

Un modèle hiérarchique pour l'analyse du sous-dénombrement local du recensement en Italie

D. Cocchi, E. Fabrizi et C. Trivisano ¹

Résumé

La comparaison des résultats des recensements et des enquêtes postcensitaires (EP) montre que les chiffres de recensement sont inexacts. En Italie, les administrations municipales jouent un rôle essentiel dans les opérations sur le terrain du recensement et de l'EPR. Dans le présent article, nous analysons l'effet des municipalités sur le taux de sous-dénombrement au recensement en Italie par modélisation des données provenant de l'EPR et d'autres sources à l'aide d'arbres de régression de Poisson et de modèles de Poisson hiérarchiques. Les arbres de régression de Poisson permettent de former des groupes homogènes de municipalités. Les modèles de Poisson hiérarchiques peuvent être considérés comme des outils pour l'estimation pour des petits domaines.

Mots clés : Sous-dénombrement au recensement; enquête postcensitaire (EP); modélisation bayésienne hiérarchique; modèles de régression Gamma-Poisson; arbres de régression de Poisson.

1. Introduction

Le recensement de la population de l'Italie, qui a lieu tous les dix ans, représente la tâche la plus importante de l'Institut national de statistique de l'Italie (ISTAT). (Les travaux qui ont abouti au présent article ont été réalisés juste avant le Recensement de l'Italie de 2001 et l'EPR subséquente. On a tenu compte des résultats obtenus pour la réalisation de l'EPR de 2001). Pour procéder au recensement, ISTAT s'appuie sur les administrations municipales qui sont chargées de toutes les opérations sur le terrain (formation des intervieweurs, planification des interviews, collecte et traitement de base des données). Durant les opérations du recensement, les municipalités travaillent indépendamment les unes des autres sous la supervision d'ISTAT. Par conséquent, l'exactitude des résultats varie considérablement d'une municipalité à l'autre, même si elles sont contiguës. En Italie, le territoire d'un bourg municipal est subdivisé en plusieurs secteurs de dénombrement (SD), qui sont affectés à un même intervieweur durant les opérations du recensement. Les SD diffèrent par leur forme, leur structure et la difficulté de dénombrement, ainsi que par l'intervieweur. Aussi est-il probable que le taux de sous-dénombrement varie fortement d'un SD à l'autre dans une même municipalité.

Après le Recensement de la population de 1991, ISTAT a réalisé une Enquête postcensitaire (EP) pour évaluer le phénomène du sous-dénombrement. Il est bien connu que les chiffres de population du recensement sont généralement incorrects parce que des personnes ne sont pas dénombrées, sont dénombrées plusieurs fois ou sont dénombrées au mauvais endroit. Les personnes non dénombrées sont la source la plus importante d'inexactitude et, habituellement, produisent un sous-dénombrement net qui peut varier selon

la région géographique ou selon le groupe social, et influencer la détermination de la taille relative des sous-populations (Abbate, Masselli et Signore 1993). Les opérations sur le terrain de l'EPR de 1991 ont été réalisées par les municipalités échantillonnées proprement dites. Les données recueillies ont été analysées par Abbate, Masselli et Signore (1993), qui ont estimé le taux national global de sous-dénombrement au moyen d'un modèle de Lincoln-Petersen (voir Wolter 1986) en utilisant des strates a posteriori de municipalités fondées sur de grandes régions géographiques (Nord, Centre, Sud). À partir des mêmes données, Fortini (1994) a estimé le sous-dénombrement national global au moyen de modèles de classes latentes.

Au lieu d'estimer le taux de sous-dénombrement pour l'ensemble du pays ou pour des domaines plus petits, nous proposons des modèles conçus pour expliquer la variation du taux de sous-dénombrement au niveau municipal. La découverte de facteurs expliquant la grandeur du sous-dénombrement net pourrait servir de base à la création de groupes homogènes de municipalités qui permettront de mieux planifier la stratification lors de futures enquêtes postérieures au recensement. En outre, le dépistage des failles dans la structure organisationnelle qui influe de façon significative sur le sous-dénombrement pourrait fournir des indications quant aux mesures à prendre pour réduire l'importance de ce dernier.

On trouve dans la littérature des études fondées sur des données d'EPR désagrégées. Alho, Mulry, Wurdeman et Kim (1993) considèrent un modèle de régression logistique pour la probabilité individuelle (ménage) d'être dénombré. En nous inspirant de Moura et Holt (1999), nous pourrions étendre le modèle afin d'y inclure les effets de municipalité ou d'autres groupes. Nous sommes parfaitement conscients que notre décision de modéliser des données municipales n'équivaut pas à analyser les enregistrements au niveau des

1. D. Cocchi, E. Fabrizi et C. Trivisano, Dipartimento di Scienze Statistiche « P. Fortunati », Università di Bologna, Italie.

ménages, puisque nombre de caractéristiques déterminant la propension individuelle à être dénombré lors du recensement s'égalisent lorsqu'on analyse des données agrégées. Dans le cas de l'Italie, une analyse complète, fondée sur des enregistrements individuels, est impossible puisque le questionnaire de l'EPR de 1991 ne contenait qu'un très petit nombre de questions s'adressant à des individus. Par ailleurs, l'EPR de 1991 fournit fort peu de données auxiliaires sur les SD, de sorte qu'on ne peut proposer des modèles fondés sur le sous-dénombrement au niveau des SD.

Notre analyse se fonde sur la combinaison de différentes sources de données. Les données auxiliaires proviennent de l'EPR de 1991 susmentionnée, de deux études sur la qualité des statistiques des municipalités réalisées par ISTAT au cours des années 1990 (Di Pietro 1998, 1999) et sur des indicateurs démographiques et sociaux tirés des résultats officiels du Recensement de 1991.

Le problème est de savoir comment utiliser efficacement l'information provenant des diverses sources de données. Nous disposons en fait d'un grand nombre de variables, dont la plupart sont nominales ou polychotomiques. Au lieu d'utiliser un algorithme de sélection de variables, nous avons décidé de former des groupes homogènes de municipalités, puis de les introduire dans le modèle à l'aide d'une matrice de plan d'expérience pour les effets aléatoires. Nous utilisons, pour construire ces groupes, des arbres de régression de Poisson (Therneau et Atkinson 1997). Cette utilisation hiérarchique de l'information constitue une base naturelle pour la formation de strates de municipalités géographiquement non contiguës.

Durant l'EPR, peu de SD sont de nouveau dénombrés à l'intérieur de chaque municipalité échantillonnée; le taux moyen d'échantillonnage des SD est de 0,001. Il s'agit d'un contexte typique d'estimation sur petits domaines où les estimations directes du taux municipal de sous-dénombrement ne sont pas fiables et doivent être remplacées par des estimations synthétiques ou composites fondées sur un modèle approprié. Le phénomène du sous-dénombrement est rare. Nos données sont constituées de dénombrements et peuvent présenter une surdispersion importante par rapport à l'hypothèse d'une loi de Poisson. Nous proposons l'utilisation de modèles de régression hiérarchiques de Poisson pour tenir compte de la surdispersion.

Les modèles hiérarchiques adoptés ici permettent de traiter explicitement la surdispersion à cause de l'hétérogénéité municipale. Une autre source de variabilité extrapoissonienne est due à l'hétérogénéité à l'intérieur des municipalités, à cause de l'existence de grappes de personnes non dénombrées dans les SD ou des grappes dues aux personnes non dénombrées dans une même famille. Cette forme de surdispersion n'est pas traitée explicitement dans les modèles.

Nous adoptons une approche entièrement bayésienne pour la spécification et l'estimation, et nous résolvons les

modèles par les méthodes de simulation de Monte Carlo par chaînes de Markov. Dans ce cadre hiérarchique, nous traitons la surdispersion en imposant une loi Gamma pour le taux du premier niveau de la loi de Poisson, donc en obtenant marginalement une loi binomiale négative. En outre, conditionnellement aux hyperparamètres, le modèle proposé présente une linéarité a posteriori et les moyennes a posteriori correspondantes des taux de sous-dénombrement municipaux sont des estimateurs linéaires composites. Donc, le degré de lissage dépend de la quantité d'information fournie par chaque échantillon municipal dans l'EPR.

Nos résultats montrent qu'on pourrait améliorer la stratification des municipalités sur laquelle repose le plan de sondage de l'EPR de 1991 (fondée sur la région géographique et la taille de population), puisque le taux de sous-dénombrement est en grande partie indépendant de la région géographique. Par contre, les variables décrivant l'efficacité statistique des administrations locales aident à faire la distinction entre les divers degrés de sous-dénombrement parmi les municipalités de mêmes taille et structure démographique. Si l'on ne modifie pas le plan de sondage de l'EPR, nos résultats pourraient fournir des indications utiles lors de l'analyse des données.

Le présent article est présenté comme suit. À la section 2, nous décrivons les caractéristiques générales de l'EPR et des autres sources de données utilisées. À la section 3, nous examinons les arbres de régression de Poisson employés pour créer des groupes homogènes de municipalités. À la section 4, nous introduisons les modèles de régression de Poisson hiérarchiques, tandis qu'à la section 5, nous discutons des résultats empiriques et comparons les modèles.

2. Données de L'EPR et information auxiliaire

2.1 Enquête postcensitaire (EP) de l'Italie

Le Recensement de la population de l'Italie de 1991 a eu lieu le 20 octobre. L'Enquête postcensitaire (EP), fondée sur un plan d'échantillonnage stratifié à deux degrés, a été réalisée quelques semaines plus tard. Les municipalités représentent les unités primaires d'échantillonnage et les secteurs de dénombrement, les unités secondaires. Un SD est le plus petit domaine en lequel le territoire municipal est divisé aux fins des opérations du recensement; chaque SD est affecté entièrement à un même intervieweur.

Les unités primaires d'échantillonnage ont été stratifiées en fonction de la région géographique (Nord-Ouest, Nord-Est, Centre, Sud, îles) et de la taille de la population (sept classes pour les municipalités comptant moins de 350 000 habitants), ce qui a produit 35 strates. Dans chaque strate, les municipalités échantillonnées ont été sélectionnées sans remise et avec probabilité proportionnelle à la taille de la population. Les 10 municipalités comptant plus de 350 000 habitants ont été incluses dans l'échantillon en tant qu'unités auto-représentatives. Les unités secondaires

d'échantillonnage ont été sélectionnées avec probabilités égales par échantillonnage systématique. L'échantillon final de l'EPR contenait 85 municipalités et 638 SD (sur un total national de 8 095 municipalités et 64 000 SD) avec une estimation fondée sur un plan d'échantillonnage national de 1,24 % (Abbate, Masselli et Signore 1993).

Le questionnaire de l'EPR, qui ne contient que quelques questions simples, a été rempli au cours d'une interview sur place. Les caractéristiques des ménages échantillonnés sont limitées au nombre et au sexe des membres du ménage. D'autres questions de l'EPR ont été conçues pour faciliter le couplage des enregistrements aux données du recensement, donc, pour réduire le nombre de cas de dénombrement au mauvais endroit et d'autres erreurs non dues à l'échantillonnage lors de l'évaluation du sous-dénombrement (voir Fortini 1994, pour plus de précisions).

2.2 Enquêtes sur la qualité des statistiques des municipalités

ISTAT a créé un ensemble de données sur la qualité des statistiques des municipalités italiennes (voir Di Pietro 1998, 1999). Cet ensemble intègre différentes sources, dont l'information provenant des enregistrements sur le rendement lors du Recensement de 1991, des registres de population municipaux et de données du ministère de l'Intérieur. Cet ensemble de données contient aussi les résultats des trois enquêtes administratives réalisées durant les années 1990 en vue d'évaluer le rendement des municipalités en ce qui concerne leurs engagements à l'égard d'ISTAT. La première enquête porte sur l'informatisation des bureaux municipaux de la statistique. La deuxième, connue sous l'acronyme POSAS, est une enquête postcensitaire (EP) fondée sur les registres de la population de résidents, classés selon l'année de naissance, l'âge et l'état civil. La troisième, connue sous l'acronyme ISCAN, vise à déterminer la mesure dans laquelle les enregistrements figurant sur les listes des registres municipaux de population sont appropriés. Ces enquêtes fournissent des données sur toutes les municipalités italiennes.

À partir de cet ensemble de données, nous avons sélectionné un sous-ensemble de variables reliées à l'activité municipale au moment du Recensement de 1991, à savoir :

- le pourcentage de champs non codés du questionnaire du recensement à l'intention des ménages qui auraient dû être remplis, après l'interview des ménages, par les bureaux municipaux de la statistique (PERCOD);
- le ratio de la population provisoirement à l'étranger à la population présente au moment du Recensement de 1991 (PERCEST);
- le ratio de la différence entre les dénombrements du Recensement de 1991 et ceux des registres de population aux dénombrements du Recensement de 1991 (PERDIFF);

- le temps nécessaire pour mettre à jour les registres de population municipaux d'après les résultats du Recensement de 1991 (IND01);
- le retard de la mise à jour des noms de rue (IND11).

2.3 Variables démographiques

Nous considérons aussi l'ensemble de ratios démographiques établis d'après les résultats du Recensement de 1991. En particulier, nous utilisons les pourcentages de ménages « ne comptant qu'un seul membre » et « comptant plus d'une famille », ainsi que les ratios de masculinité (ratios hommes-femmes) dans la municipalité. La population municipale de résidents – correspondant aux chiffres non corrigés du Recensement de 1991 – est aussi une variable fort importante. Le nombre de SD échantillonnés dans chaque municipalité pour l'EPR est un autre indicateur de l'importance de la municipalité.

3. Arbres de régression de poisson

Les sources de données disponibles nous fournissent un grand nombre de variables auxiliaires, dont beaucoup sont nominales ou polychotomiques. Avant d'ajuster les modèles hiérarchiques, nous regroupons les municipalités dont le taux de sous-dénombrement des ménages est homogène à l'aide d'arbres de régression binaires de Poisson. Les groupes établis d'après ces arbres sont inclus à titre de facteurs dans les modèles décrits à la section suivante. Notre objectif principal est de vérifier l'efficacité des méthodes habituelles de stratification, de les améliorer *ex post* à l'aide de modèles hiérarchiques contenant des covariables appropriées et de comparer les résultats ainsi obtenus à ceux de méthodes semblables fondées sur des groupements optimaux.

Les modèles de régression conditionnelle sont fondés sur le lien logarithmique canonique. Le critère de partition est la statistique habituelle de somme des carrés des écarts, ou *deviance* en anglais (Therneau et Atkinson 1997) :

$$\text{Deviance}_{\text{parent}} - (\text{Deviance}_{\text{enfant, gauche}} + \text{Deviance}_{\text{enfant, droite}})$$

L'idée fondamentale de la construction d'un arbre consiste à partir d'un grand arbre T_0 construit en appliquant une règle d'arrêt naïve et faible (comme le nombre minimal d'observations dans les nœuds finaux de l'arbre), puis à sélectionner par élagage l'arbre de taille appropriée parmi les sous-arbres de T_0 . La méthode établie pour élaguer les arbres est celle du coût-complexité, introduite pour la première fois par Breiman, Friedman, Olshen et Stone (1984). Soit D_T la somme des carrés des écarts d'un sous-arbre T de T_0 , $\text{size}(T)$ le nombre de nœuds terminaux de T et $\alpha > 0$ un paramètre de coût-complexité pour définir la mesure de coût-complexité :

$$D_T(\alpha) = D_T + \alpha \text{size}(T) \quad (1)$$

Pour un α spécifié, on peut trouver l'arbre $T(\alpha)$ qui minimise (1). On peut montrer (Breiman et coll. 1984) qu'il existe une famille emboîtée de sous-arbres $\{T_0, T_1, \dots, T_k, \dots, T_{\text{root}}\}$ de T_0 telle que chaque arbre est optimal pour une fourchette de valeurs de α .

Le problème se réduit maintenant à choisir l'un de ces sous-arbres. La sélection se fait de façon à minimiser l'erreur de prédiction définie comme étant la contribution d'une nouvelle observation à la somme des carrés des écarts. Pour estimer l'erreur de prédiction, l'existence d'un échantillon indépendant serait, théoriquement, la meilleure option, mais puisqu'il est préférable d'utiliser toutes les données afin d'« informer » l'arbre le mieux possible, on utilise une méthode de contre-vérification. Habituellement, on choisit l'arbre T_{k_0} dont l'erreur de prédiction estimée est minimale. Ici, nous utilisons une règle d'élagage plus stricte qui consiste à sélectionner le plus petit arbre ayant une erreur estimée de prédiction n'excédant pas la somme de l'erreur estimée de prédiction de T_{k_0} et de son erreur-type. Nous adoptons cette règle d'élagage, appelée « règle d'une erreur-type » (Breiman et coll. 1984), pour éviter de surajuster le modèle.

Puisque la contre-vérification des arbres de régression de Poisson peut donner, pour certains nœuds, une valeur infinie de la somme des carrés des écarts, nous utilisons des estimateurs bayésien de rétrécissement des taux réels fondés sur un simple modèle Poisson-Gamma, comme l'ont proposé Thernau et Atkinson (1997).

Nous construisons trois arbres distincts en partant de sous-ensembles différents de variables auxiliaires.

L'arbre 1 (présenté à la figure 1) s'appuie sur les variables démographiques uniquement. La première partition

sépare les municipalités de moins de 100 100 habitants de celles de plus de 100 100 habitants. Cette valeur de partition coïncide presque avec le seuil de démarcation de 100 000 utilisé pour la stratification des municipalités lors de l'EPR de 1991. La deuxième partition isole un sous-échantillon de petites municipalités pour lesquelles moins de quatre SD ont été échantillonnés pour l'EPR. Une partition supplémentaire est faite d'après le rapport de masculinité.

L'arbre 2 (figure 2) est fondé exclusivement sur des variables concernant la qualité des statistiques des municipalités. La première partition s'appuie sur le temps mis pour corriger les registres de population (IND01) : les municipalités qui procèdent le plus rapidement à cette tâche affichent les taux de sous-dénombrement les plus faibles. Les partitions de niveau inférieur mettent en relief le problème des personnes résidant provisoirement à l'étranger (PERCEST), problème qui, dans les régions caractérisées par une émigration massive, peut donner lieu à un sous-dénombrement grave de la population municipale et à des erreurs de tenue à jour des registres de population (PERDIFF). Dans cet arbre, une moitié de l'échantillon est classée dans un seul noeud qui contient vraisemblablement l'hétérogénéité résiduelle.

L'arbre 3 (figure 3) est fondé sur des variables à la fois démographiques et de qualité. La première partition est basée sur la population municipale, exactement de la même façon que pour l'arbre 1. Subséquemment, le sous-ensemble de municipalités comptant moins de 100 100 habitants est partitionné en un groupe de municipalités de petite taille et un groupe de municipalités de taille moyenne au seuil de 13 200 habitants. La variable de qualité incluse dans cet arbre est le temps mis pour corriger les registres de population (IND01).

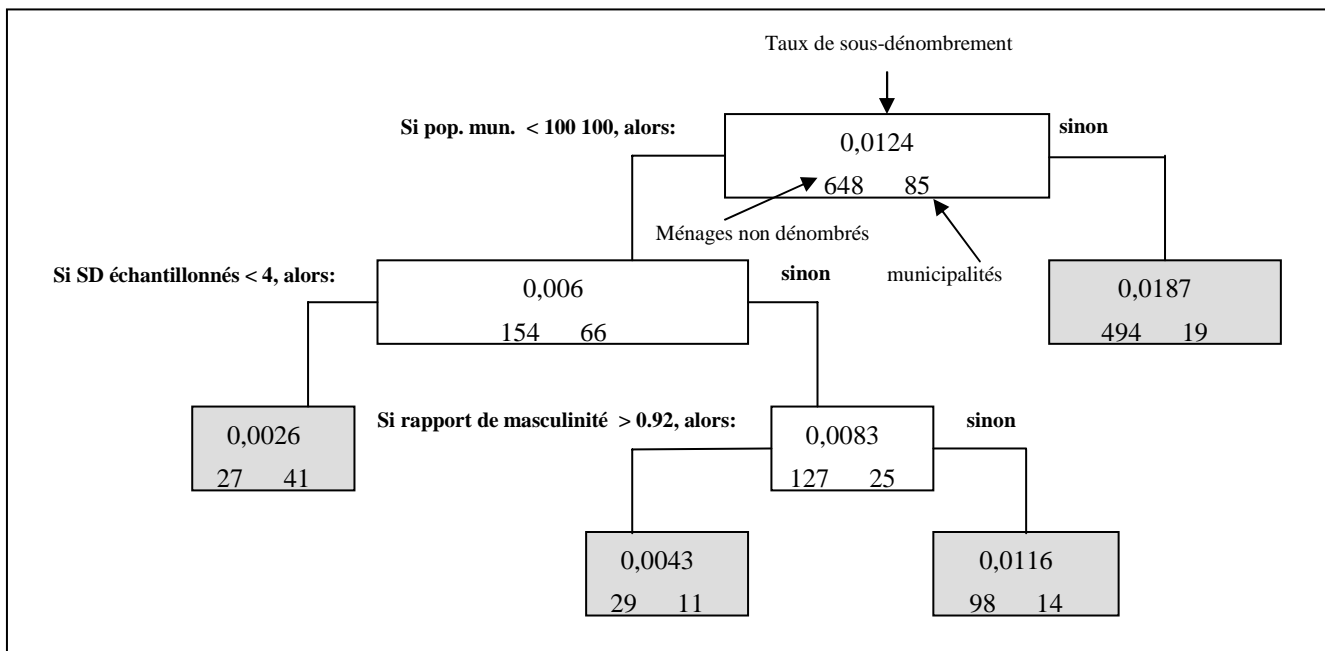


Figure 1. Arbre 1 fondé sur les variables démographiques.

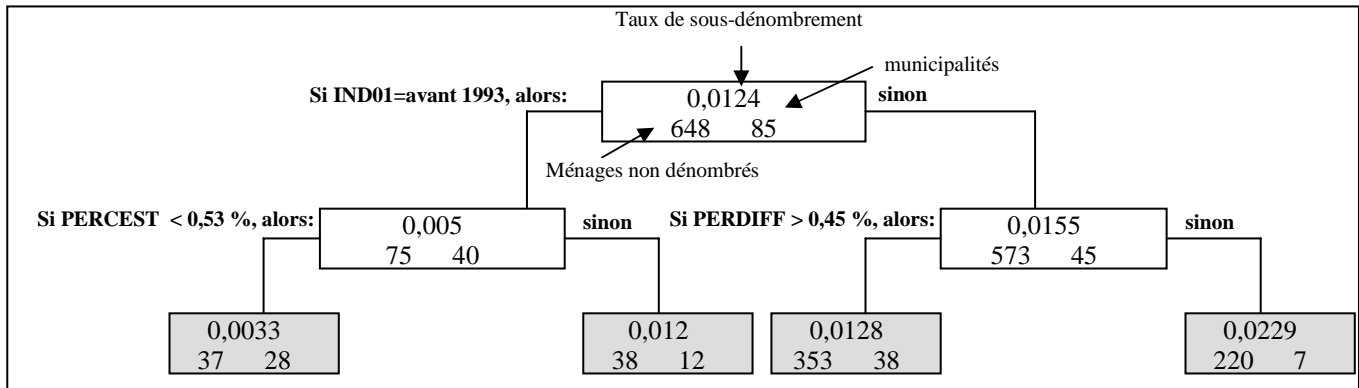


Figure 2. Arbre 2 fondé sur les variables de qualité des statistiques des municipalités.

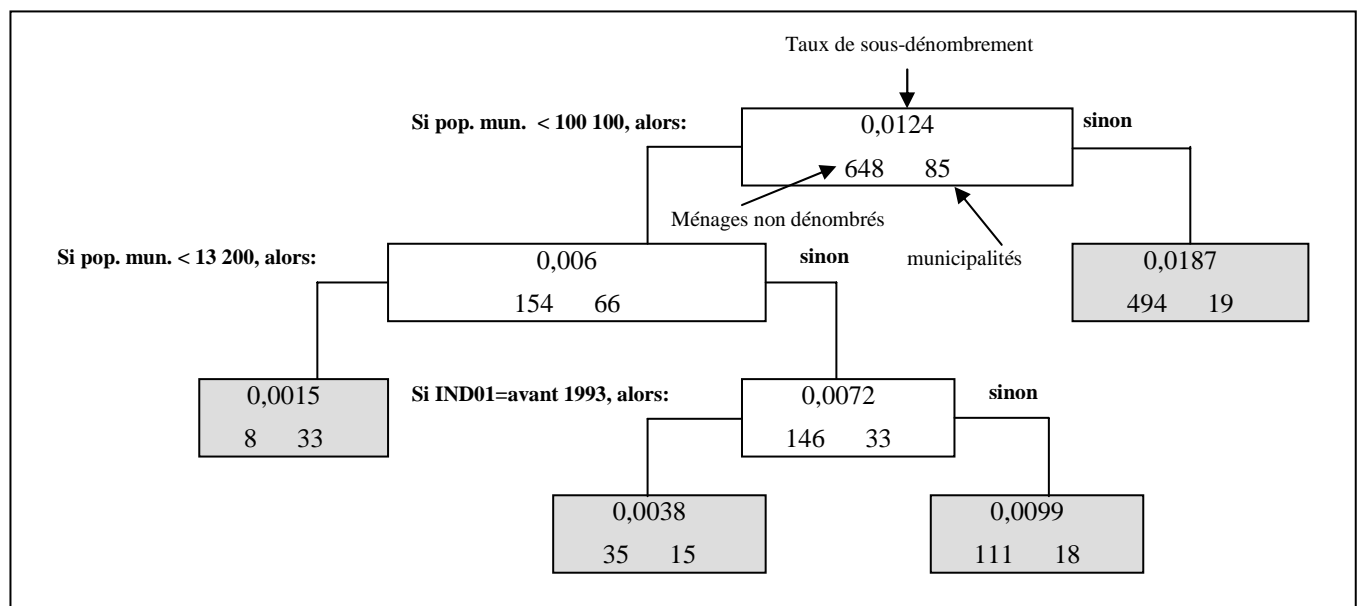


Figure 3. Arbre 3 fondé sur des variables démographiques et de qualité.

4. Modèles poisson-gamma hiérarchiques

Nous représentons le nombre de ménages non dénombrés observé dans chaque échantillon municipal par y_i ($i = 1, \dots, 85$). À titre de première approximation, nous pouvons modéliser ces dénombrements par une loi de Poisson :

$$y_i \mid \delta_i, e_i \sim \text{Pois}(\delta_i e_i) \tag{2}$$

où δ_i représente le taux de sous-dénombrement qu'il faut estimer et e_i est donné par le nombre de ménages dans les SD échantillonnés dans la municipalité. Nous exprimons la dépendance à un ensemble de variables explicatives par un lien canonique loglinéaire :

$$\ln(\delta_i e_i) = X_i \beta + Z_i \xi \tag{3}$$

où Z_i est la i^{e} ligne d'une matrice nominale de plan d'expérience introduite pour modéliser les effets de groupe. Chaque X_i est un vecteur de dimension p de variables explicatives associées à la i^{e} municipalité, et β et ξ sont les paramètres de régression.

Comparativement au nombre de ménages observés, les cas de non-dénombrement sont assez rares. Par conséquent, les données peuvent présenter une surdispersion importante. Le problème de la surdispersion peut être résolu par modélisation hiérarchique des paramètres δ_i dans (2). Si les paramètres δ_i suivent une loi Gamma(α, ν), on obtient marginalement la loi binomiale négative pour y_i par intégration des paramètres δ_i ; i.e., $y_i \mid \alpha, \nu, e_i \sim \text{NegBin}(\alpha, \nu / (\nu + e_i))$ avec les moments :

$$E(y_i \mid \alpha, e_i, \nu) = \frac{\alpha e_i}{\nu}, \quad V(y_i \mid \alpha, e_i, \nu) = \frac{\alpha e_i (\nu + e_i)}{\nu^2}$$

(voir Lawless 1987).

Au lieu de la paramétrisation susmentionnée, nous adoptons celle de la loi Gamma au deuxième niveau de la hiérarchie, conformément à la proposition de Christiansen et Morris (1997). Si nous supposons

$$\delta_i | \lambda_i, \zeta \sim \text{Gamma}(\zeta, \zeta/\lambda_i) \quad (4)$$

avec les moments $E(\delta_i | \lambda_i, \zeta) = \lambda_i$ et $V(\delta_i | \lambda_i, \zeta) = \lambda_i^2 / \zeta$, nous obtenons

$$y_i | e_i, \lambda_i, \zeta \sim \text{NegBin}\left(\zeta, \frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}\right),$$

où $V(y_i | e_i, \lambda_i, \zeta) - E(y_i | e_i, \lambda_i, \zeta) = e_i^2 \lambda_i^2 / \zeta$. À mesure que ζ tend vers l'infini, la variance de la loi binomiale négative converge vers celle de la loi de Poisson (la variance de la loi Gamma en (4) tend vers 0), tandis que les valeurs faibles de ζ indiquent une forte surdispersion.

Partant de (4), nous voyons immédiatement que :

$$E(\delta_i e_i | e_i, \lambda_i, \zeta) = \lambda_i e_i;$$

donc, l'hypothèse de dépendance (3) se réénonce en fonction de $\lambda_i e_i$ sous la forme :

$$\ln(\lambda_i e_i) = X_i \beta + Z_i \xi.$$

La loi a priori (4) est conjuguée à la vraisemblance définie par (2). Conséquemment, nous obtenons

$$\delta_i | y_i, e_i, \lambda_i, \zeta \sim \text{Gamma}(y_i + \zeta, e_i + \zeta/\lambda_i)$$

dont il découle que

$$E(\delta_i | y_i, e_i, \lambda_i, \zeta) = (1 - B_i)r_i + B_i \lambda_i \quad (5)$$

où $r_i = y_i / e_i$ et $B_i = \zeta / (\zeta + e_i \lambda_i)$.

Nous pouvons considérer chaque moyenne a posteriori (5) comme un estimateur sur petit domaine composite où les composantes directe et synthétique sont toutes deux pondérées conformément à l'information fournie par l'échantillon.

D'après (5), nous notons que la moyenne a posteriori de la distribution des paramètres de taux δ_i est une combinaison linéaire du taux de sous-dénombrement observé r_i et de la moyenne a priori λ_i . Autrement dit, le modèle présente une linéarité a posteriori. Dans (5), les deux termes sont pondérés d'après B_i , dont la valeur varie entre 0 et 1. Plus la valeur de B_i est grande, plus le poids de la moyenne a priori λ_i (estimateurs synthétiques) est grande et les estimations produites par le modèle prennent de l'importance comparativement aux taux observés. Nous constatons que chaque B_i est inversement proportionnel au terme $e_i \lambda_i$, qui exprime la quantité d'information que fournit l'échantillon de chaque domaine.

Pour achever la spécification bayésienne du modèle, nous attribuons une loi aux paramètres de troisième niveau ζ, β, ξ . D'après un critère approximatif non informatif, nous introduisons des lois a priori correctes, mais plates. Plus précisément, nous supposons que :

$$\beta_j \stackrel{\text{iid}}{\sim} N(0, 100), \quad j = 1, \dots, p \quad (6)$$

$$\xi_k \stackrel{\text{ind}}{\sim} N\left(k\bar{u}_k, \frac{1}{\tau \bar{n}_k}\right), \quad k = 1, \dots, q \quad (7)$$

où \bar{u}_k est le sous-dénombrement moyen dans le k^e groupe et \bar{n}_k est le nombre moyen de ménages échantillonnés dans les municipalités du même groupe. Les lois a priori (7), associées aux effets de groupe, sont par conséquent centrées sur les moyennes de groupes et leur précision est proportionnelle à la taille des groupes. Elles sont construites de façon à être faiblement informatives en vue d'améliorer les propriétés de stabilité et de convergence du modèle. Les lois a priori des coefficients de régression (6) associées aux autres variables explicatives sont centrées sur 0. Pour le paramètre de surdispersion ζ , nous choisissons la loi a priori

$$\zeta \sim 1000^* \text{Gamma}(0,001, 1) \quad (8)$$

en suivant la proposition de Christiansen et Morris (1997). Notons que les deux premiers moments a priori de (8) sont $E(\zeta) = 1$ et $V(\zeta) = 1000$; donc, la loi a priori est très diffuse et caractérisée par une forte asymétrie positive.

Au quatrième niveau de la hiérarchie, nous spécifions les lois a priori suivantes :

$$k \sim N(0, 100) \quad (9)$$

$$\tau \sim \text{Gamma}(0,001, 0,001). \quad (10)$$

qui sont toutes deux conçues pour avoir un effet très faible sur les inférences a posteriori.

Nous calculons les lois a posteriori de $(\delta_i | y_i, e_i)$ à l'aide d'algorithmes d'échantillonnage de Monte Carlo par chaînes de Markov (MCMC). Pour ces calculs, nous utilisons le logiciel BUGS (Spiegelhalter, Thomas, Best et Gilks 1995), qui est fondé sur l'échantillonnage de Gibbs. Puisque la résolution des modèles comportant des lois discrètes demande des calculs compliqués, nous spécifions les lois a priori (6) à (10) en choisissant des formes fonctionnelles simples bien connues, comme la loi normale et la loi Gamma, qui facilitent les calculs. L'examen de la sensibilité des moyennes a posteriori dans (6) à (10) n'a révélé aucune variation importante de ces moyennes. Donc, nous pouvons considérer les lois a priori comme étant non informatives. Pour évaluer la convergence, nous considérons la méthode à chaînes multiples proposée par Gelman et Rubin (1992), et nous exécutons trois chaînes différentes avec point de départ bien distinct pour chaque modèle. Nous considérons l'inspection visuelle du cheminement des chaînes et la statistique modifiée de Gelman et Rubin (Brooks et Gelman 1998) comme des outils élémentaires d'évaluation de la convergence. Nous avons exécuté 10 000 itérations pour chaque chaîne, en éliminant 3 000

en moyenne, à titre de « rodage » prudent, ce qui donne environ 20 000 tirages à partir de la loi a posteriori de chaque modèle.

5. Comparaison des modèles et discussion des résultats empiriques

Nous avons estimé une gamme de modèles pour diverses définitions des matrices des variables explicatives X et Z . En ce qui concerne la matrice de plan d'expérience Z , nous considérons sept cas distincts, dans lesquels les municipalités sont groupées en fonction des critères de stratification classiques (région géographique et taille de la population) ou des résultats des partitions techniques décrites à la section 3. Il s'agit des groupements fondés sur : a) la région géographique (Nord, Centre, Sud et îles), b) les classes de taille de population uniquement, c) les classes de taille de population selon la région géographique, d) les classes de taille de population et les régions géographiques, e) l'arbre 1 (fondé sur les variables démographiques), f) l'arbre 2 (fondé sur les variables de qualité), g) l'arbre 3 (fondé sur les variables démographiques et de qualité). On peut proposer deux types de variables pour la matrice X , à savoir les variables de qualité de la section 2.2 et les variables démographiques de la section 2.3. Par conséquent, la matrice X peut avoir trois compositions distinctes : I) variables de qualité uniquement, II) variables démographiques uniquement et III) variables de qualité et démographiques. En établissant la correspondance entre les diverses définitions de X et Z , nous avons estimé vingt-huit modèles distincts. Cette méthode nous permet d'introduire différents blocs de variables, au lieu de procéder à la sélection de variables.

La quantité habituellement utilisée pour comparer les modèles dans le cadre bayésien est le facteur de Bayes (BF). Une approximation en grand échantillon de $-2\ln(BF)$ est donnée par

$$\Delta BIC = -2\ln \left[\frac{\sup_{M_0} f(y | \theta_0)}{\sup_{M_k} f(y | \theta_k)} \right] - (p_k - p_0) \ln n \quad (11)$$

(voir Schwarz 1978) qui, de surcroît, ne renvoie à aucune des hypothèses a priori. Nous notons que, dans (11), les M_k ($k = 1, \dots, K$) indiquent l'ensemble de modèles concurrents et que θ_k est le paramètre dimensionnel p_k indiquant la vraisemblance associée à chaque modèle. Le modèle nul auquel sont comparés tous les autres, qui est celui comportant la seule coordonnée à l'origine, est dénoté par M_0 . Les valeurs positives et grandes de (11) appuient le modèle M_k .

Dans (11), la pénalisation de complexité dépend de la taille du sous-ensemble de paramètres de troisième niveau; autrement dit, tous les modèles sont comparés comme s'ils n'étaient pas hiérarchiques. Puisqu'ils ont une même structure hiérarchique, cette modification opérationnelle du critère d'information bayésien type BIC ne modifie pas les

résultats de la comparaison des modèles résumés au tableau 1.

Nous notons que les modèles où les effets de groupe sont fondés sur la région géographique donnent de forts mauvais résultats (ligne 1) et qu'il en est de même si l'on combine la région géographique et la taille de la population des municipalités (lignes 3 et 4). Ces résultats sont assez étonnants, puisqu'on se fonde sur les régions géographiques pour concevoir la stratification de l'échantillon de l'EPR et que l'efficacité des administrations, ainsi que d'autres indicateurs socio-économiques sont censés être regroupés en fonction des grandes subdivisions géographiques de l'Italie (Nord, Centre, Sud). Ce résultat peut être attribué au rôle prédominant que l'organisation particulière de chaque municipalité joue dans la détermination de l'efficacité des opérations du recensement sur le territoire de la municipalité.

Les modèles à effets de groupe fondés sur un arbre (lignes 5 à 7) donnent de nettement meilleurs résultats que ceux à effets de groupe fondés sur les critères de stratification habituels d'ISTAT (lignes 1 à 4). Seuls font exception les modèles basés sur l'arbre 2 (ligne 5), qui donnent des résultats assez médiocres lorsqu'on n'inclut pas la taille de la population et d'autres variables démographiques. En fait, la population municipale peut être considérée comme étant une approximation de la complexité organisationnelle de la municipalité. Il semble que les variables de qualité soient des discriminants puissants du niveau de sous-dénombrement parmi les municipalités dont les caractéristiques démographiques sont semblables, mais qu'ils aient peu de pertinence si l'on ne tient pas compte de l'effet d'un degré de complexité organisationnelle différent par introduction d'une variable de taille de population. Nous soulignons que l'ajout d'une matrice de plan d'expérience Z fondée sur des groupes de municipalités établis d'après des arbres de régression de Poisson nous permet de modéliser les relations non linéaires entre le sous-dénombrement et les variables explicatives.

En fait, les modèles fondés sur l'arbre 3 sont ceux dont les propriétés sont les meilleures. Suivent plusieurs commentaires sur le modèle dont le ΔBIC est maximal. Ce modèle comprend des variables démographiques et des variables de qualité comme variables explicatives. L'adéquation du modèle choisi est évaluée au moyen de vérifications prédictives a posteriori. Plus précisément, nous adoptons la mesure d'usage générale de la divergence par rapport à l'ajustement valide du modèle proposée par Brooks, Catchpole et Morgan (2000) en tant qu'outil approprié pour les occurrences rares, comme les sous-dénombrements au recensement :

$$D(y; \theta) = \sum_i (\sqrt{y_i} - \sqrt{\text{Exp}_i})^2, \quad (12)$$

où $\text{Exp}_i = e_i E(\delta_i | y_i, e_i)$. La probabilité de 0,46 associée à l'aire de la queue témoigne d'un bon ajustement pour le modèle choisi.

Tableau 1
 Δ BIC des modèles estimés comparativement au modèle de référence M_0

		Variables dans les modèles			
		Effets de groupe uniquement	Effets de groupe + variables de qualité	Effets de groupe + variables démographiques	Effets de groupe + variables de qualité et démographiques
Effets de groupe	Région	-4,22	-0,39	18,52	23,32
	Classes de pop. mun.	15,34	17,87	17,32	20,09
	Région* classes pop. mun.	2,08	6,13	4,91	8,45
	Région + classes pop. mun.	9,68	13,2	13,74	17,83
	Arbre 2 (variables de qualité)	11,81	8,34	23,48	26,15
	Arbre 1 (variables démographiques)	35,14	35,37	32,28	35,53
	Arbre 3 (variables de qualité + démographiques)	38,89	35,76	41,12	41,45

Nous avons de nouveau estimé l'ensemble de modèles après élimination de la municipalité la plus grande, qui constitue un cas influent éventuel. De nouveau, le modèle fondé sur l'arbre 3 avec variables démographiques et de qualité en tant que variables explicatives a été sélectionné à l'aide du critère (11). Ce modèle est caractérisé par un bon ajustement (la valeur p bayésienne associée à la mesure de divergence (12) est égale à 0,51). En outre, les estimations composites varient peu comparativement à celles obtenues au moyen de l'échantillon entier.

Afin de vérifier l'ajustement du modèle, à la figure 4, nous avons tracé les estimations composites en fonction des estimations directes du nombre de ménages non dénombrés dans chaque municipalité (les valeurs pour les 10 plus grandes municipalités sont présentées pour une échelle différente). Les estimations composites sont $w_i e_i E(\delta_i | y_i, e_i)$, tandis que les estimations directes sont $w_i y_i$, w_i étant le facteur d'expansion dû à l'échantillonnage des SD dans chaque municipalité. Les estimations composites sont les espérances a posteriori des paramètres de premier niveau et, conditionnellement aux hyperparamètres, sont des estimations composites dans lesquelles on accorde peu de poids aux prédictions fondées sur le modèle représenté par les λ_i lorsqu'on dispose de données d'échantillonnage valables. D'après (5), nous savons que cette méthode de pondération est régie par les facteurs municipaux de rétrécissement B_i . Ceux-ci pondèrent les estimations directes y_i / e_i proportionnellement à $e_i \lambda_i$, c'est-à-dire le nombre de ménages non dénombrés dans l'échantillon municipal prévu par le modèle.

Pour les municipalités comptant jusqu'à 10 000 résidents (cette valeur est assez proche de la valeur de partition de 13 200 de l'arbre 3), dans presque tous les cas, nous avons des valeurs de B_i très proches de 1; autrement dit, pour les petites municipalités, la composante basée sur le modèle joue un rôle prédominant dans la détermination de l'estimation composite. À la figure 5, les estimations composites

(et leurs intervalles de crédibilité à 95 %) sont tracées en fonction des estimations directes.

La largeur des intervalles de crédibilité dépend du niveau de sous-dénombrement et, comme il faut s'y attendre, est importante lorsque la taille de l'échantillon à l'intérieur de la municipalité est faible. Les estimations composites associées à un grand intervalle de crédibilité sont également caractérisées par un facteur de rétrécissement important, à cause du peu d'information provenant de l'échantillon. Pour certaines municipalités de taille moyenne, le fait qu'elles soient sous-échantillonnées par rapport à leur taille pourrait expliquer l'obtention d'un grand intervalle.

Dans le cas des petites municipalités, où le recensement est réalisé plus facilement, le sous-dénombrement est généralement très faible. Son estimation est difficile, puisque fort peu de SD sont échantillonnés à l'heure actuelle à partir de chaque petite municipalité, ce qui, souvent, ne donne aucune preuve de sous-dénombrement. Le cas échéant, l'estimation composite correspond essentiellement à la composante fondée sur le modèle. Par conséquent, pour la prochaine EPR, étant donné la taille globale de l'échantillon, nous recommandons de ne pas insister sur l'échantillonnage d'un grand nombre de petites municipalités, mais plutôt de réorienter l'échantillonnage vers les municipalités de taille moyenne, qui sont plus hétérogènes. En outre, on devrait augmenter le nombre de SD échantillonnés dans les petites municipalités sélectionnées.

Les résultats de la présente étude, où l'on envisage pour la première fois d'utiliser un critère de groupement des municipalités en fonction de leur rendement concernant les opérations statistiques, confirment qu'on pourrait améliorer les résultats lors de futures enquêtes similaires en modifiant le plan d'échantillonnage stratifié et en modélisant le sous-dénombrement au moyen des covariables représentant les difficultés qu'éprouvent les municipalités lors de la réalisation des recensements.

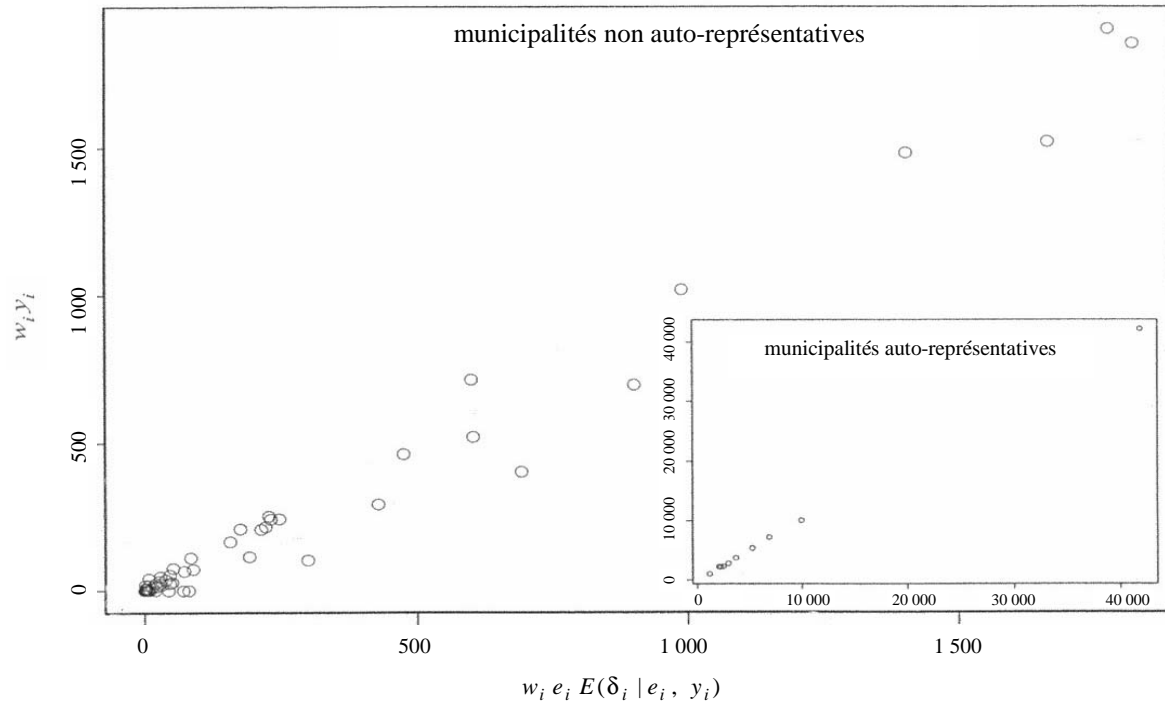


Figure 4. Estimations composites en fonction des estimations directes du nombre de ménages non dénombrés dans chaque municipalité.

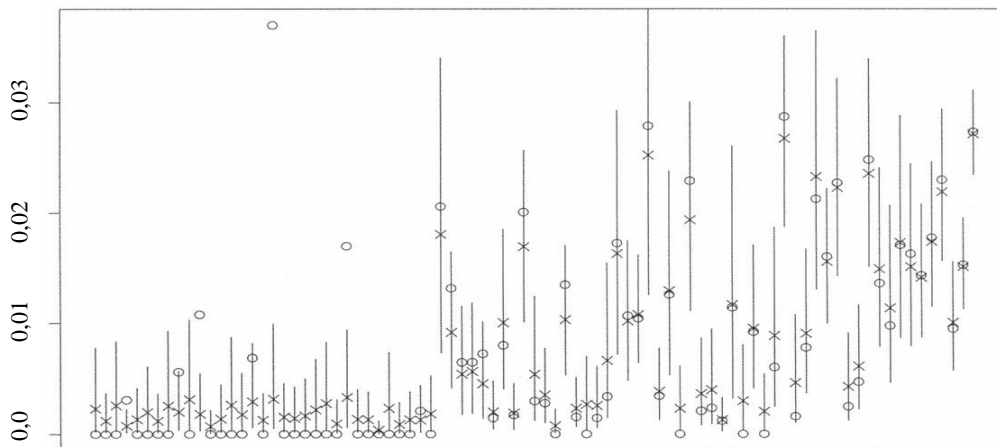


Figure 5. Estimations composites (x) et leurs intervalles de crédibilité à 95 %; (o) estimations directes. Les municipalités sont triées par taille de population.

Remerciements

Nous remercions Angela Ferruzza, Marco Fortini, Aldo Orasi et Fernanda Panizon, de l'équipe d'ISTAT travaillant sur le Recensement et l'EPR de 2001, ainsi que Mariella Dimitri et Ersilia Di Pietro, du groupe d'ISTAT s'occupant des enquêtes sur le rendement statistique des municipalités, de leurs suggestions utiles et de leur soutien continu.

Les travaux ont été financés en partie par la subvention du projet de recherche «Quality of total and partial surveys» (1999–2000) de l'Université de Bologne (60 %).

Les données de l'EPR et les archives contenant les données sur les municipalités ont pu être utilisées grâce à une entente spéciale entre ISTAT et le Département de statistique de l'Université de Bologne.

Nous tenons à remercier Francesca Bruno et Loredana Di Consiglio de leur contribution inestimable à la préparation des ensembles de données de base, et Meri Raggi de son soutien permanent et de sa discussion des sujets de la présente étude.

Nous remercions le rédacteur, un rédacteur adjoint et deux examinateurs anonymes de leurs commentaires et suggestions qui nous ont aidé à réviser et à améliorer le manuscrit.

Bibliographie

- Abbate, C., Masselli, M. et Signore M. (1993). A combined post-enumeration survey for the 1991 Italian population and industrial censuses. *Bulletin of the International Statistical Institute, Firenze, 48th Session*, Tome LV, 2, 159-173.
- Alho, J.M., Mulry, M.H., Wurdeman, K. et Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J., (1984). *Classification and Regression Trees*. Wadsworth, California.
- Brooks, S.P., Catchpole, E.A. et Morgan, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science*, 15, 357-276.
- Brooks, S.P., et Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *Journal of Computational and Graphical Statistics*, 7, 434-455.
- Christiansen, C.L., et Morris, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association*, 92, 618-632.
- Di Pietro, E. (1998). Anagrafi comunali: funzione statistica e livello di informatizzazione. *Atti Della Quarta Conferenza Nazionale di Statistica*. Tomo 1–Sessioni Plenarie, Workshop: Il progetto anagrafi. Roma. 11-13.
- Di Pietro, E. (1999). Anagrafe informatizzata e Censimenti demografici: dal censimento tradizionale al censimento basato sugli Archivi. *Società Italiana di Statistica: Atti Del Convegno "Verso i Censimenti del 2000*. Udine 7-9 giugno. 169-182.
- Fortini, M. (1994). Un'applicazione del modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione. *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*. San Remo 6-8 Aprile. 2, 423-430.
- Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*, 7, 457-72.
- Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15, 209-225.
- Moura, F.A.S., et Holt, D. (1999). Production d'estimations régionales à partir de modèles multiniveau. *Techniques d'enquête*, 25, 81-89.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Spiegelhalter, D.J., Thomas, A., Best, N. et Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*. Rapport Technique, Medical Research Council biostatistics Unit, Institute of Public Health, Cambridge University.
- Therneau, T.M., et Atkinson, E.J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*. Rapport Technique, Mayo Foundation.
- Wolter, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.