# POLICY-RELATED SMALL-AREA ESTIMATION

***Nicholas T. Longford***
SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, c/ Ramon Trias Fargas 25 – 27, 08005 Barcelona, Spain.
e-mail: *sntlnick@sntl.co.uk*

***Key words:*** Borrowing strength, composition, empirical Bayes, expected loss, exploiting similarity, small-area estimation, utility function.

***Summary:*** A method of small-area estimation with a utility function is developed. The utility characterises a policy planned to be implemented in each area, based on the area's estimate of a key quantity. It is shown by simulations that the commonly applied composite and empirical Bayes estimators are inefficient for a wide range of asymmetric utility functions. An argument is presented for a closer integration of estimation and (regional) policy making because no single small-area estimator is suitable for a wide range of purposes.

## 1. Introduction

Recent developments in small-area estimation (SAe) respond to the increasing demand for information about the divisions (districts or areas) of a country. The key methodological advance in SAe is borrowing strength (Robbins, 1955; Efron and Morris, 1972; Fay and Herriot, 1979; Ghosh and Rao, 1994), that is, exploiting the similarity of the areas, possibly after taking into account relevant auxiliary information. The goal of a typical application of SAe is to estimate a quantity associated with each area efficiently, with minimum mean squared error (MSE), and to estimate this MSE, preferably without bias (Hall and Maiti, 2006; Slud and Maiti, 2006).

When implementing a policy in the areas of a country, estimates of area-level quantities are usually treated as if they were the underlying quantities, sometimes with only cursory attention to their estimated standard errors or confidence intervals. Problems arise when the estimates are subjected to nonlinear or even discontinuous transformations, such as ranking and comparisons with a set threshold, because efficiency is not retained by such transformations (Shen and Louis, 1998; Longford, 2005b).

In this paper, we study the following problem. A national government department wishes to apply a particular course of action in every district $m$ in which the unemployment rate $\theta_m$ exceeds 20%. This threshold, $T = 0.20$, was set in a consultation of the department with the country's trade unions and the principal employers' organisation. Based on a set of recent estimates $\hat{\theta}_m$ of $\theta_m$, $m = 1, \ldots, M$, the measure is to be applied in every district in which $\hat{\theta}_m > T$, in effect, regarding the

*AMS: 62C12, 62D99, 62-07*

estimate $\hat{\theta}_m$ as if it were the population rate $\theta_m$. We show that the established composite estimator (Longford, 1999), and by implication the empirical Bayes estimator (Ghosh and Rao, 1994; Rao, 2003), which aim to minimise the MSE, are not useful in this context, and propose alternatives in which different shrinkage or adjustment is applied.

Our approach incorporates the losses (negative utilities) that quantify the consequences of inappropriate actions. It reflects the view that the ultimate role of statistics is to contribute to making intelligent decisions (in the presence of uncertainty), and inferential statements, such as estimates of the relevant quantities, or the outcomes of hypothesis tests about them, are sometimes irrelevant in this effort. Estimation and decision making have to be closely integrated. These views are influenced by DeGroot (1970) and Lindley (1985), although we do not subscribe to the Bayesian paradigm.

The utilities are elicited from the policy maker (the expert, or sponsor of the analysis) in the form of loss functions. Suppose applying the intended measure in a district with rate $\theta_m < T$, for which estimation yielded $\hat{\theta}_m > T$, that is, a false positive, is associated with loss equal to $(\hat{\theta}_m - \theta_m)^2$, and failure to apply it in a 'deserving' district (a false negative), with rate $\theta_m > T$, but for which $\hat{\theta}_m < T$, is associated with loss equal to $R(\hat{\theta}_m - \theta_m)^2$, where $R \geq 1$ is a constant called the penalty ratio. Estimation of $\theta_m$ with minimum expected loss is sought. This loss function differs from the squared error loss even for $R = 1$, because positive loss is incurred only when $\hat{\theta}_m < T \leq \theta_m$ or $\hat{\theta}_m > T \geq \theta_m$. The same threshold $T$ applies to all districts, but the development we consider is not restricted to this case, although the threshold(s) have to be known.

We show that the empirical Bayes (EB) and the related composite estimators are suboptimal solutions for this problem — the expected loss with them is higher than with some other estimators. We search for alternatives among estimators of the form

$$\tilde{\theta}_m = (1 - b_m)\hat{\theta}_m^{(\mathrm{S})} + b_m F_m \,, \tag{1}$$

where $\hat{\theta}_m^{(\mathrm{S})}$ is a direct (unbiased) estimator of $\theta_m$, which uses information only from the target district $m$ and the variable concerned; $b_m$ and $F_m$ are constants called the shrinkage coefficient and the focus of shrinkage, respectively. We assume that the sampling variances $v_m = \mathrm{var}(\hat{\theta}_m^{(\mathrm{S})})$ are known. The product $b_m F_m$ could be replaced by a single term, but we prefer the expression in equation (1) because its form, with $F_m = \theta$ or $\hat{\theta}$, where $\theta = (\theta_1 + \theta_2 + \cdots + \theta_M)/M$ and $\hat{\theta}$ is its estimator, is related to the EB estimator for normally distributed outcomes when no covariates are available.

Each district-level quantity $\theta_m$ is regarded as fixed, because it is associated with a labelled and well-identified area. As in the established sampling paradigm, the value of the outcome variable is fixed for every member of the population, and so is the division of the country to its districts. The sample selection is the sole source of variation; see Longford (2005a, Chapter 6, and 2007) for related discussion. For the targets $\theta_m$, we consider their mean $\theta$ and the (district-level) variance

$$\sigma_{\mathrm{B}}^2 = \frac{1}{M}\sum_{m=1}^{M} (\theta_m - \theta)^2 \,.$$

The covariance and correlation of two sets of district-level quantities are defined similarly. The variance or a covariance is estimated by moment matching, adjusting its naive estimator for its bias. For example,

$$\hat{\sigma}_{\mathrm{B}}^2 = \frac{1}{M}\sum_{m=1}^{M} \left(\hat{\theta}_m^{(\mathrm{S})} - \hat{\theta}_m\right)^2 - v - \frac{1}{M}\sum_{m=1}^{M} (v_m - 2c_m) \,,$$

where $c_m = \mathrm{cov}(\hat{\theta}_m^{(\mathrm{S})}, \hat{\theta})$ and $v = \mathrm{var}(\hat{\theta})$. The expectations in the definitions of $c_m$ and $v$ and elsewhere in the paper are taken over the sampling design. We use the term 'averaging' for replacing expressions involving $\theta_m$ for a specific $m$ by their averages over the districts, while holding other district-level quantities, such as $F_m$ and $v_m$, fixed. For example, by averaging $(F_m - \theta_m)^2$ we obtain $(F_m - \theta)^2 + \sigma_{\mathrm{B}}^2$.

The sole restriction that we impose on the sampling design is that the estimators $\hat{\theta}_m^{(\mathrm{S})}$ are independent. Stratified sampling with the districts or their subsets as the strata satisfy this condition. To avoid complexities that would dilute our focus, we assume that $\hat{\theta}_m^{(\mathrm{S})}$ are linear functions of the data and $\hat{\theta}$ is a linear combination of $\hat{\theta}_1^{(\mathrm{S})}, \ldots, \hat{\theta}_M^{(\mathrm{S})}$.

The next section presents the key concepts and Section 3 derives an estimator which, setting aside some approximations and estimation of $\sigma_{\mathrm{B}}^2$, has smaller expected loss than the established alternatives. Simulations in Section 4 confirm the anticipated properties of the new estimator. Section 5 extends the method to incorporating auxiliary information. The paper is concluded with a discussion.

## 2. Policy and utility

Suppose a policy calls for one of two courses of action; action A is appropriate for district $m$ if $\theta_m > T$ and action B is appropriate otherwise; the threshold $T$ is given. The *loss function* for action $d = \mathrm{A}$ or B is defined as a non-negative function $L_d(\hat{\theta}_m, \theta_m)$ of the estimate used and its target. The appropriate action results in no loss. A pair of loss functions can be expressed as a single function $L = L_{\mathrm{A}} + L_{\mathrm{B}}$ after defining $L_d = 0$ whenever action $d$ is not taken. Function $L$ is associated with the class of equivalence defined by the functions $CL$, where $C$ is an arbitrary positive constant.

The loss functions $L_{\mathrm{A}}$ and $L_{\mathrm{B}}$ are elicited from the policy maker. Instead of a single pair of functions (or classes of equivalence) we use a set (range) of plausible pairs of loss functions, one for action A and the other for B in each pair. We assume that there is an ideal pair of loss functions, and that it is one of plausible loss functions, but it cannot be identified. See Longford (2010) for a similar way of dealing with uncertainty about the (Bayes) prior and Garthwaite, Kadane and O'Hagan (2005) for a review of statistical issues in elicitation. The elicited set should be as small as possible but the policy maker has to be satisfied that all loss functions outside this set can be ruled out.

The quadratic kernel loss is a special case of power kernel loss defined as

$$\begin{aligned} L_{\mathrm{A}}(\hat{\theta}_m, \theta_m) &= \left|\hat{\theta}_m - \theta_m\right|^h \\ L_{\mathrm{B}}(\hat{\theta}_m, \theta_m) &= R\left|\hat{\theta}_m - \theta_m\right|^h, \end{aligned}$$

when $\theta_m < T < \hat{\theta}_m$ and $\hat{\theta}_m < T < \theta_m$, respectively, and as zero otherwise; $R > 0$ is the penalty ratio and $h > 0$. Only $h = 0$ (absolute kernel), $h = 1$ (linear kernel) and $h = 2$ are relevant in practice. Loss functions involving $|\hat{\theta}_m - T|$ are not suitable because the trivial estimator $\hat{\theta}_m \equiv T$ would then be optimal. When the loss depends on the magnitude of the error, $|\hat{\theta}_m - \theta_m|$, absolute kernel has little to recommend.

Other loss functions can be defined, but power kernels are relatively easy to handle. Different penalty ratios and even different kernels may be defined for distinct subsets of districts. The functions $L_{\mathrm{A}}$ and $L_{\mathrm{B}}$ do not have to be in the same class (e.g., both quadratic). Also, a few districts (a

region or the capital) may be singled out for an exceptional treatment, and the constants involved ($R$ and $T$) may be district-specific. For instance, $R_m$ may be a (linear) function of the population size of the district. In any case, the development in the next section is focused on a single district.

# 3. Policy-related estimator

The sampling distribution of the estimator $\tilde{\theta}_m$ given by equation (1) is normal, $\mathcal{N}(\gamma_m, \nu_m^2)$, with

$$\begin{aligned} (\gamma_m =) \quad \mathrm{E}\left(\tilde{\theta}_m \,|\, \theta_m\right) &= (1-b_m)\,\theta_m + b_m F_m \\ (\nu_m^2 =) \quad \mathrm{var}\left(\tilde{\theta}_m \,|\, \theta_m\right) &= (1-b_m)^2 \nu_m\,. \end{aligned}$$

We regard $\theta_m$ as fixed (related to a labelled district), unlike in the usual treatment of (exchangeable) districts in EB analysis (Ghosh and Rao, 1994; Rao, 2003). We do not assume that $\gamma_m = \theta_m$. Denote by $\phi$ the density of $\mathcal{N}(0,1)$ and by $\Phi$ its distribution function. With the quadratic kernel, the expected loss with the policy applied to district $m$ according to estimator $\tilde{\theta}_m$ is

$$\begin{aligned} (E_{\mathrm{A}} =) \quad \mathrm{E}\left\{L_{\mathrm{A}}\left(\tilde{\theta}_m, \theta_m\right)\right\} &= \frac{1}{\nu_m}\int_T^{+\infty} (y-\theta_m)^2\, \phi\left(\frac{y-\gamma_m}{\nu_m}\right) \mathrm{d}y \\ (E_{\mathrm{B}} =) \quad \mathrm{E}\left\{L_{\mathrm{B}}\left(\tilde{\theta}_m, \theta_m\right)\right\} &= \frac{R}{\nu_m}\int_{-\infty}^{T} (y-\theta_m)^2\, \phi\left(\frac{y-\gamma_m}{\nu_m}\right) \mathrm{d}y\,, \end{aligned}$$

if $\theta_m < T$ and $\theta_m > T$, respectively. Simple operations yield the identities

$$\begin{aligned} E_{\mathrm{A}} &= \nu_m^2 \left\{\left(1+z_\dagger^2\right)\Phi(\tilde{z}) + (2z_\dagger - \tilde{z})\,\phi(\tilde{z})\right\} \\ E_{\mathrm{B}} &= R\nu_m^2 \left[\left(1+z_\dagger^2\right)\left\{1-\Phi(\tilde{z})\right\} - (2z_\dagger - \tilde{z})\,\phi(\tilde{z})\right]\,, \end{aligned}$$

where $\tilde{z} = (\gamma_m - T)/\nu_m$ and $z_\dagger = (\gamma_m - \theta_m)/\nu_m$. We do not aspire to minimise $\min(E_{\mathrm{A}}, E_{\mathrm{B}})$ as a function of $v_m$ and $F_m$ directly, but seek estimators $\tilde{\theta}_m$ which have the following two properties:

- *equilibrium condition* — if district $m$ had $\theta_m = T$, the choice between actions A and B would be immaterial in expectation: $\mathrm{E}\left\{L_{\mathrm{A}}\left(\tilde{\theta}_m, T\right)\right\} = \mathrm{E}\left\{L_{\mathrm{B}}\left(\tilde{\theta}_m, T\right)\right\}$;
- *minimum averaged* MSE (aMSE).

Averaging in the second condition is similar to the step taken in EB analysis, where $\theta_m$ is regarded as random and $\sigma_{\mathrm{B}}^2$ is its district-level variance. Without taking aMSE the problem of minimising the expected loss is not tractable.

For quadratic kernel loss, the equilibrium condition, when $z_\dagger = \tilde{z}$, is equivalent to

$$(R+1)\left\{\left(1+\tilde{z}^2\right)\Phi(\tilde{z}) + \tilde{z}\,\phi(\tilde{z})\right\} - R\left(1+\tilde{z}^2\right) = 0\,. \tag{2}$$

It can be shown by simple calculus that the left-hand side, called the equilibrium function, has a single root. The root, denoted by $z^*$, is found by the Newton method.

The minimum of aMSE of $\tilde{\theta}_m$, equal to $(1-b_m)^2 v_m + b_m^2\{\sigma_B^2 + (F_m-\theta)^2\}$, is attained for

$$b_m^* = \frac{v_m}{v_m + \sigma_B^2 + (F_m-\theta)^2}; \tag{3}$$

if we ignore the equilibrium condition, the shrinkage coefficient is always within the range $(0,1)$. The composite estimator is obtained by setting $F_m = \hat{\theta}$, minimising $\mathrm{aMSE}(\tilde{\theta}_m; \theta_m)$ and substituting an estimate for $\sigma_B^2$ in equation (3). This estimator, referred to as estimator C, is given by equation (1) with $F_m = \hat{\theta}$ and

$$b_m = \frac{v_m - c_m}{v_m + v - 2c_m + \hat{\sigma}_B^2}.$$

With $v$ and $c_m$ omitted, this estimator differs from the EB estimator only by how $\sigma_B^2$ is estimated. Omission of $v$ and $c_m$ introduces a negligible error for all districts except one or two for which $v_m$ is not substantially greater than $v$. Usually, the subsample for such a district is a large fraction (20% or more) of the overall sample size.

The equilibrium condition implies that

$$F_m = T + \frac{|1-b_m|}{b_m} z^* \sqrt{v_m}. \tag{4}$$

The aMSE with this constraint is equal to

$$(1-b_m)^2 \left(1 + z^{*2}\right) v_m + b_m^2 \sigma_B^2 + b_m^2 (T-\theta)^2 + 2b_m |1-b_m| (T-\theta) z^* \sqrt{v_m},$$

and the coefficient that minimises this function of $b_m$ has to satisfy the identity

$$b_m = \frac{v_m \left(1 + z^{*2}\right) - \mathrm{sign}(1-b_m)(T-\theta) z^* \sqrt{v_m}}{v_m + \sigma_B^2 + \left\{z^* \sqrt{v_m} - \mathrm{sign}(1-b_m)(T-\theta)\right\}^2}. \tag{5}$$

The aMSE is continuous and diverges to $+\infty$ for $b_m \to \pm\infty$. Further, equation (5) implies that it cannot have more than two extremes. Hence it has a unique minimum, and it is its only extreme. The corresponding estimator is denoted by $\tilde{\theta}_m^{(\mathrm{P})}$ and referred to as estimator P. The solution $b_m^*$ may be outside $(0,1)$, and then it does not have the common interpretation as a shrinkage coefficient. It exceeds unity when

$$(\theta - T) z^* \sqrt{v_m} > \frac{\sigma_B^2 + (\theta - T)^2}{3},$$

that is, for sufficiently large $v_m$ when $T < \theta$. It is negative when $\sqrt{v_m} < \dfrac{z^*}{1+z^{*2}} (T-\theta)$, that is, for sufficiently small $v_m$ when $T > \theta$. However, $b_m^*$ is not a monotone function of $v_m$. Minimum expected loss is our sole criterion, and we pay no regard to the desirability of an interpretation of the estimators we use. Truncating $b_m^*$ at zero and unity would lead to an increase of both aMSE and the expected loss of the estimator.

No shrinkage, $b_m^* = 0$, is applied not only when $\theta_m$ is known, but also when $\sqrt{v_m} = (T - \theta) z^*/(1+z^{*2})$. For $v_m \to +\infty$, $b_m^* \to 1$ and $F_m \to T$; when we have no information about $\theta_m$, $\tilde{\theta}_m = T$ is optimal, unlike in EB estimation, where $\tilde{\theta}_m = \hat{\theta}$ in such a case. The focus $F_m$ in equation (4) is not defined when $b_m^* = 0$. However, the product $b_m^* F_m$ is then well defined by its limit, equal to $z^* \sqrt{v_m}$.

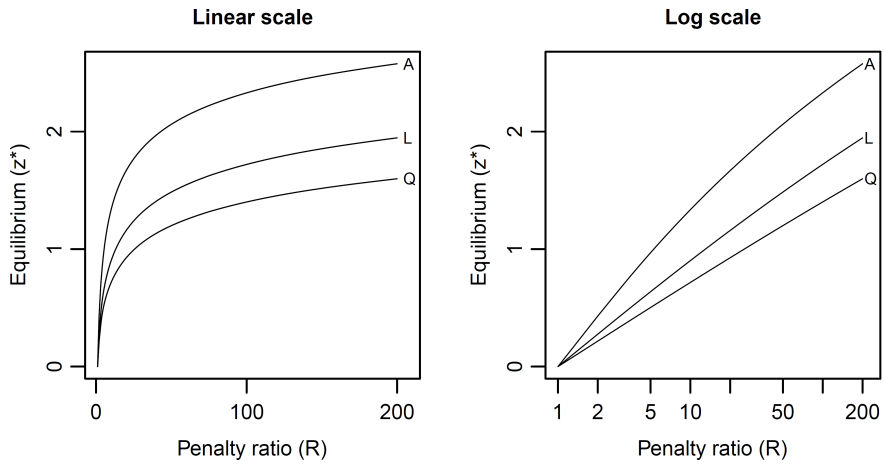


**Figure 1**: The roots of the equilibrium equations, $z^*$, as functions of the penalty ratio $R$ for the absolute (A), linear (L) and quadratic (Q) kernel loss functions, on the linear and log scales for $R$.

For the absolute and linear kernel loss functions, we have the respective equilibrium conditions

$$\begin{aligned} \Phi(\tilde{z}) &= \frac{R}{R+1} \\ (R+1)\{\tilde{z}\Phi(\tilde{z})+\phi(\tilde{z})\} &= R\tilde{z}. \end{aligned} \tag{6}$$

The latter, solved by the Newton method, has a unique solution for each $R > 0$. The equilibrium values $z^*$ as functions of $R$ are drawn in Figure 1 for the three kernels. No generality is lost by assuming that $R \geq 1$, because we could work with the outcomes $-y$, estimators $-\tilde{\theta}_m$ and $-\tilde{\theta}$, and penalty ratio $1/R$. For $R > 0$ and G = A, L or Q, $z > z^*_{\mathrm{G}}(R)$ corresponds to action A and $z < z^*_{\mathrm{G}}(R)$ to action B being preferable.

The optimal coefficients $b^*_m$ and foci $F^*_m$ are drawn in Figure 2 as functions of the variance $v_m$ of the direct estimator ($1.0 \leq v_m \leq 2.5$) for the quadratic kernel loss and penalty ratios $10 < R < 100$. The mean of the district-level means is $\theta = 16\%$, the district-level variance is $\sigma^2_{\mathrm{B}} = 6.25$ ($\%^2$), and the threshold is set to $T = 20\%$. The EB shrinkage coefficient, $v_m/(v_m + \sigma^2_{\mathrm{B}})$, is drawn by dashes in the left-hand panel. In the right-hand panel, the horizontal dashes indicate its focus, $\theta = 16\%$. The diagram shows that radically different linear combinations of $\hat{\theta}^{(\mathrm{S})}_m$ and foci $F_m$ are optimal from those in EB estimation. The focus of shrinkage is smaller than $\theta$ and decreases with the variance $v_m$. However, the shrinkage is *negative*, away from these foci.

The equilibrium conditions given by equations (2) and (6) involve $\gamma_m$ and $v_m$ only through $\tilde{z}$. Not all common classes of loss functions have this property. For example, the exponential loss does not satisfy this condition; see Longford (2014, Appendix A2).

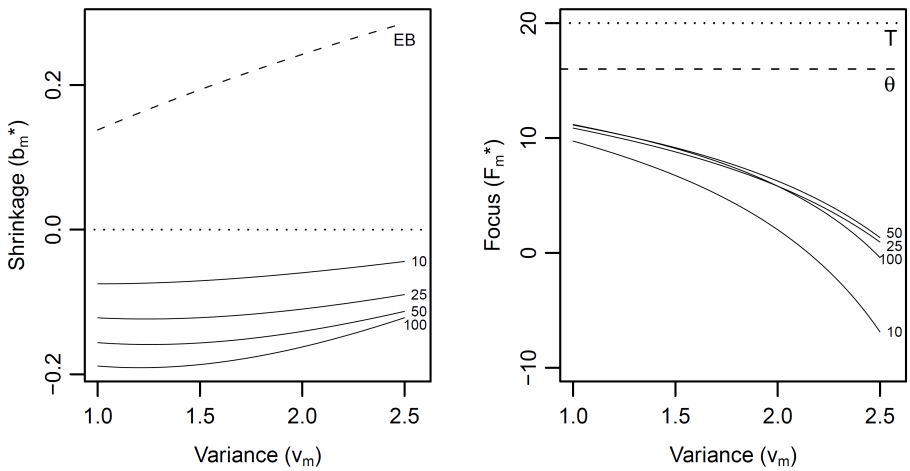


**Figure 2**: The optimal shrinkage coefficients and foci of shrinkage for quadratic kernel loss and penalty ratios $R$ = 10, 25, 50 and 100, indicated at the right margin; $\theta = 16\%$, $T = 20\%$ and $\sigma_B^2 = 6.25\%^2$. The coefficient and focus of the EB estimator is drawn by dashes (EB, $\theta$).

# 4. Simulations

We assess the properties of the estimators defined in Section 3 by simulations based on an imaginary country that comprises $M = 60$ districts with labour force sizes $N_m$ in the range 0.30 – 2.30 million, and the national total of 58.90 million. The focal variable is unemployment, a dichotomy, and the district-level (population) rates of unemployment $\theta_m$ are in the range 7.9 – 26.3%. These rates are weakly associated with the population size; more populous districts tend to have higher rates, although the most populous district (the capital), has an unemployment rate well below average. The 22 districts for which $\theta_m > T = 20\%$ account for 23.23 million members (39.4%) of the labour force. The population sizes and unemployment rates of the districts are plotted in Figure 3. The mean of the district-level unemployment rates is $\theta = 16.8\%$, and the national unemployment rate is $\theta^* = 17.3\%$. The variance of the district-level unemployment rates is $\sigma_B^2 = 27.05$ ($\%^2$).

Suppose a national survey is conducted, with a stratified sampling design using the districts as the strata, and simple random sampling design with a fixed sample size $n_m$ in district $m$. The overall sample size is $n = 17\,500$. The sample sizes $n_m$, indicated in Figure 3 by the size of the black disc, are in the range 113 – 567. They are approximately proportional to $N_m^{0.9}$. They are sufficiently large for approximate normality of all the sample rates $\hat{\theta}_m^{(S)}$. However, the sampling variances are far too large; composition (estimator C) yields substantial reduction of MSE for the smallest districts. We show that the naive classification of districts based on estimator S is associated with expected loss much greater than based on estimator P (shrinkage toward $\hat{F}_m^*$). The expected loss with estimator C is even higher than with estimator S. We assume the quadratic kernel loss with plausible penalty ratio in the range $(5, 20)$.

We replicate the processes of sampling (from a fixed population) and estimation 10 000 times and

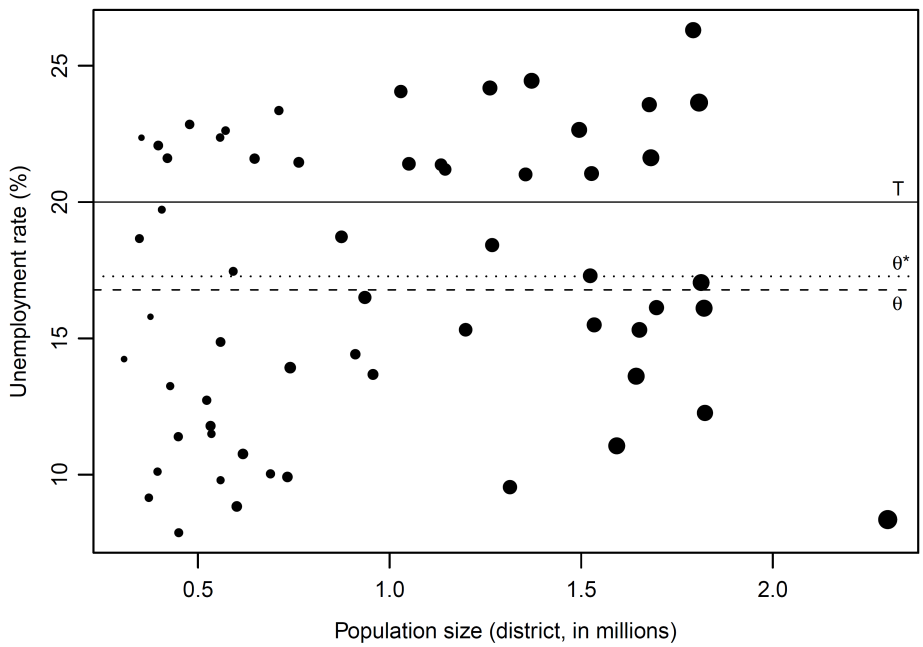


**Figure 3**: The population sizes and unemployment rates in the districts of a country. Computer-generated data used for simulation. The area of the black disc is proportional to the sample size of the district it represents.

accumulate the losses separately for each district and the three estimators. The empirical expected losses for the districts are displayed in Figure 4. They are marked by the symbols S, C and P for the three estimators. When the expected loss is smaller than 2.5, a black disc is displayed instead of the symbol. The population rates of unemployment in the districts are marked by horizontal dashes. The same scale happens to be suitable for the rates and the expected losses.

Most of the losses are incurred by false negatives, for districts with $\theta_m > T$, and among them the loss for every district is smallest for estimator P. We summarise an estimator by its weighted total expected loss, with weight equal to the size of the district's labour force (in millions). These totals are 439.2, 581.9 and 162.3 for the respective estimators S, C and P. The false positives contribute to these figures by only 19.6 (4.4%), 8.4 (1.4%) and 45.2 (27.9%), respectively. If we evaluated the losses with much smaller value of $R$ estimators S and C would remain far inferior. Estimators C and S are not sensitive to the penalty ratio and only estimator P has to be simulated again. The expected losses with C and S have the form $V + RU$, where $V$ is the expected loss for the false negatives and $U$ the expected loss for the false positives, pro-rated for unit penalty ($R = 1$). The weighted total losses for $R = 2.0$ have expectations 103.6, 123.1 and 65.4 for respective estimators S, C and P.

For $R = 10$, estimator P has the smallest expected loss in none of the eight districts with $\theta_m < T$ that have non-trivial expected losses. However, all these expected losses are much smaller than for most of the deserving districts. The shrinkage applied by the composition to minimise aMSE is

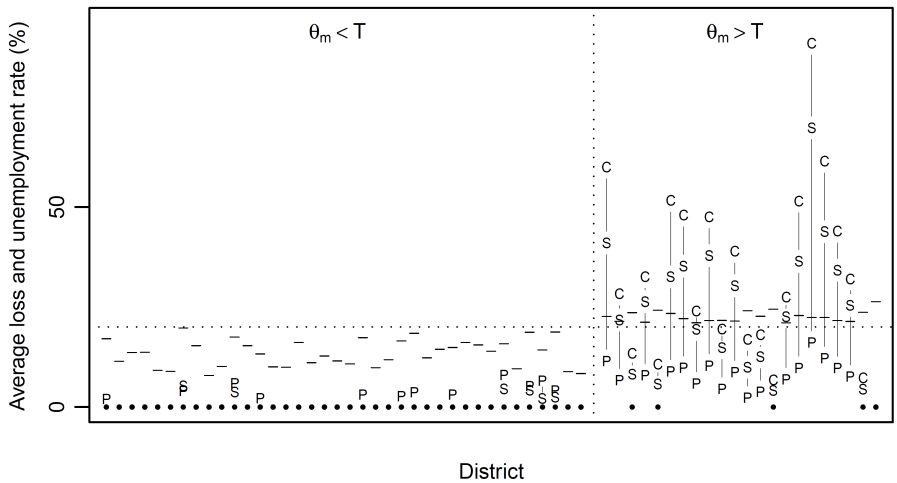


**Figure 4**: The empirical expected losses for the districts and estimators S (direct); C (composite) and P (policy-related), with penalty ratio $R = 10$. The districts are in the ascending order of labour force size, within the two groups divided by the threshold $T = 20\%$. The districts' unemployment rates are marked by horizontal ticks.

counterproductive, and a substantially smaller expected weighted total loss is obtained by estimator P. Small MSE and small expected loss with large penalty ratio $R$ are diametrically different criteria.

We repeated the simulations with $R = 5$ and $R = 20$ to confirm that estimator P remains superior to C and S. The results for penalty ratio $R = 5$ are summarised in Figure 5. They do not differ from the results for $R = 10$ substantially when the expected losses for the deserving districts are doubled. Similar conclusions are arrived at for $R = 20$. The expected losses are quite robust with respect to the specification of the penalty ratio $R$.

We conclude this section by the table of weighted totals of the (empirical) expected losses with the quadratic, linear and absolute kernel losses, displayed in Table 1. Estimator P has a distinct advantage over S and C for higher penalty ratios. For $R = 1$, its advantage is only slight for quadratic and linear kernels, and for the absolute kernel estimator S is preferable to both estimators P and C. The expected loss with estimator P increases with $R$ much slower, and estimators C and S are inferior for $R$ very close to 1.0 even with the absolute kernel loss. Even though absolute kernel loss and $R = 1$ are not a realistic combination of settings, the failure to outperform both estimators C and S suggests that there may be some scope for improvement of estimator P. Note that expected losses, or their totals, cannot be compared across the kernels, because they regard the relative losses with small and large deviations $|\hat{\theta}_m - \theta_m|$ differently.

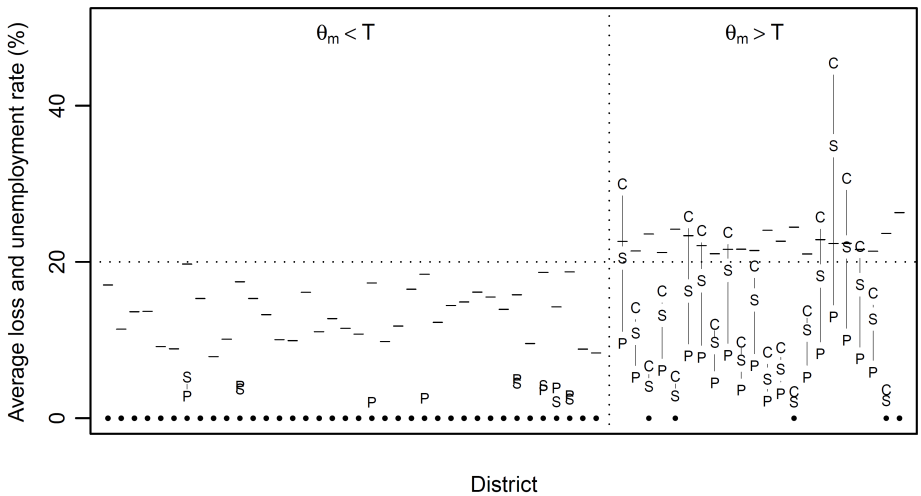


**Figure 5**: The empirical expected losses for the districts and estimators S, C and P, with penalty ratio $R = 5$. The districts are in the same order as in Figure 4.

# 5. Auxiliary information

We consider auxiliary information in the form of (column) vectors of district-level estimators or exact quantities $\hat{\boldsymbol{\xi}}_m$ for $\boldsymbol{\xi}_m$. We put no restrictions on $\boldsymbol{\xi}_m$, although summaries in $\boldsymbol{\xi}_m$ that are highly correlated with $\theta_m$ and elements of $\hat{\boldsymbol{\xi}}_m$ with small sampling variances are more useful. Common examples of elements of $\boldsymbol{\xi}_m$ are the direct estimates of the version of $\theta_m$ in the past year(s), values of a quantity prima facie closely related to $\theta_m$ obtained from an administrative register, and the values of the same summary as $\theta_m$ but estimated in a different subpopulation; see Longford (2005a, Chapter 10) for examples.

We assume that the estimators $\hat{\boldsymbol{\xi}}_m$ are unbiased for the respective $\boldsymbol{\xi}_m$. In practice, $\hat{\boldsymbol{\xi}}_m$ comprise direct estimators or exact quantities; for the latter components, $\hat{\xi}_m = \xi_m$. Denote $\boldsymbol{\theta}_m = (\theta_m, \boldsymbol{\xi}_m^\top)^\top$ and $\hat{\boldsymbol{\theta}}_m = (\hat{\theta}_m, \hat{\boldsymbol{\xi}}_m^\top)^\top$, and let $\mathbf{u} = (1,0,\ldots,0)^\top$ be the indicator of the first component, so that $\theta_m = \mathbf{u}^\top \boldsymbol{\theta}_m$. We define $\boldsymbol{\theta} = (\theta, \boldsymbol{\xi}^\top)^\top = (\boldsymbol{\theta}_1 + \cdots + \boldsymbol{\theta}_M)/M$ and $\hat{\boldsymbol{\theta}}$ as an unbiased estimator of $\boldsymbol{\theta}$, linear in each $\hat{\boldsymbol{\theta}}_m$. Let $\mathbf{V}_m = \mathrm{var}(\hat{\boldsymbol{\theta}}_m)$, $\mathbf{V} = \mathrm{var}(\hat{\boldsymbol{\theta}})$, $\mathbf{C}_m = \mathrm{cov}(\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\theta}})$ and $\boldsymbol{\Sigma}_\mathrm{B}$ be the respective multivariate versions of $v_m$, $v$, $c_m$ and $\sigma_\mathrm{B}^2$. For example,

$$\boldsymbol{\Sigma}_\mathrm{B} = \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\theta}_m - \boldsymbol{\theta})(\boldsymbol{\theta}_m - \boldsymbol{\theta})^\top .$$

The matrix $\mathbf{C}_m$ is a linear function of $\mathbf{V}_m$, and does not depend on $\mathbf{V}_{m'}$ for $m' \neq m$.

The multivariate composite estimator Longford (1999, and 2005b, Chapter 8) is defined as $\tilde{\theta}_m = (\mathbf{u} - \mathbf{b}_m)^\top \hat{\boldsymbol{\theta}}_m + \mathbf{b}_m^\top \hat{\boldsymbol{\theta}}$. The optimal vector of coefficients $\mathbf{b}_m$ is $\mathbf{b}_m^* = \mathbf{Q}_m^{-1} \mathbf{P}_m$, where $\mathbf{Q} = \mathbf{V}_m + \mathbf{V} + \boldsymbol{\Sigma}_\mathrm{B} - \mathbf{C}_m - \mathbf{C}_m^\top$ and $\mathbf{P} = \mathbf{V}_m - \mathbf{C}_m$. In practice, $\mathbf{Q}_m$, $\mathbf{P}_m$ and $\mathbf{b}_m$ are estimated, yielding the estimator $\tilde{\theta}_m = \tilde{\theta}_m(\hat{\mathbf{b}}_m)$. Univariate composition corresponds to empty $\boldsymbol{\xi}_m$ and scalar $\mathbf{u} = 1$. The variances in $\mathbf{V}$ are much smaller than in $\mathbf{V}_m$ for all $m$, unless one district's sample or population size is a large

**Table 1**: The expected total losses, weighted by the labour force size, in simulations of estimators S, C and P, with quadratic, linear and absolute kernels and penalty ratios $R = 1$, 5, 10 and 20. Based on 10 000 replications.

| | *Quadratic loss* | | | *Linear loss* | | | *Absolute loss* | | |
|---|---|---|---|---|---|---|---|---|---|
| $R$ | P | S | C | P | S | C | P | S | C |
| 1 | 58.3 | 61.6 | 65.8 | 15.1 | 15.9 | 18.3 | 6.0 | 5.0 | 6.0 |
| 5 | 123.7 | 229.4 | 295.1 | 32.8 | 60.6 | 81.9 | 8.9 | 19.3 | 26.8 |
| 10 | 162.3 | 439.2 | 581.9 | 41.0 | 116.5 | 161.3 | 10.4 | 37.2 | 52.8 |
| 20 | 207.4 | 858.8 | 1155.4 | 50.2 | 228.4 | 320.3 | 12.0 | 73.0 | 104.7 |

fraction of the entire sample in one or several surveys (data sources) on which $\hat{\boldsymbol{\theta}}_m$ are based. When there is no such dominant district the matrix $\mathbf{C}_m$ can also be ignored.

The multivariate policy-related composite estimator is defined by shrinkage toward a (multivariate) focus $\mathbf{F}_m$, with the intent to minimise the expected loss $\mathrm{E}\{L(\hat{\theta}_m, \theta_m)\}$:

$$\tilde{\theta}^*_m = (\mathbf{u} - \mathbf{b}_m)^\top \hat{\boldsymbol{\theta}}_m + \mathbf{b}_m^\top \mathbf{F}_m .$$

Details of the algorithm, based on a multivariate version of the equilibrium condition, are given in Longford (2014, Appendix A3).

## 5.1. Example continued

We simulate the setting of Section 4 with one auxiliary variable, the unemployment status in the previous year. We generate the district-level unemployment rates in the previous year by a scaled perturbation of the current rates and the districts' labour force sizes in the previous year by reducing the current year's sizes by a random percentage in the range 1.7 – 3.1%; the country's labour force increased during the year from 57.4 to 58.9 million. The districts' sample sizes in the past survey are generated by the same process as for the current survey (proportional to $N^{0.9}_{m,\mathrm{past}}$).

The district-level unemployment rates and sample sizes are plotted in Figure 6. The rectangles are centered at the districts' current and past unemployment rates and their sides are proportional to the sample sizes in the respective surveys. The two surveys, conducted in the current and the previous year, are independent. The four highlighted districts are discussed below.

The results of the simulation with 2000 replications, using quadratic kernel loss with penalty ratio $R = 10$, as in Figure 4, are summarised in Figure 7. The direct estimator (S) has the same distribution as in the simulation in Section 4, because it uses no auxiliary information. The bivariate composite estimator ($\mathrm{C}_2$) is associated with smaller expected losses than estimator S for most of the deserving districts. The reduction of aMSE, attributable to the auxiliary information, is accompanied by a substantial reduction of the expected losses for most of these districts. However, they still exceed the expected losses with the policy-related estimator, both the univariate version applied in Section

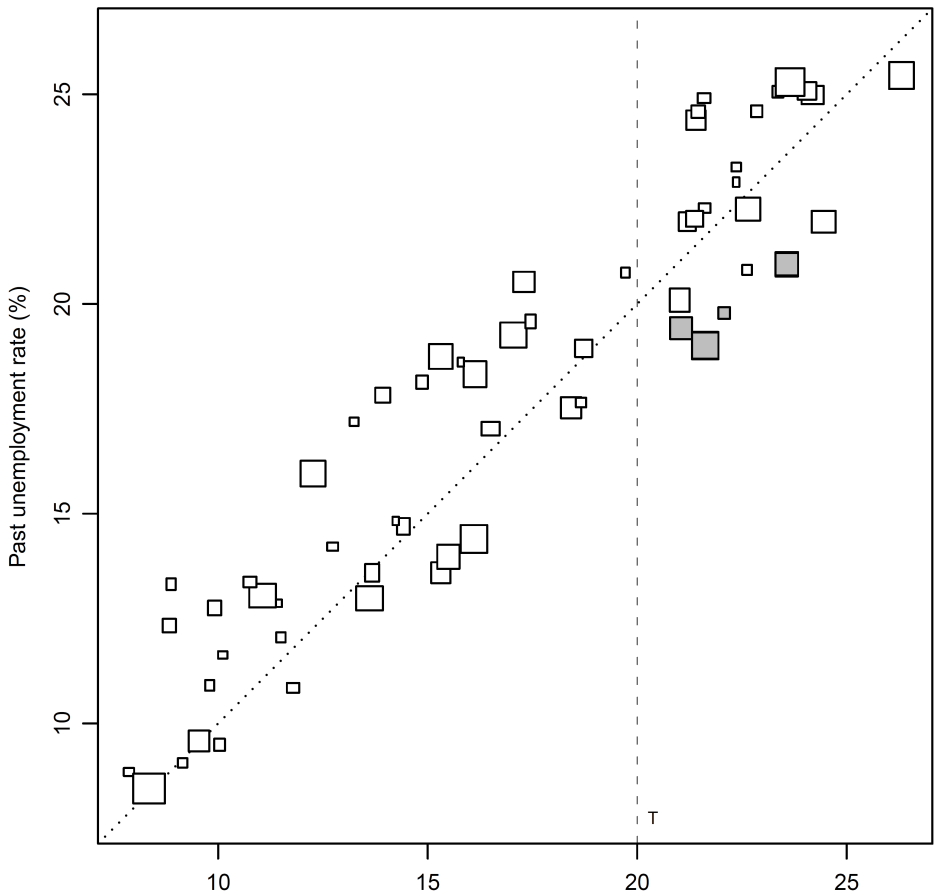


**Figure 6**: The district-level unemployment rates and sample sizes in the current and previous year. The sides of the rectangles are proportional to the sample sizes.

4, and the bivariate version ($P_2$), which exploits the auxiliary information. The weighted total of the expected losses is 436.9 (= 20.0 + 416.9) for estimator S, 400.0 (= 6.9 + 393.1) for $C_2$, and 123.9 (38.1 + 85.8) for $P_2$; the figures in parentheses are the contributions from the normal and deserving districts. For estimator $C_2$, the reduction attributable to the auxiliary information is 181.9 (31%). The reduction for $P_2$ over P, by 38.4 (24%), is more modest.

The expected loss of estimator $C_2$ is not uniformly smaller than for C. For the four deserving districts highlighted in Figure 6, auxiliary information brings about an increase of the expected loss. Their rates in the previous year are much lower than in the current year, even when related to the national trend. The auxiliary information is distracting, especially for the small district, for which substantial shrinkage takes place toward being a false negative. Auxiliary information is unhelpful also for a few normal districts. However, the inflation of their expected losses is only slight.

For linear and absolute kernels, estimator $P_2$ remains far superior to $C_2$ and S. With linear kernel loss and $R = 10$, the weighted total loss for $C_2$ is 124.5 (2.0+122.5), greater than for S, 116.1

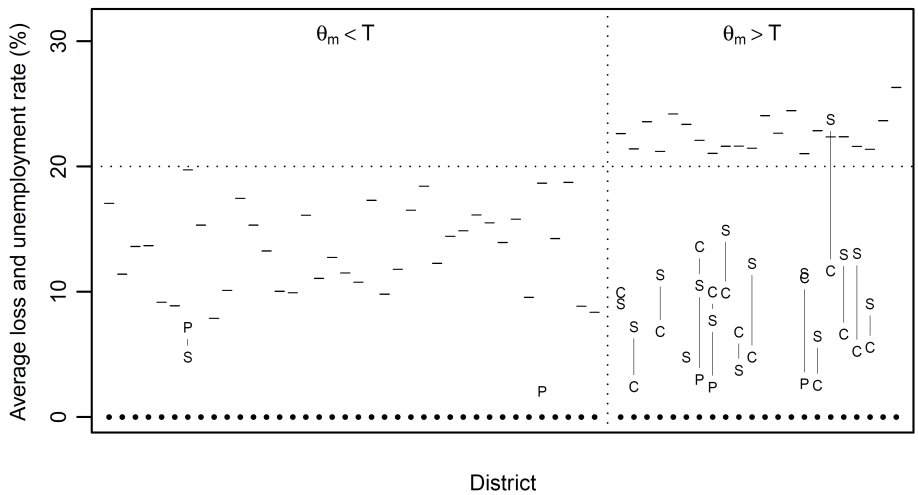


**Figure 7**: The empirical expected losses with the direct estimator (S), bivariate composite estimator $C_2$ (marked by C), using information from the previous year, and bivariate policy-related estimator $P_2$ (P); quadratic kernel loss and penalty ratio $R = 10$.

(4.8+111.3); for $P_2$ the loss is 40.4 (8.9+31.5). For more extensive auxiliary information, with several variables in $\boldsymbol{\xi}_m$, estimator C has slightly smaller empirical MSE and expected loss. Such information is detrimental for estimator P, but its weighted total expected loss remains much smaller than for estimator C.

# 6. Discussion

Simulations of the policy-related estimator developed in Sections 3 and 5 indicate that there is no single small-area estimator that is preferable to all others, because different estimators are optimal for different loss functions (policies or criteria). Shen and Louis (1998) highlight a related problem, that efficiency of small-area estimators is not retained by nonlinear transformations or summaries. Evaluation of small-area estimators has so far almost exclusively focused on MSE and aMSE. Alternatives to these criteria that reflect the objectives to be served by the analysis should be carefully considered. Elicitation of the loss function imposes some burden on the analyst and the client, but its outcome enables them to tailor the analysis closely to the needs, priorities and the perspective of the client. Instead of a single penalty ratio a plausible range can be defined, informed by the client's perspective and assessment of the damage, harm, additional expense or erosion of the intended effect caused by the inappropriate decision. As an alternative, the sets of decisions can be presented to the client for a wide range of penalty ratios, with the instruction to specify a much narrower range, so that an impasse, when both courses of action A and B are preferred for some of the plausible loss functions, would arise only for a few (or no) districts. The methods have a simple extension to $K > 2$ available courses of action; simply $K$ expected losses have to be compared.

The simulations confirm that composite (and EB) estimation is not conducive to good policy implementation when the loss function used differs radically from the (symmetric) quadratic loss. The policy-related estimator introduced in Section 3 is not the minimum expected loss estimator, because in its derivation we imposed the equilibrium condition, which has a flavour of unbiasedness, and then we minimised the (symmetric) averaged MSE instead of the expectation of the specified loss function. The class of estimators defined by equation (1) was selected by pragmatic considerations, without any reference to optimality. However, the gains made over the established estimators are large in a range of settings studied by simulations, not all of them reported here.

The simulations, conducted in R (R Development Core Team, 2009), can be adapted to other settings. The main difficulty is to specify a setting, the computer version of the country with its districts, that faithfully reflects the studied problem. One set of 10 000 (univariate) replications in Section 4 takes about 40 seconds, and one set of 2000 (bivariate) replications in Section 5.1 about 120 seconds of CPU time, so a wide range of alternative scenarios and loss structures can be explored in real time. The results are robust with respect to the details of how the loss functions are defined, although these details are very distant from the mean squared error loss used conventionally. The direct and composite (and EB) estimators have a higher expected weighted total loss (as well as unweighted total loss) than the policy-related estimator in all the simulated scenarios, many of them not described here.

We have treated the districts as isolated units and assumed that there is no interference among them. In practice, the labour force as well as employers react to government's applied or anticipated interventions, especially when crossing borders (of districts, regions, or even countries) entails little expense or inconvenience. Incorporating such a dynamic is beyond the scope of our analysis.

The policy-related estimator errs more frequently on the side of false positives because they are in general associated with smaller losses than false negatives. When the budget for implementing the policy is insufficient the policy-related estimator has to be adapted. Preliminary exploration indicates that the most effective way of satisfying the budget constraint is by cutting the expenditure in every district by the same percentage. Design of the survey on which the policy implementation is based is another challenging problem, relevant especially when the survey and the policy implementation share a single budget.

## Acknowledgements

This work was supported by the Spanish Ministry of Science and Technology [grant SEJ2006–13537] and by the Science Foundation of the Czech Republic [grant No. 402/09/0515]. Part of the work on the manuscript was concluded while the author was a Visiting Professor at CEPS/INSTEAD at Esch-Belval, Luxembourg, in May and June 2011. Comments on an earlier version of the manuscript by Aleix Ruiz de Villa and Yves Berger are acknowledged.

# References

DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. 2nd edition. McGraw Hill: New York.

EFRON, B. AND MORRIS, C. N. (1972). Limiting the risk of Bayes and empirical Bayes estimators

– Part II: The empirical Bayes case. *Journal of the American Statistical Association*, **67**, 130–139.

FAY, R. A. AND HERRIOT, R. E. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

GARTHWAITE, P. H., KADANE, J. B., AND O'HAGAN, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–700.

GHOSH, M. AND RAO, J. N. K. (1994). Small area estimation. An appraisal. *Statistical Science*, **9**, 55–93.

HALL, P. AND MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B*, **69**, 221–238.

LINDLEY, D. V. (1985). *Making Decisions*. Wiley and Sons, Chichester, UK.

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

LONGFORD, N. T. (2005b). On selection and composition in small-area and mapping problems. *Statistical Methods in Medical Research*, **14**, 3–16.

LONGFORD, N. T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, **33**, 69–79.

LONGFORD, N. T. (2010). Bayesian decision making about small binomial rates with uncertainty about the prior. *The American Statistician*, **64**, 164–169.

LONGFORD, N. T. (2014). Policy-related small-area estimation. UPF Working Papers 1427, Universitat Pompeu Fabra, Barcelona, Spain.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

ROBBINS, H. (1955). An empirical Bayes approach to statistics. *In* NEYMAN, J. (Editor) *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. University of California Press: Berkeley, CA, pp. 157–164.

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.

SLUD, E. V. AND MAITI, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society, Series B*, **69**, 238–257.

R DEVELOPMENT CORE TEAM (2009). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Manuscript received, 2014-01-28, revised, 2014-06-30, accepted, 2014-08-30.