STATISTICS IN TRANSITION new series, December 2026
Vol. 27, No. 4, pp. xx–XX, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.2026

Small-area estimation — a confusion of paradigms

Nicholas T. Longford[1]

Abstract

Small-area estimation is concerned with inference about the districts or regions of a country or another division of a geo-political entity. Of interest are aspects of the current state of these districts, such as the prevalences of particular economic conditions in the population of each district. Empirical Bayes models and their various adaptations are applied in their estimation, with the model adopted as an approximation to the studied reality.

We highlight the discrepancy between the modelling approach and the design-based paradigm implied by the survey data collection process. Models refer to a hypothetical superpopulation from which the districts and their units (elements) are drawn by a hierarchical sampling design, whereas the inferential goals relate to an existing, and therefore fixed, set of districts and their units, with the values of all their variables also fixed. The consequences of this discrepancy, or confusion, are analyzed and the misleading references to optimal properties, such as unbiasedness and efficiency of the best linear unbiased predictor, are pointed out. Reduced shrinkage is proposed as a partial remedy.

Our arguments are not against models as such, but against their uncritical assessment that is based on the same models. Greater integrity in such an assessment is attained by using the sampling design applied to the fixed finite population with its division to districts also fixed. Our conclusions are supported by a combination of analytical derivations, illustrations and simulations. They lead us to the consideration of the client's priorities and purposes that would inform how efficiency in estimation could be traded among the districts.

Key words: BLUP, design vs. model, hierarchical models, shrinkage.

1. Introduction

Since their beginnings based on theoretical foundations and motivation by Robbins (1956) and James and Stein (1961), computational and algorithmic innovations (Hartley and Rao, 1967; Dempster, Laird and Rubin, 1977), and proliferation of modern statistical software and large-scale administrative databases, hierarchical models have found a firm ground in a wide range of applications in areas where datasets with the nesting structure assumed by these models are collected and analyzed. Small-area estimation is an important example (Fay and Herriot, 1979; Ghosh and Rao, 1994; Longford, 2005; Rao and Molina, 2015), especially in official statistics related to social policy, employment, health care, education, and public service in general.

Estimation for one district can be strengthened by drawing on information in the other districts, and the extent of this strengthening, or borrowing strength, depends on the similarity of the districts. After all, if all the districts were identical with regard to the target

[1] SNTL Statistics Research and Consulting, London, UK, and Warsaw School of Economics, Warsaw, Poland. E-mail: sntlnick@sntl.co.uk. ORCID: https://orcid.org/000–0003-4129-9726.

© Nicholas T. Longford. Article available under the CC BY-SA 4.0 licence

of estimation, then the overall (national) estimator would be suitable for every district, and would be much more efficient than the direct estimator based on data solely from the district concerned. Districts are often substantially different in any outcome of interest, but they may become much more similar after an adjustment for one or a set of relevant covariates. This provides a strong rationale for a modelling approach.

The original paradigm for inference about districts in official statistics is that of sampling. That is, there is an identified country with its division to districts, fixed at a time point in the recent past for which inference is intended. The population of each district as well as the attributes of every one of its members (residents, households, businesses, or the like) are also fixed. The only source of variation is the sampling process based on the sampling design, by which a sample of units (members of the population) is selected. The values of a set of variables, including all those for which inference is sought, are measured on the units in this sample or elicited from them. Usually, the targets of such inference are subpopulation means or proportions, such as the rates of poverty in the country's districts.

Hierarchical or two-level models, for units (elements) within districts, facilitate borrowing of strength through assumptions that the districts are a random sample from a superpopulation of districts. In such models, the district-level variance, or its multivariate version, is a pivotal parameter that describes the similarity of the districts after an adjustment made by the model. Thus, the districts, fixed in the sampling paradigm, become random in the frequentist or Bayesian modelling paradigm. With random districts, replication of the data generating process would yield a different set of districts, each with a different composition of units. Inferences would then be drawn about different districts. This contravenes the original goal of estimation related to a given set of districts and their populations. In the frequentist paradigm, the standard error of an estimator is defined as the standard deviation of its values across replications. However, the estimator for any given district is well defined only in replications where the same district happens to be realized. Therefore, inference with conditioning on this district is problematic.

Model-assisted and design-consistent small-area estimators, Gao and Wakefield (2024) and references therein, incorporate features of the sampling design, the sampling weights and clustering in particular, but their properties are commonly evaluated with reference to the assumed model. Such evaluation is often without any concern for the individual districts, especially those with some exceptional features. For instance, the average bias may be close to zero but it hides the substantial biases for a few exceptional districts, which even out in the average. In a wide-ranging review of methods for small-area estimation Pfeffermann (2013) distinguishes the classes of design-based and model-based methods and notes that the former are usually evaluated within the design-based paradigm whereas the latter are evaluated with respect to the adopted model. Many contemporary approaches are best described as hybrids that combine models and design-based estimation. For example, direct estimates for districts are based entirely on the design and strength is then borrowed across districts according to a model.

In this article we argue that, when the data-generating process is governed by a sampling design, the properties of the estimators should be assessed with respect to this design, irrespective of the method of estimation. Our justification is that the model applied is not valid

even when it fits the data well and is not contradicted by any data-based diagnostic procedures. It fails in the assumptions related to the random terms in the model, a feature that cannot be observed or made transparent in the data. In contrast, design-based methods are valid, subject to perfect implementation of the design. Although implementation is rarely perfect, there is a vast array of methods that compensate for this, (multiple) imputation and benchmarking being most prominent among them.

A model-based (or model-assisted) estimator may nevertheless be more efficient than a design-based estimator, when it effectively exploits the auxiliary information and similarity of the districts. We demonstrate on simple examples that even the ideal model-based estimators are not uniformly more efficient than their design-based counterparts, when the assessment is based on the design. We identify the types of districts for which model-based estimators are less efficient than even the sample mean. We explore some adaptations which improve estimation for these districts but make estimation for some other districts less efficient. This leads to the consideration of a trade-off in efficiency, optimising a summary of the mean squared errors (MSE) of the estimators. We conclude that the priorities and purposes are an important extra-data input to the analysis that can tailor the outputs for the specific demands of the client or 'consumer' of small-area estimates.

Numerous simulation studies claim to confirm the theoretical properties of hierarchical models in the context of small-area estimation because they generate a new synthetic country in each replication. For exceptions, using enumeration of the population or a constructed database that emulates it, see Fay and Herriot, 1979, and Heady and Ralphs, 2004, respectively. The reality is that, at a set time point, the country is fixed, and the simulation scheme should use, in every replication, the same real or synthetic country, with the same division to its districts, and a fixed population of each district. This is objected to by many on the grounds that the results of the simulation do not merit any generalization. However, the preferred simulation scheme that is much more general is, in our perspective, not relevant. The districts are unambiguously fixed; at least in principle, the data-generating process could be replicated by applying, on the same population and its division to districts, the same (survey) sampling design, in which sampling is the only source of randomness.

Our objection is not to the models as such, and certainly not to the algorithms used to fit them, but to the uncritical adoption of the model as the generating process for the analyzed data. All models are wrong, but some are useful (Box, 1976). A model should be subjected to a stringent examination to appraise its usefulness and appropriateness. This is often interpreted as an instruction to inspect the fit of the model, and search for any features in the data that contradict the model, that is, to assess the goodness of fit. In small-area estimation, such a diagnosis is deficient when the various diagnostic checks are conducted on terms of the model — in such a statistical 'trial', the judge is also the jury, an anathema to impartial assessment. Model diagnostics are a very low bar for confirming usefulness, or any statistical properties, and the properties of being the best linear and unbiased estimator, or *predictor* (BLUP), in particular. We contend that an estimator derived from a model should be assessed, or diagnosed, on the terms of the implemented sampling design.

The term 'predictor' is hoisted upon us by the adaptation for which hierarchical models were originally not intended. Estimation of a set of district-level summaries is not a task

of predicting in the sense of anticipating what would, or might, happen in any particular circumstance. Simply, there is no intent to forecast the state of a district if this district were manipulated in any way or if it continued its evolution in the current circumstances. Application of a new policy in the district is an example of manipulation in this context. The target is the current or recent state of the districts, with no experimental or hypothetical intervention. The inference is for the specific districts and time point, not for any districts like the ones analyzed, nor for any time point in the future.

In the next section we give details of the model we study and contrast the inferences based on the modelling and sampling paradigms. In Section 3, we show not only that BLUP is suboptimal in the sampling paradigm, but is even less efficient than the sample mean for a nontrivial fraction of the districts. A class of estimators that use less shrinkage than BLUP is introduced and its properties are assessed in Section 4. Section 5 discusses problems that are specific to districts with small sample sizes. Section 6 illustrates the analytical results of Sections 3 and 4 on examples and simulations. Section 7 outlines some alternative solutions to the general problem of performance assessment, that is, comparing the districts on their outcomes. They combine causal inference with decision theory. Section 8 concludes with a summary, arguing that no set of small-area estimators is optimal for a wide range of purposes, and therefore the established way of publishing a single set of estimates (with associated standard errors or confidence intervals) is deficient.

2. The setting and model

We assume that there are K districts, with population sizes N_k, $k = 1, \ldots, K$. Denote by $\mathbf{1}_n$ the column vector of unities of length n. The outcome variable is denoted by Y, its $N_k \times 1$ vector of values in district k by $\mathbf{Y}_k$, and we assume that the targets of inference are the within-district means $\bar{Y}_k = \mathbf{Y}_k^\top \mathbf{1}_{N_k} / N_k$. These districts are represented in the data by $n_k \times 1$ vectors of outcomes $\mathbf{y}_k$ and $n_k \times p$ matrices $\mathbf{X}_k$ of the values of the recorded covariates. In general, $n_k \ll N_k$ for each k, and this we assume throughout, so that we are not distracted by any finite-sample correction. We do not rule out $n_k = 0$, in which case $\mathbf{y}_k$ and $\mathbf{X}_k$ are empty objects, but inference about $\bar{Y}_k$ is sought even for such districts.

In the hierarchical model for outcomes with values on an ordinal (continuous) scale, we assume the model with random variation of units' outcomes around their districts' expected outcomes, after adjustment for the covariates,

$$\mathbf{y}_k = \mathbf{X}_k \beta + \delta_k \mathbf{1}_{n_k} + \varepsilon_k \,, \tag{1}$$

where δ_k, $k = 1, \ldots, K$, are a random sample from the centered normal distribution with (unknown) variance τ^2 and the vector ε_k comprises a random sample from another centered normal distribution with variance σ^2. The $K+1$ random samples are mutually independent and the values of the variances σ^2 and τ^2 are not related. Throughout, we assume that $\sigma^2 > 0$ and that the values of $\mathbf{X}_k$ and $\mathbf{y}_k$ are free of any errors. Thus, if the data were collected for the entire population, $\mathbf{Y}_k$ would be recovered with precision, and so would be the target $\bar{Y}_k$. Randomness of each δ_k implies that any attribute of district k is also random;

for instance, the district may have high unemployment rate in one replication of the country and low unemployment rate in another. This is clearly untenable.

We use the subscripts 's' and 'm' with expectation E, variance (var) and MSE to indicate evaluation under the sampling design and model, respectively. For example, with simple random sampling within district k, the sample mean $\bar{y}_k = \mathbf{y}_k^\top \mathbf{1}_{n_k}/n_k$ has $\mathrm{var_s}(\bar{y}_k) = \sigma_k^2/n_k$, whereas $\mathrm{var_m}(\bar{y}_k) = \tau^2 + \sigma^2/n_k$ and $\mathrm{var_m}(\bar{y}_k \,|\, \delta_k) = \sigma^2/n_k$. Here σ_k^2 is the (finite-population) variance of Y in district k. The K variances σ_k^2 are unlikely to be identical even when homoscedasticity (equal within-district variances) is palatable in the model.

In the model in (1), the target for district k is $\eta_k = \mathbf{x}_k\beta + \delta_k$, where $\mathbf{x}_k$ is a $1 \times p$ (row) vector of values of the covariates; η_k is a conditional expectation that eliminates ε_k but retains δ_k. Usually, $\mathbf{x}_k$ is set to the sample or population means of the covariates in district k, but there may be a rationale for other settings, such as a vector common to all the districts; see Section 7 for an example. Estimating η_k is referred to as prediction, thus avoiding the contradiction of estimating a random quantity that involves δ_k. In fact, its conditional expectation $\mathrm{E_m}(\delta_k \,|\, \mathbf{y}_k)$ is estimated. The target in the sampling paradigm is

$$\bar{Y}_k = \eta_k + \frac{1}{N_k}\,\boldsymbol{\varepsilon}_k^{*\top}\mathbf{1}_{N_k}\,, \tag{2}$$

where $\boldsymbol{\varepsilon}_k^*$ is the $N_k \times 1$ vector of the model deviations in the population of district k. Of course, $\bar{Y}_k$ differs from η_k. In the data (sample), the residual term involving ε_k may be ignored for districts with sizeable N_k, but we should pay heed to the qualifier 'small' in 'small-area' estimation. The difference $\bar{Y}_k - \eta_k$ cannot be dismissed by a reference to the identity $\mathrm{var_m}(\bar{Y}_k - \eta_k) = \sigma^2/N_k$ because the districts have uneven influence in the analysis. In most methods of estimation, districts with more abundant representation have a stronger influence and the estimation process prefers to satisfy the maxims of the model for these 'large' districts. As a result, a (weighted) average of the residuals $\mathrm{E}(\hat{\varepsilon}_k \,|\, \mathbf{y}_k)$ vanishes, but the within-district subtotals may not.

The two paradigms differ also in some other aspects. First, ε_k is regarded as a (random) error, implying that $\mathbf{X}_k\beta + \delta_k$ is the truth. In the original perspective, $\mathbf{y}_k$ is the truth; $\mathbf{X}_k\beta + \delta_k$ is an approximation to the truth, so it entails some error, and ε_k makes up for this error. The model is wrong, but because it may be useful we treat it as if it were correct. That is a logical inconsistency. Next, δ_k is random but in the original perspective it is fixed. If δ_k were random, then in a replication of the study a given district k would have a different summary (target) η_k. Even if β were known and $\mathbf{x}_k$ fixed, the values of η_k might differ much more than what the approximation of $\bar{Y}_k$ by η_k would imply.

In summary, hierarchical models for small-area estimation have weaknesses unrelated to the quality of the model fit — they are caused by the switch from sampling design used for data collection to the adopted model. In the next section we illustrate and assess these weaknesses in a simple setting.

3. Random model, fixed reality

To avoid some distractions and simplify the discourse, we assume that $\mathbf{X}_k$ contains only the intercept, that is, $\mathbf{X}_k = \mathbf{1}_{n_k}$, and that the corresponding parameter β, the univariate version of the regression parameter vector β, together with τ^2 and σ^2 in this simplified model in (1), are known. In this model, δ_k are the only essentially unknown quantities because $\varepsilon_k = \mathbf{y}_k - (\beta + \delta_k)\mathbf{1}_{n_k}$. Denote the variance ratio $\rho = \tau^2/\sigma^2$. In the modelling paradigm, the BLUP of $\eta_k = \beta + \delta_k$ is

$$\begin{aligned} \hat{\eta}_k &= \beta + \frac{\rho}{1+n_k\rho}\left(\mathbf{y}_k - \beta\mathbf{1}_{n_k}\right)^\top \mathbf{1}_{n_k} \\ &= \beta + (1-s_k)\,(\bar{y}_k - \beta)\,, \end{aligned} \tag{3}$$

where $s_k = 1/(1+n_k\rho)$ is the shrinkage coefficient, a decreasing function of n_k. When $n_k = 0$, the equation is interpreted as $\hat{\eta}_k = \beta$.

We explore the properties of $\hat{\eta}_k$ in the sampling paradigm. We assume simple random sampling within district k and $n_k \ll N_k$, so that the finite-sample correction can be ignored. The sample mean $\bar{y}_k$ is design-unbiased for $\bar{Y}_k$ with variance σ_k^2/n_k. In the sampling paradigm, $\hat{\eta}_k$ has bias $-s_k(\bar{Y}_k - \beta)$ and variance $(1-s_k)^2\sigma_k^2/n_k$, so its MSE is

$$\mathrm{MSE_s}\,(\hat{\eta}_k \,|\, \bar{Y}_k) = s_k^2\,(\bar{Y}_k - \beta)^2 + (1-s_k)^2\,\frac{\sigma_k^2}{n_k}\,. \tag{4}$$

The model-based variance of the 'predictor' of η_k is

$$\mathrm{var_m}\,(\hat{\eta}_k) = (1-s_k)^2\left(\tau^2 + \frac{\sigma^2}{n_k}\right) = (1-s_k)\,\tau^2\,, \tag{5}$$

although it is more appropriate to consider the conditional variance

$$\mathrm{var_m}\,(\hat{\eta}_k \,|\, \eta_k) = \frac{1}{n_k}\,(1-s_k)^2\sigma^2 \tag{6}$$

because $\hat{\eta}_k$ is intended for estimating η_k. The sample mean $\bar{y}_k$ is often dismissed on the grounds that $\mathrm{var_m}(\hat{\eta}_k \,|\, \delta_k) < \mathrm{var_m}(\bar{y}_k \,|\, \delta_k)$ and $\mathrm{var_m}(\hat{\eta}_k \,|\, \delta_k) < \mathrm{var_s}(\bar{y}_k)$. These inequalities are correct, except when σ_k^2 is much smaller than σ^2, but they are irrelevant because the model-based paradigm is inappropriate. They are often used in the established orthodoxy to justify a model-based approach, possibly allowing an exception for a dominant district, such as the large capital of a relatively small country.

We regard as appropriate only a comparison in the sampling paradigm, in which $\bar{y}_k$ are mutually independent estimators. In this paradigm, $\hat{\eta}_k$ is more efficient than $\bar{y}_k$ when

$$s_k^2\,(\bar{Y}_k - \beta)^2 < \frac{1}{n_k}\left\{1 - (1-s_k)^2\right\}\sigma_k^2\,. \tag{7}$$

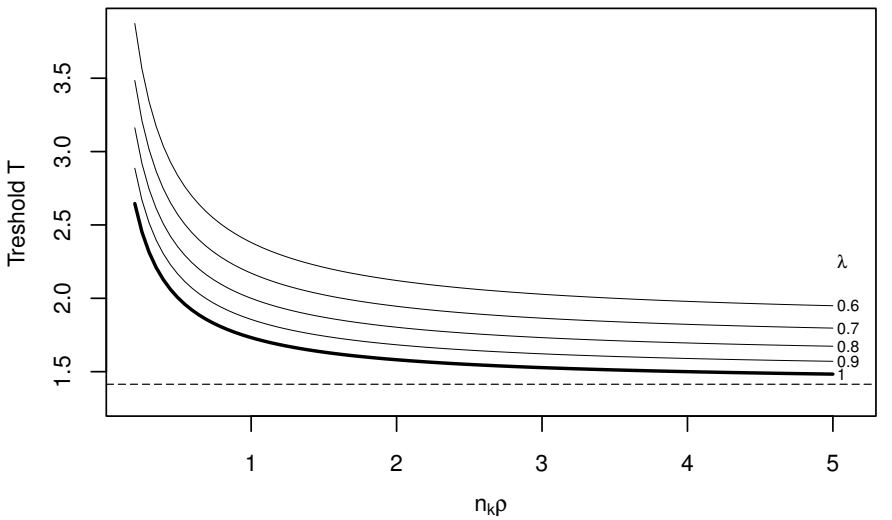


Figure 1. Thresholds for $\gamma_k = |\bar{Y}_k - \beta|/\tau$ up to which $\hat{\eta}_k$ is more efficient than $\bar{y}_k$, as a function of $n_k\rho$, in the sampling paradigm, assuming $\sigma_k^2 = \sigma^2$. The values of the shrinkage factor λ, introduced in Section 4, are printed at the right margin; BLUP corresponds $\lambda = 1$.

When $\tau^2 = 0$, then $\bar{Y}_k = \beta$ and $s_k = 1$, so this inequality is satisfied whenever $\sigma_k^2 > 0$. If $\tau^2 > 0$, the inequality is equivalent to

$$\frac{(\bar{Y}_k - \beta)^2}{\tau^2} < \frac{\sigma_k^2}{\sigma^2}\left(2 + \frac{1}{n_k\rho}\right). \tag{8}$$

Denote by γ_k the absolute scaled deviation $\frac{1}{\tau}|\bar{Y}_k - \beta|$ and set $T_k = \sqrt{2 + 1/(n_k\rho)}$. The average of γ_k^2 over the districts is 1.0. The sample mean $\bar{y}_k$ is more efficient than $\hat{\eta}_k$ when

$$\gamma_k > T_k \frac{\sigma_k}{\sigma}. \tag{9}$$

The threshold T_k, as a function of $n_k\rho$, is drawn by solid line in Figure 1. (The other lines in the plot are discussed later.) It is a decreasing function with respective limits of $+\infty$ and $\sqrt{2}$ at $n_k = 0$ and as $n_k \to +\infty$. Thus, $\hat{\eta}_k$ is more efficient for districts with small n_k or large σ_k, unless $\gamma_k\sigma/\sigma_k$ is extremely large. But it is less efficient for districts with large n_k unless $\gamma_k\sigma/\sigma_k$ is very small. For example, when $n_k\rho = 0.5$ and $\sigma_k = \sigma$, then $T_k = 2$. In this case the borderline, where $\hat{\eta}_k$ and $\bar{y}_k$ have the same MSE, is in accord with the common consideration of outliers when $\sigma_k = \sigma$. Normality implies that about 5% of such districts would be flagged as outliers. In contrast, when $n_k\rho = 2$ and $\sigma_k = \sigma$, then $T_k = 1.5$, close to the limit of $\sqrt{2}$. Assuming normality of the population means $\bar{Y}_k$, $\hat{\eta}_k$ would be less efficient for about 13% of such districts.

In summary, shrinkage estimation by BLUP is superior to the sample mean for a large fraction of the districts, but not for all of them. The weakness of $\hat{\eta}_k$ is for well-represented

districts (large n_k) that deviate from the overall mean β by more than the threshold T_k. The ratio σ_k/σ is also a factor. Note the inexactness of this statement: β is not the overall mean, but the model-related (superpopulation) expectation of the within-district (superpopulation) means. The normality assumption ameliorates this conceptual error somewhat, but not altogether. Notwithstanding this, the 'Best' (BLUP) is not the best, not even in the setting most congenial for the model. Further, $\hat{\eta}_k$ is unbiased only when $\bar{Y}_k = \beta$. In what follows, we refer to the districts for which $\bar{y}_k$ is more efficient than $\hat{\eta}_k$, that is, $\mathrm{var}_s(\bar{y}_k) < \mathrm{MSE}_s(\hat{\eta}_k)$, as *extreme*, and use the term *niche* for the settings in which district k is extreme.

Extremeness of district k, that is, inferiority of $\hat{\eta}_k$, is not related to outlier status because the threshold T_k varies with $n_k\rho$; $T_k = 2$ for $n_k\rho = 0.50$, but $T_k = 3$ for $n_k\rho = 0.14$ and $T_k = 1.5$ for $n_k\rho = 4.0$. Such values of $n_k\rho$ can easily occur in a single dataset, for example, when the ratio of the largest and smallest within-district sample size exceeds $4/0.14 \doteq 30$.

We selected a simple setting, with the trivial model (univariate β) and known variance ratio $\rho = \tau^2/\sigma^2$, to make the derivations simple, exact and involving no asymptotic arguments. From this basis, we can move towards more complex settings. First, if β and ρ are not known, the properties of $\bar{y}_k$ are unaffected whereas the properties of $\hat{\eta}_k$ are subject to a penalty owing to estimation of β and ρ. So, the niche of $\bar{y}_k$ is expanded.

By expanding the model with suitable covariates (columns of $\mathbf{X}_k$) we are likely to reduce the variances σ^2 and τ^2, and maybe also the ratio ρ, giving the BLUP a distinct edge that is partly eroded by the (increased) uncertainty about these variances as well as δ_k. To keep the contest of the two paradigms fair, we point out that estimation of $\bar{Y}_k$ can also be improved by adjustment for covariates strictly within the sampling paradigm (Särndal, Swensson and Wretman, 2003).

In summary, the empirical Bayes estimator for small areas is not more efficient for all districts; its weakness is districts with extreme values of the deviation δ_k, and it is more pronounced for districts with larger samples, when the threshold T_k is smaller, and when σ_k^2 is relatively small. The problem is compounded by us not knowing which districts are extreme, although the values of $\hat{\delta}_k$ may provide some clues. The next section proposes a solution by applying less shrinkage than used by BLUP and evaluates the trade-off between improved estimation for the extreme districts and inflation of the MSE for the rest.

4. Applying less shrinkage

The previous section uncovered a weakness of BLUP. This is not an argument against shrinkage in general but for a more measured understanding of its properties in the context of small-area estimation. In this section we explore estimators with altered shrinkage.

For districts with large deviation γ_k, the variance reduction, equal to $\mathrm{var}_s(\bar{y}_k) - \mathrm{var}_s(\hat{\eta}_k) = \sigma_k^2 s_k(2-s_k)/n_k$, may be insufficient to compensate for the bias of the shrinkage estimator, equal to $s_k\gamma_k\tau$. This suggests that we should apply less shrinkage than s_k. An obvious way to accomplish it is by replacing s_k with λs_k for a constant $\lambda \in (0,1)$. The corresponding shrinkage estimator is denoted by $\hat{\eta}_k(\lambda)$, so that $\hat{\eta}_k$ in the original notation coincides with

$\hat{\eta}_k(1)$. The MSE with this reduced shrinkage is

$$\mathrm{MSE_s}\{\hat{\eta}_k(\lambda)\} = \lambda^2 s_k^2 (\bar{Y}_k - \beta)^2 + (1-\lambda s_k)^2 \frac{\sigma_k^2}{n_k}, \tag{10}$$

and its comparison with $\mathrm{var_s}(\bar{y}_k) = \sigma_k^2/n_k$ yields the squared threshold

$$T_k^2(\lambda) = \frac{\sigma_k^2}{\sigma^2}\left(\frac{2}{\lambda} + \frac{2-\lambda}{n_k \lambda \rho}\right) = T_k^2(1) + 2\frac{\sigma_k^2}{\sigma^2}\left(\frac{1}{\lambda} - 1\right)\left(1 + \frac{1}{n_k \rho}\right). \tag{11}$$

This is a decreasing function of λ. So, the thresholds T_k increase as λ is reduced; see Figure 1, with the values of λ printed at the right-hand margin. By applying less shrinkage, λs_k instead of s_k, we raise the threshold T_k, and so improve the chances of the shrinkage estimator being more efficient than the sample mean. However, no shrinkage, $\lambda = 0$, corresponds to the direct estimator, which is not a good solution either.

Choosing $\lambda < 1$ results in a trade-off. Estimation is improved for the extreme districts at the price of poorer estimation of the districts with $\gamma_k < T_k(1)$. The equation

$$\mathrm{MSE_s}\{\hat{\eta}_k(\lambda)\} - \mathrm{MSE_s}\{\hat{\eta}_k(1)\} = 0, \tag{12}$$

quadratic in λ has two solutions, $\lambda_1 = 1$ and λ_{k2}; the latter satisfies the identity

$$\gamma_k^2 = \frac{\sigma_k^2}{\sigma^2}\frac{1}{n_k s_k \rho}\frac{2 - (\lambda_{k2}+1)s_k}{\lambda_{k2}+1}. \tag{13}$$

It is easy to check that $\gamma_k < T_k(\lambda_{k2})$. Therefore, for extreme districts, for which $\mathrm{var_s}(\bar{y}_k) < \mathrm{MSE_s}\{\hat{\eta}_k(1)\}$, $\bar{y}_k$ is less efficient than $\hat{\eta}_k(\lambda)$ for $\lambda \in (0, \lambda_{k2})$ and more efficient for $\lambda \in (\lambda_{k2}, 1)$, which includes the optimal value of λ. For districts that are not extreme, for which $\mathrm{var_s}(\bar{y}_k) \geq \mathrm{MSE_s}\{\hat{\eta}_k(1)\}$, $\hat{\eta}_k(\lambda)$ is less efficient for $\lambda < \lambda_{k2}$, and more efficient otherwise. So, by applying reduced shrinkage, $\lambda < 1$, $\hat{\eta}_k(\lambda)$ may remain superior to $\bar{y}_k$. In Section 6 we explore setting the value of λ empirically.

5. Districts with small or zero sample size

Small-area estimation appears to be particularly effective for districts with sparse or no representation in the data. The 'prediction' $\mathbf{x}_k\hat{\beta}$ offers a ready answer for districts with no sampled units, $n_k = 0$, where there are no alternatives. One weakness of this approach is that it entails extrapolation. Small districts, which tend to have small sample sizes n_k, are often outliers in the covariate space, as well as in the population size. Regression estimators and predictors are well known to be vulnerable in such a setting, and not only when the model validity is questionable.

By way of an example, consider a rural district distant from any district with an industrial centre that has many opportunities for education and employment. This district has a socio-demographic and economic profile very different from districts that contain or are part of populous cities or are adjacent to them. Some aspects of these differences can be captured

by suitably defined background variables. However, when there are not many such districts, a suitable model may include only a few variables that contrast urban with rural, high-tech services with subsistence farming, low with high percentage of residents in labour-force age and with tertiary education, and the like. The few such variables, however carefully selected, may not fully account for these differences. Data-based model diagnostics may not identify this issue because the data at hand contain no information about such districts.

This problem can be addressed, in principle, by dropping the largest districts from estimation for the smallest districts. Whilst the largest districts provide abundant information, their removal makes extrapolation less acute and the retained information relatively more relevant. This may make inference about small districts more efficient. More data is not always accompanied by more efficient estimation.

Sampling design has a much greater potential to deal with this problem, but it is predicated on awareness of the demand for small-area statistics at the planning stage. Concern about increasing costs of the survey is justified only partly because the numbers of observations added for the small districts, where they would contribute to estimation nontrivially, could be compensated by reduced sample sizes in the largest districts, for which the standard errors would be inflated only slightly. See Longford (2006) for a detailed exploration of this issue. A flip side of the presented argument is that survey costs may be higher in regions with low population density than in urban areas, although population registers and ubiquitous modes of telecommunication have greatly reduced these disparities.

6. Illustrative simulations

In this section we devise and explore schemes for setting the shrinkage factor λ for which an optimal trade-off of precisions in estimating the districts' means $\bar{Y}_k$ is reached. We generated a synthetic country with $K = 100$ districts, with sample sizes n_k generated by rounding a random sample from the Gamma distribution with shape 5.0 and rate 1/7 (expectation 35 and variance 245) and doubling the largest value, to represent the capital of the country. The districts' means $\bar{Y}_k$ are generated first as a random sample from chi-squared distribution with 10 degrees of freedom and the districts' standard deviations (σ_k) as a random sample from chi-squared distribution with 5 degrees of freedom. Then they are adjusted as follows: the means are increased by 0.04-multiple of the sample size and the standard deviations by the 0.04-multiple of the means (after the adjustment) and one-sixtieth of the sample size. These adjustments are applied so that the area-level means, standard deviations and sample sizes would be moderately correlated. The population sizes of the areas are $N_k = 100n_k$ and the values of Y in area k are generated as random samples from the normal distribution with mean $\bar{Y}_k$ and variance σ_k^2. Finally, the values of $\bar{Y}_k$ and σ_k^2 are adjusted to agree with the finite-population mean and variance of Y in area k. The populations of all areas are generated only once and are kept fixed, until stated otherwise.

In the realized country, $10 \le n_k \le 192$, $4.21 \le \bar{Y}_k \le 31.96$ and $3.46 \le \sigma_k^2 \le 30.79$. The means have correlation 0.30 with the subsample size and 0.31 with the variance. The variance and subsample size have correlation 0.41. The capital district has the largest sample size, 192, followed by a district with $n_k = 88$. The capital has the fourth largest mean, 21.45,

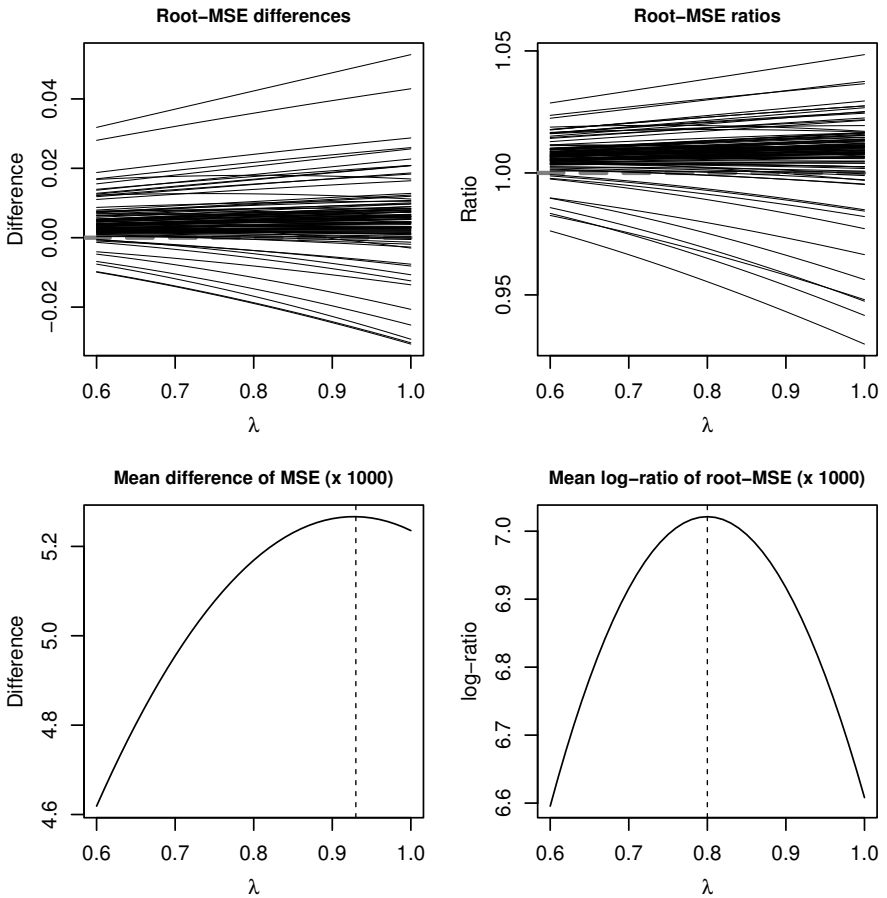


Figure 2. Panels A and B: differences and ratios of $\sqrt{\text{var}_s(\bar{y}_k)}$ and root-MSE$_s(\hat{\eta}_k)$ as functions of λ for the 100 simulated districts of a country. Panels C and D: average differences and average log-ratios of root-MSEs of $\bar{y}_k$ and $\hat{\eta}_k$.

and the largest standard deviation, 30.79. A simple random sample with specified sample size is drawn in each area k, and these samples are mutually independent.

We evaluate the MSEs under the sampling paradigm for the shrinkage estimator $\hat{\eta}_k(\lambda)$, $0.6 \leq \lambda \leq 1$. At the top of Figure 2 we plot the differences $\sigma_k/\sqrt{n_k} - \sqrt{\text{MSE}_s\{\hat{\eta}_k(\lambda)\}}$ in panel A and ratios $\sigma_k/\sqrt{n_k}/\sqrt{\text{MSE}_s\{\hat{\eta}_k(\lambda)\}}$ in panel B. For $\lambda = 1$, the sample mean $\bar{y}_k$ is more efficient than the shrinkage estimator $\hat{\eta}_k$ for 16 of the 100 districts. Panels A and B show that as we reduce λ, the differences of the two MSEs converge toward zero and their ratios to unity. The lines for six of the 16 extreme districts cross the zero-difference (and unity-ratio) line; for these districts, $\lambda_{2k} \in (0.6, 1)$ and $\hat{\eta}_k(\lambda)$ is more efficient than $\bar{y}_k$ for $\lambda < \lambda_{2k}$ and less efficient for $\lambda > \lambda_{2k}$.

By reducing λ the dispersion of the root-MSE differences and ratios is reduced — the largest gains in efficiency of $\hat{\eta}_k(1)$ are reduced, but so are the inefficiencies for the extreme districts. How should λ be chosen? This question has no simple answer even in our oracle setting of known subpopulation means $\bar{Y}_k$ and variances σ_k^2. If we regard each difference of root-MSEs as equally important, then we would choose λ for which the average difference is largest. Even though the differences are positive and increasing with λ for most districts, the average difference decreases as λ approaches 1.0. This average is largest for $\lambda = 0.93$, equal to 5.267; see panel C of Figure 2. It is equal to 5.235 at $\lambda = 1$ and to 5.262 at $\lambda = 0.90$. Erring on the side of smaller λ seems to be less consequential than setting λ too large.

An alternative way of comparing the gains is by averaging the ratios of the root-MSEs (or MSEs). We do this on the log scale, evaluating the average of the log-ratios of the root-MSEs. The largest mean of the log-ratios of the root-MSEs is attained for $\lambda = 0.80$, see panel D. The means are a nearly symmetric function of λ.

In summary, this simulation provides an example of suboptimality of BLUP in a congenial setting in which the ratio $\rho = \tau^2/\sigma^2$ and the overall mean β are known. Less shrinkage than applied by BLUP is optimal by both criteria we use. However, the choice of the criterion to use should not be taken lightly. The trade-off implied by the coefficient λ, or another scheme for reducing the shrinkage, should be carefully weighed.

This example is meant for a different (real) country, from which we would use features in generating a synthetic country and its districts. To establish to what extent the example entails uncertainty owing to random draws, we replicate the process of generating a country and its districts and evaluate $\mathrm{var}_s(\bar{y}_k)$ and $\mathrm{MSE}_s\{\hat{\eta}_k(\lambda)\}$ for $\lambda \in (0.4, 1)$. We use 1000 replications. The top panels of Figure 3 display the average differences and log-ratios of the two root-MSEs as functions of λ for a random sample of 100 of the realized replications. Panels C and D display histograms of the optimal values of λ, which are the maxima of the functions from which a random sample is drawn in respective panels A and B.

The diagram shows that the difference and log-ratio functions are widely dispersed and for a few replicates $\lambda = 1$ is optimal. This is confirmed in the panels at the bottom, which show that $\lambda = 1$ is optimal for differences in about 5% of replicates and for log-ratios in about 2%. The average of the optimal values of λ is 0.77 (standard deviation 0.12) for differences and 0.74 (standard deviation 0.11) for log-ratios. Thus, less shrinkage should be applied than by BLUP, but there is considerable uncertainty as to how much less, as controlled by the parameter λ. And the scale on which we compare MSE summaries also matters.

6.1. Focus on subpopulation totals. Purposes

So far, we were concerned about estimating the within-district means, regarding each district and its estimator as equally important. Such a position is not tenable when the district-level totals are also estimated because the estimators for larger districts may naturally be regarded as more important. The established BLUP or empirical BLUP is not responsive to the concept of trading efficiency among the districts; its recipe is to multiply each district's estimator of the mean by the population size (or its estimate).

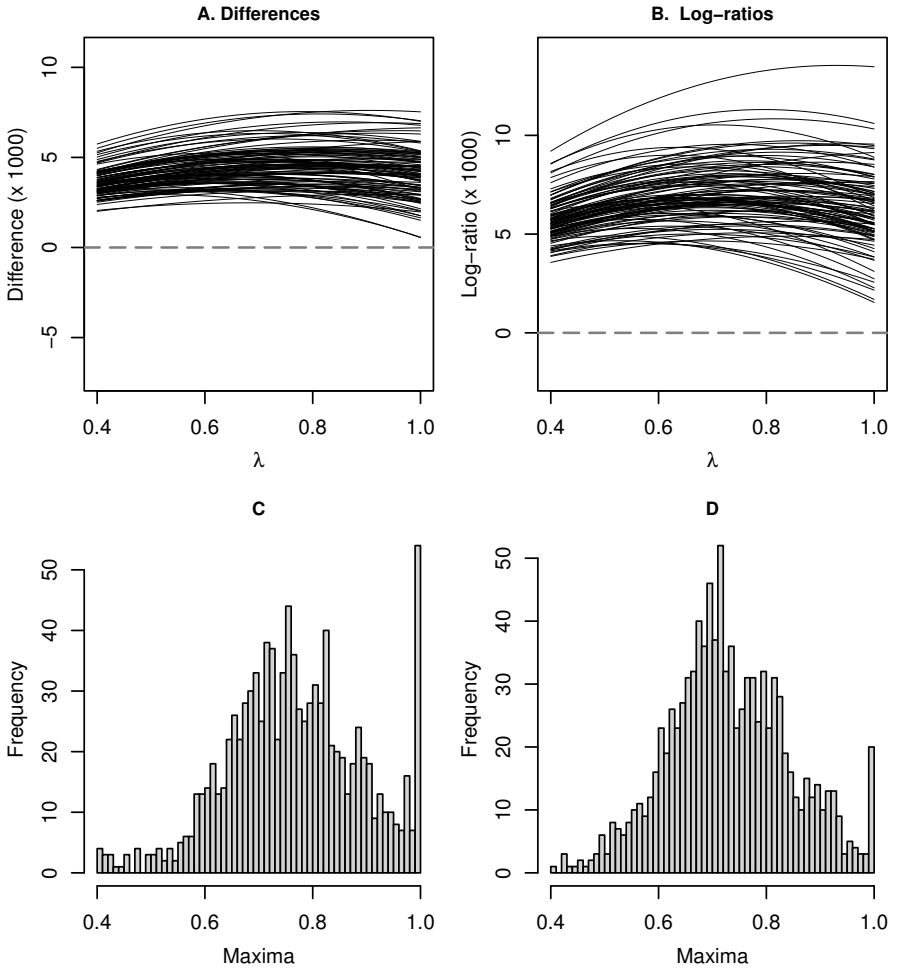


Figure 3. Panels A and B: replicate difference and log-ratio functions (root-MSEs) of λ; 100 replicates. Panels C and D: histograms of the optimal values of λ for differences and log-ratios, based on 1000 replications.

We repeat the simulation study with the population totals as the targets, and seek the optimal coefficient λ for reduced shrinkage. We assume that the districts' population sizes are proportional to the sample sizes. The log-ratios of the MSEs (and root-MSEs) are the same for the estimators of the totals as they are for the estimators of the means. In contrast, the differences are affected by the population sizes; districts with large populations have large values, and therefore potentially large errors in estimation, but these errors are tempered by larger samples. Biases in estimation are also inflated in estimation of the total, and this presents an insidious problem when we do not know which districts may be extreme.

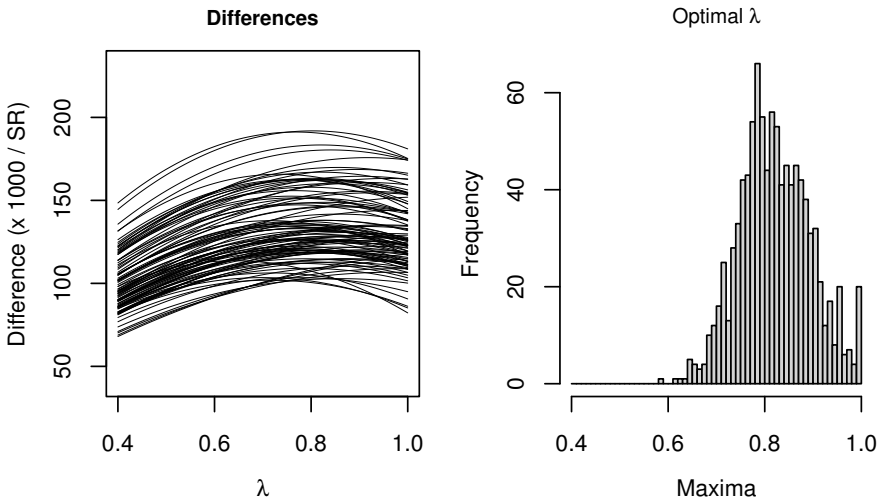


Figure 4. Replicate difference functions (root-MSEs) of λ (100 replicates) and histogram of the optimal values of λ for differences, based on 1000 replications.

Figure 4 presents the summaries for the differences of root-MSEs and the histogram of the optimal values of shrinkage reduction (λ). Note that we have to specify the population sizes only up to a constant multiplicative factor. In the diagram, we evaluate the MSEs for n_kSR, where SR is the sampling ratio, assumed to be the same for all districts. The average value of the optimal λ is 0.826, greater than for estimating the means by 0.055. The values of λ are dispersed much less than for estimating the means (standard deviation 0.074, compared to 0.124).

Estimation of the district-level means and totals is an example of *priorities*, or purposes, having a say in the choice of (small-area) estimators and, more generally, in estimating a set of related quantities. We have shown that such purposes are relevant for optimal estimation. Other common priorities are comparisons of selected pairs of districts (such as neighbours) and ranking of districts (forming league tables), discussed in the next section.

7. Comparing districts

When the role of districts is played by institutions, such as local authorities, schools, hospitals or providers of some other services, it is desirable to ensure equity of any comparisons by taking into account the background profile of their clients. In this setting, the target of inference is the quality of an institution, indicated by the outcomes of its clients.

Adjustment by regression is an established approach to such comparisons, leading naturally to the application of hierarchical linear models. Our discussion related to the paradigms has to be adapted. Now the qualities of the institutions, referred to as their *performance*, are fixed, given by their management, staffing, equipment and the like, but the clients and

their cases may well be regarded as random, being drawn from a population and distribution specific to the institution, which can be referred to as its *clientele*. The modelling paradigm remains problematic because its designation of what is random and what is fixed is opportunistic. In this section, we describe an approach motivated by causal analysis, that is largely faithful to the setting and purpose of the intended comparisons.

Two districts, k and h, may be compared by the estimate of the difference $\bar{Y}_k - \bar{Y}_h$. A criticism of such a comparison is that it is confounded by differences in the background profiles of the two districts. A more principled approach seeks an answer to the following question:

> What would be the average outcome of district k if it were occupied by the residents of district h?

It is referred to as indirect standardization. Its alternative, direct standardization, is based on a *template*, an artificial (synthetic) set of units defined by their background profile. We pose the question:

> What would be the average outcome of district k if it were occupied by the template?

These problems can be framed as an application of causal inference (Silber *et al.*, 2016, and Longford, 2018 and 2020). Direct standardization is fair — the districts are assessed by the same 'exam'. However, it is not relevant for districts that have a background profile substantially different from the template. In contrast, indirect standardization is relevant but not fair — districts are subjected to different exams. The results of these two approaches may differ. This highlights the crude nature of the comparison based on hierarchical models which have a single (superpopulation) target that caters neither for the nuances of the two kinds of standardization nor for the details of the template.

Direct standardization is suitable, in principle, for forming league tables, whereas indirect standardization is appropriate only for pairwise comparisons. However, these pairwise comparisons could be scored, akin to the games in a football league, and a league table formed according to the accumulated scores.

It is often desirable to conclude a comparison of two districts with a verdict of the 'winner' — which district has a greater mean. Hypothesis testing provides a motivation for three options: district k is 'better' than district h; district h is 'better' than district k; or the two districts differ insubstantially or not at all (failing to reject the alternative to the null hypothesis). We object to this practice on a multitude of grounds that constitute an abuse of hypothesis testing. Foremost, failure to reject the null hypothesis is appropriately characterized as a state of ignorance, not as a confirmation of the null hypothesis. Interpreting it as small or no difference is an error that is tolerated by many mainly because of its common occurrence. It confuses uncertainty with magnitude.

In an alternative approach, we specify two cutpoints, C_1 and C_2, that delineate, in terms of the reality, what is meant by the terms *Win*, *Loss* and *Draw*. If the target, a population-level difference, such as $D_{kh} = \bar{Y}_k - \bar{Y}_h$, is smaller than C_1, then the correct verdict is Loss; if $D_{kh} > C_2$, then the correct verdict is Win; otherwise, when $C_1 \leq D_{kh} \leq C_2$, it is Draw. In most settings, $C_1 = -C_2 < 0$. For each combination of reality and verdict, a rule based

on the estimated distribution of $\hat{D}_{kh}$, we specify the loss that would be incurred. These losses form the *loss matrix*. Its diagonal, which corresponds to the correct verdicts, contains zeros. We choose the verdict for which the expected loss is the smallest. A weakness of this approach would seem to be the specification of the thresholds C_1 and C_2. In the absence of experience with elicitation of these values, we have to resort to a sensitivity analysis. It comprises specification of a range of plausible pairs (C_1, C_2) and evaluation of the verdict for these pairs. It suffices to find the borderline, the value of C_1^* (and $C_2^* = -C_1^*$), for which two verdicts have the same expected loss. If this value of C_1^*, or C_2^*, lies within the plausible range (it is plausible), then we have an *impasse* because there is no one verdict that is optimal for all plausible settings. Otherwise, there is a unequivocal verdict. The entries of the loss matrix are treated similarly. See Longford (2020) for an application.

Other inferential targets include identifying outliers, the values of the extremes, position in the league table or the change in the position from past year to the current. In our approach, the starting point in these problems is the specification of the losses of the two kinds, false positive and false negative, or over-rating and under-rating.

When we seek a single district, say, the one with the largest mean, and a false positive is associated with a relatively small loss, it makes sense to list several candidates, to reduce the chances of an expensive false negative, even though we are certain to incur a positive loss.

8. Conclusion

Our derivations and illustrations show that the sampling design or, more generally, the data generating process, cannot be ignored in the assessment of small-area estimators. Another largely ignored aspect of the analysis is its purpose which we interpret as a preference for certain trade-offs of precisions among the areas. This has been ignored in the past in the mistaken belief that the commonly applied methods are efficient, or nearly so, for all the small areas simultaneously.

The main implication of this for practice is that small-area estimators should be assessed by design-based criteria. With no theoretical results available, this means using replicate datasets drawn by the sampling design or its approximation from a synthetic population. Using the data generating process implied by the adopted model, or referring to the analytical properties of the model and its estimators, leads to conclusions that are optimistic, especially about exceptional districts for which the estimators may have substantial bias.

Another implication is that no single set of estimators is optimal for a wide range of purposes. This conclusion echoes Shen and Louis (1998), who studied three specific purposes (goals) of small-area estimation within the model-based paradigm. Their conclusions apply also in the design-based paradigm. Even estimation of within-area means and totals of the same outcome variable are two different purposes because they may imply different optimal trade-offs of precision. Finding these trade-offs analytically is a challenge for future research. We explored reducing the shrinkage that is applied by the off-the-shelf methods. A practical challenge is to reorganize the production processes and outputs so that they would deliver sets of estimates and related quantities that are tailored to purposes declared a priori.

References

Box, G .E. P., (1976). Science and statistics. *Journal of the American Statistical Association*, 71, pp. 791–799.

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

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

Gao, P.A., Wakefield, J., (2024). Smoothed model-assisted small area estimation of proportions. *Canadian Journal of Statistics*, 52, pp. 337–358.

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

Hartley, H. O., Rao, J. N. K., (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93–108.

Heady, P., Ralphs, M., (2004). Some findings of the EURAREA project — and their implications for statistical policy. *Statistics in Transition*, 6, pp. 641–654.

James, W., Stein, C., (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 361–379, University of California Press, Berkeley.

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

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

Longford, N. T., (2018). Decision theory for comparing institutions. *Statistics in Medicine*, 37, pp. 437–456.

Longford, N. T., (2020). Performance assessment as an application of causal inference. *Journal of the Royal Statistical Society*, Series A, 184, 1363–1385.

Pfeffermann, D., (2013). New important developments in small area estimation. *Statistical Science*, 28, pp. 40–68.

Rao, J. N. K., Molina, I., (2015). *Small Area Estimation.* 2nd ed. Wiley, New York.

Robbins, H., (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 1954–1955, Vol. 1, 157–163. University of California Press, Berkeley.

Särndal, C.-E., Swensson, B., and Wretman, J., (2003). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

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

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., Hill, A. S., Even-Shoshan, O., Kelz, R. R. and Fleisher, L. A., (2016). Indirect standardization matching: assessing specific advantage and risk synergy. *Health Services Research*, 51, pp. 2330–2357.