qu'ils sont toujours vides de logements (logement de gardien, ou espace converti à l'habitation) ; les logements éventuellement trouvés dans de tels immeubles seraient considérés autoreprésentatifs [2] parce qu'exceptionnels ; en effet ces immeubles sont réputés ne contenir aucun logement d'habitation. L'ensemble des logements collectifs sera couvert chaque année ; un cinquième d'entre eux seront visités alors que l'effectif des quatre autres cinquièmes sera mis à jour, éventuellement par enquête téléphonique. Finalement, les immeubles d'habitation neufs, c'est-à-dire achevés et habitables depuis la dernière année d'enquête, seront recensés afin de pouvoir les prendre en compte dès que possible (donc, avec un poids de 1) ; ce recensement permettra aussi de faire le portrait statistique de ces immeubles neufs et de les insérer au mieux dans l'un des groupes de rotation (donc avec un poids d'environ 5). La Figure 5 illustre l'organisation de la collecte.

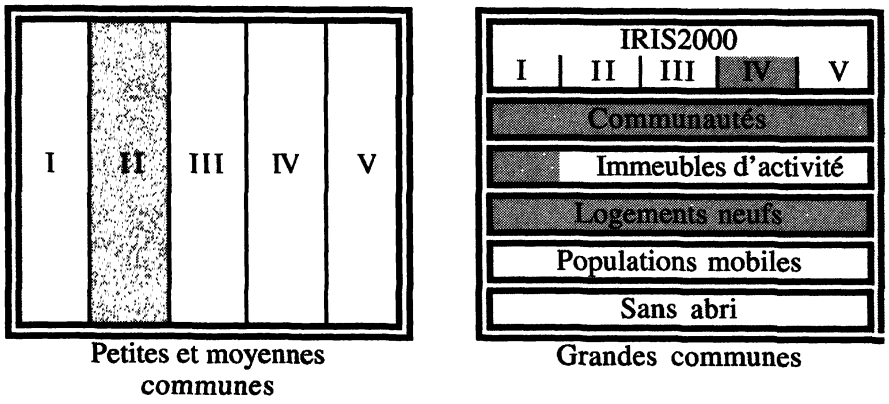


FIG 5. — Grille de collecte par groupe de rotation

Comme décrit plus haut, les groupes de rotation d'immeubles seront visités à tour de rôle au cours d'une période de 5 ans. Puis, pour les immeubles d'habitation du groupe de rotation de l'année en cours, on dressera la liste des logements qu'ils contiennent dont on tirera un échantillon au taux de 40 %. Finalement, les logements ainsi choisis seront invités à participer au recensement de l'année. En « grande commune », on aura la possibilité d'augmenter localement l'échantillon de ménages jusqu'à couvrir 100 % des logements du groupe de rotation.

En résumé, l'échantillon annuel comptera environ 8 millions de bulletins individuels, 6 millions des PMC et 2 millions des « grandes communes ».

2. Une unité de sondage est dite « autoreprésentative » si elle ne représente qu'elle-même, c'est-à-dire si son poids est 1.

2.3. Comparaison avec d'autres plans

Le plan de sondage en « grande commune » prévoit la création de cinq groupes de rotation d'immeubles. Le tirage équilibré sur les variables démographiques résulte en une dispersion de l'échantillon annuel sur tout le territoire de l'IRIS2000. L'efficacité de ce plan a été comparée à celle de six autres plans concurrents assujettis aux mêmes taux de sondage. Les résultats sont présentés aux tableaux 2 et 3. Le test a été conduit sur la commune de Sedan, commune d'environ 21.000 habitants en 1990 (environ 20.500 au RP99), constituée de onze IRIS. Il s'agit de la précision du seul sondage annuel, sans synthèse (voir section 5), exprimée par le coefficient de variation (en pourcentage). Sauf sur de très grands territoires, il n'est pas prévu à ce jour de publier les résultats directement obtenus des enquêtes annuelles en grande commune.

Pour permettre de mieux évaluer la performance de chaque plan, on a réalisé un premier tirage de logements au hasard simple au taux de 8 %, sans égards aux immeubles. Pour ce plan de référence, le Plan A, on a une précision relative de 2,1 % pour le « nombre total de personnes » et 4,5 % pour le « nombre de femmes âgées de 20 à 39 ans » pour un échantillon de taille 700 logements. Ce plan, possible lors de simulations, ne serait pas viable en réalité parce qu'un répertoire exhaustif des logements ne peut pas exister en permanence.

Le Plan B est un plan à deux phases, similaire au plan retenu pour le RRP, mais sans souci d'équilibre au premier tirage; pour ce plan donnant un échantillon final d'environ 700 logements, on atteint une précision relative de l'ordre de 7,0 % pour le « nombre total de personnes » et 9,1 % pour le « nombre de femmes âgées de 20 à 39 ans ». Le plan suivant (Plan C) est un plan à deux degrés : d'abord un tirage au hasard simple d'immeubles au taux de 1/5, puis un tirage au hasard simple de 40 % des logements de chaque immeuble choisi; si nécessaire, le taux de sondage des logements est ajusté à la taille de l'immeuble M_i :

$$m_i = \begin{cases} 1, & \text{si } M_i = 1 \\ 2, & \text{si } 2 \leqslant M_i \leqslant 5 \\ [40\ \%M_i] + 1, & \text{si } M_i > 5 \end{cases}$$

La taille d'échantillon résultante est d'environ 1 100 logements.

Pour les deux autres plans, on a créé des grappes de logements contigus (en admettant que la numérotation des immeubles sur les fichiers du recensement suive la continuité du terrain). Le découpage compte dix grappes de taille égale pour le quatrième plan (Plan D), il en compte vingt pour le cinquième (Plan E). Dans les deux cas, on sonde le cinquième des grappes au hasard simple, puis 40 % des logements des grappes. Les tailles d'échantillon résultantes sont respectivement de 685 et 700 logements. On notera que les Plans D et E supposent la fabrication a priori de blocs connexes, et que cette fabrication suppose à son tour une connaissance du nombre total de logements disponibles. Ces deux approches reposent donc sur un prérecensement complet de l'IRIS2000 avant le début de la collecte.

TABLEAU 2. — Comparaison de sept plans de sondage en grande commune, Sedan, « nombre de personnes »

IRIS	Nombre de personnes	Plan étudié						
		A	B	C	D	E	F	Equilibre
0101	855	8,7	17,2	6,1	6,7	6,6	44,0	7,0 %
0102	2 297	7,1	25,2	23,0	8,3	8,5	49,9	5,7 %
0201	1 673	7,1	17,9	14,1	10,2	8,7	39,2	5,7 %
0202	1 582	7,5	21,6	18,8	7,3	10,9	29,4	6,1 %
0301	2 226	6,5	16,8	14,5	5,7	7,2	37,7	5,2 %
0302	1 826	6,9	15,3	11,9	5,9	6,3	27,3	5,5 %
0303	1 862	6,6	20,2	18,1	6,5	9,2	48,2	5,4 %
0401	2 291	5,7	28,4	26,9	12,2	6,9	63,9	4,6 %
0402	1 777	6,5	12,0	5,4	5,2	5,1	41,1	5,2 %
0501	2 183	6,1	28,3	27,0	18,6	18,6	38,1	4,9 %
0502	2 753	6,6	21,8	19,6	9,1	12,4	52,8	5,4 %
Sedan	21 325	2,1	7,0	6,2	3,1	3,2	14,3	1,7 %

TABLEAU 3. — Comparaison de sept plans de sondage en grande commune : Sedan, « femmes de 20 à 39 ans »

IRIS	Nombre de personnes de 20 à 39 ans	Plan étudié						
		A	B	C	D	E	F	Equilibre
0101	96	29,2	32,5	17,5	31,0	28,4	56,3	23,6 %
0102	347	14,3	27,5	24,3	16,3	15,0	53,2	11,6 %
0201	247	18,4	29,8	25,1	26,9	22,7	46,1	14,8 %
0202	248	17,4	30,0	26,3	19,2	24,3	34,9	14,1 %
0301	382	14,8	22,5	19,8	14,8	14,5	40,1	11,9 %
0302	344	15,2	22,9	19,1	17,4	16,7	34,1	12,3 %
0303	371	14,1	24,0	21,4	17,8	17,0	61,2	11,4 %
0401	382	13,1	38,2	36,4	24,0	18,8	81,6	10,6 %
0402	237	18,6	21,6	13,2	18,0	18,4	45,9	15,1 %
0501	457	11,5	30,7	29,5	23,1	20,9	41,0	9,3 %
0502	517	10,7	26,9	25,3	25,6	22,7	50,8	8,6 %
Sedan	3 628	4,5	9,1	8,1	6,8	6,2	16,5	3,6 %

Enfin, le Plan F est aussi un plan sur des structures connexes puisqu'il s'agit d'un tirage d'un îlot sur cinq suivi d'un tirage de 40 % des logements des îlots choisis au premier degré.

On voit donc la qualité du plan proposé pour le RRP (intitulé « Équilibre » dans les tableaux 2 et 3) comparativement à d'autres plans, même le Plan C où l'on fait un sondage de logements dans chaque immeuble avec une taille d'échantillon finale augmentée de plus de 50 %.

2.4. Sondage par liste ou sondage aréolaire

Le choix de passer en grande commune par liste plutôt que par grappe aréolaire (c'est-à-dire l'ensemble des unités toutes incluses à l'intérieur de frontières reconnaissables sur le terrain : rues, cours d'eau, chemin de fer; c'est l'idée du recensement des unités constituant des îlots) s'inscrit dans la tradition des enquêtes auprès des ménages menées par l'INSEE (à l'exception de l'enquête-emploi). On peut par ailleurs invoquer nombre d'arguments théoriques et pratiques pour défendre cette position : réduction de la variance échantillonnale pour une taille d'échantillon donnée, accès à une liste d'immeubles à jour, meilleure maîtrise de la taille d'échantillon, constitution d'un échantillon maître, etc..

Le travail de constitution et de maintenance des listes d'immeubles n'est pas négligeable; cependant il est en partie assuré par le système de cartographie numérisée, en partie assuré par le retour des campagnes de collecte.

Le sondage par grappes élimine la nécessité d'une liste d'immeubles à jour puisqu'on prend le contenu du terrain au moment où l'on fait l'enquête. De tels plans offrent des réductions de frais de collecte par rapport à des sondages sur liste. Cependant, les grappes « naturelles » ont le désavantage de présenter peu de variation, c'est-à-dire peu d'information originale; on doit donc compenser par une augmentation, parfois substantielle, de la taille de l'échantillon. Par ailleurs, on avance souvent l'idée de l'assurance d'exhaustivité comme plaidant en faveur du sondage par aires. On doit se souvenir que les recensements traditionnels sont des enquêtes par aires, et il est habituel d'y trouver un taux de sous-dénombrement de l'ordre de 1 % à 5 %, particulièrement en milieu urbain.

Des tests sont actuellement (octobre 2000) en cours pour évaluer les pertes de précision et les coûts de constitution des grappes, si une telle option était retenue. Par ailleurs, on peut craindre un effet d'entraînement pervers, par lequel le refus d'une unité de la grappe compromet la participation des autres unités de la grappe. De telles situations ont été vécues au RP99 et lors des pré-tests de questionnaires à l'été 2000.

3. COLLECTE ET TRAITEMENT DES DONNÉES

L'option privilégiée actuellement est de collecter les *données* durant les mois de janvier et février A, pour une publication référencée au 1er janvier A.

Pour une commune donnée, la campagne de collecte ne devrait pas excéder 4 semaines en PMC ou 6 semaines en GC.

Le mode de collecte en « dépôt-retrait » qui a prévalu dans les recensements antérieurs est retenu. La possibilité de laisser le répondant retourner son (ses) questionnaire(s) par la poste en un point central dès les premières vagues est à l'étude. Cette possibilité, à l'initiative du répondant, suppose toutefois que l'agent recenseur puisse apporter son aide sur demande de répondant ; elle impose que le questionnaire soit d'une très grande lisibilité et convivialité.

Dans l'optique d'une collecte éclatée sur cinq années, il faudra modifier le questionnaire du recensement pour tenir compte de dates de collecte et de référence différentes ; ces modifications sont présentement à l'étude au sein d'un groupe de travail sur le questionnement placé sous l'égide du Conseil National de l'Information Statistique. Dans ce groupe, il est également question de la mesure des migrations sur des périodes fixes comparables à celles utilisées en Europe ou en Amérique (un ou cinq ans), conformément aux recommandations de l'ONU [1990].

Du fait de la non-simultanéité de la collecte, la prise en compte des doubles (ou multiples) résidences conduit à une refonte de cette partie du questionnaire. Dans un recensement traditionnel, cette réalité est ignorée puisque au moment du recensement, le même pour tous, chaque individu ne peut être rattaché qu'à une seule résidence principale. L'aménagement du questionnement sur la double résidence se justifie par la volonté de maîtriser d'éventuels doubles comptes. Bénéfice additionnel, cet aménagement permettrait de produire une statistique sur ce phénomène qu'on croit prendre de l'ampleur.

Encore d'autres aspects du questionnaire, bulletin individuel ou feuille de logement, méritent un examen attentif avant le début des opérations de collecte.

Parallèlement à la tenue du recensement, l'INSEE prévoit la conduite d'enquêtes « associées » au recensement. Il s'agit d'enquêtes d'intérêt général, dont la collecte devra utiliser le mode dépôt-retrait, leurs questionnaires étant distribués en même temps que les bulletins du recensement, auprès d'une partie de l'échantillon des personnes recensées une année donnée. Les enquêtes associées peuvent ne pas être obligatoires comme l'est le recensement. La teneur d'une enquête associée ne peut être telle qu'elle nuise à la bonne marche du recensement.

Comme le mode de collecte des enquêtes associées est celui du recensement, il ne peut s'agir d'une enquête dont la cible doive être identifiée sur le terrain par l'agent recenseur. Pour les enquêtes qui portent sur une population particulière et demandent une identification préalable des individus, le recensement pourra être utilisé comme filtre ou complété par une enquête-filtre permettant de réaliser une enquête spécifique quelques mois après la conclusion du recensement.

Les conditions d'ouverture et les modalités de définition des enquêtes associées et des enquêtes filtre sont actuellement à l'étude.

Les principales opérations nécessaires au traitement des données devraient suivre le modèle du recensement de 1999; il est encore trop tôt pour donner des détails précis sur les options étudiées ou à étudier.

4. ESTIMATION ET SYNTHÈSE

4.1. Estimations globales et estimations locales

La façon de collecter les renseignements au recensement rénové permet de produire deux séries d'estimations pour une année de référence donnée.

La première, dite d'**estimations globales**, produit des données de cadrage au niveau national, pour les régions et pour des grandes zones (à définir). Elles résultent de l'extrapolation des données collectées durant l'année de référence selon les règles habituelles dans l'exploitation des sondages. Par exemple une statistique portant sur l'ensemble des PMC s'obtiendra selon les règles d'extrapolation d'un sondage en grappes. Les estimations globales sont le résultat d'une seule campagne de collecte et devraient être publiées à la fin de l'année de référence; c'est-à-dire que les estimations globales pour l'année « A » sont publiées à la fin de l'année « A » à partir du sondage réalisé durant la campagne de l'année « A ».

La seconde, dite d'**estimations locales**, produit des données sur tout découpage du territoire. Les estimations locales sont les résultats amalgamés des cinq années de collecte et des modèles d'imputation massive. Disponibles elles aussi à la fin de l'année « A », les estimations locales se rapportent à l'année « $A-2$ », soit le temps nécessaire à la complétion de la collecte et à l'imputation massive des groupes de rotation collectés en « $A-4$ », « $A-3$ », « $A-1$ » et « A ».

4.2. Estimations globales

Les estimations globales pour les *GC* devraient être du type estimation par valeurs dilatées, aussi connues comme estimations de Horvitz-Thompson [Morin, 1993; Särndal *et coll.*, 1992]. En notant $y_{logement}$ la valeur prise par la variable d'intérêt Y pour un *logement* donné, $y_A, y_C, , y_M$ et y_N représentant respectivement la valeur de la variable d'intérêt pour un logement trouvé dans un immeuble d'activité, pour une communauté, pour une habitation mobile et pour un logement trouvé dans un immeuble neuf, et en notant $w_{logement}$ le poids associé à un *logement* donné, on peut écrire pour la grande commune *GC* :

$$\hat{Y}^{GC}_{annuel} = \sum_{Iris \in GC} \sum_{logement \in Iris} w_{logement} y_{logement} + \sum_{logement\ Neuf \in GC} y_{logement} + \sum_{Imm\ Activité \in GC} Y_A + \sum_{Communauté \in GC} \hat{Y}_C + \sum_{Mobiles \in GC} \hat{Y}_M$$

Dans le cas des PMC, on ne pourra produire des estimations globales que pour les communes recensées au cours de l'année; on peut noter Y_{PC} et w_{PC},

TABLEAU 4. — Précision attendue pour les estimations globales, exemple de la région Champagne-Ardenne

		Ardennes		Aube		Marne		Haute-Marne		Ensemble	
		Total	cv	Total	cv	Total	cv	Total	cv	Total	cv
HOMMES	Moins de 20 ans	42 186	12,10%	39 518	0,14%	76 153	6,97%	28 394	14,76%	186 251	0,14%
	20 à 39 ans	42 152	12,01%	42 221	0,14%	82 369	6,53%	28 912	15,01%	195 654	0,14%
	40 à 59 ans	32 256	12,23%	33 476	0,15%	60 123	7,22%	22 777	14,93%	148 632	0,15%
	60 à 74 ans	16 399	11,51%	17 124	0,20%	26 827	7,31%	12 411	13,46%	72 761	0,20%
	75 ans ou plus	6 447	12,11%	7 881	0,31%	11 256	7,58%	5 527	14,36%	31 111	0,31%
FEMMES	Moins de 20 ans	40 644	12,15%	37 596	0,14%	72 983	7,01%	26 693	14,91%	177 916	0,14%
	20 à 39 ans	40 902	11,99%	41 329	0,14%	83 231	6,25%	27 698	15,19%	193 160	0,14%
	40 à 59 ans	31 104	12,17%	32 874	0,14%	59 438	7,07%	21 748	15,18%	145 164	0,14%
	60 à 74 ans	19 419	11,71%	20 331	0,20%	32 411	7,11%	14 353	13,94%	86 514	0,20%
	75 ans ou plus	6 184	13,05%	13 883	0,24%	20 460	8,13%	9 828	15,91%	56 325	0,24%
RESIDENCE	PRINCIPALE	103 472	11,72%	111 001	0,26%	199 493	6,35%	74 351	14,71%	488 317	0,26%
	SECONDAIRE	5 907	11,45%	8 742	7,32%	7 805	11,01%	7 404	27,74%	29 858	7,32%
	VACANT	10 523	13,18%	10 409	2,45%	14 337	6,91%	7 259	13,06%	42 528	2,45%
TAILLE DES LOGEMENTS	1 ou 2 pièces	10 661	12,70%	17 274	1,81%	29 301	4,47%	9 345	15,83%	66 581	1,81%
	3 ou 4 pièces	51 669	12,51%	61 272	0,85%	97 853	6,34%	38 138	15,86%	248 932	0,85%
	de 5 à 9 pièces	41 130	11,46%	32 436	1,29%	72 329	7,79%	26 864	13,62%	172 759	1,29%
NB de VOITURES	0	27 025	13,00%	25 519	1,45%	44 244	5,59%	17 768	15,10%	114 556	1,45%
	1	54 525	12,06%	55 919	0,57%	102 778	6,67%	39 578	15,45%	252 800	0,57%
	au moins 2	21 910	11,09%	29 543	1,75%	52 455	7,41%	16 999	13,85%	120 907	1,75%

		Ardennes		Aube		Marne		Haute-Marne		Ensemble	
		Total	cv	Total	cv	Total	cv	Total	cv	Total	cv
ACTIVITE ECONOMIQUE	Agriculture	8 220	11,36%	10 500	13,57%	23 348	11,51%	6 728	12,96%	48 796	5,93%
	Industrie	30 503	15,52%	38 142	2,37%	48 104	8,86%	22 737	18,35%	139 486	2,37%
	Bâtiment génie civil	6 560	15,26%	8 320	2,67%	13 884	7,45%	4 504	14,21%	33 268	2,67%
	Services marchands	36 608	12,29%	40 716	1,31%	88 013	6,74%	26 704	17,30%	192 041	1,31%
	Services non marchands	18 012	12,17%	18 732	2,05%	39 332	7,59%	14 520	17,01%	90 596	2,05%
	Agriculteurs exploitants	6 748	11,94%	8 028	5,77%	16 084	11,43%	5 540	13,52%	36 400	5,77%
	Artisans, commerçants	7 159	14,03%	8 031	2,46%	13 691	9,62%	5 420	17,36%	34 301	2,46%
CATEGORIE SOCIALE	Cadres	6 520	11,31%	8 104	2,54%	19 816	6,52%	4 788	17,57%	39 228	2,54%
	Prof. Intermédiaires	18 148	12,42%	20 752	1,48%	41 576	6,81%	12 820	16,88%	93 296	1,48%
	Employés	29 588	12,85%	30 548	1,12%	63 392	6,94%	21 124	16,63%	144 652	1,12%
	Ouvriers	45 528	13,37%	54 139	1,30%	80 410	7,79%	32 897	16,30%	213 154	1,30%
	Retraités	42 423	11,98%	51 128	0,70%	73 940	7,45%	33 972	13,88%	201 463	0,70%
	Autres sans activité prof.	128 216	12,18%	105 796	0,53%	218 260	6,75%	82 590	14,39%	534 862	0,53%
	EMPLOI	100 037	11,80%	117 139	0,52%	213 447	6,83%	75 172	15,63%	505 795	0,52%
	CHÔMEUR	16 420	13,26%	14 190	1,71%	24 645	6,44%	8 548	14,08%	63 803	1,71%
	ETRANGER	14 623	24,22%	16 408	5,70%	21 889	5,34%	7 378	25,39%	60 298	5,70%
	FRANCAIS	269 040	11,79%	269 825	0,28%	503 362	6,90%	190 963	14,47%	1 233 190	0,28%
POPULATION TOTALE		283 663	11,96%	286 233	0,05%	525 251	6,79%	228 341	14,67%	1 293 488	0,05%

la valeur de la variable d'intérêt et le poids associés à une petite commune donnée. Ainsi, à partir des estimations globales tirées des GC et des PMC, on peut composer des estimations globales nationales, régionales

$$\hat{Y}_{annuel}^{National} = \sum_{Région} \left(\sum_{PC \in Région} W_{PC} Y_{PC} + \sum_{GC \in Région} \hat{Y}_{annuel}^{GC} \right)$$

Le Tableau 4 donne une idée de la précision à laquelle on peut s'attendre pour des estimations globales ; il faut se souvenir, en lisant le tableau, que les statistiques départementales ne seraient pas nécessairement publiées.

4.3. Estimations locales

Les résultats détaillés, ou estimations locales, relatifs à « $A-2$ » seront mis à disposition à la fin de l'année « A » ; ces résultats détaillés seront le fruit d'une combinaison entre les observations faites par sondage ou recensement et des données synthétiques. Dans le Tableau 5 (voir p. 27), on peut voir l'amélioration apportée à la précision en comparant le seul effet du sondage sur des estimations construites à partir d'une seule année de collecte et de 5 années de collecte. Rappelons que des données aussi détaillées ne seraient probablement pas publiées en n'utilisant qu'une seule année de collecte. Les estimations basées sur une année de collecte n'utiliseraient qu'un cinquième des petites et moyennes communes de la Drôme (équilibrées pour Rhône-Alpes, pas nécessairement pour ses départements) et que 8 % des logements de ses grandes communes. La précision du sondage est donnée dans les colonnes intitulées « 1 an ». En comparaison, les données détaillées construites sur 5 années de collecte utiliseraient toutes les petites communes et 40 % des logements des grandes communes. La précision relative pour cette série d'estimations est présentée dans la colonne intitulée « 5 ans ». Il faut cependant noter (voir infra) qu'il ne s'agit que d'une partie de la mesure de précision. En effet, après l'amalgame des 5 années de collecte, il est nécessaire de faire « vieillir » certaines années et d'en « rajeunir » d'autres au moyen de modèles de type régression. Une composante de variation, due à la qualité (plus précisément au manque de qualité) des modèles, viendra donc s'ajouter aux estimations présentées dans les colonnes « 5 ans » du Tableau 5. Encore une fois, la précision est une précision relative exprimée en terme de coefficient de variation.

Les données synthétiques seront obtenues à partir de la relation entre données observées et données administratives sur un même point en un même instant. À ce jour, il est prévu, sous réserve de l'avis de la CNIL, d'exploiter les fichiers administratifs à un niveau d'agrégation géographique assez détaillé (immeuble, îlot) qui renseigne sur les individus (âge, sexe d'après les fichiers de l'assurance maladie, ci-dessous CAM) ou leurs logements (fichiers de taxe d'habitation, ci-dessous TH).

En régime permanent, pour une PMC enquêtée en $A-5$ et A (voir Figure 6), on aura mesuré des variables sur les personnes (âge, sexe, activité,

TABLEAU 5. — Précision relative en utilisant 1 et 5 ans de collecte, Drôme et Romans

	Précision relative (cv en %) du sondage seul après 1 an et 5 ans de collecte					
Précision pour le domaine	Drôme 1990		Unité Urbaine de Romans 1990		Commune de Romans 1990	
Durée de la collecte	1 an	5 ans	1 an	5 ans	1 an	5 ans
Nombre de						
Hommes	14.4	0.14	47.8	0.57	3.4	0.9
Personnes de 0-19 ans	14.3	0.28	48.0	1.14	5.9	1.8
Personnes de 40-59 ans	14.1	0.24	45.2	0.93	4.1	1.5
Personnes mariées	14.6	0.15	48.3	0.62	3.3	1.0
Personnes divorcées	14.6	0.70	47.5	2.27	8.9	3.0
Pers. même logement en 82	14.4	0.20	46.3	0.85	4.1	1.3
Personnes ayant un emploi	14.6	0.16	48.3	0.67	3.5	1.0
Personnes retraitées	16.0	0.27	50.5	1.08	5.2	1.6
Personnes de nationalité française	14.8	0.10	48.4	0.41	2.9	0.6
Nombre total de personnes	14.5	0.11	47.8	0.42	2.9	0.6

Groupe de rotation					Année diffusée		Année courante
	A-6	A-5	A-4	A-3	A-2	A-1	A
GROUPE I			R		S		
GROUPE II				R	S		
GROUPE III					R		
GROUPE IV	R				S	R	
GROUPE V		R			S		R

R = recensement S = synthèse

FIG 6. — Groupes de rotation, synthèses et diffusion

profession,...) et sur les logements (taille du ménage, nombre de pièces, mode d'occupation, etc.) aux deux moments. Les fichiers administratifs fourniront des informations complémentaires à un niveau de détail assez fin. Dès lors, il sera possible de mesurer la différence entre ce qui a été observé et ce qui résulte de l'exploitation du fichier pour des objets similaires (immeubles, îlots, communes, etc.).

Cette différence, calculée sur des objets à la fois observables sur le terrain et repérables sur les fichiers administratifs, se traduira en facteur de correction à appliquer aux données administratives de sorte que la somme corrigée de celles-ci corresponde bien aux comptes ou estimations censitaires. Par exemple, si une source administrative indiquait une population de 150 personnes

pour une petite commune où le recensement en aurait observé 100 pour la même année, il faudrait appliquer une correction de 100/150 aux autres indicateurs fournis par la source administrative. De façon équivalente, le rapport ente les valeurs données par une source administrative pour deux années différentes pourrait être utilisé pour « mettre à jour » le résultat d'une observation censitaire. Par exemple, pour une année donnée, on aurait observé 100 personnes au recensement alors que la source administrative en indique 150 ; si l'année suivante, alors que la petite commune d'intérêt n'est plus dans le champ d'observation, la source administrative nous indiquait une population de 165 personnes, on pourrait « vieillir » le recensement par un facteur de 165/150. C'est l'idée de la synthèse, détaillée dans ce qui suit.

On peut tenir un raisonnement analogue en grande commune si on remplace une PMC par un « immeuble ».

Par exemple, pour l'année $A - 4$,

$$\begin{aligned} P_{A-4} &= CAM_{A-4}/CAM_{A-5} \\ \text{et } L_{A-4} &= TH_{A-4}/TH_{A-5} \end{aligned}$$

seraient les facteurs correspondant respectivement aux enregistrements censitaires « personnes » et « logements » de l'année $A - 5$. Il est facile de voir comment le modèle s'applique aux années $A - 3$ à $A - 1$.

De plus, l'estimation synthétique pour les années $A - 4$ à $A - 1$ pourrait profiter des informations recueillies durant la campagne de l'année A ; en effet, il serait possible de calculer des facteurs d'ajustement par rapport au plus récent recensement et obtenir

$$\begin{aligned} P_{A-2} &= CAM_{A-2}/CAM_A \\ \text{et } L_{A-2} &= TH_{A-2}/TH_A, \end{aligned}$$

et rétropoler sur la période intercensitaire.

Il est à peu près certain que les deux séries ne coïncideront pas. Toutefois, pour faciliter les utilisations, il est souhaitable de publier une et une seule série d'estimations pour toute zone pour tout moment. En conséquence, il apparaît naturel de produire une série « composite » dont les extrémités soient ancrées aux valeurs du recensement. Les combinaisons linéaires suivantes peuvent jouer ce rôle :

$$\begin{aligned} S_{A-4} &= 0,8 \times \text{Extrapolation}_{A-4} + 0,2 \times \text{Rétropolation}_{A-4} \\ S_{A-3} &= 0,6 \times \text{Extrapolation}_{A-3} + 0,4 \times \text{Rétropolation}_{A-3} \\ S_{A-2} &= 0,4 \times \text{Extrapolation}_{A-2} + 0,6 \times \text{Rétropolation}_{A-2} \\ S_{A-1} &= 0,2 \times \text{Extrapolation}_{A-1} + 0,8 \times \text{Rétropolation}_{A-1} \end{aligned}$$

Les figures 7 et 8 illustrent le comportement possible de ces séries, sachant que les sources administratives diffèrent du recensement d'environ 10 % en

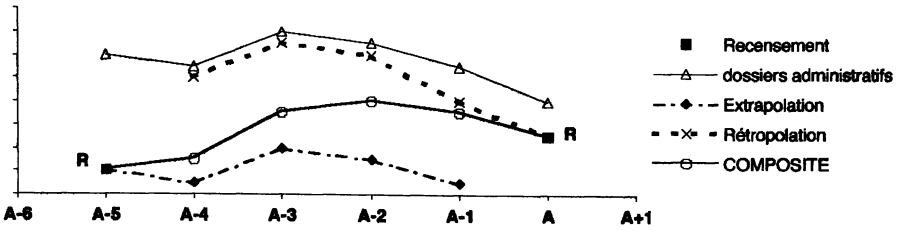


FIG 7. — Comportement des synthèses

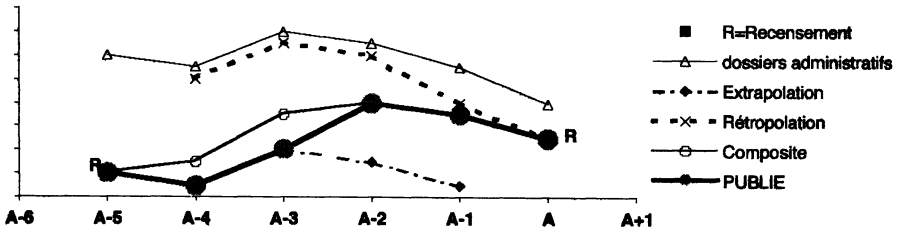


FIG 8. — Série publiée

début de cycle et de 1 % en fin de cycle, cinq ans plus tard; il s'agit ici de données fictives utilisées pour fins d'illustration.

Compte tenu de leur caractère ponctuel, les enquêtes associées ou les résultats issus des enquêtes réalisées après le filtre opéré lors du *RRP* ne peuvent pas bénéficier d'estimations par synthèse.

4.4. Imprécision due à la synthèse

Le diagramme de la Figure 9 illustre comment la production des estimations par synthèse utilise l'information amassée : d'abord une extrapolation d'un ou deux ans uniquement pour un recensement passé, pour deux groupes de rotation (I et II); ensuite l'utilisation directe des résultats du recensement pour un troisième groupe de rotation (III); enfin, la combinaison des extrapolations et rétropolations pour caler les deux derniers groupes (IV et V).

En se concentrant sur les groupes II et III aux années $A-3$ et $A-2$, on isole ainsi une partie de la Figure 9 :

	$A-3$	$A-2$	$A-1$
II	R	$\rightarrow$ S	
III		R	

On imagine utiliser le rapport de données administratives (notées F_{A-3} et F_{A-2}) sur deux années pour un groupe donné pour représenter le rapport des données du recensement (notées Y_{A-3} et Y_{A-2}) sur les deux mêmes années, $A-3$ et $A-2$; ceci nous permet de trouver la valeur synthétique pour le

GROUPE DE ROTATION					Année diffusée		Année courante
	A-6	A-5	A-4	A-3	A-2	A-1	A
GROUPE I			R		→S		
GROUPE II				R	→S		
GROUPE III					R		
GROUPE IV	R				→S S←	R	
GROUPE V		R	→S S←	→S S←	→S S←	→S S←	R

R = recensement →S = synthèse par extrapolation S← = synthèse par rétropolation

FIG 9. — Synthèse, extrapolation et rétropolation

groupe II en $A-2$, $\tilde{Y}^{\mathrm{II}}_{A-2}$:

$$\frac{F_{A-2}}{F_{A-3}} = \frac{Y^{\mathrm{II}}_{A-2}}{Y^{\mathrm{II}}_{A-3}} \Longrightarrow \tilde{Y}^{\mathrm{II}}_{A-2} = \left(\frac{F_{A-2}}{F_{A-3}}\right) Y^{\mathrm{II}}_{A-3}.$$

Après synthèse, on peut (re)construire l'estimation :

$$\begin{aligned} \tilde{Y}_{\bullet} &= \sum_{PC} \tilde{Y}_{PC} + \sum_{GC} \tilde{Y}_{GC} + \sum_{Autres} \tilde{Y}_{Autres} \\ &= 100\ \%\ PC + 10\ \%\ GC + Autres \end{aligned}$$

4.5. Un modèle pour la synthèse

Cette synthèse peut être formalisée sous l'angle d'un modèle de type non-réponse : la campagne annuelle s'apparente alors à un sondage à 100 % qui subirait 80 % de non-réponse, laquelle est palliée par le recours à l'imputation par le ratio. Si l'échantillon complet est noté s, les répondants sont notés r et les non-répondants sont notés $s-r$, on peut écrire

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{\beta} x_k & \text{si } k \in s-r \end{cases}$$

avec, en reprenant les notations usuelles, y_k une valeur observée, $y_{\bullet k}$ une valeur imputée et $\hat{\beta} = \dfrac{\bar{y}_r}{\bar{x}_r}$ le coefficient utilisé pour l'imputation. C'est-à-dire que le modèle d'imputation est décrit par

$$\xi : \begin{cases} y_k &= \beta x_k + \varepsilon_k \\ E(\varepsilon_k) &= 0 \\ v(\varepsilon_k) &= \sigma^2 x_k \end{cases}$$

Avec un tel modèle d'imputation, sous sondage aléatoire simple,

$$\hat{Y}_\bullet = \frac{N}{n}\sum y_{\bullet k} = \frac{N}{n}\left\{\sum_r y_k + \sum_{s-r}\hat{\beta}x_k\right\} = \dots$$
$$= N\frac{\bar{y}_r}{\bar{x}_r}\bar{x}_s$$

L'incertitude autour de l'estimation avec imputation dépend des aléas de sondage et de la qualité du modèle d'imputation ξ :

$$(\hat{Y}_\bullet - Y) \quad = \quad (\hat{Y} - Y) \quad + \quad (\hat{Y}_\bullet - Y)$$

$$\begin{matrix}\text{incertitude}\\ \text{totale}\end{matrix} \quad = \quad \begin{matrix}\text{incertitude}\\ \text{du sondage}\end{matrix} \quad + \quad \begin{matrix}\text{incertitude}\\ \text{du modèle}\end{matrix}$$

Cela suppose [Särndal, 1990] que l'imputation se fasse sans biais mais pas que le modèle de réponse soit non informatif :

$$E_\xi E_s E_r(\hat{Y}_\bullet - Y) = 0$$

Donc,

$$V_{totale} = E_\xi E_s E_r(\hat{Y}_\bullet - Y)^2 = \dots$$
$$= E_\xi E_s E_r(\hat{Y} - Y)^2 + E_\xi E_s E_r(\hat{Y}_\bullet - Y)^2$$
$$= E_\xi V_s + E_s E_r V_\xi$$
$$V_{totale} = V_{échantillon} + V_{imputation}$$

Pour de nombreux modèles d'imputation, l'utilisation de données imputées comme si elles avaient été observées dans le calcul de l'estimation de V_s mène à une sous-estimation de $V_{échantillon}$. En espérance,

$$E_\xi(\hat{V}_s - \hat{V}_{\bullet s}) = V_{dif}$$

Par ailleurs, on devrait pouvoir trouver un estimateur sans biais pour $V_{imputation}$.

Sous les hypothèses de tirages aléatoires simples et de modèle d'imputation ξ, on peut écrire :

$$E_{sondage} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

$$V_{imputation} = E_s E_r E_\xi\left\{(\hat{Y}_\bullet - \hat{Y})^2 | s, r\right\}$$
$$= E_s E_r \sum_r \left\{E_\xi\left((\hat{\beta}x_k - y_k)^2 | s, r\right)\right\}$$
$$= E_s E_r \left\{\text{variance de } \hat{\beta}, \text{ sous-échantillon de taille } m\right\}$$
$$= E_s E_r \left\{N^2\left(\frac{1}{m} - \frac{1}{n}\right)\frac{\bar{x}_s\bar{x}_{s-r}}{\bar{x}_r}\sigma^2\right\}$$

Pour les estimateurs de ces variances, on obtient

$$\hat{V}_{sondage} = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \{s_\bullet^2 + C_0 \hat{\sigma}^2\}$$

avec C_0 près et $\left(1 - \frac{m}{n}\right) \bar{x}_{s-r}$ et $\hat{\sigma}^2$ près de $\dfrac{\sum\limits_r e_k^2}{\sum\limits_r x_k}$ et

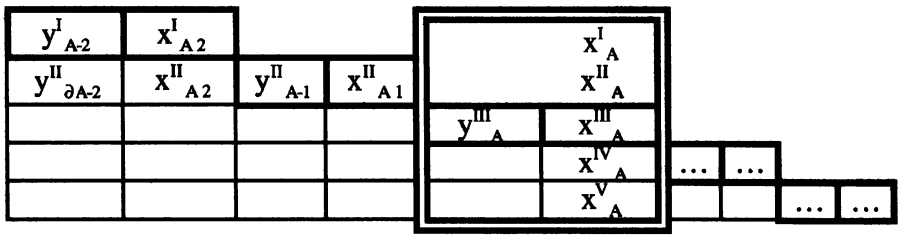
$$\hat{V}_{imputation} = N^2 \left(\frac{1}{m} - \frac{1}{n}\right) A \bar{x}_s \hat{\sigma}^2,$$

avec $A = \dfrac{\bar{x}_{s-r}}{\bar{x}_r}$, qu'on peut comprendre comme un effet de sélection des répondants. On remarque que, si $x_k \approx 1$, alors on n'impute pas et on obtient un sondage à deux phases de taille m parmi n et n parmi N. De plus, si $s = r$, $V_{totale} = V_{sondage}$.

Dans le modèle de Särndal, les x et y sont contemporains; à tout le moins, on aura observé certains des y.

y_k	x_k	m répondants
$y_{\bullet k}$	x_k	$n - m$ imputations

Dans le modèle du RRP, tout n'est pas synchrone :



En effet, $y^{II}_{\bullet A-2}, x^{II}_{A-2}, y^{II}_{A-1}$, et x^{II}_{A-1} ne sont pas toutes mesurées ou observées la même année. En fait, quand on regarde un seul groupe de rotation, on a un échantillon de taille n en $A-1$ et un échantillon identique en $A-2$ entièrement non-répondant. En conséquence, certains paramètres de l'estimation de V_{totale} ne sont plus calculables.

En revanche, en regardant le problème pour un temps donné, on a bien un échantillon de taille n répondants et $4n$ non-répondants. On pourrait approcher l'incertitude du processus d'imputation asynchrone (marqué par des pointillés) par celle du processus d'imputation synchrone (marqué par une ligne double)

Cette approche a été testée sur les petites communes de Rhône-Alpes, pour lesquelles les groupes de rotation, la Taxe d'Habitation (TH90) et le Recensement général de la population de 1990 (RGP90) sont disponibles. Des modèles de régression ont été mis en place pour estimer l'erreur additionnelle due à l'imputation massive des variables relatives au ménage ou au logement en utilisant comme régresseur le nombre de résidences principales indiqué par la TH90. Les résultats apparaissent au Tableau 6 [Kauffmann, 2000].

TABLEAU 6. — Précision relative due à l'imputation

Variable	Valeur au RP90	Erreur relative
Nombre de résidences principales	1 055 264	0,3 %
Nombre de résidences principales		
– de type maison individuelle	692 467	3,6 %
– de type logement collectif	268 784	10,8 %
– de type ferme	58 759	21,3 %
– de type autre[3]	35 254	122,5 %
– occupées par le propriétaire	686 673	11,4 %
– louées ou sous-louées	299 517	8,0 %
– occupées à titre gratuit	69 074	23,5 %
Nombre de résidences principales		
– de 1 pièce[4]	32 571	177,7 %
– de 2 pièces	88 353	77,1 %
– de 3 pièces	202 813	32,5 %
– de 4 pièces	325 906	16,5 %
– de 5 pièces ou plus	405 621	11,3 %
Nombre de résidences principales		
– ayant WC à l'extérieur	54 198	30,9 %
– ayant WC à l'intérieur	1 001 066	7,9 %
– n'ayant ni baignoire, ni douche	60 752	13,4 %
– ayant baignoire ou douche	816 435	9,5 %
– disposant d'une douche	178 077	8,4 %
– ayant le chauffage central collectif	155 318	35,2 %
– ayant le chauffage central individuel	615 603	5,2 %
– sans chauffage central	284 343	9,2 %
Nombre de ménages disposant de deux voitures ou plus	392 220	13,5 %

3. La taille de l'erreur s'explique par la présence de quatre communes dans le même groupe de rotation (GR) comptant plus de 200 telles résidences alors que les autres varient entre 0 et 160. Dans tous les GR, l'intervalle interquartile est de 10.

4. La taille de l'erreur s'explique par la présence d'une commune comptant 1 324 telles résidences alors que les autress varient entre 0 et 500. Dans tous les groupes de rotation, l'intervalle interquartile est de 6.

Population des ménages, ménages dont la personne de référence est		
– agriculteur exploitant	131 944	20,2%
– artisan, commerçant, chef entreprise	283 600	19,6 %
– cadre, profession intellectuelle supérieure	283 376	17,9 %
– professions intermédiaires	470 948	10,5 %
– employé	209 892	12,9 %
– ouvrier	874 176	6,6 %
– retraité	569 420	2,7 %
– sans activité	92 188	20,3 %

5. TRAVAUX EN COURS ET PERSPECTIVES D'ÉVOLUTIONS

Les travaux de méthodologie autour de la rénovation du recensement ne sont pas terminés. De nombreux chantiers sont ouverts. Les plus importants semblent être les suivants :

- il est nécessaire d'étudier et d'établir les protocoles de décision qui feront passer une commune de « petite »à « grande », et donc de mettre en branle toutes les activités nécessaires à la création des zones de type « IRIS », l'extension de la cartographie numérisée et du RIL, et la création des groupes de rotation d'immeubles ; réciproquement, on doit mettre au point un protocole de décision et traitement des grandes communes qui auraient passé sous le seuil des 10.000 habitants et qui devraient donc être insérées dans un groupe de rotation de PMC ;
- il est nécessaire d'étudier l'utilité et l'universalité de la stratification des petites communes en région ; la diversité des régions fait qu'il est très difficile de fixer un seul protocole de stratification des PMS efficace pour l'ensemble des régions ;
- il faut étudier la mise à jour et la maintenance des bases de sondage et des échantillons ; en particulier l'incorporation de nouveaux objets (communes en PMC, immeubles en GC) dans les groupes de rotation est actuellement à l'étude ; une méthode qui permettrait la mise à jour des groupes de rotation sans trop perturber l'équilibre des groupes devrait être proposée dans les mois qui viennent ;
- la confection d'un fichier de dépouillement pondéré annuel totalisable est commencée ; ce fichier servirait de banc test pour trouver « tous les chiffres du recensement » et simulerait une collecte sur plusieurs années et leur

« mise à jour » par synthèse. Au nombre des chantiers qui seront ouverts au courant de l'année 2001, on peut noter :

- l'étude plus détaillée des modèles utilisés pour l'imputation massive et la synthèse, et la précision qu'on peut escompter ;
- l'évaluation de la précision des estimateurs ; et
- les divers tests de prérecensement, de procédures de collecte, de maquettes de questionnaire ; des modèles de réponse et de mesure des charges de travail sur le terrain devraient avoir lieu courant 2001.

RÉFÉRENCES BIBLIOGRAPHIQUES

BERTRAND P. (2000), *Estimations annuelles dans la rénovation du recensement de la population*, note de travail interne, Département de la démographie, INSEE.

BORCHSENIUS L. (2000), « From a Conventional to a Register-based Census of Population », Les Recensements après 2001, Séminaire Eurostat-INSEE, Paris.

DEVILLE J.C., Tillé Y.,(1999) *Balanced Sampling by Means of the Cube Method*, CREST-ENSAI, document interne, soumis pour publication.

DEVILLE J.C., TILLÉ Y. (2000), « Échantillonnage équilibré par la méthode du cube et estimation de variance », *Journées de Méthodologie*, décembre 2000, INSEE, Paris.

JACOD M., DEVILLE J.C. (1996), « Replacing the Traditional French Census by a Large Scale Continuous Population Survey », *Annual Research Conference Proceedings*, USBC, Washington.

KISH L. (1990), « Recensement par étapes et échantillons avec renouvellement complet », Techniques d'enquêtes, Vol 16, N° 1, pp. 67-86, Statistique Canada, Ottawa, juin 1990.

KAUFFMANN B. (2000), *Estimation de la précision due au modèle de synthèse*, note de travail interne, Département de la démographie, INSEE.

LAIHONEN A. (2000), « 2001 Round Population Censuses in Europe », *Les Recensements après 2001*, Séminaire Eurostat-INSEE, Paris.

MORIN H. (1993), *Théorie de l'échantillonnage*, Presses de l'Université Laval, Québec.

NATIONAL ACADEMY of SCIENCES (1994), « Radical Alternatives », *Modernizing the U.S. Census*, B. Edmonston et C. Schultze, éditeurs ; Panel on Census Requirements in the Year 2000 and Beyond, National Research Council, National Academy Press, pp. 59-74.

ONU (1990), *Principes et recommandations complémentaires concernant les recensements de la population et de l'habitat*, Études statistiques, ST/ESA/STA/série M/67, New York.

SÄRNDAL C.E.(1990), « Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation », *Recueil du Symposium 90 de Statistique Canada : Mesure et amélioration de la qualité des données*, Ottawa, octobre 1990, pages 369-380.

SÄRNDAL C.E., B. SWENSSON, J. WRETMAN (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.