Survey Methodology, June 2007
Vol. 33, No. 1, pp. 69-79
Statistics Canada, Catalogue No. 12-001

On standard errors of model-based small-area estimators

Nicholas Tibor Longford [1]

Abstract

We derive an estimator of the mean squared error (MSE) of the empirical Bayes and composite estimator of the local-area mean in the standard small-area setting. The MSE estimator is a composition of the established estimator based on the conditional expectation of the random deviation associated with the area and a naïve estimator of the design-based MSE. Its performance is assessed by simulations. Variants of this MSE estimator are explored and some extensions outlined.

Key Words: Composite estimation; Empirical Bayes estimation; Shrinkage; Small-area estimation.

1. Introduction

Design-based methods have over the years been proven to be inefficient for small-area estimation because, unlike empirical Bayes and related methods, they cannot make effective use of auxiliary information. However, the assumptions associated with the models that are applied remain a weakness of model-based methods because inferences based on them have the ubiquitous caveat of 'If the model is valid …'. In the application of empirical Bayes models to small-area estimation, the local areas (districts) are associated with random effects. In the design-based perspective, this assumption is not valid because in a hypothetical replication of the survey the same districts would be realised (except for some districts that happen not to be represented in the sample drawn), and the target quantities associated with them would also be the same. That is, the districts should be associated with fixed effects. The lack of validity in this aspect of empirical Bayes models has no adverse impact on estimation of small-area quantities (means, totals, proportions, and the like). Associating small areas with random effects is key to borrowing strength from or exploiting the similarity of the areas, as well as to doing so across variables, time points, surveys and other data sources, but it distorts the assessment of the precision of the estimators. Some composite estimators and estimators of their mean squared errors have the same deficiency.

In the next section we diagnose this problem in detail, and in Section 3 propose a solution, which is then illustrated and assessed in Section 4 by simulations using a set of examples. These range from the simplest and most congenial (agreeing with most of the assumptions made) to more complex and realistic but least congenial, so as to explore the robustness of the method. Its fuller potential is discussed in the concluding section.

2. Fixed and random

By sampling variance of a general estimator $\hat{\theta}$ based on a given data-generating (sampling) process χ we understand the variation of the values of $\hat{\theta}(\mathbf{X})$ in replications of the processes that generate datasets $\mathbf{X}$ and apply $\hat{\theta}$ to them. In the design-based perspective, the replication of a survey of a country with its division to D districts yields the same district-level population quantities $\theta_d, d = 1, ..., D$; these D quantities are *fixed*. In contrast, each replication in the model-based perspective, using empirical Bayes models, starts by generating a fresh set of D values θ_d, independently of the previous replications.

We regard the design-based perspective as appropriate, because, in principle, each quantity θ_d could be established with precision and a hypothetical replication of the survey would draw a sample from the same population, with the same division of the country into its districts and the same values of the recorded variables for each member of the population. Most established design-based methods are valid when the survey is based on a perfect sampling frame, which contains no duplicates and is exclusive for the studied population, and the sampling design is implemented with perfection, without any departures from the protocol. That is, the estimators they yield are (approximately) unbiased, the expressions for their sampling variances are correct, or nearly so, and these variances are estimated with small or no bias.

In contrast, model-based methods carry a much heavier burden of assumptions that often cannot be verified. Various model diagnostic procedures are available, but they are all subject to uncertainty. Interpreting failure to find a contradiction as evidence of absence of any contradiction is a commonly committed logical inconsistency. It can be overcome only by quoting properties of estimators when the assumptions are not valid, but such methods are difficult to develop because of a wide range of model violations that

1. Nicholas Tibor Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: NTL@SNTL.co.uk.

one would have to take into account. Yet, despite these drawbacks, model-based methods have proven their worth in small-area estimation and are nowadays rightly regarded as indispensable (Ghosh and Rao 1994; Rao 2003; and Longford 2005).

The EURAREA project (EURAREA Consortium 2004) carried out a large-scale simulation study involving sampling from artificially generated populations that resemble the human populations of several European countries and application of several classes of estimators. It confirmed the superiority of model-based estimators, with several qualifications, but reported rather disappointing results regarding estimators of their standard errors. We trace this problem to an averaging applied in deriving the standard errors of shrinkage estimators.

Suppose a population is divided into D districts, each of them of population size that can for all practical purposes be regarded as infinite, and independent simple random sampling schemes are applied in the districts. We assume that within each district d the outcome variable Y has the normal distribution with mean μ_d and the same variance σ_{W}^2, $N(\mu_d, \sigma_{\mathrm{W}}^2)$. For the within-district population means μ_d, we assume the superpopulation model $\mu_d \sim N(\mu^*, \sigma_{\mathrm{B}^*}^2)$, but we want to make inferences about a fixed set of (realised) means $\{\mu_d\}$. In Section 5, we discuss the more general regression setting defined by the within-district models

$$(Y|d) \sim N(\mathbf{X}_d\boldsymbol{\beta} + \delta_d, \sigma_{\mathrm{W}}^2),$$

in which $\mathbf{X}_d$ are the within-district regression matrices, $\boldsymbol{\beta}$ the set of corresponding regression parameters common to the districts, and δ_d is the deviation of the within-district regression from the typical regression defined by $\delta_d = 0$. In the superpopulation, δ_d are a random sample from $N(0, \sigma_{\mathrm{B}^*}^2)$, but we want to make inferences about the fixed (realised) set $\{\delta_d\}$. Thus, we use model-based estimators, but assess their properties by design-based criteria.

Denote by μ the (national) mean of the quantities μ_d and by σ_{B}^2 the district-level variance, $\sigma_{\mathrm{B}}^2 = D^{-1}\Sigma_d (\mu_d - \mu)^2$. Note that they differ from their respective superpopulation counterparts μ^* and $\sigma_{\mathrm{B}^*}^2$. We assume first that $\sigma_{\mathrm{B}}^2, \sigma_{\mathrm{W}}^2$ and μ are known. Let $\hat{\mu}_d$ and $\hat{\mu}$ be the sample means of the variable of interest in district d and in the whole domain (country). They are based on samples of respective sizes n_d and $n = n_1 + \cdots + n_D$. When no covariates are used the empirical Bayes (shrinkage) estimator of μ_d is

$$\tilde{\mu}_d = \left(1 - \frac{1}{1 + n_d\omega}\right)\hat{\mu}_d + \frac{1}{1 + n_d\omega}\hat{\mu}, \qquad (1)$$

where $\omega = \sigma_{\mathrm{B}}^2/\sigma_{\mathrm{W}}^2$ is the variance ratio. The model-based conditional variance of μ_d, given the data, $\mu, \sigma_{\mathrm{W}}^2$ and σ_{B}^2, equal to $\sigma_{\mathrm{B}}^2/(1 + n_d\omega)$, is often regarded as the sampling variance of $\tilde{\mu}_d$; the origins of this practice can be traced to the application of the EM algorithm. A more careful derivation acknowledges that in the design-based perspective $\tilde{\mu}_d$ is biased for μ_d,

$$\mathrm{E}(\tilde{\mu}_d|\mu_d) - \mu_d = -\frac{\mu_d - \mu}{1 + n_d\omega},$$

and its mean squared error is

$$\begin{aligned}\mathrm{MSE}(\tilde{\mu}_d; \mu_d) &= \left(1 - \frac{1}{1 + n_d\omega}\right)^2 \mathrm{var}(\hat{\mu}_d) + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2} \\ &= \sigma_{\mathrm{W}}^2 \frac{n_d\omega^2}{(1 + n_d\omega)^2} + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2}, \qquad (2)\end{aligned}$$

assuming, for simplicity, that $\hat{\mu} \equiv \mu$. To emphasise that MSE depends on the target, we include both the estimator and the target in its argument. In particular, $\mathrm{MSE}(\hat{\mu}; \mu) \neq \mathrm{MSE}(\hat{\mu}; \mu_d)$, unless $\mu_d = \mu$. An inconvenient feature of the identity in (2) is that it involves μ_d, the target of estimation. If we replace $(\mu_d - \mu)^2$ with its expectation over the districts, σ_{B}^2, we obtain the more familiar identity

$$\overline{\mathrm{MSE}}(\hat{\mu}_d; \mu_d) = \frac{\sigma_{\mathrm{B}}^2}{1 + n_d\omega}, \qquad (3)$$

the EM-related conditional model-based variance of μ_d. The bar over MSE indicates expectation (averaging) of $(\mu_d - \mu)^2$, the numerator in the last term of (2), over the districts, with the sample sizes n_d intact. Throughout, we condition on the within-district sample sizes $n_d, d = 1, \ldots, D$, even though in the sampling design each of them may be variable. $\overline{\mathrm{MSE}}$ can be interpreted as model expectation, although the expectation or average of the squared deviations $(\mu_d - \mu)^2$ could be considered and estimated for a given set of districts without any reference to a model. The conditional variance in (3) is appropriate for districts with μ_d in the 'typical' distance, σ_{B}, from the national mean μ. When $|\mu_d - \mu| \neq \sigma_{\mathrm{B}}$, an unbiased estimator of the conditional variance $\sigma_{\mathrm{B}}^2/(1 + n_d\omega)$ is biased for $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$. As the bias is related to the population quantity $\mu_d - \mu$, it is not reduced by increasing the sample size n_d.

3. Composite estimation of MSE

To estimate $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$, we reuse the idea of shrinkage and combine the alternative estimators, $\sigma_{\mathrm{B}}^2/(1 + n_d\omega)$ and a naïve estimator of the MSE in (2). This composite estimator can be motivated as follows. If $n_d = 0$, and therefore $\tilde{\mu}_d = \hat{\mu}$, we have no direct information about μ_d, so we cannot improve on $\sigma_{\mathrm{B}}^2/(1 + n_d\omega)$ as an estimator of $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$. When n_d is large, μ_d is estimated with precision sufficient for using $(\tilde{\mu}_d - \hat{\mu})^2$, possibly with an adjustment for bias, as an estimator of $(\mu_d - \mu)^2$. For

intermediate sample sizes, we search for a composition (compromise) of these two alternatives that are suitable in the extreme settings, when $n_d = 0$ and as $n_d \to +\infty$. We therefore derive expressions for their MSEs and then for the MSE of their combination.

We regard the constant $\sigma_{\mathrm{B}}^2/(1+n_d\omega)$ as an estimator, and refer to it as the *averaged* estimator of MSE. Although it has no variance, it is biased, with mean squared error

$$\begin{aligned}
&\mathrm{MSE}\left\{\frac{\sigma_{\mathrm{B}}^2}{1+n_d\omega}; \mathrm{MSE}(\tilde{\mu}_d;\mu_d)\right\} \\
&= \left\{\frac{\sigma_{\mathrm{B}}^2}{1+n_d\omega} - \frac{\sigma_{\mathrm{W}}^2 n_d \omega^2}{(1+n_d\omega)^2} - \frac{(\mu_d-\mu)^2}{(1+n_d\omega)^2}\right\}^2 \\
&= \left\{\frac{\sigma_{\mathrm{B}}^2 - (\mu_d-\mu)^2}{(1+n_d\omega)^2}\right\}^2. \qquad (4)
\end{aligned}$$

The squared deviation $(\mu_d-\mu)^2$, involved in (2), is estimated naïvely by $(\hat{\mu}_d-\hat{\mu})^2$ with bias equal to $\sigma_{\mathrm{W}}^2(n_d^{-1}-n^{-1}) \doteq \sigma_{\mathrm{W}}^2/n_d$ and, assuming that $\hat{\mu}_d$ is normally distributed,

$$\begin{aligned}
&\mathrm{MSE}\{(\hat{\mu}_d-\hat{\mu})^2; (\mu_d-\mu)^2\} \\
&= \mathrm{var}\{(\hat{\mu}_d-\hat{\mu})^2 \,|\, \mu_d\} \\
&\qquad + [\mathrm{E}\{(\hat{\mu}_d-\hat{\mu})^2 - (\mu_d-\mu)^2 \,|\, \mu_d\}]^2 \\
&\doteq \frac{2\sigma_{\mathrm{W}}^4}{n_d^2} + 4(\mu_d-\mu)^2\frac{\sigma_{\mathrm{W}}^2}{n_d} + \frac{\sigma_{\mathrm{W}}^4}{n_d^2} \\
&= \frac{\sigma_{\mathrm{W}}^2}{n_d}\left\{\frac{3\sigma_{\mathrm{W}}^2}{n_d} + 4(\mu_d-\mu)^2\right\}, \qquad (5)
\end{aligned}$$

derived from the properties of the non-central χ^2 distribution and an approximation by letting $n \to +\infty$. As an alternative, $\tilde{\mu}_d$ may be used instead of $\hat{\mu}_d$; elementary operations yield the approximations

$$\mathrm{E}\{(\tilde{\mu}_d-\hat{\mu})^2 \,|\, \mu_d\} \doteq (1-b_d)^2\left\{\frac{\sigma_{\mathrm{W}}^2}{n_d} + (\mu_d-\mu)^2\right\}$$

$$\mathrm{var}\{(\tilde{\mu}_d-\hat{\mu})^2 \,|\, \mu_d\} \doteq \frac{(1-b_d)^4}{n_d^2}\sigma_{\mathrm{W}}^2\{2\sigma_{\mathrm{W}}^2 + 4n_d(\mu_d-\mu)^2\},$$

where $b_d = 1/(1+n_d\omega)$, and so

$$\begin{aligned}
&\mathrm{MSE}\{(\tilde{\mu}_d-\hat{\mu})^2; (\mu_d-\mu)^2\} \\
&= \mathrm{var}\{(\tilde{\mu}_d-\hat{\mu})^2 \,|\, \mu_d\} + [\mathrm{E}\{(\tilde{\mu}_d-\hat{\mu})^2 - (\mu_d-\mu)^2 \,|\, \mu_d\}]^2 \\
&\doteq (1-b_d)^4\frac{3\sigma_{\mathrm{W}}^4}{n_d^2} \\
&\qquad + 2(1-b_d)^2(2-6b_d+3b_d^2)\frac{\sigma_{\mathrm{W}}^2(\mu_d-\mu)^2}{n_d} \\
&\qquad + b_d^2(2-b_d)^2(\mu_d-\mu)^4. \qquad (6)
\end{aligned}$$

This approximation is valid only for $b_d = 1/(1+n_d\omega)$, so further approximation is involved when we substitute a possibly suboptimal choice or an estimate of b_d based on an estimate of ω. In general, the coefficient b_d that minimises the MSE in (6) differs from $1/(1+n_d\omega)$ because the shrinkage with $b_d = 1/(1+n_d\omega)$ is optimal only for targets that are linear transformations of μ_d (Shen and Louis 1998). We do not pursue this avenue because the solution, being a complicated function of the parameters, is likely to be sensitive to the error in estimation of the parameters. The estimator $(\hat{\mu}_d-\hat{\mu})^2$ could be corrected for its bias in estimating $(\mu_d-\mu)^2$, although this may result in a negative estimate, especially when n_d is small.

Finally, we combine the two (biased) estimators of $\mathrm{MSE}(\tilde{\mu}_d;\mu_d)$, the averaged estimator $\sigma_{\mathrm{B}}^2/(1+n_d\omega)$ and the naïve estimator derived from the identity in (2), using $(\hat{\mu}_d-\hat{\mu})^2$ as an estimator of $(\mu_d-\mu)^2$. The MSEs of these two estimators depend on $(\mu_d-\mu)^2$, so we replace the relevant terms by their expectations across the districts d. We replace $(\mu_d-\mu)^2$ with σ_{B}^2, and $(\mu_d-\mu)^4$ with $3\sigma_{\mathrm{B}}^4$ or, in general, with $\kappa\sigma_{\mathrm{B}}^4$, where κ is the kurtosis of the (district-level) distribution of μ_d. Although it may at first appear that we have not gained anything, because we still have to remove the dependence of MSE on $(\mu_d-\mu)^2$ by using σ_{B}^2 instead, now we make this step at a later stage. In the simulations in Section 4, we show that this reduces the undesirable impact of averaging.

Thus, we search for the coefficient c_d that minimises the expected MSE of the composite estimator of the MSE,

$$\begin{aligned}
&\widetilde{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) \\
&= (1-c_d)\widehat{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) + c_d\,\overline{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d) \\
&= (1-c_d)\left\{(1-b_d)^2\frac{\sigma_{\mathrm{W}}^2}{n_d} + b_d^2(\hat{\mu}_d-\hat{\mu})^2\right\} + c_d b_d \sigma_{\mathrm{B}}^2. \qquad (7)
\end{aligned}$$

To evaluate the MSE of this MSE estimator, as a function of c_d, we use the expressions

$$\overline{\mathrm{MSE}}\{b_d\sigma_{\mathrm{B}}^2; \mathrm{MSE}(\tilde{\mu}_d;\mu_d)\} = 2b_d^4\sigma_{\mathrm{B}}^4,$$

$$\overline{\mathrm{MSE}}\{(\hat{\mu}_d-\hat{\mu})^2; (\mu_d-\mu)^2\} \doteq \frac{\sigma_{\mathrm{W}}^4}{n_d^2}(3+4n_d\omega),$$

$$\begin{aligned}
&\overline{\mathrm{MSE}}\{(\tilde{\mu}_d-\hat{\mu})^2; (\mu_d-\mu)^2\} \\
&\qquad \doteq \frac{\sigma_{\mathrm{W}}^4}{n_d^2}\{3(1-b_d)^4 + 3b_d^2(2-b_d)^2 n_d^2\omega^2 \\
&\qquad\qquad + 2(1-b_d)^2(2-6b_d+3b_d^2)n_d\omega\},
\end{aligned}$$

derived by averaging of the respective equations (4), (5) and (6); $(\mu_d-\mu)^2$ is replaced by σ_{B}^2 and $(\mu_d-\mu)^4$ by $3\sigma_{\mathrm{B}}^4$.

Assuming that the district-level targets μ_d are normally distributed, the MSE of the composite estimator in (7) is

$$
\begin{aligned}
&\mathrm{E}\{(1-c_d)(1-b_d)^2\frac{\sigma_{\mathrm{W}}^2}{n_d}+(1-c_d)b_d^2(\hat{\mu}_d-\hat{\mu})^2\\
&\quad+c_d b_d\sigma_{\mathrm{B}}^2-b_d^2\sigma_{\mathrm{W}}^2 n_d\omega^2-b_d^2(\mu_d-\mu_d)^2\}^2\\
&=b_d^4\ \mathrm{E}\{(1-c_d)\sigma_{\mathrm{B}}^2 n_d\omega+(1-c_d)(\hat{\mu}_d-\hat{\mu})^2\\
&\qquad+c_d\sigma_{\mathrm{B}}^2(1+n_d\omega)-\sigma_{\mathrm{B}}^2 n_d\omega-(\mu_d-\mu)^2\}^2\\
&=b_d^4\ \mathrm{E}\{(1-c_d)(\hat{\mu}_d-\hat{\mu})^2+c_d\sigma_{\mathrm{B}}^2-(\mu_d-\mu)^2\}^2\\
&\doteq b_d^4\left[(1-c_d)^2\left\{\frac{2\sigma_{\mathrm{W}}^4}{n_d^2}+\frac{4\sigma_{\mathrm{W}}^2}{n_d}(\mu_d-\mu)^2\right\}\right]\\
&\quad+b_d^4\left[(1-c_d)\frac{\sigma_{\mathrm{W}}^2}{n_d}+c_d\{\sigma_{\mathrm{B}}^2-(\mu_d-\mu)^2\}\right]^2,
\end{aligned}
$$

using the identities $(1-b_d)^2=b_d^2 n_d^2\omega^2$ and $\sigma_{\mathrm{B}}^2=\sigma_{\mathrm{W}}^2\omega$ to extract the factor b_d^4. By taking the expectation over the districts, keeping the sample sizes intact, we obtain

$$
\begin{aligned}
&\overline{\mathrm{MSE}}\,\{\widetilde{\mathrm{MSE}}(\tilde{\mu}_d;\mu_d)\}\\
&\qquad\doteq\frac{b_d^4}{n_d^2}\{(1-c_d)^2(3+4n_d\omega)\sigma_{\mathrm{W}}^4+2c_d^2 n_d^2\sigma_{\mathrm{B}}^4\}.
\end{aligned}
$$

The minimum of this quadratic function of c_d is attained for

$$
c_d^*=\frac{3+4n_d\omega}{3+4n_d\omega+2n_d^2\omega^2}.
$$

This choice of a coefficient c_d agrees with our expectations. For $n_d=0, c_d^*=1$ and we rely solely on the averaged MSE estimator, equal to σ_{B}^2. Further, c_d^* is a decreasing function of n_d, converging to zero as n_d diverges to $+\infty$; for large n_d we rely on the naïve estimator of MSE. It is also a decreasing function of ω; for $\omega=0$, that is, $\sigma_{\mathrm{B}}^2=0, c_d^*=1$ for every district *d*, confirming that $\mu_d\equiv\mu$ and μ_d would be estimated precisely *if* μ were known. With increasing $\omega, \sigma_{\mathrm{B}}^2/(1+n_d\omega)$ becomes less and less useful because the squared deviations $(\mu_d-\mu)^2$ are widely spread (around σ_{B}^2).

If we adjust $(\hat{\mu}_d-\hat{\mu})^2$ for its bias in estimating $(\mu_d-\mu)^2$, the expected MSE of the shrinkage estimator is minimised for

$$
c_d^\dagger=\frac{1+2n_d\omega}{(1+n_d\omega)^2}.
$$

It is easy to check that

$$
c_d^*-c_d^\dagger=\frac{n_d^2\omega^2}{(1+n_d\omega)^2}\,\frac{1}{3+4n_d\omega+2n_d^2\omega^2},
$$

so the bias-adjusted estimator derived from (2) is assigned greater weight (equal to $1-c_d^\dagger$) than the naïve estimator would be. But the difference is small for all values of $n_d\omega$.

The composite MSE estimator based on $(\tilde{\mu}_d-\hat{\mu})^2$ is derived similarly, but the resulting expression is much more complex. The optimal shrinkage coefficient is

$$
\begin{aligned}
c_d^{*\prime}&=3(1-b_d)^4+2(1-b_d)^2 f(b_d)n_d\omega-b_d(2-b_d)f(b_d)n_d^2\omega^2\\
&\quad\times[3(1-b_d)^4+2(1-b_d)^2 f(b_d)n_d\omega-\\
&\qquad\{2-4b_d(2-b_d)+3b_d^2 f(b_d)\}n_d^2\omega^2],
\end{aligned}
$$

where $f(b_d)=2-6b_d+3b_d^2$. The dependence on b_d is particularly problematic, because in practice b_d is estimated and the properties of the MSE estimator based on estimated $c_d^{*\prime}$ are bound to be affected by the uncertainty about b_d. In the derivations, we used the identity $b_d=1/(1+n_d\omega)$, so this expression could not be used when the values of b_d are set *a priori*.

4. Simulations

Properties of the composite estimator of MSE cannot be derived analytically, and so we resort to simulations. We consider the artificial setting of a national survey with a stratified sampling design, with strata coinciding with the country's 100 districts for which estimates of the means of a variable Y are sought. Simple random sampling is applied within each stratum, assumed to be of practically infinite population size. We have generated the values of the means μ_d from the normal distribution $N(\mu=20,\sigma_{\mathrm{B}}^2=8)$, and the sample sizes n_d from scaled conditional beta distributions, given the means μ_d, so as to inject a modicum of dependence of the means on the sample sizes. With this adjustment, the assumption underlying the averaged MSE estimator is false, but this could not be detected by a diagnostic procedure or a hypothesis test, not even with μ_d known. The sample size of one district was altered to be much greater than the rest, to represent the capital of the fictitious country. The within-stratum distributions of *Y* are $N(\mu_d,\sigma_{\mathrm{W}}^2=100)$. The district-level means and sample sizes are fixed in the replications. For orientation, they are plotted in Figure 1. The districts are assigned order numbers from 1 to 100 in the ascending order of their sample sizes. The smallest sample size is $n_1=15$ and the overall sample size is 3,698.

In the simulations, comprising 1,000 replications, we generate the direct estimates $\hat{\mu}_d$ as independent random draws from $N(\mu_d,\sigma_{\mathrm{W}}^2/n_d)$ and the within-district corrected sums of squares as independent draws from the appropriately scaled χ^2 distributions with n_d-1 degrees of freedom. Then we evaluate the shrinkage estimator $\tilde{\mu}_d$ for each district *d*, followed by evaluation of the averaged, naïve and the two composite MSE estimators using the coefficients c_d^* and $c_d^\dagger$ or their naïve estimates.

In the first set of replications, we assume that μ, σ_W^2 and σ_B^2 are known, so that the simulation reproduces the theoretically derived results and enables us to assess the quality of the composite MSE estimators without the interference of uncertainty about the shrinkage coefficient $b_d = 1/(1 + n_d\omega)$. The results are summarised graphically in Figure 2. The empirical biases (their absolute values) of the four MSE estimators are plotted in the left-hand panel. Circles and black dots are used for the averaged and naïve estimators, respectively, and the biases of the composite estimators are connected by solid lines. The absolute values of the empirical biases are plotted, to highlight their strong association with the sample size for the naïve estimator. For 60 districts (60%), the composite estimator of MSE has a positive bias. For the naive estimator, this count or percentage is higher (78), and for the averaged estimator lower (52). Throughout, the main contributor to the bias of the averaged MSE estimator is the deviation of the squared distance $(\mu_d - \mu)^2$ from the district-level variance σ_B^2. The two composite estimators, based on $(\hat{\mu}_d - \hat{\mu})^2$ and on its bias-adjusted version, differ so little that their biases cannot be distinguished in the plot. The diagram shows that the averaged estimator of MSE entails substantial bias for a few districts, including several with large sample sizes. The biases of the naïve and composite estimators are without such extremes.

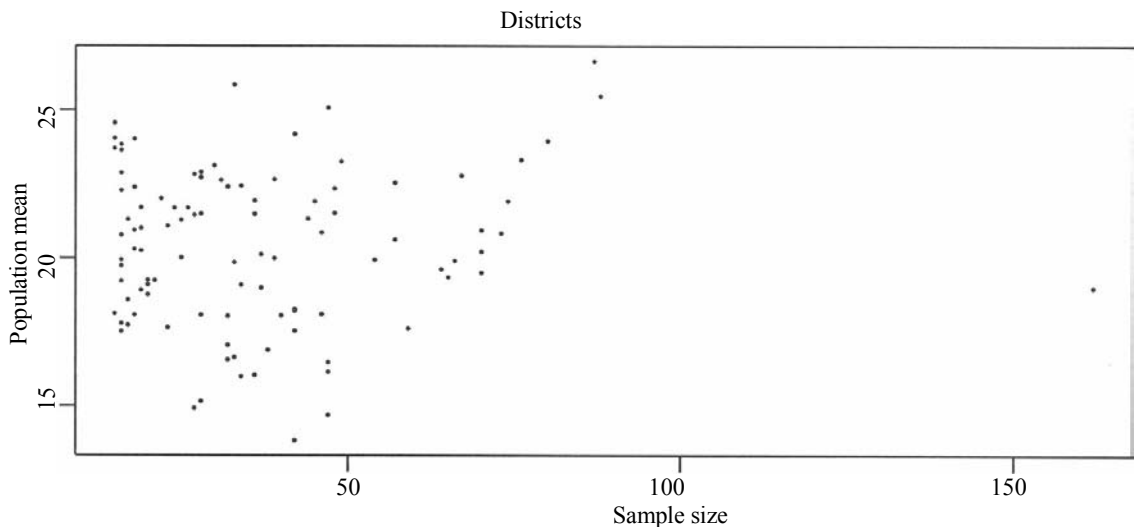

Figure 1 The district-level sample sizes and population means of *Y*. Artificially generated values

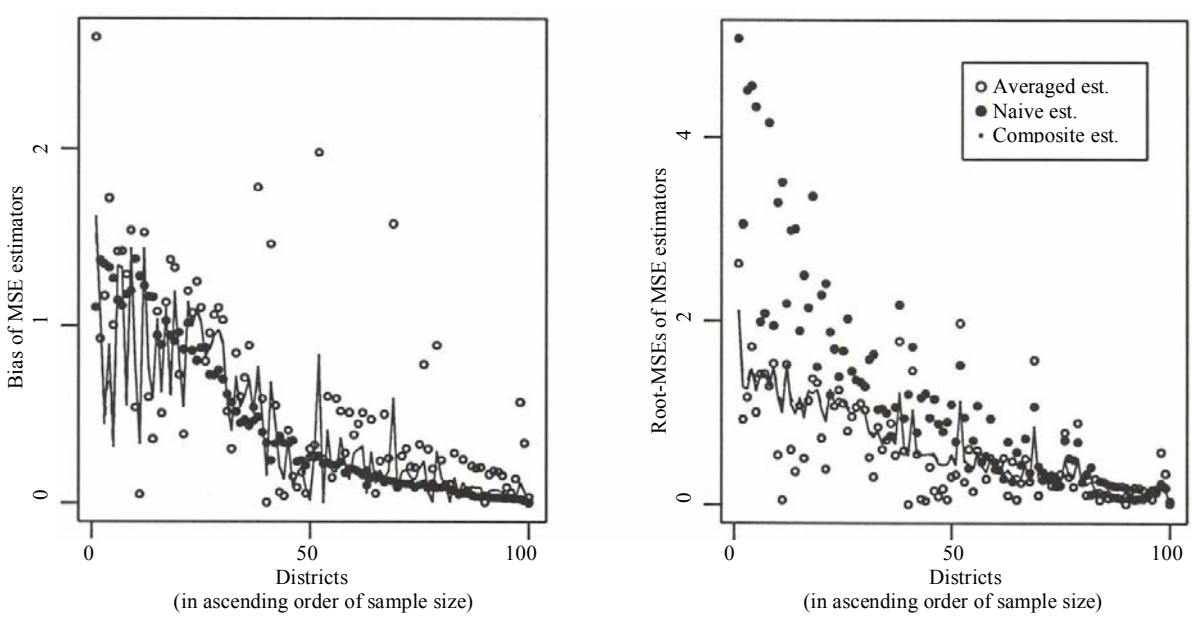

Figure 2 The bias and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. Based on simulations with an artificial setting. The bias and root-MSE of the composite estimators are connected by solid lines

In the right-hand panel, the root-MSEs of the MSE estimators are plotted, using the same symbols and layout. The diagram shows that the naïve estimator is inefficient, especially for districts with the smallest sample sizes, whereas the averaged estimator is very efficient for some but inefficient for some other districts, without any apparent relation to their sample sizes. In fact, apart from sample size, high efficiency is associated with proximity of $(\mu_d - \mu)^2$ to σ_B^2 and low efficiency with the smallest and largest values of $(\mu_d - \mu)^2$. For example, the empirical root-MSE of the averaged MSE estimator for district 1, with $n_1 = 15$, is 2.63, whereas its counterpart for district 11 $(n_{11} = 16)$ is 0.049. Their population means are $\mu_1 = 24.55$, exceeding $\mu + \sigma_B$ by 1.72, and $\mu_{11} = 22.87$, differing from $\mu + \sigma_B$ by only 0.04. The root-MSEs of the naïve estimator are 5.08 and 3.51, and those of the composite estimator are 2.10 and 1.00 for the respective districts 1 and 11. The composite MSE estimator performs much more evenly, moderating the deficiencies of the averaged and naïve estimators.

All three estimators are conservative (have positive biases) for districts with relatively small MSE of $\tilde{\mu}_d$. The averaged estimator has negative biases when the MSEs are relatively large. The composite estimator also has negative biases for such districts, but they tend to be smaller in absolute value. For districts with the smallest sample sizes, the composite estimator is not very effective because the naïve estimator is very inefficient. For a few of these districts, the composition is counterproductive, as a result of averaging, but such districts cannot be identified from a single realisation of the survey.

Next we study a less congenial setting, in which the normality assumptions of μ_d across the districts and of the elementary observations y_{id} within the districts are still satisfied, but the global parameters, μ, σ_W^2 and σ_B^2, are not known and are estimated. We use the same means μ_d and sample sizes n_d as in Figure 1. The results of the simulations are summarised in Figure 3. In the left-hand panel, the empirical means of the MSE estimators are plotted, using the same symbols as in Figure 2, together with the empirical MSEs (crosses '+') of the shrinkage estimators $\tilde{\mu}_d$. The empirical means of the averaged estimators have a regular pattern because the estimates in each replication depend only on the sample size n_d and the estimated variance ratio $\hat{\omega}$. For biases, the naive estimators have a regular pattern, similar to their pattern in Figure 2. The naïve estimators have positive biases that decline with the sample size. The averaged estimators are far too conservative; their means do not veer from the smooth trend. The composite MSE estimators deviate from this trend in the appropriate direction, but not to full degree. Their average bias is positive, equal to 0.22, or 10% (2.42 *vs.* 2.20), and they overestimate the target MSE for 70 out of the 100 districts.

The right-hand panel displays the root-MSEs of the MSE estimators. The naïve estimator is inefficient, whereas the averaged estimator is very efficient for some and rather inefficient for other districts. The composite MSE estimator is more efficient than either naïve or averaged estimator for 36 districts; it is more efficient than the averaged estimator in exactly half of the districts, but it does not have its glaring weaknesses. As in the congenial setting (Figure 2), the differences due to bias adjustment of $(\hat{\mu}_d - \hat{\mu})^2$ in composite MSE estimation (using coefficients c_d^* or $c_d^{*\prime}$) are negligible.

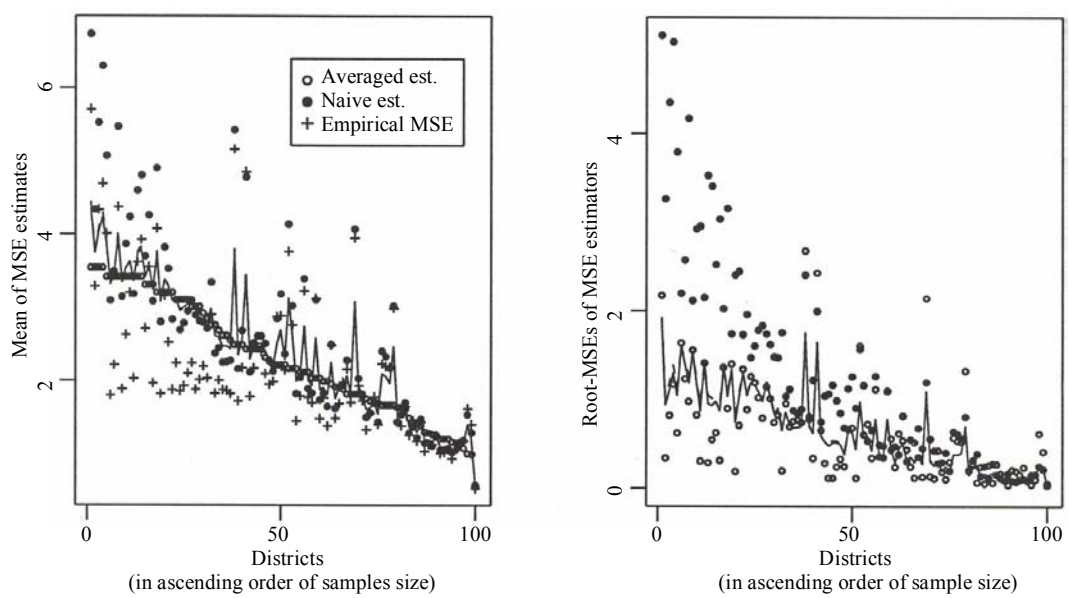

Figure 3 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. The global parameters μ, σ_W^2 and σ_B^2 are estimated

Next we compare the MSE estimators for the district-level means of $Y^2/100$, denoted by ν_d. The assumptions of normality both within and across districts are no longer appropriate. We apply the methods that rely on the normality assumptions, to assess the robustness of the composite estimators, but also to contrast the deficiencies of the averaging with the consequences of using 'incorrect' models. We chose the square transformation because the within-district expectations are known, equal to $(\mu_d^2 + \sigma_W^2)/100$, and could be estimated by

$$\tilde{\nu}_d^* = \frac{\tilde{\mu}_d^2 - \widehat{\mathrm{MSE}}(\tilde{\mu}_d) + \hat{\sigma}_W^2}{100}. \qquad (8)$$

We denote by $\tilde{\nu}_d$ the empirical Bayes estimators applied to $y_{id}^2/100$.

The results of the simulations based on the values of $y_{id}^2/100$ are presented in Figure 4, using the same layout and symbols as in Figure 3. The same conclusions about the biases and root-MSEs are arrived at as before, except that the naïve estimator is even more inefficient and the performance of the averaged estimator even more erratic - it is both very efficient and inefficient for more districts than in the more congenial setting of Figure 3. The naïve estimator is conservative, but for some districts with small n_d far too much so, and its MSEs for these districts are very large.

We contrast these conclusions with a comparison of estimating the district-level means of $Y^2/100$ by $\tilde{\nu}_d^*$, transforming the estimates $\tilde{\mu}_d$ according to (8). The estimator $\tilde{\nu}_d^*$ is more efficient than $\tilde{\nu}_d$ for most districts (90, in fact), and when less efficient, the relative difference of their MSEs is less than 4%. For a few districts, the difference in efficiency is perceptible, exceeding 20% for ten districts. However, the differences in the MSEs are small in comparison with the biases in estimating these MSEs, as shown in Figure 5. The biases and MSEs of $\tilde{\nu}_d$ are marked by black dots connected to their counterparts for $\tilde{\nu}_d^*$.

Part of the lack of efficiency of $\tilde{\nu}_d$ is due to its bias; the bias of $\tilde{\nu}_d$ exceeds the bias of $\tilde{\nu}_d^*$ for all but two districts, but the difference is non-trivial only when both estimators are positively biased. Thus, little efficiency is gained by arranging the analysis so that the distributional assumptions are satisfied. The gains are modest in comparison with the increase in the difficulty of estimating the efficiency, as expressed by $\mathrm{MSE}(\tilde{\nu}_d^*; \nu_d)$. Although the sampling variation of $\hat{\sigma}_W^2$ is trivial in large-scale surveys, the contribution of $\widehat{\mathrm{MSE}}(\tilde{\mu}_d; \mu_d)$ to $\mathrm{MSE}(\tilde{\nu}_d^*; \nu_d)$ cannot be ignored.

Figure 6 compares the composite MSE estimator with the naïve estimator of MSE of $\tilde{\mu}_d$ based on the empirical Bayes estimator of μ_d; it is derived by substituting $\tilde{\mu}_d$ for μ_d in (2). For brevity, we refer to it as the EB-naïve estimator. As anticipated in Section 3, it tends to underestimate its target. It is more efficient than the composite estimator of MSE for about half the districts (52 out of 100), but its performance is more uneven than that of the composite MSE estimator. In principle, the EB-naïve estimator could be improved by combining it with the averaged estimator; however, only minor improvement is made even in the congenial setting (known μ, σ_W^2 and σ_B^2), and the composition is detrimental for several districts in the less congenial settings. Details are omitted.

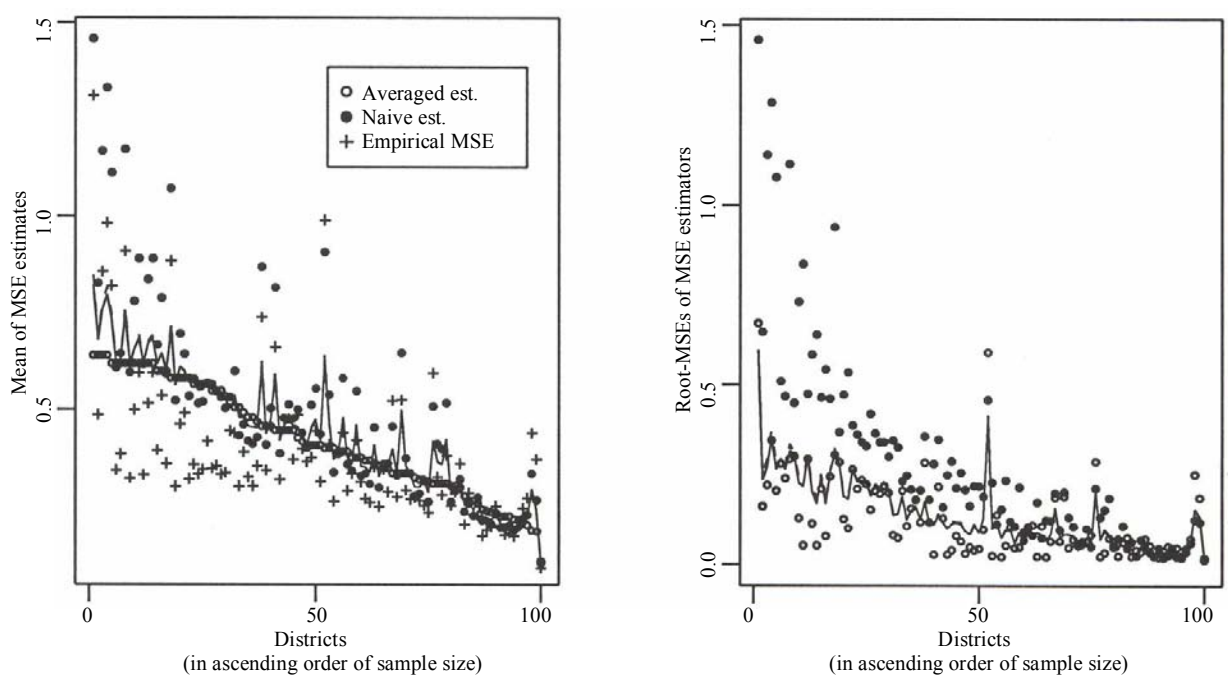

Figure 4 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators; estimation of the means of $Y^2/100$

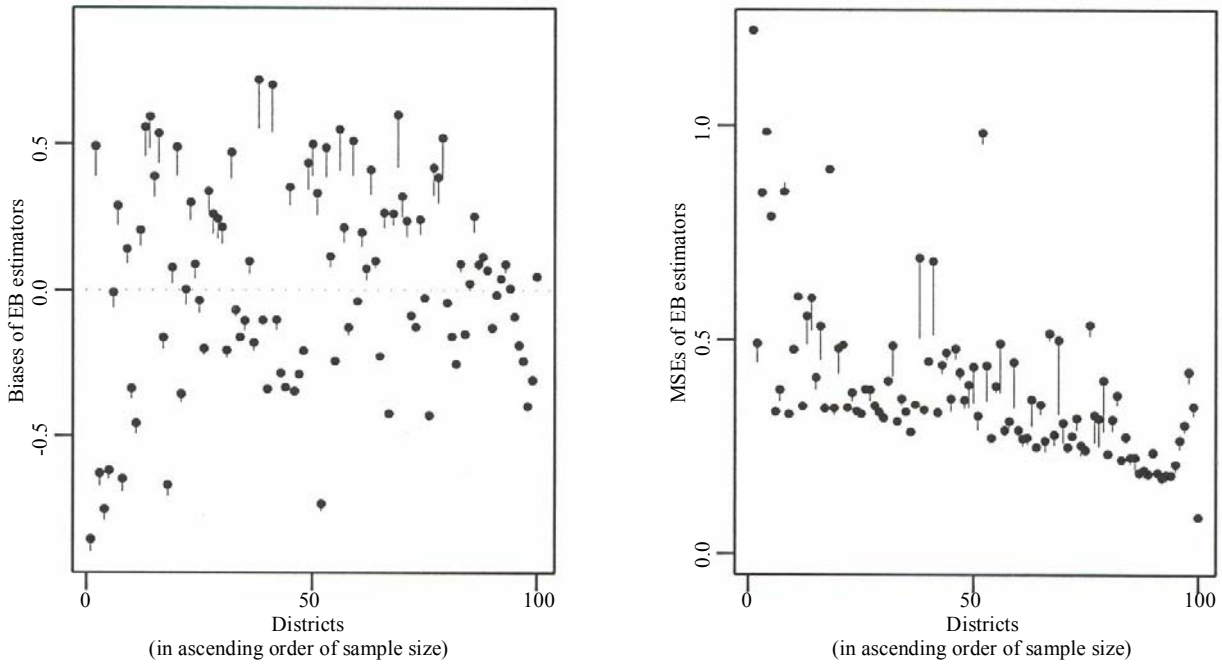

Figure 5 The biases and MSEs of estimators of ν_d. The vertical segments connect the quantities associated with $\tilde{\nu}_d^*$ and $\tilde{\nu}_d$. The quantities associated with $\tilde{\nu}_d$ are marked by black dots

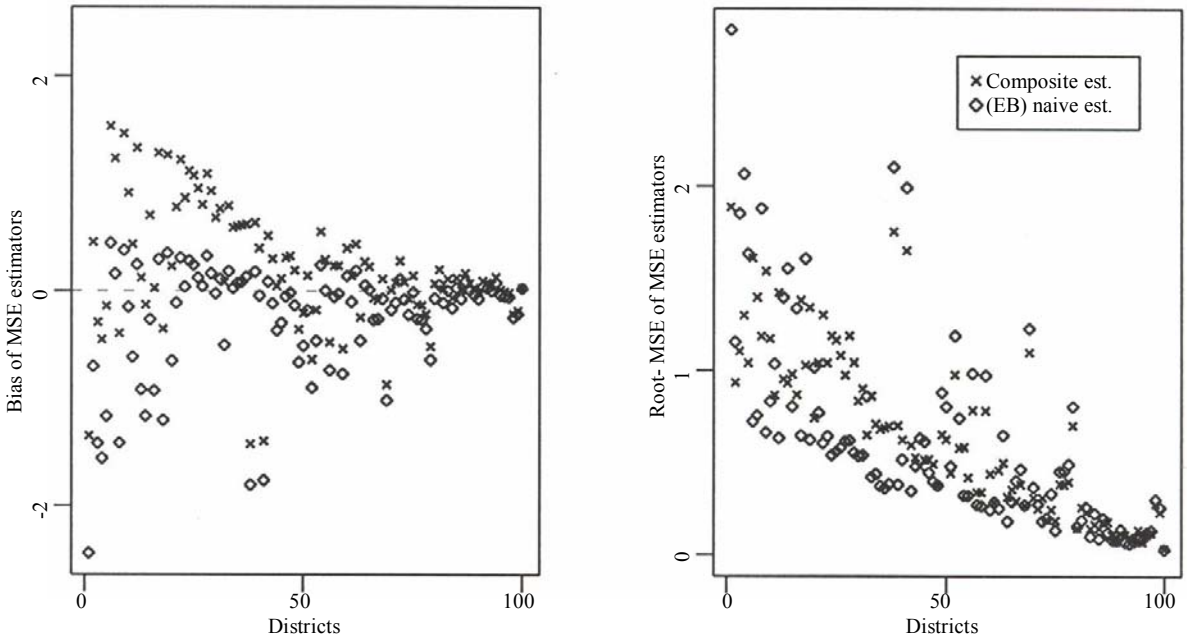

Figure 6 The bias and root-MSE of the composite and empirical-Bayes naïve estimators of the MSE of $\tilde{\mu}_d$

As a final simulation, we consider a binary outcome variable that indicates whether $Y < 5$, so that the district-level percentages are in the range 1.5-18.8 and the dependence of the percentage on the variance within districts is substantial. The mean of the district-level percentages is 6.85; the substantial skew of these percentages (skewness coefficient equal to 1.01 and kurtosis to 3.78) provides a stern test of the method.

In the simulation, the district-level percentages are estimated by the univariate version of the shrinkage method described in Longford (1999 and 2005, Chapter 8). The results are summarised in Figure 7. The MSE is over-estimated by all three estimators for most districts, except for a minority for which the empirical MSE is several times higher than for the rest. The naïve estimator has a substantial bias for most districts. The averaged estimator is less

regimented than for normally distributed outcomes because the shrinkage coefficient depends also on the estimated proportion, which is truncated from below at 2% to avoid zero estimated variance $\hat{p}_d(1-\hat{p}_d)/n_d$. The graph of the composite MSE estimates has the spikes for the appropriate districts, but the spikes are far too short to reduce the bias substantially.

The MSEs of the averaged estimator are satisfactory for most, but are very large for several districts. For the latter districts, the naïve MSE estimator is even less efficient. The composite MSE estimator is less efficient than the averaged estimator for many districts, but the difference is rather small, compensated by the gains in efficiency for districts for which the averaged estimator is less efficient. The EB-naïve MSE estimator resembles in many features the naïve MSE estimator; it is not plotted in the diagram.

In conclusion, this simulation shows that when one of the MSE estimators, in this case the naïve estimator, is very inefficient, it nevertheless contributes, even if very modestly, to the efficiency of the composite MSE estimator. The composite estimator draws on the best that the constituent estimators, averaged and naïve, have to offer, even in uncongenial settings. A remaining challenge is to combine the naïve and averaged estimators to satisfy a particular criterion which trades off the precision for districts that are estimated with high precision for higher precision in estimating in the districts with low precision. For example, we may be less concerned about estimation of the MSEs for districts with abundant representation in the sample and much more about the sparsely represented districts. Also, some districts (*e.g.*, those in a particular region) may be of specific interest, unrelated to their representation. Of course, the first step in this is the definition of one or a class of criteria that reflect the inferential priorities, and this is bound to be specific to each survey and client. See Longford (2006) for some proposals.

4.1 Refinements and extensions

Several elements of realism can be incorporated in the derivation of the composite MSE estimator. First, uncertainty about μ can be reflected by acknowledging that $\hat{\mu}_d$ and $\hat{\mu}$ are correlated. Thus, $\mathrm{var}(\hat{\mu}_d - \hat{\mu}) = \sigma_W^2\ (1/n_d - 1/n)$ and the approximation in (5) becomes equality when both instances of σ_W^2/n_d are replaced by $\sigma_W^2(1/n_d - 1/n)$. This brings about only a slight change when $n_d \ll n$, the case for most districts. If the country has a dominant district, with sample size that is a large fraction of the overall sample size, then this adjustment might be relevant, but it has a negligible impact on MSE estimation because even direct estimation of the mean for the district is nearly efficient.

A similar refinement can be applied to the empirical Bayes estimator of μ_d. It amounts to replacing n_d with $1/(n_d^{-1} - n^{-1}) = n_d n/(n - n_d)$ in the coefficient $b_d = 1/(1 + n_d\omega)$. The change is not trivial only for a dominant district, but for such a district shrinkage yields only minute improvement over direct estimation with or without this adjustment.

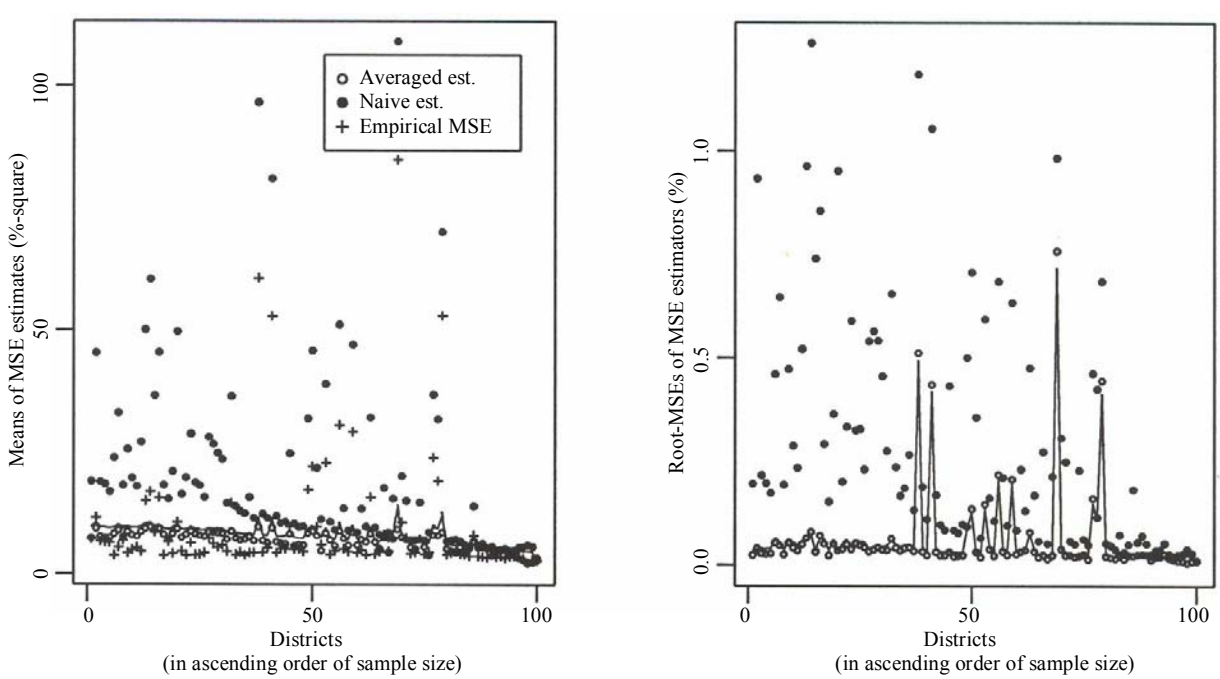

Figure 7 The mean and root-MSE of the composite naïve and averaged estimators of the MSEs of district-level percentages

Accommodating sampling designs that differ from stratified random sampling, and which associate subjects with sampling weights, generates in composite estimation no problems additional to direct estimation with such designs and weights, because we require only the sampling variances of $\hat{\mu}_d, \hat{\mu}$ and functions of these. Similarly, exploiting auxiliary information by applying (empirical Bayes) regression

$$y_{jd} = \mathbf{x}_{jd}\boldsymbol{\beta} + \delta_d + \varepsilon_{jd},$$

with independent random samples $\delta_d \sim N(0, \sigma_{\mathrm{B}}^2)$ and $\varepsilon_{jd} \sim N(0, \sigma_{\mathrm{W}}^2)$, amounts to replacing $\hat{\mu}$ in (1) with the prediction $\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}$, where $\hat{\mathbf{x}}_d$ is the vector of means of the regressors for district d and $\hat{\boldsymbol{\beta}}$ is the vector of regression parameter estimates. To see this, we express the empirical Bayes fit for district d as

$$\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}} + \frac{n_d\omega}{1+n_d\omega}(\hat{\mu}_d - \hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}) = \frac{n_d\omega}{1+n_d\omega}\hat{\mu}_d + \frac{1}{1+n_d\omega}\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}.$$

Pfeffermann *et al.* (1998) discuss issues related to fitting empirical Bayes models to observations with sampling weights. Composite estimation uses direct estimators $\hat{\boldsymbol{\mu}}_d$ and $\hat{\boldsymbol{\mu}}$ for the vectors of all the variables involved and their estimated sampling variance matrices; their evaluation is a standard task in sampling theory. An outstanding problem with empirical Bayes estimators arises when $\hat{\mathbf{x}}_d$ is based on very few observations because the uncertainty about μ_d is then inflated, even when the model fit is very good; if the vector of means $\mathbf{x}_d$ were known (available from sources external to the survey), μ_d would be estimated much more efficiently using $\mathbf{x}_d\hat{\boldsymbol{\beta}}$. Composite estimation bypasses this problem by searching for the combination of district-level means of auxiliary variables, whether known or estimated from the survey or from other sources, aiming directly to minimise the MSE of the combination (Longford 1999).

The approach developed in Section 3 can be adapted to distributions other than normal straightforwardly, so long as the kurtoses required for evaluating the district-level variance of $(\mu_d - \mu)^2$ and the sampling variance of $(\hat{\mu}_d - \mu)^2$ are known. In practice, kurtosis depends on the mean μ_d, creating difficulties that can be overcome only by approximations or averaging. Estimating proportions p_d with dichotomous data is a case in point. We have

$$\mathrm{var}\{(\hat{p}_d - p)^2\} = \frac{v_d}{n_d^3}(1 - 3p_d + 3p_d^2) + \frac{4v_d}{n_d^2}(1-2p_d)(p_d - p) + \frac{6v_d}{n_d}(p_d - p)^2 - \frac{v_d^2}{n_d^2},$$

where $v_d = p_d(1-p_d)/n_d$ and p is the national proportion. The complex dependence on the poorly estimated p_d presents an analytical challenge that does not have a universal solution.

Throughout, we assumed that the value of the variance ratio ω is known. In practice, ω is estimated. It is difficult to take account of the uncertainty about ω analytically, but its impact on estimation of μ_d and $\mathrm{MSE}(\tilde{\mu}_d; \mu_d)$ can be assessed by sensitivity analysis which repeats the simulations described in Section 4 for a range of plausible values of ω. As one set of simulations takes about one minute of CPU time, this is a manageable computational task. One difficulty in such an assessment is that with an altered assumed value of ω the estimator $\tilde{\mu}_d$ is changed, and so the target of the composite MSE estimator is also changed. An alternative informal approach considers the consequences of under- and over-stating the value of ω. In estimating μ_d it is advisable to err on the side of greater ω, giving more weight to the direct estimator $\hat{\mu}_d$ (Longford 2005, Chapter 8). For estimating the MSE of $\tilde{\mu}_d$, we may prefer to err on the side of the more stable averaged estimator. That corresponds to increasing the value of the coefficient c_d^* and, as c_d^* is a decreasing function of ω, to reducing the value of ω used for setting c_d^*. Of course, this should be done in moderation, not to discard the contribution of the naïve estimator of MSE altogether.

5. Conclusion

The approach developed in this paper applies the general idea of shrinkage to estimation of MSE of small-area estimators and reduces the impact of averaging, regarded as undesirable when viewed from the design-based perspective, in which the country's districts have fixed population quantities μ_d. We have focussed on improvement in estimation of the MSE for each district separately. In practice, improvement of estimation for some districts is more important than for others. Many surveys are designed for inferences other than small-area estimation, or take small areas into account in planning only peripherally, and so they may yield more than satisfactory estimators for some districts, typically the most populous ones, and less satisfactory for others, often the sparsely populated districts. In such a setting, relatively higher inferential priority should be ascribed to the latter districts. Shrinkage estimators of small-area means and proportions have this property, and the simulations documented in Section 4 indicate that composite estimation of MSE has a similar property, at least in relation to the averaged estimator.

For a given size of the bias in estimating an MSE, we prefer the positive bias, because we regard understating the precision as statistically 'dishonest', whereas overstating it merely fails to present the estimate in the light it deserves - we undersell the results of our analytical effort. With this perspective, the optimal coefficient c_d in (7) should not be

sought by minimising the MSE of the combination, but by a criterion that regards underestimation of MSE as an error more severe than its overestimation by the same amount. Finding a suitable criterion for this, for which optimisation is tractable, is an open problem. The composite MSE estimator derived in Section 3 tends to overestimate the MSE, but this is not by our design.

We have experimented with ML and REML estimation; in the setting used for the simulations, the differences between the two approaches are minute. The advantage of unbiased estimation of the variance σ_B^2 is lost when $\hat{\sigma}_B^2$ is subjected to a non-linear transformation, and efficiency is maintained by transformations only asymptotically. However, small-area estimation is a quintessentially small-sample problem.

The approach presented in this paper illustrates the universality of the general idea of combining alternative estimators. The composite estimator exploits the strengths and reduces the drawbacks of the constituent estimators. Applying it is not detrimental when one of the estimators is far inferior to the other. As a form of averaging is involved even in the composite MSE estimator, it contributes to its robustness by ameliorating departures from the assumptions made in the theoretical development, such as heteroscedasticity and asymmetric (non-normal) within-district distributions.

Incorporating inferential priorities, in effect, redistributing the precision in estimating the MSEs for the small areas, is an open problem. A similar problem, designing surveys for small-area estimation so as to ensure sufficient precision in the model-based perspective (with averaging) is addressed by Longford (2006).

Acknowledgement

Partial support for the work on this manuscript by Grants SEC2003-04476 and SAB2004-0190 from the Spanish Ministry of Education and Science is acknowledged. Insightful and constructive comments of two referees and an Associate Editor are acknowledged.

References

EURAREA Consortium. (2004). EURAREA Project Final Reference Volume. Enhancing Small-Area Estimation Techniques to Meet European Needs. Office for National Statistics, London. Available from http://www.statistics.gov.uk/eurarea.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

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

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. New York: Springer-Verlag.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1988). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40.

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

Shen, W., and Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society*, Series B, 60, 455-471.