



**Conseil économique  
et social**

Distr.  
GÉNÉRALE

ECE/CES/AC.6/2008/9  
4 mars 2008

FRANÇAIS  
Original: ANGLAIS

---

**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

**VÉRIFICATION ET VALIDATION DES DONNÉES**

**Aperçu général des méthodes de contrôle et d'imputation utilisées  
pour les recensements à venir en Italie**

Note de l'Institut national italien de statistique

**Résumé**

L'Institut national italien de statistique prépare actuellement la prochaine campagne de recensement de la population et des logements qui aura lieu en 2011. Le présent document décrit brièvement les principales méthodes de vérification des données qui seront utilisées pour le recensement de 2011. Il examine aussi les incidences sur le contrôle et la validation de certaines innovations qui devraient être adoptées à l'occasion de ce recensement.

## I. INTRODUCTION

1. Les données de recensement présentent des caractéristiques qui rendent la phase de contrôle et de validation de ces données très complexe. Elles partagent certaines caractéristiques avec les données issues d'autres enquêtes sur les ménages, comme la structure hiérarchisée des unités employées aux fins de collecte et d'analyse des données (ménages et individus, immeubles et logements), ou l'association de nombreuses variables qualitatives et quantitatives. Les caractéristiques propres aux données de recensement sont par exemple l'important volume de données exigeant des solutions de calcul hautement efficaces, et la présence dans la population de groupes peu nombreux, mais importants, qu'il est difficile de recenser – personnes âgées, bébés, travailleurs frontaliers, étrangers, etc. – et pour lesquels il faut procéder à des estimations par imputation qui reflètent leur distribution effective. Une stratégie bien conçue de contrôle et de validation, adaptée aux caractéristiques des données de recensement, doit donc être prévue.

2. La stratégie de contrôle et de validation adoptée dans le cadre du recensement italien de 2001 a été marquée par des innovations méthodologiques et techniques visant à remédier à certains problèmes de contrôle. En particulier, de nouveaux procédés et de nouveaux logiciels ont été conçus, appliqués, et utilisés avec succès en vue d'améliorer la qualité du résultat final.

3. Pour l'heure, il n'est pas encore possible de donner tous les détails de la stratégie de contrôle et de validation pour 2011, car on étudie actuellement la possibilité d'adopter pour ce recensement certaines innovations qui auraient des conséquences pour la phase de contrôle et de validation.

4. On peut dire, d'ores et déjà, que la stratégie de validation et de contrôle de 2011 sera élaborée sur la base des enseignements dégagés du recensement de 2001: un processus de contrôle et de validation mixte sera mis en place en associant et hiérarchisant plusieurs procédures qui tendent à résoudre des problèmes spécifiques. En particulier, certaines procédures de contrôle utilisées en 2001 seront adaptées au questionnaire de 2011 et réutilisées, tandis que d'autres seront améliorées et/ou modifiées pour tenir compte des innovations qui seront adoptées. L'utilisation de logiciels communs permettra d'ajuster facilement les données et d'adapter rapidement les procédures à la nouvelle stratégie.

5. La stratégie de contrôle et de validation pour 2011 visera également à réduire les délais de traitement des données collectées de façon que les résultats soient diffusés rapidement (Eurostat, 2008). Les facteurs suivants seront déterminants à cet égard: un matériel informatique technologiquement à jour, des algorithmes hautement efficaces, et des procédures de contrôle et de validation bien planifiées, exécutées et gérées.

6. Le présent document fait ressortir les composantes essentielles du processus de contrôle et de validation de 2011 en signalant certains aspects à prendre en compte lors de la phase préparatoire. Plus particulièrement, la section II décrit les principales procédures de contrôle qui seront utilisées pour le recensement de 2011, et la section III précise les incidences sur la stratégie de contrôle et de validation des innovations envisagées lors de la conception de l'enquête. Enfin, la section IV fait état d'éléments de réflexion supplémentaires sur les questions actuellement à l'étude et sur les travaux futurs.

## II. PRINCIPALES PROCÉDURES DE CONTRÔLE POUR LE RECENSEMENT DE 2011

7. La stratégie de contrôle du recensement de 2001 avait pour objet de réaliser des imputations plausibles qui préservent la plus grande quantité possible de renseignements collectés. Pour ce faire, on avait cherché à isoler les problèmes liés au contrôle et à trouver une solution appropriée pour chacun. En conséquence, un processus de contrôle composé de plusieurs procédures – interdépendantes – remédiant à des sous-problèmes spécifiques, et faisant appel à des solutions méthodologiques adaptées, avait été mis en place (Bianchi *et al.*, 2004 et 2005).

8. Pour des raisons liées aux calculs et d'ordre pratique, les opérations de contrôle relatives aux variables incluses sur la feuille de collecte d'informations sur les ménages privés<sup>1</sup> du recensement de 2001 ont été réalisées en deux étapes successives. Le contrôle portant sur les variables démographiques (Année de naissance, Sexe, Lien avec la personne de référence du ménage, Situation matrimoniale et mariage) est intervenu avant celui qui concerne les autres variables, appelées variables individuelles (Nationalité, Présence sur le territoire et logement, Degré d'instruction et formation, Situation professionnelle et non professionnelle, Type de travail, Lieu d'étude ou de travail). De ce fait, il pouvait arriver que des réponses individuelles soient effacées par erreur (ou mal imputées) en fonction des valeurs imputées pour les variables démographiques. Ce problème a été résolu en utilisant une variable auxiliaire appropriée (voir sous-section A) reposant sur l'identification du chemin du recensé.

9. Les techniques d'optimisation ont joué un rôle important dans le traitement des variables démographiques. Un nouveau logiciel basé sur ces techniques a en fait été conçu pour appliquer l'approche pilotée par les données et la méthode du changement minimum (théorique), lesquelles avaient été utilisées pour traiter les variables démographiques lors du recensement de 2001 (voir sous-sections B et C).

10. L'identification du chemin du recensé a aussi constitué une étape essentielle de la stratégie de contrôle retenue pour les variables individuelles. Les chemins permettent de définir des classes de personnes qui partagent certaines caractéristiques (voir sous-sections A et D). Ces classes ont été utilisées comme des strates pour repérer les erreurs et opérer les imputations. La méthode déterministe a été appliquée principalement pour détecter les erreurs concernant les variables individuelles, tandis que les nouvelles valeurs ont été affectées en se servant d'une méthode d'imputation à base de modèles (voir sous-section D).

11. Une attention particulière a été prêtée au contrôle des données relatives à certains groupes restreints de population, comme les centenaires, les bébés et les travailleurs frontaliers. Des procédures spécifiques ont été élaborées pour ces groupes, en faisant aussi appel à la corrélation avec les données du recensement de 1991 (voir sous-section E).

---

<sup>1</sup> Bulletin de recensement employé pour recueillir des données sur les personnes qui résident habituellement dans le logement considéré.

12. La stratégie de contrôle du recensement de 2011 sera fondée sur celle qui avait été adoptée en 2001. En particulier, le même ordre de traitement (variables démographiques, puis variables individuelles) sera suivi, et les principales procédures employées pour traiter les deux types de variables seront adaptées et réutilisées. On trouvera dans les sous-sections suivantes une brève description des aspects méthodologiques de ces procédures.

### **A. Identification du chemin du recensé**

13. D'après la théorie des graphes (Picard, 1980), un questionnaire peut être représenté comme un graphe connexe où les sommets seraient les variables et où les réponses définiraient les arêtes. Lorsque la réponse correspondant à une variable n'est nécessaire que pour certaines valeurs attribuées à une variable prise en compte précédemment, la variable antérieure est appelée «filtre», et la suivante «dépendante». Par exemple, Situation matrimoniale est une variable filtre pour la variable dépendante Années de mariage. Les variables filtres sont représentées par des sommets d'où partent plusieurs arêtes. Chacune de ces arêtes rejoint un sommet successif qui représente une variable dépendante. Deux sommets sont adjacents s'ils sont reliés par une arête. Un chemin est une séquence de sommets adjacents distincts. Un chemin permet d'identifier une classe de recensés qui partagent des caractéristiques. Il peut donc être considéré comme une variable synthétique utile aux fins d'identifier les sous-populations de recensés.

14. Les réponses manquantes et/ou les valeurs erronées peuvent entacher d'incertitude l'identification du chemin exact de certains recensés, et du même coup celle des bonnes classes de recensés. En fait, un échec de la procédure de contrôle peut donner un résultat erroné (non admissible), c'est-à-dire un chemin qui n'est pas compatible avec les règles de compilation du questionnaire. Dans ces cas de figure, il est nécessaire de trouver le (les) chemin(s) admissible(s) les plus probables.

15. Pour le recensement de 2001, une procédure automatique avait été mise en œuvre pour l'identification du chemin le plus probable de chaque recensé. Cette procédure permettait de repérer le chemin le plus vraisemblable parmi l'ensemble de ceux qui étaient admissibles à partir de l'analyse des réponses données aux questions filtres et dépendantes. Si deux possibilités ou plus étaient admissibles (solutions multiples), un chemin était sélectionné aléatoirement en fonction de la distribution de fréquence observée des chemins les plus probables du recensé.

16. Le chemin du recensé a été utilisé pour:

a) Calculer la nouvelle variable auxiliaire Sous-ensemble de valeurs admissibles (SVA) de la variable Année de naissance (Manzari *et al.*, 2002). Le SVA de l'Année de naissance (abrégé ci-après «SVA») a été calculé pour chaque membre du ménage et a défini un sous-domaine de l'Année de naissance correspondant au plus grand nombre de réponses données. Il a été utilisé pour identifier des strates de donneurs lors du traitement des variables démographiques: un membre d'un ménage pour lequel le contrôle avait réussi était un donneur convenable pour un membre d'un ménage dans le cas duquel le contrôle avait échoué (le receveur), si, et seulement si, son Année de naissance était comprise dans le SVA du receveur. Comme les variables individuelles étaient traitées au cours de la deuxième étape du contrôle, les valeurs correspondantes pouvaient être déterminées par la valeur imputée de la variable Année de naissance, et des suppressions erronées pouvaient se produire. L'utilisation du SVA a permis d'imputer une Année de naissance conforme au nombre le plus élevé de

valeurs individuelles, de sorte que les pertes d'information dues à des suppressions erronées ont été grandement réduites;

b) Procéder aux imputations requises en cas d'absence de réponse ou d'incohérence pour les variables Situation professionnelle, Type de travail et Lieu d'étude ou de travail. En particulier, pour chaque cas d'échec du contrôle, le chemin le plus probable a d'abord été sélectionné, puis la valeur à imputer a été choisie en tenant compte du chemin du recensé, du Lieu d'habitation, de la Classe d'âge, du Sexe et d'autres variables de la strate.

### **B. Utilisation combinée de l'approche pilotée par les données et de la méthode du changement dans le système DIESIS**

17. Les données de recensement de la population sont recueillies auprès des ménages, et sont collectées pour chaque membre du ménage. Les variables démographiques sont corrélées entre les différents membres du ménage au moyen de ce qu'il est convenu d'appeler les contrôles interpersonnels, et au niveau de chaque personne au moyen d'un contrôle individuel. Ainsi, la conservation des relations entre variables démographiques peut être garantie si les variables sont traitées en même temps pour tous les membres du ménage, c'est-à-dire au niveau des ménages. À cette fin, il faut employer un système d'édition capable de réaliser les contrôles aussi bien individuels qu'interpersonnels. Comme certains contrôles démographiques sont exprimés par des inégalités linéaires, le système doit aussi être capable de traiter simultanément les variables qualitatives et quantitatives.

18. Au cours de la préparation du recensement de 2001, un nouveau logiciel, Data Imputation and Edit System – Italian Software (DIESIS), a été mis au point conjointement par l'Institut national italien de statistique (ISTAT) et des chercheurs du Département d'informatique et de science des systèmes de l'Université de Rome «La Sapienza» (Bruni *et al.*, 2001). Le système DIESIS permet de traiter les variables qualitatives et quantitatives simultanément, au niveau tant des ménages que des individus. Après une évaluation statistique rigoureuse de ses performances (Manzari et Reale, 2001), le système DIESIS a été appliqué avec succès, pour l'imputation en cas de non-réponse et la solution des problèmes d'incohérence dans les réponses, en ce qui concerne les variables démographiques retenues en 2001.

19. Deux méthodes de contrôle sont appliquées dans le système DIESIS: l'approche pilotée par les données et la méthode du changement minimum (théorique), au moyen d'algorithmes «donneurs/champs» et «champs/donneurs».

20. L'algorithme «donneurs/champs» identifie d'abord un sous-ensemble de donneurs potentiels, puis détermine le nombre minimum de variables dont la valeur doit être imputée en fonction de ces donneurs. Les donneurs potentiels sont les ménages validés (contrôle réussi) qui sont aussi semblables que possible aux ménages non validés (pour lesquels le contrôle a échoué). Le degré de similitude entre chaque ménage non validé et chaque ménage validé est calculé par une fonction définie comme la somme pondérée des écarts (s'agissant des variables quantitatives) ou des similitudes (s'agissant des variables qualitatives) pour chaque variable applicable aux ménages et en considérant l'ensemble des personnes. L'algorithme sélectionne, parmi les donneurs potentiels, l'ensemble minimum (pondéré) de valeurs à imputer pour que le ménage non validé satisfasse, après ajustement, à tous les contrôles (ce qui implique donc

un changement minimum). En utilisant cet algorithme, les valeurs imputées pour un ménage proviennent d'un seul ménage donneur.

21. L'algorithme «champs/donneurs» détermine d'abord le nombre minimum (pondéré) de variables dont les valeurs doivent être imputées et repère les donneurs potentiels (selon la procédure décrite plus haut). Ensuite, pour chaque individu receveur, il prélève les valeurs à imputer auprès du donneur, en choisissant celui qui est le plus semblable possible au receveur. Cet algorithme permet de procéder aux imputations requises pour les variables intéressant une personne à la fois. Si possible, on impute dans le même temps les valeurs des variables propres à la personne. Il convient de noter que les valeurs imputées pour un ménage peuvent provenir de deux ménages donneurs ou plus.

22. Les deux algorithmes ont été appliqués, conjointement, au traitement des variables démographiques, afin de concilier, d'une part, la plausibilité des imputations et, d'autre part, la conservation des renseignements collectés. L'algorithme «donneurs/champs» a été sélectionné par défaut, tout en laissant la possibilité d'opter pour l'algorithme «champs/donneurs» dans les cas où la procédure de contrôle avait échoué pour un ménage donné, et où le nombre de changements proposés par le premier algorithme était trop élevé par comparaison à celui des changements proposés par le second (la marge de variation a été fixée en fonction de la taille du ménage).

23. L'algorithme «donneurs/champs» a été principalement utilisé pour le traitement des données relatives aux ménages partageant la même structure: il s'agit habituellement des ménages de plus petite taille. Pour ces ménages, il s'est avéré possible en règle générale de trouver suffisamment de donneurs potentiels. En revanche, lors du traitement des données relatives aux ménages présentant une structure peu courante, c'est-à-dire, dans la majorité des cas, des ménages de taille importante, peu de donneurs étaient généralement disponibles et, bien souvent, ils n'étaient pas très semblables au ménage pour lequel le contrôle avait échoué. Dans ce genre de situation, la méthode d'imputation pilotée par les données aurait exigé un grand nombre de changements pour obtenir un ménage ajusté qui satisfasse aux contrôles et on a donc préféré appliquer la méthode du changement minimum.

### **C. Identification de la personne de référence et des couples potentiels au sein du ménage**

24. Deux procédures importantes ont été mises en œuvre avant de traiter les variables démographiques. L'une visait à valider la personne de référence au sein du ménage et l'autre était conçue pour identifier les couples potentiels parmi les membres du ménage en question.

25. L'une des principales variables démographiques est le lien avec la personne de référence du ménage. Il s'agit de la variable de base déterminante pour tous les contrôles interpersonnels et la plupart des contrôles individuels. Par ailleurs, il est nécessaire de définir le noyau familial et, par conséquent, la structure de la famille. Dans chaque ménage, la personne de référence doit être identifiée (avant de procéder à la vérification des variables démographiques) pour permettre à tous les autres membres du ménage d'indiquer leur lien avec cette personne. Trois cas d'erreur éventuels peuvent se présenter en ce qui concerne la personne de référence:

a) Une personne a déclaré qu'elle était la personne de référence; or, son Sous-ensemble de valeurs admissibles (SVA) pour l'Année de naissance révèle qu'elle n'a que 17 ans ou moins et il n'est donc pas compatible avec cette fonction;

b) Plusieurs personnes ont déclaré être la personne de référence;

c) Aucun membre du ménage n'a déclaré être la personne de référence.

26. La procédure utilisée pour valider la personne de référence du ménage, lors du recensement de 2001, s'appuyait sur des techniques d'optimisation et a été appliquée en adaptant l'algorithme «champs/donneurs» du système DIESIS à ce problème particulier. La procédure comprenait deux grandes phases. Au cours de la première, les personnes de référence potentielles ont été repérées. Plus précisément:

- i) Dans les cas de figure a) et c) mentionnés plus haut, les personnes de référence potentielles étaient tous les membres du ménage dont le SVA était compatible avec la fonction de personne de référence, c'est-à-dire celles qui étaient âgées de 18 ans ou plus;
- ii) Dans le cas de figure b), les personnes de référence étaient potentiellement toutes celles qui avaient déclaré être la personne de référence pour autant que leur SVA leur permette de remplir cette fonction; sinon, les personnes de référence potentielles étaient tous les membres du ménage dont le SVA était compatible avec la fonction de personne de référence.

27. Si aucun membre du ménage n'avait un SVA compatible avec cette fonction, les personnes de référence potentielles étaient toutes les personnes qui composaient le ménage, abstraction faite de la condition imposant que leur SVA soit compatible.

28. Durant la seconde phase, la procédure a permis de sélectionner la personne de référence potentielle et de rétablir ainsi la cohérence des informations aux fins du contrôle pour le ménage considéré, en changeant le nombre minimum (pondéré) de valeurs des variables démographiques.

29. La définition du noyau familial, et, par conséquent, de la structure familiale, est fondée sur l'analyse des couples au sein du ménage. Certains couples sont liés par une relation à caractère unique avec la personne de référence (personne de référence, épouse/mari; père, mère; beau-père, belle-mère). D'autres couples ont un lien qui n'a pas un caractère unique avec la personne de référence (fils/fille, gendre/belle-fille, par exemple). Les personnes qui constituent des couples à caractère unique sont bien identifiées dans le ménage, si elles sont présentes. En 2001, le contrôle des informations se rapportant à ces personnes a été effectué au moyen d'une série de contrôles interpersonnels spéciaux (contrôles des couples). Le contrôle des couples qui n'ont pas un caractère unique pose davantage de difficultés parce que les personnes qui les constituent doivent être identifiées au préalable.

30. Pour conserver autant que possible les observations recueillies sur les couples qui n'ont pas un caractère unique, on a, lors du recensement de 2001, procédé en deux étapes en s'inspirant de la stratégie suivie au Canada (Bankier *et al.*, 1997; Bankier, 1999). Dans un premier temps, les

couples qui n'avaient pas un caractère unique ont été repérés avant de procéder au contrôle (couples potentiels). Plus exactement, l'identification des couples potentiels s'est faite en attribuant une note à chaque paire de personnes possible au sein du ménage et en sélectionnant les paires qui avaient obtenu la note la plus élevée. La note était fondée sur les réponses communiquées aux questions portant sur les variables démographiques et reflétait la probabilité que la paire considérée forme un couple.

31. Dans un second temps, les procédures de contrôle des couples ont été appliquées exclusivement aux personnes formant les couples potentiels (composantes). On s'est servi, dans ce but, d'une variable auxiliaire dérivée spéciale. Un couple potentiel pouvait être soit conservé, soit éliminé par imputation. Dans le premier cas, cela signifiait que les composantes considérées présentaient les valeurs requises en ce qui concerne les variables Relation avec la personne de référence du ménage, Âge, Situation matrimoniale et Années de mariage, à l'instar des composantes des couples uniques. Dans le second cas, il n'y avait, pour les composantes en question, aucune contrainte déterminant qu'il s'agissait d'un couple.

#### **D. Contrôle des variables individuelles**

32. Les variables individuelles sont corrélées uniquement au niveau de la personne grâce aux contrôles individuels, de sorte qu'elles ont été traitées à ce niveau (à l'exception de la variable Nationalité). Le processus de contrôle consistait en différentes procédures spécialement adaptées à l'unité pour laquelle les données avaient été recueillies (ménage privé, ménage institutionnel, personnes résidant temporairement dans le pays) et la caractéristique spécifique mesurée (Lieu de naissance, Nationalité, Présence et logement, Degré d'instruction et formation, Situation professionnelle et non professionnelle, Type de travail, Lieu d'étude ou de travail).

33. Une analyse des erreurs a d'abord été réalisée pour classer les types d'erreur (erreurs de compilation ou de codage systématique, erreurs aléatoires, par exemple). Ensuite, il a été procédé à l'imputation des valeurs dans les cas de non-réponse ou d'incohérence à l'aide de diverses méthodes. Plus précisément, les erreurs de codage systématique ont été traitées par la méthode d'imputation déterministe tandis que la méthode de rejet normalisée a été utilisée pour l'imputation des valeurs en cas de non-réponse ou de valeurs aberrantes dues à des erreurs aléatoires. Les valeurs imputées ont été obtenues à partir de fonctions de distribution calculées sur la base des informations qui avaient satisfait aux critères de contrôle, regroupées en strates définies également par le chemin du recensé (voir sous-section A).

#### **E. Validation des centenaires**

34. Les centenaires représentent une proportion relativement faible mais importante de l'ensemble de la population. La méthode appliquée pour valider les centenaires dénombrés lors du recensement de 2001 associe plusieurs méthodes directement appliquées aux microdonnées brutes saisies par lecture optique (Nuccitelli *et al.*, 2006). Les principales étapes clefs de cette démarche sont les suivantes:

- 1) Appariement automatique des informations se rapportant aux individus nés durant la période 1888-1901 et dénombrés au cours du recensement de 1991 avec les informations relatives aux individus nés au cours de la même période et



dénombrés lors du recensement de 2001, en se servant d'une méthode d'appariement exacte (autrement dit «déterministe»);

- 2) Contrôle automatique de la cohérence interne des informations non appariées au cours de la première étape;
- 3) Contrôle de cohérence manuel, à l'aide des images correspondantes du questionnaire saisies par lecture optique, de certains cas ambigus répertoriés au cours de la deuxième étape.

35. Au stade de l'appariement, deux informations ont été considérées comme étant mises en correspondance pour autant que les conditions ci-après soient satisfaites:

- a) Les informations enregistrées coïncidaient parfaitement, pour des champs communs aux deux ensembles de données (variables d'appariement);
- b) La clef d'appariement utilisée identifiait les liens de façon unique et ne contenait pas de valeurs manquantes.

36. Les individus (dénombrés en 2001), pour lesquels les informations enregistrées avaient été mises en correspondance, ont été considérés comme étant validés. Les individus pour lesquels les informations enregistrées n'avaient pas pu être appariées durant la première étape mais qui avaient satisfait au contrôle de cohérence automatique de la deuxième étape ont été également considérés comme validés.

### **III. PROCESSUS DE VÉRIFICATION ET DE VALIDATION À LA LUMIÈRE DES INNOVATIONS QU'IL EST ENVISAGÉ D'INTRODUIRE DANS LA CONCEPTION DU RECENSEMENT**

37. Quelques innovations dans la conception du recensement de 2011 sont actuellement à l'étude à l'ISTAT (Crescenzi et Fortini, 2007).

38. Parmi les innovations qui seront très probablement introduites, celles qui auront le plus d'impact sur les phases de vérification et de validation sont les suivantes:

- a) Recensement à l'aide de questionnaires en version abrégée et version longue;
- b) Disponibilité de données issues de registres:
  - i) Données de registres locaux se rapportant aux individus, dont le lieu de résidence habituelle est situé dans la commune, et à la manière dont ils sont organisés en ménages privés ou institutionnels (informations tirées des registres locaux de la population);
  - ii) Données de registres intégrés se rapportant aux individus, dont le lieu de résidence habituelle n'est pas situé dans la commune mais qui s'y trouvent temporairement, et à la manière dont ils sont organisés en ménages privés ou

institutionnels (informations obtenues à l'aide de données provenant de sources auxiliaires locales ou centrales);

iii) Listes locales d'adresses résidentielles;

c) Utilisation, en parallèle, de plusieurs modes de collecte de données: questionnaires rassemblés par les agents recenseurs, transmis par courriel ou accessibles via le Web, ou collecte des données sous forme d'enquête par téléphone assistée par ordinateur (CATI).

39. Lors du recensement de la population de 2011, un nombre restreint de questions (portant principalement sur des caractéristiques démographiques) pourraient être posées à chaque personne et pour une unité d'habitation donnée, tandis que d'autres questions (ayant trait aux caractéristiques socioéconomiques) ne seraient adressées qu'à un échantillon de personnes et d'unités d'habitation. Dans ce cas, on aurait besoin de deux formulaires de recensement. La version abrégée pourrait comprendre des questions portant sur les variables démographiques (Sexe, Année de naissance, Situation matrimoniale, Lien avec la personne de référence, Nationalité, Présence et logement, Degré d'instruction). En revanche, le questionnaire dans sa version longue pourrait contenir à la fois des questions portant sur les variables incluses dans la version abrégée et des questions concernant les variables suivantes: Inscription dans un établissement d'enseignement, Diplôme et formation professionnelle, Situation professionnelle, Type de travail, Lieu d'étude ou de travail et déplacements entre le domicile et le lieu d'étude ou de travail. Chaque ménage recevrait soit la version abrégée, soit la version longue du questionnaire.

40. La décision de traiter les variables démographiques (toutes incluses dans la version abrégée du questionnaire) avant les variables individuelles s'inscrit dans la ligne de la nouvelle stratégie de recensement. Toutefois, il pourrait s'avérer nécessaire de déterminer le SVA (voir sect. II.A) selon deux méthodes différentes: l'une s'appliquant à la version abrégée et l'autre à la version longue du questionnaire. Il convient de noter que, dans le cas des personnes qui recevraient la version abrégée, le SVA pourrait être moins performant, parce que la quantité de renseignements disponibles pour le calculer se trouverait alors sensiblement réduite.

41. La distribution de la version longue du questionnaire à un échantillon de recensés pourrait avoir une incidence négative sur la disponibilité (et par conséquent sur le choix) des donneurs pour les petits groupes de population. La précision des méthodes d'imputation (sur la base des donneurs et sur la base de la distribution) risque d'en être affectée dans la mesure où elle dépend beaucoup de la disponibilité d'un grand nombre de donneurs potentiels. Les taux élevés de non-réponse à une question conjugués à la rareté des donneurs pourraient nuire sensiblement à la qualité des estimations finales. Cet effet négatif pourrait être évité en gérant avec un soin particulier les phases de collecte des données et de sélection des donneurs possibles.

42. En outre, la phase de validation des variables individuelles sera rendue plus complexe du fait de la collecte par sondage. Des coefficients de pondération correcteurs devraient donc être immédiatement disponibles pour pouvoir à la fois tester et ajuster les procédures de contrôle et comparer les données finales et les macrodonnées provenant de sources extérieures ou de fichiers administratifs. À ce propos, l'efficacité de la phase de validation pourrait être renforcée si l'on avait sous la main un ensemble détaillé d'indicateurs (dispositif de contrôle) qui faciliterait l'analyse approfondie des incohérences entre les données de recensement et d'autres sources

d'information disponibles. Par conséquent, un tel dispositif de contrôle devrait être planifié avec précision et mis en œuvre comme il se doit.

43. Certaines informations (données sur le sexe et la date de naissance, par exemple) extraites des listes et registres locaux fournis par les communes (voir plus haut, points b i) à iii)) pourraient contribuer de manière appréciable à l'amélioration de la qualité des données finales. En règle générale, les informations tirées des registres sont correctes et pourraient donc être utilisées pour:

a) Améliorer le contrôle quantitatif des formulaires, c'est-à-dire vérifier la correspondance entre le nombre de questionnaires effectivement renvoyés et le nombre «escompté» de renvois, dans le but de réduire les lacunes dans le dénombrement;

b) Compléter le fichier de recensement par l'imputation des valeurs manquantes ou destinées à remplacer les valeurs aberrantes.

44. Il convient donc d'entreprendre certaines études visant à définir des stratégies efficaces en matière d'utilisation des macrodonnées et microdonnées provenant de registres.

45. Plus particulièrement, la mise en correspondance des enregistrements issus des recensements, d'une part, et des registres, d'autre part, permettrait de procéder à une imputation précise en cas de valeurs manquantes ou aberrantes. Les difficultés soulevées par la mise en correspondance des données de recensement et données de registres pourraient être dues essentiellement à l'absence de données d'identification fiables (lors des précédents recensements, on n'a pas relevé le code fiscal de chaque individu et l'ISTAT ne disposait pas du nom complet), ainsi qu'au temps de traitement nécessaire. Si le code fiscal était relevé au moyen du formulaire pour 2011, on pourrait s'en servir comme identificateur unique (variable d'appariement) pour établir les correspondances avec les données de registres. Dans cette éventualité et puisque les questionnaires seront, dans la plupart des cas, remplis par les recensés eux-mêmes, le nom complet devrait être communiqué pour pouvoir vérifier si la variable d'appariement est correcte. Sinon, des erreurs pourraient survenir au cours de la phase d'appariement de sorte qu'il serait inutile, voire dommageable, de se servir des données mises en correspondance à des fins d'imputation.

46. Une autre solution pourrait consister à ajouter les données de registres aux informations relatives aux donneurs et à adopter une méthode d'imputation s'appuyant sur les donneurs. Cette démarche aurait pour but d'améliorer le degré de similitude entre les informations concernant les receveurs et celles qui se rapportent aux donneurs (celles-ci sont parfois très différentes de certaines informations relatives aux receveurs) et, de ce fait, l'efficacité de l'imputation. Une étude spéciale devrait être entreprise pour évaluer la précision et l'efficacité des deux utilisations envisagées des données de registres (imputation fondée sur la mise en correspondance des données et imputation élargie à partir d'informations se rapportant aux donneurs).

47. Enfin, l'effet de l'emploi de plusieurs modes de collecte des données sur le contrôle quantitatif des formulaires est loin d'être négligeable. Il convient de mettre en place un processus de vérification pour s'assurer que des questionnaires ne sont pas renvoyés en double exemplaire, en raison de l'utilisation de divers modes de collecte, et ce, pour éviter le surdénombrement.

#### IV. CONCLUSIONS

48. Les processus de contrôle et de validation pour le prochain recensement de la population seront principalement définis en fonction des enseignements dégagés du recensement de 2001. Toutefois, les stratégies doivent être conçues de manière à tenir compte, comme il se doit, des innovations qui pourraient être introduites dans la conception de l'enquête. Il faudra prêter une attention particulière aux éléments ci-après:

- a) Les incidences du mode de collecte au moyen d'un questionnaire long sur la détermination du chemin du recensé, sur le choix des donneurs et sur la phase de validation;
- b) La possibilité de se servir de macrodonnées et microdonnées tirées des registres locaux de la population, ou d'autres sources intégrées;
- c) La disponibilité de listes locales d'adresses résidentielles;
- d) L'emploi, en parallèle, de plusieurs modes de collecte des données.

49. En particulier, les éléments visés sous les points b) à d) ci-dessus permettent de procéder à des contrôles au cours du processus de collecte de données. Par conséquent, la mise en place de systèmes d'appui appropriés s'impose.

50. Ces innovations auront un impact significatif sur l'ensemble du processus de contrôle et de validation. À ce stade, certaines questions ont déjà été analysées et des solutions ont été proposées, mais d'autres questions doivent encore être examinées. En outre, des études sont en cours dans le but de réduire le temps de calcul qu'exige le contrôle quantitatif et d'améliorer la détection des erreurs systématiques. Comme il est nécessaire, par ailleurs, d'accélérer la diffusion des résultats, le travail de contrôle et de validation constitue assurément une tâche ambitieuse.

## RÉFÉRENCES

- Bankier M., Houle A. et Luc M. (1997), 1996, Canadian Census Demographic Variables Imputation, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, République tchèque (Prague).
- Bankier M. (1999), Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extension for future Censuses, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Italie (Rome).
- Bianchi G., Pezone A., Reale A., Saporito G. (2004), Metodi e Procedure per il Controllo e la Correzione delle Variabili Demografiche Familiari del Censimento della Popolazione 2001, *document interne (en italien seulement)*, ISTAT.
- Bianchi G., Manzari A., Pezone A., Reale A., Saporito G. (2005), New procedures for editing and imputation of demographic variables, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Canada (Ottawa).
- Bruni R., Reale A., Torelli R. (2001), Optimization Techniques for Edit Validation and Data Imputation, communication *présentée lors du Symposium 2001 de Statistique Canada intitulé «La qualité des données d'un organisme statistique: une perspective méthodologique»*, XVIII<sup>e</sup> Symposium international sur les questions de méthodologie.
- Crescenzi F. et Fortini M. (2007), Due strategie per l'uso censuario di dati anagrafici, *document interne (en italien seulement)*, ISTAT.
- Eurostat (2008) Proposition de règlement du Parlement européen et du Conseil concernant les recensements de la population et du logement. Bruxelles, 11 janvier 2008.
- Manzari A. et Reale A. (2001), Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, dans *IASS Proceedings 53rd Session of The International Statistical Institute, August 22-29, 2001, Seoul*, p. 634 à 655.
- Manzari A., Pezone A., Reale A. (2002), Evaluation of a new approach for edit and imputation of social and demographical data with hierarchical structure, *Atti della XLI Riunione Scientifica SIS*, Milano, 5-7 Giugno 2002, Sessioni spontanee, p. 689 à 692.
- Nuccitelli A., Pezone A., Reale A., (2006), The Validation of the Census Micro-Data on the Oldest Old Living in Italy, *Proceedings of Q2006 European Conference on Quality in Survey Statistics*, Royaume-Uni (Cardiff).
- Picard C. F. (1980), *Graphs and questionnaires*, North-Holland, Pays-Bas.

-----