Discussion Papers No. 285, August 2000
Statistics Norway, Department of Coordination and Development

Li-Chun Zhang

On dispersion preserving estimation of the mean of a binary variable from small areas

Abstract:
Over-shrinkage is a common problem in small area (or domain) estimation. It happens when the estimated small-area parameters have less between-area variation than their true values. To deal with this problem, Louis (1984), Ghosh (1992) and Spjøtvoll and Thomsen (1987) have proposed various constrained empirical and hierarchical Bayes methods. In this paper we study two non-Bayesian methods based on, respectively, the synthetic estimator and a variance-component model. We show first that the synthetic estimator entails loss of dispersion in general, from which it follows that the coverage level of the confidence intervals could be far below the nominal level of confidence, when these are derived from the sampling error alone. A bivariate variance-component model at the area-level, as well as its simplification, can greatly improve the efficiency of the confidence intervals. However, super-population approaches as such are unable to capture the distribution of the true area-parameters. We develop a finite-population approach based on an empirical finite-population distribution function of the area-parameters, which provides the necessary adjustment. The various methods will be illustrated using the data of the Census 1990. Finally, we notice that several European countries will base the upcoming Census on their administrative register systems, instead of collecting the information in the field. Improved small area estimation methods may prove to be valuable for assessing the quality of such Register Counting.

Keywords: Over-shrinkage, synthetic estimator, variance-component model

Acknowledgement: I wish to thank Johan Heldal and Rolf Aaberge for helpful comments.

Address: Li-Chun Zhang, Statistics Norway, Department of Coordination and Development.
E-mail: li.chun.zhang@ssb.no

Discussion Papers comprise research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborate than a standard journal article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
N-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

Over-shrinkage is a common problem in small area (or domain) estimation. It happens when the estimated small-area parameters have less between-area variation than their true values, which makes the small areas look more like each other than they actually are. In Louis (1984), Ghosh (1992) and Spjøtvoll and Thomsen (1987) various constrained empirical and hierarchical Bayes methods have been developed. Judkins and Liu (2000) compared these methods in details. Over-shrinkage occurs also with many non-Bayesian methods. Take for instance the synthetic estimator (Gonzalez, 1973). When combined with post-stratification, this amounts to a group-mean model (Holt, Smith, and Tomberlin, 1979). Since the group-mean, or the post-stratum mean, will actually vary from one area to another, assuming them to be constant generally leads to loss of variation in the resulting estimates. Modeling the mean of a binary variable through the logistic regression models presents a similar case. Here over-shrinkage of the estimates is often referred to as over-dispersion of the true area-means (Cox and Snell, 1989). The random-effect approach of the generalized linear mixed model can be very helpful (Breslow and Clayton, 1993; Jiang, 2000). However, the data in small area estimation can be absent or extremely sparse in a large number of areas, which makes it impossible to estimate the random-effects in these areas from the sample.

We shall develop the methods of *dispersion preserving estimation* from a non-Bayesian point of view, short-handed as *DISPREE* similarly as SPREE for the structure preserving estimation (Purcell and Kish, 1980). We begin in Section 2 by defining dispersion to be a finite-population characteristic which measures the variation of the small area parameters. Through a decomposition of the dispersion, we will show that the post-stratification based synthetic estimator entails loss of dispersion in general. Moreover, its error consists of two components. The first one of these arises from the sampling error, and tends to zero in probability under suitable regularity conditions. Whereas the second one, which we call the dispersion error, is a characteristic of the population, and will eventually dominate the sampling error. It follows that confidence intervals based on the sampling error alone, though valid under the group-mean model, asymptotically lead to increasing under-coverage. That is, the proportion of the true area-parameters which fall within these intervals will be farther and farther below the nominal level of confidence as the sample grows larger and larger. We apply the DISPREE based on the synthetic estimator to the Employment data collected in the Census 1990. Having estimated the loss of dispersion of the synthetic municipality Employment Rate estimates, we derive the asymptotic confidence intervals of the area-parameters assuming normally distributed dispersion errors. The intervals turn out to be unnecessarily long. That is, the nominal level of confidence now becomes lower than the true level of coverage. This is because the correlation, between the Census-Employment and the auxiliary Register-Employment, is much weaker at the unit-level than at the municipality-level.

In Section 3 we construct a bivariate variance-component model directly at the area-level which, similar to the multivariate components of variance model of Fuller and Harter (1987), contains both area-level and unit-level random effects. The variances of the random effects is derived directly from a few parameters of the population. When applied to the data of the Census 1990, the model provides confidence intervals with correct coverage level. In fact, we may simplify the model to contain area-level random effects alone, and it works almost as well. Neither model, however, produces satisfactory estimation of the distribution of the true area-parameters. We argue that this is because super-population models as such fail to recognize the finiteness of the population. In the rest of Section 3 we shall develop a finite-population DISPREE approach through a concept of empirical finite-population distribution function (EFPDF). We demonstrate the method on the data of the Census 1990, which preserves the distribution of the true municipality Census-Employment Rates, in addition to producing confidence intervals with correct coverage level. We discuss how the method can be applied to the updated Labour Force Survey (LFS) situation. Finally, Section 4 provides a short summary. We notice that several European countries will base the upcoming Census on their administrative register systems, instead of collecting the information in the field. Improved small area estimation methods may prove to be valuable for assessing the quality of such Register Counting.

2 DISPREE based on the synthetic estimator

Denote by a the *area index*, $a = 1, ..., A$. Denote by h the *post-stratum index*, $h = 1, ..., H$, based on auxiliary information of Sex, Age and so on. Denote by U_{ah} the *population-stratum* cross-classified by a and h. Let N_{ah} be the size of U_{ah}, and n_{ah} the size of the corresponding sub-sample. Let $N_a = \sum_h N_{ah}$, and $N_h = \sum_a N_{ah}$, and so on. Let $u_{ah} = N_{ah}/N_a$ be the marginal distribution of the post-strata within area a. Denote by p_{ah} the mean of a binary *survey* variable from U_{ah}. Denote by an overbar the arithmetic average of a variable over a, such that $\overline{u_{ah}} = \sum_a u_{ah}/A$ and $\overline{p_{ah}} = \sum_a p_{ah}/A$. Define the *(finite-population) co-dispersion* of $\{u_{ah}\}_{a=1}^{A}$ and $\{u_{aj}\}_{a=1}^{A}$ as

$$\Psi(u_{ah}, u_{aj}) = \overline{u_{ah}u_{aj}} - \overline{u_{ah}} \cdot \overline{u_{aj}} = \overline{(u_{ah} - \overline{u_{ah}})(u_{aj} - \overline{u_{aj}})}.$$

Define the *(finite-population) dispersion* of $\{p_a\}$, denoted by $\Psi_2(p_a)$, as the co-dispersion of $\{p_a\}$ and itself, i.e. $\Psi_2(p_a) = \Psi(p_a, p_a)$. Let $\tau_{hj} = \Psi(u_{ah}, u_{aj})$ and $\sigma_{hj} = \Psi(p_{ah}, p_{aj})$. We have,

$$\begin{aligned}\Psi_2(p_a) &= \Psi(\sum_h u_{ah}p_{ah}, \sum_h u_{ah}p_{ah}) = \sum_{h,j} \overline{u_{ah}p_{ah}u_{aj}p_{aj}} - \sum_{h,j} \overline{u_{ah}p_{ah}} \cdot \overline{u_{aj}p_{aj}} \\ &= \sum_{h,j} \overline{u_{ah}u_{aj}} \cdot \overline{p_{ah}p_{aj}} - \sum_{h,j} \overline{u_{ah}} \cdot \overline{p_{ah}} \cdot \overline{u_{aj}} \cdot \overline{p_{aj}} \\ &= \sum_{h,j} (\tau_{hj} + \overline{u_{ah}} \cdot \overline{u_{aj}})(\sigma_{hj} + \overline{p_{ah}} \cdot \overline{p_{aj}}) - \sum_{h,j} \overline{u_{ah}} \cdot \overline{u_{aj}} \cdot \overline{p_{ah}} \cdot \overline{p_{aj}} \\ &= \sum_{h,j} \overline{p_{ah}} \cdot \overline{p_{aj}} \cdot \tau_{hj} + \sum_{h,j} \overline{u_{ah}u_{aj}} \cdot \sigma_{hj},\end{aligned}$$

provided that $\Psi(u_{ah},\ p_{ah}) = 0$ and $\Psi(u_{ah}u_{aj},\ p_{ah}p_{aj}) = 0$. We notice that, while these two assumptions greatly simplifies the expression of $\Psi_2(p_a)$, their validity need to be checked in practice.

Define *synthetic area-means* to be of the form $\sum_h u_{ah}p_h$, for $a = 1, ..., A$, where we set p_{ah} to be some constant p_h regardless of a. In particular, denote by $\tilde{p}_a$ the synthetic mean where $p_h = \overline{p_{ah}}$, so that $\tilde{p}_a = \sum_h u_{ah}\overline{p_{ah}}$. It follows that $\Psi_2(\tilde{p}_a) = \sum_{h,j} \overline{p_{ah}} \cdot \overline{p_{aj}} \cdot \tau_{hj}$. Conditional on u_{ah}, we have $\Psi_2(p_a|u_{ah}) = \sum_{h,j} u_{ah}u_{aj}\sigma_{hj}$, and

$$\Psi_2(p_a) = \Psi_2(\tilde{p}_a) + \overline{\Psi_2(p_a|u_{ah})}. \qquad (1)$$

The *decomposition of dispersion* (1) makes it clear that the synthetic area-mean $\tilde{p}_a$ generally entails loss of dispersion, or *over-shrinkage*, which is measured by the second term on the right-hand side.

Let us from now on concentrate on the case where p_a is the municipality Labour Force Survey (LFS) Employment Rate for two reasons: (a) it simplifies the discussions, and (b) it is the type of data which we shall use to illustrate our methods. Denote by q_a the municipality Register-Employment Rate from area a, which is constructed from the administrative registers independent of the LFS, and can be linked to the LFS at the unit-level. Let $u_{a1} = q_a$, i.e. the Register-Employed, and $u_{a2} = 1 - q_a$, i.e. the Register-Unemployed, and $H = 2$.

Example: The LFS of the 4th quarter in 1997. This quarterly LFS was arbitrarily chosen. First of all, we have $\Psi_2(\tilde{p}_a) = \Psi_2\{q_a\overline{p_{a1}}+(1-q_a)\overline{p_{a2}}\} = (\overline{p_{a1}}-\overline{p_{a2}})^2\cdot\Psi_2(q_a)$, so that $\tilde{p}_a$ entails loss of dispersion compared to q_a in general. As a matter of fact, the bigger the difference between $\overline{p_{a1}}$ and $\overline{p_{a2}}$, the less the loss of dispersion. It is more difficult to check on the assumptions $\Psi(u_{ah}, p_{ah}) = 0$ and $\Psi(u_{ah}u_{aj}, p_{ah}, p_{aj}) = 0$. We divide the LFS into 19 sub-samples according to which county a person comes form. We then treat the 19 sub-sample Register-Employment Rate as u_{a1}, and the 19 pairs of sub-sample post-stratum means as (p_{a1}, p_{a2}). This gives us $\Psi_2(u_{a1}) = \Psi_2(u_{a2}) = 1.03 \times 10^{-3}$, and $\Psi_2(p_{a1}) = 2.19 \times 10^{-4}$, and $\Psi_2(p_{a2}) = 5.61 \times 10^{-4}$, and $\Psi(u_{a1}, p_{a1}) = 5.09 \times 10^{-6}$, and $\Psi(u_{a2}, p_{a2}) = -4.83 \times 10^{-5}$. We have $\Psi(u_{ah}, p_{ah})/\sqrt{\Psi_2(u_{ah})\Psi_2(p_{ah})} = 0.01$ for $h = 1$ and -0.06 for $h = 2$. Similarly, we obtain $\Psi(u_{ah}u_{aj}, p_{ah}p_{aj})/\sqrt{\Psi_2(u_{ah}u_{aj})\Psi_2(p_{ah}p_{aj})} = 0.01$ for $(h, j) = (1, 1)$, and -0.06 for $(h, j) = (1, 2)$ or $(2, 1)$, and -0.01 for $(h, j) = (2, 2)$.

Let the synthetic estimator be based on post-stratification according to the Register-Employment Status alone. Let $\hat{p}_a = q_a\hat{p}_1 + (1 - q_a)\hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are the corresponding overall sample post-stratum mean. Since we do not have enough data to estimate p_{ah} directly, we need assumptions in order to evaluate the expectation of $\hat{p}_a$. Let us for the moment call the within-area post-stratum means $\{p_{ah}\}_{a=1}^{A}$ *favorable* to the sample if, for $h = 1, 2$,

$$\overline{\epsilon_{ah}} = \sum_a (n_{ah}/n_h)\epsilon_{ah} = 0 \quad \Leftrightarrow \quad \overline{p_{ah}} = \sum_a (n_{ah}/n_h)p_{ah} \qquad \text{where} \quad \epsilon_{ah} = p_{ah} - \overline{p_{ah}}.$$

Given favorable $\{p_{ah}\}$, we have $E[\hat{p}_a|n_{ah}] = \tilde{p}_a$, provided equal inclusion probability within U_{ah}. Although exact favorability is seldom attainable, approximate favorability is by no means unusual.

Example: The LFS of the 4th quarter in 1997 (continued). First of all, we notice that $\overline{q_a} = 0.700 = \sum_a(N_a/N)q_a$, so that the Register area-means are favorable to the self-weighting sample. Moreover, we have $\hat{p}_1 = 0.931$ and $\hat{p}_2 = 0.141$. The synthetic estimator is such that $\Psi_2(\hat{p}_a)/\Psi_2(q_a) = 0.625$. Whereas favorable $\{p_{ah}\}$ implies that $\Psi_2(\tilde{p}_a)/\Psi_2(q_a) \approx (0.931 - 0.141)^2 = 0.624$. It seems therefore plausible that $\{p_{a1}\}$ and $\{p_{a0}\}$ are approximately favorable to the present sample.

Given favorable within-area post-stratum means, we may decompose the error of $\hat{p}_a$ as

$$\hat{p}_a - p_a = (\hat{p}_a - \tilde{p}_a) + (\tilde{p}_a - p_a) = e_a + b_a. \tag{2}$$

The first component e_a arises from the sampling error, and tends to 0 in probability as the sample proportionally grows to infinite. We call the second component b_a the *dispersion error*. Being a population characteristic, b_a does not depend on the sample. It follows that the dispersion error eventually dominates the sampling error as the sample grows larger. In other words, the coverage level of the confidence intervals of p_a, when derived from the sampling error alone, would be farther and farther below the nominal level of confidence. Finally, since $\overline{b_a} = 0$, we have

$$\Psi_2(b_a) = \overline{b_a^2} = \Psi_2(\tilde{p}_a) + \Psi_2(p_a) - 2\Psi(\tilde{p}_a, p_a) = \Psi_2(p_a) - \Psi_2(\tilde{p}_a) = \overline{\Psi_2(p_a|u_{ah})}.$$

In this way, the error decomposition (2) attributes the asymptotic loss of dispersion of the synthetic estimator to each area, provided favorable within-area post-stratum means.

To be able to describe the dispersion error b_a in probability terms, we need a statistical model for it. Now that p_{ah} is the within-area post-stratum mean of a binary variable, multivariate normality may not be unreasonable. More explicitly, for σ_{hj} as defined in (1), let

$$Z_h \sim N(0, \sigma_{hh}) \qquad \text{and} \qquad Cov(Z_h, Z_j) = \sigma_{hj} \quad \text{for} \quad h,\ j = 1,\ 2.$$

The dispersion error $b_a = \sum_h u_{ah}\epsilon_{ah}$ is a linear combination of $\epsilon_{ah} = p_{ah} - \overline{p_{ah}}$. Assume (i) favorable $\{p_{ah}\}$, and (ii) $(\epsilon_{a1}, \epsilon_{a2})$ as iid replicates of (Z_1, Z_2), we have, as n_{ah} grows proportionally to infinite,

$$E[\hat{p}_a|u_{ah}] = \tilde{p}_a \qquad \text{and} \qquad Var(\hat{p}_a|u_{ah}) \xrightarrow{P} \sum_{h,j} u_{ah}u_{aj}\sigma_{hj}.$$

We may now derive the asymptotic confidence interval of p_a based on $\hat{p}_a$ which preserves any aprior dispersion of p_a. Assume $\Psi_2(p_a)$, the nominal 95%-confidence interval of p_a is given as

$$(\hat{p}_a - 1.96s,\ \hat{p}_a + 1.96s) \qquad \text{where} \quad s^2 = \Psi_2(p_a) - \Psi(\hat{p}_a). \tag{3}$$

Notice that it is generally unrealistic to estimate σ_{hj} directly from the sample. Neither is the last Census necessarily of much help here due to developments or changes in the auxiliary information.

Example: Census 1990. Let p_a be the municipality Census-Employment Rate, where $A = 435$. Notice that the definition of the Census-Employment differs from that of the LFS-Employment. Neither is q_a of the same quality as the present one due to improvements in the Registers. In any case, we have $\Psi_2(q_a) = 0.270 \times 10^{-2}$ and $\Psi_2(p_a) = 0.235 \times 10^{-2}$. Based on the 2nd quarter LFS in 1990, we obtain $\hat{p}_1 = 0.941$, $\hat{p}_2 = 0.227$, and $\Psi_2(\hat{p}_a) = 0.138 \times 10^{-2}$. To account for the definition differences, we adjust the mean of $\hat{p}_a$ to be the same as that of p_a, in which case the error, i.e. $\hat{p}_a - p_a$, varies from -8.8% to 5.8%. The sample post-strata sizes are $(n_1, n_2) = (12915, 7760)$, based on which we could derive the confidence interval of p_a, assuming the validity of the group-mean model. However, the coverage level of the resulting nominal 95%-confidence intervals is only 19.6%. Whereas that of the dispersion preserving 95%-confidence intervals by (3) is 98.7%, where $s = 0.031$ (Figure 1).

The concept of favorability in the development above should largely be taken heuristically. Conditionally, we have $\hat{p}_a - p_a = (\hat{p}_a - E[\hat{p}_a|n_{ah}]) + (E[\hat{p}_a|n_{ah}] - \tilde{p}_a) + (\tilde{p}_a - p_a)$. Favorable sample simplifies it to (2), whereas approximate favorable sample implies that the two are close. In any case, this is not the main reason why the confidence intervals based on the synthetic estimator are unnecessarily conservative. As noted before, the synthetic estimator amounts to a group-mean model at the unit-level, since p_{ah} here is interpreted as the probability of a person's being Census-Employment given his Register-Employment Status. Whereas the interest of inference, i.e. the municipality Census-Employment Rate, is an area-level variable. While the correlation coefficient between the binary Register- and LFS-Employment Status is 0.736 in the LFS of the 2nd quarter in 1990, the similar coefficient at the area-level, i.e. $\Psi(q_a, p_a)/\sqrt{\Psi_2(q_a)\Psi(p_a)}$, is 0.905 in the Census 1990. Notice that the area-level correlation coefficient should also be 0.736, had the population been homogeneous.

3 Finite-population DISPREE

3.1 A bivariate variance-component model

Consider first a pure area-level bivariate normal distribution of $(q_a, p_a)^T$, i.e.

$$\begin{pmatrix} q_a \\ p_a \end{pmatrix} \sim N(\mu, \Sigma) \quad \text{where} \quad \mu = \begin{pmatrix} \overline{q_a} \\ \overline{p_a} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Psi_2(q_a) & \Psi(q_a, p_a) \\ \Psi(p_a, q_a) & \Psi_2(p_a) \end{pmatrix}.$$

Notice that this is in fact a simplification of a more elaborate variance-component model. Let q_a be the convolution of two random components where, for the same μ as above,

$$q_a = \mu_q + \eta_a + \gamma_a \quad \text{where } E[\eta_a] = E[\gamma_a] = 0 \text{ and } Var(\gamma_a|\eta_a) = (\mu_q + \eta_a)(1 - \mu_q - \eta_a)/N_a.$$

In other words, we consider η_a to be an area-level random component, and $\mu_q + \eta_a$ the latent area-mean. Conditional to η_a, we consider $N_a q_a \sim$ Binomial(N_a, $\mu_q + \eta_a$), and γ_a the mean of the unit-level deviations from $\mu_q + \eta_a$. We may similarly define the variance components for p_a, denoted by (η'_a, γ'_a). The covariance between q_a and p_a involves both the area-level and the unit-level random effects. Assume $Cov(\eta_a, \gamma'_a) = Cov(\eta'_a, \gamma) = 0$. Let $\rho_a = Corr(\eta_a, \eta'_a)$ at the area-level, and $\rho = Corr(\gamma_a, \gamma'_a)$ at the unit-level, we obtain the variance/covariance structure of $(q_a, p_a)^T$ as

$$Var(q_a) = \frac{N_a - 1}{N_a} Var(\eta_a) + \frac{\mu_q(1-\mu_q)}{N_a} \qquad Var(p_a) = \frac{N_a - 1}{N_a} Var(\eta'_a) + \frac{\mu_p(1-\mu_p)}{N_a}$$

$$Cov(q_a, p_a) = \rho_a\{Var(\eta_a)Var(\eta'_a)\}^{\frac{1}{2}} + \rho\{(\frac{\mu_q(1-\mu_q)}{N_a} - \frac{Var(\eta_a)}{N_a})(\frac{\mu_p(1-\mu_p)}{N_a} - \frac{Var(\eta'_a)}{N_A})\}^{\frac{1}{2}},$$

The area-level components η_a and η'_a clearly dominate the overall variation in q_a and p_a; and we obtain the pure area-level model as N_a tends to infinity for all the areas. However, the effect of γ_a and γ'_a remain to be felt as long as there are a few really small areas, where N_a is only about a few hundred. In either case, we derive the 95% confidence interval of p_a as

$$(\hat{p}_a - 1.96\sigma_a,\ \hat{p}_a + 1.96\sigma_a) \qquad \text{where} \quad \hat{p}_a = E[p_a|q_a] \quad \text{and} \quad \sigma_a^2 = Var(p_a|q_a). \tag{4}$$

Example: Census 1990 (continued). All the parameters of the pure area-level model are known from the Census. We obtain, from (4), $\Psi_2(\hat{p}_a) = 0.192 \times 10^{-2}$ and $\sigma_a = 0.021$, where the coverage level of the 95%-confidence intervals is 94.4% (Figure 1). The error $\hat{p}_a - p_a$ varies from -10.5% to 5.7%. Improvements are evident compared to the DISPREE based on the synthetic estimator. The parameters of the variance-component model are not self-evident. We set $\rho = 0.736$ based on the LFS. We obtain a method of moment estimate $Var(\eta) = 0.259 \times 10^{-2}$ as the solution of

$$\Psi_2(q_a) = \overline{\frac{N_a - 1}{N_a}} Var(\eta_a) + \overline{\frac{\mu_q(1-\mu_q)}{N_a}},$$

and $Var(\eta'_a) = 0.223 \times 10^{-2}$ similarly. Substituting these into $\Psi_2(q_a, p_a) = \overline{Cov(q_a, p_a)}$, we obtain $\rho_a = 0.913$. These give us $\Psi_2(\hat{p}_a) = 0.192 \times 10^{-2}$, and $\overline{\sigma_a} = 0.021$, and a coverage level of 94.6%, which are almost identical with those under the simplified area-level model (Figure 1). The error $\hat{p}_a - p_a$ varies form -10.5% to 5.6%. Notice that the area-estimates under both models still contain about 20% loss of dispersion now that $Corr(q_a, p_a) \approx 0.910$. More importantly, no matter how much we may improve the Register, $Corr(q_a, p_a)$ shall remain less than unity. A super-population approach, i.e. $\hat{p}_a = E[p_a|q_a]$, will never capture the distribution of p_a since $\Psi_2(\hat{p}_a)$ will always be less than $\Psi_2(p_a)$.

3.2 Empirical finite-population distribution function (EFPDF) and finite-population DISPREE using normal approximation

Let us first give a finite-population definition of the distribution of the area-parameters, denoted by θ_a for $a = 1, ..., A$. Denote by $\{\theta_{(a)}\}$ the order statistic of $\{\theta_a\}$, where $\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(A)}$. We define the *empirical finite-population distribution function (EFPDF)* of θ_a to be

$$F_\theta(t) = \frac{1}{A}\sum_{a=1}^{A} I_{\theta_a \leq t} \qquad \text{where} \quad I_{\theta_a \leq t} = 1 \text{ if } \theta_a \leq t \quad \text{and} \quad I_{\theta_a \leq t} = 0 \text{ if } \theta_a > t. \tag{5}$$

The EFPDF is thus equivalent to $\{\theta_{(a)}\}$. Notice that the EFPDF is numerically identical with the empirical culmulative distribution function (ECDF) when $\{\theta_a\}$ is considered an iid sample. The ECDF is a nonparametric approximation to the true distribution that has generated the iid sample. However, the randomness in the area-parameters $\{\theta_a\}$ given the EFPDF F_θ is entirely different from the randomness of an iid sample $\{\theta_a\}$ given F_θ as their estimated identical distribution. In fact, conditional to the EFPDF, any admissible set of $\{\theta_a\}$ must by definition be a permutation of $\{\theta_{(1)}, ..., \theta_{(A)}\}$, in which sense the area-parameters are now dependent of each other.

By restricting $\{\hat{\theta}_a\}$ to the permutations of $\{\theta_{(a)}\}$, we ensure that they all have the same distribution F_θ and, in particular, the same dispersion. However, not all the permutations are equally probable. That depends on the distribution of $\{p_a\}$ conditional to $\{q_a\}$, such as that under the variance-component model earlier. We propose a finite-population DISPREE procedure as follows:

1. generate p^*_a from the corresponding normal distribution (4) of p_a conditional to q_a under either the pure area-level model or the variance-component model, for $a = 1, ..., A$;

2. identify the order of $\{p^*_1, ..., p^*_A\}$, denoted by $\{r_1, ..., r_A\}$, such that $p^*_a = p^*_{(r_a)}$;

3. set $p^{(1)}_a = p_{(r_a)}$ where $\{p_{(1)}, ..., p_{(A)}\}$ are given by the true EFPDF of p_a.

Independent repetitions of Step 1 - 3 give us the approximate joint distribution of $(p_1, ..., p_A)$ conditional to both $(q_1, ..., q_A)$ and F_p. Under either model, the order of $E[p_a|q_a]$ coincides with the order of q_a. A method of moment estimator of p_a is therefore given by

$$\hat{p}_a = p_{(r_a)} \qquad \text{where} \quad q_a = q_{(r_a)}. \tag{6}$$

We could now use the sample percentile interval of $\{p^{(1)}_a, ..., p^{(B)}_a\}$, where B is the number of resamples, as the estimated confidence interval of p_a. Or, to obtain confidence intervals which vary more smoothly over the areas, we could calculate, at the nominal 95%-level,

$$(\mu_a^* - 1.96 s_a^*,\ \mu_a^* + 1.96 s_a^*) \quad \text{where} \quad \mu_a^* = \frac{1}{B}\sum_{j=1}^{B} p_a^{(j)} \quad \text{and} \quad s_a^* = \{\frac{1}{B}\sum_{j=1}^{B}(p_a^{(j)} - \mu_a^*)^2\}^{\frac{1}{2}}. \tag{7}$$

Example: Census 1990 (continued). Due to the finiteness of the population, the simulation of the coverage level has a precision modulus of $1/A = 0.2\%$. Nevertheless, repeated simulations at the same value of B suggest that the finite-population adjustments of (6) and (7) are negligible here, both in terms of the confidence levels and the first-order error $\hat{p}_a - p_a$. The apparent improvement lies in the preservation of the distribution of p_a. The results under the pure area-level model have been plotted in Figure 1.

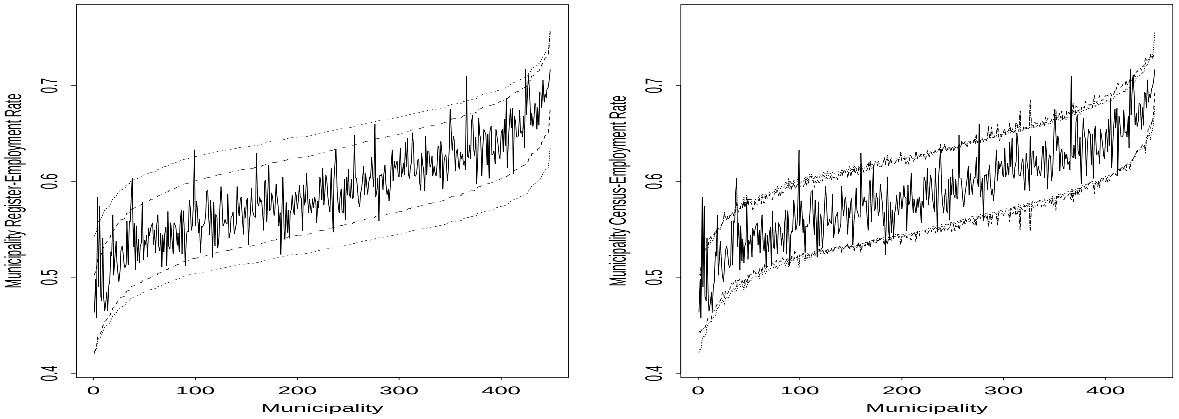

Figure 1: DISPREE based on the Census 1990 data. Left panel: Municipality Census-Employment Rate (solid), 95%-confidence intervals based on synthetic estimator (dotted), and under the area-level model (dashed). Right panel: Municipality Census-Employment Rate (solid), 95%-confidence intervals under the variance-component model — super-population approach (dotted) and finite-population approach (dashed).

3.3 Finite-population DISPREE of the LFS data

Asymptotic theories of the order statistics from general parametric distributions are available (e.g. Cox and Hinkley, 1974, Appendix 2). In particular, Blom (1958) suggested, for $Z_1, ..., Z_A \overset{iid}{\sim} N(0,1)$,

$$\xi_{(a)} = E[Z_{(a)}] = \Phi^{-1}(k) \qquad \text{and} \quad k = (a - 3/8)/(A + 1/4), \tag{8}$$

where Φ^{-1} denotes the inverse of the standard normal CDF. We obtain from (8) the asymptotic expectation of the order statistics of arbitrary $N(\mu, \sigma^2)$-distribution as $\mu + \sigma\xi_{(a)}$. Assume for the moment $\overline{p_a}$ and $\Psi_2(p_a)$ to be known. Provided the normal approximation to F_p, we could apply formula (8) directly, using $\overline{p_a}$ as the mean and $\Psi_2(p_a)$ as the variance. Notice that the resulting $\hat{F}_p$ is always symmetric about $\overline{p_a}$. On the other hand, denote by F_θ some other known EFPDF to which normal approximation is valid. We may derive $\hat{F}_p$ as a *parallel shift* of F_θ, i.e.

$$\hat{p}_{(a)} = \overline{p_a} + R(\theta_{(a)} - \overline{\theta_a}) \qquad \text{where} \quad R^2 = \Psi_2(p_a)/\Psi_2(\theta_a),$$

which generally is asymmetric about $\overline{p_a}$. Possible choice of θ_a could be the Register q_a or the synthetic $\hat{p}_a$. Since θ_a is known, it is easy to check whether its normal approximation is valid.

Example: Census 1990 (continued). We derive the asymptotic expectations of $q_{(a)}$ and $p_{(a)}$ by (8) based on, respectively, $\{\overline{q_a}, \Psi_2(q_a)\}$ and $\{\overline{p_a}, \Psi_2(p_a)\}$, and compare them to the true $q_{(a)}$ and $p_{(a)}$. In addition, we derive $\hat{F}_p$ as parallel shifts of F_q and $F_{\hat{p}_a}$, where $\hat{p}_a$ is the synthetic estimator. All of them have been plotted in Figure 2. The difference between the normal approximation and the true value varies from -0.6% to 1.5% for $q_{(a)}$. For the approximations of $p_{(a)}$, it varies from -1.4% to 1.8%. In all the three cases, the $\hat{F}_p$ is dispersion preserving, and yields similar confidence intervals as the true F_p for the Census 1990 data.

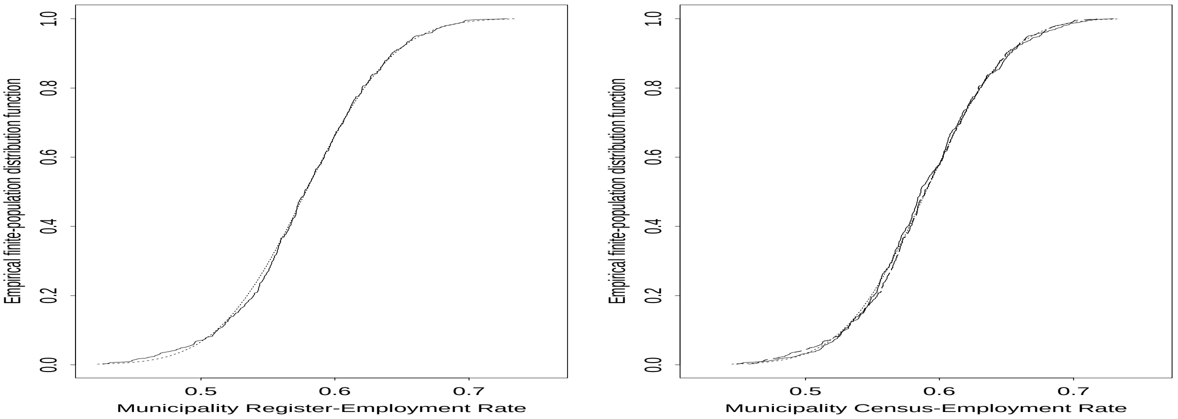

Figure 2: Empirical finite-population distribution functions based on the Census 1990 data and their normal approximations. Left panel: Municipality Register-Employment Rate (solid) and normal approximation (dotted). Right panel: Municipality Census-Employment Rate (solid), normal approximation (dotted), parallel shift of Register-Rate (dashed) and parallel shift of synthetic estimator (long-dashed).

In order to apply the method of finite-population DISPREE to the present LFS data, we could use the mean of the synthetic estimator as an estimate of $\overline{p_a}$. We need also some plausible aprior value of $\Psi_2(p_a)$, which may for instance be based on the dispersion of the present q_a as well as those of (q_a, p_a) from the last Census. The area-level correlation coefficient $Corr(q_a, p_a)$ is perhaps more difficult to set. For instance the improvements in the Register sources has raised the unit-level correlation coefficient from 0.736 in the 2nd quarter of 1990 to 0.782 in the 4th quarter of 1997. The values from the Census 1990, i.e. 0.905 of the area-level model and 0.913 of the variance-component model, are therefore likely to be the lower bounds of the present ones. Also it is reasonable to seek advice form the subject-matter experts regarding the choices of $\Psi_2(p_a)$ and $Corr(q_a, p_a)$.

Example: Census 1990 (continued). Provided the normal approximation to F_p, the error in $\hat{F}_p$ is directly determined by those of $\overline{p_a}$ and $\Psi_2(p_a)$. The error in the coverage level of the estimated confidence intervals, on the other hand, does not have a closed form. We have therefore performed a simple sensitivity analysis of both the area-level and the variance-component models. Since the finite-population adjustment is negligible with respect to the coverage level, we could use the super-population approach (3) here. Let RHO $= Corr(q_a, p_a)$ and $\Psi_2(p_a)$ vary over a grid of values. We found the coverage level of the nominal 95%-confidence intervals in each case (Table 1). The results are rather similar under both of the models. They suggest that conservative choice of $Corr(q_a, p_a)$ is quite capable of safe-guarding moderate under-estimations of $\Psi_2(p_a)$.

	Dispersion of p_a (all dispersions $\times 10^{-2}$)						
RHO	0.205	0.215	0.225	0.235	0.245	0.255	0.265
0.875	95.3 (95.1)	95.8 (95.1)	96.2 (95.8)	96.4 (96.0)	96.7 (96.7)	96.9 (96.9)	96.9 (96.9)
0.890	94.4 (94.2)	94.9 (94.9)	95.3 (95.1)	95.8 (95.3)	96.0 (96.0)	96.2 (96.2)	96.7 (96.2)
0.905	92.6 (92.9)	93.7 (93.5)	94.4 (94.0)	94.4 (94.6)	94.9 (95.1)	951 (95.1)	96.0 (95.5)

Table 1: Coverage level $(\%)$ of the 95%-confidence intervals at the various choices of $\Psi_2(p_a)$ and $Corr(q_a, p_a)$. The area-level model (without parentheses) and the variance-component model (within parentheses).

4 Summary and discussions

We have studied two non-Bayesian methods in dealing with over-shrinkage of small area estimators. The first one was based on the synthetic estimator. We began by defining dispersion as a finite-population measure of the variation of the small area parameters. Through two decompositions of the dispersion, we showed that the post-stratification based synthetic estimator entails loss of dispersion in general, and that the coverage level of the confidence intervals could be far below the nominal level of confidence, when these are derived from the sampling error alone. We derived the dispersion preserving confidence intervals, which turned out to be unnecessarily conservative due to a much weaker correlation between the survey and auxiliary variables at the unit-level than at the area-level. We therefore put up a bivariate variance-component model, as well as its simplification, directly at the area-level. This improved the efficiency of the confidence intervals. However, the super-population approach was unable to capture the distribution of the true area-parameters. We introduced the empirical finite-population distribution function of the area-parameters, conditional to which a finite-population DISPREE procedure provided the necessary adjustment. The various methods were illustrated using the data of the Census 1990. We also examined the possibility of applying the method to the updated LFS situation. Sample-based estimates of the true dispersion of the municipality LFS-Employment Rate, and the area-level correlation coefficient between the municipality Register- and LFS-Employment Rate are generally unreliable. We need to set there values *a priori*. Simple sensitivity analysis suggest that conservative choices of the area-level correlation are quite capable of safe-guarding moderate under-estimations of the true dispersion. Finally, we notice that several European countries will base the upcoming Census on their administrative register systems, instead of collecting the information in the field. Improved small area estimation methods may prove to be valuable for assessing the quality of such Register Counting.

References

Blom, G. (1958). *Statistical Estimates and Transformed Beta-variables*. New York: Wiley.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9–25.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Fuller, W.A. and Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*, ed. by R. Platek, J. Rao, C.-E. Särndal, and M. Singh, pp. 103–123. Wiley, New York.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *J. Amer. Statist. Assoc.*, **87**, 533–540.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimators. In *Proceedings of the Social Statistics Section*, pp. 33–36. Amer. Statist. Assoc., Washington, DC.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *J. Amer. Statist. Assoc.*, **74**, 405–410.

Jiang, J. (2000). Conditional inference about generalized linear mixed models. *Ann. Statist.*, **27**, 1974–2007.

Judkins, D.R. and Liu, J. (2000). Correcting the bias in the range of a statistic across small areas. *J. Off. Statist.*, **16**, 1–13.

Louis, T. (1984). Estimating a population of parameter values using bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.*, **79**, 393–398.

Purcell, N.J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *Int. Statist. Rev.*, **48**, 3–18.

Spjøtvoll, E. and Thomsen, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, **2**, 435–449.