Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Small area estimation of the number of firms' recruits by using multivariate models for count data

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

by Maria Rosaria Ferrante and Carlo Trivisano

December 2010

Small area estimation of the number of firms' recruits by using multivariate models for count data

Maria Rosaria Ferrante and Carlo Trivisano [1]

Abstract

The number of people recruited by firms in Local Labour Market Areas provides an important indicator of the reorganisation of the local productive processes. In Italy, this parameter can be estimated using the information collected in the Excelsior survey, although it does not provide reliable estimates for the domains of interest. In this paper we propose a multivariate small area estimation approach for count data based on the Multivariate Poisson-Log Normal distribution. This approach will be used to estimate the number of firm recruits both replacing departing employees and filling new positions. In the small area estimation framework, it is customary to assume that sampling variances and covariances are known. However, both they and the direct point estimates suffer from instability. Due to the rare nature of the phenomenon we are analysing, counts in some domains are equal to zero, and this produces estimates of sampling error covariances equal to zero. To account for the extra variability due to the estimated sampling covariance matrix, and to deal with the problem of unreasonable estimated variances and covariances in some domains, we propose an "integrated" approach where we jointly model the parameters of interest and the sampling error covariance matrices. We suggest a solution based again on the Poisson-Log Normal distribution to smooth variances and covariances. The results we obtain are encouraging: the proposed small area estimation model shows a better fit when compared to the Multivariate Normal-Normal (MNN) small area model, and it allows for a non-negligible increase in efficiency.

Key Words: Multivariate Poisson-Log Normal distribution; Zero counts; Generalized Variance Function; Hierarchical Bayesian models.

1. Introduction

The number of people recruited by firms for a certain period can be taken as a key indicator of ongoing changes in the economic system. To highlight the dynamic of the demand for local labour, we consider the number of people recruited by firms in Local Labour Market Areas (LLMAs), these last grouped according to i) productive specialization, ii) firms' size classes and iii) industrial sector. Domains are defined by cross-classifying these three variables. In order to emphasise the signals of the reorganisation of the productive process, we focus on the numbers of "recruits replacing employees leaving the firm (substitute recruits – SR)" and "recruits filling new positions (new recruits – NR)". In Italy, information about firms' recruits is collected by the Excelsior Survey co-sponsored by the Union of Italian Chambers of Commerce (UNIONCAMERE), the Ministry of Labour and the European Union. Unfortunately, this survey does not provide reliable estimates of firms' recruits for each of these domains due to small domain sample size. As a consequence, a small area estimation (SAE) technique has to be adopted in order to obtain estimates with an acceptable degree of variability.

In this paper, we propose a SAE approach for the estimation of counts. Due to data constraints, we adopt an aggregated area-level model.

Since we aim at estimating SR and NR, we adopt a multivariate SAE model that borrows strength not only from areas but also from the correlations between the NR and SR true values. In order to estimate the median income of different sized groups of families, Fay (1987) proposed a multivariate regression model in an Empirical Bayes context. Multivariate SAE approaches have also been developed by Ghosh, Nangia and Kim (1996) and Datta, Fay and Ghosh (1991), Datta, Ghosh, Nangia and Natarajan (1996) and Datta, Lahiri, Maiti and Lu (1999) for continuous data in the hierarchical cross-section time series model framework. Fabrizi, Ferrante and Pacei (2005, 2008) adopted multivariate area level models to estimate a vector of continuous poverty parameters. As in the univariate Fay-Herriot model (Fay and Herriot 1979), all of the papers mentioned above assume the use of small area normal sampling and linking models.

Since the sampling correlations between SR and NR estimators are mainly negative, we propose a SAE model based on the Multivariate Poisson-Log Normal (MPLN) distribution. Unlike other multivariate distributions for counts proposed in the literature, this particular distribution allows for unconstrained (that is, both positive and negative) correlations (Aitchison and Ho 1989).

We also deal with the instability of estimators of sampling error variances and covariances. An approximately unbiased estimate of the variance of direct estimators is

[1] Maria Rosaria Ferrante, Department of Statistics - University of Bologna, Italy. E-mail: maria.ferrante@unibo.it; Carlo Trivisano, Department of Statistics - University of Bologna, Italy. E-mail: carlo.trivisano@unibo.it.

usually available in SAE. However, in area-level models it is customary to assume that the sampling variance is known and equal to its estimate (Rao 2003; page 76). This assumption is commonly stated and largely accepted in the case of large samples, whereas both the variance estimator and direct point estimators suffer from instability in the case of small samples. As a partial solution, sampling variance estimates are often smoothed through the generalized variance functions (GVF) approach (Wolter 1985). In You, Rao and Gambino (2003), sampling variances and covariances were smoothed over areas and times. In order to consider the extra variability associated with the estimated sampling variances, Arora and Lahiri (1997) proposed an integrated Hierarchical Bayes (HB) smoothing approach for continuous data. See You and Chapman (2006), Liu, Lahiri and Kalton (2007) and You (2008) for different extensions of Arora and Lahiri (1997).

Due to the rarity of recruits in certain domains, a further problem arises that is linked to the instability of sampling error variances and covariances estimators. When direct estimates of SR or NR (or both) are equal to zero, estimated sampling error variances and covariances are also equal to zero. Note that observing estimated variances equal to zero does not necessarily imply that the estimates have a high degree of accuracy. This problem was encountered in previous small area estimation problems (*e.g.*, Elazar 2004; Chattopadhyay, Lahiri, Larsen and Reimnitz 1999). Chen (2001) proposed a unit level hierarchical modeling to handle the problem. Moreover, some studies (Cohen 2000) use the logarithmic transformation of the mean (or total) direct estimates of the count data in order to adopt a linear SAE model, simply discarding the estimates equal to zero. Although this solution overcomes the "zero variance" problem, it also leads to biased estimates and neglects a portion of the sample.

In order to deal with the instability of variances and covariances estimators as well as the problem of estimated sampling variances equal to zero, we suggest an "integrated" approach in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). Within an HB framework, we jointly model the parameters of interest and the sampling error covariance matrices by adopting a smoothing covariance solution based once again on the Poisson-Log Normal distribution.

The layout of this paper is as follows. The data set employed is described in section 2, while section 3 presents direct domain estimation and its associated sampling error variances and covariances. In section 4, we describe the multivariate SAE model we propose for estimating counts as well as the solution we suggest for overcoming the instability of sampling error variances and covariances estimators in the presence of zero counts. Section 5 reports the results obtained by measuring the performance of the adopted SAE model. Details on the Poisson-Log Normal distribution are given in the Appendix.

2. The excelsior survey

The Excelsior Survey is one of the most complete Italian statistical sources for labour demand data, providing estimates of the number of people recruited by Italian firms. Each year, a stratified simple random sample of about 100,000 firms with at least one employee is contacted and asked about the number of people it plans to hire in the short term. The factors used for stratification are the firm's industrial sector and size class. The allocation of the sample in the strata satisfies a constraint on the maximum estimated standard error corresponding to a 95% significance level (Baldi, Bellisai, Fivizzani and Sorrentino 2007). By focusing on local geographical details, the survey is designed to produce reliable estimates for the administrative provinces (NUTS3, following the "Nomenclature of Units for Territorial Statistics" reported in http://europa.eu.int/comm/eurostat/ramon/nuts). This geographical unit, singled out on the basis of administrative criteria, does not appear to be the best choice when analysing the dynamics of the local labour demand. In order to shed some light on the signals of the reorganization of the local productive process, a better territorial subdivision would be LLMAs (following the OECD definition). LLMAs are groups of municipalities sharing the same labour market conditions (for the location of LLMAs in Italy, see Sforzi 1991). In Italy, following the strategy proposed by Sforzi and Lorenzini (2002) and adopted by the Italian Statistical Institute (ISTAT), certain LLMAs are labelled "industrial districts" (IDs). IDs are geographically defined productive systems characterized by a dominant specialization. In the 1990s, these were considered to be the main stimulus for the growth of the Italian economic system (Becattini 1992).

Estimating the number of substitute and new recruits in firms operating within/outside of IDs can help us verify whether IDs are still a source of dynamism for the Italian economy as a whole. In order to refer to types of ID, we group them according to their productive specialization. Similarly, LLMAs not labelled as IDs can be classified according to their economic vocation (LLMAs can be characterized by a specific manufacturing activity, tourist area, city, *etc*.). Moreover, the comparison between ID and non-ID firms makes economic sense if the industrial sector and size of the firms are also taken into account. Finally, as already noted, domains of interest are defined by cross-classifying: i) groups of LLMAs obtained according to their productive specialization, ii) firm's industrial sector and iii) firm's size.

This paper focuses on the manufacturing sector characterising the IDs' economic activity. The analysis is limited to two Italian regions containing a large quantity of IDs, namely Tuscany and Emilia-Romagna, and to firms with fewer than 100 employees (as censuses are taken for the other size classes). The target population consists of 54,089 firms employing a total of 809,059 people.

3. Direct estimates

Table 1 provides details of the categories defining the 208 domains of interest. Note that the number of domains is less than that expected due to the absence of a number of domains within the population. The domains are unplanned since they are formed grouping LLMAs contained in the same planned stratum. For the sake of simplicity, in the following we avoid using the stratum subscription wherever possible.

Let θ_{i1} and θ_{i2} be the true number of NR and SR for domain $i\ (i = 1, \ldots, 208)$, respectively. We shall first define a direct estimator of $\theta_{ij}\ (i = 1, \ldots, 208;\ j = 1, 2)$. Let y_{ijl} be the response of the l^{th} unit related to the j^{th} variable in the i^{th} domain ($l = 1, \ldots, n_i$, where n_i is the sample size in domain i; $i = 1, \ldots, 208$; $j = 1, 2$). As design based (direct) estimator we use a ratio domain estimator defined as $\hat{\theta}_{ij} = \sum_{l=1}^{n_i} y_{ijl} / (n_i / N_i)\ N_i / \hat{N}_i$, where N_i and n_i are respectively the population size and the sampling size referred to domain i, and $\hat{N}_i = n_i / n_{t \ni i}\ N_{t \ni i}$, where $N_{t \ni i}$ and $n_{t \ni i}$ are respectively the population size and the sampling size of the stratum t containing the domain i (Särndal, Swensson and Wretman 1992; page 391).

Since we are estimating the number of occurrences of rare events, in 50 of the 208 domains, direct estimates of NR and/or of SR are equal to zero, that is, $\hat{\theta}_{i1} = 0$ and/or $\hat{\theta}_{i2} = 0$. Zero point estimates imply that $\hat{V}(\hat{\theta}_{i1}) = 0$ and/or $\hat{V}(\hat{\theta}_{i2}) = 0$, where $\hat{V}(\hat{\theta}_{i1})$ and $\hat{V}(\hat{\theta}_{i2})$ are the standard design-based variance estimates of $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$, respectively. This result gives a false impression of high accuracy, whereas the exact opposite is more likely to be true in a small area context. Moreover, design based estimates of NR and/or of SR equals to zero produce $\widehat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$, where $\widehat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$ denotes the standard design-based estimate of the design-based covariance between $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$. As a result, covariances also need to be smoothed in a multivariate SAE model.

We hereafter refer to the set of the 50 small areas having one or both zero estimated variances and zero covariances as the "Zero Count" (ZC) set. The complementary set of 158 domains, where $\hat{V}(\hat{\theta}_{i1}) > 0$ and $\hat{V}(\hat{\theta}_{i2}) > 0$, is named the "Non Zero Count" (NZC) set.

Considering the data generating process and the nature of the outcome variables, we expect mainly negative correlations between θ_{i1} and θ_{i2}. Briefly, we need a suitable distribution for both smoothing covariance matrices and modeling small area parameters that allows for an unrestricted covariance matrix, that is, for both positive and negative correlations.

Table 1
Variables defining domains of interest

LLMAs grouped by productive specialization	Firm size [(b)]	Industrial sector[(a)]
Industrial district[(a,c)]	1-9	1 Food, beverages and tobacco
Food, beverages and tobacco	10-49	2 Textiles and clothing
Textiles and clothing	50-99	3 Paper products, printing and publishing
Paper products, printing and publishing	≥ 100	4 Machinery
Machinery		5 Chemicals and basic metals
Jewellery, musical instruments, games, *etc.*		6 Leather and footwear
Leather and footwear		7 Wood, furniture and household equipment
Wood, furniture and household equipment		8 Jewellery, musical instruments, games, *etc.*
LLMAs not defined as district [(c)]		9 Builders, contractors
Non-specialised manufacturing		10 Other manufacturing
Non-specialized, excluding manufacturing		
Tourist		
Cities		

(a) As defined by the 2-digit ATECO 91-ISIC 3 level classification and by Sforzi (1991).
(b) Defined according to the number of employees.
(c) Defined in accordance with Istat (1997).

4. An integrated multivariate small area model for count data

Multivariate count data can have a non-trivial correlation structure. In general, the modeling of this structure significantly affects the estimators' efficiency and the computation of correct standard errors. A number of multivariate models for count data have been proposed in the literature, such as the Multivariate Poisson, Multivariate Negative Binomial and Multivariate Poisson-Gamma Mixture models (for a review of such models, see Winkelmann 2003). Unfortunately, these distributions are not suitable for modeling our data since they are based on the hypothesis that correlation is the result of an individual factor that does not vary across outcomes, thus implying a covariance structure restricted to non-negative correlations. In the bivariate case, a more flexible covariance structure is provided by the Latent Poisson Normal distribution (van Ophem 1999); however, any extensions to higher dimensional multivariate data appear impractical.

Aitchison and Ho (1989) proposed a d-variate distribution that allows for an unrestricted covariance structure, the Multivariate Poisson-Log Normal distribution (MPLN). No closed form exists for this distribution, but it can be represented as a simple mixture allowing for parameter estimation in an MCMC approach (Chib and Winkelmann 2001). Details of the MPLN distribution are provided in the Appendix.

4.1 Smoothing sampling covariance matrices

As previously mentioned, the instability of standard errors in SAE is usually dealt with using a GVF approach. In this section, we present a GVF model with a regression function inspired by the MPLN distribution.

Let $\mathbf{y}_{il} = [y_{i1l}, y_{i2l}]'$ be the vector of the two outcome variables referring to the l^{th} unit in the i^{th} domain. Let $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \perp \mathbf{y}_{il'} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i$ and $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \sim \text{PLN}_2(\boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i)$, $\forall i,\ \forall l$. Under these hypotheses, the moments leading up to the second order can be expressed as follows:

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

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

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

where $\sigma_{i,\,jh}$ denotes the (j, h), $j, h = 1, 2,$ element of $\boldsymbol{\Sigma}_i$.

To deal with the problem of smoothing covariance matrices, Otto and Bell (1995), suggested an approach based on a Wishart distributional assumption; specifically, they used smoothed estimates in a small area Normal-Normal model. In the same spirit, we propose a Bayesian approach using the following GVF strategy. Under simple random sampling, let us assume that the sampling covariance matrix in domain i, $\mathbf{C}_i$ follows a Wishart distribution with $n_i - 1$ degrees of freedom:

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

where $\boldsymbol{\Gamma}_i = E(\mathbf{C}_i | n_i, \boldsymbol{\Gamma}_i)$, $i = 1, 2, ..., 158$, and elements (j, h) of $\mathbf{C}_i$ are defined as $C_{i,\,jh} = n_i^{-1}\sum_{i=1}^{n_i}(y_{ijl} - \bar{y}_{ij})(y_{ijh} - \bar{y}_{ih})$, where $\bar{y}_{ij} = n_i^{-1}\sum_{i=1}^{n_i} y_{ijl}$.

If ζ_{ij} parameters are known, then $E(\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i)$ only depends on elements of the $\boldsymbol{\Sigma}_i$ matrix. We propose to estimate ζ_{ij} using the design based estimator $\hat{\zeta}_{ij} = N_i^{-1}\hat{\theta}_{ij}$. Thus, we can express each element of the $\boldsymbol{\Gamma}_i$ matrix as a function of estimates $\hat{\zeta}_{ij}$ and of the elements of the $\boldsymbol{\Sigma}_i$ matrix:

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

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

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

where $\sigma_{i,11} = \bar{\boldsymbol{\sigma}}_{11}'\mathbf{Z}_i$, $\sigma_{i,22} = \bar{\boldsymbol{\sigma}}_{22}'\mathbf{Z}_i$, $\sigma_{i,12} = \bar{\boldsymbol{\sigma}}_{12}'\mathbf{Z}_i$, being $\mathbf{Z}_i$ is a 3×1 vector of dummy variables identifying the firm's size class in the domain i, and

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

that is, we assume that parameters $\boldsymbol{\Sigma}_i$ are equal for domains belonging to the same firm size class.

We estimate $\bar{\boldsymbol{\sigma}}_{11}, \bar{\boldsymbol{\sigma}}_{22}, \bar{\boldsymbol{\sigma}}_{12}$ parameters on NZC data. Since we are following a Bayesian approach, prior specifications for $\bar{\sigma}_{k,\,jj}$ and $\bar{\sigma}_{k,12}$ $k = 1, 2, 3$ are needed. We use the following prior specifications: $\bar{\sigma}_{k,11}^{1/2} \sim U^+$, $\bar{\sigma}_{k,22}^{1/2} \sim U^+$, $\bar{\rho}_k \sim U(-1,1)$, where $\bar{\sigma}_{k,12} = \bar{\rho}_k(\bar{\sigma}_{k,11}\bar{\sigma}_{k,22})^{1/2}$ and U^+ denotes a uniform distribution over a subset of R^+ with a large but finite length. In section 4.3, we show how these estimates can be used to integrate the SAE model with a model for sampling error covariance matrices.

4.2 A Multivariate Normal-Poisson-Log Normal small area model

In this section, we propose a multivariate SAE model based on the MPLN distribution in order to jointly estimate SR and NR using the NZC set.

Let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T$ be the vector of the two parameters of interest for the i^{th} domain in the set of NZC data $(i = 1, ..., 158)$, and let $\hat{\boldsymbol{\theta}}_i$ be the corresponding vector of direct estimates. The SAE model consists of two separate models. The first model is a sampling model:

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

As in Lahiri and Rao (1995), we justify the normality assumption in (1) using the central limit argument. It is standard practice to assume that sampling error covariance matrices $\boldsymbol{\Psi}_i$ are known, and a GVF method is generally used to estimate $\boldsymbol{\Psi}_i$. Here, as a smoothed estimation of $\boldsymbol{\Psi}_i$ we adopt $\hat{\boldsymbol{\Psi}}_i = E(\boldsymbol{\Gamma}_i|\mathbf{C}_i, n_i)K_i$, where $K_i = N_i(N_{tэi}/n_{tэi} - 1)$. From this point on we will refer to $\hat{\boldsymbol{\Psi}}_i$ as Smoothed Sampling Error Covariance matrix (SMSEC).

The second component of the SAE model is a linking model that relates $\boldsymbol{\theta}_i$ to area specific auxiliary data:

$$\boldsymbol{\theta}_i \sim \text{ind PLN}_2(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_\text{v}), \quad i = 1, ..., 158,$$

where (2)

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\beta}\mathbf{Z}_i x_i$$

$\mathbf{Z}_i$ is a 3×1 vector of dummy variables identifying the firm's size class in the domain i and $x_i = \log(x_i^*)$, where x_i^* is the number of employees in the domain i.

At the end, $\boldsymbol{\Sigma}_\text{v}$ is the covariance matrix related to the area-specific random effects:

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

and

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

From here on, we refer to this small area model as "Multivariate Normal-Poisson-Log Normal" (MNPLN).

We adopt a fully hierarchical Bayesian approach. In this framework, relatively complex (*e.g.*, multivariate) models can be implemented easily; in addition, posterior distributions can be approximated using MCMC algorithms. Computing small area multivariate estimates, and estimates of their MSE in particular, can be difficult within a frequentist approach. The specification of priors for the described model is as follows:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim N_2(\mathbf{0}, a\mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \sim N_2(\mathbf{0}, g_{k'}\mathbf{I}_2) \qquad k' = 2,3,$$

$$\begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} \sim N_2(\mathbf{0}, b_k\mathbf{I}_2) \qquad k = 1, 2, 3,$$

$$\boldsymbol{\Sigma}_\text{v}^{-1} \sim W(s, \mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \perp \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix},$$

where $s = 3$ and a, $g_{k'}$, b_k are large compared with the scale of the data. This is to reflect the lack of prior information about model parameters, thus defining diffuse but proper specification of priors. The posterior means $\hat{\boldsymbol{\theta}}_i^{\text{HB}} = E(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ are taken as estimators of the area parameters, while the posterior variance $V(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ is used as a measure of uncertainty.

For the sake of comparison, we take the standard Multivariate Normal-Normal (MNN) model as a benchmark, where the sampling model is defined as in (1) and the linking model is defined as follows:

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

where $\boldsymbol{\mu}_i^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^*\mathbf{Z}_i + \boldsymbol{\beta}^*\mathbf{Z}_i x_i^*$. Parameters $\boldsymbol{\alpha}^*$, $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ and their prior distributions are defined as $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the previous model.

4.3 An integrated MNPLN small area model

In order to account for the extra variability due to the estimated covariance matrices of sampling errors, as well as to overcome the zero variances and covariances problem, we suggest a solution in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). We integrate the model for sampling error covariance matrices of section 4.1 into SAE models (1) and (2). Thus, we here refer to the whole set of 208 domains.

In this context, the small area sampling model is formulated as usual, that is, $\hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i \sim \text{ind } N_2(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i^*)$, $i =$ 1, ..., 208. Under the hypotheses regarding $\mathbf{y}_{il}$ formulated in section 4.1, assuming that the $\boldsymbol{\Sigma}_i$s are known and assuming that $\theta_{ij} = N_i\zeta_{ij}$, the elements of the sampling error covariance matrix $\boldsymbol{\Psi}_i^*$ can be expressed as follows:

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

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

where $\hat{\bar{\boldsymbol{\sigma}}}'_{jj}$ $j = 1, 2$ and $\hat{\bar{\boldsymbol{\sigma}}}'_{12}$ are posterior means of parameters $\bar{\boldsymbol{\sigma}}_{jj}$ and $\bar{\boldsymbol{\sigma}}_{12}$, respectively, computed using the model of section 4.1.

Since the sampling error covariance matrices are expressed as a function of the $\boldsymbol{\theta}_i$ parameters, here they can be considered Model Based Sampling Error Covariances (MBSEC). The posterior means $\hat{\boldsymbol{\theta}}_i^{\text{HB}} = E(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i)$ are taken as estimators of $\boldsymbol{\theta}'_i$s, while the posterior variance $V(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i)$ is used as a measure of uncertainty.

We note that the MNN model cannot be implemented following the integrated approach described above. In fact, (3) does not ensure the positivity of $\boldsymbol{\theta}_i$ nor of the diagonal elements of $\boldsymbol{\Psi}_i$ as a result.

5. Data analysis

In section 5.1, we compare the MNPLN model with the benchmark MNN model and their univariate counterparts. We assume SMSEC for both models; we thus refer to the two strategies as MNPLN-SMSEC and MNN-SMSEC from here on. Since these models do not allow us to deal with the zero count problem, we refer this analysis to the NZC set. In section 5.2, we compare the SAE integrated strategy based on the MNPLN model and MBSEC (MNPLN-MBSEC), which we presented in Section 4.3, with the strategy based on the MNPLN-SMSEC. We limit the analysis to the NZC set in order to evaluate the two strategies under the same conditions. Finally, in section 5.3 we evaluate the overall performance of the proposed SAE model MNPLN-MBSEC for the whole data set (NZC+ZC).

Posterior distributions of parameters were obtained for all models, using Monte Carlo integration via the Gibbs sampling algorithm. We used the MCMC software WinBUGS (Spiegelhalter, Thomas, Best and Gilks 1995) to run three parallel chains (each with 25,000 runs), the starting point being drawn from an over-dispersed distribution. WinBUGS codes are available at the URL http://www2.stat.unibo.it/trivisano/. The convergence of the Gibbs sampler was monitored by visual inspection of the chains' plots and of autocorrelation diagrams, and by means of the potential scale reduction factor proposed by Gelman and Rubin (1992). Although all models displayed fast convergence, we discarded the first 5,000 iterations from each chain. In multivariate models, the fairly strong autocorrelation of chains is reduced by thinning the chain (1 out of every 3 values has been considered for posterior summaries). See Rao (2003, pages 228-232) for details.

The performances of the small area models discussed in sections 4.2 and 4.3 are compared using various measures. In order to choose among competing models, we computed the Deviance Information Criterion (DIC). The DIC is a model selection criterion according to which a model's performance is evaluated as the sum of a measure of fit (the posterior mean of the deviance $\bar{D}$) and a measure of complexity obtained as the difference between $\bar{D}$ and the deviance evaluated at the parameters' posterior mean. In this way, a model is preferred if it displays a lower DIC value (Spiegelhalter, Best, Carlin and Van der Linde 2002).

In order to verify the strength of the multivariate approach to SAE, we use as a benchmark the univariate versions of models discussed in sections 4.2 and 4.3, defined as follows. For all models, we set $\sigma_{\nu,12} = 0$ in $\boldsymbol{\Sigma}_{\nu}$, and we assume $\sigma_{\nu,11} \perp \sigma_{\nu,22}$, $\sigma_{\nu,jj}^{1/2} \sim U(0, U^{+})$, $j = 1, 2$. For SMSEC models, we set $\boldsymbol{\Psi}_i = \text{diag}(\hat{\boldsymbol{\Psi}}_i)$, while for MBSEC models we set $\sigma_{1,12} = 0$ in (5). In addition, a new set of estimates for parameters $\bar{\boldsymbol{\sigma}}_{11}$ and $\bar{\boldsymbol{\sigma}}_{22}$ is obtained by setting $\bar{\rho}_k = 0$ in the model of section 4.1.

Table 2 reports the DIC results for the whole set of small area models.

Table 2
Model comparison using DIC statistic

Model	Data set	DIC
MNN-SMSEC	NZC	2,742.2
(univariate version)	NZC	2,745.4
MNPLN-SMSEC	NZC	2,656.9
(univariate version)	NZC	2,661.0
MNPLN-MBSEC	NZC	2,623.6
(univariate version)	NZC	2,638.1
MNPLN-MBSEC	NZC+ZC	3,202.7
(univariate version)	NZC+ZC	3,214.3

All the multivariate models considered perform better in terms of DIC than their univariate counterparts (Table 2). In addition, for all multivariate models we find that posterior credibility intervals of $\rho_{\nu} = \sigma_{\nu,12} / \sqrt{\sigma_{\nu,11}\sigma_{\nu,22}}$ do not contain zero. We thus focus on multivariate models in the following paragraphs.

We checked the adequacy of the specified multivariate models using posterior predictive checks. Simulated values of a suitable discrepancy measure are generated from the posterior predictive distribution and are then compared with the values of the same measure computed from observed data. Let $\hat{\boldsymbol{\theta}}_{\text{obs}}$ and $\hat{\boldsymbol{\theta}}_{\text{new}}$ denote the observed and generated data, respectively. The posterior predictive p-value is defined as $p = P\{d(\hat{\boldsymbol{\theta}}_{\text{new}}, \boldsymbol{\theta}) > d(\hat{\boldsymbol{\theta}}_{\text{obs}}, \boldsymbol{\theta}) \mid \hat{\boldsymbol{\theta}}_{\text{obs}}\}$. We consider a discrepancy measure proposed in Datta *et al.* (1999), which is defined as

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

Computing the p-value is straightforward using the MCMC output. Extreme values of the probability p indicate a given model's lack of fit. Following Rao (2003, page 245-246) and You and Rao (2002), we computed two statistics that are useful in order to assess model fit at the individual domain level. The first statistic, $p_{ij}^{*} = P(\hat{\theta}_{ij,\text{new}} < \hat{\theta}_{ij,\text{obs}} | \hat{\boldsymbol{\theta}}_{\text{obs}})$, provides information about the degree of consistent overestimation or underestimation of $\hat{\theta}_{ij,\text{obs}}$.

The second statistics is defined as

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

where expectation and variance are under the posterior predictive distribution. Table 3 summarizes results relative to $p,\ p_{ij}^{*}$ and d_{ij}^{*}.

To further check the consistency of the data, we calculated direct and model-based estimates of ${}_{A}\theta_{sj}, s = 1, \ldots, 10,$ that is, the total number of NR and SR for the ten domains identified by classifying firms only according to the industrial sector. Let $w_{is} = 1$ if the number of recruits in the domain i refers to the industrial sector s and $w_{is} = 0$; otherwise, then

$$ {}_{A}\theta_{sj} = \sum_{i} \theta_{ij} w_{is}. \tag{7}$$

At this level of aggregation, direct estimates can be considered accurate. Consequently, given two sets of model-based estimates referring to these large domains, we prefer the one that agrees with the direct estimates. Domains identified by industrial sectors are planned in the Excelsior Survey; each industrial sector is stratified according to firm size. Therefore, direct estimates ${}_{A}\hat{\theta}_{sj}$ for each industrial sector are calculated using the standard Horwitz-Thompson estimator. Aggregated model-based estimates are computed based on the MCMC output. For models referring to NZC data, we aggregated following (7) at each MCMC step $t, t = 1, \ldots, T,$ with samples ${}^{t}\theta_{ij}^{*}$ and ${}^{t}\theta_{ij}^{**}$ generated respectively from the posterior distribution of θ_{ij} for domains belonging to the NZC set and from the predictive distribution of θ_{ij} for domains belonging to the ZC set. The HB estimator is defined as ${}_{A}\hat{\theta}_{sj}^{\text{HB}} = T^{-1}\sum_{t=1}^{T}(\sum_{i \in NZC} {}^{t}\theta_{ij}^{*} w_{is} + \sum_{i \in ZC} {}^{t}\theta_{ij}^{**} w_{is})$. Otherwise, for the model on NZC+ZC data, we aggregated following (7) MCMC samples from the posterior distributions of θ_{ij}. In this case, the HB estimator is defined as ${}_{A}\hat{\theta}_{sj}^{\text{HB}} = T^{-1}\sum_{t=1}^{T}(\sum_{i \in NZC} {}^{t}\theta_{ij}^{*} w_{is})$. Table 4 reports summaries of ${}_{A}\hat{\theta}_{sj}$ and ${}_{A}\hat{\theta}_{sj}^{\text{HB}}$.

For all the multivariate models, we examined the following variants of the prior distributions: independent non-informative flat prior distributions were used for the elements of vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^{*}, \boldsymbol{\beta}^{*},$ and $\boldsymbol{\gamma}^{*}; \sigma_{\nu,\, jj}^{1/2} \sim U^{+},\ j = 1, 2,\ \rho_{\nu} \sim U(-1,1),\ \sigma_{\nu,12} = \rho_{\nu}\,(\sigma_{\nu,12}\,\sigma_{\nu,12})^{1/2}$. We do the same for the elements of matrix $\boldsymbol{\Sigma}^{*}$ in the MNN model. We did not find any relevant changes in the posterior distributions of parameters of interest.

5.1 Comparing the MNPLN-SMSEC and MNN-SMSEC models on the NZC set

We find that the MNPLN-SMSEC model largely outperforms the MNN-SMSEC one in terms of DIC (Table 2). This last model shows a lack of fit as it displays a p-value equal to 0.034 (Table 3), whereas a value of 0.65 suggests the adequacy of the MNPLN-SMSEC model. This finding is confirmed when p_{ij}^{*} and d_{ij}^{*} measures (Table 3) for the two models are compared. For the MNN-SMSEC model, p_{ij}^{*} ranges over domains from 0.000 to 0.995 for NR $(j = 1)$ and from 0.003 to 0.993 for SR $(j = 2)$, respectively, indicating overestimation and underestimation in some domains. In addition, summaries of the standardized residuals d_{ij}^{*} indicate that there are predicted values outside two standard deviations of the corresponding observed values. The same measures for the MNPLN-SMSEC model indicate an adequate fit.

We also find that the MNPLN-SMSEC model outperforms the MNN-SMSEC models when performances are evaluated with reference to estimates for large domains (Table 4). In fact, credibility intervals for the MNN-SMSEC only cover 2 aggregated direct estimates for NR and 4 for SR, while credibility intervals under the MNPLN-SMSEC cover 6 aggregated direct estimates for NR and 6 for SR.

Table 3
Posterior predictive checks; summaries of p_{ij}^{*} and d_{ij}^{*} calculated with respect to i

Model	Data set	p		p_{i1}^{*}	p_{i2}^{*}	d_{i1}^{*}	d_{i2}^{*}
			min	0.000	0.003	-3.764	-2.867
MNN-SMSEC	NZC	0.034	median	0.591	0.616	0.257	0.295
			max	0.995	0.993	2.656	-2.515
			min	0.154	0.129	-0.965	-1.165
MNPLN-SMSEC	NZC	0.65	median	0.535	0.561	0.124	0.149
			max	0.891	0.912	1.216	1.286
			min	0.090	0.134	-1.085	-0.983
MNPLN-MBSEC	NZC	0.78	median	0.515	0.519	-0.084	-0.085
			max	0.916	0.914	1.401	1.787
			min	0.072	0.111	-1.164	-0.945
MNPLN-MBSEC	NZC+ZC	0.79	median	0.506	0.523	-0.076	-0.094
			max	0.903	0.913	1.301	1.778

Table 4
Direct and HB estimates for industrial sectors; in italic HB estimates whose credibility intervals cover direct estimates

	Direct estimates		**HB estimates**												
			MNN-SMSEC (NZC)			**MNPLN-SMSEC (NZC)**			**MNPLN-MBSEC (NZC)**			**MNPLN-MBSEC (NZC+ZC)**		
s	${}_A\hat{\theta}_{s1}$	$\text{se}({}_A\hat{\theta}_{s1})$	${}_A\hat{\theta}_{s1}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s1}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s1}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s1}^{\text{HB}}$	95% cred int.	
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*
	${}_A\hat{\theta}_{s2}$	$\text{se}({}_A\hat{\theta}_{s2})$	${}_A\hat{\theta}_{s2}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s2}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s2}^{\text{HB}}$	95% cred int.		${}_A\hat{\theta}_{s2}^{\text{HB}}$	95% cred int.	
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*

5.2 Comparing the MNPLN-SMSEC and MNPLN-MBSEC models on the NZC set

Values of p, p_{ij}^* and d_{ij}^* are approximately comparable for the MNPLN-SMSEC and MNPLN-MBSEC models (Table 3). Likewise, model-based estimates produced by MNPLN-SMSEC assume values very close to those obtained using MNPLN-MBSEC; in fact, the correlation between the posterior means of θ_{i1} under the two models is equal to 0.98, while the same measure referring to θ_{i2} is equal to 0.94. The same results arise for the correlation between posterior standard errors, which are 0.92 and 0.94, respectively. Performances of the MNPLN-MBSEC model in terms of agreement with direct estimates of large domains (Table 4) are slightly better than those of the MNPLN-SMSEC model: respectively, 7 direct estimates of *NR* and 8 of SR are covered by the credibility interval calculated under this model.

Given these results, we conclude that the fit of the MNPLN-MBSEC model is adequate.

5.3 Evaluating the performances of MNPLN-MBSEC models on the NZC+ZC set

We observe that the performances of the MNPLN-MBSEC model on the whole dataset in terms of p, p_{ij}^* and d_{ij}^* measures are satisfactory and comparable with those of the same model on the NZC data set (Table 3). Obviously, DIC values for the two models cannot be compared as the two models are estimated on different data sets.

As can be seen in Table 4, all the credibility intervals calculated using this model cover direct estimates referring to large domains; in other words, the agreement of HB estimates with direct estimates is very satisfactory. This result can be explained by noting that zero counts are more probable in small domains, which are characterized by a small number of employees (the covariate in all models). Therefore, estimating models on NZC data can lead to biased estimates of parameter $\boldsymbol{\beta}$. We conclude that integrating a sampling covariance model into the MNPLN small area model leads to an appreciable increase in the reliability of small area estimates. To describe the efficiency gain of the HB estimates, we computed on the NZC set the average percent CV reduction (You 2008), defined as the average of the difference of the direct CV and HB CV (the ratio of the square root of the posterior variance and the posterior mean) relative to direct CV. The average CV reduction is 23.1% for NR and 29.1% for SR.

Acknowledgements

The authors would like to thank the Editor, Associate Editor and Referee for their helpful comments and suggestions. The research leading to this paper was partially supported by Miur-PRIN 2003/2003133249 and Miur-Prin 2008/2008CEFF37-001.

Appendix

The Multivariate Poisson-Log Normal distribution

Let $\mathbf{y} = (y_1, y_2, ..., y_j, ..., y_d)$ be a d-dimensional vector of counts, and suppose that $y_j|\tau_j \sim \text{Po}(\tau_j)$, with $y_j|\tau_j \perp y_{j'}\,|\,\tau_{j'}\,(j \neq j')$. Let the vector of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_j, ..., \tau_d)$ follow a multivariate Log Normal, that is, $\boldsymbol{\tau}\,|\,\boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{LN}_d\,(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$, where $\boldsymbol{\lambda} = E(\log \boldsymbol{\tau})$ and $\boldsymbol{\Sigma} = \text{COV}(\log \boldsymbol{\tau})$. Then the marginal distribution of $\mathbf{y}$ is a Multivariate Poisson-Log Normal (MPLN) distribution, which is a log normal mixture of d independent $\text{Po}(\tau_j)$, that is, $\mathbf{y}\,|\,\boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{PLN}_d\,(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. By denoting the $(j, h), j, h = 1, 2, ..., d$ element of $\boldsymbol{\Sigma}$ as σ_{jh}, marginal moments can be obtained easily through conditional expectation results and the standard properties of the Poisson and Log Normal distributions:

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

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

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

Note that the MPLN model allows for overdispersion provided that $\sigma_{jj} > 0$, thus leading to $V(y_j\,|\,\boldsymbol{\lambda}, \boldsymbol{\Sigma}) > E(y_j\,|\,\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. Moreover, the correlation structure of counts is unrestricted, since $\text{COV}(y_j, y_h\,|\,\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ can be either positive or negative depending on the sign of σ_{jh}. Aitchison and Ho (1989), as well as Good and Pirog-Good (1989), studied a bivariate MPLN distribution, albeit exclusively in cases without covariates. However, the same model can easily be extended to take covariates into consideration (Chib and Winkelmann 2001).

References

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

Arora, V., and 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. and 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. In *Industrial Districts and International Co-operation in Italy*, (Eds., F. Pyke, G. Becattini and W. Sengenberger). Internation Labor Office, Geneva.

Chattopadhyay, M., Lahiri, P., Larsen, M. and Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas. *Survey Methodology*, 25, 81-86.

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

Chib, S., and 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. and 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. and Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. In *Bayesian Analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner and J.M. Geweke). New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. and 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. and 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. and 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. In *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: John Wiley & Sons, Inc., 91-102.

Fay, R.E., and 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., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.

Ghosh, M., Nangia, N. and 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., and 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., and 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. and 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., and 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. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.

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

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

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

Spiegelhalter, D.J., Thomas, A., Best, N.G. and 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). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You, Y., and 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. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.