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

Article

Utilisation de modèles multivariés pour l'estimation sur petits domaines du nombre de recrues dans les entreprises

par Maria Rosaria Ferrante et Carlo Trivisano

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 185-195
Statistique Canada, N° 12-001-X au catalogue

Utilisation de modèles multivariés pour l'estimation sur petits domaines du nombre de recrues dans les entreprises

Maria Rosaria Ferrante et Carlo Trivisano [1]

Résumé

Le nombre de recrues dans les entreprises des zones locales de marché du travail est un important indicateur de la réorganisation des processus de production locaux. En Italie, ce paramètre peut être estimé au moyen des données de l'Enquête Excelsior, bien que celle-ci ne fournisse pas d'estimations fiables pour les domaines d'intérêt. Dans le présent article, nous proposons une méthode d'estimation sur petits domaines multivariée appliquée à des données de comptage et basée sur la loi multivariée Poisson-Log-normale. Cette méthode servira à estimer le nombre de personnes recrutées par les entreprises pour remplacer les employés qui quittent ainsi que pour doter de nouveaux postes. Dans le cadre de l'estimation sur petits domaines, on suppose habituellement que les variances et les covariances d'échantillonnage sont connues. Cependant, ces dernières, de même que les estimations ponctuelles directes, sont instables. Étant donné la rareté du phénomène que nous analysons, les dénombrements dans certains domaines sont nuls, ce qui produit des estimations nulles des covariances des erreurs d'échantillonnage. Afin de tenir compte de la variabilité supplémentaire due à la matrice de covariance d'échantillonnage estimée et de résoudre le problème des variances et covariances insensées dans certains domaines, nous proposons une approche « intégrée » suivant laquelle nous modélisons conjointement les paramètres d'intérêt et les matrices de covariance des erreurs d'échantillonnage. Nous suggérons une solution de nouveau fondée sur la loi Poisson-Log-normale pour lisser les variances et les covariances. Les résultats que nous obtenons sont encourageants : le modèle d'estimation sur petits domaines proposé donne de meilleurs résultats que le modèle d'estimation sur petits domaines fondé sur la loi multivariée normale-normale (MNN) et il rend possible une augmentation non négligeable de l'efficacité.

Mots clés : Loi multivariée Poisson-Log-normale ; dénombrements nuls ; fonction de variance généralisée ; modèles hiérarchiques bayésiens.

1. Introduction

Le nombre de personnes recrutées par les entreprises pendant une période déterminée peut être considéré comme un indicateur clé des changements en cours dans le système économique. Afin de mettre en relief la dynamique de la demande de main-d'œuvre locale, nous considérons le nombre de personnes recrutées par les entreprises dans les zones locales de marché de travail (LLMA pour *Local Labour Market Areas*), ces dernières étant regroupées selon i) la spécialisation de production, ii) la catégorie de taille des entreprises et iii) le secteur industriel. Les domaines sont définis par recoupement de ces trois variables. Afin de mettre en relief les signes de réorganisation du processus de production, nous nous concentrons sur le nombre de « recrues remplaçant des employés quittant l'entreprise (recrues de substitution – RS) » et de « recrues occupant de nouveaux postes (nouvelles recrues – NR) ». En Italie, l'information au sujet des personnes recrutées par les entreprises est recueillie dans le cadre de l'Enquête Excelsior coparrainée par l'Union des chambres de commerce italiennes (UNIONCAMERE), le ministère du Travail et l'Union européenne. Malheureusement, cette enquête ne fournit pas d'estimations fiables du nombre de personnes recrutées par les entreprises pour tous les domaines à cause de la petite taille d'échantillon pour les petits domaines. Par conséquent, une méthode d'estimation sur petits domaines (EPD) a été adoptée afin d'obtenir des estimations dont le degré de variabilité est acceptable.

Dans le présent article, nous proposons une approche d'estimation sur petits domaines pour l'estimation de dénombrements. En raison de contraintes liées aux données, nous adoptons un modèle agrégé au niveau du domaine.

Puisque nous cherchons à estimer les nombres de RS et de NR, nous adoptons un modèle d'estimation sur petits domaines multivarié qui emprunte de l'information non seulement aux domaines, mais aussi aux corrélations entre les valeurs réelles des nombres de NR et de RS. Afin d'estimer le revenu médian de groupes de familles de diverses tailles, Fay (1987) a proposé un modèle de régression multivariée dans un contexte bayésien empirique. Des approches multivariées d'estimation sur petits domaines ont également été élaborées par Ghosh, Nangia et Kim (1996), par Datta, Fay et Ghosh (1991), par Datta, Ghosh, Nangia et Natarajan (1996) et par Datta, Lahiri, Maiti et Lu (1999) pour des données continues dans le cadre des modèles hiérarchiques transversaux et chronologiques. Fabrizi, Ferrante et Pacei (2005, 2008) ont adopté des modèles multivariés au niveau du domaine pour estimer un vecteur de paramètres continus de la pauvreté. Comme dans

1. Maria Rosaria Ferrante, Département de statistique, Université de Bologne, Italie. Courriel : maria.ferrante@unibo.it ; Carlo Trivisano, Département de statistique, Université de Bologne, Italie. Courriel : carlo.trivisano@unibo.it.

le modèle univarié de Fay-Herriot (Fay et Herriot 1979), tous les articles susmentionnés reposent sur l'hypothèse de l'utilisation de l'échantillonnage normal dans les petits domaines et de modèles de lien.

Puisque les corrélations d'échantillonnage entre les estimateurs des nombres de RS et de NR sont principalement négatives, nous proposons un modèle d'estimation sur petits domaines basé sur la loi multivariée Poisson-Log-normale (MPLN). Contrairement à d'autres lois multivariées pour les données de comptage proposées dans la littérature, cette loi particulière permet des corrélations non contraintes, c'est-à-dire positives ainsi que négatives (Aitchison et Ho 1989).

Nous traitons également de la stabilité des estimateurs des variances et covariances des erreurs d'échantillonnage. Une estimation approximativement sans biais de la variance des estimateurs directs est habituellement disponible dans le contexte de l'estimation sur petits domaines. Cependant, dans le cas des modèles au niveau du domaine, il est habituel de supposer que la variance d'échantillonnage est connue et est égale à son estimation (Rao 2003, page 76). Cette hypothèse est habituellement énoncée et largement acceptée dans le cas des grands échantillons, tandis que l'estimateur de variance ainsi que les estimateurs ponctuels directs souffrent d'instabilité dans le cas des petits échantillons. En guise de solution partielle, les estimations de la variance d'échantillonnage sont souvent lissées suivant l'approche des fonctions de variance généralisées (FVG) (Wolter 1985). Dans You, Rao et Gambino (2003), les variances et covariances d'échantillonnage ont été lissées sur les domaines et au cours du temps. Afin de tenir compte de la variabilité supplémentaire associée aux variances d'échantillonnage estimées, Arora et Lahiri (1997) ont proposé une approche de lissage hiérarchique bayésienne (HB) intégrée pour les données continues. Voir You et Chapman (2006), Liu, Lahiri et Kalton (2007) et You (2008) pour diverses extensions des travaux d'Arora et Lahiri (1997).

La rareté des personnes recrutées dans certains domaines pose un problème supplémentaire qui est lié à l'instabilité des estimateurs des variances et covariances des erreurs d'échantillonnage. Quand les estimations directes de RS ou de NR (ou des deux) sont nulles, les variances et covariances estimées des erreurs d'échantillonnage sont également nulles. Notons que l'observation de variances estimées nulles ne signifie pas nécessairement que les estimations ont un haut degré d'exactitude. Cette question a été soulevée dans le cas de problèmes antérieurs d'estimation sur petits domaines (par exemple Elazar 2004 ; Chattopadhyay, Lahiri, Larsen et Reimnitz, 1999). Chen (2001) a proposé une modélisation hiérarchique au niveau de l'unité pour traiter le problème. De surcroît, certaines études (Cohen 2000) s'appuient sur la transformation logarithmique des estimations directes de la moyenne (ou du total) des données de comptage afin d'adopter un modèle d'estimation sur petits domaines linéaire, en écartant simplement les estimations nulles. Bien que cette solution permette de contourner le problème des « variances nulles », elle produit aussi des estimations biaisées et néglige une partie de l'échantillon.

Afin de traiter l'instabilité des estimateurs des variances et covariances, ainsi que le problème des variances d'échantillonnage estimées nulles, nous proposons une approche « intégrée » dans l'esprit de celle proposée par Arora et Lahiri (1997), Liu et coll. (2007) et You (2008). Dans un cadre hiérarchique bayésien, nous modélisons conjointement les paramètres d'intérêt et les matrices de covariance des erreurs d'échantillonnage en adoptant une solution de lissage des covariances basée une fois de plus sur le modèle de mélange Poisson-Log-normale.

Le plan de l'article est le suivant. À la section 2, nous décrivons l'ensemble de données utilisé, et à la section 3, nous présentons l'estimation directe de domaine et ses variances et covariances d'erreurs d'échantillonnage. À la section 4, nous décrivons le modèle d'estimation sur petits domaines multivarié que nous proposons pour estimer les dénombrements, ainsi que la solution recommandée pour surmonter le problème d'instabilité des estimateurs des variances et covariances d'erreurs d'échantillonnage en présence de dénombrements nuls. À la section 5, nous présentons les résultats obtenus en mesurant les propriétés du modèle retenu d'estimation sur petits domaines. Des détails sur la loi multivariée Poisson-Log-normale sont fournis en annexe.

2. L'Enquête Excelsior

L'Enquête Excelsior est, en Italie, l'une des sources statistiques les plus complètes de données sur la demande de main-d'œuvre, fournissant des estimations du nombre de personnes recrutées par les entreprises italiennes. Chaque année, un échantillon aléatoire simple stratifié d'environ 100 000 entreprises ayant au moins un employé sont contactées afin de leur demander le nombre de personnes qu'elles prévoient embaucher dans le court terme. Les facteurs utilisés pour la stratification sont le secteur industriel et la catégorie de taille de l'entreprise. La répartition de l'échantillon entre les strates satisfait à une contrainte imposée à la valeur maximale de l'erreur-type estimée pour un seuil de signification de 95 % (Baldi, Bellisai, Fivizzani et Sorrentino, 2007). En étant axée sur le niveau de détail géographique local, l'enquête est conçue pour produire des estimations fiables pour les provinces administratives (NUTS3, conformément à la Nomenclature des unités territoriales statistiques, à l'adresse http://europa.eu.int/comm/eurostat/ramon/nuts). Cette unité géographique, choisie sur la base de critères administratifs, ne semble pas être le choix idéal pour analyser la dynamique de la demande locale de main-d'œuvre. Afin d'apporter des éclaircissements sur les signes de réorganisation

du processus de production locale, une meilleure subdivision territoriale serait celle des zones locales de marché du travail (*Local Labour Market Areas* ou LLMA), conformément à la définition de l'OCDE. Les LLMA sont des groupes de municipalités dans lesquelles les conditions du marché du travail sont les mêmes (pour l'emplacement des LLMA en Italie, voir Sforzi 1991). En Italie, selon la stratégie proposée par Sforzi et Lorenzini (2002) et adoptée par l'Institut italien de statistique (ISTAT), certaines LLMA sont appelées « districts industriels » (DI). Les DI sont des systèmes de production définis géographiquement caractérisés par une spécialisation dominante. Au cours des années 1990, ces districts étaient considérés comme le principal aiguillon de la croissance du système économique italien (Becattini 1992).

L'estimation du nombre de recrues de substitution et de nouvelles recrues dans les entreprises exploitées à l'intérieur et à l'extérieur des DI peut nous aider à vérifier si les districts industriels sont encore une source de dynamisme de l'économie italienne dans son ensemble. Afin de faire référence aux types de districts industriels, nous les regroupons en fonction de leur spécialisation de production. De même. Nous classons les LLMA non considérées comme des DI en fonction de leur vocation économique (les LLMA peuvent être caractérisées par une activité manufacturière particulière, une région touristique, une ville, *etc.*). En outre, sur le plan économique, la comparaison entre les entreprises situées dans les DI et celles non situées dans les DI est logique si le secteur industriel et la taille des entreprises sont également pris en compte. Enfin, comme nous l'avons déjà mentionné, les domaines d'intérêt sont définis par recoupement : i) des groupes de LLMA obtenus en fonction de leur spécialisation de production, ii) du secteur industriel des entreprises et iii) de la taille de l'entreprise.

Le présent article porte sur le secteur de la fabrication qui caractérise l'activité économique des districts industriels. L'analyse est limitée à deux régions de l'Italie contenant une grande quantité de districts industriels, à savoir la Toscane et l'Émilie-Romagne, et aux entreprises ayant moins de 100 employés (puisque des recensements sont effectués pour les autres catégories de taille). La population cible comprend 54 089 entreprises employant, en tout, 809 059 personnes.

3. Estimations directes

Le tableau 1 donne des renseignements sur les catégories qui définissent les 208 domaines d'intérêt. Notons que le nombre de domaines est inférieur à celui prévu en raison de l'absence d'un certain nombre de ces domaines dans la population. Les domaines ne sont pas planifiés, puisqu'ils sont formés en regroupant les LLMA contenues dans une même strate planifiée. Par souci de simplicité, dans la suite de l'exposé, nous évitons d'utiliser dans la mesure du possible l'indice inférieur de strate.

Soit θ_{i1} et θ_{i2} les nombres réels de NR et de RS pour le domaine i ($i = 1, ..., 208$), respectivement. Nous définissons d'abord un estimateur direct de θ_{ij} ($i = 1, ..., 208$; $j = 1, 2$). Soit y_{ijl} la réponse de la l^{e} unité relative à la j^{e} variable dans le i^{e} domaine ($l = 1, ..., n_i$, où n_i est la taille d'échantillon dans le domaine i; $i = 1, ..., 208$; $j = 1, 2$). En tant qu'estimateur (direct) fondé sur le plan de sondage, nous utilisons un estimateur par le ratio au niveau du domaine défini comme étant $\hat{\theta}_{ij} = \sum_{l=1}^{n_i} y_{ijl}/(n_i/N_i)\ N_i/\hat{N}_i$, où N_i et n_i sont, respectivement, la taille de population et la taille de l'échantillon se rapportant au domaine i, et $\hat{N}_i = n_i / n_{t\ni i}\ N_{t\ni i}$, où $N_{t\ni i}$ et $n_{t\ni i}$ sont, respectivement, la taille de population et la taille d'échantillon de la strate t contenant le domaine i (Särndal, Swensson et Wretman 1992, page 391).

Puisque nous estimons le nombre d'occurrences d'événements rares, dans 50 des 208 domaines, les estimations directes des nombres de NR et/ou de RS sont nulles, c'est-à-dire $\hat{\theta}_{i1} = 0$ et/ou $\hat{\theta}_{i2} = 0$. Des estimations ponctuelles nulles impliquent que $\hat{V}(\hat{\theta}_{i1}) = 0$ et/ou $\hat{V}(\hat{\theta}_{i2}) = 0$, où $\hat{V}(\hat{\theta}_{i1})$ et $\hat{V}(\hat{\theta}_{i2})$ sont les estimations habituelles de variance par rapport au plan de sondage de $\hat{\theta}_{i1}$ et $\hat{\theta}_{i2}$, respectivement. Ce résultat donne une fausse impression de grande précision, alors que, dans le contexte des petits domaines, c'est plus probablement l'exact opposé qui est vrai. En outre, des estimations fondées sur le plan de sondage des nombres de NR et/ou de RS qui sont égales à zéro produisent $\hat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$, où $\hat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$ est l'estimation standard fondée sur le plan de sondage de la covariance sous le plan entre $\hat{\theta}_{i1}$ et $\hat{\theta}_{i2}$. Par conséquent, les covariances doivent également être lissées dans un modèle multivarié d'estimation sur petits domaines.

Dans la suite de l'exposé, l'ensemble des 50 petits domaines dont la variance estimée, ou la covariance, ou les deux sont nulles est nommé « ensemble à dénombrement nul » (DN). L'ensemble complémentaire de 158 domaines pour lesquels $\hat{V}(\hat{\theta}_{i1}) > 0$ et $\hat{V}(\hat{\theta}_{i2}) > 0$, est nommé « ensemble à dénombrement non nul » (DNN).

Compte tenu du processus de génération des données et de la nature des variables de résultat, nous nous attendons à observer principalement des corrélations négatives entre θ_{i1} et θ_{i2}. En bref, nous avons besoin, pour le lissage des matrices de covariance ainsi que pour la modélisation des paramètres de petits domaines, d'une loi appropriée qui permet l'existence d'une matrice de covariance non contrainte, c'est-à-dire des corrélations positives ainsi que négatives.

Tableau 1
Variables définissant les domaines d'intérêt

LLMA regroupées selon la spécialisation de production	Taille de l'entreprise[b]	Secteur industriel[a]
District industriel[a,c]	1 à 9	1 Aliments, boissons et tabac
Aliments, boissons et tabac	10 à 49	2 Textiles et habillement
Textiles et habillement	50 à 99	3 Produits du papier, imprimerie et édition
Produits du papier, imprimerie et édition	≥ 100	4 Machinerie
Machinerie		5 Produits chimiques et métaux de base
Bijoux, instruments de musique, jeux, *etc*.		6 Cuir et chaussures
Cuir et chaussures		7 Bois, mobilier et équipement ménager
Bois, mobilier et équipement ménager		8 Bijoux, instruments de musique, jeux, *etc*.
LLMA non définies comme un district[c]		9 Constructeurs, entrepreneurs
Fabrication non spécialisée		10 Autres activités de fabrication
Non spécialisée, à l'exclusion de la fabrication		
Tourisme		
Villes		

(a) Tel que défini par la classification au niveau à deux chiffres de l'ATECO 91-CITI 3 et par Sforzi (1991).
(b) Définie en fonction du nombre d'employés.
(c) Défini conformément à Istat (1997).

4. Un modèle multivarié intégré d'estimation sur petits domaines pour les données de comptage

Les données de comptage multivariées peuvent avoir une structure de corrélation non négligeable. En général, la modélisation de cette structure a une incidence importante sur l'efficacité des estimateurs et sur le calcul des erreurs-types correctes. Un certain nombre de modèles multivariés pour données de comptage ont été proposés dans la littérature, comme le modèle de Poisson multivarié, le modèle binominal négatif multivarié et les modèles de mélange Poisson et Gamma multivariés (pour une revue de ces modèles, voir Winkelmann 2003). Malheureusement, ces lois ne conviennent pas pour modéliser nos données, puisqu'elles sont basées sur l'hypothèse que la corrélation est due à un facteur individuel qui ne varie pas d'un résultat à l'autre, ce qui implique une structure de covariance restreinte à des corrélations non négatives. Dans le cas bivarié, une structure de covariance plus souple est fournie par le modèle de mélange Poisson et Normale latente (van Ophem 1999) ; cependant, toute extension à des données multivariées de plus grande dimensionnalité semble difficilement applicable.

Aitchison et Ho (1989) ont proposé une loi *d*-variée qui permet une structure de covariance non contrainte, à savoir la loi multivariée Poisson-Log-normale (MPLN). Aucune forme analytique n'existe pour cette dernière, mais elle peut être représentée par un mélange plus simple qui permet l'estimation des paramètres selon la méthode de Monte Carlo appliquée aux chaines de Markov (MCMC) (Chib et Winkelmann 2001). Des renseignements concernant la loi MPLN sont fournis en annexe.

4.1 Lissage des matrices de covariance d'échantillonnage

Comme nous l'avons mentionné plus haut, l'approche des fonctions de variance généralisées (FVG) est généralement adoptée pour traiter l'instabilité des erreurs-types dans l'estimation sur petits domaines. À la présente section, nous présentons un modèle FVG avec une fonction de régression inspirée de la loi MPLN.

Soit $\mathbf{y}_{il} = [y_{i1l},\ y_{i2l}]'$ le vecteur des deux variables de résultat se rapportant à la l^{e} unité dans le i^{e} domaine. Soit $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \perp \mathbf{y}_{il'} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i$ et $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \sim \text{PLN}_2(\boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i),\ \forall i, \forall l$. Sur la base de ces hypothèses, les moments aboutissant au deuxième ordre peuvent être exprimés de la façon suivante :

$$E(y_{ijl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \exp(\lambda_{ij} + \sigma_{i,\,jj}/2) = \zeta_{ij}$$

$$V(y_{ijl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{ij} + \zeta_{ij}^2[\exp(\sigma_{i,\,jj}) - 1]$$

$$\text{COV}(y_{ijl}, y_{ihl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{i1}\zeta_{i2}[\exp(\sigma_{i,\,jh}) - 1],\ j \neq h$$

où $\sigma_{i,\,jh}$ désigne l'élément (j, h), $j, h = 1, 2$ de $\boldsymbol{\Sigma}_i$.

Pour résoudre le problème du lissage des matrices de covariance, Otto et Bell (1995) ont proposé une approche fondée sur l'hypothèse d'une loi de Wishart ; plus précisément, ils ont utilisé des estimations lissées dans un modèle normal-normal d'estimation sur petits domaines. Dans le même esprit, nous proposons une approche bayésienne s'appuyant sur la stratégie FVG qui suit. Sous échantillonnage aléatoire simple, supposons que la matrice de covariance d'échantillonnage dans le domaine i, $\mathbf{C}_i$ suit une loi de Wishart avec $n_i - 1$ degrés de liberté :

$$\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i \sim W_2(n_i - 1, \boldsymbol{\Gamma}_i)$$

où $\boldsymbol{\Gamma}_i = E(\mathbf{C}_i | n_i, \boldsymbol{\Gamma}_i)$, $i = 1, 2, ..., 158,$ et les éléments (j,h) de $\mathbf{C}_i$ sont définis comme $C_{i,\,jh} = n_i^{-1}\sum_{i=1}^{n_i}(y_{ijl} - \bar{y}_{ij})(y_{ijh} - \bar{y}_{ih})$, où $\bar{y}_{ij} = n_i^{-1}\sum_{i=1}^{n_i} y_{ijl}$.

Si les paramètres ζ_{ij} sont connus, alors $E(\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i)$ dépend uniquement des éléments de la matrice $\boldsymbol{\Sigma}_i$. Nous proposons d'estimer ζ_{ij} en utilisant l'estimateur fondé sur le plan $\hat{\zeta}_{ij} = N_i^{-1}\hat{\theta}_{ij}$. Donc, nous pouvons exprimer chaque élément de la matrice $\boldsymbol{\Gamma}_i$ comme une fonction des estimations $\hat{\zeta}_{ij}$ et des éléments de la matrice $\boldsymbol{\Sigma}_i$:

$$\Gamma_{i,11} = \hat{\zeta}_{i1} + \hat{\zeta}_{i1}^2(\exp(\sigma_{i,11}) - 1)$$
$$\Gamma_{i,22} = \hat{\zeta}_{i2} + \hat{\zeta}_{i2}^2(\exp(\sigma_{i,22}) - 1)$$
$$\Gamma_{i,12} = \hat{\zeta}_{i1}\hat{\zeta}_{i2}(\exp(\sigma_{i,12}) - 1)$$

où $\sigma_{i,11} = \bar{\boldsymbol{\sigma}}_{11}'\mathbf{Z}_i$, $\sigma_{i,22} = \bar{\boldsymbol{\sigma}}_{22}'\mathbf{Z}_i$, $\sigma_{i,12} = \bar{\boldsymbol{\sigma}}_{12}'\mathbf{Z}_i$, étant donné que $\mathbf{Z}_i$ est un vecteur 3×1 de variables indicatrices indiquant la catégorie de taille de l'entreprise dans le domaine *i*, et :

$$\bar{\boldsymbol{\sigma}}_{11} = \begin{pmatrix} \bar{\sigma}_{1,11} \\ \bar{\sigma}_{2,11} \\ \bar{\sigma}_{3,11} \end{pmatrix}, \bar{\boldsymbol{\sigma}}_{22} = \begin{pmatrix} \bar{\sigma}_{1,22} \\ \bar{\sigma}_{2,22} \\ \bar{\sigma}_{3,22} \end{pmatrix}, \bar{\boldsymbol{\sigma}}_{12} = \begin{pmatrix} \bar{\sigma}_{1,12} \\ \bar{\sigma}_{2,12} \\ \bar{\sigma}_{3,12} \end{pmatrix}$$

c'est-à-dire que nous supposons que les paramètres $\boldsymbol{\Sigma}_i$ sont égaux pour les domaines appartenant à la même catégorie de taille d'entreprise.

Nous estimons les paramètres $\bar{\boldsymbol{\sigma}}_{11}, \bar{\boldsymbol{\sigma}}_{22}, \bar{\boldsymbol{\sigma}}_{12}$ sur les données de l'ensemble DNN. Puisque nous suivons une approche bayésienne, nous devons spécifier des priors pour $\bar{\sigma}_{k,jj}$ et $\bar{\sigma}_{k,12}$ $k = 1, 2, 3$. Nous utilisons les spécifications de priors suivantes : $\bar{\sigma}_{k,11}^{1/2} \sim U^+$, $\bar{\sigma}_{k,22}^{1/2} \sim U^+$, $\bar{\rho}_k \sim U(-1,1)$, où $\bar{\sigma}_{k,12} = \bar{\rho}_k(\bar{\sigma}_{k,11}\bar{\sigma}_{k,22})^{1/2}$ et U^+ désignent une loi uniforme sur un sous-ensemble de R^+ dont la longueur est grande mais finie. À la section 4.3, nous montrons comment ces estimations peuvent être utilisées pour intégrer le modèle d'estimation sur petits domaines à un modèle pour les matrices de covariance des erreurs d'échantillonnage.

4.2 Un modèle Normal-Poisson-Log-normal multivarié d'estimation sur petits domaines

À la présente section, nous proposons un modèle multivarié d'estimation sur petits domaines basé sur la loi MPLN afin d'estimer conjointement les nombres de RS et de NR en utilisant l'ensemble DNN.

Soit $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T$ le vecteur des deux paramètres d'intérêt pour le i^{e} domaine dans l'ensemble de données DNN $(i = 1, ..., 158)$, et soit $\hat{\boldsymbol{\theta}}_i$ le vecteur correspondant d'estimations directes. Le modèle d'estimation sur petits domaines est constitué de deux modèles distincts. Le premier est un modèle d'échantillonnage :

$$\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \sim \text{ind}\ N_2(\boldsymbol{\theta}_i \mid \boldsymbol{\Psi}_i), \quad i = 1, ..., 158. \qquad (1)$$

Comme dans Lahiri et Rao (1995), nous justifions l'hypothèse de normalité dans (1) en utilisant l'argument de la limite centrale. Il est d'usage, en pratique, de supposer que les matrices de covariance des erreurs d'échantillonnage $\boldsymbol{\Psi}_i$ sont connues et, généralement, d'utiliser une méthode FVG pour estimer $\boldsymbol{\Psi}_i$. Ici, comme estimation lissée de $\boldsymbol{\Psi}_i$, nous adoptons $\hat{\boldsymbol{\Psi}}_i = E(\boldsymbol{\Gamma}_i | \mathbf{C}_i, n_i)K_i$, où $K_i = N_i(N_{t\text{ə}i}/n_{t\text{ə}i} - 1)$. À partir d'ici, nous donnerons à $\hat{\boldsymbol{\Psi}}_i$ le nom de matrice de covariance des erreurs d'échantillonnage lissée (MCEEL).

La deuxième composante du modèle d'estimation sur petits domaines est un modèle de lien qui relie $\boldsymbol{\theta}_i$ aux données auxiliaires propres au domaine :

$$\boldsymbol{\theta}_i \sim \text{ind}\ \text{PLN}_2(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_\nu), \quad i = 1, ..., 158,$$
où
$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\beta}\mathbf{Z}_i x_i. \qquad (2)$$

$\mathbf{Z}_i$ est un vecteur 3×1 de variables indicatrices indiquant la catégorie de taille de l'entreprise dans le domaine *i* et $x_i = \log(x_i^*)$, où x_i^* est le nombre d'employés dans le domaine *i*.

En fin ce compte, $\boldsymbol{\Sigma}_\nu$ est la matrice de covariance reliée aux effets aléatoires propres au domaine :

$$\boldsymbol{\Sigma}_\nu = \begin{pmatrix} \sigma_{\nu,11} & \sigma_{\nu,12} \\ \sigma_{\nu,21} & \sigma_{\nu,22} \end{pmatrix}$$

et

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} 0 & \gamma_{12} & \gamma_{13} \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}.$$

À partir d'ici, nous donnerons à ce modèle d'estimation sur petits domaines le nom de modèle « multivarié Normal-Poisson-Log-normal » (MNPLN).

Nous adoptons une approche entièrement hiérarchique bayésienne. Dans ce cadre, des modèles relativement complexes (par exemple multivariés) peuvent être facilement appliqués ; en outre, on peut approximer les lois a posteriori en utilisant des algorithmes MCMC. Le calcul d'estimations multivariées sur petits domaines, et des estimations de leur EQM en particulier, peut être difficile dans une approche fréquentiste. La spécification des priors pour le modèle décrit est la suivante :

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim N_2(\mathbf{0}, a\mathbf{I}_2),$$
$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \sim N_2(\mathbf{0}, g_{k'}\mathbf{I}_2) \qquad k' = 2, 3,$$
$$\begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} \sim N_2(\mathbf{0}, b_k\mathbf{I}_2) \qquad k = 1, 2, 3,$$
$$\boldsymbol{\Sigma}_\nu^{-1} \sim W(s, \mathbf{I}_2),$$
$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \perp \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix},$$

où $s = 3$ et a, $g_{k'}$, b_k sont grands comparativement à l'échelle des données. Il en est ainsi pour refléter le manque

d'information a priori au sujet des paramètres du modèle, donc pour établir une spécification diffuse, mais correcte, des priors. Les moyennes a posteriori $\hat{\boldsymbol{\theta}}_i^{\mathrm{HB}} = E(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ sont prises comme estimateurs des paramètres de petit domaine, tandis que la variance a posteriori $V(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ est utilisée comme mesure de l'incertitude.

Par souci de comparaison, nous prenons comme référence le modèle multivarié Normal-Normal (MNN) standard, dans lequel le modèle d'échantillonnage est défini comme en (1) et le modèle de lien, comme suit :

$$\boldsymbol{\theta}_i \sim \text{ind } N_2(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_{\mathrm{v}}^*), \tag{3}$$

où $\boldsymbol{\mu}_i^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^* \mathbf{Z}_i + \boldsymbol{\beta}^* \mathbf{Z}_i x_i^*$. Les paramètres $\boldsymbol{\alpha}^*$, $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ et leurs lois a priori sont définis comme $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ et $\boldsymbol{\beta}$ dans le modèle précédent.

4.3 Modèle d'estimation sur petits domaines MNPLN

Afin de tenir compte de la variabilité supplémentaire due aux matrices de covariance estimées des erreurs d'échantillonnage, ainsi que pour surmonter le problème des variances et covariances nulles, nous proposons une solution dans l'esprit de celle proposée par Arora et Lahiri (1997), Liu et coll. (2007) ainsi que You (2008). Nous intégrons le modèle pour les matrices de covariance des erreurs d'échantillonnage de la section 4.1 dans les modèles d'estimation sur petits domaines (1) et (2). Donc, ici, nous faisons référence à l'ensemble complet de 208 domaines.

Dans ce contexte, le modèle d'échantillonnage sur petits domaines est formulé comme d'habitude, c'est-à-dire $\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \sim \text{ind } N_2(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i^*)$, $i = 1, \ldots, 208$. Conformément aux hypothèses $\mathbf{y}_{il}$ relatives aux équations formulées à la section 4.1, en supposant que les $\boldsymbol{\Sigma}_i$ sont connues et en supposant que $\theta_{ij} = N_i \zeta_{ij}$, les éléments de la matrice de covariance des erreurs d'échantillonnage $\boldsymbol{\Psi}_i^*$ peuvent être exprimés de la façon suivante :

$$\Psi_{i,jj}^* = K_i[\theta_{ij}/N_i + \theta_{ij}^2/N_i^2 \,(\exp(\hat{\bar{\boldsymbol{\sigma}}}'_{jj}\mathbf{Z}_i) - 1)] \tag{4}$$

$$\Psi_{i,12}^* = K_i[N_i^{-2}\theta_{i1}\theta_{i2} \,(\exp(\hat{\bar{\boldsymbol{\sigma}}}'_{12}\mathbf{Z}_i) - 1)] \tag{5}$$

où $\hat{\bar{\boldsymbol{\sigma}}}'_{jj}$ $j = 1, 2$ et $\hat{\bar{\boldsymbol{\sigma}}}'_{12}$ sont les moyennes a posteriori des paramètres $\bar{\boldsymbol{\sigma}}_{jj}$ et $\bar{\boldsymbol{\sigma}}_{12}$, respectivement, calculées, en utilisant le modèle de la section 4.1.

Puisque les matrices de covariance des erreurs d'échantillonnage sont exprimées en fonction des paramètres $\boldsymbol{\theta}_i$, elles peuvent être considérées ici comme des matrices de covariance des erreurs d'échantillonnage fondées sur un modèle (MCEEM). Les moyennes a posteriori $\hat{\boldsymbol{\theta}}_i^{\mathrm{HB}} = E(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i)$ sont utilisées comme estimateurs des $\boldsymbol{\theta}'_i$, tandis que la variance a posteriori $V(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i)$ est utilisée comme mesure de l'incertitude.

Nous notons que le modèle MNN ne peut pas être mis en œuvre en suivant l'approche intégrée décrite plus haut. En fait, (3) ne garantit pas la positivité de $\boldsymbol{\theta}_i$ ni, par conséquent, des éléments diagonaux de $\boldsymbol{\Psi}_i$.

5. Analyse des données

À la section 5.1, nous comparons le modèle MNPLN au modèle MNN de référence et à leurs analogues univariés. Pour les deux modèles, nous supposons que les matrices de covariance des erreurs d'échantillonnage sont lissées (MCEEL) ; nous désignons donc les deux stratégies par MNPLN-MCEEL et MNN-MCEEL dans la suite. Puisque ces modèles ne nous permettent pas de résoudre le problème des dénombrements nuls, nous faisons référence dans cette analyse à l'ensemble de données DNN. À la section 5.2, nous comparons la stratégie intégrée d'estimation sur petits domaines basée sur le modèle MNPLN et la matrice de covariance des erreurs d'échantillonnage fondée sur un modèle MCEEM (MNPLN-MCEEM), que nous avons présentés à la section 4.3, avec la stratégie basée sur le MNPLN-MCEEL. Nous limitons l'analyse à l'ensemble de données DNN afin d'évaluer les deux stratégies sous les mêmes conditions. Enfin, à la section 5.3, nous évaluons le rendement global du modèle d'estimation sur petits domaines proposé MNPLN-MCEEM pour l'ensemble complet de données (DNN+DN).

Pour tous les modèles, nous avons obtenu les lois a posteriori des paramètres par intégration Monte Carlo au moyen de l'algorithme d'échantillonnage de Gibbs. Nous avons utilisé le logiciel de simulation MCMC WinBUGS (Spiegelhalter, Thomas, Best et Gilks 1995) pour exécuter trois chaînes parallèles (chacune comportant 25 000 exécutions) en tirant le point de départ d'une distribution surdispersée. Les codes de WinBUGS sont disponibles à l'adresse URL http://www2.stat.unibo.it/trivisano/. Nous avons surveillé la convergence de l'échantillonneur de Gibbs par inspection visuelle des graphiques des chaînes et des diagrammes d'autocorrélation, ainsi qu'au moyen du facteur de réduction d'échelle potentielle proposé par Gelman et Rubin (1992). Bien que la convergence ait été rapide pour tous les modèles, nous avons écarté les 5 000 premières itérations de chaque chaîne. Dans les modèles multivariés, l'autocorrélation relativement forte des chaînes est réduite en amincissant la chaîne (une valeur sur trois a été considérée pour les sommaires a posteriori). Voir Rao (2003, pages 228-232) pour des précisions.

Les propriétés des modèles d'estimation sur petits domaines dont nous avons discuté aux sections 4.2 et 4.3 sont comparées en utilisant diverses mesures. Afin de choisir parmi les modèles concurrents, nous avons calculé le critère d'information de déviance (DIC pour *Deviance Information Criterion*). Le DIC est un critère de sélection de modèle en vertu duquel le rendement d'un modèle est évaluée comme la somme d'une mesure d'ajustement (la

moyenne a posteriori de la déviance $\bar{D}$) et d'une mesure de complexité correspondant à la différence entre $\bar{D}$ et la déviance évaluée à la moyenne a posteriori du paramètre. De cette façon, un modèle aura la préférence s'il donne une valeur du DIC plus faible (Spiegelhalter, Best, Carlin et Van der Linde 2002).

Afin de vérifier la force de l'approche multivariée de l'estimation sur petits domaines, nous utilisons comme référence les versions univariées des modèles discutés aux sections 4.2 et 4.3, définies comme il suit. Pour tous les modèles, nous posons que $\sigma_{v,12} = 0$ dans $\boldsymbol{\Sigma}_v$, et nous supposons que $\sigma_{v,11} \perp \sigma_{v,22}$, $\sigma_{v,jj}^{1/2} \sim U(0, U^+)$, $j = 1, 2$. Pour les modèles MCEEL, nous posons que $\boldsymbol{\Psi}_i = \text{diag}(\hat{\boldsymbol{\Psi}}_i)$, tandis que pour les modèles MCEEM, nous posons que $\sigma_{1,12} = 0$ dans (5). En outre, nous obtenons un nouvel ensemble d'estimations pour les paramètres $\bar{\boldsymbol{\sigma}}_{11}$ et $\bar{\boldsymbol{\sigma}}_{22}$ en posant que $\bar{\rho}_k = 0$ dans le modèle de la section 4.1.

Le tableau 2 donne les valeurs du critère DIC pour l'ensemble complet de modèles d'estimations sur petits domaines.

Tableau 2
Comparaison des modèles en utilisant la statistique DIC

Modèle	Ensemble de données	DIC
MNN-MCEEL	DNN	2 742,2
(version univariée)	DNN	2 745,4
MNPLN-MCEEL	DNN	2 656,9
(version univariée)	DNN	2 661,0
MNPLN-MCEEM	DNN	2 623,6
(version univariée)	DNN	2 638,1
MNPLN-MCEEM	DNN+DN	3 202,7
(version univariée)	DNN+DN	3 214,3

Tous les modèles multivariés considérés donnent de meilleurs résultats en ce qui concerne le DIC que leurs analogues univariés (tableau 2). En outre, pour tous les modèles multivariés, nous constatons que les intervalles de crédibilité a posteriori de $\rho_v = \sigma_{v,12}/\sqrt{\sigma_{v,11}\sigma_{v,22}}$ ne contiennent pas la valeur zéro. Par conséquent, dans les paragraphes qui suivent, nous nous concentrons sur les modèles multivariés.

Nous avons vérifié l'adéquation des modèles multivariés spécifiés par des vérifications prédictives a posteriori. Nous avons généré les valeurs simulées d'une mesure de divergence appropriée à partir de la loi prédictive a posteriori et nous les avons comparées aux valeurs de la même mesure calculée d'après les données observées. Soit $\hat{\boldsymbol{\theta}}_{obs}$ et $\hat{\boldsymbol{\theta}}_{sim}$ les données observées et générées, respectivement. La valeur p prédictive a posteriori est définie comme $p = P\{d(\hat{\boldsymbol{\theta}}_{sim}, \boldsymbol{\theta}) > d(\hat{\boldsymbol{\theta}}_{obs}, \boldsymbol{\theta}) \mid \hat{\boldsymbol{\theta}}_{obs}\}$. Nous envisageons une mesure de divergence proposée dans Datta et coll. (1999), définie comme suit :

$$d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{N} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)' \, \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i). \qquad (6)$$

Le calcul de la valeur p au moyen du résultat de la simulation MCMC est simple. Les valeurs extrêmes de la probabilité p indiquent un manque d'ajustement d'un modèle donné. À l'exemple de Rao (2003, pages 245-246) et de You et Rao (2002), nous calculons deux statistiques utiles pour évaluer l'ajustement d'un modèle au niveau du domaine individuel. La première statistique $p_{ij}^* = P(\hat{\theta}_{ij,sim} < \hat{\theta}_{ij,obs} \mid \hat{\boldsymbol{\theta}}_{obs})$, renseigne sur le degré de surestimation ou de sous-estimation systématique de $\hat{\theta}_{ij,obs}$.

La deuxième statistique est définie ainsi :

$$d_{ij}^* = [E(\hat{\theta}_{ij} \mid \hat{\boldsymbol{\theta}}_{obs}) - \hat{\theta}_{ij,obs}] \Big/ \sqrt{V(\hat{\theta}_{ij} \mid \hat{\boldsymbol{\theta}}_{obs})},$$

où l'espérance et la variance sont calculées en fonction de la loi prédictive a posteriori. Le tableau 3 résume les résultats relatifs à p, p_{ij}^* et d_{ij}^*.

En guise de vérification supplémentaire de la cohérence des données, nous avons calculé les estimations directes et celles fondées sur le modèle de ${}_A\theta_{sj}$, $s = 1, ..., 10$, c'est-à-dire le nombre total de NR et de RS pour les dix domaines déterminés en classant les entreprises uniquement selon le secteur industriel. Soit $w_{is} = 1$ si le nombre de recrues dans le domaine i se rapporte au secteur industriel s et $w_{is} = 0$; autrement ; alors :

$${}_A\theta_{sj} = \sum_i \theta_{ij} w_{is}. \qquad (7)$$

À ce niveau d'agrégation, les estimations directes peuvent être considérées comme exactes. Par conséquent, étant donné deux ensembles d'estimations fondées sur un modèle se rapportant à ces grands domaines, nous préférons celui qui concorde avec les estimations directes. Les domaines correspondant au secteur industriel sont planifiés dans l'Enquête Excelsior ; chaque secteur industriel est stratifié en fonction de la taille de l'entreprise. Par conséquent, les estimations directes ${}_A\hat{\theta}_{sj}$ pour chaque secteur industriel sont calculées en utilisant l'estimateur classique d'Horwitz-Thompson. Les estimations fondées sur un modèle agrégé sont calculées en se basant sur les données de sortie MCMC. Pour les modèles ayant trait aux données DNN, nous avons effectué l'agrégation selon (7) à chaque étape $t, t = 1, ..., T$, de la simulation MCMC, avec les échantillons ${}^t\theta_{ij}^*$ et ${}^t\theta_{ij}^{**}$ tirés respectivement de la loi a posteriori de θ_{ij} pour les domaines appartenant à l'ensemble DNN et de la loi prédictive de θ_{ij} pour les domaines appartenant à l'ensemble DN. L'estimateur HB est défini comme ${}_A\hat{\theta}_{sj}^{HB} = T^{-1}\sum_{t=1}^{T}(\sum_{i \in NZC} {}^t\theta_{ij}^* w_{is} + \sum_{i \in ZC} {}^t\theta_{ij}^{**} w_{is})$. Sinon, pour le modèle sur les données DNN+DN, nous avons agrégé selon (7) les échantillons MCMC des lois a posteriori de θ_{ij}. Dans ce cas, l'estimateur HB est défini

comme ${}_A\hat{\theta}_{sj}^{\mathrm{HB}} = T^{-1}\sum_{t=1}^{T}(\sum_{i\in NZC_s}{}^t\theta_{ij}^* w_{is})$. Le tableau 4 donne les résultats sommaires pour ${}_A\hat{\theta}_{sj}$ et ${}_A\hat{\theta}_{sj}^{\mathrm{HB}}$.

Pour tous les modèles multivariés, nous avons examiné les variantes qui suivent des lois a priori : nous avons utilisé des lois a priori uniformes non informatives indépendantes pour les éléments des vecteurs $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$, et $\boldsymbol{\gamma}^*$; $\sigma_{v,jj}^{1/2} \sim U^+$, $j = 1, 2$, $\rho_v \sim U(-1,1), \sigma_{v,12} = \rho_v (\sigma_{v,12}\,\sigma_{v,12})^{1/2}$. Nous avons fait la même chose pour les éléments de la matrice $\boldsymbol{\Sigma}^*$ dans le modèle MNN. Nous n'avons relevé aucun changement pertinent dans les lois a posteriori des paramètres d'intérêt.

5.1 Comparaison des modèles MNPLN-MCEEL et MNN-MCEEL sur l'ensemble de données DNN

Nous constatons que le modèle MNPLN-MCEEL surpasse largement le modèle MNN-MCEEL en ce qui concerne le DIC (tableau 2). Ce dernier modèle présente un manque d'ajustement, sa valeur p étant égale à 0,034 (tableau 3), tandis qu'une valeur p de 0,65 donne à penser que le modèle MNPLN-MCEEL est adéquat. Ce résultat est confirmé quand nous comparons les mesures p_{ij}^* et d_{ij}^* (tableau 3) pour les deux modèles. Dans le cas du modèle MNN-MCEEL, p_{ij}^* varie, selon le domaine, de 0,000 à 0,995 pour les nouvelles recrues NR $(j = 1)$ et de 0,003 à 0,993 pour les recrues de substitution RS $(j = 2)$, respectivement, ce qui témoigne d'une surestimation et d'une sous-estimation dans certains domaines. En outre, les statistiques sommaires pour les résidus standardisés d_{ij}^* indiquent que certaines valeurs prédites se situent à plus de deux écarts-types des valeurs observées correspondantes. Les mêmes mesures pour le modèle MNPLN-MCEEL indiquent que l'ajustement est adéquat.

Nous constatons aussi que le modèle MNPLN-MCEEL surpasse le modèle MNN-MCEEL quand le rendement est évalué en fonction des estimations pour les grands domaines (tableau 4). En fait, les intervalles de crédibilité pour le modèle MNN-MCEEL couvrent seulement deux estimations directes agrégées pour les NR et quatre pour les RS, tandis que sous le modèle MNPLN-MCEEL, les intervalles de crédibilité couvrent six estimations directes agrégées pour les NR et six pour les RS.

5.2 Comparaison des modèles MNPLN-MCEEL et MNPLN-MCEEM sur l'ensemble de données DNN

Les valeurs de p, p_{ij}^* et d_{ij}^* sont approximativement comparables pour les modèles MNPLN-MCEEL et MNPLN-MCEEM (tableau 3). De même, les estimations fondées sur un modèle produites par MNPLN-MCEEL prennent des valeurs très proches de celles obtenues en utilisant MNPLN-MCEEM ; en fait, la corrélation entre les moyennes a posteriori de θ_{i1} sous les deux modèles est égale à 0,98, tandis que la même mesure ayant trait à θ_{i2} est égale à 0,94. Les mêmes résultats sont obtenus pour la corrélation entre les erreurs-types a posteriori, qui sont de 0,92 et de 0,94, respectivement. Les propriétés du modèle MNPLN-MCEEM en ce qui concerne la concordance avec les estimations directes sur les grands domaines (tableau 4) sont un peu meilleures que celles du modèle MNPLN-MCEEL : sept estimations directes des NR et huit des RS sont couvertes, respectivement, par l'intervalle de crédibilité calculé sur ce modèle.

Étant donné ces résultats, nous concluons que l'ajustement du modèle MNPLN-MCEEM est adéquat.

5.3 Évaluation du rendement du modèle MNPLN-MCEEM sur l'ensemble de données DNN+DN

Nous observons que le rendement du modèle MNPLN-MCEEM sur l'ensemble complet de données en ce qui concerne p, p_{ij}^* et d_{ij}^* est satisfaisante et comparable à celui du même modèle sur l'ensemble de données DNN (tableau 3). De toute évidence, les valeurs du critère DIC obtenues pour les deux modèles ne peuvent être comparées, les deux modèles étant estimés sur différents ensembles de données.

Comme le montre le tableau 4, tous les intervalles de crédibilité calculés au moyen de ce modèle couvrent des estimations directes sur de grands domaines ; autrement dit, la concordance des estimations HB avec les estimations directes est très satisfaisante. Ce résultat peut s'expliquer en observant que la probabilité de dénombrements nuls est plus grande dans les petits domaines, qui sont caractérisés par un petit nombre d'employés (la covariable dans tous les modèles). Par conséquent, le fait d'estimer les modèles sur l'ensemble de données DNN peut produire des estimations biaisées du paramètre $\boldsymbol{\beta}$. Nous concluons que l'intégration d'un modèle de covariance d'échantillonnage dans le modèle d'estimation sur petits domaines MNPLN accroît sensiblement la fiabilité des estimations sur petits domaines. Afin de rendre compte du gain d'efficacité des estimations HB, nous avons calculé sur l'ensemble de données DNN la réduction moyenne en pourcentage du CV (You 2008), définie comme la moyenne de l'écart entre le CV direct et le CV HB (le ratio de la racine carrée de la variance a posteriori et de la moyenne a posteriori) par rapport au CV direct. La réduction moyenne du CV est de 23,1 % pour les NR et de 29,1 % pour les RS.

Tableau 3
Vérification prédictive a posteriori ; sommaires de p^*_{ij} et de d^*_{ij} calculés par rapport à *i*

Modèle	Ensemble de données	p		p^*_{i1}	p^*_{i2}	d^*_{i1}	d^*_{i2}
MNN-MCEEL	DNN	0,034	Min.	0,000	0,003	-3,764	-2,867
			Médiane	0,591	0,616	0,257	0,295
			Max.	0,995	0,993	2,656	-2,515
MNPLN-MCEEL	DNN	0,65	Min.	0,154	0,129	-0,965	-1,165
			Médiane	0,535	0,561	0,124	0,149
			Max.	0,891	0,912	1,216	1,286
MNPLN-MCEEM	DNN	0,78	Min.	0,090	0,134	-1,085	-0,983
			Médiane	0,515	0,519	-0,084	-0,085
			Max.	0,916	0,914	1,401	1,787
MNPLN-MCEEM	DNN+DN	0,79	Min.	0,072	0,111	-1,164	-0,945
			Médiane	0,506	0,523	-0,076	-0,094
			Max.	0,903	0,913	1,301	1,778

Tableau 4
Estimations directes et HB pour les secteurs industriels ; en italique, estimations HB dont les intervalles de crédibilité couvrent les estimations directes

	Estimations directes		Estimations HB											
			MNN-MCEEL (DNN)			MNPLN-MCEEL (DNN)			MNPLN-MCEEM (DNN)			MNPLN-MCEEM (DNN+DN)		
s	${}_A\hat{\theta}_{s1}$	se$({}_A\hat{\theta}_{s1})$	${}_A\hat{\theta}^{HB}_{s1}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s1}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s1}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s1}$	Int. de crédit. à 95 %	
1	1 702,0	41,3	1 077,0	964,3	1 201,0	1 266,0	1 055,0	1 509,0	*1 649,0*	*1 434,0*	*1 906,0*	*1 630,0*	*1 406,0*	*1 899,0*
2	1 758,8	41,9	1 936,0	1 793,0	2 091,0	*1 720,0*	*1 441,0*	*2 011,0*	*1 975,0*	*1 665,0*	*2 347,0*	*1 908,0*	*1 598,0*	*2 291,0*
3	725,0	26,9	557,8	460,6	662,7	534,6	435,8	642,3	*696,6*	*573,3*	*842,3*	*682,8*	*575,5*	*811,8*
4	373,9	19,3	202,7	123,0	294,8	192,1	129,1	277,0	*370,0*	*291,1*	*471,4*	*319,8*	*252,1*	*408,3*
5	142,4	11,9	*158,2*	*66,5*	*258,2*	*146,0*	*98,4*	*205,7*	235,6	164,3	326,9	*149,7*	*108,3*	*205,0*
6	5 624,1	75,0	4 134,0	3 800,0	4 484,0	*5 235,0*	*4 814,0*	*5 670,0*	*5 537,0*	*5 136,0*	*5 963,0*	*5 594,0*	*5 187,0*	*6 029,0*
7	887,7	29,8	659,9	549,1	783,7	629,6	526,4	743,4	*872,7*	*761,7*	*1 003,0*	*844,6*	*732,3*	*980,3*
8	223,9	15,0	*263,3*	*188,2*	*340,6*	*260,6*	*182,8*	*351,3*	362,0	262,8	494,1	*288,7*	*203,1*	*410,8*
9	661,5	25,7	893,7	790,3	999,4	*777,6*	*624,7*	*948,7*	931,0	754,8	1 150,0	*803,3*	*638,7*	*1 017,0*
10	1 792,6	42,3	1 460,0	1 334,0	1 598,0	*1 579,0*	*1 381,0*	*1 798,0*	*1 847,0*	*1 650,0*	*2 074,0*	*1 813,0*	*1 610,0*	*2 053,0*
	${}_A\hat{\theta}_{s2}$	se$({}_A\hat{\theta}_{s2})$	${}_A\hat{\theta}^{HB}_{s2}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s2}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s2}$	Int. de crédit. à 95 %		${}_A\hat{\theta}^{HB}_{s2}$	Int. de crédit. à 95 %	
1	942,7	300,2	482,0	428,5	531,3	503,7	413,3	600,4	*832,6*	*706,4*	*987,6*	*817,8*	*686,0*	*980,0*
2	920,0	135,7	*883,9*	*798,7*	*967,4*	*849,8*	*694,8*	*1 022,0*	*949,8*	*778,9*	*1 161,0*	*922,3*	*747,6*	*1 167,0*
3	253,2	35,6	*249,2*	*209,2*	*292,1*	*254,1*	*202,1*	*309,9*	338,8	269,2	423,1	*284,7*	*226,2*	*354,5*
4	150,5	36,0	84,4	53,3	120,4	84,7	56,8	119,2	*160,6*	*116,7*	*218,0*	*131,5*	*97,0*	*179,6*
5	39,8	16,6	*66,7*	*31,2*	*104,2*	*62,0*	*37,3*	*89,3*	116,3	74,3	173,0	*60,9*	*38,4*	*90,5*
6	2 304,0	131,5	1 869,0	1 692,0	2 054,0	2 070,0	1 856,0	2 282,0	*2 273,0*	*2 060,0*	*2 508,0*	*2 297,0*	*2 079,0*	*2 542,0*
7	532,7	105,8	293,0	247,7	345,6	299,0	245,9	357,2	*471,5*	*402,8*	*553,2*	*443,3*	*377,2*	*538,3*
8	80,8	32,3	115,7	85,7	143,5	*100,5*	*67,7*	*140,3*	*139,5*	*76,7*	*210,4*	*98,0*	*58,5*	*156,9*
9	362,7	66,3	*407,0*	*358,6*	*453,0*	*361,0*	*285,8*	*438,8*	*432,1*	*335,4*	*552,9*	*360,4*	*274,7*	*476,2*
10	856,3	70,7	661,1	598,1	722,6	714,4	614,0	824,7	*855,4*	*740,5*	*984,6*	*832,7*	*719,8*	*964,5*

Remerciements

Les auteurs remercient le rédacteur en chef, le rédacteur associé et l'examinateur de leurs commentaires et suggestions utiles. Les travaux de recherche à l'origine du présent article ont été financés, en partie, par les subventions Miur-PRIN 2003/2003133249 et Miur-Prin 2008/2008CEFF37-001.

Annexe

La loi multivariée Poisson-Log-normale

Soit $\mathbf{y} = (y_1, y_2, ..., y_j, ..., y_d)$ un vecteur de dénombrements de dimension d, et supposons que $y_j|\tau_j \sim \text{Po}(\tau_j)$, avec $y_j|\tau_j \perp y_{j'}|\tau_{j'}\ (j \neq j')$. Soit le vecteur de paramètres $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_j, ..., \tau_d)$ qui suit une loi Log-normale multivariée, c'est-à-dire $\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{LN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$, où $\boldsymbol{\lambda} = E(\log \boldsymbol{\tau})$

et $\boldsymbol{\Sigma} = \mathrm{COV}(\log \boldsymbol{\tau})$. Alors, la loi marginale de $\mathbf{y}$ est une loi multivariée Poisson-Log-normale (MPLN), qui est un mélange Log-normal de d lois de Poisson $\mathrm{Po}(\tau_j)$ indépendantes, c'est-à-dire $\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \mathrm{PLN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. En désignant le (j, h), $j, h = 1, 2, ..., d$ élément de $\boldsymbol{\Sigma}$ par σ_{jh}, nous pouvons obtenir facilement les moments marginaux par la voie des résultats d'espérance conditionnelle et des propriétés standard des lois de Poisson et Log-normale :

$$E(y_j \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \exp(\lambda_j + \sigma_{jj}/2) = \zeta_j$$

$$V(y_j \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j + \zeta_j^2 \left[\exp(\sigma_{jj}) - 1\right]$$

$$\mathrm{COV}(y_j, y_h \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j \zeta_h [\exp(\sigma_{jh}) - 1], \quad j \neq h.$$

Notons que le modèle MPLN tient compte de la surdispersion fournie par $\sigma_{jj} > 0$, ce qui mène à $V(y_j \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}) > E(y_j \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma})$. En outre, la structure de corrélation des dénombrements n'est pas contrainte, puisque $\mathrm{COV}(y_j, y_h \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma})$ peut être positive ou négative selon le signe σ_{jh}. Aitchison et Ho (1989), ainsi que Good et Pirog-Good (1989), ont étudié une loi MPLN bivariée, bien qu'uniquement dans les cas où il n'existait pas de covariables. Cependant, le même modèle peut facilement être étendu en vue de prendre les covariables en considération (Chib et Winkelmann 2001).

Bibliographie

Aitchison, J., et Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Baldi, C., Bellisai, D., Fivizzani, S. et Sorrentino, M. (2007). Production of job vacancy statistics: Coverage. Contributi Istat, Istituto Nazionale di Statistica.

Becattini, G. (1992). The Marshallian industrial district as a socio-economic notion. Dans *Industrial Districts and International Co-operation in Italy*, (Éds., F. Pyke, G. Becattini et W. Sengenberger). Internation Labor Office, Genève.

Chattopadhyay, M., Lahiri, P., Larsen, M. et Reimnitz, J. (1999). Estimation composite de la prévalence des drogues pour des zones infraétats. *Techniques d'enquête*, 25, 91-97.

Chen, S. (2001). Empirical best prediction and hierarchical Bayes methods in small area estimation. Thèse de doctorat, Department of Mathematics and Statistics, University of Nebraska, Lincoln.

Chib, S., et Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19, 428-435.

Cohen, M.L. (2000). Evaluation of Census Bureau's small-area poverty estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 62-68.

Datta, G.S., Fay, R.E. et Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. *Proceedings of Bureau of the Census 1991 Annual Research Conference*, U. S. Bureau of the Census, Washington, DC, 63-79.

Datta, G.S., Ghosh, M., Nangia, N. et Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. Dans *Bayesian Analysis in Statistics and Econometrics*, (Éds., D.A. Berry, K.M. Chaloner et J.M. Geweke). New York : John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 488, 1074-1082.

Elazar, D. (2004). Small area estimation of disability in Australia. *Statistics in Transition*, 6, 5, 667-684.

Fabrizi, E., Ferrante, M.R. et Pacei, S. (2005). Estimation of poverty indicators at sub-national level using multivariate small area models. *Statistics in Transition*, 7, 3, 587-608.

Fabrizi, E., Ferrante, M.R. et Pacei, S. (2008). Measuring sub-national income poverty by using a small area multivariate approach. *Review of Income and Wealth*, 54, 4, 597-615.

Fay, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*, (Éds., R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York : John Wiley & Sons, Inc., 91-102.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.

Ghosh, M., Nangia, N. et Kim, D. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Good, D.H., et Pirog-Good, M.A. (1989). Models for bivariate count data with an application to teenage delinquency and paternity. *Sociological Methods and Research*, 17, 4, 409-431.

Istat (1997). I sistemi locali del lavoro 1991. *Argomenti*, Roma 1997, 10.

Lahiri, P., et Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

Liu, B., Lahiri, P. et Kalton, G. (2007). Hierarchical Bayes modeling of survey weighted small area proportions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3181-3186.

Otto, M.C., et Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey : John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York : Springer-Verlag.

Sforzi, F. (1991). I distretti industriali marshalliani nell'economia italiana. Dans *Distretti industriali e cooperazione fra imprese in Italia*, (Éds., F. Pyke, G. Becattini et W. Sengenberger). Quaderni di Studi e Informazioni, 34.

Sforzi, F., et Lorenzini, F. (2002). I distretti industriali. Dans *Ministero delle Attività Produttive-IPI*, *L'esperienza italiana dei distretti industriali*, Roma, IPI.

Spiegelhalter, D.J., Best, N., Carlin, B.P. et Van der Linde, A. (2002). Bayesian measures of model complexity and fit (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 64, 583-639.

Spiegelhalter, D.J., Thomas, A., Best, N.G. et Gilks, W.R. (1995). *BUGS*: *Bayesian Inference Using Gibbs Sampling*. Version 0.50, Medical Research Council Biostatistics Unit, Cambridge.

Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228-237.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer, Berlin.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 1, 21-31.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.