DONNÉES PRODUITES PAR LE RECENSEMENT RÉNOVÉ DE LA POPULATION

Philippe BERTRAND, Guillaume CHAUVET,
Barbara CHRISTIAN, Jean-Marie GROSBRAS

Insee, Programme de rénovation du recensement
Maîtrise d'œuvre Méthodologie

Introduction : la publication des résultats statistiques pour les communes

Dès la fin du premier cycle quinquennal des enquêtes de recensement et en régime de croisière, l'Insee publiera chaque année des résultats statistiques détaillés aux niveaux communal et infra-communal. 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, de A-4 à A, agrégées pour produire des estimations à l'année médiane de la période de cinq ans (cf Dumais [1]).

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 illustratif et fera l'objet de développements ultérieurs.

1. Les communes de moins de 10 000 habitants.

1.1. 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 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 :

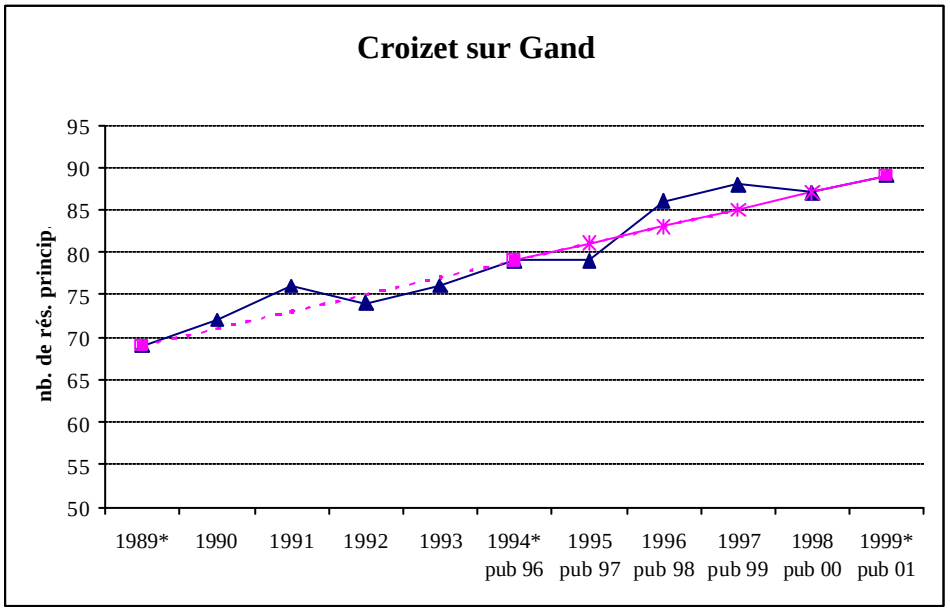
Recensement	**Action**
A-4 (donc aussi A-9)	On extrapole à A-2 la droite (A-9 -- A-4)
A-3 (donc aussi A-8)	On extrapole à A-2 la droite (A-8 -- A-3)
A-2	On garde le recensement
A-1 (donc aussi A-6)	On interpole sur la droite (A-6 – A-1)
A (donc aussi A-5)	On interpole sur la droite (A-5 – A)

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 poids 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.

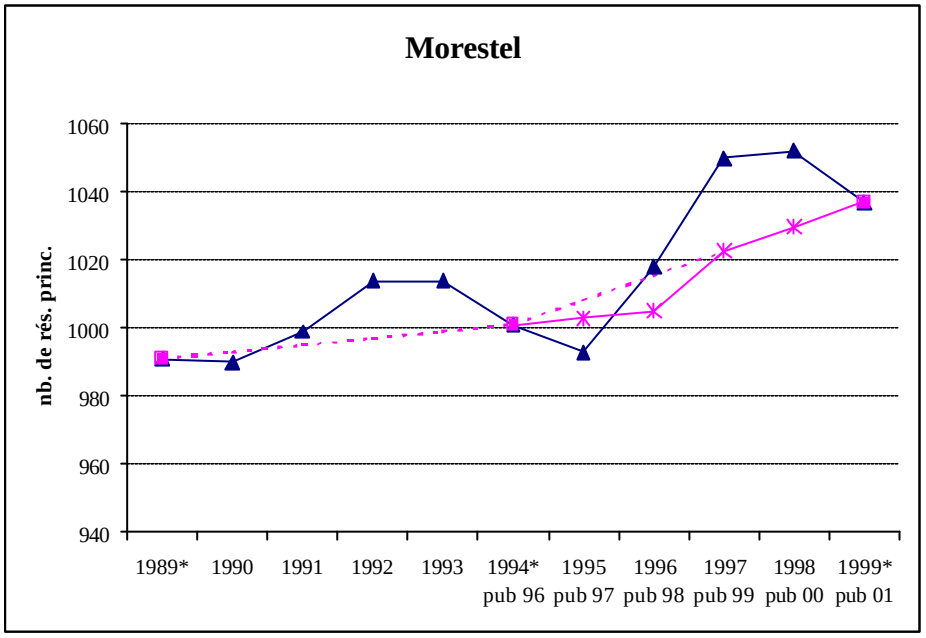
Pour illustrer la méthode et analyser ses risques, il faudrait disposer de séries annuelles de population exactes. En l'absence d'une telle source on a eu recours aux fichiers de la taxe d'habitation, en considérant que le nombre de logements qu'ils contiennent est en étroite corrélation avec l'effectif 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.

1.2. 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 pour 1994 du recensement qui s'est opéré 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.

1.3. 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.

1.4. Conclusion.

Dans la grande majorité des cas, les procédures d'estimation donneront des résultats « raisonnables ». Les cas les plus fragiles 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 plus satisfaisant que dans le système de recensements traditionnels où le délai de prise en compte était 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, indépendamment du fait qu'elle ne rend pas compte de l'évolution de l'ensemble de la population. C'est aussi dans ces cas que l'amélioration des estimations à l'aide d'évolutions constatées sur des sources administrative (taxe d'habitation par exemple) pourra être bénéfique.

2. Les communes de 10 000 habitants ou plus

2.1. L'approche par les sommes mobiles.

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 le problème est de rassembler, pour A-2, les données annuelles récoltées pendante la période. A la fin de l'année A+1, on établira les résultats, millésimés A-1, des enquêtes des années A-3 à A+1, et ainsi de suite.

Pour décrire le plus simplement possible 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, qu'il n'y a pas d'adresses nouvelles pendant la période et que la strate des autres adresses comprend donc exactement 90% des logements partageables en cinq groupes égaux.

En rassemblant 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 d'un coefficient égal à l'inverse du taux de sondage (par exemple 3 si le taux de sondage est 1/3). Ainsi le total d'une variable *Y* se calcule par :

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

où $c_i = 1$ pour les données recueillies dans la strate des grandes adresses et $c_i = 1/t$ pour les données recueillies dans la strate des autres adresses, *t* étant le taux de sondage.

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

- ils dépendront tout d'abord du poids respectifs des strates. En effet, 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 et s'il ny a pas d'adresses nouvelles, le coefficient des données observées dans la strate des autres adresses sera égal à 2,714 (*explication : sur 100 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 plus fiables à l'infra-communal, l'ajustement pourra se faire pour chaque IRIS2000.

Les adresses nouvelles seront traitées avec le même principe que les adresses de grande taille.

2.2. Estimations annuelles produites : études 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 1990. 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 effectif Y , on calcule les 500 estimations $\hat{Y}_j$, $j = 1,...,500$, puis l'écart quadratique moyen et le coefficient de variation :

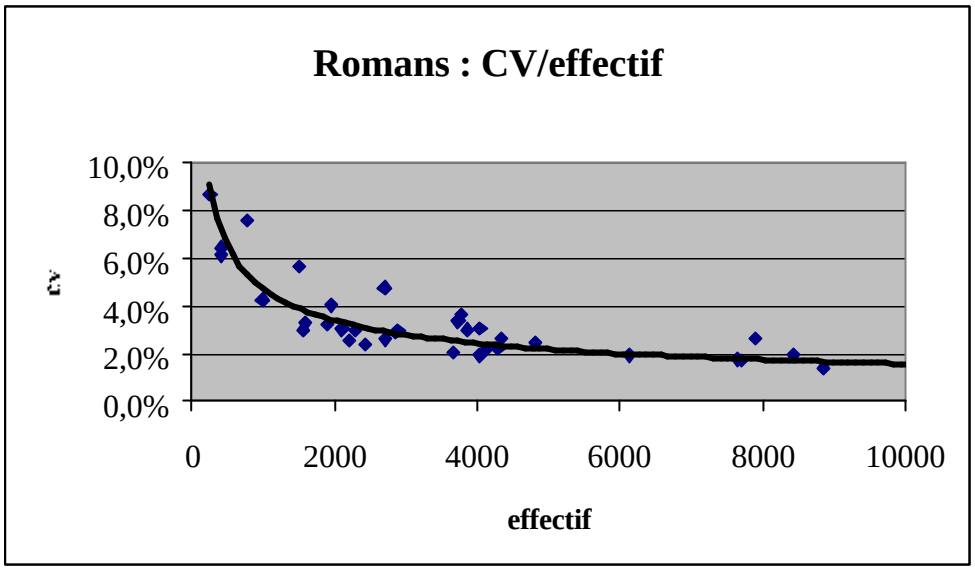
$$V(\hat{Y}) = \frac{\sum_{j=1}^{500} (\hat{Y}_j - Y)^2}{500} \text{ et } CV(\hat{Y}) = \frac{\sqrt{V(\hat{Y})}}{Y} .$$

Le tableau et le graphique suivants montrent les coefficients de variation, **avant calage**, 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. (*Nota : les premières études par simulation des estimateurs après calage montrent des gains moyens de précision entre 0,5 et 1 point par rapport à ce qui est affiché ci dessous*).

Tableau 1. Résultats pour Romans

	effectif	coefficient de variation
hommes de 60 à 74 ans	2106	2,7%
femmes de moins de 20 ans	3857	3,0%
femmes de 60 à 74 ans	2705	2,5%
femmes de 75 ans et plus	1900	3,2%
français	29031	1,0%
étrangers	2697	5,0%
ayant un emploi	10725	1,5%
chômeurs	2855	2,9%
inactifs	18127	1,4%
ouvriers qualifiés	1582	3,4%
ingénieurs, cadres	410	6,4%
résidant.ds autre dép.	3734	3,6%
mariés	12177	1,5%
divorcés	2292	2,9%
loge.occ. ou rés.secondaires	255	8,9%
logements vacants	1501	4,2%
logts dt st.occ.=propriétaire	6128	1,8%
logts dt st.occ.=locataire	7654	1,5%
logts de quatre pièces et plus	8847	1,4%

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



Pour illustrer les résultats concernant les statistiques de structure, le tableau suivant donne, par exemple, les écarts-types du taux de 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 2. Part des moins de 20 ans par IRIS

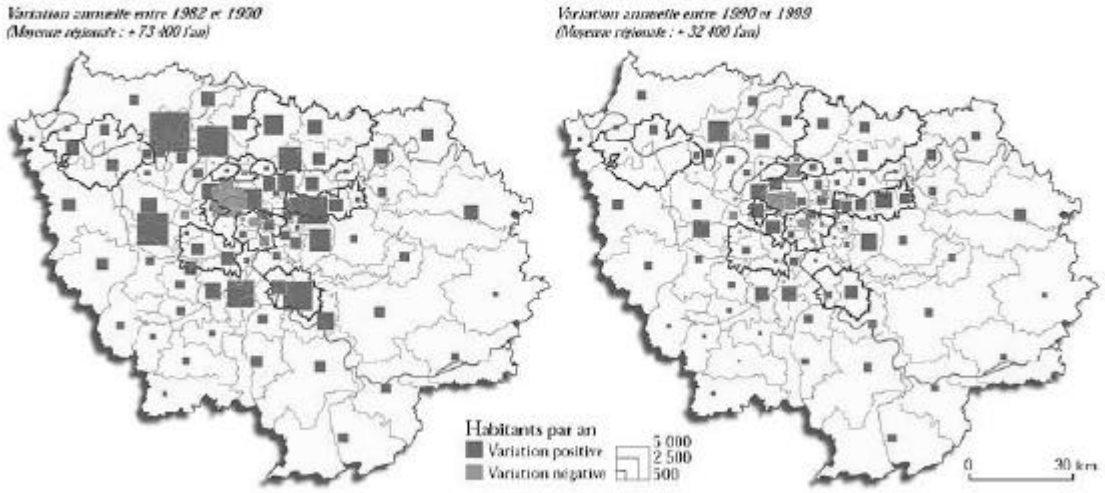
	Part des moins de 20 ans		
IRIS2000	vraie valeur	écart-type	écart-type sas
0103	20,0%	1,6%	1,1%
0104	22,8%	1,5%	0,9%
0201	22,3%	1,8%	1,1%
0202	20,6%	1,1%	0,9%
0301	23,5%	1,6%	1,0%
0302	22,3%	1,4%	1,0%
0303	18,3%	1,6%	0,9%
0304	18,8%	1,3%	0,9%
0401	20,8%	1,5%	1,0%
0402	21,0%	2,1%	1,1%
0403	20,0%	1,8%	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. Les séries annuelles.

3.1. 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 Île-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 recensement rénové pourront apporter des éléments chiffrés à ces questions : plutôt que des photographies espacées dans le temps, le RRP fournira un diaporama annuel, même si le « grain » de chaque diapositive est moins fin que celui d'une photographie exhaustive mais susceptible d'être rapidement obsolète.

Évolution de la population par secteur du SDRIF



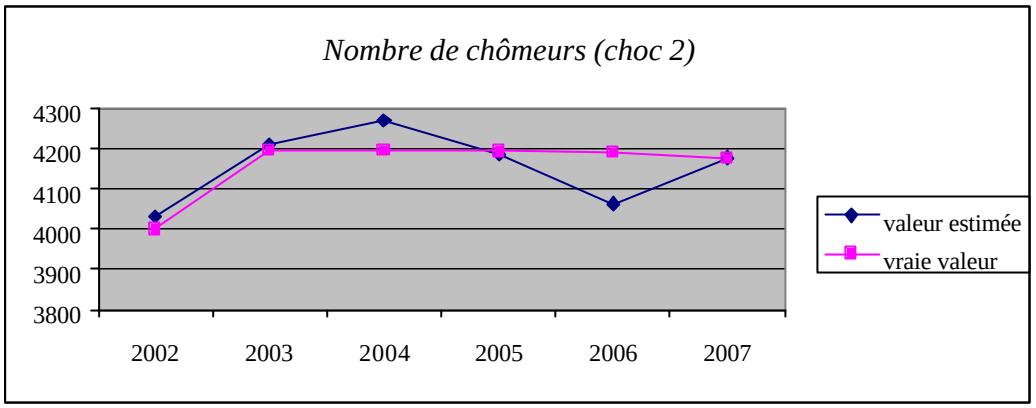
3.2. Travaux de simulation entrepris.

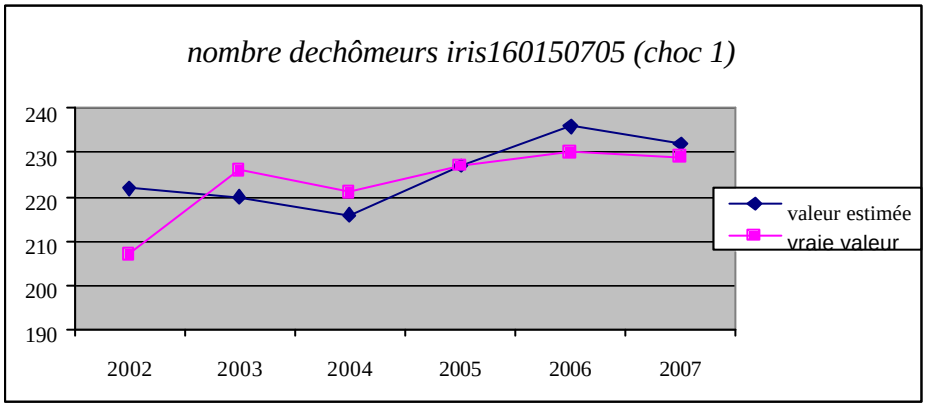
Dans un premier temps, nous sélectionnons des communes avec des quartiers assez typés, d'après les données du recensement de 1999 et imaginons des scénarios d'évolution sur plusieurs années : scénario d'évolution régulière avec des taux démographiques moyens (natalité, etc.), puis introduction de chocs tels que la fermeture d'une usine entraînant une élévation du nombre de chômeurs dans le quartier de l'usine et les quartiers environnants (choc 1), rénovation urbaine avec destruction d'immeubles et relogement, non uniformément réparti des habitants dans le reste de la commune (choc 2), 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 ses simulations est d'examiner comment les séries estimées reflètent la réalité que l'on a introduit dans les scénarios. 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).

Les graphiques suivants montrent ce que peuvent produire les simulations.





Bibliographie

[1] Dumais J., Bertrand Ph., Kauffmann B., « 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