Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Prédiction de la population de petits domaines au moyen de modèles hiérarchiques

par Debora F. Souza, Fernando A.S. Moura et Helio S. Migon

Décembre 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2009
Vol. 35, N° 2, pp. 221-234
Statistique Canada, N° 12-001-X au catalogue

Prédiction de la population de petits domaines au moyen de modèles hiérarchiques

Debora F. Souza, Fernando A.S. Moura et Helio S. Migon [1]

Résumé

Le présent article décrit une méthode de prédiction pour petits domaines fondée sur des données tirées d'enquêtes périodiques et de recensements. Nous appliquons cette méthode pour obtenir des prédictions démographiques pour les municipalités non échantillonnées dans l'enquête annuelle sur les ménages du Brésil (PNAD), ainsi que pour accroître la précision des estimations fondées sur le plan de sondage obtenues pour les municipalités échantillonnées. En plus des données fournies par la PNAD, nous utilisons des données démographiques provenant des recensements de 1991 et de 2000, ainsi que d'un dénombrement complet de la population effectué en 1996. Nous proposons et comparons des modèles de croissance hiérarchiquement non structurés et spatialement structurés qui gagnent en puissance en s'appuyant sur toutes les municipalités échantillonnées.

Mots clés : Méthode de Monte Carlo par chaines de Markov (MCMC) ; projection démographique ; modèles spatiaux.

1. Introduction

Comme dans de nombreux autres pays, la demande de statistiques détaillées et à jour sur des petits domaines a augmenté régulièrement au Brésil. La nécessité de brosser un tableau plus précis des sous-régions, en vue de résoudre des problèmes de distribution, d'équité et de disparité, est à l'origine de cet accroissement de la demande. Par exemple, certaines sous-régions ou certains sous-groupes pourraient, à certains égards, être à la traîne par rapport à la moyenne globale. Par conséquent, il est nécessaire de repérer ces régions et d'obtenir des renseignements statistiques à ce niveau géographique avant de pouvoir prendre toute mesure éventuelle de correction. Outre ces exigences nationales, les autorités locales doivent disposer d'estimations fiables, comme les cacactéristiques démographiques, à des fins d'analyse, de planification et d'administration.

Au Brésil, un exemple important de demande d'estimations fiables a trait à la façon dont le revenu fédéral, qui doit être partagé en vertu de la constitution, est réparti entre les diverses municipalités (le Brésil est une république fédérée constituée d'États et du District fédéral. Les États sont subdivisés en municipalités, qui partagent des caractéristiques des villes et des comtés – elles peuvent contenir plus d'une région urbaine, mais elles ne sont dotées que d'un seul maire et d'un seul conseil municipal). Le nombre prédit d'habitants de la municipalité est utilisé par le gouvernement fédéral comme critère pour affecter les fonds. D'où la nécessité d'obtenir des prévisions fiables de la population municipale afin d'appliquer équitablement ce critère, réglementé par la loi fédérale.

L'enquête annuelle sur les ménages (PNAD) est une source importante de données démographiques. Toutefois, elle n'est pas conçue pour produire des estimations au niveau municipal. Autrement dit, à part quelques municipalités, les tailles des échantillons municipaux ne sont pas suffisamment grandes pour produire des erreurs-types acceptables quand sont utilisées des estimations par sondage directes. De surcroît, un nombre important de municipalités ne sont pas échantillonnées du tout.

L'approche courante consiste à obtenir des estimations démographiques municipales en se basant d'abord sur une prédiction pour une plus grande région, puis en utilisant des données auxiliaires pour répartir la population totale prédite entre les municipalités. À son tour, la prédiction pour une région plus grande est produite en supposant que les taux de natalité, de mortalité et de migration sont les mêmes pour toutes les municipalités. Le principal inconvénient de cette méthode est qu'elle dépend de l'évolution hypothétique du modèle. Elle ne tient pas compte de toutes les incertitudes et, en général, ne fournit pas de mesures de l'erreur des estimations.

Le problème de l'estimation sur petits domaines a suscité de l'intérêt dans la littérature statistique à cause de la demande croissante d'information statistique détaillée des secteurs public et privé. Un excellent exposé à jour sur les méthodes d'estimation sur petits domaines et leurs applications peut être consulté dans Rao (2003). Les données sur les petits domaines proviennent principalement d'enquêtes périodiques dont les tailles d'échantillon ne sont pas suffisamment grandes pour pouvoir produire des estimations fiables pour les domaines. Un moyen d'aborder le problème

1. Debora F. Souza, Département des méthodes et de la qualité, IBGE, Rio de Janeiro, 20031-170. Courriel : debora.souza@ibge.gov.br ; Fernando A.S. Moura, Universidade do Brasil-UFRJ, Rio de Janeiro, Brésil. Courriel : fmoura@im.ufrj.br ; Helio S. Migon, Universidade do Brasil-UFRJ, Rio de Janeiro, Brésil. Courriel : migon@im.ufrj.br.

consiste à renforcer l'estimation en empruntant de l'information à tous les domaines et à d'autres sources de données connexes. Comme l'a énoncé Pfeffermann (2002), les sources de données appropriées pour cette tâche peuvent être classées en deux catégories, à savoir les données provenant d'autres domaines semblables en ce qui concerne les caractéristiques d'intérêt, les données antérieures obtenues pour la caractéristique d'intérêt et l'information auxiliaire. Dans notre contexte démographique, la principale source de données connexes comprend les recensements de 1991 et de 2000, ainsi qu'un dénombrement complet de la population exécuté en 1996.

Le but de la présente étude est d'obtenir des estimations des populations municipales fondées sur des données d'enquête fournies par la PNAD et sur des données de recensement. Nous proposons un modèle hiérarchiquement non structuré et nous évaluons la qualité de son ajustement et son pouvoir prédictif. Nous envisageons également un modèle hiérarchique spatialement structuré dans l'esprit de Moura et Migon (2002), puisque le chiffre de population par domaine et son profil de croissance pourraient être reliés au développement des domaines voisins. Par souci de simplicité, dans la suite de l'exposé, nous donnons respectivement aux modèles hiérarchique non structuré et hiérarchique structuré spatialement les noms de modèle hiérarchique et modèle spatial.

À la section 2, nous décrivons les principales sources de données utilisées dans nos travaux. À la section 3, nous présentons les modèles proposés, ainsi qu'un critère de sélection de modèle. À la section 4, nous présentons des applications à des données réelles, ainsi qu'à des données simulées. Enfin, à la section 5, nous résumons brièvement l'étude et décrivons dans les grandes lignes nos futurs travaux de recherche.

2. Ensemble de données

Les données d'entrées pour les modèles présentés à la section 3 proviennent des cycles de 1992 à 1999 de l'enquête annuelle sur les ménages (PNAD), des recensements de 1991 et de 2000, et d'un dénombrement complet de la population effectué en 1996. Afin d'évaluer l'approche proposée, nous considérons les municipalités de l'État de São Paulo comme étant les domaines d'intérêt.

À la présente section, nous décrivons brièvement les sources des données, en mentionnant leurs principaux avantages et limites. Nous avons tiré de la PNAD les estimations démographiques directes pour les municipalités échantillonnées. Comme nous l'expliquons à la section 3, ces estimations servent de données d'entrée pour l'inférence au sujet de nos paramètres cibles. Nous utilisons aussi dans notre application les deux recensements et le dénombrement de population de 1996.

Le recensement démographique du Brésil est la source principale d'information au sujet de la population. Il est effectué tous les dix ans, habituellement au début de la décennie. Bien que l'objectif consiste à compter tous les membres de la population, certaines erreurs de recensement sont découvertes. L'importance des erreurs est évaluée au moyen d'une enquête postcensitaire exécutée peu après l'achèvement du recensement.

L'enquête annuelle sur les ménages (PNAD) est conçue pour produire des renseignements de base sur la situation socioéconomique du pays. L'unité étudiée est le ménage, pour lequel sont recueillis des renseignements annuels sur le nombre de membres, leur sexe, leur niveau d'études, leur situation d'emploi, *etc*. L'enquête n'est pas exécutée durant les années de recensement et n'a pas été réalisée en 1994 pour des raisons administratives. L'échantillon est sélectionné selon un plan d'échantillonnage en grappes à trois degrés. Les unités primaire et secondaire d'échantillonnage sont respectivement la municipalité et le secteur de dénombrement (qui compte, en moyenne, 250 ménages). Les municipalités sont stratifiées en fonction de la taille de leur population déterminée d'après le dernier recensement. Au premier degré, toutes les municipalités appartenant aux régions métropolitaines et aux capitales des États (lesquelles, au Brésil, sont normalement les plus grandes villes dans les États respectifs) sont échantillonnées. Les municipalités dont la population est supérieure à une certaine valeur seuil sont également incluses dans l'échantillon avec une probabilité de un. Celles qui restent sont stratifiées et deux d'entre elles sont échantillonnées dans chaque strate avec une probabilité proportionnelle à la taille de leur population.

Les secteurs de dénombrement sont échantillonnés avec probabilité proportionnelle au nombre de ménages résidant dans le secteur au moment du dernier recensement. Enfin, au dernier degré, les ménages sont échantillonnés systématiquement avec probabilité égale au moyen d'une liste qui est mise à jour au début de l'enquête. Les mêmes municipalités et secteurs de dénombrement sont gardés pour toutes les enquêtes exécutées durant une décennie particulière, tandis que les ménages sont échantillonnés chaque année.

Puisque chaque secteur est échantillonné avec une probabilité proportionnelle à son nombre respectif de ménages, on pourrait soutenir que le mécanisme d'échantillonnage est informatif en ce qui concerne la population du secteur. Toutefois, puisque la variable réponse effectivement utilisée dans la présente étude est la densité de population par domaine, il est raisonnable de supposer que le mécanisme de sélection de l'échantillon n'est pas pertinent. Donc, nous n'abordons pas cette question dans

l'étude. En ce qui concerne l'inférence pour petits domaines sous échantillonnage informatif, Pfeffermann et Sverchkov (2007) est une bonne référence. Nous recommandons également de consulter Pfeffermann, Moura et Silva (2006) aux lecteurs qui aimeraient savoir comment suivre une approche de modélisation hiérarchique bayésienne sous échantillonnage informatif.

3. Spécification du modèle

3.1 Modèle de croissance exponentielle

Soit y_t les valeurs d'échantillon d'une loi appartenant à une famille exponentielle dont la valeur prévue est donnée par $\pi_t = E(y_t \mid \theta_t)$ où θ_t est un vecteur de paramètres inconnus.

Une classe importante et vaste de modèles de croissance exponentielle paramétrisés par $(\alpha, \beta, \gamma, \phi)$ est définie par :

$$\pi_t = [\alpha + \beta \exp(\gamma t)]^{1/\phi}. \qquad (1)$$

Certains cas particuliers bien décrits dans la littérature sont les distributions :

1) logistique : avec $\phi = -1$, $\pi_t^{-1} = \alpha + \beta \exp(\gamma t)$;
2) Gompertz : avec $\phi = 0$, en définissant (1) comme $\log(\pi_t) = \alpha + \beta \exp(\gamma t)$;
3) exponentielle modifiée : avec $\phi = 1$, $\pi_t = \alpha + \beta \exp(\gamma t)$.

Le principal avantage de l'utilisation du modèle (1) est qu'il est possible de garder l'échelle originale des observations y_t et de ne changer que la trajectoire de π_t, ce qui facilite l'interprétation. De surcroît, les intervalles de temps ne doivent pas être tous de la même longueur, si bien que les données peuvent provenir de diverses sources de référence (voir la section 4 pour plus de renseignements).

Quand $\psi = \exp(\gamma) < 1$, le processus est non explosif, ce qui implique que π_t converge vers $\alpha^{1/\phi}$ quand $t \to \infty$, sous la condition que, pour $\phi = 0$, cette quantité est égale à $\log(\alpha)$. Quand $\psi > 1$, les courbes sont concaves pour $\phi \geq 0$ et $\beta > 0$, donnant lieu à un processus explosif. Cette classe de modèles est celle des modèles généralisés de croissance exponentielle. Migon et Gamerman (1993) montrent comment le modèle de croissance exponentielle peut être vu comme un cas particulier d'un modèle dynamique général.

3.2 Modèles de croissance hiérarchiques

Dans le présent article, les principaux paramètres d'intérêt π_{it} sont les fonctions de croissance exponentielle non linéaire dont certains paramètres sont hiérarchiquement ou spatialement structurés. Les modèles spatialement structurés fournissent d'autres moyens de relier des domaines voisins semblables. Nous supposons en outre que la variance d'échantillonnage σ_{it}^2 suit un modèle qui dépend de la taille d'échantillon dans la municipalité en question. Dans la présente étude, nous ajustons et comparons des modèles hiérarchiques et spatiaux.

Nous supposons que les tailles de population sont disponibles pour chacune des m municipalités de l'État de São Paulo pour les recensements de 1991 et 2000, ainsi que pour le dénombrement complet de la population de 1996. Dans la suite de l'exposé, nous les appelons simplement données de recensement. Afin d'améliorer l'hypothèse d'échangeabilité des paramètres décrivant la moyenne du processus, nous prenons comme variables réponses les estimations des densités de population des municipalités échantillonnées au lieu des estimations de population municipale. Voir aussi la fin de la section 2 pour d'autres raisons d'utiliser les densités.

Pour chaque période, les estimations de ces quantités ne sont disponibles que pour $k < m$ municipalités correspondant aux unités de premier degré de l'échantillon de la PNAD. Afin d'estimer la densité de population municipale, nous divisons simplement l'estimation de la population totale par la superficie de la municipalité.

Soit y_{it} la densité de population obtenue d'après les données de recensement ou estimées d'après la PNAD à la période t, $t = 1, \ldots n$ pour la i^{e} municipalité, $i = 1, \ldots, m$. Notre but est de faire des inférences au sujet de la densité de population réelle π_{it} pour la population de toutes les municipalités, y compris celles qui ne sont pas échantillonnées. À la section suivante, nous modélisons les densités réelles de population municipale π_{it} au moyen d'une fonction de croissance hiérarchique non linéaire stochastique. Nous supposons que les quantités aléatoires y_{it} suivent une loi normale de moyenne π_{it} et de variance σ_{it}^2.

Nous adoptons une approche bayésienne pour cette étude. Par conséquent, les prédictions sont décrites par des lois de probabilité, ce qui donne aux utilisateurs l'occasion d'analyser les incertitudes que comporte le processus de décision. Ce fait est l'un des avantages, parmi de nombreux autres, de l'utilisation de ce genre d'approche.

Les valeurs de y_{it} ne sont obtenues pour toutes les municipalités de l'État de São Paulo que pour les années de recensement. Bien que l'on s'efforce, durant le recensement, d'obtenir le dénombrement complet de toute la population, des erreurs de couverture peuvent avoir lieu. Par conséquent, nous émettons l'hypothèse du modèle qui suit pour les données de recensement ainsi que pour celles provenant de la PNAD, à l'exception des variances σ_{it}^2, qui sont fixées à une valeur plus faible pour les données de

recensement (voir la section 3.4 ainsi que les remarques finales à la section 5) :

$$\begin{aligned} y_{it} &= \pi_{it} + \varepsilon_{it},\ \varepsilon_{it} \sim N(0,\ \sigma_{it}^2) \\ \pi_{it} &= \{\alpha_i + \beta \exp(\gamma_i t)\}^{1/\phi} \\ \alpha_i &= \alpha + \xi_{\alpha_i},\ \xi_{\alpha_i} \sim N(0,\ \sigma_{\xi_\alpha}^2) \\ \gamma_i &= \gamma + \xi_{\gamma_i},\ \xi_{\gamma_i} \sim N(0,\ \sigma_{\xi_\gamma}^2) \end{aligned} \tag{2}$$

où les lois a priori de α, β et γ sont données par : $\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\beta \sim N(\mu_\beta, \sigma_\beta^2)$, $\gamma \sim N(\mu_\gamma, \sigma_\gamma^2)$. Il convient de souligner que l'information provenant de tous les domaines est obtenue au moyen de la structure hiérarchique des paramètres α_i, et γ_i. Une autre façon de permettre l'emprunt d'information entre les municipalités consiste à supposer que les α_i sont spatialement structurés (voir la section 3.3). Si nous supposons que la moyenne π_{it} est non explosive, nous pouvons considérer le paramètre $\alpha^{1/\phi}$ comme étant la valeur à laquelle la population municipale moyenne se stabilise. Les paramètres β et γ affectent l'évolution de la densité de population au cours du temps. Les lois a priori de α, β et γ peuvent être choisies en profitant d'une certaine connaissance démographique a priori de l'évolution prévue de la population. Dans notre application, nous fixons $\phi = 1$, ce qui implique que, pour $t = 0$, la valeur réelle de la densité de population dans chaque municipalité est donnée par $\alpha_i + \beta$. La structure hiérarchique imposée aux paramètres α_i implique que la valeur espérée de la densité réelle pour toute municipalité à la période $t = 0$ est $\alpha + \beta$. Supposer que les paramètres de croissance, γ_i, possèdent une structure hiérarchique signifie que les densités ont des taux de croissance différents, mais ont en commun la même moyenne. Une petite étude en simulation (voir la section 4.1) nous dicte de maintenir le paramètre β fixe pour tous les domaines, sans aucune perte de généralité, puisque les niveaux diffèrent encore pour différentes municipalités. Dans tous les modèles considérés dans notre application, nous supposons que $\tau_\alpha^2 = \sigma_\alpha^{-2} \sim G(a_\alpha, b_\alpha)$, $\tau_\gamma^2 = \sigma_\gamma^{-2} \sim G(a_\gamma, b_\gamma)$. Afin d'attribuer des priors vagues, à la section 4.2, nous donnons des valeurs faibles aux paramètres reliés à ces lois a priori de la précision.

L'hypothèse selon laquelle la fonction moyenne π_{it} est donnée par une courbe de croissance exponentielle permet d'effectuer une correction en cas d'accroissement ou de diminution de la densité de la population. Les sources de données utilisées ont des périodes de référence différentes et les données ne sont pas réparties de la même façon dans le temps. Dans chaque cas, l'utilisation d'une courbe de croissance exponentielle offre un avantage supplémentaire, puisque nous pouvons simplement produire une échelle de temps afin de nous conformer aux différentes sources de données, comme nous l'expliquons dans la description de l'application à la section 4.

3.3 Modèle spatial

Dans le modèle hiérarchique présenté à la section précédente, l'information provenant de tous les domaines est combinée afin de prédire la population d'un domaine particulier. Cependant, il est raisonnable de supposer que les densités démographiques de deux municipalités voisines sont plus semblables que celles de deux autres choisies arbitrairement. La structure régionale est représentée par la loi a priori conjointe des effets spatiaux aléatoires. Nous considérons que deux domaines sont voisins s'ils ont une limite commune.

Dans le modèle que nous proposons, la densité démographique dans un domaine i à la période t, π_{it}, est affectée par les domaines voisins par ajout d'effets spatiaux aléatoires δ_{α_i} aux paramètres α_i, c'est-à-dire $\alpha_i = \alpha + \delta_{\alpha_i}$, où α est un terme représentant l'ordonnée à l'origine. Par conséquent, α_i varie uniquement avec l'effet spatial, ce qui représente un effet local, tandis que les paramètres de croissance γ_i sont considérés comme semblables dans tous les secteurs (effet global).

La relation entre les domaines voisins est définie dans les lois a priori des δ_{α_i}. La loi conjointe a priori de $\delta_\alpha = (\delta_{\alpha_1}, ..., \delta_{\alpha_m})'$ sachant l'hyperparamètre σ_α^2, est définie comme dans Mollié (1996) :

$$p(\delta_\alpha \mid \sigma_\alpha^2) \propto \frac{1}{\sigma_\alpha^{m/2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2} \sum_{i=1}^{m} \sum_{k<i} w_{ik} (\delta_{\alpha_i} - \delta_{\alpha_k})^2\right\} \tag{3}$$

où w_{ik} représente les poids associés à la structure régionale. Les poids sont choisis de manière que $w_{ik} = 1$, si i et k sont contigus, et $w_{ik} = 0$, autrement. La loi de $\delta_\alpha \mid \sigma_\alpha^2$ est évidemment incorrecte, puisque nous pouvons ajouter n'importe quelle constante à tous les δ_{α_i} sans que $p(\delta_\alpha \mid \sigma_\alpha^2)$ soit affectée. Donc, nous devons imposer une contrainte pour nous assurer que le modèle est identifiable. Nous posons que $\sum_{i=1}^{m} \delta_{\alpha_i} = 0$ et attribuons à l'ordonnée à l'origine α une loi a priori uniforme sur l'ensemble de la droite réelle. Il n'est pas difficile de voir que cette procédure donne lieu à une densité de vraisemblance $(m-1)$-dimensionnelle appropriée. Voir Besag et Kooperang (1995) pour plus de renseignements.

La loi a priori conditionnelle de δ_{α_i}, sachant les effets δ_{α_k} des secteurs restants et l'hyperparamètre σ_α^2, est normale de moyenne et de variance données par :

$$E[\delta_{\alpha_i} \mid \delta_{\alpha_k},\ k \in \partial i,\ \sigma_\alpha^2] = \overline{\delta}_{\alpha_i}$$

$$\mathrm{Var}[\delta_{\alpha_i} \mid \delta_{\alpha_k},\ k \in \partial i,\ \sigma_\alpha^2] = \frac{\sigma_\alpha^2}{w_{i+}}$$

où $\overline{\delta}_{\alpha_i}$ désigne la moyenne arithmétique des δ_{α_j}, pour $k \in \partial i$ (les domaines contigus de i), et $w_{i+} = \sum_{k=1}^{m} w_{ik}$ est le nombre de municipalités voisines de i.

La figure 1 montre les densités démographiques des municipalités de São Paulo en 1991. Ces municipalités ont tendance à être concentrées géographiquement en fonction de classes de densité, ce qui donne à penser que l'application du modèle spatial peut être fructueuse.

3.4 Modélisation des variances d'échantillonnage

Comme nous utilisons des données provenant de deux sources différentes, il est logique de supposer que les variances d'échantillonnage varient au cours du temps. En outre, nous pouvons également considérer que les variances varient selon le domaine.

Dans le cas des années pour lesquelles des données sont fournies par la PNAD, nous supposons que les variances d'échantillonnage sont données par le modèle suivant :

$$\log(\sigma_{it}^2) = \eta_0 + \eta_1 . (1/n_i) \qquad (4)$$

avec n_i représentant le nombre de secteurs de dénombrement échantillonnés dans le i^{e} domaine. Ce modèle traduit l'espérance que la variance diminue à mesure que la taille de l'échantillon augmente.

Pour les années durant lesquelles les recensements ont été exécutés, nous supposons que σ_{it}^2 est connue et que $\log(\sigma_{it}^2) = \log(v_{it})$ où v_{it} est calculée de telle manière que l'erreur de couverture au recensement soit égale à 5 % pour tous les domaines. Cette hypothèse implique que, dans chaque domaine pour les années de recensement, la population réelle est comprise dans l'intervalle donné par la population observée au recensement plus ou moins 5 % de cette valeur. Par conséquent, pour les années de recensement, nous fixons l'écart-type à $\sigma_{it} = 0{,}05 * (y_{it}/2)$. Supposer que la variance est connue pour les années de recensement est un moyen d'accorder plus de poids aux données de recensement, puisque l'on pourrait s'attendre à ce qu'un recensement complet fournisse des renseignements plus fiables que des données d'enquête. Nous supposons que les paramètres η_0 et η_1 suivent des lois normales indépendantes : $\eta_k \sim N(\mu_{\eta_k}, \phi_{\eta_k})$; $k = 0{,}1$. Afin d'attribuer des priors vagues aux η, nous donnons, pour chaque prior, une valeur nulle à la moyenne et de grandes valeurs aux ϕ_η. Voir la section 4.2 pour plus de renseignements.

3.5 Sommaire des modèles

Les lois a priori des paramètres communs des modèles spatial et hiérarchique sont les mêmes que celles déjà décrites pour le premier. Les lois suivies par les effets spatiaux aléatoires sont spécifiées à la section 3.3. La variance σ_{it}^2 a été énoncée de la même façon dans le modèle spatial que dans le modèle hiérarchique. Le tableau 1 résume les modèles utilisés à la section 4. Pour simplifier, nous avons exécuté l'application en fixant $\phi = 1$ dans les deux modèles.

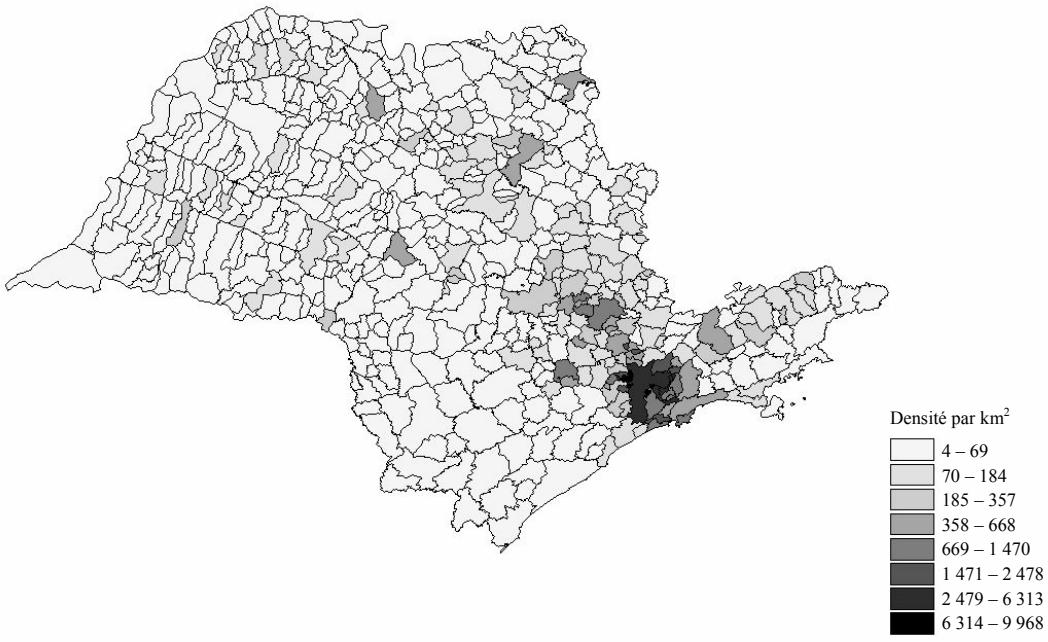

Figure 1 Densité de population des municipalités de São Paulo en 1991

3.6 Problèmes de calcul

Les lois a posteriori des paramètres des modèles proposés ne peuvent pas être exprimées sous une forme explicite. Il est donc nécessaire de recourir à des méthodes d'approximation numérique. Une option, souvent utilisée et facile à appliquer, consiste à produire des échantillons de ces lois en se servant de l'algorithme Monte Carlo par chaînes de Markov (MCMC). Puisque les lois conditionnelles complètes de tous les paramètres des modèles possèdent une forme explicite, sauf pour le vecteur $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)$, nous nous sommes servi de l'algorithme de l'échantillonneur de Gibbs avec une étape de l'algorithme d'acceptation-rejet pour l'échantillonnage à partir du vecteur $\boldsymbol{\gamma}$. Soit π_{it} la densité de population dans le i^{e} domaine à la période t. Les étapes qui suivent résument comment tirer des échantillons de la loi a posteriori de π_{it}:

1. Produire $\alpha_i^{(l)}, \beta^{(l)}, \gamma_i^{(l)}, \alpha^{(l)}, \gamma^{(l)}, \tau_\alpha^{2(l)}, \tau_\gamma^{2(l)}, \eta_0^{(l)}$ et $\eta_1^{(l)}$ pour $l = 1, ..., M$, où M est le nombre d'échantillons générés à partir des lois conditionnelles complètes de tous les paramètres du modèle, y compris les effets aléatoires;
2. calculer $\pi_{it}^{(l)} = \alpha_i^{(l)} + \beta^{(l)}\exp(\gamma_i^{(l)} t)$;

Durant l'ajustement des modèles que nous proposons, nous avons appliqué trois vérifications informelles, basées sur des techniques graphiques, pour évaluer la convergence. Elles consistent à observer l'histogramme, la trace et la fonction d'autocorrélation pour chacune des valeurs échantillonnées calculées. L'analyse de l'histogramme nous permet de repérer les écarts éventuels par rapport à la convergence, tel que la présence de modes multiples. La trace des chaînes multiples simulées en parallèle, chacune avec un point de départ différent et surdispersées par rapport à la loi cible, donne une idée grossière du comportement stationnaire quand les suites de valeurs ont tendance à osciller dans la même région. La représentation graphique de la fonction d'autocorrélation permet de déterminer si l'échantillon peut être considéré comme indépendant.

En plus de ces vérifications informelles, nous avons appliqué d'autres critères plus formels. Le critère introduit par Brooks et Gelman (1998) et implémenté dans WinBugs 1.4 (Spiegelhalter, Thomas, Best and Lunn 2004) permet de diagnostiquer si la dispersion à l'intérieur des chaînes est plus grande que la dispersion entre les chaînes. Considérons I chaînes parallèles et un paramètre d'intérêt λ. Soit λ_i^j la j^{e} valeur de la i^{e} chaîne, pour $i = 1, ..., K$ et $j = 1, ..., J$. Les variances entre les chaînes $\hat{B}$ et à l'intérieur des chaînes $\hat{W}$ sont alors données par

$$\hat{B} = J(K-1)^{-1}\sum_{i=1}^{K}(\bar{\lambda}_i - \bar{\lambda})^2$$

et

$$\hat{W} = \{K(J-1)\}^{-1}\sum_{i=1}^{K}\sum_{j=1}^{J}(\lambda_i^j - \bar{\lambda}_i)^2$$

où $\bar{\lambda}_i$ et $\bar{\lambda}$ sont, respectivement, la moyenne des observations de la chaîne i, $i = 1, ..., K$ et la moyenne globale. Sous convergence, ces KJ valeurs sont toutes tirées de la loi a posteriori de λ et la variance de λ peut être estimée de manière convergente par $\hat{B}$, $\hat{W}$ et la moyenne pondérée $\hat{\sigma}_\lambda^2 = (1 - 1/J)\,\hat{W} + (1/J)\,\hat{B}$.

Si les chaînes n'ont pas encore convergé, les valeurs initiales influenceront encore les trajectoires et $\hat{\sigma}_\lambda^2$ surestimera σ_λ^2 jusqu'à ce qu'un état stationnaire soit atteint. Par ailleurs, avant la convergence, $\hat{W}$ aura tendance à sous-estimer σ_λ^2. En suivant ce raisonnement, Brooks et Gelman (1998) ont proposé une approche graphique itérée qui est implémentée dans WinBugs 1.4. Elle permet de vérifier si i) la variance a posteriori pondérée estimée $\hat{\sigma}_\lambda^2$ et la variance à l'intérieur des chaînes $\hat{W}$ se stabilisent sous forme d'une fonction de J et ii) le facteur de réduction de la variance, $\hat{R} = \hat{\sigma}_\lambda^2/\hat{W}$, s'approche de 1.

Tableau 1
Sommaire des modèles utilisés

Modèle	Paramètres	Variance	Loi a priori
Hiérarchique	$\alpha_i = \alpha + \xi_{\alpha_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1(1/n_i)$,	$\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$
	β	pour les données d'enquête	$\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$
	$\gamma_i = \gamma + \xi_{\gamma_i}$	σ_{it}^2 est supposée connue pour les données de recensement	
Spatial	$\alpha_i = \alpha + \delta_{\alpha_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1(1/n_i)$	$\delta_{\alpha_i} \mid \delta_{\alpha,-i}, \tau_\alpha^2 \sim N(\bar{\delta}_{\alpha_i}, \tau_\alpha^2/w_{i+})$
	β	dans l'enquête	$\sum_{i=1}^{m}\delta_{\alpha_i} = 0$
	$\gamma_i = \gamma + \xi_{\gamma_i}$	σ_{it}^2 est supposée connue pour les données de recensement	$\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$
			$\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$

4. Application

Nous présentons maintenant deux applications de notre approche, la première en nous servant d'un ensemble de données simulées et la deuxième, de l'ensemble de données réelles qui a motivé la présente étude. L'étude par simulation a pour but de vérifier si les paramètres d'intérêt sont estimés correctement et de procéder à une analyse de sensibilité en ce qui concerne la forme des lois a priori utilisées pour ajuster le modèle.

4.1 Application aux données simulées

Nous avons exécuté une petite étude en simulation de l'ajustement des modèles hiérarchique et spatial présentés à la section 3. Nous avons fixé les hyperparamètres réels du modèle reliés à la courbe de croissance à $\alpha = 40$, $\beta = 25$, $\gamma = 0,05$. Donc, nous considérons une situation où la taille de la population double approximativement en 25 ans. Nous avons fixé les paramètres associés au modèle de la variance d'échantillonnage à $\eta_0 = 6,5$, $\eta_1 = 0,5$. Enfin, nous avons fixé les paramètres de précision à $\tau_\alpha^2 = 0,0001$ et $\tau_\gamma^2 = 400$, respectivement. Nous avons fixé les précisions τ_α^2 et τ_γ^2 de façon qu'elles concordent avec les échelles des quantités qu'elles mesurent respectivement. La variation relative de l'ordonnée à l'origine entre les domaines est plus importante que celle du paramètre de croissance, ce à quoi nous nous attendons dans les situations pratiques.

Puisqu'il est généralement reconnu que la forme des priors a plus d'incidence sur la composante des paramètres de variance que les paramètres fixes, nous avons ajusté le modèle aux données simulées en utilisant deux priors vagues pour les paramètres liés à la variance : pour l'écart-type, nous avons choisi la loi uniforme, qui est l'un des priors recommandés par Gelman (2006) pour les modèles hiérarchiques linéaires, et pour la précision, la loi gamma, fréquemment utilisée comme loi par défaut dans certains progiciels. Dans le premier cas, nous avons assigné $\sigma_\alpha \sim U(0, 1,000)$ et $\sigma_\gamma \sim U(0, 100)$, où $\sigma_\alpha = 1/\tau_\alpha$ et $\sigma_\gamma = 1/\tau_\gamma$. Dans le deuxième cas, nous avons considéré $\tau_\alpha^2 \sim G(0,001, 0,001)$ et $\tau_\gamma^2 \sim G(0,001, 0,001)$. Pour les autres paramètres, nous avons fixé $\alpha \sim U(-\infty, +\infty)$ pour le modèle spatial (voir la section 3.3 pour plus de précisions) et $\alpha \sim N(0, 10^6)$ pour le modèle hiérarchique. Pour les autres paramètres, nous avons fixé $\beta \sim N(0, 10^6)$, $\gamma \sim N(0, 10^2)$, $\eta_0 \sim N(0, 10^4)$ et $\eta_1 \sim N(0, 10^4)$ pour les deux modèles. Nous avons aussi étudié l'effet du nombre de petits domaines. Nous avons simulé des données distinctes provenant des modèles hiérarchique et spatial avec $m = 60$ et $m = 100$ domaines dans chaque cas. Pour chaque combinaison du nombre de domaines et du modèle employé, nous avons généré 200 ensembles de données. Par conséquent, nous avons simulé en tout 800 ensembles de données artificielles. La distribution des tailles d'échantillon dans les domaines est la même pour les ensembles de données simulés comptant 60 et 100 domaines. Le tableau 2 donne les fréquences relatives des tailles d'échantillon de petit domaine pour les deux ensembles de données simulés. Ces tailles d'échantillon sont fort semblables à celles provenant des données réelles qui sous-tendent cette étude par simulation. Le nombre de voisins utilisés dans le modèle spatial varie de 1 à 12, et chaque domaine possède en moyenne cinq voisins. Nous avons considéré une période totale de $n = 9$ années.

Tableau 2
Fréquences relatives des tailles d'échantillon de petit domaine pour les deux ensembles de données simulés

Taille d'échantillon	Fréquence relative
2	0,05
5	0,20
8	0,25
10	0,25
12	0,20
15	0,05

Afin d'éliminer la corrélation entre les chaînes, nous avons généré 20 000 échantillons après avoir écarté les 10 000 premiers. Nous ne dégageons aucune preuve de non-convergence des paramètres des modèles hiérarchique et spatial. Une analyse minutieuse de certaines données de sortie provenant des échantillons MCMC pour certains ensembles de simulation donne à penser que la convergence a été atteinte pour tous les paramètres du modèle. Nous avons évalué les propriétés statistiques des estimations de la densité de population (π_{it}) en examinant l'erreur relative absolue moyenne (ERA) des estimations et l'erreur quadratique moyenne (EQM), respectivement données par :

$$\mathrm{ERA}_{i,t} = \frac{1}{200}\sum_{l=1}^{200}\frac{|\hat{\pi}_{i,t}^{(l)} - \pi_{i,t}^{(l)}|}{\pi_{i,t}^{(l)}}$$

et

$$\mathrm{EQM}_{i,t} = \frac{1}{200}\sum_{l=1}^{200}(\hat{\pi}_{i,t}^{(l)} - \pi_{i,t}^{(l)})^2,$$

$i = 1, ..., m$, $t = 1, ..., n$. Si l'on s'en tient aux valeurs de l'ERA, la variation est faible. Pour les deux modèles ajustés et les deux tailles d'échantillon de petit domaine testées, les valeurs de l'ERA sont de l'ordre de 1,5 %.

Le tableau 3 résume les valeurs de l'EQM obtenues dans les simulations exécutées sous les modèles spatial et hiérarchique avec 60 et 100 domaines et en assignant respectivement un prior gamma et un prior uniforme à la précision et à l'écart-type des paramètres associés à la variance. L'examen du tableau 3 révèle que les EQM ne sont pas affectées par l'utilisation de divers priors vagues. Il convient de souligner que l'accroissement du nombres de domaines, pour passer de 60 à 100, réduit légèrement, de 6 %, la médiane de l'EQM pour le modèle spatial. Par contre, dans le cas du modèle hiérarchique, la diminution est d'environ 13 %.

Tableau 3
Sommaire de la distribution de l'erreur quadratique moyenne pour les modèles spatial et hiérarchique

Modèle	N^bre de domaines	Prior gamma			Prior uniforme		
		1^er quart.	Médiane	3^e quart.	1^st quart.	Médiane	3^e quart.
Spatial	60	0,398	1,741	3,574	0,394	1,737	3,595
	100	0,525	1,637	3,538	0,524	1,641	3,517
Hiérarchique	60	0,542	2,218	6,262	0,646	2,223	6,278
	100	0,594	1,959	5,593	0,596	1,960	5,619

Nous avons également examiné le pourcentage de couverture des intervalles de crédibilité à 95 % nominaux. Les résultats sont présentés au tableau 4. En ce qui concerne la présente étude par simulation, les intervalles pour les paramètres d'intérêt possèdent en général les pourcentages de couverture corrects pour les deux modèles étudiés et ces résultats ne varient pas selon que l'on utilise 60 ou 100 domaines. Cependant, si le nombre de domaines est petit, nous pourrions rencontrer des problèmes de convergence à moins de choisir des priors plus stricts pour les hyper-paramètres. L'étude par simulation révèle que la prédiction de la population n'est pas affectée par les formes des priors vagues assignés à la variance de l'ordonnée à l'origine.

Tableau 4
Taux de couverture des intervalles de crédibilité à 95 % nominaux pour les densités de population

Modèle	N^bre de domaines	Couverture du prior gamma (%)	Couverture du prior uniforme (%)
Spatial	60	96	96
	100	96	96
Hiérarchique	60	94	94
	100	95	95

Nous avons analysé la qualité de l'ajustement du modèle quand le modèle correct ou le modèle incorrect est ajusté aux données générées au moyen d'un modèle. La figure 2 donne l'erreur quadratique moyenne pour les situations suivantes : a) données générées au moyen du modèle spatial auxquelles sont ajustés les modèles spatial et hiérarchique, et b) données générées au moyen du modèle hiérarchique auxquelles sont ajustés les modèles spatial et hiérarchique. Puisque la forme des priors assignés aux paramètres associés à la variance n'ont pas d'incidence sur l'inférence, nous utilisons des priors uniformes pour les deux modèles. Les mesures de l'ERA sont présentées à la figure 3.

La figure 2 révèle que, si les données sont générées au moyen du modèle le plus simple (hiérarchique), l'efficacité des procédures d'estimation complexe (modèle spatial) n'est pas réduite de manière appréciable. Par ailleurs, quand les données sont générées au moyen du modèle plus complexe (spatial), certaines propriétés de l'estimateur plus simple (hiérarchique) deviennent moins bonnes. Cependant, ce résultat ne tient pas pour les mesures de l'ERA. La figure 3 montre que le fait d'ajuster le modèle qui n'a pas été utilisé pour générer les données donne lieu à un accroissement appréciable du biais relatif. Comme il faut s'y attendre, l'ajustement du modèle et les tests diagnostiques jouent un rôle crucial dans l'obtention d'une prédiction appropriée pour la population des petits domaines.

4.2 Application aux données réelles

Les ensembles de données de la PNAD pour la période allant de 1992 à 1999 (sauf 1994 et 1996) et les données des recensements de la population de 1991, 1996 et 2001 ont été utilisées dans notre application. Les domaines d'intérêt sont toutes les municipalités de l'État de São Paulo, soit en tout 572 domaines, parmi lesquels 111 ont été échantillonnés pour l'enquête PNAD. La figure 4 montre les domaines échantillonnés pour la PNAD, classés selon la définition d'échantillonnage : domaines appartenant aux régions métropolitaines et domaines autoreprésentés (échantillonnés avec une probabilité de 1) et domaines non autoreprésentés. Il convient de mentionner que les périodes de référence des recensements et de la PNAD diffèrent. Nous avons fixé $t = 0$ pour le recensement de 1991. Donc, les valeurs de t pour les données fournies par la PNAD sont égales au nombre d'années entre la période de référence du recensement de 1991 et l'année du cycle en question de la PNAD. Par exemple, une donnée d'enquête fournie par la PNAD 18 mois après le recensement de 1991 correspond à $t = 1,5$.

La figure 5 donne le coefficient de variation estimé de l'estimateur direct selon la taille d'échantillon de domaine. Ces estimations sont fondées sur les données de la PNAD. On peut voir que les coefficients de variation varient considérablement selon le domaine et ont tendance à diminuer à mesure qu'augmente la taille d'échantillon. Les valeurs élevées de ces coefficients montrent qu'il est difficile d'utiliser uniquement l'estimateur direct pour produire des estimations pour les municipalités. En outre, nous ne pouvons faire aucune prédiction pour les domaines non échantillonnés en utilisant uniquement les estimateurs directs.

4.3 Spécification des lois a priori

Pour assigner la moyenne des lois a priori normale des paramètres α, β et γ, reliés à l'évolution de la population, nous avons d'abord développé la fonction $\alpha + \beta\exp(\gamma t)$ autour de zéro par un développement en série de Taylor jusqu'au deuxième ordre, puis nous avons égalé l'expression résultante aux valeurs de la densité moyenne aux recensements de 1991 et 2000, ainsi qu'au dénombrement de la population de 1996. En l'absence d'information a priori, nous avons considéré une valeur raisonnablement grande (10^6) pour les variances a priori de α, β et γ. Donc, nous avons posé $\alpha \sim U(-\infty, +\infty)$ (voir la section 3.3 pour plus de renseignements) pour le modèle spatial et $\alpha \sim N(370, 10^6)$ pour le modèle hiérarchique, ainsi que $\beta \sim N(726, 10^6)$, $\gamma \sim N(0{,}04, 10^6)$ pour les deux modèles. Cet ajustement a pour but d'obtenir une valeur raisonnable, mais essentiellement vague, des moyennes a priori. En ce qui concerne les précisions et η_0, η_1, nous avons attribué des priors relativement vagues : $\tau_\alpha^2 \sim \text{Ga}(0{,}001,\ 0{,}001)$, $\tau_\gamma^2 \sim \text{Ga}(0{,}001,\ 0{,}001)$, $\eta_0 \sim N(0,\ 10^6)$ et $\eta_1 \sim N(0,\ 10^6)$.

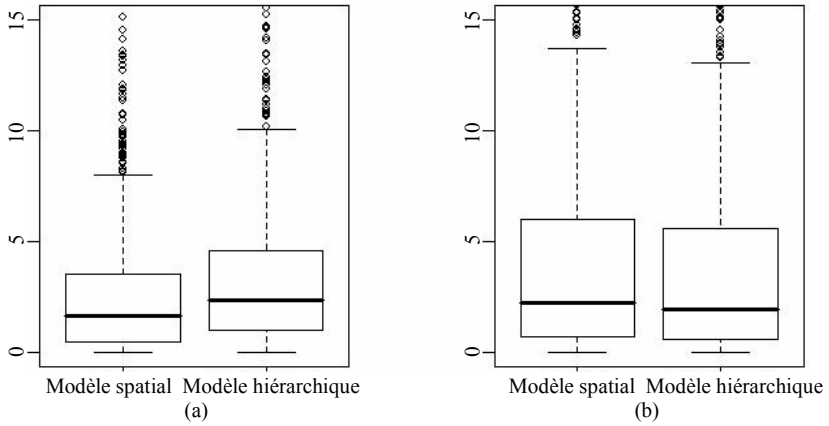

Figure 2 Boîtes à moustaches de l'erreur quadratique moyenne (EQM) pour les cas suivants : a) données générées au moyen du modèle spatial auxquelles sont ajustés respectivement les modèles spatial et hiérarchique et b) données générées au moyen du modèle hiérarchique auxquelles sont ajustés respectivement les modèles spatial et hiérarchique

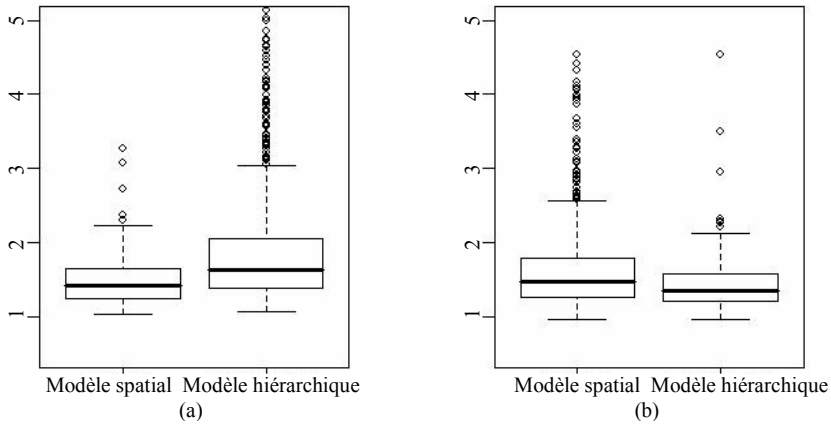

Figure 3 Boîtes à moustaches de l'erreur relative absolue (ERA) pour les cas suivants : a) données générées au moyen du modèle spatial auxquelles sont ajustés respectivement les modèles spatial et hiérarchique et b) données générées au moyen du modèle hiérarchique auxquelles sont ajustés respectivement les modèles spatial et hiérarchique

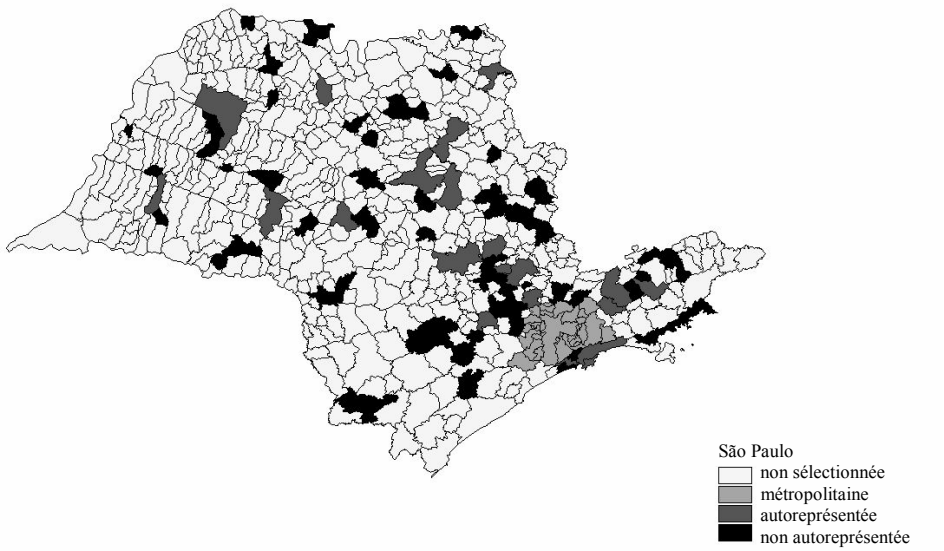

Figure 4 Municipalités de São Paulo échantillonnées pour la PNAD, classifiées selon la définition d'échantillonnage

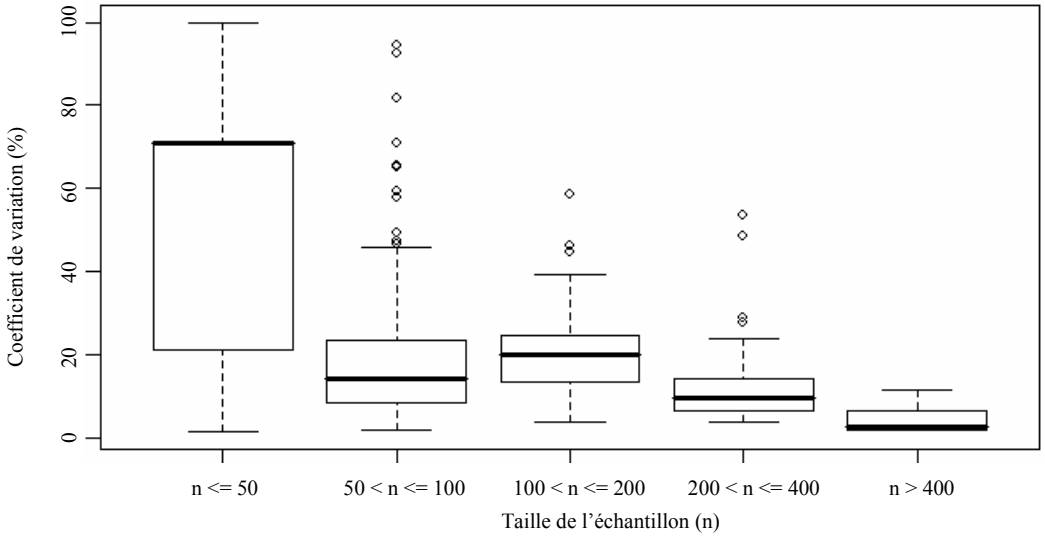

Figure 5 Boîtes à moustaches des coefficients de variation des estimations de population directe

4.4 Certains résultats

Nous avons généré 20 000 échantillons après avoir écarté les 5 000 premiers. Nous ne dégageons aucune preuve de non-convergence des paramètres des modèles hiérarchique et spatial. Une analyse minutieuse des données de sortie MCMC donne à penser que la convergence a été atteinte pour tous les paramètres des modèles. Nous résumons les résultats obtenus par ajustement du modèle hiérarchique (3) aux données fournies par l'enquête PNAD. Nous avons utilisé les moyennes a posteriori des paramètres du modèle comme estimations ponctuelles. Le tableau 5 donne ces estimations ainsi que les racines carrées respectives de la variance a posteriori. L'examen du tableau montre que l'estimation de η_1 est significativement positive, ce qui concorde avec le résultat attendu selon l'équation 4 : plus la taille d'échantillon est grande, plus σ_{it}^2 est petite.

Tableau 5
Résumé des lois a posteriori des paramètres du modèle (2)

Paramètre	Moyenne a posteriori	É.-T. a posteriori
α	892,500	202,000
β	105,700	1,278
γ	0,072	0,008
η_0	10,620	0,133
η_1	3,185	0,484
τ_α^2	2,174E-7	2,961E-8
τ_γ^2	139,000	19,560

La figure 6 montre que les moyennes a posteriori des paramètres α et γ qui indicent le modèle hiérarchique semble être distribuées spatialement. Les paramètres des domaines voisins paraissent plus semblables que ceux des domaines éloignés, ce qui suggère d'appliquer le modèle spatial.

4.5 Choix du modèle

Nous nous sommes servi de la mesure de l'écart par rapport à la prédiction attendue (EPA) (Gelfand et Ghosh 1998) pour choisir le modèle le plus approprié. La mesure de l'EPA est égale à la somme de deux termes. Le premier, désigné par G, peut être interprété comme une mesure de la qualité de l'ajustement et le deuxième, désigné par P, est un terme de pénalité pour les modèles sous-ajustés ainsi que surajustés. Les expressions pour G et P sont données, respectivement, par $G=\sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it}-E(y_{it}^{\text{rep}}|M))^2$ et $P=\sum_{i=1}^{m}\sum_{t=1}^{n}V(y_{it}^{\text{rep}}|M)$, où les espérances et les variances sont calculées par rapport à la loi prédictive a posteriori associée à une future observation (y_{it}^{rep}) de y_{it} générée sous le modèle hypothétique (M). Selon ce critère, le modèle est d'autant meilleur que la valeur est petite. Comme le montre le tableau 6, le critère EPA favorise légèrement le modèle spatial.

4.6 Analyse des résultats

Le niveau le plus agrégé pour lequel la PNAD fournit des estimations précises est la région métropolitaine, qui correspond à un ensemble de municipalités contigues. Afin de valider les résultats obtenus au moyen du modèle spatial, nous avons comparé des estimations de population pour la grande région métropolitaine de São Paulo aux projections statistiques officielles. La loi a posteriori de $\mu_t=\sum_{i=1}^{r}\pi_{it}*A_i$ s'obtient facilement en ajoutant $\mu_t^{(l)}=\sum_{i=1}^{r}\pi_{it}^{(l)}*A_i$ à l'algorithme MCMC, où μ_t représente la population totale de la région métropolitaine à la période t et r est le nombre de municipalités appartenant à cette région métropolitaine.

Tableau 6
Mesures pour le choix des modèles de densité démographique

Modèle	*G*	*P*	EPA
Hiérarchique	1,37E+09	6,14E+09	7,51E+09
Spatial	1,05E+09	6,19E+09	7,24E+09

À la figure 7, nous comparons les estimations de population (μ_t) de la région métropolitaine de São Paulo obtenues au moyen du modèle spatial aux statistiques officielles. Les courbes en trait plein représentent les limites des intervalles de crédibilité à 95 % de μ_t, tandis que la courbe en trait interrompu représente les estimations ponctuelles respectives. Le symbole (+) représente les statistiques officielles observées. Soulignons que certaines projections statistiques officielles se situent en dehors de la limite inférieure de crédibilité (y compris pour le recensement de 1991). Une enquête plus approfondie devrait donc être faite afin de découvrir les raisons de ces divergences. Cependant, si nous faisons la comparaison au niveau des municipalités, la conclusion générale est que les prédictions du modèle et les statistiques officielles concordent raisonnablement. Les intervalles de crédibilité à 95 % contiennent 92,4 % des projections statistiques officielles. La moyenne de l'erreur relative absolue (ERA) entre la densité de population estimée et les projections statistiques officielles est de 3 %. Ces mesures de l'ERA sont, en moyenne, presque les mêmes pour les municipalités sélectionnées et non sélectionnées.

À la figure 8 nous comparons les estimations ponctuelles des tailles de population (μ_{it}) aux projections statistiques officielles et aux tailles de population officielles selon le recensement pour une municipalité échantillonnée. La méthode de calcul des projections officielles s'appuie sur l'hypothèse qu'un ensemble de petits domaines et un domaine plus grand, qui les contient, ont la même courbe de croissance démographique. La population du domaine plus grand est projetée par la méthode des composantes, puis est répartie proportionnellement entre les petits domaines. La méthode des composantes s'appuie sur des données provenant du recensement le plus récent, ainsi sur les nombres de naissances et de décès et les chiffres de migration nette tirés des dossiers administratifs. La méthode des composantes consiste à projeter la population pour une période t en ajoutant à la population durant une période antérieure le nombre de naissances et le chiffre de migration nette, et en soutrayant le nombre de décès survenus durant le même intervalle de temps.

Les courbes en trait plein représentent les limites des intervalles de crédibilité à 95 % pour μ_{it} obtenues au moyen du modèle spatial, tandis que la droite en trait interrompu montre les moyennes a posteriori respectives. Le symbole (+) représente la projection de population officielle pour la période intercensitaire et la population observée durant les années de recensement. Il convient de souligner que les estimations ponctuelles sont relativement proches des statistiques projetées officielles et du chiffre de population obtenu l'année du recensement. L'utilisation du modèle proposé semble donc produire des estimations fiables au niveau de la municipalité, avec l'avantage supplémentaire de fournir une mesure de l'erreur respective.

Nous analysons aussi les estimations obtenues pour certaines municipalités non échantillonnées dans la PNAD. La figure 9 donne les prédictions du modèle, les intervalles de crédibilité à 95 %, les statistiques projetées officielles et les valeurs de population observées aux recensements pour une municipalité non échantillonnée (+). Nous voyons que

les prédictions obtenues au moyen du modèle spatial concordent raisonnablement avec les chiffres officiels.

5. Remarques finales

Le modèle utilisé dans le présent article permet de dégager la tendance de croissance de la population des municipalités. Des estimations raisonnables des populations municipales sont obtenues pour les années pour lesquelles il existe des données d'enquête, de même que pour celles pour lesquelles on dispose de données de recensement. Les estimations ponctuelles ont une bonne précision et concordent raisonnablement avec les estimations obtenues pour les domaines plus grands en utilisant d'autres techniques. L'information antérieure peut être mise à jour aussitôt que des estimations fondées sur un nouveau recensement ou une nouvelle enquête deviennent disponibles. De surcroît, l'approche proposée fournit la densité de probabilité de la quantité d'intérêt, ce qui facilite le processus de prise de décision.

D'autres travaux devraient être effectués afin de tenir compte de l'autocorrélation des paramètres d'intérêt au cours du temps. Les renseignements supplémentaires sur les estimations de la variance d'échantillonnage des estimateurs directs pourraient également être considérés comme des données additionnelles. L'hypothèse selon laquelle l'erreur de couverture au recensement est distribuée symétriquement autour de zéro pourrait être relâchée en lui appliquant une distribution non symétrique. Il faut néanmoins pour cela bien connaître la forme de la distribution, ce qui pourrait être difficile en pratique.

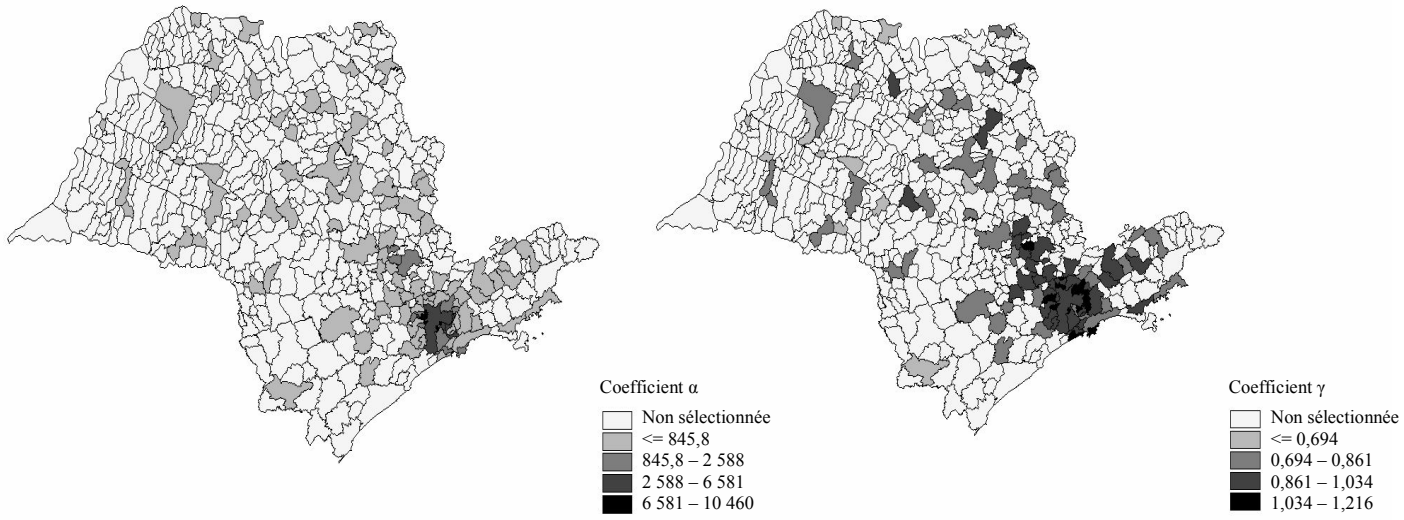

Figure 6 Moyennes a posteriori des paramètres α et γ obtenus au moyen du modèle hiérarchique

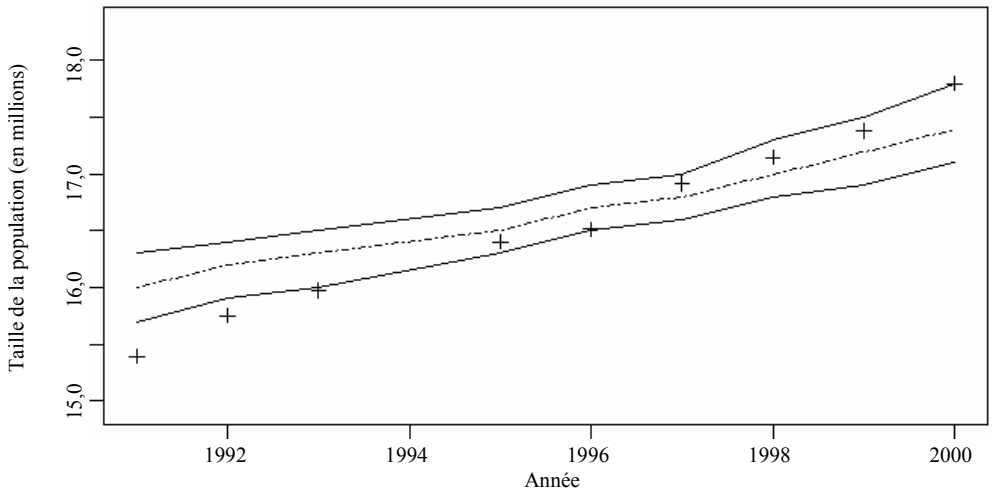

Figure 7 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour la région métropolitaine

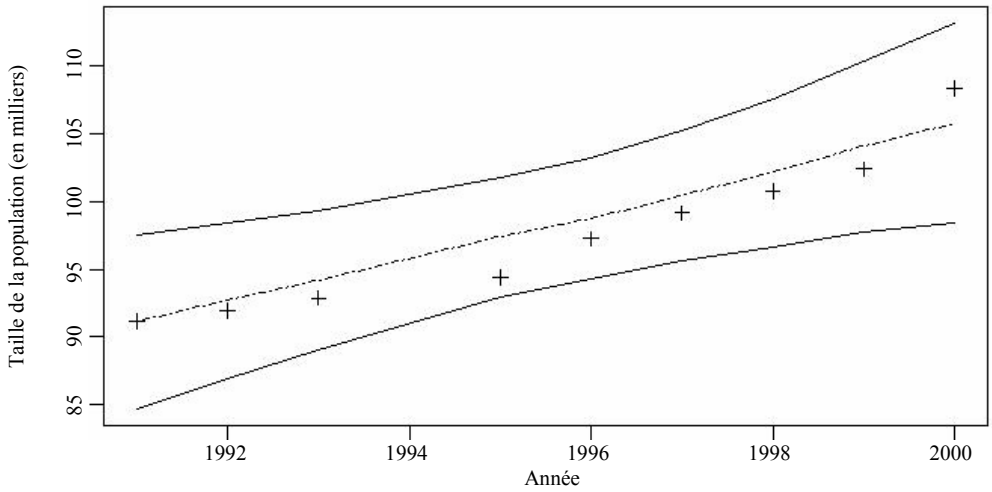

Figure 8 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour une municipalité échantillonnée

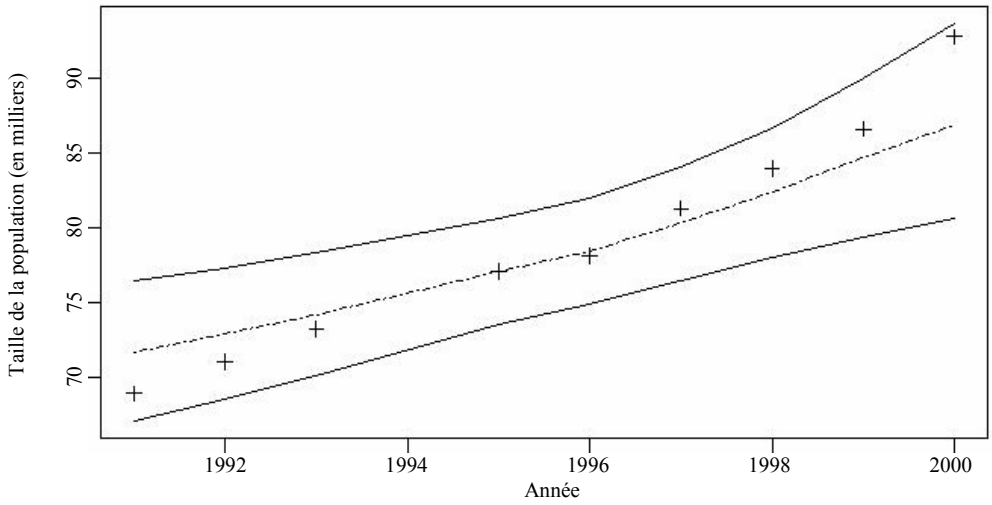

Figure 9 Comparaison entre les tailles de population prédites par le modèle spatial et les statistiques officielles (+) pour une municipalité non échantillonnée

Remerciements

Les auteurs remercient un rédacteur associé et deux examinateurs de leurs commentaires et suggestions constructifs. Les travaux de Fernando Moura et Helio Migon ont été financés en partie par une subvention de recherche du Conseil national brésilien pour le développement de la science et de la technologie (CNPq).

Bibliographie

Besag, J., et Kooperang, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

Brooks, S.P., et Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 4, 434-455.

Gelfand, A.E., et Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1, 1-11.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 3, 515-533.

Migon, H., et Gamerman, D. (1993). Generalized exponential growth models: A bayesian approach. *Journal of Forecasting*, 12, 573-584.

Mollié, A. (1996). Bayesian mapping of disease. Dans : Gilks, W.R.; Richardson, S.; Spieghalter, D.J. Markov Chain Monte Carlo in Practice. New York : Chapman & Hall, 359-379.

Moura, F.A.S., et Migon, H.S. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modeling*: *An International Journal*, 2, 3, 183-201.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 1, 125-143.

Pfeffermann, D., Moura, F.A.S. et Silva, P.L.N. (2006). Multi-level modeling under informative sampling. *Biometrika*, 93, 4, 943-959.

Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of American Statistical Association*, 102, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

Souza, D.F. (2004). Estimação de População em Nível Municipal via Modelos Hierárquicos e Espaciais. Unpublished master's dissertation. Universidade Federal do Rio de Janeiro.

Spiegelharter, D.J., Thomas, A., Best, N. et Lunn, D. (2004). WinBUGS User Manual Version 1.4. MRC Biostatistics Unit, Cambridge.