



Conseil économique et social

Distr. générale
18 août 2009
Français
Original: anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Douzième réunion

Genève, 28-30 octobre 2009

Point 4 de l'ordre du jour provisoire

Qualité des recensements et contrôle de la divulgation

Évaluation de la précision des hypercubes au niveau NUTS 2 avec l'adoption d'une stratégie de sondage pour le recensement italien de la population en 2011

Note de l'Institut national de la statistique, Italie

I. Introduction

1. L'Institut national italien de la statistique (Istat) envisage d'avoir recours à des techniques de sondage dans le but d'adopter une stratégie associant des formulaires longs et des formulaires abrégés pour le recensement italien de la population en 2011. Il utilise un plan probabiliste simple pour la sélection des échantillons de ménages privés choisis dans des registres de la population et les estimateurs calibrés.
2. Étant donné que l'adoption d'une stratégie de sondage entraîne l'introduction d'erreurs d'échantillonnage, des essais ont été effectués et des études menées afin d'évaluer la qualité des estimations par sondage et la précision des hypercubes utilisés pour la diffusion des données.
3. La principale contrainte affectant la définition de la stratégie de sondage est la précision des estimations obtenues à différentes échelles territoriales: plus le territoire de référence est vaste, plus grande devrait être la précision, d'une part, des estimations concernant les variables sur lesquelles porte le questionnaire long et, d'autre part, de leurs classifications croisées avec d'autres variables appartenant à la même classe ou au domaine démographique (hors échantillonnage).
4. En particulier, pour un territoire donné et un hypercube particulier, il est possible de déterminer à la fois le pourcentage de cellules pour lesquelles on pourrait estimer la fréquence absolue avec un faible degré de précision et le pourcentage de personnes recensées dans ces cellules critiques. Ce dernier indicateur exprime l'information estimée avec un faible degré de précision; des pourcentages relativement peu élevés de personnes recensées dans des cellules critiques indiquent une bonne qualité des données comprises dans cet hypercube.

5. Les évaluations de l'incidence des erreurs d'échantillonnage sur la qualité des données qui seront diffusées ont porté sur certains hypercubes d'Eurostat comprenant différentes caractéristiques et se situant au niveau 2 de la Nomenclature des unités territoriales statistiques (NUTS 2).

6. Au cours de la programmation du recensement italien de la population pour 2011, il a été tenu compte à la fois des points critiques du dernier recensement et de la possibilité d'introduire des innovations méthodologiques conformément aux recommandations internationales¹. Afin d'améliorer la qualité des opérations d'enquête, de réduire le travail à la charge des communes et d'alléger autant que faire se peut le travail statistique lié au dénombrement, de nombreuses solutions ont été envisagées. Les plus importantes font appel aux registres de population, au publipostage des questionnaires de recensement et à divers moyens combinés de collecte des données qui utilisent principalement les messageries et l'Internet.

7. Comme seuls des taux élevés de réponses spontanées détermineront le succès du prochain recensement, la stratégie de sondage qui va être adoptée prévoit l'utilisation de formulaires abrégés et de formulaires longs. Elle associera les deux versions du questionnaire. Le formulaire abrégé contribue à réduire la charge de travail tandis que l'échantillon de formulaire long préserve la multiplicité des informations fournies par le recensement. De la sorte, le sondage peut être considéré comme faisant suite aux innovations introduites au cours de la programmation du recensement.

8. L'adoption d'une stratégie de sondage pour la prochaine série de recensements permet implicitement de réaliser des économies, et de surcroît la gestion d'une plus petite quantité de données offre l'occasion d'en améliorer la qualité d'une manière générale. Il sera possible de mettre en place et de réaliser un plus grand nombre de vérifications des formulaires de recensement et des suivis sur le terrain afin de réduire les erreurs hors échantillonnage. Cette stratégie offre également un autre avantage en ce qu'elle permet de mieux respecter les délais, lesquels constituent une contrainte étant donné que les données du recensement de 2011 doivent être communiquées à Eurostat pour le 1^{er} avril 2014.

9. L'introduction d'erreurs d'échantillonnage constitue un inconvénient à prendre en compte, et c'est la raison pour laquelle il importe d'évaluer la qualité des estimations pouvant être obtenues à partir d'un sondage et la précision des hypercubes utilisés pour la diffusion des données (tableaux statistiques établis par classification croisée des variables du recensement).

II. La stratégie de sondage

10. Cette stratégie consistera à utiliser simultanément des formulaires abrégés et des formulaires longs: le formulaire abrégé ne portera que sur des variables démographiques et du logement tandis que le formulaire long comportera tout l'ensemble des variables considérées pour le recensement, y compris le degré d'instruction, la situation dans la profession et les migrations pendulaires. De la sorte, les données démographiques seront réunies pour l'ensemble de la population alors que les informations concernant les autres variables seront prélevées auprès d'un échantillon de ménages (uniquement des ménages *privés*).

¹ CEE (2006). «Recommandations pour les recensements de la population et des logements de 2010. Commission économique des Nations Unies pour l'Europe et Office statistique des Communautés européennes». Conférence des statisticiens européens. ECE/CES/STAT/NONE/2006/4.

11. La stratégie de sondage s'appliquera pour les communes de plus de 20 000 habitants; dans les communes de moins de 5 000 habitants, il est prévu d'appliquer une démarche classique en soumettant le formulaire long à l'ensemble de la population. Lorsque les communes comptent de 5 000 à 20 000 habitants, l'adoption de la stratégie dépendra de la qualité des estimations.

12. Considérant la nécessité d'adopter une stratégie très simple, des essais ont été effectués et des études menées afin d'évaluer la qualité des estimations par sondage selon différents plans de sondage et méthodes d'estimation (directes et indirectes).

13. Les résultats des études empiriques ont donné à penser qu'il conviendrait d'adopter la stratégie suivante: un sondage aléatoire simple des ménages choisis dans les registres de population et des estimateurs calibrés² employant des coefficients de pondération finals dûment modifiés dans le but de rendre l'échantillon plus représentatif. Il a également été suggéré d'utiliser chaque fois que possible, pour constituer l'échantillon, des estimations bien précises des districts de recensement qui sont des divisions infracommunales comprenant environ 15 000 habitants.

14. Un point important à considérer pour le choix du taux d'échantillonnage est le fait que les estimations sont d'autant plus précises que l'échantillon est plus grand. Il est envisagé d'utiliser un taux d'échantillonnage de 33 % dans le but de préserver autant que possible la multiplicité des informations fournies par le recensement. En cas de fortes contraintes budgétaires, il serait possible de réduire le taux d'échantillonnage en envisageant un compromis entre la nécessité de réaliser des économies sur le plan financier et la précision requise aux différentes échelles territoriales.

15. S'agissant du choix des méthodes d'estimation, des estimateurs calibrés garantissent la cohérence entre les estimations et les données démographiques réunies pour l'ensemble de la population. Il serait possible d'adopter des méthodes indirectes basées sur des techniques d'estimation pour de petites zones afin de produire des estimations plus précises concernant à la fois les plus petites échelles territoriales et les populations rares. Les premiers résultats des expérimentations³ en cours semblent être encourageants étant donné que l'on a obtenu, pour des fréquences absolues inférieures à 150 unités, une réduction de 40 à 80 % du coefficient de variation.

III. Précision des estimations

16. Une simulation a été réalisée à partir des données du recensement de la population de 2001⁴. On a considéré un ensemble de 40 communes qui avaient des effectifs de population différents et se trouvaient situées dans diverses régions de niveau NUTS 2 de l'Italie, afin de tenir compte des fortes différences entre les communes. La simulation a pris en compte un peu plus de 10 % des ménages et un peu moins de 10 % des personnes dénombrées en Italie lors du dernier recensement.

² Deville J. C., Särndal C. E. (1992) «Calibration Estimators in Survey Sampling». *Journal of the American Statistical Association*, vol. 87, p. 367 à 382.

³ Borrelli F., Carbonetti G., De Felici L., Solari F. (2008) «Metodologie di stima per piccole aree applicabili a variabili di censimento rilevabili tramite questionario long form». Vingt-neuvième Conférence italienne sur les sciences régionales, Bari (Italie), septembre 2008.

⁴ Carbonetti G., Fortini M. (2008) «Précision attendue des résultats obtenus à partir d'échantillons lors du recensement italien de la population et des logements». Réunion commune CEE/Eurostat sur les recensements de la population et des habitations, ONU, Genève (Suisse), mai 2008. ECE/CES/AC.6/2008/4.

17. On a estimé quelque 90 décomptes de cellule de tableaux de recensement à plusieurs entrées avec des méthodes de calibrage. Des décomptes ont été effectués pour chacune des 497 divisions infracommunales considérées comprenant de 5 000 à 15 000 personnes. Les propriétés des estimations établies à partir d'un sondage ont été évaluées au moyen d'un coefficient de variation statistique calculé sur 1 000 répliques pour chaque taux d'échantillonnage considéré. Des simulations selon la méthode de Monte Carlo des espaces d'échantillons ont été réalisées pour un sondage aléatoire simple des ménages avec différents taux d'échantillonnage et pour un sondage aréolaire.

18. Le tableau 1 montre un certain nombre de résultats⁵ de la simulation de sondage aléatoire simple des ménages choisis dans les registres administratifs. Il indique, pour chaque classe de décomptes de cellule et pour chaque taux d'échantillonnage considéré, les coefficients de variation moyen et maximum estimés pour les divisions infracommunales.

Tableau 1

Distribution des coefficients de variation (CV) moyen et maximum exprimés en pourcentage dans les classes de décomptes de cellule pour les trois taux d'échantillonnage considérés (sondage aléatoire simple des ménages choisis dans les registres administratifs)

| Classe de décompte | Taux d'échantillonnage = 10 % | | Taux d'échantillonnage = 20 % | | Taux d'échantillonnage = 33 % | |
|--------------------|----------------------------------|----------|----------------------------------|----------|----------------------------------|----------|
| | CV moyen % | CV max % | CV moyen % | CV max % | CV moyen % | CV max % |
| <10 | 143,3 | 191,8 | 101,4 | 123,7 | 66,5 | 95,8 |
| 10 -30 | 75,9 | 85,1 | 48,4 | 54,6 | 33,8 | 38,5 |
| 30 -50 | 51,8 | 57,1 | 31,8 | 37,1 | 23,4 | 25,6 |
| 50 -100 | 38,6 | 41,3 | 22,3 | 28,4 | 17,4 | 19,1 |
| 100 -250 | 25,4 | 28,5 | 15,7 | 19,6 | 11,4 | 12,8 |
| 250 -500 | 16,1 | 18,3 | 10,4 | 12,5 | 7,5 | 8,1 |
| 500 -1 000 | 11,8 | 12,8 | 7,5 | 8,2 | 5,3 | 5,9 |
| 1 000 -2 500 | 7,5 | 8,9 | 4,7 | 5,9 | 3,3 | 3,9 |
| 2 500 -5 000 | 4,9 | 5,4 | 3,0 | 3,6 | 2,0 | 2,5 |
| 5 000 -10 000 | 3,2 | 3,8 | 2,0 | 2,5 | 1,3 | 1,9 |

19. Par exemple, dans le cas du taux d'échantillonnage égal à 10 %, les coefficients de variation moyens tombent en dessous de 10 % lorsque les décomptes dépassent le millier; lorsque le taux d'échantillonnage passe à 33 % le seuil s'abaisse à 250. Les valeurs du coefficient de variation seront probablement plus élevées pour les petites fréquences. Toutefois, des valeurs plus élevées du coefficient de variation lorsque les décomptes sont peu élevés correspondent à de plus petites différences en valeur absolue. Comme on pouvait s'y attendre, ce sont les taux d'échantillonnage les plus élevés qui fournissent les estimations les plus précises. En général, l'amélioration de la qualité exprimée par la différence relative des coefficients de variation est d'environ 33 à 38 % lorsque le taux d'échantillonnage passe de 10 à 20 %, et d'environ 53 à 58 % lorsque ce taux passe de 10 à 33 %.

⁵ Carbonetti G., Fortini M., Solari F. (2008) «Innovations on methods and survey process for the 2011 Italian population census». Conférence européenne sur la qualité des statistiques officielles, Rome (Italie), 2008.

IV. Incidence des erreurs d'échantillonnage sur les hypercubes utilisés pour la diffusion des données

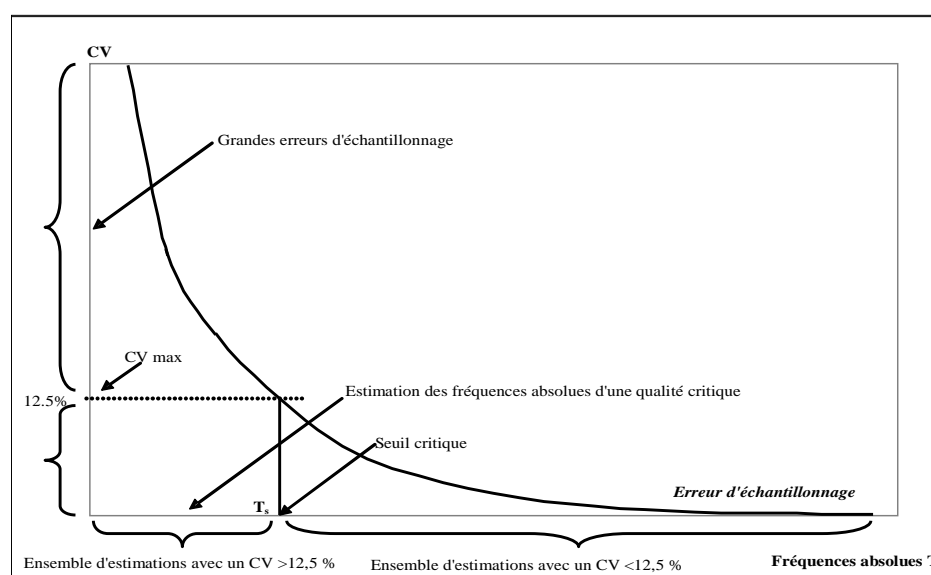
20. La principale contrainte associée à la définition de la stratégie de sondage est la précision des estimations pour différentes échelles territoriales: plus le territoire de référence est vaste et plus les estimations doivent être précises, s'agissant des variables des *formulaire longs* et de leurs classifications croisées avec d'autres variables qui appartiennent soit à la même classe soit au domaine démographique (hors échantillon). La présente section se rapporte à l'étude de l'incidence que la stratégie de sondage peut avoir sur la qualité des hypercubes utilisés pour la diffusion des données.

21. En corrélation avec une valeur acceptable du coefficient de variation, la courbe des erreurs d'échantillonnage établie à partir des résultats de la simulation⁶ permet de fixer le *seuil* des fréquences absolues de sorte que chaque fréquence qui lui est plus élevée est estimée avec un coefficient de variation plus petit que la valeur fixée.

22. Le graphique 1 montre comment il est possible de déterminer les ensembles d'estimations avec une précision critique. Pour un coefficient de variation fixé à 12,5 %, il est possible de déterminer, au moyen de la courbe des erreurs d'échantillonnage, le seuil critique en dessous duquel toutes les fréquences absolues sont estimées avec un coefficient de variation supérieur à la valeur fixée. D'une part ces fréquences peuvent être considérées comme des cas critiques dus à une grande erreur d'échantillonnage; d'autre part les fréquences supérieures au seuil correspondront à une plus petite erreur d'échantillonnage.

Graphique

Détermination des ensembles d'estimations avec une «précision critique» au moyen de la courbe des erreurs d'échantillonnage et du seuil critique T_s associé à un coefficient de variation (CV) déterminé



⁶ Carbonetti G., Dardanelli S., Fiorello E., Mastroluca S., Verrascina M. (2008) «Ipotesi di innovazione per il censimento della popolazione del 2011: una valutazione degli effetti su un possibile piano di diffusione»; vingt-neuvième Conférence italienne sur les sciences régionales, Bari (Italie), septembre 2008.

23. Si des estimations se rapportent à des domaines partiellement pris en compte, toute la population ne remplissant pas les conditions requises pour faire partie de l'échantillon, la courbe des erreurs va en diminuant ce qui favorise une amélioration de la qualité et des abaissements du seuil critique.

24. De la sorte, considérant une zone territoriale déterminée et un hypercube pour la diffusion des données, il est possible de calculer à la fois le pourcentage de cellules dans lesquelles la fréquence absolue estimée est inférieure au seuil critique observé et le pourcentage de personnes recensées dans ces cellules critiques. En particulier, ce dernier indicateur exprime la quantité de données dont on estime qu'elles présentent un faible degré de précision; des pourcentages peu élevés de personnes recensées dans des cellules critiques dénotent une bonne qualité des informations dont il est fait état dans l'hypercube en question. Par exemple, 10 % peut être considéré comme une valeur acceptable de cet indicateur.

V. Qualité escomptée des hypercubes au niveau NUTS 2

25. Les évaluations⁷ de l'incidence de la stratégie de sondage sur la qualité des données qui seront diffusées ont porté sur certains hypercubes comportant des données détaillées différentes et correspondant au niveau NUTS 2. La présente partie comprend une analyse du choix des hypercubes utilisés dans le programme de diffusion d'Eurostat, suivie de la présentation d'un certain nombre de résultats quant à la qualité escomptée.

A. Hypercubes considérés

26. On a sélectionné, dans la version préliminaire du programme de diffusion qui doit être approuvée pour 2011, huit hypercubes (tableau 2) qui contiennent des caractéristiques utilisées dans le recensement italien de 2001 mais aussi, et surtout, des caractéristiques qui étaient alors ventilées, pour ce qui est des effectifs et du contenu informationnel, comme elles le seront approximativement pour le prochain cycle de recensement, et dont les définitions sont à peu près les mêmes que celles prévues au niveau international. Par exemple, nous avons exclu les hypercubes contenant la caractéristique «résidence antérieure à l'étranger» que l'Italie utilisera pour la première fois en 2011, mais également les caractéristiques et ventilations en rapport avec les ménages et les familles, dont la décomposition est parfois différente en Italie de celle appliquée au niveau international pour le prochain recensement.

27. Pour utiliser des hypercubes comportant des caractéristiques en rapport avec les ménages, les familles et la position dans le ménage et dans la famille, il aurait fallu établir une nouvelle définition et faire un nouveau comptage des ventilations qui s'y rapportent, étant donné que les nouvelles définitions et ventilations sont, dans certains cas, différentes de celles utilisées pour le programme italien de diffusion des données, ce qui aurait demandé plus de temps pour nos études. Nous avons sélectionné les hypercubes reprenant uniquement les caractéristiques de la population.

28. Les hypercubes au niveau LAU2 ont été écartés parce qu'ils comprennent des caractéristiques (démographiques) que nous envisageons d'inclure dans le formulaire abrégé, de sorte que les informations seront disponibles pour l'ensemble de la population.

⁷ Carbonetti G. (2009) «Use of sampling strategy in the Italian population census and accuracy of estimates for different territorial domains». ITACOSM09 – Première Conférence italienne sur les méthodes d'enquête, Sienne (Italie), juin 2009.

Nous nous sommes donc concentrés sur les caractéristiques indispensables uniquement au niveau NUTS 2 et que nous envisageons d'inclure dans le formulaire long, puis sur les hypercubes correspondant au niveau géographique NUTS 1 et NUTS 2 du pays. Nous nous sommes efforcés de sélectionner des hypercubes différents les uns des autres et englobant les diverses caractéristiques proposées pour l'enquête. Un ensemble de caractéristiques (sexe et âge) est commun à tous les hypercubes, encore que l'âge soit ventilé par année dans certains hypercubes et suivant une classification plus globale dans d'autres (M ou S).

Tableau 2

Hypercubes d'Eurostat considérés dans l'étude de l'incidence de la stratégie de sondage sur les tableaux de diffusion des données

| <i>Numéro de l'hypercube</i> | <i>Code</i> | <i>Intitulé</i> | <i>Base du dénombrement</i> | <i>Caractéristiques (ventilations)</i> |
|------------------------------|-------------|--|--|--|
| 1 ^{er} hypercube | H.B1.E0.R1 | Âges, par année – «situation au regard de l'activité du moment» | Population totale | <ul style="list-style-type: none"> • Sexe (M) • Âge (L) • Situation au regard de l'activité du moment (M) |
| 2 ^e hypercube | H.B1.E0.R2 | Âges, par année – «profession» | Population totale | <ul style="list-style-type: none"> • Sexe (M) • Âge (L) • Profession (M) |
| 3 ^e hypercube | H.B1.E0.R3 | Âges, par année – «branche d'activité économique» | Population totale | <ul style="list-style-type: none"> • Sexe (M) • Âge (L) • Branche d'activité économique (M) |
| 4 ^e hypercube | H.B1.E0.R4 | Âges, par année – «situation dans la profession» | Personnes ayant un emploi | <ul style="list-style-type: none"> • Sexe (M) • Âge (L) • Situation dans la profession (M) |
| 5 ^e hypercube | H.B1.E0.R5 | Âges, par année – «niveau d'instruction» | Population totale | <ul style="list-style-type: none"> • Sexe (M) • Âge (L) • Niveau d'instruction (niveau d'études le plus élevé) (M) |
| 14 ^e hypercube | H.B1.E1.R2 | Emploi (au lieu de résidence) – «profession» | Population totale à son lieu de résidence habituelle | <ul style="list-style-type: none"> • Sexe (M) • Âge (M) • Situation au regard de l'activité du moment (L) • Profession (M) • Niveau d'instruction (niveau d'études le plus élevé) (M) |

| <i>Numéro de l'hypercube</i> | <i>Code</i> | <i>Intitulé</i> | <i>Base du dénombrement</i> | <i>Caractéristiques (ventilations)</i> |
|------------------------------|-------------|--|--|---|
| 15 ^e hypercube | H.B1.E1.R3 | Emploi (au lieu de résidence) – «branche d'activité économique» | Population totale à son lieu de résidence habituelle | <ul style="list-style-type: none"> • Sexe (M) • Âge (M) • Situation au regard de l'activité du moment (M) • Branche d'activité économique (M) • Niveau d'instruction (niveau d'études le plus élevé) (M) |
| 16 ^e hypercube | H.B1.E1.R4 | Emploi (au lieu de résidence) – «profession» par «branche d'activité économique» | Population totale à son lieu de résidence habituelle | <ul style="list-style-type: none"> • Sexe (M) • Âge (S) • Profession (M) • Branche d'activité économique (M) • Niveau d'instruction (niveau d'études le plus élevé) (M) |

29. On a initialement calculé pour chaque hypercube un nombre de cellules qui correspond au produit du nombre de catégories par caractéristique reprise dans l'hypercube. On a alors introduit une nouvelle dimension des hypercubes, la «dimension adéquate», que l'on a calculée en excluant dès l'abord le nombre de catégories correspondant aux totaux, aux sous-totaux et aux catégories «pas de réponse».

30. On a calculé (tableau 3) la «dimension adéquate» des hypercubes sélectionnés par multiplication du nombre de catégories utilisées dans le programme italien de diffusion des données pour le recensement de 2001 et obtenu un certain nombre de cellules potentielles. On a alors exclu les «nuls structurels», c'est-à-dire les cases qui sont certainement vides parce qu'elles correspondent à des résultats impossibles. On a ainsi obtenu le nombre de cellules acceptables en excluant les cas où les croisements sont impossibles (par exemple Âge = 15 ans et Niveau d'instruction = CITE, niveau 5a).

Tableau 3

Nombre de cellules potentielles et de cellules acceptables pour chaque hypercube considéré

| <i>Code</i> | <i>Nombre de cellules potentielles</i> | <i>Nombre de cellules acceptables</i> |
|-------------|--|---------------------------------------|
| H.B1.E0.R1 | 1 212 (2x101x6) | 1 062 |
| H.B1.E0.R2 | 2 020 (2x101x10) | 1 922 |
| H.B1.E0.R3 | 3 434 (2x101x17) | 3 126 |
| H.B1.E0.R4 | 1 212 (2x101x6) | 1 032 |
| H.B1.E0.R5 | 1 414 (2x101x7) | 1 342 |
| H.B1.E1.R2 | 23 520 (2x21x8x10x7) | 3 810 |
| H.B1.E1.R3 | 29 988 (2x21x6x17x7) | 5 574 |
| H.B1.E1.R4 | 30 940 (2x13x10x17x7) | 26 350 |

31. Les cinq premiers hypercubes montrent qu'il est tout à fait possible d'associer des caractéristiques démographiques avec une seule caractéristique socioéconomique en utilisant un nombre limité de cellules potentielles. Ce nombre diminue encore lorsque l'on ne tient compte que des cellules acceptables, c'est-à-dire celles pour lesquelles on peut penser qu'il existe une fréquence. Par contre, pour la deuxième série d'hypercubes, même si l'âge est moins ventilé (il ne s'agit plus d'une année après l'autre mais de tranches de 5 ou 10 ans), le nombre de croisements augmente, de même que celui des caractéristiques socioéconomiques qui entrent en jeu. De ce fait, la dimension potentielle des hypercubes augmente beaucoup, même si elle diminue fortement lorsque l'on considère les cellules acceptables. L'hypercube qui a la plus grande dimension est le dernier qui associe, abstraction faite du sexe et de l'âge (tranches de 10 ans), la profession, la branche d'activité économique et le niveau d'instruction.

B. Quelques résultats

32. On trouvera ci-après un certain nombre de résultats de l'analyse qualitative concernant les hypercubes les plus complexes parmi ceux considérés dans la présente étude et reprenant des données fournies par le recensement italien de la population de 2001.

33. Le premier exemple fait référence à l'hypercube H.B1.E1.R3 qui associe le sexe, l'âge, la situation au regard de l'activité du moment, la branche d'activité économique et le niveau d'instruction. Les *cellules acceptables* sont au nombre de 5 574 (aucun nul structurel n'est pris en compte). S'agissant de cet hypercube, le tableau 4 montre, pour chaque coefficient d'échantillonnage considéré (en adoptant le sondage aléatoire simple des ménages choisis dans les registres de population) et pour trois zones italiennes au niveau NUTS 2 et de taille différente (*Molise* est l'une des plus petites, *Marche* correspond à une moyenne et *Sicilia* est l'une des plus grandes), les seuils critiques, les pourcentages de cellules critiques et les pourcentages d'individus dans les cellules critiques. On constate aisément qu'avec l'amélioration de la qualité le coefficient d'échantillonnage augmente et que les zones les plus vastes présentent la meilleure qualité.

34. Par exemple, si l'on considère *Sicilia* avec un coefficient d'échantillonnage de 33 %, le seuil critique est alors de 100, 59,4 % des cellules présentent des fréquences absolues inférieures à 100 mais ces cellules comptent 1,0 % seulement d'individus remplissant les conditions requises.

Tableau 4

Hypercube H.B1.E1.R3. Indicateurs de qualité (CV = 12,5 %) pour trois zones de niveau NUTS 2 en Italie: Molise, Marche et Sicilia

| Taux d'échantillonnage | Molise | | | Marche | | | Sicilia | | |
|------------------------|----------------|------------------------|--------------------|----------------|------------------------|--------------------|----------------|------------------------|--------------------|
| | Seuil critique | % d'individus dans les | | Seuil critique | % d'individus dans les | | Seuil critique | % d'individus dans les | |
| | | de cellules critiques | cellules critiques | | de cellules critiques | cellules critiques | | de cellules critiques | cellules critiques |
| 10 % | 100 | 79,2 | 10,7 | 250 | 78,8 | 6,9 | 500 | 75,9 | 4,2 |
| 20 % | 50 | 71,0 | 5,8 | 100 | 68,4 | 3,0 | 250 | 68,7 | 2,1 |
| 33 % | 30 | 63,6 | 3,4 | 50 | 59,8 | 1,5 | 100 | 59,4 | 1,0 |

35. Le tableau 5 montre les résultats en rapport avec un hypercube plus complexe (H.B1.E1.R4) qui associe le sexe, l'âge, la profession, la branche d'activité économique et le niveau d'instruction. Les *cellules acceptables* sont au nombre de 26 350.

Tableau 5
Hypercube H.B1.E1.R4. Indicateurs de qualité (CV = 12,5 %) pour trois zones de niveau NUTS 2 en Italie: Molise, Marche et Sicilia

| Taux d'échantillonnage | Molise | | | Marche | | | Sicilia | | |
|------------------------|----------------|---|--------------------|----------------|---|--------------------|----------------|---|--------------------|
| | Seuil critique | % d'individus dans les cellules critiques | | Seuil critique | % d'individus dans les cellules critiques | | Seuil critique | % d'individus dans les cellules critiques | |
| | | de cellules critiques | cellules critiques | | de cellules critiques | cellules critiques | | de cellules critiques | cellules critiques |
| 10 % | 100 | 91,9 | 14,9 | 250 | 91,1 | 11,2 | 500 | 91,8 | 7,3 |
| 20 % | 50 | 86,5 | 9,3 | 100 | 84,4 | 6,0 | 250 | 87,6 | 4,5 |
| 33 % | 30 | 81,3 | 6,5 | 50 | 77,1 | 3,4 | 100 | 79,4 | 2,2 |

36. Les observations sont analogues à celles exprimées au sujet du cas précédent. La différence tient à une moindre qualité étant donné que cet hypercube compte un plus grand nombre de cellules. Toutefois, la précision diminue peu et la qualité globale du tableau statistique demeure acceptable.

37. Le tableau 6 présente succinctement les résultats concernant le pourcentage d'individus dans des cellules critiques, pour l'ensemble des 20 zones italiennes au niveau NUTS 2, pour tous les hypercubes étudiés et pour les trois taux d'échantillonnage considérés.

Tableau 6
Distribution des zones italiennes au niveau NUTS 2 selon le pourcentage d'individus recensés dans les cellules critiques (CV = 12,5 %) pour certains hypercubes d'Eurostat

| Nombre de zones au niveau NUTS 2 | Taux d'échantillonnage = 33 % | | Taux d'échantillonnage = 20 % | | | Taux d'échantillonnage = 10 % | | | | |
|--|---|--------|-------------------------------------|--------|---------|-------------------------------------|--------|---------|---------|-------|
| | | | | | | | | | | |
| Hypercubes d'Eurostat (cellules acceptables) | Pourcentage d'individus dans les cellules critiques | | | | | | | | | |
| | <55 % | 5-10 % | <5 % | 5-10 % | 10-15 % | <5 % | 5-10 % | 10-15 % | 15-20 % | >20 % |
| H.B1.E0.R1 (1 062) | 20 | 0 | 20 | 0 | 0 | 19 | 1 | 0 | 0 | 0 |
| H.B1.E0.R2 (1 922) | 20 | 0 | 20 | 0 | 0 | 15 | 4 | 1 | 0 | 0 |
| H.B1.E0.R3 (3 126) | 20 | 0 | 17 | 3 | 0 | 11 | 4 | 4 | 1 | 0 |
| H.B1.E0.R4 (1 032) | 20 | 0 | 16 | 4 | 0 | 7 | 8 | 5 | 0 | 0 |
| H.B1.E0.R5 (1 342) | 20 | 0 | 20 | 0 | 0 | 15 | 5 | 0 | 0 | 0 |
| H.B1.E1.R2 (3 810) | 20 | 0 | 18 | 2 | 0 | 12 | 6 | 2 | 0 | 0 |
| H.B1.E1.R3 (5 574) | 20 | 0 | 17 | 3 | 0 | 10 | 5 | 5 | 0 | 0 |
| H.B1.E1.R4 (26 350) | 15 | 5 | 8 | 10 | 2 | 1 | 9 | 6 | 3 | 1 |

38. On observe que, pour un taux d'échantillonnage égal à 33 % dans 7 des 8 hypercubes considérés, le pourcentage d'individus recensés dans des cellules critiques est inférieur à 5 % dans toutes les zones; par contre, dans le cas du plus grand hypercube H.B1.E1.R4, cet indicateur est inférieur à 5 % dans 15 zones et se situe entre 5 et 10 % dans les 5 zones restantes.

39. Un taux d'échantillonnage de 20 % entraîne automatiquement une légère diminution de la qualité; en fait, les pourcentages d'individus recensés dans des cellules critiques correspondant aux différentes zones demeurent inférieurs à 10 % sauf pour le dernier hypercube, auquel cas ils se situent entre 10 et 15 % dans deux zones (les plus petites).

40. La qualité pourrait être acceptable si l'on adopte une stratégie de sondage prévoyant un taux d'échantillonnage de 10 %. On peut alors observer, pour un grand nombre de zones, des indicateurs de qualité inférieurs à 10 % pour les différents hypercubes; un certain nombre de problèmes pourraient apparaître dans le cas d'hypercubes plus complexes et de plus petites zones.

VI. Conclusions

41. Les résultats obtenus militent en faveur de l'introduction de techniques de sondage pour le recensement italien de la population en 2011. La stratégie consistant à adopter un sondage aléatoire simple des ménages choisis dans des registres de population et des estimateurs calibrés aboutit à des estimations précises et permet de réduire les erreurs indépendantes de l'échantillonnage. Les estimateurs de petites zones peuvent améliorer la précision des estimations en rapport avec des domaines de petite taille ou des populations rares.

42. Les définitions et la classification prévues pour le recensement italien sont tout à fait conformes aux prescriptions d'Eurostat de sorte qu'il est possible de produire tous les hypercubes indispensables. Néanmoins, l'adoption d'une stratégie de sondage s'accompagnera d'une certaine variabilité dans une partie des tableaux établis à partir du recensement. Plus précisément, des formulaires abrégés suffiront pour tous les hypercubes indispensables au niveau LAU 2 de sorte qu'il n'y aura, dans ce cas, aucune variabilité d'échantillonnage; par contre, les hypercubes correspondant aux niveaux NUTS 1 et NUTS 2 du pays, qui comprennent des caractéristiques pour lesquelles l'échantillonnage sera réalisé au moyen de formulaires longs, seront assortis d'une variabilité d'échantillonnage et de ce fait accompagnés d'estimations de la variance d'échantillonnage.

43. Les expérimentations réalisées pour évaluer l'incidence de la stratégie de sondage sur la qualité de certains hypercubes correspondant au niveau 2 de la NUTS ont mis en lumière l'utilité de la nouvelle démarche adoptée pour le recensement italien de la population. La principale conclusion qui en a été tirée est qu'il est possible, avec des hypercubes comprenant plus de 20 000 cellules, d'établir des estimations avec un pourcentage peu élevé d'unités dans les cellules critiques, même lorsque le coefficient d'échantillonnage est moins élevé. On observe certains problèmes uniquement pour les plus petites zones (par exemple Val d'Aoste).

44. Puisqu'il faut parvenir à concilier la dimension de l'échantillon et les coûts de la collecte des données, le choix final dépendra de l'équilibre que l'on trouvera entre les considérations budgétaires et le degré de précision minimum requis aux diverses échelles territoriales.

45. Un autre point qui s'y rapporte concerne la possibilité de produire des données plus diversifiées à une échelle territoriale moins élevée. Différentes démarches envisagent l'utilisation de *ventilations* associées à un moindre nombre de catégories ou bien l'adoption d'autres indicateurs indirects. En fait, les méthodes consistant à utiliser de petites zones améliorent, semble-t-il, la précision des estimations pour les plus petites échelles territoriales car l'échantillon pourrait alors ne pas être représentatif et pour les très petits décomptes de cellule dans les domaines plus vastes.