



**Conseil économique
et social**

Distr.
GÉNÉRALE

ECE/CES/AC.6/2008/8
4 Mars 2008

Original: FRANÇAIS

COMMISSION ÉCONOMIQUE POUR L'EUROPE

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Réunion commune CEE/Eurostat sur les recensements
de la population et des habitations

Onzième réunion
Genève, 13-15 mai 2008
Point 4 de l'ordre du jour provisoire

EDITION ET VALIDATION DES DONNEES

La validation des données du recensement en France

Note par l'institut national de la statistique et des études économiques, France

Résumé

Ce document décrit les procédures développées par l'institut national de la statistique et des études économiques pour la validation des données du recensement en France. La méthode adoptée pour calculer la population des communes est rappelé d'abord. Après, les opérations de validation sont présentées pour tous les sources de données qui sont utilisées pour le calcul de la population des communes.

I. RAPPEL DE LA METHODE DE CALCUL DES CHIFFRES DE POPULATION ET DES RESULTATS STATISTIQUES.

1. L'institut national de la statistique et des études économiques (INSEE) doit satisfaire deux conditions pour la publication des chiffres de population:
 - (a) publier tous les ans la population de toutes les communes;
 - (b) pour des raisons d'égalité de traitement des communes entre elles et de qualité pour la population de zones supra-communales (cohérence des chiffres à additionner), cette population doit être relative à la même année pour tout le monde. On ne peut pas retenir pour une commune le chiffre de 2004 et pour une autre le chiffre de 2008.
2. L'INSEE calcule ces populations de la manière suivante:

A. LA POPULATION DES MENAGES

3. Pour les communes de 10,000 habitants et plus, le chiffre repose sur une moyenne mobile établie à partir des échantillons de cinq années successives. Par agrégation des cinq échantillons des années A-4 à A, on calcule une population moyenne par logement, représentative de la situation du milieu de période (l'année A-2). On multiplie ensuite ce ratio par le nombre de logements au début A-2 tiré du Répertoire d'Immeubles Localisés (RIL), pour obtenir la population des ménages de la commune.
4. Parallèlement, pour les communes de moins de 10,000 habitants, il faut ramener le chiffre de la population à l'année médiane du cycle quinquennal, pour être cohérent avec les communes de 10,000 habitants ou plus.
5. On s'appuie sur les enquêtes de recensement, selon le schéma suivant.

A-4	A-3	A-2	A-1	A
Recensement		→		
	Recensement			
		Recensement		
			← Recensement	
				← Recensement

6. Pour les communes recensées en A-2, on retient le résultat de l'enquête de recensement.
7. Pour les communes recensées en A-1 et A, on obtient la population en A-2 par interpolation entre l'enquête de recensement et le dernier résultat publié. Pour les communes

recensées en A-4 et A-3, on procède par extrapolation entre le résultat de l'enquête de recensement et A-2 ; cette extrapolation s'appuie sur les données de la Taxe d'Habitation (impôt local assis sur les logements), qui fournissent une indication sur l'évolution du nombre de logements par commune. Elle est affinée pour tenir compte du différentiel entre la croissance des logements et la croissance du nombre de personnes. Ce différentiel, mesuré entre les derniers recensements, est appliqué à l'évolution mesurée par la TH pour donner l'évolution du nombre de personnes.

B. LA POPULATION HORS MENAGES

8. La population des communautés est ramenée au 1 janvier 2006, soit à la même date que la population des ménages. Si les communautés de la commune ont été recensées en 2006, le chiffre issu de la collecte est retenu. Si elles ont été recensées en 2004 ou 2005, la population des communautés est actualisée en ajoutant la population des communautés nouvelles et en retranchant celle des communautés disparues. Ces évolutions sont réalisées à partir du répertoire des communautés. Si elles ont été recensées en 2007 ou 2008, la population des communautés est actualisée par interpolation entre le dernier chiffre publié et celui fourni par le recensement des communautés.

9. Les populations des personnes sans abri, vivant dans des habitations mobiles ou dont la résidence principale est à l'hôtel, ne sont pas actualisées. Elles sont reportées à l'identique durant les quatre années qui suivent l'année de la collecte. La nouvelle collecte remplace la précédente dès qu'elle est disponible.

II. LA VALIDATION, PROCESSUS PAR PROCESSUS

10. Les ingrédients suivants sont donc nécessaires pour le calcul de la population des communes :

- (a) les fichiers issus des différentes collectes ;
- (b) le stock de logements tiré du répertoire d'immeubles localisés (RIL);
- (c) le répertoire des communautés;
- (d) la taxe d'habitation.

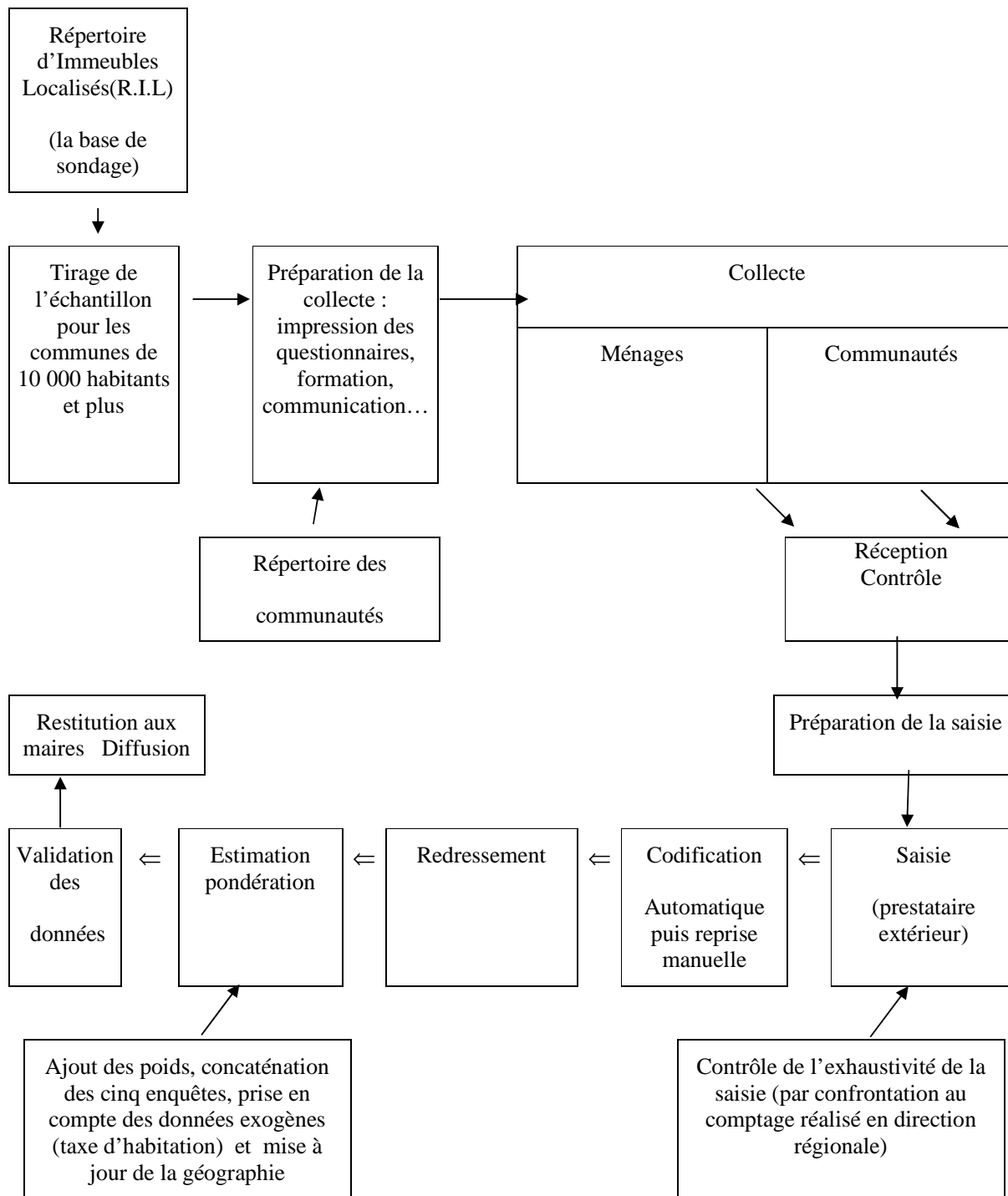
11. La validation portera sur chacun de ces ingrédients. Elle est menée au fur et à mesure de leur élaboration ou de leur traitement, car les erreurs sont d'autant plus faciles à corriger qu'elles sont détectées précocement.

12. La validation comprendra également un examen de deux aspects pouvant jouer sur la qualité ou l'interprétation des données : les évolutions dues aux différences de concepts d'un recensement à l'autre et l'effet des redressements. Dans les grandes communes, du fait d'une étape de pondération et d'extrapolation des résultats du sondage, la validation portera également sur l'effet des « grandes adresses », qui peuvent générer des effets de grappe.

13. La validation a deux objectifs : d'une part s'assurer de la qualité statistique des données que l'on va publier. D'autre part, comprendre les évolutions, estimer leur robustesse et en déduire des précautions d'emploi. Cet aspect est particulièrement important si des différences de concepts affectent particulièrement une commune donnée.

14. Le diagramme ci-dessous montre l'enchaînement des différentes étapes du recensement.

Diagramme. Les étapes du recensement



A. Le stock de logements tiré du RIL

15. La base de sondage, dans chaque grande commune, est constituée par son Répertoire d'Immeubles Localisés (RIL). Ce répertoire est une liste des immeubles (qu'ils soient d'habitation, administratifs, industriels ou commerciaux), identifiés et localisés à leur adresse, grâce à un Système d'Information Géographique.

16. Le RIL a été initialisé en partant du recensement général de 1999 et mis à jour depuis au moyen de fichiers administratifs (permis de construire, fichiers de gestion de la fiscalité locale) ou postaux (fichiers d'adresses de La Poste). Il est soumis chaque année à l'expertise des communes et validé in fine par l'INSEE.

17. Il y a donc une étape de validation importante de ce RIL, juste avant sa mise en œuvre comme base de sondage.

18. Les directions régionales établissent ensuite un « rapport d'édition » pour chaque commune; il s'agit d'un rapport d'analyse de la qualité du répertoire, il comporte des appréciations sur la participation de la commune aux travaux d'expertise, les sources de mise à jour, etc.

19. Une fois le RIL validé, la direction générale procède à une mise à disposition d'éléments d'information complémentaires pour les DR (séries de nombre de logements, information sur les grandes adresses, écart collecte-Ril) dans l'année qui suit la livraison, pour leur permettre de mieux cerner l'impact statistique des évolutions du RIL.

20. Au niveau national, voire régional, la mesure de la qualité du Répertoire d'immeubles localisés (RIL) est réalisée grâce à une enquête annuelle, menée sur le terrain par des enquêteurs de l'INSEE.

21. Son objectif est de mesurer l'exhaustivité du RIL en terme de déficit et d'excédent d'adresses de catégorie habitation, et de logements associés. Elle vise à s'assurer que toutes les adresses rencontrées sur le terrain sont bien répertoriées dans le RIL et inversement que toutes les adresses du répertoire existent sur le terrain.

Tableau 1. Déficits et excédents d'adresses d'habitation et de logements du RIL

		Ril 2003	Ril 2004	Ril 2005	Ril 2006	Ril 2007
Adresses	En déficit	2,3%	1,9%	1,3%	1,3%	1,4%
	En excédent	2,0%	1,9%	1,8%	1,7%	1,4%
	Solde excédent/déficit	-0,3%	0%	0,5%	0,4%	0%
Logements	En déficit	2,3%	1,9%	1,1%	0,9%	1%
	Excédent	2,1%	2,3%	1,8%	1,4%	1%
	Solde excédent/déficit	-0,2%	0,4%	0,7%	0,5%	0%

Source : INSEE, enquêtes qualité du RIL 2003 à 2007.

B. Les fichiers tirés de la collecte

22. La qualité de ces fichiers tient à deux facteurs: la qualité de la collecte et celle de la saisie. Par ailleurs, les redressements effectués visent à améliorer la qualité, par correction des non-réponses ou des réponses incohérentes. L'étape de codage fait également l'objet d'un contrôle de qualité.

23. La collecte est souvent le maillon le plus délicat du processus statistique. Sa qualité repose sur des procédures précises et rigoureuses, mais également sur l'adhésion des répondants (qui conditionne la sincérité de leurs réponses) et l'engagement de ceux qui recueillent l'information, c'est à dire les agents recenseurs. Une préparation rigoureuse (repérages, formations) et des contrôles permettent d'en assurer la qualité. Des contrôles sont réalisés par les communes et l'INSEE pendant la collecte. Après la collecte, le contrôle est du ressort de l'INSEE ; il permet de vérifier les données collectées et le cas échéant de les corriger.

24. Pour les résidences principales que l'on n'a pas réussi à enquêter (personnes impossibles à joindre, absentes de longue durée ou qui refusent de répondre), l'agent recenseur remplit une Fiche de Logement Non Enquêté (FLNE), en mentionnant le nombre de personnes supposées habiter ce logement, ce qui permettra de redresser les chiffres de population. Les risques d'omission sont donc réduits.

Après la collecte: les contrôles menés par l'INSEE

25. Après la réception à l'INSEE des documents transmis par les communes, des vérifications sont menées en bureau, puis sur le terrain. Le protocole de contrôle est adapté à la nouvelle configuration des enquêtes de recensement. Par rapport aux recensements généraux, des vérifications plus approfondies sont rendues possibles pour chaque commune par le moindre volume de données collectées chaque année et le plus faible nombre de communes. Le fichier de la Taxe d'Habitation joue un rôle important dans les contrôles en bureau et permet de vérifier l'exhaustivité de la collecte sur des communes ou des zones de collecte données. Ces contrôles, en bureau ou sur le terrain, sont menés « au plus près de la collecte ». Les erreurs qu'ils permettent de détecter sont corrigées.

26. Ils sont sélectifs car ils résultent de « tamis » successifs:

- (a) un premier contrôle lors de la réception et de l'enregistrement des questionnaires permet d'établir une appréciation de la qualité pour chacune des communes ; il comporte notamment un comptage des documents reçus, par « flashage » des codes-barres présents sur chacun des documents. Ce comptage est confronté d'une part à celui que la commune a réalisé lors de la clôture de la collecte, d'autre part à un nombre attendu de documents (au regard du recensement précédent et d'indications de tendance). Le comptage est disponible à un niveau infra communal : le secteur d'agent recenseur, ce qui permet de cibler, si nécessaires, les contrôles de cohérence et de qualité;
- (b) cette appréciation débouche, pour environ 10 pour cent, sur un contrôle approfondi en bureau, portant sur tout ou partie de la commune;
- (c) enfin, les cas non résolus en bureau font l'objet de contrôles menés sur le terrain par l'INSEE (environ 28,000 contrôles effectués sur le terrain en 2007). Un

logement enquêté sur 200 fait l'objet d'un contrôle sur le terrain. Ces contrôles confirment la collecte dans la plupart des cas. Ils permettent également de détecter des erreurs. Ces erreurs sont corrigées et les documents erronés sont remplacés par des documents sans erreur, qui peuvent ainsi intégrer la chaîne de saisie.

27. Un effort particulier est mené sur le nombre et la qualité des FLNE. Il convient de réduire le taux de FLNE à quelques pourcents, de manière à ne pas fragiliser les estimations. Les taux sont actuellement d'environ 3 pour cent, grâce à de gros efforts réalisés lors de la collecte auprès des ménages.

28. Il convient également de s'assurer que les FLNE sont bien établies pour des résidences principales, et que le nombre de personnes indiqué donne une bonne estimation de la taille des ménages. Les contrôles de terrain permettent de vérifier cette information.

La validation de la saisie

29. La saisie est effectuée par un prestataire externe. Il y a saisie optique des questionnaires, qui sont scannés puis soumis à un logiciel de reconnaissance des formes. Une saisie des libellés mentionnés en clair complète le processus. Un cahier des charges précis définit ce qu'on attend du prestataire sur la qualité. Ceci concerne à la fois la sécurité des données, pendant les transports et chez le prestataire (sécurité physique et informatique, engagement de chaque personne à ne pas divulguer les informations traitées) et la qualité des données (nombre de documents saisis par zone de collecte, taux d'erreur maximal par variable). Enfin il y a comptage des documents, et contrôle des écarts avec le premier comptage réalisé à l'INSEE lors du flashage (cf. ci-dessus).

30. Au cours du processus, les documents (bulletins individuels et feuilles de logement) sont donc comptés trois fois: une fois par la commune lors du bouclage de la qualité, une fois par l'INSEE lors de la phase de contrôle, une fois par le prestataire de saisie. La confrontation de ces trois comptages permet de garantir un niveau très élevé de qualité sur le nombre de bulletins.

31. Enfin, la qualité de la saisie est mesurée par une double saisie (cf. encadré)

Une double saisie pour vérifier la qualité des variables saisies.

Les erreurs de saisie sont évaluées par double saisie sur un échantillon de bulletins, chez un second prestataire, qui travaille à partir des images scannées en utilisant les mêmes règles de saisie. Les divergences sont analysées à l'INSEE, de façon à établir des taux d'erreur sur le codage de chacune des variables. La double saisie est réalisée au fil de la campagne de saisie, ce qui permet, éventuellement, de corriger les protocoles de saisie pour les lots suivants. Des contacts et des réunions régulières sont organisés avec le prestataire, de manière à anticiper les problèmes ou à les résoudre au plus vite. Des ajustements ou des améliorations sont demandés au prestataire, en cours de campagne ou en période inter-campagnes s'il s'agit d'améliorations plus lourdes, suite aux enseignements de la double saisie. Par exemple, l'optimisation des outils de reconnaissance de forme a permis de diminuer significativement le taux d'erreur de saisie des dates de naissance. Il faut noter à ce sujet que, même sur des données très simples comme les dates de naissance, il est illusoire d'espérer un taux d'erreur zéro : l'information étant manuscrite au départ, des erreurs d'interprétation sont inévitables.

32. Les redressements. L'enjeu est d'améliorer la qualité des fichiers tirés de la collecte. Il s'agit d'analyser les non réponses, totales ou partielles, de manière à les imputer, mais aussi les bulletins qui présentent des incohérences (par exemple une personne de 95 ans qui déclare être active). La procédure souvent utilisée est celle du hot-deck, qui consiste à compléter l'information manquante en s'aidant de la réponse d'un « donneur » qui lui ressemble. Là encore, les redressements s'affinent d'année en année à partir des résultats précédents, ce qui constitue un système efficace d'amélioration de la qualité.

33. En particulier, le redressement des FLNE permet d'imputer, pour le calcul de la population des communes, la population des résidences principales pour lesquelles le contact avec l'agent recenseur n'a pas eu lieu (absences de longue durée, personnes impossibles à joindre, refus). Pour les résidences principales dont le nombre de personnes est renseigné (plus de 8 sur 10), on « impute » pour le logement autant d'individus que la FLNE en indique dans le ménage. Pour les autres FLNE, on procède également à une imputation, selon une méthode visant à reproduire la structure, par taille de ménage, observée en moyenne nationale sur les FLNE renseignées. Pour tenir compte du fait qu'un petit nombre de FLNE sont remplies à tort pour des résidences non principales, ce redressement est fait à concurrence de la part des résidences principales dans la commune.

C. Le répertoire des communautés

34. Comme le RIL, le répertoire a été constitué avec le Recensement de 1999 et il est mis à jour avec des sources administratives (fichier des établissements sanitaires et sociaux, fichier des internats, fichier des établissements pénitentiaires ou relevant de la Protection Judiciaire de la Jeunesse,...). Comme le RIL, il est transmis aux communes pour expertise juste avant l'enquête de recensement. Les communes reçoivent la liste de leurs communautés et il leur est demandé d'indiquer les communautés qui auraient disparu, celles dont l'INSEE n'aurait pas connaissance et celles dont l'adresse aurait changé. L'INSEE valide ensuite les propositions de création ou de suppression du répertoire formulées par les communes.

35. Au démarrage de la collecte, une ultime validation de la liste des communautés est demandée à l'agent recenseur des communautés.

36. Enfin, l'INSEE confronte chaque année le RIL et le répertoire des communautés. Il s'agit d'une part de bien géolocaliser chacune des communautés du répertoire ; le RIL est alors utilisé comme liste de référence pour les adresses et leurs caractéristiques géographiques (coordonnées X,Y, appartenance à un quartier, à un canton, à une zone infra-communale). Il s'agit d'autre part de vérifier l'absence de doublons ou d'omissions entre ces deux répertoires (notamment quand une communauté se transforme en ensemble de logements ordinaires ou inversement).

D. Le fichier de taxe d'habitation

37. L'INSEE reçoit chaque année de la part de la direction générale des impôts (DGI) un fichier récapitulant l'ensemble des locaux taxés au titre de la taxe d'habitation. Un premier traitement consiste à écarter les locaux qui ne correspondent pas à des locaux d'habitation (garages, parkings notamment).

38. Le fichier est ensuite utilisé à deux fins:

- (a) suivi et contrôle de la collecte, pour les communes concernées
- (b) établissement de séries de nombre de logements totaux, résidences principales, résidences secondaires. Ces séries sont utilisées pour le calcul des populations des communes de moins de 10,000 habitants.

39. Les communes reçoivent par ailleurs de la DGI l'information sur le nombre de rôles émis et sur l'assiette de l'impôt. On peut penser que, toute omission se traduisant par de moindres recettes fiscales, la commune signale les omissions à la DGI. Par ailleurs, tout excédent (local figurant à tort dans la liste) se traduisant pour le contribuable concerné par une imposition supplémentaire, les excédents sont également signalés à la DGI.

40. L'INSEE analyse systématiquement le profil des séries de logements ainsi constituées et, en cas d'évolution atypique (par exemple une forte baisse suivie d'une forte hausse) interroge la DGI pour détecter d'éventuels artefacts de gestion. Les artefacts confirmés par la DGI sont corrigés.

E. Les extrapolations en grande communes

41. Le chiffre de population des grandes communes est obtenu en extrapolant l'échantillon cumulé sur cinq années de collectes successives. La validation du chiffre obtenu repose sur des indicateurs de robustesse, obtenus en analysant chacune des enquêtes annuelles et chacun des « cumuls ». Les évolutions constatées représentent un indicateur empirique de la robustesse de l'estimation d'ensemble. D'autres critères entrent en compte pour la validation : la cohérence de l'estimation avec les résultats du recensement précédent et les indicateurs exogènes d'évolution donnent également des indications sur la qualité du chiffre : les communes pour lesquelles les données sont jugées les plus surprenantes (par le niveau national ou les directions régionales) font l'objet d'une analyse approfondie, destinée à déceler d'éventuels problèmes de qualité du RIL ou d'effets de grappe non traités. Les communes pour lesquelles apparaissent ce type de problème peuvent faire l'objet d'estimations spécifiques destinées à corriger ces défauts.

42. La fiche jointe en annexe a été utilisée pour la validation des estimations provisoires fondées sur quatre enquêtes. Elle est complétée par une analyse infra-communale visant à mettre en évidence des quartiers dont l'évolution, très particulière, pèserait particulièrement sur le résultat final.

F. Les écarts de concepts entre les deux recensements

43. Ces écarts peuvent jouer sur l'analyse des évolutions de la population d'une commune. Il convient donc d'établir une liste des communes a priori fortement impactées par ces écarts pour éclairer la validation de ces communes et les analyses statistiques reposant sur ces communes (ou des zones englobantes). Par exemple, le fait de prendre en compte, au nouveau recensement, les élèves internes majeurs dans la commune de leur internat et non plus dans la commune de leurs parents fait évoluer fortement la population municipale de la commune de l'internat. Cet effet doit être signalé aux valideurs et aux utilisateurs des données.

G. La validation finale

44. Chacun des ingrédients étant validé au fur et à mesure de sa mise en place, il n'y a pas à proprement parler d'étape finale de validation. C'est surtout l'occasion de rassembler tous les éléments qualitatifs et quantitatifs, qui permettent d'expliquer le chiffre de population. C'est essentiel pour répondre aux éventuelles questions que peuvent être amenés à poser les communes, mais aussi d'autres utilisateurs des chiffres de population.

III. LA VALIDATION DES DONNEES DETAILLEES

45. Au-delà du seul chiffre de population, la validation concerne l'ensemble des données détaillées. Certes, la validation de la qualité du chiffre total garantit que la plupart des problèmes de qualité (couverture, effets liés à l'échantillonnage, etc.) ont été détectés et qu'il en a été tenu compte. En complément il est procédé à deux validations complémentaires.

A. La validation de la codification des questionnaires

46. Le codage comprend d'abord une phase automatique de codification selon des nomenclatures d'activité et de profession notamment, puis une phase de reprise par des opérateurs en direction régionale des cas non codés. On développe actuellement un dispositif de contrôle de qualité sur ce processus. Le principe est de procéder à une seconde codification sur un échantillon de bulletins représentatifs des différents modes de codification (automatique et reprise), puis d'analyser les divergences.

47. L'objectif est de mesurer la qualité du codage automatique et celui de la reprise manuelle (obtenir d'abord un pourcentage global de cas bien codés, puis à terme des pourcentages par grandes modalités des variables, par exemple pourcentage de cadres supérieurs bien codés). Un deuxième objectif de l'outil de contrôle qualité du codage est de mesurer un taux de bulletins individuels non codables (exemple, un intitulé vague de type fonctionnaire, sans autre indication), ce qui permettra de se donner une cible réaliste en matière de qualité et d'éviter la « sur-qualité ».

48. Ce dispositif de mesure et de contrôle permet aussi d'enrichir les fichiers d'apprentissage qui servent à la codification automatique, de manière à en améliorer la performance, de mieux cibler la formation des codeurs de la reprise, et d'assurer finalement un bon management de la qualité. L'apport de cette étape a été sensible dès la campagne suivante.

B. Un contrôle de vraisemblance sur les données détaillées.

49. Les données détaillées sont analysées, à partir de « tris à plat » des différentes variables (pyramide des âges, état matrimonial, structures familiales, emploi, parc de logement, etc.). Il s'agit d'une part de confronter ces données avec des données provenant d'autres sources, d'autre part d'apprécier leur robustesse en analysant leurs évolutions d'une enquête à l'autre. L'analyse des niveaux peut permettre d'évaluer la qualité des procédures de redressement. Couplée avec des estimations de la précision, l'analyse de la robustesse doit permettre de définir le niveau de finesse de l'information diffusée (par exemple diffuse-t-on des âges quinquennaux ou des âges

décennaux?). La validation est aussi l'occasion de cerner et de quantifier si possible l'effet des changements de questionnaires. C'est une information essentielle pour les utilisateurs qui veulent analyser des évolutions par rapport au recensement précédent.

50. Les analyses menées sur les premières années ont permis de déceler certains effets indésirables de la procédure de redressement (que l'on a limités pour les enquêtes suivantes) ou de protocoles de saisie trop imprécis (erreur rectifiée à partir de 2005). Elles ont également donné lieu à des analyses des effets du changement de questionnaire (notamment sur l'emploi et sur les structures familiales).

51. Actuellement, un « crible » analyse les évolutions des principales variables statistiques et invite les directions régionales à analyser les cas de problèmes.
