Les données produites : principes sur la mise à disposition annuelle de données et sur leur utilisation

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

La présentation d'aujourd'hui est 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 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 présentera les données pour les communes selon leur taille.

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

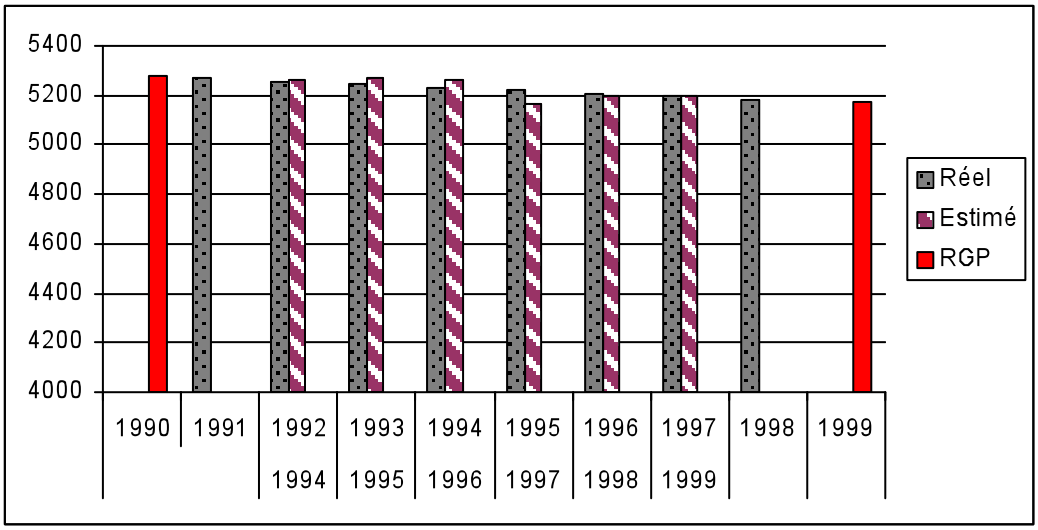
Au fil des années, avec le nouveau recensement, on collecte l'information suivante :

	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**
ESTIME			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-dessous 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 :

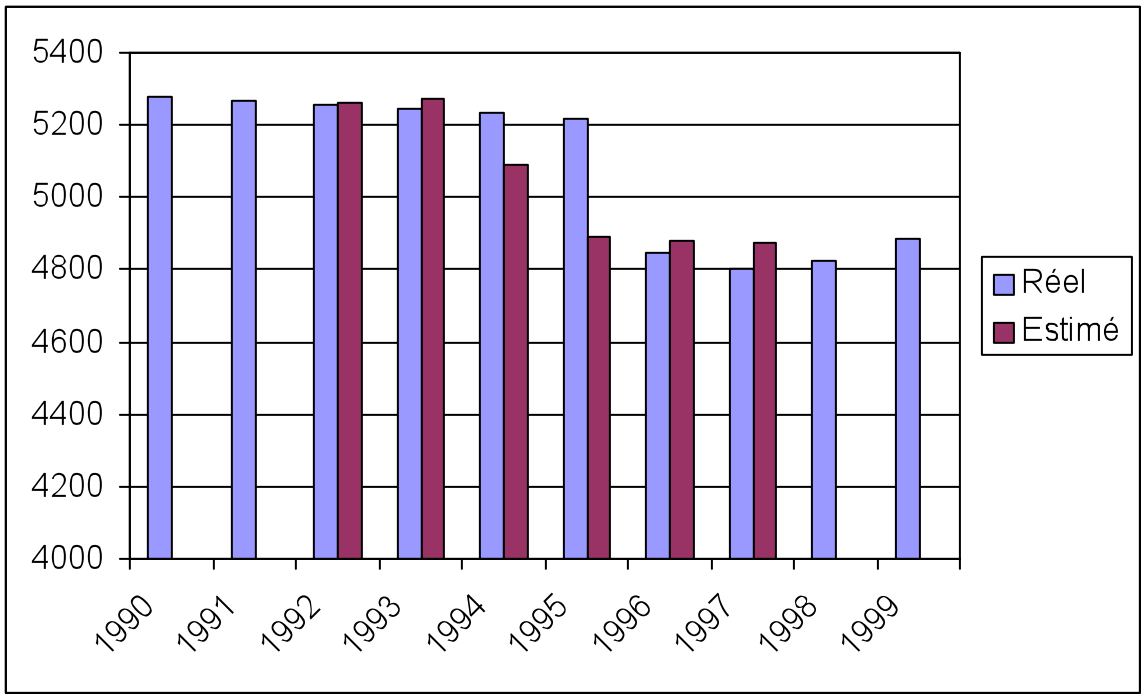
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Groupe 1	1 186	1 185	1 183	1 182	1 181	1 179	1 000	998	1 002	1 015
Groupe 2	2 357	2 337	2 317	2 297	2 277	2 257	2 150	2 120	2 125	2 135
Groupe 3	693	699	705	711	717	722	695	689	702	714
Groupe 4	299	300	301	303	304	305	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%

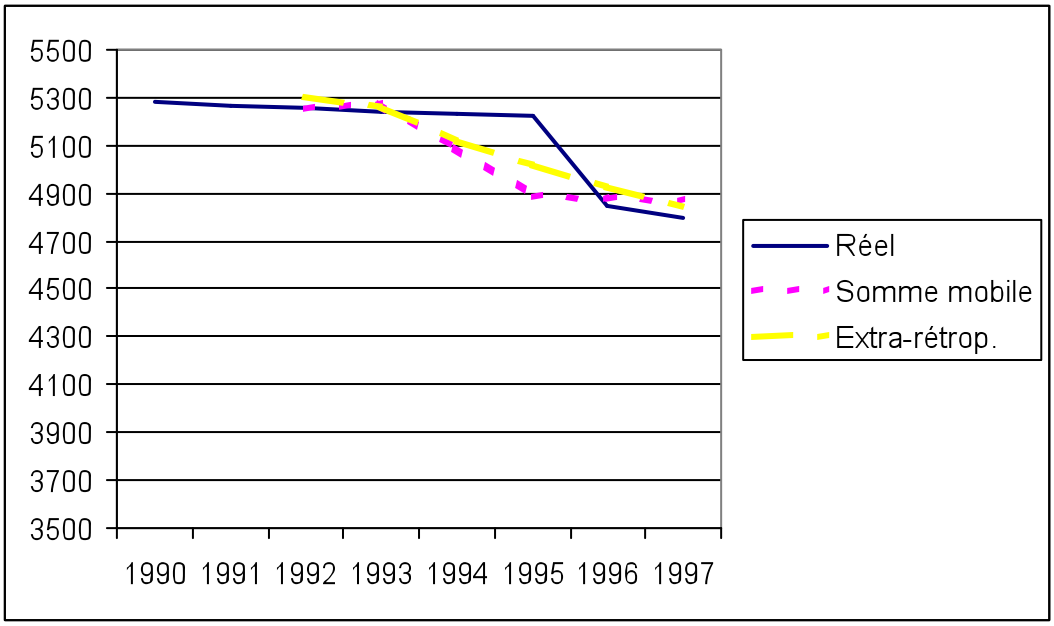


Il se produit une anticipation de la baisse en 1994, en raison du positionnement des groupes plus nombreux après la date de la rupture ; ils sur-pondèrent donc la baisse dans les estimations de 1994. L'effet serait inverse s'ils étaient enquêtés juste avant la rupture. On constate également l'absence 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.