JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

B. RIANDEY
MH. GENSBITTEL

Séminaire méthodologique SFdS-INSEE sur la rénovation du recensement de la population (RRP). Actes de la séance du 25 juin 2002

Journal de la société française de statistique, tome 143, n° 3-4 (2002), p. 111-153

<http://www.numdam.org/item?id=JSFS_2002__143_3-4_111_0>

SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE SUR LA RÉNOVATION DU RECENSEMENT DE LA POPULATION (RRP)

Actes de la séance du 25 juin 2002

RÉSUMÉ

La loi n° 2002-276 du 27 février 2002 relative à la démocratie de proximité constitue désormais le socle juridique du recensement de la population en France.

Pour les communes dont la population est inférieure à 10 000 habitants, les enquêtes sont exhaustives et ont lieu chaque année par roulement au cours d'une période de cinq ans. Pour les autres communes, une enquête par sondage est effectuée chaque année ; la totalité du territoire de ces communes est prise en compte au terme de la même période de cinq ans.

Les échantillons successifs, appelés *groupes de rotation*, présentent des garanties de représentativité aussi bien pour des estimations nationales et régionales annuelles, fondées sur la collecte de l'année, que pour des estimations détaillées fondées sur cinq années successives de collecte.

Le dossier analyse en détail les méthodes d'échantillonnage pour la catégorie des communes de moins de 10 000 habitants et pour celle des communes de 10 000 habitants ou plus. Il présente, en s'appuyant sur des exemples et des simulations, les résultats disponibles aux différents niveaux géographiques. Il fournit enfin des premières indications sur la précision des résultats et illustre l'intérêt de disposer de séries annuelles de données.

ABSTRACT

The law n° 2002-276 of February 27, 2002, relating to the democracy of proximity, is from now the legal basis of the population census in France.

For the communes whose population is lower than 10 000 inhabitants, the investigations are exhaustive and take place each year by rotation over a five years period. For the other communes, a sample survey is carried out each year; the whole territory of these communes is taken into account at the end of the same five years period.

The successive samples, called *groups of rotation*, present guarantees of representativeness for annual estimates at national or regional level founded on the data collection of the year, as well as for detailed estimates founded on five successive years of data collection.

The set of papers examines in detail the sampling procedures for both categories of communes. By means of examples and simulations, it gives the results available to the various geographical levels. Finally, it provides first indications on the precision of the results, and illustrates the interest to have series of annual data to one's disposal.

Introduction

La rénovation du recensement de la population, déjà évoquée dans ce Journal à deux reprises, continue de susciter un vif intérêt chez les statisticiens.

La réforme présentée est ambitieuse et ingénieuse : répartir la charge d'enquête sur cinq ans en fournissant des données fiables chaque année nécessite une technique de pointe et une organisation sans faille.

Mais les utilisateurs, s'ils sont rassurés de savoir que l'édifice reposera sur des bases solides, s'intéressent d'abord aux résultats et aux données produites. Le but du séminaire organisé le 25 juin 2002 était donc aussi de leur donner un aperçu complémentaire de ce que le recensement rénové leur offrirait.

Les textes présentés ci-après reprennent les exposés et les débats qui ont eu lieu lors de ce séminaire, et viennent compléter ceux qui ont été antérieurement publiés dans le présent Journal de la SFdS (vol 140, n° 4 et vol 142, n° 3). Ils présentent, outre l'actualité de la rénovation, le nouveau plan de sondage (tenant compte des tests réalisés antérieurement) et les principes de mise à disposition, chaque année, des données relatives aux différents niveaux géographiques.

La Société Française de Statistique est heureuse de pouvoir offrir un cadre à ces séminaires, qui permettent aux statisticiens et aux utilisateurs de s'informer et de s'exprimer ; elle remercie vivement l'équipe de l'INSEE responsable de la rénovation du recensement pour son importante contribution.

B. Riandey
MH. Gensbittel

LES ACTUALITÉS TECHNIQUES ET MÉTHODOLOGIQUES DU PROGRAMME

Jean-Michel DURR *

Il s'agit ici de relater les événements importants intervenus dans la vie du programme de rénovation du recensement de la population depuis le séminaire méthodologique SFdS - INSEE sur la rénovation du recensement de la population d'octobre 2001, dont les actes ont été publiés dans le Journal de Société Française de Statistique (Volume 142, n° 3).

L'événement majeur est la promulgation de la loi n° 2002-276 du 27 février 2002 relative à la démocratie de proximité, dont le titre V (« Des opérations de recensement », articles 156 à 158) constitue désormais le socle juridique du recensement de la population en France.

Le législateur a amendé sur deux points le texte qui lui était proposé par le gouvernement :

– d'une part, il a introduit dans la loi le seuil de 10 000 habitants à partir duquel les communes seront recensées au moyen d'enquêtes annuelles par sondage et non pas de manière exhaustive tous les cinq ans ;

– d'autre part, il a créé une commission spéciale du Conseil national de l'information statistique (CNIS) dont l'avis [1] sera pris pour la détermination des modalités de réalisation des enquêtes par sondage préalablement à la signature du décret d'application de la loi.

L'Insee se trouve donc maintenant en phase de réalisation pour lancer les premières enquêtes de recensement en janvier et février 2004. Cela impose la définition de priorités et la réalisation d'arbitrages.

Les priorités se sont organisées autour :

– du plan de sondage dans les communes de 10 000 habitants ou plus :

lors du séminaire du 24 octobre 2001, j'avais expliqué comment les tests réalisés par sondage de logements avaient conduit l'Insee à décider de sonder à l'adresse, par immeubles entiers (*cf.* « Les tests de la collecte menée en

* INSEE, programme de rénovation du recensement de la population, 18 boulevard A. Pinard, 75675 PARIS CEDEX 14
e-mail . jean michel durr@insee.fr

1 Les travaux de la commission spéciale du CNIS se sont achevés en septembre 2002 et le président de la commission, M GIBLIN, a adressé le rapport au ministre de l'économie, des finances et de l'industrie, président du Conseil national de l'information statistique, ainsi qu'au vice-président du CNIS et aux membres de la commission.

2001 : résultats et décisions», J-M DURR, Journal de la SfdS, *op. cit.*). Jean-Marie GROSBRAS explique, dans sa contribution (*cf. infra*), les modalités du nouveau plan de sondage et les dispositions prises pour limiter les effets de grappe ;

– de la base de sondage pour les échantillons d'adresses :

cette base est constituée par le répertoire d'immeubles localisés (RIL), qui est constituée à partir des informations collectées au cours du recensement de la population de mars 1999 et qui est ensuite mis à jour, en collaboration avec les communes de plus de 10 000 habitants. L'objectif est de tirer les échantillons d'adresses des premières enquêtes de recensement dans un RIL validé avec elles en juin 2003 ;

– de la mise au point des procédures de collecte dans les ménages :

pour ce faire, des tests sont réalisés cette année encore en métropole et dans les DOM, à la fois dans des communes de moins de 10 000 habitants et dans des communes dépassant cette taille. Ces tests sont riches d'enseignements. En particulier, ils confortent l'idée de réaliser une tournée de reconnaissance préalable au dépôt des questionnaires, au cours de laquelle l'agent recenseur note le nombre de logements qu'il localise à chaque adresse qui lui est confiée. Cette tournée améliore manifestement la qualité de la collecte ;

– de la mise au point de la collecte dans les communautés :

cette collecte sera prise en charge par l'Insee, qui constitue à cet effet un répertoire dont la mise à jour sera effectuée en continu.

La réalisation des premières enquêtes de recensement début 2004 impose de conduire à leur terme, d'ici là, les multiples projets du programme de rénovation :

– projets juridiques, puisqu'il faut préparer et obtenir les accords gouvernementaux sur un décret en Conseil d'État, des décrets simples et de nombreux arrêtés, notamment pour définir les traitements de données individuelles qui feront l'objet de demandes d'avis à la Commission nationale de l'informatique et des libertés (CNIL) ;

– projets organisationnels, pour arrêter de bonnes procédures de collecte à proposer aux communes et les meilleures organisations au sein des directions régionales de l'Insee ;

– applications informatiques nombreuses ;

– communication accompagnant les enquêtes de recensement pour que celles-ci soient bien accueillies par la population ;

– impression et routage des questionnaires à faire parvenir chaque année à plus de 8 000 communes ; saisie des questionnaires collectés ;

etc.

Au début de 2003, un test à grande échelle est programmé avec une centaine de communes dispersées dans une douzaine de régions (cette opération permettra, d'une part, de roder les procédures et, d'autre part, de tester l'application informatique de pilotage de la collecte : suivi des tâches des différents acteurs, enregistrement d'indicateurs, communication entre acteurs, documentation).

Les premières enquêtes de recensement ayant lieu en 2004, les premières exploitations démarreront aussitôt et les résultats détaillés seront livrés pour la première fois fin 2008, puis régulièrement chaque année ensuite.

On se propose d'examiner dans ce qui suit :

– en premier lieu, le détail des plans de sondage adoptés de part et d'autre du seuil de 10 000 habitants ;

– en second lieu, l'état actuel de nos travaux sur les données produites par le recensement rénové : que nous disent-elles, indépendamment des informations complémentaires qui seront apportées par les fichiers administratifs ? Ces travaux sont loin d'être achevés, mais l'urgence est moindre puisque les premiers résultats détaillés seront produits après cinq années de collecte, c'est-à-dire fin 2008.

SÉMINAIRE MÉTHODOLOGIQUE SFdS INSEE

LES PLANS DE SONDAGE

Jean-Marie GROSBRAS *

1. Introduction

La loi n° 2002-276 du 27 février 2002 définit dans son titre V (« Des opérations de recensement ») le cadre général des enquêtes de recensement et de la production des chiffres de la population, en particulier au paragraphe VI de l'article 156 :

VI. – Les dates des enquêtes de recensement peuvent être différentes selon les communes.

Pour les communes dont la population est inférieure à 10 000 habitants, les enquêtes sont exhaustives et ont lieu chaque année par roulement au cours d'une période de cinq ans. Pour les autres communes, une enquête par sondage est effectuée chaque année ; la totalité du territoire de ces communes est prise en compte au terme de la même période de cinq ans. Chaque année, un décret établit la liste des communes concernées par les enquêtes de recensement au titre de l'année suivante.

L'idée de base repose donc sur l'observation, chaque année, d'une fraction de la population, choisie grâce à des méthodes de sondage plutôt que sur une observation exhaustive tous les 7 à 9 ans. Un des objectifs de la rénovation est de continuer à publier les résultats du recensement sur toute portion du territoire (même infracommunale), dans les limites imposées par le respect du secret statistique, comme dans un recensement traditionnel. L'apport de données administratives et la modélisation permettront d'affiner la précision des estimations effectuées à partir des enquêtes de recensement.

La modification des méthodes employées pour le recensement de la population doit s'effectuer à coût constant par rapport aux recensements traditionnels. Un calcul simple montre alors que la rénovation du recensement devra se faire à raison d'environ 8,4 millions de bulletins individuels par année, soit 60 millions de bulletins en sept ans, ce qui correspond à l'effort de collecte d'un recensement général chaque sept ans.

La loi précise que les communes de moins de 10 000 habitants (PMC) sont enquêtées exhaustivement par roulement sur cinq ans. Ces quelque 36 000 communes comprenaient en 1999 environ 30 millions d'habitants (29 900 000).

* INSEE, programme de rénovation du recensement de la population, 18 boulevard A. Pinard, 75675 PARIS CEDEX 14
e-mail · jean marie.grosbras@insee.fr

Journal de la Société Française de Statistique, tome 143, n° 3-4, 2002

Il y aura donc, chaque année, 6 millions de bulletins recueillis. Restent donc 2,4 millions pour les autres communes (GC), soit un taux de sondage annuel de 8 % et de 40 % en cinq ans, puisque ces communes sont actuellement au nombre de 860 avec une population totale de près de 29 millions d'habitants (28 800 000) :

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

On obtient bien un taux moyen de sondage annuel de l'ordre de 1/7 des logements. Au terme d'une période quinquennale, on aura collecté environ 41 500 000 bulletins, soit tout près de 70 % de la population, c'est-à-dire la totalité des communes de moins de 10 000 habitants et 40 % de la population des communes dépassant cette taille.

En régime de croisière seront produits à la fin de l'année n deux ensembles de résultats :

– les résultats du recensement, estimations détaillées aux niveaux communal et infra-communal, à valeur pour le premier janvier de l'année $n-2$, c'est-à-dire au point médian d'un cycle de cinq ans ;

– des estimations nationales et régionales, appelées autrefois estimations globales, basées sur la seule collecte de l'année, à valeur pour le premier janvier de l'année n.

Les échantillons successifs, appelés *groupes de rotation*, doivent donc présenter des garanties de « représentativité » annuelle pour l'objectif des estimations nationales et régionales et de « représentativité » quinquennale pour les résultats du recensement ou estimations détaillées. La suite du document présente les méthodes d'échantillonnage pour la strate des communes de moins de 10 000 habitants et pour celle des communes de 10 000 habitants ou plus.

2. Plans pour les communes de moins de 10 000 habitants : les groupes de rotation

2.1. La question posée

Toutes les communes de moins de 10 000 habitants sont enquêtées exhaustivement, par roulement, à raison d'une sur cinq chaque année. L'objectif est de définir les « groupes de rotation » annuels, constitués de façon à assurer une qualité optimale des estimations globales annuelles, concernant les données sur la population et les logements.

2.2. La méthode

La méthode statistique utilisée est celle des échantillons équilibrés. Généralisant la notion de stratification, la méthode consiste à choisir des structures de référence, et à construire des échantillons reproduisant, le plus fidèlement possible, ces structures.

Dans le cas présent, les structures de référence sont à choisir parmi les variables démographiques et les catégories de logements. Les valeurs cibles sont établies à partir du recensement de 1999 (RP99). En d'autres termes, on fait l'hypothèse, par exemple, qu'un ensemble de communes dont la population, en 1999, a une structure par âge identique à celle de l'ensemble des communes de moins de 10 000 habitants conservera, au moins pendant un certain temps, une bonne qualité de représentativité sur ce critère.

À quel niveau géographique peut-on assurer une bonne représentativité ? Le problème statistique posé s'exprime en termes de « degrés de liberté ». Pour faire image, on peut se figurer un ensemble de cinq balances dont on veut équilibrer les plateaux à la même hauteur. Les poids placés dans les plateaux sont les communes, et elles sont de tailles disparates. Intuitivement, on voit qu'un équilibrage correct suppose que l'on ait suffisamment de poids à répartir, c'est-à-dire suffisamment de degrés de liberté.

La contrainte de degrés de liberté ne peut être satisfaite au niveau départemental. Elle peut l'être correctement dans des départements pourvus d'un grand nombre de communes de moins de 10 000 habitants mais pas dans les autres. Pour appliquer un principe homogène sur le territoire, on a donc retenu un niveau d'équilibrage régional.

2.3 Les variables de référence

Pour équilibrer les groupes de rotation des communes de moins de 10 000 habitants dans chacune des régions, les variables suivantes, issues du recensement de population de 1999, ont été retenues :

le nombre de logements ;

le nombre de logements en immeuble collectif ;

pour chacun des départements de la région, la population totale ;

la population des personnes de moins de 20 ans ;

la population des personnes de 20-39 ans ;

la population des personnes de 40-59 ans ;

la population des personnes de 60-74 ans ;

la population des personnes de 75 ans ou plus ;

la population des femmes ;

la population des hommes.

Les variables de type « logement » permettent d'obtenir l'équilibre entre groupes de rotation sur la proportion de logements dans le collectif. Cela a une influence sur la répartition des « grandes petites communes » dans les groupes de rotation. Cela permet également d'obtenir des groupes de rotation qui évolueront de façon plus homogène.

Les variables de « stratification » au niveau départemental permettent de mieux répartir les populations départementales dans les groupes de rotation. Dans chacun d'eux, chaque département est représenté proportionnellement à son poids de population vivant dans des communes de moins de 10 000

habitants. Enfin, les tranches d'âge et le sexe, variables de population essentielles, améliorent l'homogénéité des groupes de rotation pour la structure de population.

Les tableaux suivants présentent, d'une part, des résultats relatifs à la répartition des communes de la région de Basse-Normandie et, d'autre part, un premier indicateur de la stabilité des groupes régionaux entre 1990 et 1999.

Exemple d'équilibrages pour la région de Basse-Normandie.

	Effectif				
Groupes	**1**	**2**	**3**	**4**	**5**
Femmes de 0 à 19 ans	26 209	25 334	26 354	26 377	25 820
Femmes de 40 à 59 ans	26 943	26 248	26 913	27 257	26 344
Femmes de 60 à 74 ans	17 414	17 042	17 702	17 780	17 204
Femmes de 75 ans ou plus	10 709	10 527	10 994	11 193	10 621
Hommes de 60 à 74 ans	15 567	15 222	15 399	15 706	15 197
Population totale	210 432	204 570	212 081	213 705	206 910
Population active ayant un emploi en 99	82 868	79 910	83 366	83 130	81 275
Nb résid. princ. avec baignoire douche 99	79 021	75 614	77 786	79 456	75 995
Nombre total de chômeurs en 1999	9 840	9 808	9 894	10 757	9 861
Population étrangère en 1999	2 145	2 194	1 712	1 970	1 864
Population active féminine en 1999	41 842	40 216	42 196	42 220	41 113
Population active masculine en 1999	50 866	49 502	51 064	51 667	50 023
Nombre de logements en 1999	107 594	107 535	107 200	115 848	109 074
Résidences principales louées ou sous-louées en 1999	25 952	24 548	24 843	27 587	24 769
Population arrivée dans la commune entre 90-99	77 306	73 901	75 820	76 593	74 241
Résidences principales de 2 pièces	7 227	6 767	6 580	7 341	6 791
Résidences principales de 5 pièces ou plus	34 398	32 717	34 066	33 400	33 140
Résidences principales en propriété	53 491	51 629	53 495	52 552	51 958
Résidences principales sans chauffage central	18 657	18 383	18 884	18 459	18 945
Actifs travaillant dans la commune de résidence	5 615	5 422	5 748	6 362	5 846
Logements vacants	25 797	25 314	25 921	27 920	24 730
Ménages sans voiture	11 294	10 793	11 161	12 159	10 663
Ménages disposant de deux voitures	30 916	29 457	31 097	30 039	29 581

Taux d'évolution de la population des groupes entre 1990 et 1999.

Groupes	1	2	3	4	5
Ile-de-France	10 %	12 %	10 %	10 %	11 %
Champagne-Ardenne	−1 %	−1 %	0 %	0 %	1 %
Picardie	3 %	3 %	3 %	4 %	4 %
Haute-Normandie	5 %	3 %	4 %	3 %	5 %
Centre	5 %	5 %	4 %	3 %	5 %
Basse-Normandie	5 %	3 %	4 %	3 %	4 %
Bourgogne	1 %	3 %	1 %	0 %	1 %
Nord - Pas-de-Calais	0 %	0 %	2 %	0 %	2 %
Lorraine	1 %	1 %	1 %	0 %	1 %
Alsace	8 %	8 %	10 %	8 %	8 %
Franche-Comté	2 %	3 %	5 %	2 %	2 %
Pays de la Loire	6 %	6 %	5 %	5 %	5 %
Bretagne	4 %	4 %	3 %	5 %	4 %
Poitou-Charentes	4 %	4 %	3 %	3 %	2 %
Aquitaine	4 %	3 %	4 %	4 %	6 %
Midi-Pyrénées	5 %	6 %	4 %	4 %	5 %
Limousin	−1 %	−2 %	2 %	3 %	−3 %
Rhône-Alpes	9 %	9 %	10 %	9 %	10 %
Auvergne	−1 %	0 %	0 %	−1 %	0 %
Languedoc-Roussillon	11 %	12 %	12 %	10 %	11 %
Provence - Côte-d'Azur	15 %	12 %	15 %	13 %	12 %

Des ajustements à la marge pourront être apportés pour améliorer le lissage des évolutions.

3. Plans pour les communes comptant 10 000 habitants ou plus

3.1. Unités échantillonnées et base de sondage

Le plan de sondage dans ces communes est un plan « à l'adresse », toute adresse échantillonnée étant enquêtée de façon exhaustive. Cette contrainte est forte si l'on doit prendre en compte le niveau infracommunal, niveau auquel il faut pouvoir obtenir des estimations détaillées ayant une précision acceptable.

Le sondage utilisera comme base de sondage le « répertoire d'immeubles localisés » (RIL). Ce répertoire est une liste d'édifices (résidentiels, institutionnels ou commerciaux) repérés individuellement de façon à créer une cartographie numérisée. Le RIL a d'abord été alimenté par les résultats du RP99, permettant ainsi de décrire statistiquement chaque immeuble résidentiel.

Le RIL est maintenant mis à jour en continu à partir des permis de construire, des permis de démolir, de renseignements fournis par les administrations locales et l'observation directe sur le terrain.

3.2. Séparer les adresses de grande taille des autres

Le problème majeur est la variance de la taille des unités à échantillonner. En effet, la présence d'une adresse contenant parfois jusqu'à plusieurs dizaines de logements pose un problème d'effet de grappe : les estimations communales et infra-communales pour certaines variables peuvent être très sensibles à la présence ou non de ces adresses dans l'échantillon. Par exemple, dans le cas d'un échantillon aléatoire simple de grappes, on sait que la variance de l'estimateur du total d'une variable Y s'écrit :

$$V(\widehat{T}(Y)) = M^2(1-t)\frac{S_g^2}{m}$$

où M est le nombre total de grappes, m le nombre de grappes de l'échantillon, t le taux de sondage et S_g^2 la variance inter-grappes, c'est-à-dire :

$S_g^2 = \frac{1}{M-1}\sum_i (Y_i - \overline{Y})^2$ où Y_i est le total de la variable dans la grappe i.

On voit sur cette formule que la variance peut être élevée si la taille des grappes est très disparate.

C'est pourquoi il a été décidé de créer une strate particulière constituée de ces adresses. Cette strate sera enquêtée exhaustivement au cours d'un cycle de 5 ans. Il n'y aura donc pas de composante due à l'échantillonnage dans le calcul de la variance au sein de cette strate pour les estimations détaillées. Cette stratégie a pour avantage principal d'améliorer fortement la précision dans les IRIS2000 contenant des adresses de grande taille. À budget global constant, la contrepartie est un taux d'échantillonnage un peu moindre dans la strate « autres adresses ».

Nous aurons ainsi dans chaque commune deux strates principales :

– la strate des « adresses de grande taille » (en termes de nombre de logements) ;

– la strate des « autres adresses ».

Les adresses de grande taille sont déterminées à partir de la distribution de la taille des adresses au niveau communal. Le critère de scission entre les deux strates est le suivant : font partie des « adresses de grande taille » les adresses comptant le plus grand nombre de logements telles que leur ensemble représente environ la proportion p des logements de la commune. Ce seuil a été déterminé à l'examen des simulations faites.

Les adresses de la strate « adresses de grande taille » sont réparties en 5 groupes de rotation. Chacun des groupes est enquêté exhaustivement au cours du cycle.

Les adresses de la strate « autres adresses » sont au départ réparties en 5 groupes de rotation. Lors de l'initialisation de chacune des campagnes de collecte, un des groupes de rotation est échantillonné en prenant en compte le niveau infracommunal dans l'équilibrage, en tenant compte de la démographie des adresses (adresses nouvelles ou disparues) et de la répartition en logements individuels ou collectifs. Il s'agit donc dans ce cas d'un sondage en deux phases.

À l'issue du cycle de cinq ans, toutes les « adresses de grande taille » et une partie des « autres adresses » ont été enquêtées de façon à s'ajuster au taux global de collecte fixé.

Des travaux de simulation ont été effectués dans le but de déterminer le meilleur compromis pour fixer le seuil de définition des adresses de grande taille. Les simulations semblent indiquer que le seuil raisonnable qui peut être proposé est celui qui consiste à déclarer comme adresses de grande taille dans une commune donnée celles qui représentent 10 % du nombre de logements de la commune, avec un plancher de 50 ou 60 logements.

3.3. Initialisation des groupes de rotation

Pour les adresses de grande taille, l'équilibrage se fait uniquement sur le nombre de logements. Les groupes d'adresses sont rendus de taille aussi égale que possible en nombre des logements.

Pour les autres adresses, les variables d'équilibrage pour l'initialisation des groupes de rotation sont les mêmes que celles ayant servi à la constitution des groupes de rotation des communes de moins de 10 000 habitants : tranches d'âge, sexe, nombre de logements, répartition en termes de logements individuels ou collectifs. Le niveau géographique d'équilibrage est la commune.

3.4. Échantillonnage des adresses de grande taille

Dans la strate des adresses de grande taille, l'échantillonnage se fait en une seule phase. La probabilité d'inclusion d'une adresse dans le groupe de rotation G est en moyenne de $1/5$. Étant donné le faible effectif de cette strate, quelques dizaines d'adresses tout au plus, l'équilibre entre groupes de rotation sera parfois approché. Cela peut avoir un impact sur les estimations globales mais sera neutre dans le cadre des résultats du recensement puisque les estimations détaillées s'appuient sur l'agrégation des cinq groupes. Chaque année on enlève les adresses détruites et on répartit les adresses nouvelles en cherchant toujours à égaliser au mieux le nombre de logements par groupe.

3.5. Échantillonnage des autres adresses

Dans cette strate « autres adresses », on doit faire un échantillon de deuxième phase. Cet échantillonnage doit permettre de produire des estimations au niveau infracommunal. C'est pourquoi sont utilisées, lors du tirage de deuxième phase, des variables « nombre de logements » spécifiques à chacun

des IRIS2000. Cette méthode doit permettre d'obtenir des résultats acceptables à des niveaux infracommunaux tout en effectuant l'équilibrage au niveau communal.

Pour équilibrer les échantillons de deuxième phase, on utilise :

– le nombre de logements ;

– une variable par IRIS2000, égale au nombre de logements de l'adresse si elle appartient à l'IRIS2000 et à zéro dans le cas contraire.

Ainsi, au cours du tirage de l'année, on s'assure que l'échantillon est bien réparti sur tout le territoire de la commune.

SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE

COMPTE RENDU DES ÉCHANGES SUR LA PARTIE PLAN DE SONDAGE

Jean-Marie GROSBRAS, Jean-Michel DURR
et Dominique ALLAIN

Question : pour chaque commune, la strate des adresses de grande taille et celle des autres adresses représentent au total 40 % en taux de sondage. Pour les adresses de grande taille, le choix du programme de rénovation a été de les définir sur le critère de leur poids relatif en nombre de logements par commune en prenant en définitive un seuil relatif identique. Or, on peut se demander s'il n'aurait pas été meilleur d'essayer d'optimiser ce seuil commune par commune ?

Réponse : pour déterminer ce seuil, de nombreuses simulations ont été réalisées sur une quarantaine de variables dans des communes très typées. Les travaux étaient réalisés sur de nombreuses communes et par IRIS. Il n'y a pas eu de calcul d'optimum de seuil en tant que tel mais il a été constaté que le seuil de 10 % apparaissait comme un bon compromis la plupart du temps.

Question : la qualité de la base de sondage est essentielle. Trois critères sont à prendre en compte : son exhaustivité, son actualisation régulière et la précision de la localisation des adresses. La qualité de ce dernier critère est difficile à respecter car de nombreuses adresses n'ont pas de numéro, il y a des lieux-dits, etc. Dans ces conditions, pourquoi avoir abandonné la notion d'îlots ?

Réponse : le travail d'échanges et de mise à jour prévu avec les communes doit permettre d'avoir une base la plus exhaustive possible. Cette base sera actualisée régulièrement. Pour les enquêtes de recensement, l'Insee tirera son échantillon dans une base expertisée par les communes quelques mois avant la collecte (environ six mois). Mais, il faut garder à l'esprit que les constructions neuves ne représentent qu'un faible pourcentage du total des adresses, que l'annualisation de la collecte permettra de tenir compte des mouvements récents un an après et que les résultats prennent en compte les cinq années consécutives de collecte. Pour ce qui est de la difficulté de localisation des adresses, on peut à ce stade indiquer qu'elle est généralement plus fréquente dans les communes de moins de 10 000 habitants que dans les autres. Par ailleurs, revenir à une notion d'îlots et non d'adresses amplifierait les problèmes de grappe que l'on cherche à limiter en créant la strate des adresses de grande taille.

Question : y a-t-il une obligation à enquêter les communes de moins de 10 000 habitants tous les cinq ans ? Faut-il gérer des changements de groupes pour les petites communes ?

Réponse : enquêter toutes les communes de moins de 10 000 habitants tous les cinq ans relève d'une obligation légale. Mais il n'est pour autant pas question de « rebattre » toutes les cartes tous les cinq ans. Des règles simples sont prévues pour les changements de groupes et elles seront expliquées. Derrière le terme « changement de groupe », on a trois cas de figure : les franchissements du seuil de 10 000 habitants, les fusions et les scissions de communes.

Question : sans parler de rebattre les cartes tous les cinq ans, on peut se demander comment se comporte dans le temps l'équilibre réalisé à l'origine sur les groupes de rotation ?

Réponse : pour les communes de moins de 10 000 habitants, les répartitions ont été équilibrées à partir des données du recensement de 1999. On a donc pu considérer les groupes selon les données du recensement de 1990 et on a vérifié qu'ils avaient à peu près les mêmes qualités de représentativité neuf ans plus tôt, ce qui laisse à penser que les équilibrages au niveau régional ont une bonne stabilité.

Pour les communes de 10 000 habitants ou plus, on a réalisé des simulations sur plusieurs communes selon la méthode suivante. On a considéré comme base de sondage, dans le fichier du recensement de 1999, l'ensemble des logements construits avant 1990 et, dans cette base, on a constitué des groupes de rotation d'adresses équilibrés pour les variables prévues par le plan de sondage. Puis on a fait évoluer ces groupes en incorporant successivement les logements construits en 1990, 1991, etc., jusqu'à 1998 selon la méthode prévue pour la prise en compte des adresses nouvelles et, en même temps, on a simulé une fonction de « mortalité » réaliste pour les adresses anciennes. On a pu ainsi vérifier si les équilibrages réalisés dans la base « 1990 » se retrouvaient après neuf ans d'évolution, selon les données du recensement de 1999. Le constat est que, dans les cas examinés, les équilibrages sont restés stables, même par IRIS : pour les variables servant de test, si d'un groupe à l'autre les effectifs initiaux étaient presque identiques, ils pouvaient varier de 1 à 2,5 % au bout de neuf ans.

Même si l'exercice est incomplet puisque l'on n'a pas simulé des mouvements de migration de population, la conclusion apparaît que des refontes des distributions peuvent, en régime général, s'opérer à un rythme de 10 ou 15 ans.

Question : la redistribution entre les groupes de rotation devra-t-elle se faire en continu ou à un moment donné, ce qui aurait pour effet de ne plus avoir de statistiques sur cinq ans ?

Réponse : non, la redistribution doit se faire d'un coup mais, comme elle est réalisée par ajustement, on peut continuer à produire des statistiques.

Question : est-il possible de préciser ce que l'on a perdu avec les dernières évolutions sur le plan de sondage ?

Réponse : en passant d'un sondage logement à un sondage à l'adresse, on n'a rien perdu ou du moins la perte est toute théorique puisqu'il était matériellement impossible, et ce en termes d'organisation, de le réaliser. Il est toutefois possible d'estimer cette perte théorique. On a ainsi procédé par simulations pour évaluer le taux de chômage dans certains IRIS de Lyon. Les résultats montrent que la perte n'excède pas 0,5 point en écart-type pour un taux de chômage de 8 à 10 %.

Question : l'unité de base qu'est la commune n'est plus suffisante de nos jours et on parle de plus en plus d'agglomérations et d'aires urbaines. La prise en compte de l'intercommunalité est aussi importante.

Réponse : l'unité de base pour la collecte que représente la commune est mentionnée dans la loi et n'est en rien gênante quant aux résultats à produire sur les regroupements supra-communaux. On a une enquête tous les ans et il est tout à fait possible de publier des résultats annuels sur les regroupements intercommunaux et les agglomérations.

Question : quelles seront les modalités d'enquête pour les personnes vivant en habitation mobile et pour les personnes sans domicile fixe ?

Réponse : les modalités précises sont en cours d'examen. Les personnes résidant habituellement dans des habitations mobiles pourraient être interrogées une fois tous les cinq ans. Pour les personnes sans abri, on pourrait conserver le mode d'interrogation retenu précédemment avec une interrogation des services d'hébergement (et de restauration si l'on peut utiliser les méthodes de l'enquête auprès des personnes sans abri) et une interrogation dans la rue. L'enquête pourrait avoir lieu une fois tous les cinq ans.

Question : l'effort de pédagogie pour présenter le recensement est apprécié par plusieurs participants mais il doit se poursuivre et notamment auprès des élus. Les élus doivent être tout particulièrement sensibilisés à l'importance de la qualité du RIL mais aussi à l'utilisation faite des IRIS 2000.

Réponse : nous souhaitons poursuivre dans la voie de la pédagogie mais ce doit être un effort partagé. Il est important que la pédagogie auprès des élus des communes soit un travail conjoint des services des communes et de l'Insee. C'est également un effort conjoint des communes et de l'Insee qui permettra d'avoir un RIL de qualité. Les IRIS 2000 seront la brique de base de la diffusion, cela a été dit mais peut-être pas suffisamment. Il est proposé de se donner un temps d'appréciation des limites actuelles des IRIS et notamment de disposer de quelques données pour réfléchir à une éventuelle révision, qui serait en tout état de cause à négocier avec la CNIL. La stabilité des découpages est également souhaitable pour assurer commodément des comparaisons dans le temps.

Question : que se passe-t-il quant aux données produites pour les communes de moins de 10 000 habitants ? Les fait-on évoluer tous les ans ? Toutes ces communes ne sont pas recensées en même temps et on a alors des degrés d'ancienneté différents. Comment justifie-t-on le choix réalisé, sachant que la population est une base de péréquation importante ?

Réponse : les populations seront actualisées tous les ans pour toutes les communes. La population étant effectivement utilisée comme base de péréquation, le Conseil d'État l'a préconisé dans son avis pour des questions d'égalité de traitement entre les communes. Le cycle de cinq ans, qui est décalé d'un an par rapport à la période séparant deux élections municipales successives, permet en outre qu'aucune commune de moins de 10 000 habitants ne soit à une distance temporelle constante des élections municipales.

Question : lorsqu'une adresse est détruite puis remplacée, est-ce pour vous la même adresse ou une autre ?

Réponse : c'est considéré comme une nouvelle adresse.

SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE

LES DONNÉES PRODUITES : PRINCIPES SUR LA MISE À DISPOSITION ANNUELLE DE DONNÉES ET SUR LEUR UTILISATION

Jean-Michel DURR *

On trouvera ici une première approche des changements entraînés pour les utilisateurs par la mise à disposition tous les ans de résultats issus du recensement rénové. C'est aussi pour nous l'occasion d'évoquer les simulations et travaux qu'il nous reste à mener sur cet important chantier.

Dès la fin du premier cycle quinquennal des enquêtes de recensement et en régime de croisière, l'Insee publiera chaque année les populations légales de toutes les communes conformément à la loi du 27 février 2002 ; l'Insee publiera également des résultats statistiques détaillés aux niveaux communal et infra-communal.

Lors du dernier séminaire, un des modèles possibles d'estimation combinant collecte et données administratives vous avait été présenté (*cf.* « Le fonctionnement de l'estimation détaillée : théorie et pratique », J. Dumais, Journal de la SFdS, *op. cit.*) et plusieurs participants avaient trouvé la présentation complexe. La démarche qui vous est donc proposée aujourd'hui est progressive et pédagogique. Il s'agit de dérouler le travail à faire sur des données collectées à des dates différentes. Avant de se préoccuper du modèle d'estimation et de données auxiliaires, nous commencerons par regarder ce que disent en elles-mêmes les données collectées. Par la suite, il faudra définir le modèle d'estimation définitif selon la taille des communes avec des critères qui feront intervenir la qualité statistique mais aussi la robustesse et la lisibilité, puis apprécier alors les perfectionnements possibles comme l'utilisation des données administratives pour encadrer les évolutions.

Je présenterai les principes de base qui seront retenus sur les données produites et des résultats obtenus sur un ensemble de plusieurs communes. Dans sa communication, Jean-Marie GROSBRAS présente les données pour les communes selon leur taille (*cf. infra* : les données produites par commune et leur utilisation).

* INSEE, programme de rénovation du recensement de la population, 18 boulevard A. Pinard, 75675 PARIS CEDEX 14
e-mail : jean-michel.durr@insee.fr

1. Les principes de base

Avant de présenter comment les données collectées se comportent, je voudrais insister sur deux éléments qui m'apparaissent majeurs dans la mise à disposition et dans l'utilisation des données.

Quelle que soit la taille de la commune, les résultats du recensement porteront sur l'année médiane des cinq dernières années de collecte, de façon à limiter les actualisations. Pour autant, il faut garder à l'esprit que la fabrication des résultats proviendra d'une démarche adaptée au mode de collecte et donc à la taille des communes. En effet, dans les communes de 10 000 habitants ou plus, de l'information est collectée tous les ans alors que ce n'est le cas qu'une année sur cinq dans les autres communes. Pour les communes de 10 000 habitants ou plus, une des questions importantes se situe donc dans le choix des pondérations selon la date de recueil des observations collectées puisque, tous les ans, on ramène une information nouvelle. Pour les communes de moins de 10 000 habitants, la production de résultats sur l'année médiane nous permettra de nous ancrer sur deux recensements.

Je voudrais aussi préciser à l'ensemble des utilisateurs que, comme avant, il y aura production d'un fichier de données détail permettant toutes les tabulations possibles sur les différentes variables aux différents niveaux géographiques existants.

2. Le supra-communal

Le premier exemple porte sur un canton. Ce canton est composé de communes de moins de 10 000 habitants recensées à des dates différentes sans équilibrage des groupes de rotation.

Le principe de la simulation est d'examiner comment se comportent les données recueillies au cours de 5 ans lorsqu'on les combine simplement, en l'absence de toute modélisation ou d'incorporation d'information auxiliaire. Pour ce faire, la donnée concernant une année est obtenue en additionnant les données recueillies pendant les cinq années encadrant cette année.

Prenons l'exemple du canton d'Asfeld dans les Ardennes; sa population est de 5 173 habitants en 1999, en légère diminution par rapport à 1990, date à laquelle elle était de 5 279 habitants. Il se compose de 18 communes, dont la population s'échelonne de 36 habitants à 976. Commençons par affecter chacune de ces communes aléatoirement à un groupe de rotation. Il ne s'agit pas ici d'assurer un quelconque équilibrage : celui-ci étant assuré au niveau régional, il est bien évident que pour un canton, la répartition est quelconque.

Supposons que l'évolution de la population de chacune des communes, et donc des groupes correspondants, a été régulière entre 1990 et 1999 : on peut construire le tableau suivant représentant les populations « réelles » correspondantes :

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Groupe 1	1 186	1 185	1 183	1 182	1 181	1 179	1 178	1 177	1 175	1 174
Groupe 2	2 357	2 337	2 317	2 297	2 277	2 257	2 237	2 217	2 197	2 177
Groupe 3	693	699	705	711	717	722	728	734	740	746
Groupe 4	299	300	301	303	304	305	306	308	309	310
Groupe 5	744	746	749	751	754	756	759	761	764	766
CANTON	5 279	5 267	5 255	5 244	5 232	5 220	5 208	5 197	5 185	5 173

Au fil des années, avec le nouveau recensement, on collecte l'information suivante :

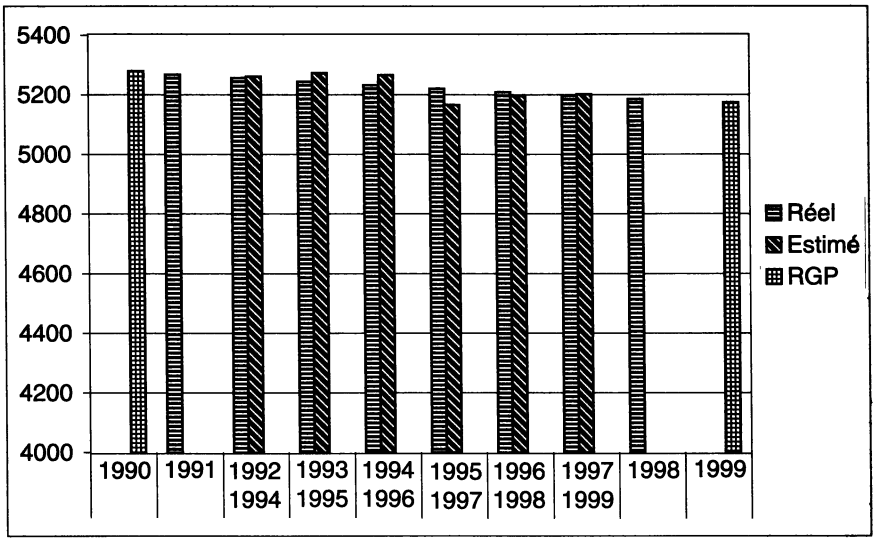
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Groupe 1		1 185					1 178			
Groupe 2			2 317					2 217		
Groupe 3				711					740	
Groupe 4					304					310
Groupe 5	744					756				

En faisant chaque année la somme des 5 années l'encadrant, ou si l'on préfère chaque année de collecte la somme des 5 dernières années, ramenée à l'année $N-2$, milieu de cette période de 5 ans, on obtient les résultats suivants :

	1990	1991	1992	1993	1994	1995	1996	1997
ESTIMÉ			5 261	5 273	5 266	5 166	5 195	5 201
REEL	5 279	5 267	5 255	5 244	5 232	5 220	5 208	5 197
Différence estimé-réel			4	29	34	–54	–13	4
En %			0,1 %	0,5 %	0,6 %	–1,0 %	–0,3 %	0,1 %

Le graphique ci-après présente les résultats obtenus par les recensements généraux (RGP) de 1990 et 1999 et l'estimation par somme mobile à partir des données collectées annuellement, comparés à la « réalité » telle que supposée. Sur l'axe des abscisses, l'année de production est indiquée sous l'année de référence des résultats.

Ainsi, il y a une bonne estimation de la population au fil des ans et une absence de dérive, même si la répartition annuelle des groupes n'est pas favorable. Par exemple, les groupes 1 et 2 étant de grande taille par rapport aux autres, les années au cours desquelles on les recense seront donc surpondérées. Comme nous avons une baisse régulière de la population sur la période, les estimations produites sur les années pour lesquelles les groupes 1 et 2 sont avant sont



surestimées, car le passé est surpondéré, alors que les estimations produites sur les années ou les groupes 1 et 2 sont après sont sous-estimées, car cette fois c'est le futur qui est surpondéré. Il y donc rattrapage systématique.

En introduisant un choc...

Simulons à présent un choc sur la tendance sous la forme d'une baisse sensible (−7 %) de la population dans toutes les communes au cours de l'année 1995, en raison par exemple de la fermeture d'une grosse entreprise du canton entraînant un départ massif d'actifs. Supposons une légère reprise les années suivantes. Soit le tableau des données réelles :

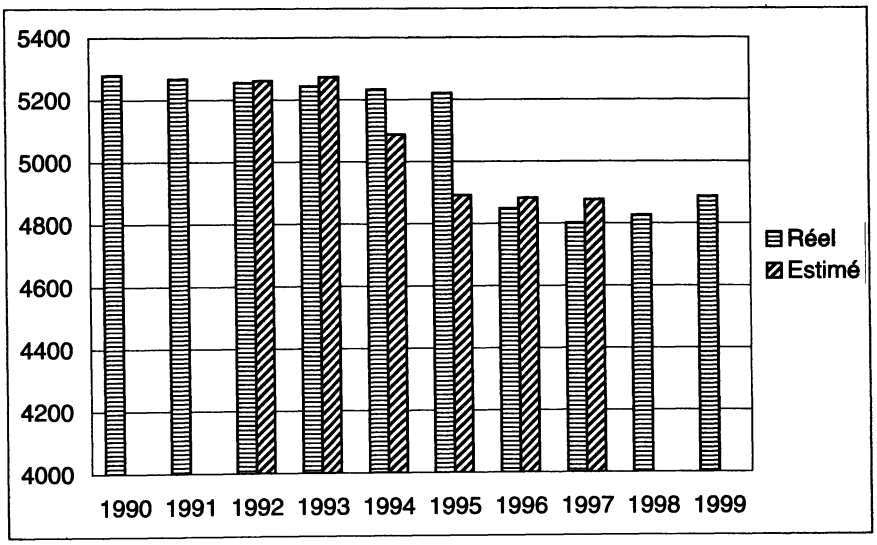
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Groupe 1	1 186	1 185	1 183	1 182	1 181	1 179	1 000	998	1 002	1 015
Groupe 2	2 357	2 337	2 317	2 297	2 277	2 257	2 150	2 120	2 125	2 135
Groupe 3	693	699	705	711	717	722	695	689	702	714
Groupe 4	299	300	301	303	304	305	280	276	281	298
Groupe 5	744	746	749	751	754	756	723	718	716	725
CANTON	5 279	5 267	5 255	5 244	5 232	5 220	4 848	4 801	4 826	4 887
							−7,1 %			

La nouvelle collecte annuelle est donc :

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Groupe 1		1 185					1 000			
Groupe 2			2 317					2 120		
Groupe 3				711					702	
Groupe 4					304					298
Groupe 5	744					756				

On obtient donc les résultats suivants en appliquant les sommes mobiles :

	1990	1991	1992	1993	1994	1995	1996	1997
Estimé			5 261	5 273	5 088	4 891	4 882	4 876
Réel	5 279	5 267	5 255	5 244	5 232	5 220	4 848	4 801
E-R			4	29	-144	-329	34	75
%			0,1 %	0,5 %	-2,8 %	-6,3 %	0,7 %	1,6 %



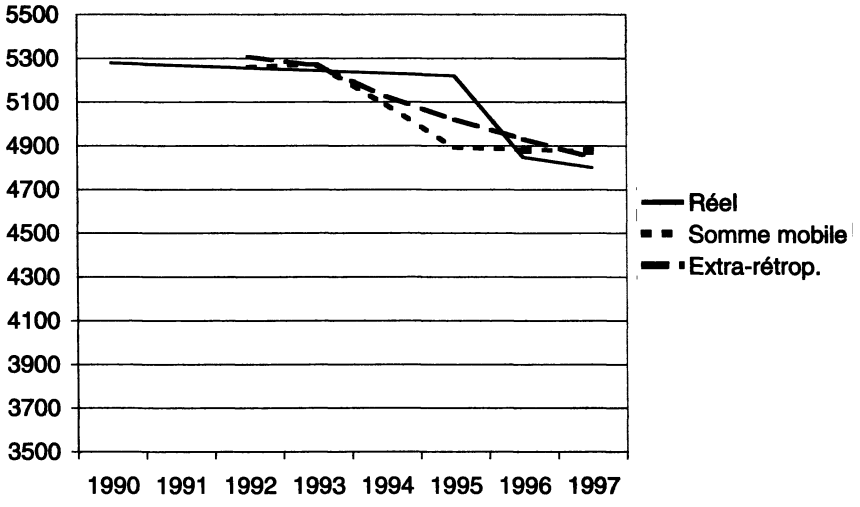
Il se produit une anticipation de la baisse en 1994, en raison du positionnement des groupes plus nombreux après la date de la rupture; ils sur-pondèrent donc la baisse dans les estimations de 1994. L'effet serait inverse s'ils étaient enquêtés juste avant la rupture. On constate également l'absence de dérive : le mouvement d'ensemble est correctement estimé, seul le timing est approximatif. Il s'agit cependant d'effets bruts : la modélisation, par apport

d'information auxiliaire comme des données administratives, peut améliorer la prise en compte de ce type de rupture.

Ces premières simulations très simples permettent déjà de vérifier que le principe de la collecte annuelle n'introduit pas de dérive. Même si une évolution brusque n'est pas détectée immédiatement, elle l'est dans un délai de deux ans au maximum et les erreurs d'estimation sous-jacentes sont alors compensées.

Ce premier point étant établi, il est possible d'améliorer les estimations en prenant en compte la dynamique propre aux données. En se rappelant que les résultats détaillés sont produits sur l'année médiane d'un cycle de 5 ans, on peut utiliser un modèle d'estimation par extrapolation-rétropolation tel que décrit dans le papier de Jean-Marie Grosbras. Compte tenu du décalage de deux ans entre année de référence et année de production, il s'agit de prolonger les tendances observées lors des recensements précédents pour les deux années qui suivent un recensement, et d'interpoler pour les années comprises entre deux recensements. Les résultats sont alors les suivants :

	1992	**1993**	**1994**	**1995**	**1996**	**1997**
Estimé	5 307	5 267	5 125	5 020	4 930	4 848
Réel	5 255	5 244	5 232	5 220	4 848	4 801
E-R	52	23	−107	−200	82	47
%	0,99 %	0,44 %	−2,05 %	−3,83 %	1,69 %	0,98 %



Cette méthode évite les écueils induits par la sommation brute des données. Elle constitue une première étape dans le modèle d'estimation, la suivante étant l'intégration d'information auxiliaire issue de sources administratives. En effet, si l'interpolation entre deux valeurs observées apporte une bonne précision pour les groupes recensés les deux dernières années, l'extrapolation menée à partir des recensements précédents pour actualiser les groupes recensés les deux années précédant l'année de référence est plus fragile, notamment en cas de rupture. Il peut donc être intéressant de corriger l'extrapolation par l'évolution constatée dans une source administrative, à condition qu'elle soit fiable et absente d'effets de gestion. Pour une année N, au cours de laquelle on produira les résultats portant sur l'année $N-2$, il est alors nécessaire de disposer des données administratives des années $N-4$, $N-3$ et $N-2$. Ces développements seront présentés dans une communication future.

SÉMINAIRE MÉTHODOLOGIQUE SFdS INSEE

LES DONNÉES PRODUITES PAR COMMUNE ET LEUR UTILISATION

Jean-Marie GROSBRAS *

1. Introduction : la publication des résultats pour les communes

Le principe est que les données publiées l'année A sont issues d'estimation à valeur pour l'année $A-2$. Par exemple, pour une commune de 10 000 habitants ou plus, les données publiées en A incorporent les résultats de cinq collectes annuelles successives agrégées pour produire des estimations à l'année médiane de la période de cinq ans.

Les méthodes mises en œuvre diffèrent selon les strates auxquelles appartiennent les communes, c'est-à-dire selon qu'elles ont moins de 10 000 habitants ou plus. Dans le premier cas, les communes sont recensées exhaustivement tous les cinq ans et le travail consiste à estimer des données entre deux recensements ; dans le second, il s'agit en particulier d'utiliser cinq enquêtes annuelles successives, ce qui n'est pas le même problème.

On abordera successivement trois points :

– ce qui peut être envisagé pour les communes de moins de 10 000 habitants ;

– une méthode pour les communes de 10 000 habitants ou plus, avec une première idée de la précision des résultats ;

– l'utilisation des données annuelles en tant que séries temporelles pour observer des tendances et évolutions.

Ce dernier point est ébauché et fera ultérieurement l'objet de plus amples développements.

2. Les communes de moins de 10 000 habitants

2.1 La méthode de base

L'idée la plus simple est d'utiliser, en tant que de besoin, la tendance observée pour une commune aux recensements les plus proches la concernant. Ainsi, la population A-2 publiée l'année A sera établie selon le groupe auquel appartient

* INSEE, programme de rénovation du recensement de la population, 18 boulevard A. Pinard, 75675 PARIS CEDEX 14
e-mail : jean-marie.grosbras@insee.fr

la commune, c'est-à-dire selon qu'elle a été recensée en A-4, A-3, A-2, A-1 ou A. La règle est alors la suivante :

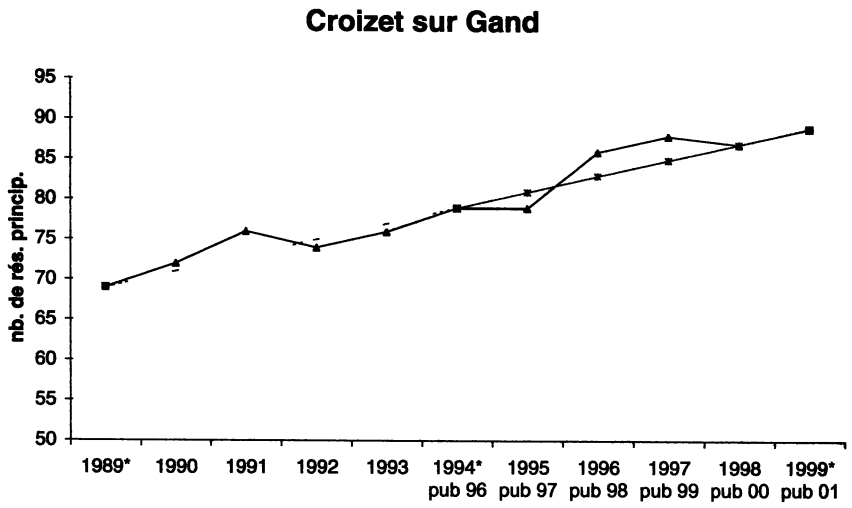
Recensement	Action
A-4 (donc aussi A-9)	On extrapole à A-2 la droite (A-9 → A-4)
A-3 (donc aussi A-8)	On extrapole à A-2 la droite (A-8 → A-3)
A-2	On garde le recensement
A-1 (donc aussi A-6)	On interpole entre A-6 et A-1
A (donc aussi A-5)	On interpole entre A-5 et A

On voit ainsi que l'« horizon » des extra-interpolations est au maximum de deux ans. Pour le démarrage des estimations, c'est-à-dire en 2008 pour des estimations en 2006, le point de départ des extra-interpolations sera le recensement de 1999.

Il reste à constituer le fichier « détail » destiné aux exploitations statistiques. Prenons, par exemple, le cas d'une commune recensée en A, avec une population de 105, alors qu'elle avait une population de 100 en A-5. Par interpolation, la population en A-2 est donc estimée à 103. Le fichier détail de A-2 sera celui du plus récent recensement, c'est-à-dire celui de A, dont toutes les unités seront pondérées par le coefficient $103/105 = 0,98$. Les pondérations servent essentiellement aux études portant sur des ensembles de plusieurs communes, appartenant à des groupes de rotation différents.

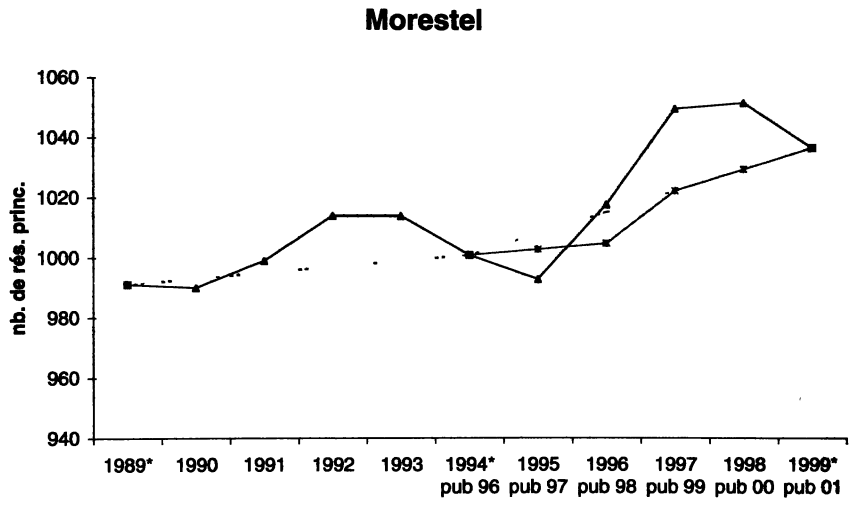
Pour illustrer la méthode et analyser toutes les questions qui se posent, il faudrait disposer de séries annuelles de population exactes. En l'absence d'une telle source, on a eu recours aux fichiers de la taxe d'habitation, en considérant que le nombre de logements qu'ils contiennent est en étroite corrélation avec l'effectif de la population. Disposant des données annuelles de 1989 à 1999, on va procéder aux calculs d'extra-interpolation comme si les recensements se situaient en 1989, 1994 et 1999 et comparer aux « vraies » valeurs.

2.2. Exemple d'une évolution régulière



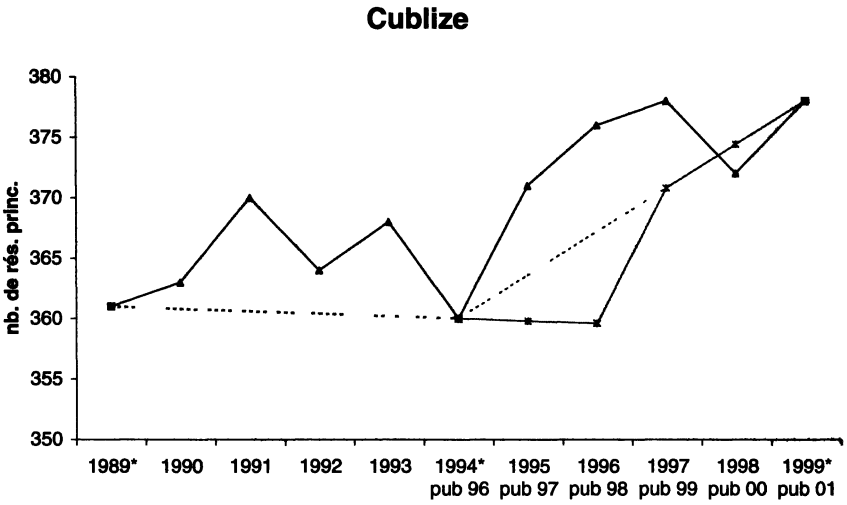
Les triangles indiquent les « vraies » valeurs, les carrés les valeurs observées aux recensements, les étoiles les valeurs estimées. Ainsi, en 1996, on publiera le résultat de 1994, c'est-à-dire le résultat du recensement réalisé cette année-là, puis on publiera en 1997 et 1998 les estimations pour 1995 et 1996, établies en prolongeant la droite reliant 1989 à 1994; enfin, en 1999 et 2000 on publiera les estimations valant pour 1997 et 1998 établies en interpolant les « recensements » de 1994 et 1999. Comme, ici, il s'agit d'une commune à évolution régulière, les estimations sont très proches de la réalité.

2.3. Exemple d'évolution moins régulière



L'évolution entre 1994 et 1999 n'est pas régulière et l'interpolation linéaire sous-estime un peu le pic des années 1997 et 1998.

2.4. Exemple de changement de tendance



Dans ce cas, on joue de malchance : un gros lotissement s'est créé juste après le recensement de 1994 et les estimations relatives aux années 1995 et 1996 l'ignorent. C'est à partir de l'estimation de l'année suivante (qui peut bénéficier de l'information plus fraîche récoltée en 1999) que la bonne tendance est récupérée.

2.5. Conclusion

Dans la grande majorité des cas, les procédures d'estimation donneront des résultats satisfaisants. Les cas les plus délicats sont ceux où des événements importants se produisent peu après les recensements. En ce cas les estimations de base mettent deux ans à les incorporer, ce qui est néanmoins nettement plus satisfaisant que le système de recensements traditionnels où le délai de prise en compte est beaucoup plus long. Les communes avaient toujours le moyen de recourir à la procédure des recensements complémentaires mais cette procédure a elle-même un délai non négligeable.

Autre élément à prendre en compte pour ce cas de figure, ce sont les outils dont on peut disposer pour améliorer les estimations. En effet, on peut dans ces cas recourir aux informations complémentaires issues des fichiers administratifs, fiscaux notamment. C'est ainsi que la taxe d'habitation, prise cette fois comme source administrative, peut servir à alerter sur les ruptures de série. Pour reprendre l'exemple de Cublize, en 1997, des données de la taxe d'habitation de 1995 et des années antérieures sont susceptibles de montrer ce qui s'est passé entre 1994 et 1995 et, en ce cas, l'estimation de base peut être améliorée pour tenir compte de cette information plus fraîche.

3. Les communes de 10 000 habitants ou plus

3.1. L'approche moyenne mobile

Il s'agit ici de consolider cinq enquêtes de recensement successives, de A-4 à A pour produire des résultats millésimés à l'année médiane A-2. La difficulté vient de ce que, comme on l'a vu dans l'exposé sur les plans de sondage, l'enquête d'une année donnée s'exécute avec une base de sondage actualisée par rapport à l'année précédente, c'est-à-dire intégrant la démographie des logements. Une méthode « simple » pour traiter cette question est de construire, pour A-2, la moyenne des cinq estimations annuelles de la période. À la fin de l'année A+1, on calculera la moyenne, millésimée A-1, des enquêtes des années A-3 à A+1, et ainsi de suite.

Pour décrire les opérations, on va supposer, dans un premier temps, que chaque année la strate des adresses de grande taille fait exactement 10 % des logements de la commune, qu'elle peut être partagée en cinq groupes de rotation comprenant exactement le même nombre de logements et que la strate des autres adresses comprend donc exactement 90 % des logements partageables en cinq groupes égaux.

En ce cas, pour une année donnée, l'extrapolation à l'ensemble de la strate des adresses de grande taille est simple : il suffit d'affecter un coefficient de 5 aux observations effectuées dans le groupe de l'année. De même, les données recueillies par sondage dans le groupe de rotation des autres adresses sont affectées du coefficient d'extrapolation de 15.

En moyennant cinq années consécutives, on établit que les données issues de la strate des adresses de grande taille sont affectées du coefficient 1 (la strate est exhaustivement enquêtée), et les données issues des groupes de rotation des autres adresses sont affectées du coefficient 3. Ainsi le total d'une variable Y se calcule par :

$$\widehat{T}(Y) = \sum_i c_i Y_i$$

où $c_i = 1$ pour les données recueillies dans la strate des grandes adresses et $c_i = 3$ pour les données recueillies dans la strate des autres adresses.

Dans la réalité, les coefficients ne seront pas exactement ceux indiqués ci-dessus pour deux raisons principales :

– ils dépendront tout d'abord du poids respectif des deux strates, la strate des adresses de grande taille ne représente pas, en général, exactement 10 % des logements, d'autant que l'on a mis un plancher en termes de nombre de logements minimum pour ces adresses. Si, par exemple, la strate représente 5 % des logements de la commune, le coefficient des données observées dans la strate des autres adresses sera égal à 2,714 [2]) ;

2 Sur 100 adresses, les adresses de grande taille représentent 5 ; pour avoir un échantillon total de 40, il faut donc tirer 35 autres adresses dans le stock de 95 ; le ratio 95/35 vaut bien 2,714.

– ils dépendront aussi des calages éventuellement pratiqués. Par exemple, l'application de la formule précédente à la variable nombre de logements n'aboutira pas exactement au nombre de logements présents au moment de l'enquête A-2 (à cause de la démographie des logements). On peut donc souhaiter que l'estimation A-2 reconstitue le nombre de logements A-2 (grandeur connue). Les coefficients sont donc ajustés, par règle de trois, pour retrouver le bon total. Pour avoir des statistiques encore plus fiables à l'infra communal, l'ajustement pourrait se faire pour chaque IRIS 2000.

3.2. Estimations annuelles produites : étude par simulation

Pour avoir une approche de la précision des résultats des sondages on a procédé aux simulations suivantes sur quelques communes (Grenoble, Romans, arrondissements Lyon 6 et Lyon 8). Pour chacune d'elle on dispose des fichiers du recensement de 1999. On a donc constitué les strates d'adresses puis les groupes de rotation, équilibrés selon la méthode prévue, puis on tire cinq échantillons successifs, un par groupe de rotation d'adresses et on applique les méthodes d'estimation.

L'opération est répétée 500 fois et on analyse, pour un certain nombre de variables témoin, la répartition des 500 résultats par rapport aux vraies valeurs. Les indicateurs de dispersion sont le coefficient de variation pour les statistiques en niveau (nombre de ...) et l'écart-type pour les statistiques de structure (proportion de ...). Par exemple, pour un total Y, on calcule les 500 estimations $\widehat{Y}_j$, $j = 1, \ldots, 500$, puis l'écart quadratique moyen et le coefficient de variation :

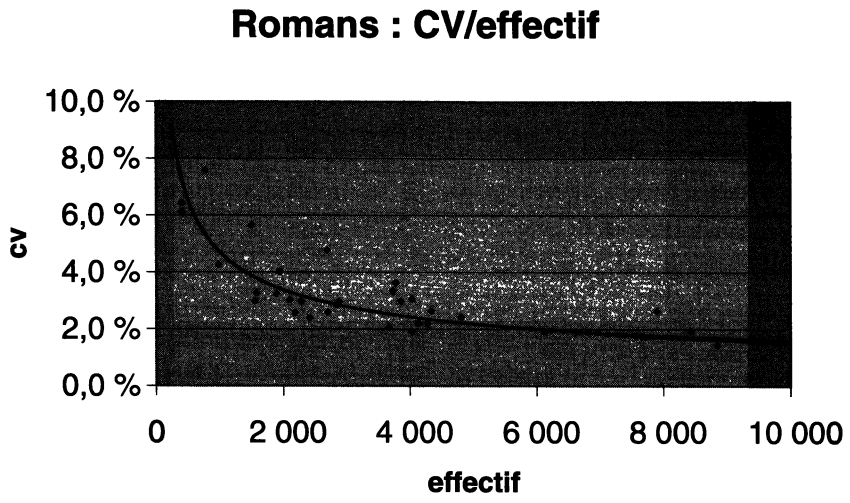
$$V(\widehat{Y}) = \frac{\sum_{j=1}^{500} (\widehat{Y}_j - Y)^2}{500} \text{ et } CV(\widehat{Y}) = \frac{\sqrt{V(\widehat{Y})}}{Y}.$$

Le tableau et le graphique suivants montrent les coefficients de variation issus des simulations sur la ville de Romans. On y trouve la conclusion attendue que la précision varie avec la taille de la « cible » : plus l'effectif est élevé, plus gros est l'échantillon et donc meilleure est la précision du sondage.

Tableau de la ville de ROMANS

	Effectif	**Coefficient de variation**
Personnes de moins de 20 ans	7 894	2,6 %
Personnes de 20 à 39 ans	8 428	2,0 %
Personnes de 40 à 59 ans	7 706	1,8 %
Personnes de 60 à 74 ans	4 811	2,2 %
Personnes de 75 ans ou plus	2 889	3,0 %
Hommes de moins de 20 ans	4 037	3,2 %
Hommes de 60 à 74 ans	2 106	2,7 %
Hommes de 75 ans et plus	989	4,2 %
Femmes de moins de 20 ans	3 857	3,0 %
Femmes de 60 à 74 ans	2 705	2,5 %
Femmes de 75 ans et plus	1 900	3,2 %
Français	29 031	1,0 %
Etrangers	2 697	5,0 %
Personnes ayant un emploi	10 725	1,5 %
Chômeurs	2 855	2,9 %
Inactifs	18 127	1,4 %
Ouvriers qualifiés	1 582	3,4 %
Ingénieurs, cadres	410	6,4 %
Personnes résidant ds autre dép.	3 734	3,6 %
Personnes mariées	12 177	1,5 %
Personnes divorcées	2 292	2,9 %
Logts occas. ou rés. secondaires	255	8,9 %
Logements vacants	1 501	4,2 %
Logts dt statut occ.= propriétaire	6 128	1,8 %
Logts dt statut occ.= locataire	7 654	1,5 %
Logts d'une pièce	772	7,4 %
Logts de quatre pièces ou plus	8 847	1,4 %
Adresses dt date ach.<1949	2 201	2,8 %

Graphique des coefficients de variation en fonction de l'effectif de base :



Remarque : ces résultats en matière de précision sont directement issus des simulations *avant calage*; les premières indications montrent que le calage sur le nombre de logements de l'année médiane fait généralement gagner de 0,5 à 1 point en coefficient de variation.

Pour illustrer les résultats concernant les statistiques de structure, le tableau suivant donne, par exemple, les écarts-types du taux des moins de 20 ans dans la population des IRIS2000 du sixième arrondissement de Lyon. On a également fait figurer l'écart-type que donnerait un échantillon aléatoire simple de 40 % des individus des IRIS de façon à illustrer la perte de précision due à l'effet de grappe résultant du sondage à l'adresse.

Tableau IRIS2000 de Lyon 6e arrondissement

	Part des personnes de moins de 20 ans		
IRIS2000	**vraie valeur**	**écart-type**	**écart-type sas[3]**
0103	20,0 %	1,6 %	1,1 %
0104	22,8 %	1,5 %	0,9 %
0201	22,3 %	1,8 %	1,1 %
0202	20,6 %	1,1 %	0,9 %
0301	23,5 %	1,6 %	1,0 %
0302	22,3 %	1,4 %	1,0 %
0303	18,3 %	1,6 %	0,9 %
0304	18,8 %	1,3 %	0,9 %
0401	20,8 %	1,5 %	1,0 %
0402	21,0 %	2,1 %	1,1 %
0403	20,0 %	1,8 %	1,1 %
0501	18,9 %	1,6 %	1,0 %
0502	20,0 %	1,5 %	1,1 %
0503	19,2 %	1,4 %	0,9 %
0601	20,0 %	1,5 %	1,2 %
0602	15,9 %	1,3 %	0,9 %
0603	15,3 %	1,6 %	1,0 %
0701	20,0 %	2,3 %	1,1 %
0702	20,5 %	1,8 %	0,9 %

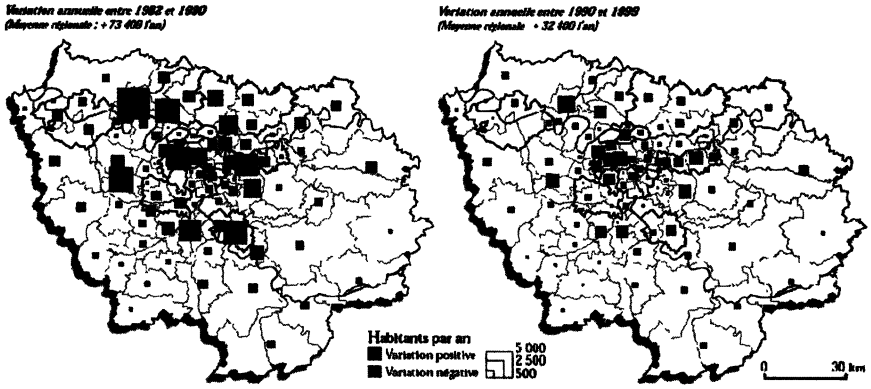
[3] SAS : échantillon aléatoire simple.

4. Les séries annuelles

4.1. L'intérêt des séries annuelles

L'exemple suivant est extrait d'une étude réalisée par l'Institut d'aménagement et d'urbanisme de la région d'Ile-de-France (IAURIF). Elle présente les évolutions de population entre les recensements de 1982, 1990 et 1999 dans le zonage du schéma directeur de la région. Les cartes montrent des différences sensibles dans les évolutions. On aimerait bien disposer d'éléments pour analyser les mouvements entre les deux photographies prises à 9 ans d'intervalle. Y a-t-il eu des points d'inflexion ? Quand ? Les mouvements ont-ils été les mêmes sur tous les secteurs ? Etc. Les données produites par le nouveau recensement pourront apporter des éléments chiffrés à ces questions : plutôt que des photographies espacées dans le temps, le nouveau recensement fournira un diaporama annuel, même si le « grain » de chaque diapositive est moins fin

Évolution de la population par secteur du SDRIF



que celui d'une photographie exhaustive mais susceptible d'être rapidement obsolète.

4.2. Travaux de simulation entrepris

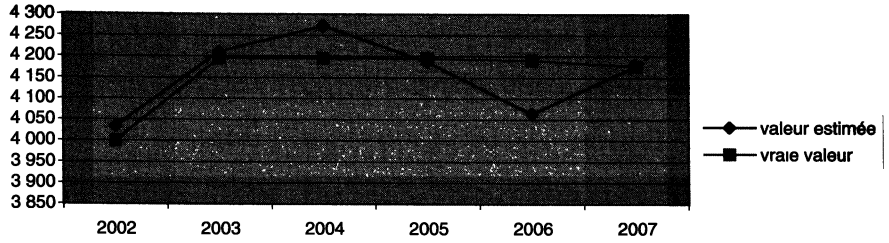
Dans un premier temps, nous sélectionnons des communes avec des quartiers assez typés, d'après les données du recensement de 1999 et imaginons des scénarios d'évolution sur plusieurs années : scénario d'évolution régulière avec des taux démographiques moyens (natalité, etc.), puis introduction de chocs tels que la fermeture d'une usine entraînant une élévation du nombre de chômeurs dans le quartier de l'usine et les quartiers environnants, rénovation urbaine avec destruction d'immeubles et relogement des habitants dans le reste de la commune (relogement non uniforme), etc.

On définit dans ces communes des plans de sondage annuels de sorte que le taux de sondage en cinq ans soit de 40 % et on procède aux estimations par moyenne mobile pour certaines variables (nombre de chômeurs, ...), estimations à la commune et à l'IRIS. Le but de ces simulations est d'examiner comment les séries estimées reflètent la réalité que l'on a introduite dans les scénarios. Ce travail n'en est qu'à ses débuts et des résultats détaillés seront présentés ultérieurement. Voici des exemples de séries produites.

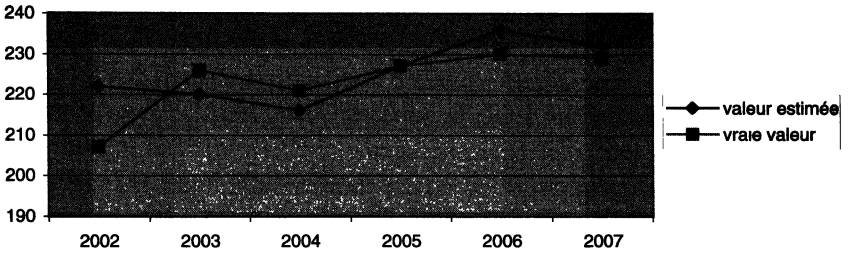
Il s'agit d'une commune de 43 000 habitants à laquelle on a fait subir une fermeture d'usine et une destruction d'immeuble (avec relogement dans le quartier et les quartiers alentours).

Au cours des prochains mois, des simulations de cette sorte seront réalisées en concertation avec un groupe de travail comprenant des utilisateurs des données du recensement, notamment en région. Elles permettront de se familiariser avec le nouveau système, avec son flux annuel de chiffres produits, sur les principaux thèmes d'étude : caractéristiques démographiques et sociales, mobilités résidentielles, trajets domicile-travail, etc. Les résultats feront l'objet de prochaines communications.

Nombre de chômeurs (fermeture d'usine)



nombre de chômeurs iris160150705 (destruction d'immeuble)



SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE

LES DEMANDES DE PRÉCISION OU D'INVESTISSEMENTS COMPLÉMENTAIRES SUR LES DONNÉES PRODUITES

Jean-Marie GROSBRAS, Jean-Michel DURR
et Dominique ALLAIN

Question : Il serait intéressant de travailler rapidement sur le taux de renouvellement de l'échantillon à cinq ans. Y a-t-il un intérêt ou non à avoir un taux de renouvellement faible ?

Réponse : Cela fait effectivement partie des investissements à réaliser.

Question : les résultats sont probants sur des populations importantes mais on peut être inquiet pour les travaux infra-communaux. La précision commence à se dégrader autour de 2 000 observations ; or, dans des villes de 40 000 à 50 000 habitants, on travaille au-dessous de 2 000 habitants.

Réponse : Les simulations actuelles sur la précision instantanée nous montrent que le décrochage ou le coude de la courbe, si l'on préfère, se situe plutôt autour de 1 000 observations que de 2 000. Par ailleurs, cet élément est à mettre en regard d'un recensement traditionnel, qui n'apportait de nouvelles informations que tous les sept à neuf ans.

Question : Pour les personnes qui travaillent sur le spatial, je comprends qu'on perd une continuité de l'espace pour gagner une continuité dans le temps. Qu'est-ce que cette rupture signifie exactement lorsqu'on travaille sur des aires géographiques composites avec des zones où la fiabilité est différente ? Pour les données de migration, la même remarque peut être faite. En outre, peut-on réellement se contenter d'être à 5 % près pour des flux de commune à commune ? Le souhait d'une présentation spécifique pour les données de flux est émis.

Réponse : Il y a continuité dans le temps mais aussi dans l'espace avec le nouveau recensement. Les données sont produites chaque année à tous les niveaux géographiques. Brigitte BACCAÏNI a montré au précédent séminaire (2001) que les questions de migration pouvaient se traiter comme les questions relatives aux autres variables.

Question : La réflexion sur les pondérations reste à creuser. Les exemples donnés portent sur les effectifs des communes. Mais prend-on la même pondération pour la population active au lieu de travail que pour la population

résidente ? Et que fait-on si on a des changements en termes de structure ? Et cette question des pondérations devient particulièrement délicate car il nous faut combiner ces pondérations aux différents poids des sondages liés aux interpolations.

Réponse : La question des pondérations reste effectivement à approfondir. Ce que l'on peut dire aujourd'hui est que l'on a une bonne connaissance du parc des logements l'année de l'observation ; on a donc tout intérêt à faire un calage au milieu de la période sur l'effectif de la commune. Concernant les déformations de structure, les interpolations se font sur les effectifs des variables ; or, on connaît l'inertie des déformations sur les structures et l'on n'a que deux ans de dérive possible.

SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE

CONCLUSION

Alain GODINOT *

Je me félicite de la tenue de ces séminaires, qui nous permettent d'entendre les préoccupations des utilisateurs, et j'en remercie vivement la SFdS.

En tant que responsable du programme de rénovation, je sais que 2004 c'est demain, et que notre principale priorité est d'avoir une bonne première collecte.

Mais j'ai entendu aujourd'hui la préoccupation des utilisateurs de données. En termes d'utilisation des données, nos habitudes mentales devront évoluer compte tenu de ce que le nouveau recensement produira chaque année en régime de croisière. Pour cela, il reste de nombreux investissements à faire à l'Insee et celui-ci doit penser à entretenir une démarche pédagogique autour d'eux. À l'Insee, les chefs des services d'études et de diffusion y travaillent avec les responsables du programme de rénovation du recensement. Pour les utilisateurs externes, un groupe de travail du CNIS sur les données produites sera lancé fin 2002. Cela nous permettra ainsi d'avancer avec eux.

* INSEE, programme de rénovation du recensement de la population, 18 boulevard A. Pinard, 75675 PARIS CEDEX 14
e-mail : alain.godinot@insee.fr

SÉMINAIRE MÉTHODOLOGIQUE SFdS-INSEE

BIBLIOGRAPHIE

ALEXANDER C. (2002), « Les échantillons successifs de Leslie Kish et l'American Community Survey », *Techniques d'enquête*, Vol. 28, n° 1, pp. 39-46, Statistique Canada, Ottawa.

DURR J.M., DUMAIS J. (2002), « La rénovation du recensement français », *Techniques d'enquête*, Vol. 28, n° 1, pp. 47-54, Statistique Canada, Ottawa.

BERTRAND P., CHAUVET G., CHRISTIAN B., GROSBRAS J.M. (2002), « Les plans de sondage du nouveau recensement », *Journées de méthodologie statistique*, décembre 2002, INSEE, Paris.

BERTRAND P., CHAUVET G., CHRISTIAN B., GROSBRAS J.M. (2002), « Données produites par le recensement rénové de la population », *Journées de méthodologie statistique*, décembre 2002, INSEE, Paris.

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

DUMAIS J., ISNARD M. (2000), « Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population », *Actes des VIIe Journées de Méthodologie Statistique*, Paris, 4 et 5 décembre 2000, Tome 1, pp. 37-50, INSEE.

DUMAIS J., BERTRAND Ph., KAUFFMANN B. (2000), « Sondage, estimation et précision dans la rénovation de recensement de la population », *Actes des VIIe Journées de Méthodologie Statistique*, Tome 1, pp. 51-75, INSEE.

DUMAIS J. (2001), « Quelques aspects méthodologiques du recensement rénové de la population en France », in *Enquêtes, modèles et applications*, pp. 467-479.

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. (1981), « Population Counts from Cumulated Samples », Congressional Research Service, *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, 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.

(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.

(1999) Journal de la Société Française de Statistique, vol. 140, n° 4.

(2001) Journal de la Société Française de Statistique, vol. 142, n° 3 (Actes de la séance du 24 octobre 2001 du séminaire méthodologique SFdS - INSEE sur la rénovation du recensement de la population).