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

Article

# Estimates for small area compositions subjected to informative missing data


by Li-Chun Zhang

December 2009

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 191-201
Statistics Canada, Catalogue No. 12-001-X



# Estimates for small area compositions subjected to informative missing data

**Li-Chun Zhang** [1]

Abstract

Estimation of small area (or domain) compositions may suffer from informative missing data, if the probability of missing varies across the categories of interest as well as the small areas. We develop a double mixed modeling approach that combines a random effects mixed model for the underlying complete data with a random effects mixed model of the differential missing-data mechanism. The effect of sampling design can be incorporated through a quasi-likelihood sampling model. The associated conditional mean squared error of prediction is approximated in terms of a three-part decomposition, corresponding to a *naive* prediction variance, a positive correction that accounts for the hypothetical parameter estimation uncertainty based on the latent complete data, and another positive correction for the extra variation due to the missing data. We illustrate our approach with an application to the estimation of Municipality household compositions based on the Norwegian register household data, which suffer from informative under-registration of the dwelling identity number.

Key Words: Conditional MSE of prediction; EMPQL algorithm; Generalized SPREE; Not missing-at-random; Two-way contingency table.

## 1. Introduction

Small area (or domain) population counts cross-classified by various social-economic characteristics are increasingly demanded for fund allocation, regional planning and social-economic research. Purcell and Kish (1980) outlined the so-called "Structure preserving estimation" (SPREE), which operates by modifying the small area estimates in a way so that they vary from one area to another in accordance with the variation that exists in another known auxiliary table of the same dimension. Typically the auxiliary table is obtained from a previous census, or some administrative register containing similar information. Zhang and Chambers (2004) developed a generalized SPREE (GSPREE) approach. Both fixed effects and random effects mixed models were introduced, and the restricted log-linear model underlying SPREE was shown to be a special case. This provides means for reducing the potential bias of the traditional SPREE estimates. We refer to Ghosh, Natarajan, Stroud and Carlin (1998) and Longford (1999) for alternative hierarchical and empirical Bayes approaches to this type of data.

In this paper we extend the GSPREE approach to situations subjected to missing data. This can be useful in sample surveys where nonresponse is unavoidable. We concentrate on *small area compositions* that can be arranged in a two-way table, where one of the two dimensions refers to the small areas and the other refers to the categories of interest. The cell counts summarize to a fixed area total that may or may not be known. For instance, each person between 16 and 74 years of age can be classified according to the labour force status "employed", "unemployed" and "not in the labour force". The sum of the three counts inside a small area is the total number of persons between 16 and 74 years of age within this area.

In the context of small area composition we say that the missing-data mechanism is *informative* provided it varies across the categories of interest. As such it is also *not* missing-at-random (Rubin 1976). In addition, the overall rate of missing differs across the areas. Differential missingness as such leads to distortion of the underlying complete data, and bias if the estimation is carried out as if the observed data were complete. We propose a double mixed modeling approach that combines the random effects mixed model for the underlying complete data with a random effects mixed model of the missing-data mechanism. The double-smoothing approach is outlined in Section 2.

It should be noted that national statistical offices that conduct large scale surveys will have accounted for missing data by weighting adjustments or imputation. This, however, will have been done at levels that are significantly higher than the small areas, and will be for variables that do not necessarily correspond to those of interest for the small areas. When available, the adjusted totals can be incorporated into the GSPREE as marginal totals for iterative proportional fitting (IPF). But modeling of the differential probabilities of missing across the small areas will generally remain a matter of interest.

It should also be noticed that informative missing data as such makes it less straightforward to assess the potential bias of any estimation approach. SPREE may be biased on two accounts: (i) the underlying restricted log-linear assumptions are likely to be unrealistic, (ii) direct IPF may fail to account for the differential probabilities of missing

1. Li-Chun Zhang, Statistics Norway, Kongensgate 6, PB 8131 Dep. N-0033 Oslo, Norway. E-mail: lcz@ssb.no.

adequately. The proposed double mixed modeling approach deals with problem (i) by GSPREE modeling of the underlying complete data, and it deals with problem (ii) by introducing a more flexible missing-data model, as we shall discuss in Section 2.2. Nevertheless, bias is likely to persist to a certain extent. Since the estimation of model parameters and random effects is more complicated under the double mixed modeling approach, alternative estimation methods that are able to preserve the computational simplicity of SPREE, while making more adequate adjustment for informative missing data, are worth investigating in future.

When it comes to the assessment of estimation uncertainty, Booth and Hobert (1998) argued for the conditional mean squared error of prediction (CMSEP) given the observed data. We extend their approach and derive approximate CMSEP in the current multivariate incomplete-data situation. This results in a three-part decomposition of the CMSEP, corresponding to a *naive* prediction variance, a positive correction that accounts for the hypothetical estimation uncertainty of the parameters based on the latent complete data, and another positive correction for the extra variation due to the missing data. The details are given in Section 3.

Estimation procedures for the parameters, the CMSEP and the small area compositions are described in Section 4. In Section 5 we apply our approach to derive estimates of the Municipality household compositions based on the Norwegian household register, which suffers from informative under-registration of the dwelling identity number (DIN). A summary is given in Section 6.

## 2. Double mixed modeling

### 2.1 Random effects mixed model in the complete-data case

#### 2.1.1 Models for finite population

The small area counts can be arranged in a two-way contingency table, denoted by $\mathbf{X} = \{X_{ak}\}$, where $a = 1, \ldots, A$ indexes the small areas and $k = 1, \ldots, K$ the categories of interest. The interest of estimation is the within-area proportions given by

$$\theta_{ak}^X = X_{ak} / X_{a.} = X_{ak} \Big/ \sum_{j=1}^{K} X_{aj}$$

referred to as compositions since $\sum_k \theta_{ak}^X = 1$. Typically under the GSPREE approach we assume that the marginal totals $\{X_{a.}\}$ and $\{X_{.k}\}$, also known as the allocation structure, are either known or can be reliably estimated, in which case estimating $\{\theta_{ak}^X\}$ is equivalent to estimating $\{X_{ak}\}$. For simplicity we then make no distinction between counts and compositions in the exposition. Otherwise, without the allocation structure, one can still use our approach to estimate $\{\theta_{ak}^X\}$ but not $\{X_{ak}\}$.

Assume that we have available an auxiliary table of the same dimension, denoted by $\mathbf{X}^0 = \{X_{ak}^0\}$, and the corresponding within-area proportions $\{\theta_{ak}^0\}$. To model $\theta_a^X = (\theta_{a1}^X, \ldots, \theta_{aK}^X)^T$ we use the *multinomial standardized-log (mslog)* link function, given by

$$\mu_{ak}^X = \log \theta_{ak}^X - K^{-1} \sum_{j=1}^{K} \log \theta_{aj}^X \tag{1}$$

and similarly for $\mu_{ak}^0$ and $\theta_{ak}^0$. Zhang and Chambers (2004) introduced the following generalized linear structural mixed model (GLSMM)

$$\mu_{ak}^X = \lambda_k + \beta \mu_{ak}^0 + v_{ak} \tag{2}$$

where

$$\sum_{k=1}^{K} \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^{K} v_{ak} = 0$$

and $\mathbf{v}_{a(1)} = (v_{a2}, \ldots, v_{aK})^T$ assumes a multivariate normal distribution with covariance matrix $G = G(\delta)$, where $\delta$ contains the variance parameters. Notice that there is no area-specific term in (2) because $\sum_k \mu_{ak} = \sum_k \mu_{ak}^0 = 0$. The term "structural" refers to the fact that this is a model of the finite-population parameters $\{\theta_{ak}^X\}$ directly, although the emphasis is not common in the small area estimation literature. For instance, the well-known Fay-Herriot model (Fay and Herriot 1979) is "structural" in the same sense.

There is an important interpretation of the model (2) in terms of the log-linear interactions of $\{\theta_{ak}\}$ due to the choice of the link function (1), *i.e.*,

$$\mu_{ak}^X = \alpha_k + \alpha_{ak}^X \tag{3}$$

where by the standard theory of log-linear models (*e.g.*, Agresti 2002), we have

$$\log X_{ak} = \log X_{a.} + \log \theta_{ak}^X = \alpha_0^X + \alpha_a^X + \alpha_k^X + \alpha_{ak}^X$$

for $\alpha_0^X = (AK)^{-1} \sum_{a,k} \log X_{ak}$, and $\alpha_a^X = K^{-1} \sum_k \log X_{ak} - \alpha_0^X$, and $\alpha_k^X = A^{-1} \sum_a \log X_{ak} - \alpha_0^X$, and $\alpha_{ak}^X = \log X_{ak} - \alpha_a^X - \alpha_k^X - \alpha_0^X$, such that $\sum_a \alpha_a^X = \sum_k \alpha_k^X = \sum_a \alpha_{ak}^X = \sum_k \alpha_{ak}^X = 0$. We refer to (3) as the log-linear identity, and we refer to the log-linear parameters $\alpha_{ak}^X$ as the (first-order) interactions of the compositions $\theta_{ak}^X$ as well as the counts $X_{ak}$. Similar identity holds for $\mu_{ak}^0$. Zhang and Chambers (2004) showed that the GLSMM is equivalent to the following *proportional interactions mixed model* (*PIMM*)

$$\alpha_{ak}^X = \beta \alpha_{ak}^0 + v_{ak} + O_p(A^{-1/2}). \tag{4}$$

The parameters $\lambda_k$'s in (2) do not entail any model restriction beyond the PIMM, and they do not affect the

interactions. The parameter $\beta$ is called the proportionality coefficient. Clearly, SPREE based directly on the association structure $\{X^0_{ak}\}$ amounts to setting $\beta \equiv 1$ and $v_{ak} \equiv 0$. We therefore refer to the model (2) as a GSPREE model, which contain both fixed and random effects extensions of the SPREE model.

### 2.1.2 Model for sample

To complete the model specification we assume sample classifications $\mathbf{x} = \{x_{ak}\}$. Let

$$\mathbf{t}_a = (t_{a1}, \ldots, t_{aK})^T = (t_1(\mathbf{x}_a), \ldots, t_K(\mathbf{x}_a))^T$$

be such that $E(t_{ak} | \mathbf{v}) = E(t_{ak} | \mathbf{X}) = \theta^X_{ak}$, where $\mathbf{v} = \{v_{ak}\}$. The expectation is typically with respect to the sampling design. However, it can also be taken under a suitable model of the sampling distribution, such as a multinomial model for $\mathbf{x}_a$ provided simple random sampling within each area. We therefore make no distinction in the notation.

We assume that $\mathbf{t}_a$ is independent of $\mathbf{t}_{a'}$ for $a \neq a'$, and put

$$V(t_{ak}) = \nu_1 \omega_k(\mathbf{X}_a) \quad \text{and} \quad \text{Cov}(t_{ak}, t_{aj}) = \nu_1 \omega_{kj}(\mathbf{X}_a) \quad (5)$$

where $\omega_k(\cdot)$ and $\omega_{kj}(\cdot)$ are specified variance and covariance functions, and $\nu_1$ is the dispersion parameter that may or may not be known. This is essentially the quasi-likelihood set-up for dependent data (McCullagh and Nelder 1989). The dependence on $\mathbf{X}_a$ allows us to incorporate the sampling design effect, in which case the expectations in (5) may be evaluated with respect to the sampling distribution. This is an important reason why we do not directly assume that the distribution of $\mathbf{t}_a$ belongs to the exponential family, as *e.g.*, in the generalized linear mixed models (Breslow and Clayton 1993).

### 2.1.3 Parameter estimation

Zhang and Chambers (2004) outline an iterative weighted least square (IWLS) algorithm for the GLSMM (2), which is a variation of the PQL approach (Schall 1991; Breslow and Clayton 1993). Let $\mu_a = (\mu^X_{a1}, \ldots, \mu^X_{aK})^T$. The GLSMM (2) can formally be given by

$$\mu_a = g(\theta_a) = H_a \zeta + B \mathbf{v}_{a(1)}$$

where $g(\theta_a)$ is the mslog link function, and $\zeta = (\lambda_2, \ldots, \lambda_K, \beta)^T$, and $\mathbf{v}_{a(1)} = (v_{a2}, \ldots, v_{aK})^T$. The $K \times K$ design matrix $H_a$ and $K \times (K-1)$ design matrix $B$ are, respectively,

$$H_a = [B_{K \times K-1} \quad \mu^0_a] \quad \text{and} \quad B = \begin{pmatrix} -\mathbf{1}^T_{K-1} \\ I_{K-1 \times K-1} \end{pmatrix}$$

where $\mathbf{1}$ is a vector of 1 and $I$ is an identity matrix. Define the working variables

$$\mathbf{z}_a \overset{\text{def.}}{=} \mu_a + \mathbf{e}_a = H_a \zeta + B \mathbf{v}_a + \mathbf{e}_a \text{ and } \mathbf{e}_a = Q(\mathbf{t}_a - \theta^X_a) \quad (6)$$

where $Q = \partial \mu^X_a / \partial \theta^X_a$ is the Jacobian matrix of partial derivatives. Denote by $R_a$ the conditional covariance matrix of $\mathbf{t}_a$ given $\theta^X_a$ defined by (5). Under the PQL approach we assume that $\mathbf{e}_a$ has an approximate multivariate normal distribution with covariance matrix $QR_aQ^T$, and apply standard methods for linear mixed models (LMM) to the linearized data (6). Variants of the PQL approach differ in the estimation of the variance parameters $\delta$. The details are omitted here.

### 2.1.4 On model hierarchy

The GLSMM (2) is specified at the finite population level. More generally, we may consider the finite population $\{X_{ak}\}$ to be randomly generated from an infinite super-population. Let $\theta_{ak}$ be the within-area probability that a unit of the super-population belongs to the cell $(a, k)$, where $\sum_k \theta_{ak} = 1$. Conditional on $X_{a.} = \sum_k X_{ak}$, the within-area counts $(X_{a1}, \ldots, X_{aK})^T$ follow the multinomial distribution with parameters $(\theta_{a1}, \ldots, \theta_{aK})^T$. A *multinomial standardized-log mixed model* (*MSLMM*) of $\{\theta_{ak}\}$ is given by

$$\mu_{ak} = \lambda_k + \beta \mu^0_{ak} + v_{ak}$$

where (7)

$$\sum_{k=1}^K \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^K v_{ak} = 0$$

where $\mu_{ak}$ is given by $\theta_a$ through the mslog link function.

Unlike the GLSMM (2), the equation (7) defines a regression model. There are then three different hierarchy one may choose from in the sample survey situation:

1. Assume the GLSMM (2) for the finite population and the quasi-likelihood model (5) for the sample, yielding the GSPREE approach of Zhang and Chambers (2004).

2. Assume the MSLMM (7) for the super-population and model sample data $\mathbf{t}_a$ based on $\theta_a$ directly, yielding a purely model-based two-level approach.

3. Assume the MSLMM (7) for the super-population, and assume that the finite population totals $\mathbf{X}_a$ follow the multinomial distribution given $\theta_a$, and assume the quasi-likelihood model (5) given $\mathbf{X}_a$, yielding a general three-level model.

Provided the finite population is large, it makes little difference in practice to adopt the GSPREE approach, in

which case one does not have to deal explicitly with one extra level of hierarchy. But the distinction between (2) and (7) becomes necessary if the areas are so small that the stochastic variation in $\mathbf{X}_a$ is not negligible compared to the sampling variation in $\mathbf{x}_a$ (or $\mathbf{t}_a$). In our application later, we have register data that would have given us the interested population counts $\{X_{ak}\}$ had they not suffered from missing data. And the small area level of aggregation is so detailed that the stochastic variation in $\mathbf{X}_a$ can not be ignored. We therefore adapt the GSPREE approach by (a) adopting the MSLMM (7) instead of the GLSMM (2), and (b) modeling $\mathbf{X}_a$ as a 'sample', albeit a very large one, from the super-population directly.

### 2.2 A random effects mixed model of missing data

Missing data add another level of stochastic variation on top of the underlying complete data. In the exposition below, we consider the sample counts $\{x_{ak}\}$ as the complete data, which is the most common situation in practice. Our application later in Section 5 can be viewed as a special case where $\mathbf{X} = \mathbf{x}$.

Denote by $\mathbf{y}_a = (y_{a1}, \ldots, y_{aK})^T$ the observed cell counts, for $a = 1, \ldots, A$. Suppose that, conditional on $x_{ak}$ and a random effect $b_a$,

$$E(y_{ak} \mid x_{ak}, b_a) = x_{ak} p_{ak}$$

and (8)

$$V(y_{ak} \mid x_{ak}, b_a) = \nu_2 c_{ak} p_{ak}(1 - p_{ak})$$

where $c_{ak}$ is a known constant, and $\nu_2$ is the dispersion parameter. We assume that $y_{ak}$ is independent of $y_{aj}$ for $k \neq j$, *i.e.*, missing data are independent from one cell to another. Let the units in the complete sample cell $(a, k)$ be indexed by $i = 1, \ldots, n_{ak}$. Let $r_{i,ak} = 1$ if the $i^{\text{th}}$ unit is observed, and $r_{i,ak} = 0$ if it is missing. The parameter $p_{ak}$ is the assumed probability of $r_{i,ak} = 1$ inside cell $(a, k)$. To see this, let $x_{i,ak}$ be the contribution of the $i^{\text{th}}$ unit to $x_{ak}$, *i.e.*, $x_{ak} = \sum_{i=1}^{n_{ak}} x_{i,ak}$, such that $y_{ak} = \sum_{i=1}^{n_{ak}} r_{i,ak} x_{i,ak}$ and

$$E(y_{ak} \mid x_{1,ak}, \ldots, x_{n_{ak},ak}, b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} E(r_{i,ak} \mid b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} P(r_{i,ak} = 1 \mid b_a) = x_{ak} p_{ak}.$$

Notice that $p_{ak}$ does not depend on the value of $x_{i,ak}$, but only the position of the unit in the two-way table. We assume that $p_{ak}$ depends on $b_a$ through the logistic link function given by

$$\eta_{ak} = \log(p_{ak}/(1 - p_{ak})) = \xi_k + b_a$$

where (9)

$$b_a \sim N(0, \sigma^2).$$

The fixed effects $\xi_k$'s allow the probability of missing to depend on the categories of interest, the area-level random effect $b_a$ allows it to vary across the areas in addition.

Obviously, under the assumptions (8) and (9), the missing data cause bias in the estimates of the $\lambda_k$'s, if the observed table $\mathbf{y}$ is treated as if it were complete. Moreover, it distorts the estimation of the first-order interactions $\{\alpha_{ak}^X\}$. We have,

$$\log p_{ak} = (\xi_k + b_a) - \gamma_{ak} \text{ where } \gamma_{ak} = \log(1 + \exp(\xi_k + b_a)).$$

The first-order interactions of $\{p_{ak}\}$ are then given by $\alpha_{ak}^p = -\tilde{\gamma}_{ak} = -(\gamma_{ak} - \overline{\gamma}_{a.} - \overline{\gamma}_{.k} + \overline{\gamma}_{..})$, for the row and column means $\overline{\gamma}_{a.}$ and $\overline{\gamma}_{.k}$ and the overall mean $\overline{\gamma}_{..}$. These are non-zero unless $\xi_k = \xi$. By (8) the interactions of the expected observed table are given by

$$\alpha_{ak}^{E(\mathbf{y}|\mathbf{x},\mathbf{b})} = \alpha_{ak}^x + \alpha_{ak}^p = \alpha_{ak}^x - \tilde{\gamma}_{ak} \neq \alpha_{ak}^x$$

such that the estimates of $\{\alpha_{ak}^X\}$ will be biased if $\mathbf{y}$ is treated as $\mathbf{x}$.

It is worth noting that, as far as the estimation of the interactions is concerned, it is in principle possible to treat the observed table $\mathbf{y}$ as if it were the complete table $\mathbf{x}$ under a particular missing-data model given by

$$\log p_{ak} = \xi'_k + b'_a. \tag{10}$$

This is because the first-order interactions of $\{p_{ak}\}$ are all zero under (10), in which case we have $\alpha_{ak}^{E(\mathbf{y}|\mathbf{x})} = \alpha_{ak}^{\mathbf{x}}$. Disregarding the range restrictions, the assumption (10) defines an informative missing-data mechanism where the probability of missing varies across the categories of interest, while the area effect modifies all the within-area probabilities of missing by a factor $\exp(b'_a)$, such that $p_{ak}/\sum_{j=1}^K p_{aj} = \exp(\xi'_k)/\sum_j \exp(\xi'_j)$ remains constant. The model (9), however, is more flexible since it allows the random effects to affect the interactions. Both (9) and (10) will be examined in Section 5.

Finally, we notice that allowing for component-wise random effects in the model (9) may cause identification problems. For instance, assume simple random sampling from the finite population, in which case the interactions of the expected complete table are given by $\alpha_{ak}^{E(\mathbf{x}|\mathbf{X})} = \alpha_{ak}^X$. With component-wise $b_{ak}$ in the model (9) we have $\log p_{ak} = \xi_k + b_{ak} + \gamma_{ak}$, where $\gamma_{ak} = \log(1 + \exp(\xi_k + b_{ak}))$. It follows from (4) and (8) that the interactions of the expected table $E(\mathbf{y} \mid \mathbf{x}, \mathbf{b})$ is given by $\beta\alpha_{ak}^0 + v_{ak} + b_{ak} - \tilde{\gamma}_{ak}$. But there is no information in the observed data to distinguish between the two random effects $v_{ak}$ and $b_{ak}$.

## 3. Conditional mean squared errors of prediction

We adopt the approach of Booth and Hobert (1998) and use the CMSEP as a measure of the uncertainty in prediction. Like them we consider the CMSEP on the linear-predictor scale. In vector form the $\mu_{ak}$'s in (1) belong to the following class of linear functions

$$\mu_a = H_a\zeta + B_a\mathbf{v}_a \tag{11}$$

where $\mu_a$ is the area-specific vector of linear predictors, and $\zeta$ is the vector of fixed effects, and $\mathbf{v}_a$ is the vector of area-specific random effects, and $H_a$ and $B_a$ are the corresponding design matrices. All the quantities have been specified in (6) for the GLSMM (2), where we actually have $B_a = B$. But we shall adopt the slightly more general formulation (11) in the following. Let $\hat{\zeta}$ and $\hat{\mathbf{v}}_a$ be, respectively, the estimates of $\zeta$ and $\mathbf{v}_a$ based on observations subjected to missing data, denoted by $\mathbf{y}_a$ for $a = 1, \ldots, A$. The CMSEP of $\hat{\mu}_a = H_a\hat{\zeta} + B_a\hat{\mathbf{v}}_a$ is defined as

$$\text{CMSEP}_a = E\{(\hat{\mu}_a - \mu_a)(\hat{\mu}_a - \mu_a)^T \mid \mathbf{y}_a\}.$$

We introduce first a decomposition through the hypothetical best predictor (BP) based on $\mathbf{x}_a$, given by $\dot{\mu}_a = E(\mu_a \mid \mathbf{x}_a, \zeta, \delta) = H_a\zeta + B_aE(\mathbf{v}_a \mid \mathbf{x}_a, \zeta, \delta)$, when the parameters are known. We have

$$\begin{aligned}\text{CMSEP}_a &= E\{E((\dot{\mu}_a - \mu_a)(\dot{\mu}_a - \mu_a)^T \mid \mathbf{x}_a) \mid \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\} \\ &= E\{B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \mid \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\}\end{aligned}$$

because $\dot{\mu}_a - \mu_a$ and $\hat{\mu}_a - \dot{\mu}_a$ are conditionally independent of each other given $\mathbf{x}_a$: $\dot{\mu} - \mu$ depends on the random effects $\mathbf{v}_a$, whereas $\hat{\mu}_a - \dot{\mu}_a$ depends on random variations in the other areas. Next, for the second term on the right-hand side, we introduce a decomposition through the hypothetical estimated best predictor (EBP) based on the complete data $\mathbf{x}$, denoted by $\tilde{\mu}_a = H_a\tilde{\zeta} + B_a\tilde{v}_a$, where $(\tilde{\zeta}, \tilde{\delta})$ are the parameter estimates based on $\mathbf{x}$, and $\tilde{v}_a = E(\mathbf{v}_a \mid \mathbf{x}_a, \tilde{\zeta}, \tilde{\delta})$. We have

$$\begin{aligned}E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\} &\approx E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T\} \\ &= E\{E((\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{x})\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x})\} \\ &= E\{(\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x})\}.\end{aligned}$$

The first approximation is correct to the order of $O_p(A^{-1})$, and can be justified as the number of areas tends to infinity. Intuitively, this makes sense if the information from any single area is asymptotically negligible compared to the information from all the other areas together. Next, the decomposition follows because $\tilde{\mu}_a - \dot{\mu}_a$ and $\hat{\mu}_a - \tilde{\mu}_a$ are independent of each other given $\mathbf{x}$: the former is a constant given $\mathbf{x}$.

In this way, we obtain an approximate CMSEP with a three-part decomposition

$$\text{CMSEP}_a \approx h_{1a}(\mathbf{x}_a; \zeta, \delta) + h_{2a}(\zeta, \delta) + h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi)$$

where $\psi$ contains the parameters of the conditional distribution of $\mathbf{y}_a$ given $\mathbf{x}_a$, and

$$h_{1a}(\mathbf{x}_a; \zeta, \delta) \overset{\text{def.}}{=} B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \tag{12}$$

$$h_{2a}(\zeta, \delta) \overset{\text{def.}}{=} E\{(\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T\} \tag{13}$$

$$h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi) \overset{\text{def.}}{=} E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x}\}. \tag{14}$$

The three *h*-terms correspond, respectively, to a conditional prediction variance due to the random effects, a positive correction that accounts for the uncertainty in the estimation of the parameters based on the latent complete data, *i.e.*, the sampling variation, and another positive correction for the extra variation due to the randomness in the missing data. Alternative approximations are possible. For instance, one might use $E\{B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \mid \mathbf{y}_a\}$ instead of $h_{1a}$, or replace $h_{3a}$ with the unconditional $E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T\}$. The expressions (12) - (14) are chosen because they produce a clean separation between the sampling variation in the complete data and the extra variation owing to the missingness given the complete data. The difference from the CMSEP in the complete-data case (Booth and Hobert 1998) comes down to the third term $h_{3a}$.

## 4. Estimation

### 4.1 Parameter estimation

The structure of the data suggests an iterative procedure similar to the EM algorithm (Dempster, Laird and Rubin 1977). Given the current values of the parameters and the random effects, we calculate at the E-step the conditional expected complete two-way table $E(\mathbf{x} \mid \mathbf{y}, \mathbf{m})$. At the M-step we estimate the two random effects mixed models separately by some maximum penalized quasi-likelihood (MPQL) procedures. Iterations between the two yield an EMPQL algorithm.

For the E-step, let $I_{i,ak}=1$ if the sample unit *i* belongs to the $(a,k)^{\text{th}}$ cell, and $I_{i,ak}=0$ otherwise. The value is observed provided $r_{i,ak}=1$, but is unknown if $r_{i,ak}=0$. Let $\theta_{ak}$ be the generic compositions, depending of the adopted model. Suppose that

$$P[I_{i,ak}=1 \mid i\in s]=d_{ak}\theta_{ak}$$

where *s* denotes the complete sample, and $d_{ak}$ is some known constant which accounts for the eventual sampling design effect. For example, simple random sampling implies that $d_{ak}=1$ for all $(a,k)$. An example of $d_{ak}\neq 1$ is when the sampling units are households, which are selected by a probability proportional to the household size. Let $m_{ak}=x_{ak}-y_{ak}=\sum_{i;\,r_{i,ak}=0} I_{i,ak}x_{i,ak}$. We have $E(x_{ak}\mid \mathbf{y}_a, m_{a.})=y_{ak}+E(m_{ak}\mid m_{a.})$, where

$$\begin{aligned}E(m_{ak}\mid m_{a.}) &= \sum_{i;r_{i,ak}=0} E(I_{i,ak}\mid r_{i,ak}=0)x_{i,ak}\\ &= m_{a.}P[I_{i,ak}=1\mid r_{i,ak}=0]\\ &= m_{a.}(1-p_{ak})d_{ak}\theta_{ak}\Big/\Big\{\sum_j (1-p_{aj})d_{aj}\theta_{aj}\Big\}. \quad (15)\end{aligned}$$

Having thus 'completed' the sample data, we move to the MPQL-step, where we apply the IWLS algorithm outlined in Section 2.1.3, respectively, to the complete-data model and the missing-data model conditional on the complete data.

### 4.2 Estimation of CMSEP

Evaluating the CMSEP at the estimated parameter values yields a plug-in estimate of the CMSEP. Of the three *h*-terms, $h_{1a}$ is of the order $O_p(1)$, whereas both $h_{2a}$ and $h_{3a}$ are of the order $O_p(A^{-1})$, when the number of areas tends to infinity while the within-area sample sizes remain bounded. The results of Booth and Hobert (1998) and Prasad and Rao (1990), obtained in the univariate complete-data case, suggest that the bias in the plug-in estimate $\hat{h}_{1a}$ is of the same order as $\hat{h}_{2a}$ and $\hat{h}_{3a}$. These authors developed second-order correction through the Taylor expansion. We do not pursue such second-order asymptotics in this paper. Approximate expressions of the *h*-terms that accompany the EMPQL algorithm are given below.

Take first $h_{1a}$ by (12). Based on the linearized data (6), the covariance matrix $\mathrm{Cov}(\mathbf{v}_a,\mathbf{v}_a\mid \mathbf{z}_a)$ does not depend on either $\mathbf{z}_a$ or $\mathbf{x}_a$. This is convenient because we then have

$$\begin{aligned}h_{1a}(\mathbf{x}_a;\zeta,\delta) &\approx B_a\mathrm{Cov}(\mathbf{v}_a,\mathbf{v}_a\mid \mathbf{z}_a)B_a^T\\ &= B_a(G-GB_a^TV_a^{-1}B_aG)B_a^T \quad (16)\end{aligned}$$

where $V_a=B_aGB_a^T+QR_aQ^T$ is the marginal covariance matrix of $\mathbf{z}_a$.

Next, take $h_{2a}$ by (13). Let $\phi=(\zeta^T,\delta^T)^T$. Expanding $\tilde{\phi}$ around $\phi$ yields $\tilde{\mu}_a-\dot{\mu}_a\approx \dot{\mu}_a'(\tilde{\phi}-\phi)$, where $\dot{\mu}_a'=\partial\dot{\mu}_a/\partial\phi$, such that

$$h_{2a}\approx \dot{\mu}_a'\,\mathrm{Cov}(\tilde{\phi},\tilde{\phi})\dot{\mu}_a'^T. \quad (17)$$

Based on (6) we derive $\dot{\mu}_a=H_a\zeta+D_a\dot{\mathbf{u}}_a$, where $D_a=B_aGB_a^TV_a^{-1}$ and $\dot{\mathbf{u}}_a=\mathbf{z}_a-H_a\zeta$. Denote by *I* the identity matrix. The partial derivatives in $\dot{\mu}'_a$ are given by

$$\partial\dot{\mu}_a/\partial\zeta=(I-D_a)H_a$$

and

$$\partial\dot{\mu}_a/\partial\delta_j=(\partial D_a/\partial\delta_j)\dot{\mathbf{u}}_a=(I-D_a)B_a(\partial G/\partial\delta_j)B_a^TV_a^{-1}\dot{\mathbf{u}}_a$$

where $\delta_j$ is the $j^{\text{th}}$ variance parameter in the covariance matrix $G(\delta)$ of $\mathbf{v}_a$. To obtain $\mathrm{Cov}(\tilde{\phi},\tilde{\phi})$, suppose that the PQL approach is based on the following quasi log-likelihood

$$\ell=\sum_a \ell_a$$

and

$$\ell_a=-\frac{1}{2}\log|V_a|-\frac{1}{2}(\mathbf{z}_a-H_a\zeta)^TV_a^{-1}(\mathbf{z}_a-H_a\zeta).$$

The so-called sandwich formula yields then

$$\mathrm{Cov}(\tilde{\phi},\tilde{\phi})=\left(-\frac{\partial^2\ell}{\partial\phi^2}\right)^{-1}\left\{\sum_{a=1}^{A}\left(\frac{\partial\ell_a}{\partial\phi}\right)\left(\frac{\partial\ell_a}{\partial\phi}\right)^T\right\}\left(-\frac{\partial^2\ell}{\partial\phi^2}\right)^{-1}.$$

Finally, take $h_{3a}$ by (14). Similarly as above we have $\tilde{\mu}_a=(I-\tilde{D}_a)H_a\tilde{\zeta}+\tilde{D}_a\tilde{\mathbf{z}}_a$ evaluated at $\phi=\tilde{\phi}$, and $\hat{\mu}_a=(I-\hat{D}_a)H_a\hat{\zeta}+\hat{D}_a\hat{\mathbf{z}}_a$, where $\hat{\mathbf{z}}_a$ is derived from $\hat{\mathbf{t}}_a=\mathbf{t}(\hat{\mathbf{x}}_a)$ for $\hat{\mathbf{x}}_a=E(\mathbf{x}_a\mid \mathbf{y}_a,m_{a.};\hat{\phi},\hat{\psi})$. Expanding $\hat{\phi}$ around $\tilde{\phi}$ and retain only the leading term, we obtain

$$\hat{\mu}_a-\tilde{\mu}_a\approx \breve{\mu}_a-\tilde{\mu}_a=\tilde{D}_a(\breve{\mathbf{z}}_a-\tilde{\mathbf{z}}_a)$$

where $\breve{\mu}_a=(I-\tilde{D}_a)H_a\tilde{\zeta}+\tilde{D}_a\breve{\mathbf{z}}_a$, and $\breve{\mathbf{z}}_a$ is derived from $\breve{\mathbf{t}}_a=\mathbf{t}(\breve{\mathbf{x}}_a)$ for $\breve{\mathbf{x}}_a=E(\mathbf{x}_a\mid \mathbf{y}_a,m_{a.};\tilde{\phi},\hat{\psi})$. That is, we ignore the terms involving $\hat{\phi}-\tilde{\phi}$. The remaining variation in $\breve{\mathbf{z}}_a$ is due to the estimation of the missing-data model alone. Expanding $\hat{\psi}$ around $\psi$, we obtain, by the chain rule,

$$h_{3a}\approx C_a\mathrm{Cov}(\hat{\psi},\hat{\psi}\mid \mathbf{x})C_a^T \quad (18)$$

and

$$C_a=\left\{D_a\left(\frac{\partial\mathbf{z}_a}{\partial\mathbf{t}_a}\right)\left(\frac{\partial\mathbf{t}_a}{\partial\mathbf{x}_a}\right)\left(\frac{\partial\mathbf{x}_a}{\partial\mathbf{p}_a}\right)\left(\frac{\partial\mathbf{p}_a}{\partial\eta_a}\right)\left(\frac{\partial\eta_a}{\partial\psi}\right)\right\}_{\phi=\tilde{\phi},\psi}$$

where we assume that $E(\hat{\psi}\mid \mathbf{x})=\psi$ and $E[\breve{\mathbf{z}}_a\mid \mathbf{x}]=\tilde{\mathbf{z}}_a$. Whereas the sandwich formula yields $\mathrm{Cov}(\hat{\psi},\hat{\psi}\mid \mathbf{x})$ under the conditional model of **y** given **x**, similarly to $\mathrm{Cov}(\tilde{\phi},\tilde{\phi})$ above.

### 4.3 Estimation of small area compositions

Suppose first that the GLSMM, defined by (2) and in combination with (5), has been estimated, upon which we obtain $\hat{\mu}_a^X$, and $\hat{\theta}_{ak}^X = \exp(\hat{\mu}_{ak}^X)/\sum_j \exp(\hat{\mu}_{aj}^X)$.

When the marginal totals $X_{a.}$ and $X_{.k}$ are known, it makes sense to apply the IPF, starting with the estimated table $\{\hat{\theta}_{ak}^X\}$. The difference from SPREE, which starts with the auxiliary table $\mathbf{X}^0$, is that the interactions have been re-estimated. On convergence we obtain the estimated small area counts, denoted by $\hat{\mathbf{X}} = \{\hat{X}_{ak}\}$, and the corresponding compositions, denoted by $\theta_{ak}^{\hat{X}} = \hat{X}_{ak}/\sum_j \hat{X}_{aj}$, which are different from the direct model estimates $\hat{\theta}_{ak}^X$ that have provided the starting values for the IPF.

Often in practice, while the area totals $\{X_{a.}\}$ may be known, the marginal totals $\{X_{.k}\}$ need to be estimated based on the survey data available, separately using a method that is appropriate for the aggregated level. The IPF is still worth considering as long as these estimated marginal totals are judged to be more reliable and/or less biased than the aggregated small area estimates $\sum_a X_{a.}\hat{\theta}_{ak}^X$. The reason is that the estimated interactions $\hat{\alpha}_{ak}^X$ are preserved in the IPF, *i.e.*, $\alpha_{ak}^{\hat{X}} = \hat{\alpha}_{ak}^X$. By the log-linear identity (3), the difference between the direct model estimate $\hat{\theta}_{ak}^X$ and final estimate $\theta_{ak}^{\hat{X}}$ is due to the difference in the estimates of the main effects $\{\alpha_k^X\}$. Thus, less biased estimates of $\{X_{.k}\}$ are expected to yield less biased estimates of $\{\alpha_k^X\}$ and, thereby, less biased estimates of $\{\theta_{ak}^X\}$.

Suppose next that the MSLMM (7) combined with (5) have been estimated. We may express the interest of estimation, *i.e.*, $\{\mu_{ak}^X\}$, in terms of $\mathbf{z}_a$ defined as

$$\begin{aligned}\mathbf{z}_a &= H_a\zeta + B_a\mathbf{v}_a + \mathbf{e}_a = H_a\zeta + B_a\mathbf{v}_a + \mathbf{e}_a^X + \mathbf{e}_a^{x|X} \\ &= \mu_a^X + \mathbf{e}_a^{x|X} = H_a\zeta + \mathbf{v}_a^X + \mathbf{e}_a^{x|X}\end{aligned}$$

where $\mathbf{e}_a^X = Q(\theta_a^X - \theta_a)$ and $\mathbf{e}_a^{x|X} = Q(\mathbf{t}_a - \theta_a^X)$. In accordance we have $R_a = R_a^X + R_a^{x|X}$, where $R_a^X = \mathrm{Cov}(\theta_a^X, \theta_a^X \mid \theta_a)$ and $R_a^{x|X} = \mathrm{Cov}(\mathbf{t}_a, \mathbf{t}_a \mid \theta_a^X)$. It follows that

$$\hat{\mu}_a^X = H_a\hat{\zeta} + (B_a\hat{G}B_a^T + \hat{Q}\hat{R}_a^X\hat{Q}^T)\hat{V}_a^{-1}(\hat{\mathbf{z}}_a - H_a\hat{\zeta}). \quad (19)$$

The rest follows as above where $\mu_a^X$ is estimated directly under the GLSMM.

## 5. Example: Register-based small area household compositions

### 5.1 Register household data

Register-based household data have undergone considerable development in Norway. One of the goals is to produce detailed household statistics that traditionally are only available from the census. For this purpose the registration of a unique dwelling identity number (DIN) was initiated in the last census in 2001. The work is not yet completed, and the DIN is still missing for about 6% of the people residing in the country. The rate of missing is differential as it varies over the household type as well as across the Municipalities, the latter of which is a reflection of the overall effort of the local administration regarding the registration of the DINs.

A household register can be complied in a year after the census based on a number of data sources. The most important ones include the central population register (CPR), the DIN-register and the census household file (CH01). Even without the DIN a register household can be compiled based on the other information available. But the result suffers from informative under-registration of the DIN. For instance, a typical source of bias is cohabitants living without children, because such a couple appear as two single-person households in the CPR, unless they have already been identified as a household in the CH01. Nevertheless, historic as well as cross-country comparisons suggest that the national totals are acceptable. A more urgent problem lies on lower levels of aggregation. For example, changes from the census in 2001 are unlikely large in certain Municipalities, including the capital city Oslo where the increase in the proportion of single-person households is almost three times as high as it is in the rest of the country - see top-left plot in Figure 1. And a large part of the problem in Oslo can be explained by a combination of high proportion of cohabitants living without children and low DIN-registration rate (indeed, the lowest in the country).

### 5.2 Set-up of data

We shall illustrate our approach using these register household data. The target population contains all persons living at multiple-dwelling addresses at the beginning of year 2005, who do not belong to households of married people or registered partners; the latter household types are excluded because the DIN is not critical for compiling the households of these people. There is no distinction between the finite population and the sample in this case, *i.e.*, $\mathbf{X} = \mathbf{x}$. The households that have registered DINs are treated as the 'observed' sample $\mathbf{y}$, whereas the households that do not have registered DINs are viewed as the missing. In this way the population consists of 713,387 persons, of which 558,136 persons have registered DINs. The overall rate of missing is about 22%.

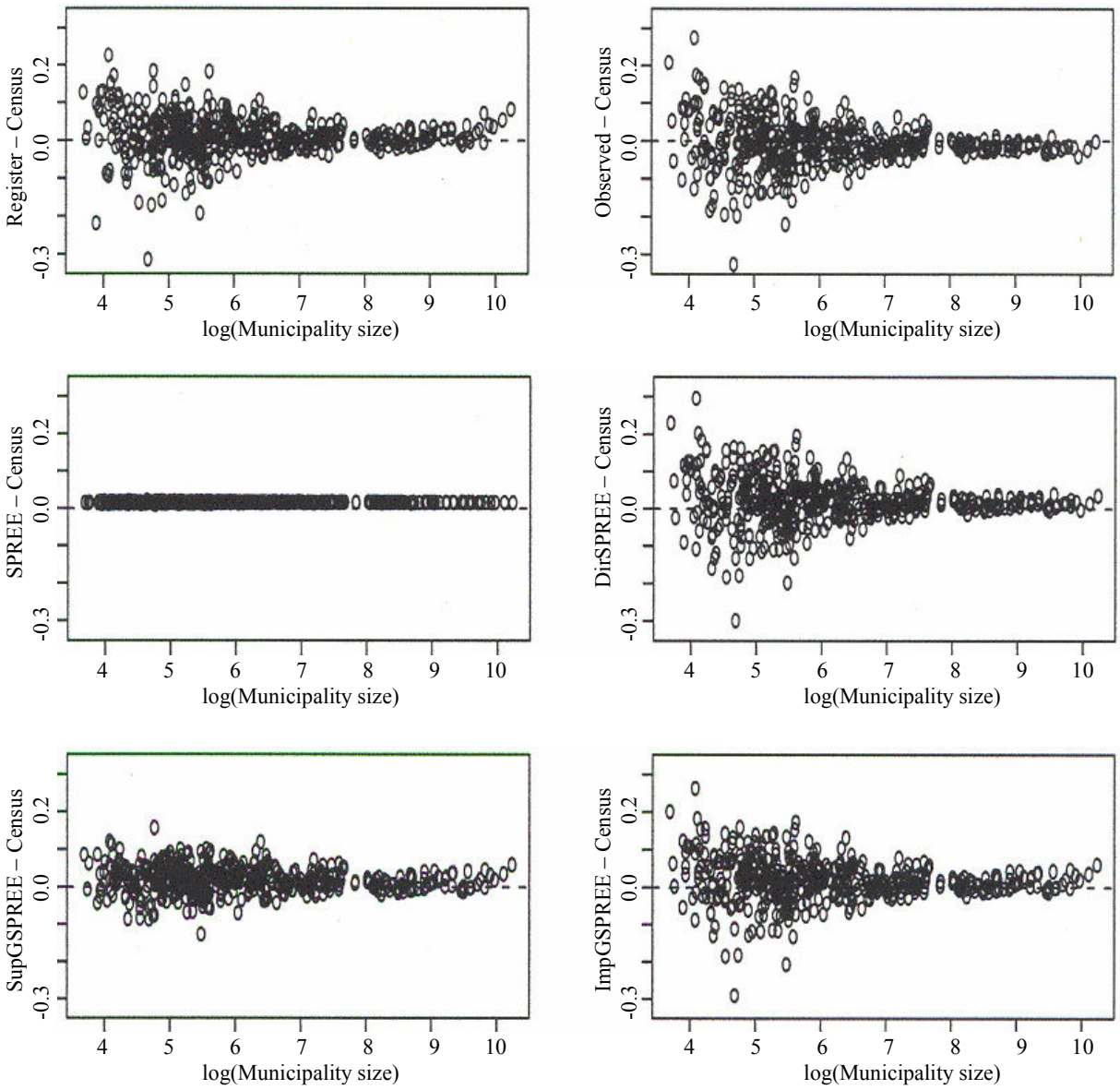


**Figure 1** **Difference between estimates of proportion of Single-person households and census counts in 2001 against log Municipality size: Register households (top-left), Households with registered DINs (top-right), SPREE based on census (middle-left), DirSPREE based on households with registered DINs (middle-right), SupGSPREE of super-population proportions (bottom-left), and ImpGRSREE of imputed finite-population proportions (bottom-right). The dashed line marks no difference**

Let the Municipalities be the small areas of this study, where $A = 433$. The households are classified into 4 categories: $k = 1$ for "Single-person", $k = 2$ for "Single-parent", $k = 3$ for "Cohabitants", and $k = 4$ for "Other", *i.e.*, $K = 4$. Let $i$ index the households, and let $x_i$ be the number of persons living in the household. Let $X_{ak} = x_{ak}$ be the number of persons in the $(a,k)^{\text{th}}$ cell in the population, and let $y_{ak}$ be the corresponding 'observed' cell count. Let $N_{ak}$ be the number of households in the $(a,k)^{\text{th}}$ cell, and let $n_{ak}$ be the corresponding number of 'observed' households. Notice that only the total number of persons is known in each area, but not the total number of households. However, provided cell-specific probability of DIN-registrations, an estimator of $N_{ak}$ based on $\hat{X}_{ak}$ is given by $\hat{N}_{ak} = n_{ak}\hat{X}_{ak}/y_{ak}$. We shall therefore concentrate on the estimation of $X_{ak}$ here.

Let $\{X^0_{ak}\}$ be the corresponding cell counts from the last census in 2001. Let $X'_{ak} = y_{ak} + m'_{ak}$ be the register counts in 2005, where $m'_{ak}$ is the number of persons without the DIN. A register household can be considered as a form of imputed household that may suffer from informative missing of DINs. The register area total is correct, *i.e.*,

$X'_{a.} = X_{a.}$, and the national totals $\{X'_{.k}\}$ are considered acceptable. The question is whether estimates of $\{X_{ak}\}$ can be derived, based on the 'observed' **y** and the allocation structure $\{X_{a.}\}$ and $\{X'_{.k}\}$, that better accounts for the differential missing DINs.

### 5.3 Set-up of model

Scatter plots of the register first-order interactions $\{\alpha_{ak}^{X'}\}$ against the census interactions $\{\alpha_{ak}^{0}\}$ provide motivation for the PIMM (4). To chose between the GLSMM (2) and the MSLMM (7), we look at the difference between the register proportion $\theta_{ak}^{X'}$ and the corresponding census proportion $\theta_{ak}^{0}$, *i.e.*, $\theta_{ak}^{X'} - \theta_{ak}^{0}$, plotted against $\log X_{a.}$: the case of $k = 1$ is shown in the top-left plot of Figure 1. Clearly, the variance of the difference increases as $X_{a.}$ decreases, and is not constant of $X_{a.}$. Notice that we are dealing with estimation at a very low level of aggregation here, where *e.g.*, the median value of all $\{X'_{ak}\}$ is only 70. We therefore adopt the model (7) for $\theta_{ak}$, the quasi-likelihood (5) for $X_{ak} = x_{ak}$, and the quasi-likelihood (8) and the model (9) for $y_{ak}$.

For the quasi-likelihood (5) we assume $\nu_1 = 1$. Let $t_{ak} = X_{ak}/X_{a.}$. We have

$$V(t_{ak}) = N_{a.}^{-1}\theta_{ak}(1-\theta_{ak})\bar{X}_a^{(2)}/\bar{X}_{a.}^2$$

and

$$\mathrm{Cov}(t_{ak}, t_{aj}) = -N_{a.}^{-1}\theta_{ak}\theta_{aj}\bar{X}_a^{(2)}/\bar{X}_{a.}^2.$$

where $\bar{X}_a^{(2)} = \sum_{i=1}^{N_{a.}} x_i^2/N_{a.}$ and $\bar{X}_{a.} = X_{a.}/N_{a.}$. Since $x_i \geq 1$, we have $\bar{X}_a^{(2)} \geq \bar{X}_{a.}^2$, and over-dispersion compared to the Multinomial-$(N_{a.}, \theta_a)$ distribution. We calculate the factor $\bar{X}_a'^{(2)}/\bar{X}_{a.}'^2$ based on the register data, which is then used as $\bar{X}_a^{(2)}/\bar{X}_{a.}^2$ in the estimation below. Moreover, for the quasi-likelihood (8) we assume $\nu_2 = 1$, and

$$V(y_{ak}|n_{ak}) = V\left(\sum_{i=1}^{n_{ak}} r_{i,ak}x_{i,ak}\right) = \left(\sum_i x_{i,ak}^2\right)V(r_{i,ak}) \Rightarrow c_{ak} = \left(\sum_i x_{i,ak}^2\right).$$

### 5.4 Estimation results

Six different estimators of the proportion of Single-person households (*i.e.*, for $k = 1$) are illustrated in Figure 1.

To start with, we have the direct register proportions $\theta_{a1}^{X'}$ in the top-left plot, and the 'observed' proportions $\theta_{a1}^{y}$ in the top-right plot. On average the proportion is increased based on the entire register compared to the census in 2001, whereas it is slightly decreased according to the 'observed' part only. This demonstrates that the missing DINs are informative, as explained before. Inclusion of the register households without the DINs raises the proportion of Single-person households. But the result is implausible in some of the largest Municipalities. Of course, large bias also exists among the smaller Municipalities, but these are not easily detectable in a plot like this one.

Next, in the middle-left plot of Figure 1, estimates are obtained by SPREE using the census counts $\{X_{ak}^0\}$ as the starting values. For the simple two-way table here, this yields an almost constant adjustment of the census proportions, with negligible change in the between-area variation. In the middle-right plot, estimates are obtained by SPREE using the 'observed' table $\{y_{ak}\}$ as the starting values. Notice that, to start with the observed sample counts would be too unstable to be useful in usual survey sampling situations, but it is a viable option here because of the large amount of 'observed' data. To distinguish from the standard SPREE we shall refer to it as the *direct* SPREE (DirSPREE). As noted earlier, DirSPREE is unbiased under the assumption (10) of informative missingness. Indeed, it is seen to lead to useful adjustments for the largest Municipalities.

In the bottom-row plots of Figure 1, estimates are obtained using the double-mixed modeling approach. The estimates of the bottom-left plot are obtained by the IPF starting with the estimated super-population compositions $\{\hat{\theta}_{ak}\}$, denoted by *SupGSPREE*. The extreme post-censal development in the largest Municipalities are reduced. But the changes from the census-proportions are clearly over-shrunk towards to the population average for the smaller areas. The variation is *e.g.*, much less than that of $\theta'_{ak} - \theta^0_{ak}$ in the top-left plot. The estimates of the bottom-right plot are derived from the imputed finite-population counts, denoted by *ImpGSPREE*, which are calculated at the E-step of the EMPQL algorithm. The estimates for the largest Municipalities are similar to those of SupGSPREE, and the variation in the changes from the census-proportions is similar to that of DirSPREE.

### 5.5 Estimation of CMSEP

Approximate CMSEP of the ImpGSPREE compositions can be derived similarly as in Section 3. Denote by $\hat{X}_{ak}$ the ImpGSPREE count, and by $\dot{X}_{ak}$ the BP based on known conditional distribution of $\mathbf{X}_a$ given $(\mathbf{y}_a, m_{a.})$. We have

$$\mathrm{CMSEP}(\hat{\mathbf{X}}_a) \approx E\{(\dot{\mathbf{X}}_a - \mathbf{X}_a)(\dot{\mathbf{X}}_a - \mathbf{X}_a)^T \mid \mathbf{y}_a, m_{a.}\} + E\{(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\}.$$

Moreover, let $\tilde{\phi}$ be the hypothetical estimate of $\phi$ based on the complete data $\mathbf{x} = \mathbf{X}$, and let $\hat{\psi}$ be the estimate of $\psi$ based on the observed data. Let $Q_1$ and $Q_2$ be, respectively, the Jacobian matrix of partial derivatives $\partial\dot{\mathbf{X}}_a/\partial\phi$ and $\partial\dot{\mathbf{X}}_a/\partial\psi$. We have

$$\begin{aligned} E\{(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\} &\approx E\{(\tilde{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\tilde{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\} \\ &\quad + E\{(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)^T \mid \mathbf{X}\} \\ &\approx Q_1\mathrm{Cov}(\tilde{\phi}, \tilde{\phi})Q_1^T + Q_2\mathrm{Cov}(\hat{\psi}, \hat{\psi} \mid \mathbf{X})Q_2^T. \end{aligned}$$

Together, these lead to a three-part decomposition of the CMSEP similar to (12) - (14). In the estimation of the CMSEP below we ignore the effect of IPF. This is justified in our case because the IPF essentially amounts to a constant multiplicative adjustment very close to unity, as can be seen in the middle-left plot in Figure 1.

The CMSEP of a DirSPREE count is calculated as a 'sampling' variance that is induced by missing-at-random within each cell of the two-way table, plus a squared bias term which is estimated by the squared difference between the ImpGSPREE count and the corresponding DirSPREE count, provided the assumption (9) is a more appropriate model for the missing data than the assumption (10).

The estimated root CMSEPs (rcmsep) are given in Figure 2. On average both are decreasing as the Municipality size increases. However, for some of the largest Municipalities, the CMSEP of the DirSPREE proportion is abnormally large for Single-person and Cohabitants households due to the bias term. On the whole the CMSEP of the ImpGSPREE composition is clearly smaller than that of the DirSPREE. The $h_{1a}$ -term, corresponding to the prediction variance of $\mathbf{X}_a$, is by far the dominating contribution to the CMSEP (over 99% in many areas). This is understandable since there are over 550 thousand people in the 'observed' sample, such that the uncertainty in parameter estimation is comparatively negligible. But the quoted percentage will be lower in a sample survey situation, as the estimation uncertainty summarized in terms $h_{2a}$ and $h_{3a}$ increases.

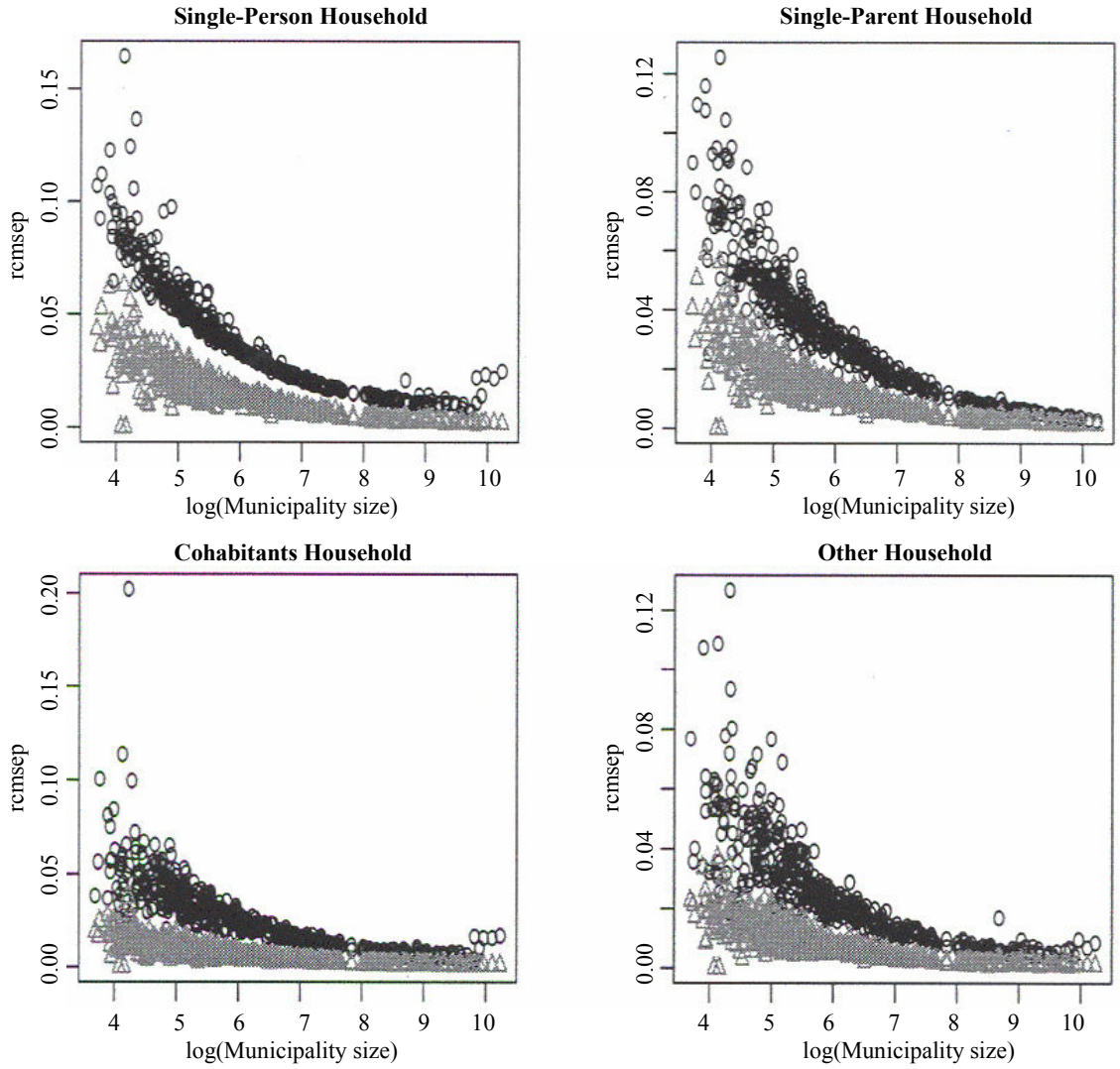


**Figure 2 Estimated root conditional mean squared error of prediction (rcmsep) of DirSPREE (circle) and ImpGSPREE (triangle) of Municipality household proportions**

## 6. Summary

In the above we outlined a double-mixed modeling approach that extends the GSPREE methodology to estimation of small area compositions subjected to differential missing data. An approximate CMSEP was derived which contains a three-part decomposition, corresponding to the prediction variance of the unknown random effect, the sampling variance in the absence of missing data, and the extra variance due to the missing data, respectively. The approach was applied to the Norwegian register household data, which yielded useful adjustments for informative missing of dwelling identity numbers.

## Acknowledgements

I am thankful to the referees and the Associate Editor for comments and suggestions that have helped to improve the presentation.

## References

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Booth, J.G., and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

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*, 93-273-282.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Longford, N. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society*, Series A, 162, 227-245.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. London: Champman and Hall.

Prasad, N.G., and Rao, J.N.K. (1990). The estimation of mean square errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.

Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society*, Series B, 66, 479-496.