CEPS INSTEAD

WORKING PAPERS

Policy-related small-area estimation

Nicholas T. LONGFORD[1]

SNTL and Department of Economics and Business, University Pompeu Fabra, Spain[1]

Working Paper No 2011-44
July 2011

Policy-related small-area estimation*

Nicholas T. Longford
SNTL and Department of Economics and Business
University Pompeu Fabra, Barcelona, Spain

July 2011

Abstract

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 that the commonly applied empirical Bayes and composite estimators are inefficient for a wide range of utility functions. Adaptations for limited budget to implement the policy are explored. An argument is presented for a closer integration of estimation and (regional) policy making.

Keywords: Composition; empirical Bayes; expected loss; borrowing strength; exploiting similarity; small-area estimation; utility function

JEL classification codes: C13 ; C14 ; C18 ; C44 ; C83

*Research for this paper was supported by Grants SEJ2006–13537 from the Spanish Ministry of Science and Technology and GAČR 402/09/0515 from the Science Foundation of the Czech Republic. 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 are acknowledged.

1 Introduction

Recent developments in small-area estimation (SAe) respond to the increasing demand for information about the regions, provinces or districts (subdomains, or areas) of a country (the domain). Together with censuses and administrative registers, large-scale national surveys are important sources of such information. The key methodological advance in SAe is borrowing strength (Robbins, 1955; Efron and Morris, 1972; Fay and Herriot, 1979; and Ghosh and Rao, 1994), that is, exploiting the similarity of the areas, possibly after taking into account relevant auxiliary information. The explicitly stated or implied goal of a typical problem in SAe is to estimate a quantity associated with each area efficiently, with minimum mean squared error (MSE), and to estimate the MSE of this estimator, preferably without bias (Hall and Maiti, 2006, and Slud and Maiti, 2006).

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

We present a perspective in which different estimators are optimal, depending on the purpose for which the estimates are to be used. We refer to this purpose as the *policy*. For example, a national government department may wish to apply a particular course of action (a measure or an intervention) to every district m in which the unemployment rate θ_m exceeds the threshold $T = 0.20$ (20%). Based on a set of recent estimates $\hat{\theta}_m$ of the rates θ_m, $m = 1, \ldots, M$, the plan may be to apply the measure in every district in which $\hat{\theta}_m > T$, in effect, regarding the estimate $\hat{\theta}_m$ as if it were the population rate θ_m. We show that the established composite estimator (Longford, 1999), and by implication the empirical Bayes estimator (Ghosh and Rao, 1994, and Rao, 2003), which aim to minimise the mean squared error of the estimator, are not useful in this context, and explore alternatives in which different shrinkage is applied.

A novel element of our approach is the incorporation of the negative utilities (losses) that quantify the consequences of inappropriate action. This 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 (p-values), are at best an intermediate and sometimes an irrelevant goal in this effort. We show that estimation of key quantities cannot be divorced from decision making; the two activities have to be closely integrated for the latter to be effective. We argue by example that decision making is within the remit of statistics because it requires statistical evaluations. These views are influenced by DeGroot (1970) and Lindley (1985 and 1992), 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 (negative-utility) functions. Suppose applying the intended measure in a district with rate $\theta_m < T$, for which the survey-based 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 $\hat{\theta}_m < T$ was obtained, is associated with loss equal to $R(\hat{\theta}_m - \theta_m)^2$, where $R \geq 1$ is a constant. In this setting, estimation with minimum expected loss is desired. Note that even for $R = 1$ the loss function in this example differs from the squared loss, defined as $(\hat{\theta}_m - \theta_m)^2$ for all pairs $\hat{\theta}_m$ and θ_m, because positive loss is incurred only when $\hat{\theta}_m < T < \theta_m$ or $\hat{\theta}_m > T > \theta_m$. The mean squared error corresponds to the quadratic kernel with $R = 1$ and T set to θ_m. In a typical application, the same threshold T applies to all districts, but the development we consider is not restricted to this case, although the threshold has to be known; θ_m is not known.

We show that the empirical Bayes 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 that have the form

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

where $\hat{\theta}_m$ is a direct (unbiased) estimator of θ_m, which uses information only from the focal district m and the variable concerned, and b_m and F_m are constants, called the shrinkage coefficient and the focus of shrinkage, respectively. Empirical Bayes estimators have this form with $F_m = \hat{\theta}$ for all districts, where $\hat{\theta}$ is an estimator of the mean of the district-level means (or rates), $\theta = (\theta_1 + \theta_2 + \cdots + \theta_M)/M$. We consider first the setting with no auxiliary variables. That is, the sole information we have about θ_m is in the values of the focal variable, y_{im}, on subjects $i = 1, \ldots, n_m$ in districts $m = 1, \ldots, M$, and the corresponding sampling weights w_{im}. To avoid complexities that would dilute our focus, we assume that $\hat{\theta}_m$ are linear statistics

in y_{im} and $\hat{\theta}$ is a linear combination of $\hat{\theta}_1, \ldots, \hat{\theta}_M$.

The next section gives formal definitions of the key concepts and Section 3 derives an estimator which, setting aside some approximations and estimation, 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. Section 6 explores adaptations necessary when the budget for implementing the policy is limited. The paper is concluded with a discussion.

2 Policy and utility

A policy is formally defined as an algorithm for selecting one of a given finite set $\mathcal{A}$ of courses of action, based on the available information. When all the information is contained in the estimator $\hat{\theta}_m$, the policy is defined as $d_m = D(\hat{\theta}_m)$, $m = 1, \ldots, M$, with actions in $\mathcal{A}$ as its possible values. We refer to $d_m^* = D(\theta_m)$ as the *ideal version* of the policy. The inverse images, $\mathcal{T}_d = \{\eta; D(\eta) = d\}$ for $d \in \mathcal{A}$, partition the parameter space into subspaces according to the actions. We assume that the policy function D is completely formulated by the policy maker, and if θ_m were available, d_m^* would be established immediately. That is, incomplete information about θ_m is the (policy maker's) sole problem.

We consider a policy that calls for one of two courses of action; $\mathcal{A} = (\mathrm{A}, \mathrm{B})$. Action A is appropriate in district m if $\theta_m \in \mathcal{T}_\mathrm{A} = \mathcal{T}$ and action B is appropriate otherwise. The set $\mathcal{T}$ is given. The two actions are exclusive (it is impossible to apply both of them) and complementary (one of them has to be applied). In the example in Section 1, $\mathcal{T} = (T, 100]$ and $T = 20\%$.

The *loss* function for an action d is defined as a non-negative function of the estimate and the target, $L_d(\hat{\theta}_m, \theta_m)$. We drop the subscript d when we refer to the actual policy; that is, $L(\hat{\theta}_m, \theta_m) = L_{D(\hat{\theta}_m)}(\hat{\theta}_m, \theta_m)$. Action d is said to be appropriate for district m if it is associated with no loss. We assume that one of the two actions is appropriate for every district, and if θ_m were known this action, for which $L_d(\theta_m, \theta_m) = 0$, would be readily identified. The assumption that one action is appropriate for each value of θ_m is not restrictive, because in practice only the difference $L_\mathrm{A}(\hat{\theta}_m, \theta_m) - L_\mathrm{B}(\hat{\theta}_m, \theta_m)$ matters. Further, we can associate any pair of loss functions $L_d(\hat{\theta}_m, \theta_m)$, d = A or B, with the class of equivalence defined by

the functions CL_d, where $C > 0$ is an arbitrary constant, common to L_{A} and L_{B}. If the loss is expressed in a particular currency, such as \$US, then C is the conversion rate to another currency.

By choosing action $D(\hat{\theta}_m)$, treating the estimate as if it were the population quantity, two kinds of error may be committed: choosing A when B is appropriate, when $\hat{\theta}_m \in \mathcal{T}$ and $\theta_m \notin \mathcal{T}$, and choosing B when A is appropriate, when $\hat{\theta}_m \notin \mathcal{T}$ and $\theta_m \in \mathcal{T}$. Parallels can be drawn with hypothesis testing, where we also have two kinds of error, but in our approach the related probabilities are relevant only in some special cases. Our point of departure from hypothesis testing is that the losses we consider, interpretable as the consequences of making the two kinds of bad decision, may depend on the magnitude of the error, $|\,\hat{\theta}_m - \theta_m\,|$, and some estimation errors are associated with no loss. It is not appropriate to assess the magnitude of the error by $|\,\hat{\theta}_m - T\,|$ or its increasing transformation, because the trivial estimator $\hat{\theta}_m \equiv T$ would then be optimal.

The loss functions L_{A} and L_{B} should be elicited from the policy maker. This is an activity similar to eliciting a (Bayes) prior, although we do not expect the elicitation process to conclude with a single pair of functions (or classes of equivalence) L_{A} and L_{B}. Instead, we work with 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 loss function for each action, and that it is contained in the set of plausible loss functions, but it cannot be identified. See Longford (2010) for a similar approach to dealing with uncertainty about the (Bayes) prior and Garthwaite, Kadane and O'Hagan (2005) for a comprehensive review of statistical issues in elicitation, although their focus is on elicitation of prior distributions. We want the elicited set to be as small as possible, but we do not want to generate any discomfort in the elicitation process by forcing the choice of the set of loss functions to be too narrow, or even reduced to a single pair, which might not include, or might differ from, the ideal pair of loss functions.

For $\mathcal{T} = (T, +\infty)$, we give three examples of (pairs of) loss functions

1. $L_{\mathrm{A}}(\hat{\theta}_m\,, \theta_m) = R(\hat{\theta}_m - \theta_m)^2$ when $\hat{\theta}_m < T < \theta_m$, and $L_{\mathrm{A}}(\hat{\theta}_m\,, \theta_m) = 0$ otherwise; $L_{\mathrm{B}}(\hat{\theta}_m\,, \theta_m) = (\hat{\theta}_m - \theta_m)^2$ when $\hat{\theta}_m > T > \theta_m$, and $L_{\mathrm{B}}(\hat{\theta}_m\,, \theta_m) = 0$ otherwise.

2. $L_{\mathrm{A}}(\hat{\theta}_m\,, \theta_m) = R(\theta_m - \hat{\theta}_m)$ when $\hat{\theta}_m < T < \theta_m$, and $L_{\mathrm{A}}(\hat{\theta}_m\,, \theta_m) = 0$ otherwise; $L_{\mathrm{B}}(\hat{\theta}_m\,, \theta_m) = \hat{\theta}_m - \theta_m$ when $\hat{\theta}_m > T > \theta_m$, and $L_{\mathrm{B}}(\hat{\theta}_m\,, \theta_m) = 0$ otherwise.

3. $L_{\rm A}(\hat{\theta}_m\,,\theta_m) = R$ when $\hat{\theta}_m < T < \theta_m$, and $L_{\rm A}(\hat{\theta}_m\,,\theta_m) = 0$ otherwise;
   $L_{\rm B}(\hat{\theta}_m\,,\theta_m) = 1$ when $\hat{\theta}_m > T > \theta_m$, and $L_{\rm B}(\hat{\theta}_m\,,\theta_m) = 0$ otherwise.

We refer to these pairs of loss functions as having quadratic, linear and absolute kernel, respectively, and to the constant R as the *penalty ratio*. A pair of loss functions in $1-3$ can be expressed as a single function as

$$L(\hat{\theta}_m\,,\theta_m) \;=\; L_{\rm A}\Big(\hat{\theta}_m\,,\theta_m\Big) \;+\; L_{\rm B}\Big(\hat{\theta}_m\,,\theta_m\Big)\;;$$

at most one of the contributions is positive for any $\hat{\theta}_m$ and θ_m. The absolute kernel has some afinity to hypothesis testing, in that the expected losses are related to probabilities. Unlike in hypothesis testing, where we fix one (conditional) probability (the size of the test), and maximise the other (the power), we aim with the absolute kernel loss for their magnitudes to be in proportion $1 : R$. When the loss depends on the magnitude of the error, absolute kernel has little to recommend.

Loss functions other than $1-3$ can be defined, although these three cases are relatively easy to handle. For example, the penalty ratio need not be constant and other kernels can be defined; an example is given in Section 3. Different loss functions may be defined for distinct subsets of districts by using different penalty ratios, or even different kernels. The functions $L_{\rm A}$ and $L_{\rm B}$ 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 focussed on a single district.

3 Policy-related estimator

Suppose the sampling distribution of $\hat{\theta}_m$ is normal with expectation γ_m and variance ν_m^2, that is, $\hat{\theta}_m \sim \mathcal{N}(\gamma_m\,,\nu_m^2)$. We do not assume that $\gamma_m = \theta_m$. Denote by ϕ the density of the standard normal distribution $\mathcal{N}(0,1)$ and by Φ its distribution function. With the quadratic kernel, the expected loss with the policy applied to district m according to estimator $\hat{\theta}_m$ is

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

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

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

where $\tilde{z} = (\gamma_m - T)/\nu_m$ and $z_\dagger = (\gamma_m - \theta_m)/\nu_m$. We do not aspire to minimise these two functions of γ_m and ν_m directly, but seek estimators $\hat{\theta}_m$ which have the following two well motivated properties:

- *equilibrium condition* — for a district with $\theta_m = T$, the choice between actions A and B is immaterial in expectation:

$$\mathrm{E}\left\{ L_{\mathrm{A}}\left(\hat{\theta}_m, T\right) \right\} = \mathrm{E}\left\{ L_{\mathrm{B}}\left(\hat{\theta}_m, T\right) \right\} ;$$

- *minimum averaged* MSE.

Averaging in the second condition refers to marginalisation over the distribution estimated or assumed to underlie the values $\theta_1, \theta_2, \ldots, \theta_m$, as applied in empirical Bayes analysis. Averaging removes the dependence of the solution on θ_m.

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

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

We refer to the left-hand side as the equilibrium function (of $\tilde{z}$). It has a single root for all R. To prove this, we show that the function is increasing; its limits as $\tilde{z} \to \pm\infty$ are $\pm\infty$, respectively. Its first and second derivatives are $2(R+1)\{\tilde{z}\Phi(\tilde{z}) + \phi(\tilde{z})\} - 2R\tilde{z}$ and $2(R+1)\Phi(\tilde{z}) - 2R$, respectively. The latter is increasing in $\tilde{z}$ and its root is $\tilde{z}^\circ = \Phi^{-1}\{R/(R+1)\}$. At this root, the first derivative attains its minimum, equal to

$$2(R+1)\, \phi\left(\tilde{z}^\circ\right) > 0 .$$

Therefore, the first derivative is positive throughout. The equilibrium value z^*, the solution of (2), is found by the Newton method.

The expectation and variance of $\tilde{\theta}_m$ in (1) are

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

assuming that the direct estimator $\hat{\theta}_m$ is unbiased with variance v_m. In the arguments of E and var we add conditioning on the value of θ_m, to emphasise that we regard it as fixed (related to a well-specified and labelled district), unlike in the usual treatment of (exchangeable) districts in empirical Bayes analysis (Ghosh and Rao, 1994; and Rao, 2003). See Longford (2005b, Chapter 6, and 2007) for related discussion. In the simulations in Section 4, the country's districts are also treated as fixed, constant across replications. The sample selection (independently in each district) is the sole source of variation.

The MSE of $\tilde{\theta}_m$ is

$$\text{MSE}\left(\tilde{\theta}_m\,;\,\theta_m\right) \;=\; (1-b_m)^2\,v_m + b_m^2\,(F_m-\theta_m)^2\;.$$

The dependence on θ_m is avoided by averaging over the district-level distribution of θ_m, $m = 1,\ldots M$, which has mean θ and variance σ^2_{B}. The averaged MSE (aMSE) is

$$(1-b_m)^2\,v_m + b_m^2\left\{\sigma^2_{\text{B}} + (F_m-\theta)^2\right\}\;,$$

and its minimum is attained for

$$b^*_m \;=\; \frac{v_m}{v_m + \sigma^2_{\text{B}} + (F_m-\theta)^2}\,; \tag{3}$$

if we ignore the equilibrium condition, the shrinkage coefficient is always within the range $(0,1)$.

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

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

and the coefficient that minimises this quantity has to satisfy the identity

$$b_m \;=\; \frac{v_m\left(1+z^{*2}\right) - \text{sign}(1-b_m)\,(T-\theta)\,z^*\sqrt{v_m}}{v_m + \sigma^2_{\text{B}} + \left\{z^*\sqrt{v_m} - \text{sign}(1-b_m)\,(T-\theta)\right\}^2}\,, \tag{5}$$

where the sign function is defined as $\text{sign}(x) = 1$ for $x > 0$, $\text{sign}(x) = -1$ for $x < 0$, and $\text{sign}(0) = 0$. The aMSE is continuous and diverges to $+\infty$ for $b_m \to \pm\infty$, so it has an odd number of extremes. Equation (5) implies that it cannot have more than two minima. Hence it has a unique minimum, and it is its only extreme.

The solution b_m^* may be outside $(0, 1)$, and then it does not have the common interpretation of a shrinkage coefficient. It exceeds unity when

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

that is, for sufficiently large v_m when $T < \theta$. It is negative when

$$\sqrt{v_m} \;<\; \frac{z^*}{1 + z^{*2}}\,(T - \theta)\,,$$

that is, for sufficiently small v_m when $T > \theta$. However, b_m^* is not a monotone function of v_m.

No shrinkage is applied when θ_m is known, and $b_m^* \to 0$ as $v_m \to 0$, but $b_m^* = 0$ 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 θ_m, $\tilde{\theta}_m = T$ is optimal, unlike in empirical Bayes estimation, where $\tilde{\theta}_m = \hat{\theta}$ in such a case. When $b_m^* = 0$, we have an anomaly because the corresponding value of F_m in (4) is not defined. However, the product $b_m^* F_m$ is well defined by its limit, equal to $z^* \sqrt{v_m}$, so the estimator in (1) is well defined.

Symmetric loss, with $R = 1$, corresponds to $z^* = 0$ and $F_m = T$ for all districts m. The coefficient b_m^* in (5) coincides with its empirical Bayes counterpart only when $R = 1$ and $T = \theta$. Without the averaging, such coincidence would arise in the unimplementable condition $T_m = \theta_m$ (an unknown threshold, specific to each district), which is closer to the intent of estimating with minimum MSE than with minimum aMSE.

For the linear kernel loss function, we have the equilibrium condition

$$(R + 1)\,\{\tilde{z}\Phi\,(\tilde{z}) + \phi\,(\tilde{z})\} \;=\; R\tilde{z}\,, \tag{6}$$

and for the absolute kernel,

$$\Phi(\tilde{z}) \;=\; \frac{R}{R + 1}\,. \tag{7}$$

The former equation is solved by the Newton method; it 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. When $\hat{\theta}_m$ has a symmetric distribution, no generality is lost by assuming that $R \geq 1$, because we could work with the outcomes $-y$, estimators $-\hat{\theta}_m$ and $-\hat{\theta}$, and penalty ratio $1/R$. For each function $z_{\mathrm{G}}^*(R)$, G = A, L or Q, the subspace above the function corresponds to action A and the subspace below to action B being preferable.

Comparisons of the functions z_{A}^*, z_{L}^* and z_{Q}^* are in general not meaningful. Nevertheless, the (uniform) inequality $z_{\mathrm{A}}^* > z_{\mathrm{L}}^* > z_{\mathrm{Q}}^*$ can be interpreted as follows. For fixed R, we should be

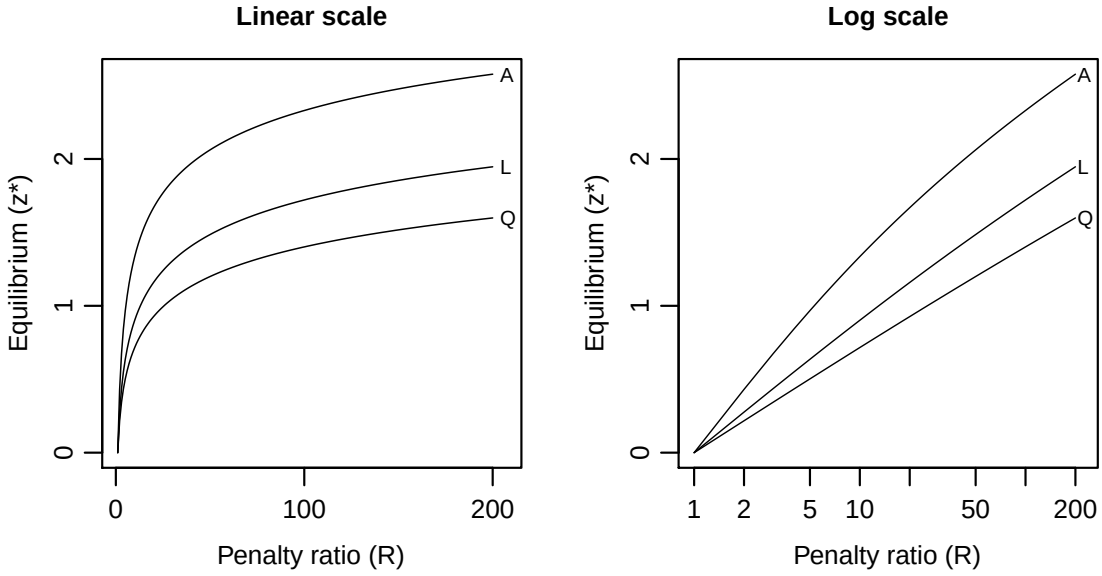

Figure 1: The roots of the equilibrium equations, z^*, as functions of the penalty ratio R for the absolute (A), linear (L) and quadratic (Q) kernel loss functions, on the linear and log scales for R.

disposed toward action A more favourably with the quadratic than with the other two kernels. After all, with action A we rule out false negatives which tend to be associated with relatively high losses.

The equilibrium conditions (2), (6) and (7) involve γ_m and ν_m only through $\tilde{z}$. This is not the property of any easy-to-identify class of loss functions. For example, for the exponential kernel, given by the functions

$$\begin{aligned} L_{\mathrm{A}}\left(\hat{\theta}_m\,,\theta_m\right) &= R\,\exp\left(\theta_m-\hat{\theta}_m\right)-R \\ L_{\mathrm{B}}\left(\hat{\theta}_m\,,\theta_m\right) &= \exp\left(\hat{\theta}_m-\theta_m\right)-1\,, \end{aligned}$$

for $\hat{\theta}_m < T < \theta_m$ and $\hat{\theta}_m > T > \theta_m$, respectively, the expected losses are

$$\begin{aligned} E_{\mathrm{A}} &= R\exp\left(\theta_m-\gamma_m+\frac{\nu_m^2}{2}\right)\left\{1-\Phi\left(\tilde{z}-\gamma_m\right)\right\}-R+R\Phi(\tilde{z}) \\ E_{\mathrm{B}} &= \exp\left(\gamma_m-\theta_m+\frac{\nu_m^2}{2}\right)\Phi\left(\tilde{z}+\nu_m\right)-\Phi(\tilde{z})\,, \end{aligned}$$

and the equilibrium solution is not a function solely of $\tilde{z}$.

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 ($1.0 \le v_m \le 2.5$) for the quadratic kernel loss and penalty ratios ranging from $R = 10$ to $R = 100$. The mean of the district-level means is $\theta = 16\%$, the

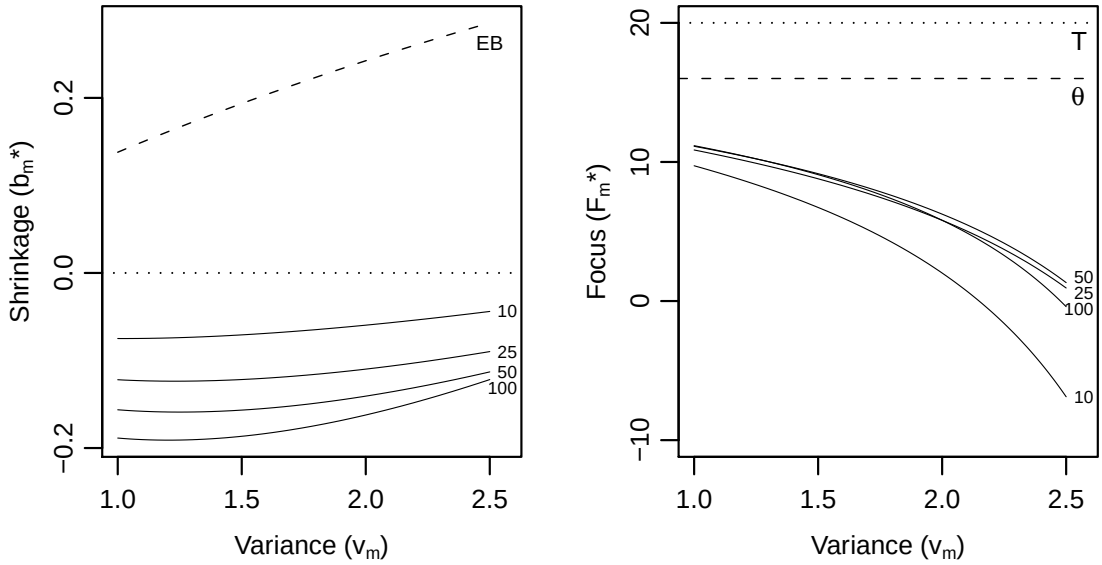

Figure 2: The optimal shrinkage coefficients and foci of shrinkage for quadratic kernel loss and penalty ratios R = 10, 25, 50 and 100, indicated at the right margin; $\theta = 16$, $T = 20$ and $\sigma_{\mathrm{B}}^2 = 6.25$. The coefficient and focus of the empirical Bayes estimator is drawn by dashes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

district-level variance is $\sigma_{\mathrm{B}}^2 = 6.25$ (%^{2}), and the threshold is set to $T = 20$%. The shrinkage coefficient of the empirical Bayes estimator (EB), $v_m/(v_m + \sigma_{\mathrm{B}}^2)$, 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 $\hat{\theta}_m$ and focus F_m are optimal than in empirical Bayes estimation. The focus of shrinkage is smaller than θ and decreases with the variance v_m. However, the shrinkage is *negative*, away from these foci. We emphasise that we search for estimators that lead to the best implementation of a policy in expectation, and do not insist on any appealing interpretation. We regard negative shrinkage as acceptable, so long as the resulting policy is optimal or at least superior to the alternatives we have, in the sense defined to reflect the policy maker's assessment of the utilities.

4 Simulations

In the derivations in Section 3, we made several simplifying assumptions, such as the knowledge of the global parameters θ and σ_{B}^2, and applied averaging to minimise aMSE at $\theta_m = T$, instead of minimising the expected loss directly. Without this compromise, the problem would be intractable. Note that we did not assume (superpopulation) normality of the district-level

summaries θ_m. We assess the properties of the estimators derived 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. The labour force of the whole country is 58.90 million. The focal variable is unemployment, a dichotomy, and the district-level (population) rates of unemployment 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, which comprises the country's capital, has an unemployment rate well below average. The correlation of the district-level population sizes and unemployment rates is 0.18, but when the capital is removed, the correlation of the remaining 59 districts is 0.27. Twenty-two districts have unemployment rates in excess of the threshold set at $T = 20\%$; these districts account for 23.23 million members of the labour force (39.4%). 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 $\theta^* = 17.3\%$. They are marked in the diagram by horizontal dashes and dots, respectively. The variance of the district-level unemployment rates is $\sigma_{\mathrm{B}}^2 = 27.05$ (%^{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 in each district. The district-level sample sizes n_m, indicated in Figure 3 by the size of the black disc, are in the range 113 – 567, sufficiently large for the normal approximation to be satisfactory for all the sample rates $\hat{\theta}_m$. The sample sizes are approximately proportional to $N_m^{0.9}$, so that the least populous districts tend to have higher sampling fractions. The overall sample size is $n = 17\,500$.

We assume the quadratic kernel loss function with plausible penalty ratio in the range $(5, 20)$. For motivation, suppose the ideal penalty ratio is $R = 10$, but the policy maker is not sure about it. The elicitation started with a very wide range of penalty ratios, and after several reductions it reached the point at which the expert was not willing to narrow the range down any further.

For orientation, we discuss the results for a single replication of sampling and estimation. Independent samples of fixed sizes n_m are drawn within the districts from Bernoulli distributions with respective probabilities θ_m, and the sample rates $\hat{\theta}_m$, composite estimates $\tilde{\theta}_m$ (shrinkage toward $\hat{\theta}$), and the policy-related estimates $\tilde{\theta}_m^*$ (shrinkage toward $\hat{F}_m^*$) are evaluated,

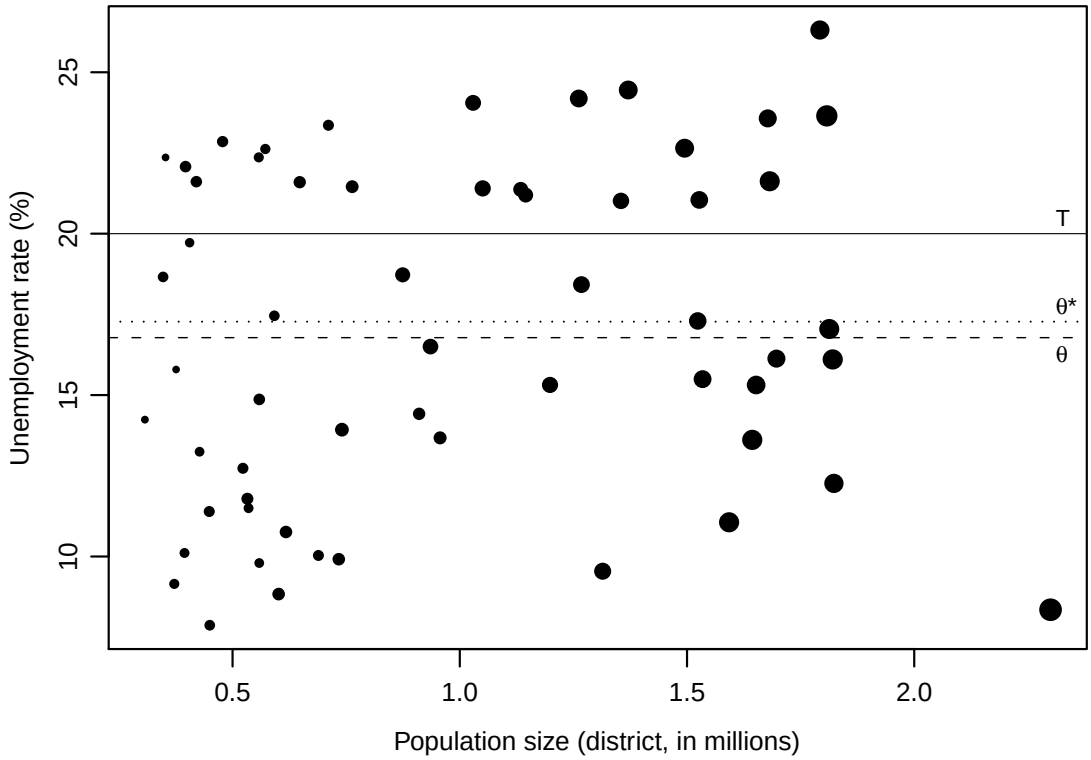

Figure 3: The population sizes and unemployment rates in the districts of a country. Computer-generated data used for simulation. The area of the black disc is proportional to the sample size of the district it represents.

based on the estimates of θ and σ_{B}^2 . The latter variance is estimated by moment-matching; see Longford (1999 and 2005b, Section 8.3.2). The loss (if any) is evaluated for each estimate and district.

In a particular replication, the policy based on the obtained sample rates (henceforth estimator S) would lead to the inappropriate action in five districts that have a total labour force of 4.47 million (7.6%), and are represented in the survey by 1349 subjects (7.7%). Four cases are false negatives, $\hat{\theta}_m < T < \theta_m$ ($\hat{\theta}_m = 18.9\%$ vs. $\theta_m = 21.0\%$, 17.9 vs. 21.6, 19.3 vs. 21.5 and 19.6 vs. 21.0) and one is a false positive, $\hat{\theta}_m > T > \theta_m$ (21.4 vs. 19.7). It is not meaningful to add up the losses L_{A} and L_{B} , because they are not comparable across the districts. We evaluate instead their weighted total, with the population sizes N_m (in millions) as the weights. This weighted total is equal to 194.2 for the four false negatives, and to only 1.2 for the one false positive. The largest loss, 71.8, arises for a district (the case 18.9 vs. 21.0) with labour force of 1.53 million, about 50% above average. The second largest district among those with losses, with labour force of 1.35 million, is also a false negative (19.6 vs. 21.0), but the loss is only

27.0, because the estimation error is smaller.

The composite estimator $\tilde{\theta}_m$ with shrinkage toward the estimated average district-level rate (estimator C) leads to a poorer policy, with total weighted loss of 364.4; inappropriate action is taken in all five districts mentioned earlier, and in two others, both of them false negatives, by narrow margins (19.1 vs. 22.6% and 19.5 vs. 21.2%). The sole false positive contributes to the total loss by only 0.04, because its estimate is shrunk to 20.05%, very close to the threshold of 20% and to the target, 19.7%. In all other cases, shrinkage is in the direction in which the loss is increased; shrinkage for the two new cases moves the estimates across the threshold.

With the policy-related composite estimator (estimator P), aimed to minimise the expected loss (for each district), inappropriate action is taken in five districts, one false negative and four false positives. Only two of these districts, the false negative and a false positive, contribute to losses also with the other estimates. The weighted total loss is only 72.4. The reason for this large reduction is that the shrinkage applied has eliminated all but one false negative (19.5 vs. 21.6%), and even for the latter the loss is greatly reduced. Simply, smaller loss is incurred in total by erring on the side of overestimating θ_m, even if some additional false positives are created in the process. Table 1 displays the estimates and losses associated with the districts discussed.

In the simulations, we replicate this process 10 000 times and accumulate the losses separately for each district and the three estimators. The expected loss for each district is estimated by the corresponding average loss. The results are summarised in Figure 4. The average losses (not multiplied by N_m), evaluated with the three estimators, are marked by the symbols C, P and S, and are connected by vertical segments when the average losses differ by more than 2.5. When the average 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. It is a coincidence that the same scale is suitable for the rates and the average losses.

The diagram shows that most of the losses are incurred by false negatives, for districts with $\theta_m > T$, and even among them the loss for one district dominates for estimators C and S. The weighted total loss has expectations 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, such as 2.0, using the same estimators (based on $R = 10$), the composite and direct estimators

Table 1: The districts associated with losses based on the sample rate $\hat{\theta}_m$, the standard composite estimator $\tilde{\theta}_m$ and the policy-related composite estimator $\tilde{\theta}^*_m$. Based on the first replication of the simulation study.

	Design			*Estimates*			*Losses ($\times N_m$)*		
m	N_m	n_m	θ_m	$\hat{\theta}_m$	$\tilde{\theta}_m$	$\tilde{\theta}^*_m$	$L(\hat{\theta}_m)$	$L(\tilde{\theta}_m)$	$L(\hat{\theta}^*_m)$
2	0.572	175	22.62	20.00°	19.14	21.10	0.00	69.41	0.00
9	1.145	304	21.20	20.07	19.50	21.08	0.00	32.87	0.00
13	0.406	154	19.72	21.43	20.05	21.94	1.18	0.04	1.99
19	1.527	392	21.05	18.88	18.59	20.00^+	71.77	92.42	0.00
20	0.420	207	21.61	17.87	17.59	19.55	58.66	67.90	17.86
29	0.763	244	21.46	19.26	18.75	20.51	36.72	55.91	0.00
32	1.524	386	17.30	18.91	18.61	20.04	0.00	0.00	11.44
39	0.911	231	14.42	18.61	18.21	20.04	0.00	0.00	28.76
47	1.355	352	21.02	19.60	19.17	20.66	27.05	45.89	0.00
52	0.307	116	14.24	18.97	18.17	20.58	0.00	0.00	12.34
Totals							195.38	364.44	72.39

Notes: N_m — the size of the labour force in district m (in millions); n_m — the sample size for district m; L — the loss, multiplied by the size of the labour force; $^+$ — exact value greater than 20.00; $^\circ$ — exactly equal to 20.00; no loss incurred.

would remain far inferior; the weighted total losses would have averages 103.6, 123.1 and 68.7. Estimators C and S are insensitive to the penalty ratio so the same estimates are obtained when we set $R = 2$ for them, whereas for estimator P a smaller value of the expected loss, 65.4, is obtained. The expected losses with C and S have the form $M + RU$, where M is the expected loss for the false negatives and U the expected loss for the false positives, pro-rated for unit penalty ($R = 1$).

Only one of the 22 deserving districts incurs small average losses with all three estimators, and four other districts have small average losses only with estimator P. For every deserving district, the average loss is the highest for estimator C, followed by S and P. This ordering is not maintained for the 38 districts with $\theta_m < T$; in neither of the eight districts that have non-trivial average losses, is estimator P associated with the smallest average loss. However, all these average losses are much smaller than for most of the deserving districts.

In summary, the simulations show that the shrinkage applied by the composition for minimising aMSE is counterproductive, and a substantially reduced expected weighted total loss

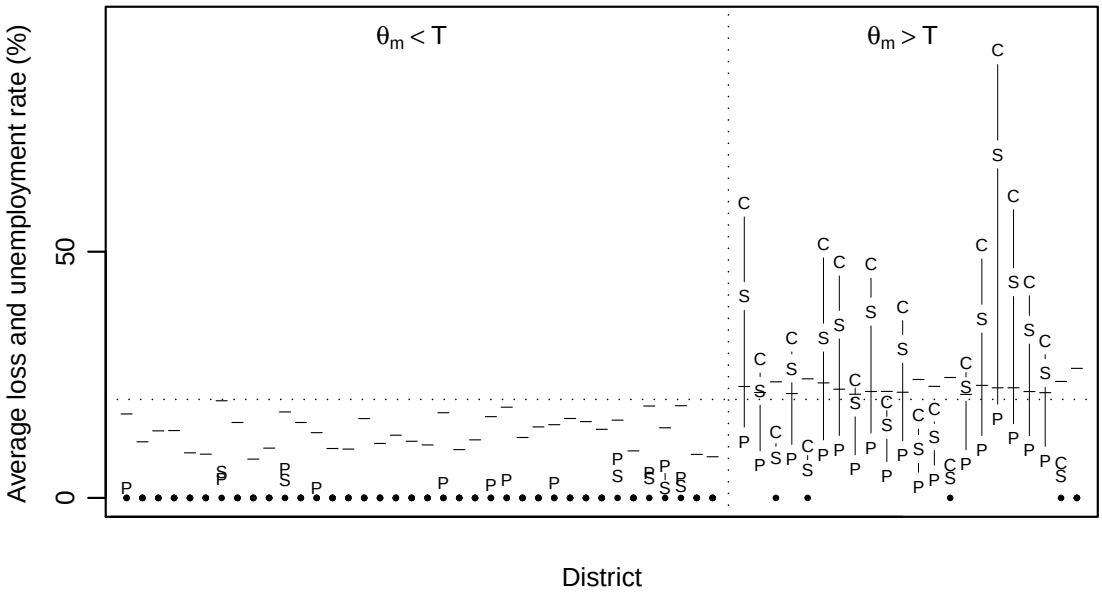

Figure 4: The empirical expected (average) losses for the districts and estimators (direct — S; composite — C; and policy-related — P), with penalty ratio $R = 10$. The districts are in the ascending order of population size, within the two groups divided by the threshold $T = 20\%$. The districts' unemployment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

is obtained with the policy-related shrinkage scheme. We repeated the simulations with $R = 5$ and $R = 20$ to confirm that estimator P based on $R = 10$ is 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. Figure 6 compares the expected losses with the two penalty ratios more directly by plotting the average losses with estimator P for the two sets of districts, normal ($\theta_m < T$, E_{B}) and deserving ($\theta_m > T$, E_{A}/R), in separate panels. The values plotted are pro-rated for unit loss (not multiplied by R), to make the two sets of expected losses comparable. The diagram shows that the relative loss is greater with $R = 5$ for every deserving district and smaller for every normal district; with higher penalty ratio we are more averse to having false positives, even after discounting the factor R. A mirror-image of this conclusion is drawn from the simulations for $R = 20$, with the roles of the normal and deserving districts reversed; the details are omitted. We conclude that there is considerable robustness of the expected losses with respect to the specification of the penalty ratio R.

The policy-related estimator is not superior for every district. One reason for this is the averaging applied to obtain a shrinkage coefficient that does not depend on θ_m. Averaging

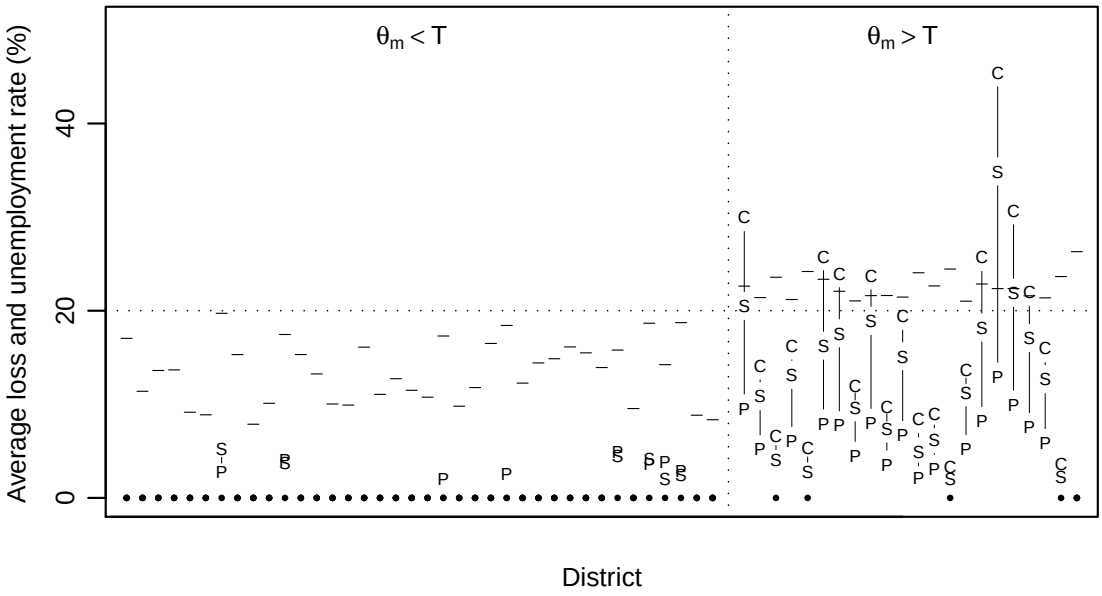

Figure 5: The empirical expected (average) losses for the districts and estimators (direct — S; composite — C; and policy-related — P), with penalty ratio $R = 5$. The districts are in the same order as in Figure 4.

introduces large errors for districts for which $(\theta_m - F_m)^2$ differs a lot from $\sigma_{\mathrm{B}}^2 + (F_m - \theta)^2$. This happens for a few districts with the smallest unemployment rates, but the expected losses for them are very small because their rates are distant also from the threshold T, and the inappropriate action for each of them has a small probability. In fact, losses due to false positiveness are non-trivial only for eight districts (out of 38); P is not the minimum-loss estimator for either of them.

We conclude this section by a summary of the simulations with the quadratic, linear and absolute kernel losses displayed in Table 2. The table of weighted totals of the (empirical) expected losses shows that the policy-related estimator (P) has a distinct advantage over the direct (S) and the established composite estimator (C) for higher penalty ratios. For $R = 1$, the advantage of estimator P is only slight for quadratic and linear kernels, and for the absolute kernel the direct estimator is preferable to both composite 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 the policy-related estimator.

Note that expected losses, or their totals, cannot be compared across the kernels, because

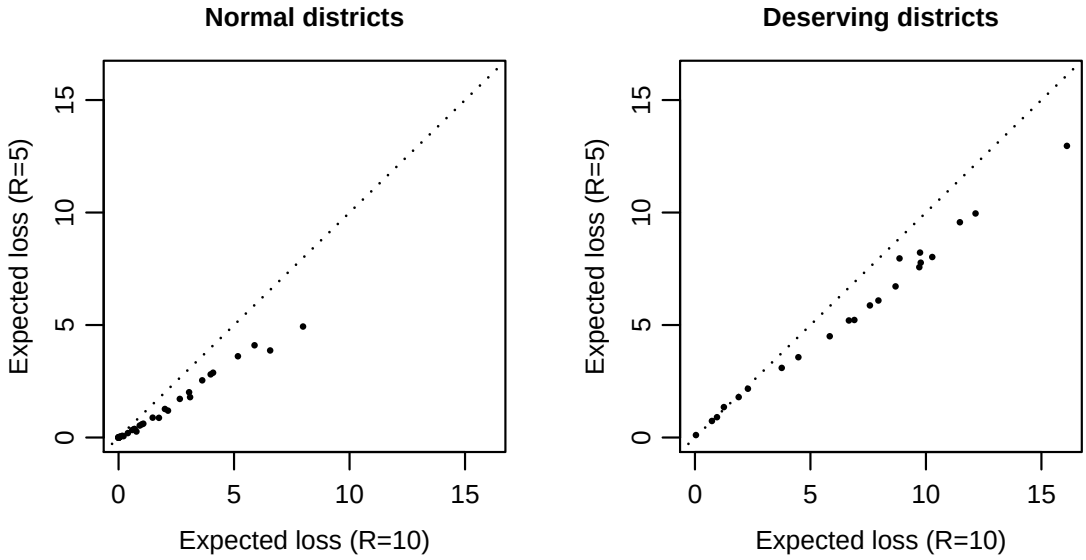

Figure 6: The average losses incurred by the policy-related estimators with $R = 10$ (horizontal axis) and $R = 5$ (vertical axis).

Table 2: The expected total losses, weighted by the population size, in simulations with quadratic, linear and absolute kernels and penalty ratios $R = 1$, 5, 10 and 20. Based on 10 000 replications. The estimators used are: P — policy-related; S — direct; C — composite (empirical Bayes).

	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

they regard the relative losses with small and large deviations $|\,\hat{\theta}_m - \theta_m\,|$ differently.

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 (similar to) θ_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 θ_m in the past year(s), values of a quantity prima facie closely related to θ_m obtained from an administrative register, and the values of the same summary as θ_m but estimated in a different subpopulation; see Longford (2005b, 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 $\boldsymbol{\Sigma}_{\mathrm{B}} = \mathrm{var}_m\,(\boldsymbol{\theta}_m)$. The latter variance matrix is for variation over the districts, in parallel with σ^2_{B} in Section 3; the other variances and covariances refer to sampling (estimation). The covariance 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 vector of coefficients $\mathbf{b}_m$ has the ideal version

$$\mathbf{b}^*_m \;=\; \mathbf{Q}_m^{-1} \mathbf{P}_m\,,$$

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, $\mathbf{Q}_m$ and $\mathbf{P}_m$ have to be estimated, yielding the vector $\hat{\mathbf{b}}_m = \hat{\mathbf{Q}}_m^{-1} \hat{\mathbf{P}}_m$ and the estimator $\tilde{\theta}_m = \tilde{\theta}_m(\hat{\mathbf{b}}_m)$. This is a generalisation of the univariate composite estimator, which is obtained for 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 fraction of the entire sample in one or several surveys on which $\hat{\boldsymbol{\theta}}_m$ are based. When there is no such 'dominating' 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 $\mathrm{E}\{L(\hat{\theta}_m\,, \theta_m)\}$:

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

We search for suitable vectors $\mathbf{b}_m$ and $\mathbf{F}_m$, the multivariate versions of the shrinkage coefficient b_m and focus F_m, respectively, that satisfy the conditions of equilibrium for $\theta_m = T$ and have minimum aMSE. For the former, we have to specify an entire vector $\mathbf{T} = (T, \boldsymbol{\xi}_{\mathrm{T}}^\top)^\top$. We set the auxiliary part of $\mathbf{T}$, $\boldsymbol{\xi}_{\mathrm{T}}$, to its conditional expectation given the first component,

$$\boldsymbol{\xi}_{\mathrm{T}} \;=\; \mathrm{E}\,(\boldsymbol{\xi}\,|\,T) \;=\; \frac{T-\theta}{\sigma^2_{\mathrm{B},1}}\,\boldsymbol{\Sigma}_{\mathrm{B},-1,1}$$

where $\sigma^2_{\mathrm{B},1}$ is the (1,1)-element of $\boldsymbol{\Sigma}_{\mathrm{B}}$ and $\boldsymbol{\Sigma}_{\mathrm{B},-1,1}$ is the first column of $\boldsymbol{\Sigma}_{\mathrm{B}}$, with its first element removed.

The condition of equilibrium at $\mathbf{T}$ is

$$\mathbf{b}_m^\top\,(\mathbf{F}_m \;-\mathbf{T}) \;=\; s z^*\,, \tag{8}$$

where $s = \sqrt{(\mathbf{u}-\mathbf{b}_m)^\top \mathbf{V}_m (\mathbf{u}-\mathbf{b}_m)}$. The MSE of a multivariate composite estimator $\tilde{\theta}_m$ is $s^2 + \{\mathbf{b}_m^\top(\mathbf{F}_m - \mathbf{T})\}^2$ and its aMSE, the expectation over the districts, is

$$s^2(\mathbf{b}_m) + \mathbf{b}_m^\top \left\{ \boldsymbol{\Sigma}_{\mathrm{B}} + (\mathbf{F}_m - \boldsymbol{\theta})\,(\mathbf{F}_m - \boldsymbol{\theta})^\top \right\} \mathbf{b}_m\,.$$

The argument $\mathbf{b}_m$ is added to s to indicate the dependence. By substituting the condition in (8) we obtain the expression

$$\begin{aligned} \mathrm{aMSE}(\tilde{\theta}_m\,; \theta_m\,|\,\mathbf{T}) \;&=\; \mathbf{b}_m^\top \boldsymbol{\Lambda} \mathbf{b}_m - 2\left(1 + z^{*2}\right) \mathbf{b}_m^\top \mathbf{V}_m \mathbf{u} + \mathbf{u}^\top \mathbf{V}_m \mathbf{u} \\ &\quad + 2 s\,(\mathbf{b}_m)\, z^*\, \mathbf{b}_m^\top\,(\mathbf{T} - \boldsymbol{\theta})\;, \end{aligned} \tag{9}$$

where $\boldsymbol{\Lambda} = (1 + z^{*2})\mathbf{V}_m + \boldsymbol{\Sigma}_{\mathrm{B}} + (\mathbf{T} - \boldsymbol{\theta})(\mathbf{T} - \boldsymbol{\theta})^\top$. The minimum of this function, with estimates substituted for $\mathbf{V}_m$, $\boldsymbol{\Sigma}_{\mathrm{B}}$ and the relevant components of $\boldsymbol{\theta}$ and $\mathbf{T}$, is found by the Newton-Raphson method. With the last term in (9) removed, the aMSE is a quadratic function of $\mathbf{b}_m$, which attains its minimum for

$$\mathbf{b}_m^{(\circ)} \;=\; \left(1 + z^{*2}\right) \boldsymbol{\Lambda}^{-1} \mathbf{V}_m \mathbf{u}\,;$$

it can be used as the initial solution for the Newton-Raphson iterations.

The first and second-order partial differentials of aMSE in (9) are

$$\begin{aligned}
\frac{\partial \mathrm{aMSE}}{\partial \mathbf{b}_m} &= 2\left\{\mathbf{\Lambda}\mathbf{b}_m - \left(1+z^{*2}\right)\mathbf{V}_m\mathbf{u} + sz^*(\mathbf{T}-\boldsymbol{\theta}) - \frac{z^*}{s}\mathbf{b}_m^\top(\mathbf{T}-\boldsymbol{\theta})\mathbf{V}_m(\mathbf{u}-\mathbf{b}_m)\right\} \\
\frac{\partial^2 \mathrm{aMSE}}{\partial \mathbf{b}_m \partial \mathbf{b}_m^\top} &= 2\left\{\mathbf{\Lambda} - \frac{z^*}{s^3}\mathbf{b}_m^\top(\mathbf{T}-\boldsymbol{\theta})\mathbf{V}_m(\mathbf{u}-\mathbf{b}_m)(\mathbf{u}-\mathbf{b}_m)^\top\mathbf{V}_m\right\} \\
&\quad - 2\frac{z^*}{s}\left\{\mathbf{b}_m^\top(\mathbf{T}-\boldsymbol{\theta})\mathbf{V}_m - (\mathbf{T}-\boldsymbol{\theta})(\mathbf{u}-\mathbf{b}_m)^\top\mathbf{V}_m - \mathbf{V}_m(\mathbf{u}-\mathbf{b}_m)(\mathbf{T}-\boldsymbol{\theta})^\top\right\}.
\end{aligned} \tag{10}$$

In each iteration t, this vector and matrix, $\mathbf{h}_t$ and $\mathbf{H}_t$, are evaluated at the current (provisional) solution $\hat{\mathbf{b}}_m^{(t-1)}$, and the new solution is defined as

$$\hat{\mathbf{b}}_m^{(t)} = \hat{\mathbf{b}}_m^{(t-1)} - \mathbf{H}_t^{-1}\mathbf{h}_t\,.$$

The iterations are terminated when the Euclidean norm of $\mathbf{H}_t^{-1}\mathbf{h}_t$ is smaller than 10^{-6}. The aMSE is evaluated at every iteration, and a warning is issued whenever its new value is higher than its value in the previous iteration. The change in the successive values of aMSE can be incorporated in the convergence criterion. The algorithm converges fast, rarely requiring more than six and never more than twelve iterations in the simulations described next and in Section 6.

5.1 Example continued

We simulate the setting of Section 4 with one auxiliary variable, equal to 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, the districts' sample sizes in the past survey by the same process as for the current survey (closely related to $N_{m,\mathrm{past}}^{0.9}$), and the population sizes in the previous year are reduced from the current year 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 district-level unemployment rates and sample sizes are plotted in Figure 7. Each district is represented by a rectangle with its centre at the current and past unemployment rates and sides 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 later in this section.

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 8. The direct estimator (S)

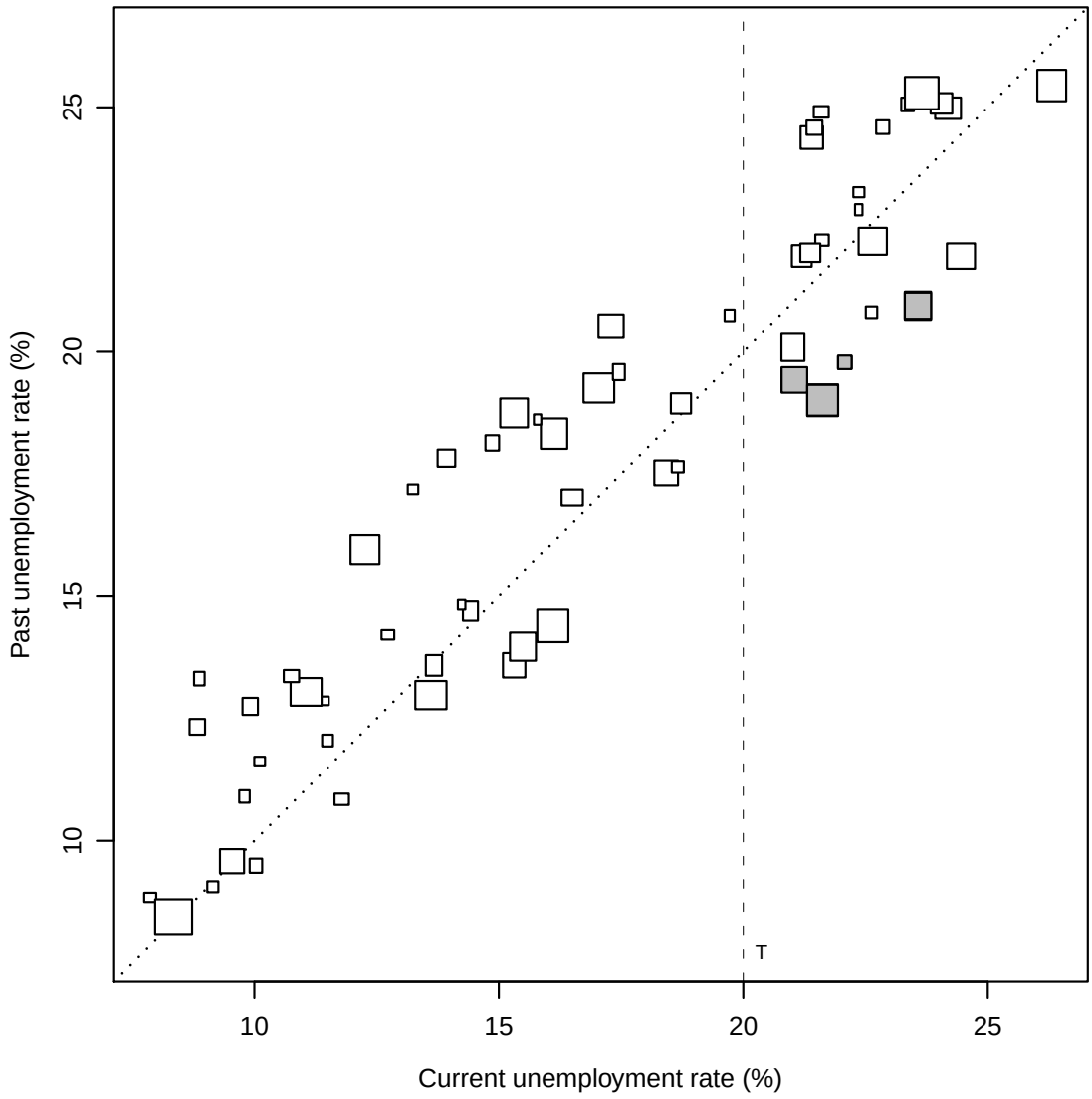

Figure 7: The district-level unemployment rates and sample sizes in the current and previous year. The sides of the rectangles are proportional to the sample sizes.

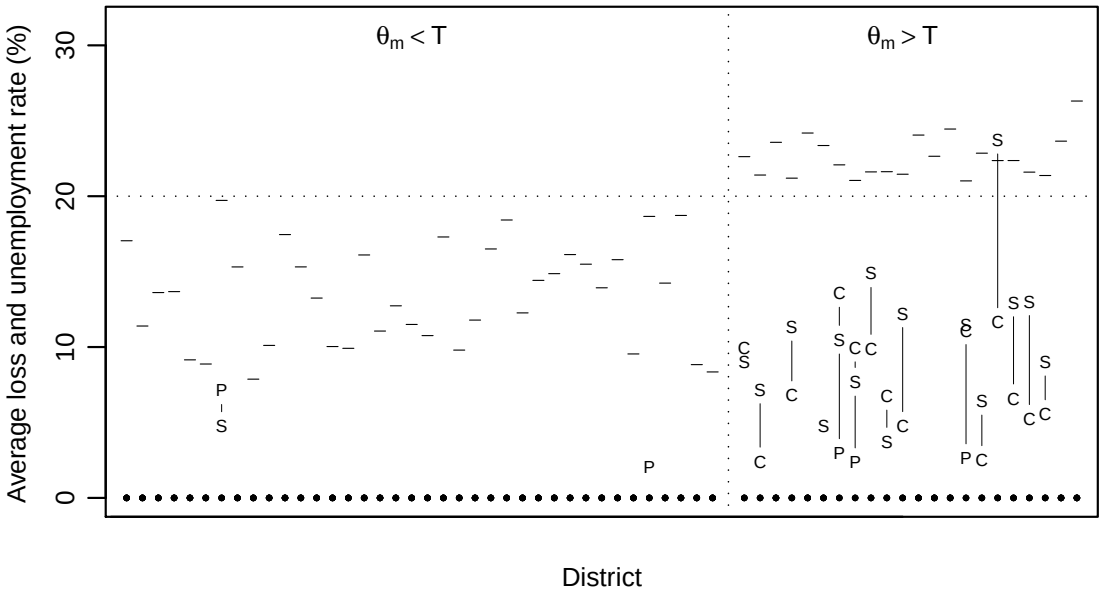

Figure 8: The empirical expected losses with the direct estimator (S), bivariate composite estimator (C, using information from the previous year) and bivariate policy-related estimator (P); quadratic kernel loss and penalty ratio $R = 10$.

has the same distribution as in the simulation in Section 4, because it does not use any auxiliary information. Some small differences between the two sets of results are present, mainly for the deserving districts (for which $\theta_m > T$), because the resampling variation of the losses for some districts is very large; the distribution for a deserving district is a mixture of (less than 50%) zeros and some large values.

The composite estimator (C) is now associated with smaller average losses than the direct estimator 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 average losses with the policy-related estimator, both the univariate version applied in Section 4, and the bivariate version which exploits the auxiliary information. The weighted-total of the average losses is 436.9 (= 20.0 + 416.9) for estimator S, 400.0 (= 6.9 + 393.1) for C, and 123.9 (38.1 + 85.8) for P; the figures in parentheses are the respective contributions from the normal and deserving districts. For estimator C, the reduction that can be attributed to the auxiliary information is by 181.9 (31%). The reduction for P, by 38.4 (24%), is more modest.

The reduction of the average loss with estimator C is not uniform among the deserving districts. For four districts, auxiliary information brings about an increase of the expected loss.

These districts are highlighted in Figure 7; one has small and the other three medium-to-large sample sizes in both surveys. Their rates in the previous year are much lower than in the current year, even after taking the national trend into account, so the auxiliary information is counterproductive (distracting), especially for the small district, for which substantial shrinkage takes place toward being a false negative. Some other districts also have rates in the previous year that deviate from the trend, but this does not cause their average losses to increase. Auxiliary information is counterproductive also for a few normal districts. However, the inflation of the losses is very small in all these cases, for both estimators C and P.

For linear and absolute kernels, estimator P remains far superior to C and S. Even though C and S are insensitive to the loss function, we evaluate the expected loss on a scale different from the quadratic kernel. With linear kernel loss and $R = 10$, the weighted-total loss for C is 124.5 (2.0+122.5), greater than for S, 116.1 (4.8+111.3); for P the loss is 40.4 (8.9+31.5). The figure for S differs from the corresponding entry in Table 2, 116.5, because it is based on a different set of replications.

For more complex auxiliary information, with several variables, the composite estimator makes only small gains, in both the values of empirical MSE and expected loss, whereas such information is detrimental to the policy-related estimator. However, the inflation of the weighted-total expected loss is only slight, and the expected losses with the composite estimator remain much higher.

6 Limited budget

Every responsible government and all its departments and programmes operate within limited budgets. In contrast, the policy-related estimator imposes no limit on the extent to which the intervention (action A) is applied. With a large penalty ratio, it prefers generating false positives, so action A is applied liberally, to many districts, with no regard for the costs of its implementation.

In the context of the previous sections, suppose a fixed overall amount of funds F has been allocated for action A in the selected districts. Suppose implementing action A in a district with labour force N_m and estimated unemployment rate $\hat{\theta}_m$ would require $HN_m(\hat{\theta}_m - T)_+$ units of funds, where H is a known constant and $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise.

That is, H is the cost pro-rated for a member of the labour force above the threshold level of unemployment, T, which should trigger action A. The units considered here (F and H), related to the cost of implementation, are different from the units associated with the losses in earlier sections, which quantify the consequences of inappropriate action (e.g., of ignoring the problems of very high unemployment). No generality is lost by assuming that $H = 1$.

If the funds are sufficient,

$$\sum_{m=1}^{M} N_m \, (\hat{\theta}_m - T)_+ \, \leq \, F \, , \tag{11}$$

then the programme is implemented as intended. Otherwise provisions have to be made, in effect, to shortchange some or all the districts that were adjudged to be in need of action A. Denote by G the funds required to implement the policy based on a set of estimates $\hat{\theta}_m$, $m = 1, \ldots, M$. We may consider any of the following options:

1. share the shortfall equally among all the districts for which action A was selected;

2. cut the expenditure by the same percentage in each district for which action A was selected;

3. raise the threshold from T to the smallest value T' for which the budget would be sufficient;

4. withdraw action A from a minimum of districts necessary for the budget to be sufficient for the rest.

Assuming known population rates θ_m, provision 1 is obtained by minimising the weighted total of the squared shortfalls, $\sum_m N_m \, s_m^2$, subject to the condition of limited budget, that is, $\sum_m N_m \, s_m = (G - F)_+$.

As soon as we contemplate provisions 1 – 4, we have to admit that the options are not merely actions A and B, but a continuum of partial implementations of action A. Therefore, we have to specify the loss associated with such an incomplete action. It is natural to associate the award of p% of the intended amount $N_m(\theta_m - T)_+$ with the (quadratic kernel) loss $Rp^2(\theta_m - T)^2$, but this choice should by no means be automatic, because even a small shortfall may be associated with a loss that is out of proportion, and the losses may differ from district to

district, not necessarily related to the district size. Establishing these factors requires another round of elicitation.

We set these issues aside and assume that the losses are proportional to the shortfall. That is, for a correctly identified positive ($\hat{\theta}_m > T$ and $\theta_m > T$), there is no loss if the amount allocated to district m, denoted by $G_m(\hat{\theta}_m)$, exceeds $N_m\,(\theta_m - T)$; otherwise the loss with action A implemented partially is

$$L_{\mathrm{A}}\left(\hat{\theta}_m\,;\theta_m\right)\left\{1-\frac{G_m\left(\hat{\theta}_m\right)}{N_m\,(\theta_m-T)}\right\}^2 .$$

If the funds are allocated inappropriately (to a false positive), the losses are reduced in the case of a shortfall, although, of course, the allocated funds would have been better spent in some deserving districts.

In the ideal implementation, action A would require a total of

$$G \;=\; \sum_{m=1}^{M} N_m(\theta_m - T)_+ \;=\; 64.55$$

units. Suppose only $F = 55.0$ units are available, so the shortfall is 9.55. In simulations, we apply all four provisions and use the auxiliary information throughout. We report the average losses only with the bivariate estimator P and quadratic kernel loss. In a replication, a typical shortfall is greater than $G - F = 9.55$, because of the liberal nature of the estimator, preferring to err on the side of false positives. The histogram of the amounts required for action A in 2000 replications with quadratic kernel and $R = 10$ is drawn in Figure 9. The vertical lines indicate the amount F available (solid line) and the amount G that is necessary for the ideal version of the policy (dashes). Only 30 values (1.5%) are smaller than G and only one of them is smaller than F. The diagram represents one component of the cost of incomplete information; in expectation, the implementation of action A based on the estimates $\hat{\theta}_m$ would be much more expensive than if all θ_m were known. The other component is due to misclassifying some districts.

The results of the simulation with the quadratic kernel loss, penalty ratio $R = 10$ and budget $B = 55.0$ are plotted in Figure 10. The symbols 1 – 4 represent the four provisions for implementing the budget constraint. We need to be concerned only with the deserving districts, which account for most of the overall loss. Black discs are drawn at height 0 for

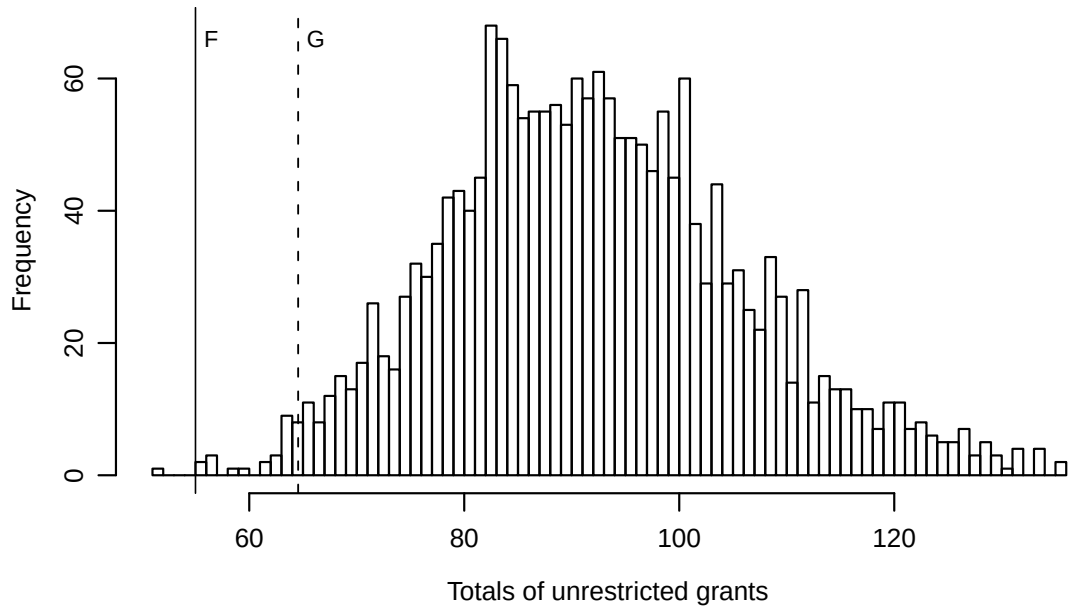

Figure 9: Empirical distribution of the total amounts $\hat{G}$ required to implement action A fully according to the policy-related estimator with quadratic kernel loss and $R = 10$.

districts that would have small expected losses if the budget were unlimited. The provisions 1 – 4 are associated with respective weighted-total average losses 510.3, 429.3, 527.2 and 772.5, compared to 123.9 if the budget were not limited. Provision 2, which might seem to be the most equitable, entails the lowest and provision 4, arguably the least equitable, the highest expected loss for all but two deserving districts that have the highest unemployment rates, 26.3% and 24.5%, and, after the capital, the highest population sizes, around 1.8 million.

If more resources were available for implementing action A, the weighted-total expected losses would be reduced. For example, with budget $F = 70.0$, they would be 350.6, 286.4, 353.1 and 526.4, each smaller by about 32% than with the budget of $F = 55.0$ units.

Increasing the size of the survey may be a more effective alternative to increasing the budget for implementing action A. If the sample sizes in the current survey were doubled in every district, without altering the sample sizes in the past survey, the weighted-totals of the expected losses would be 52.0 with no limit on the budget and 347.0, 293.2, 348.2 and 569.3 with the respective provisions 1 – 4. These values are similar to their counterparts with the original survey design and higher budget, except for provision 4, which is even poorer in relation to its alternatives. Thus, greater expenditure on the survey can be converted to more effective policy implementation.

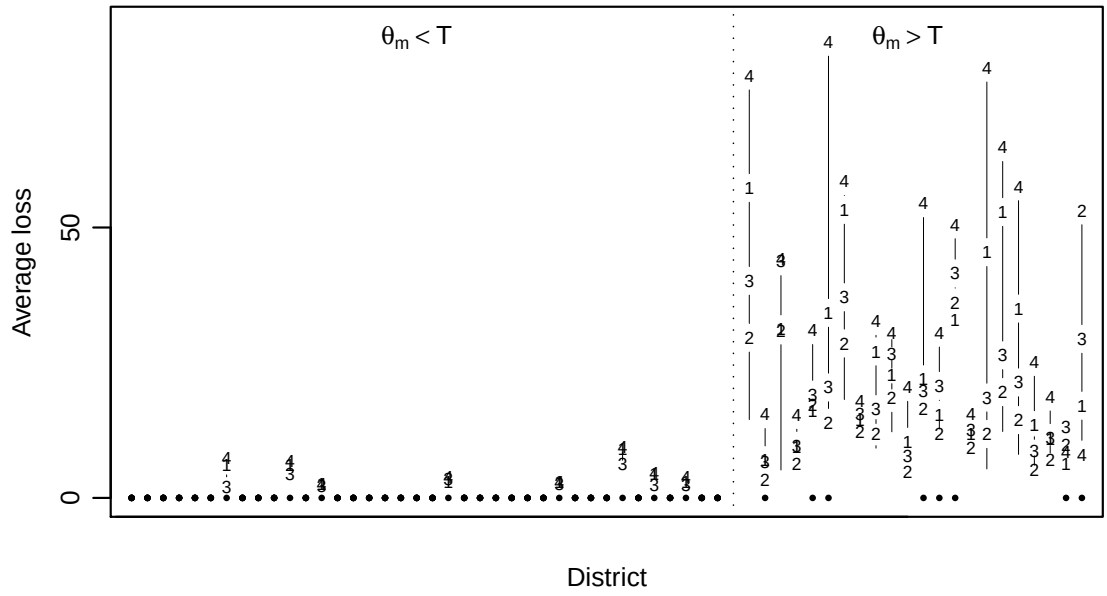

Figure 10: The empirical expected losses with the policy-related estimator with the quadratic kernel loss and penalty ratio $R = 10$, subject to budget limited to 55.0 units; 2000 replications.

With the larger survey ($n = 35\,000$), the expenditure on implementing action A has a smaller expectation and dispersion, 86.9 and 9.0, respectively, compared to 92.1 and 13.8 with $n = 17\,500$. In principle, a compromise could be found between the costs of conducting the survey and losses due to imperfect implemention of action A. In practice, this is often difficult because both activities require long-term planning and dealing with the uncertainty about the future costs and policies. Also, a typical national survey has a multitude of users (clients) whose requirements have to be satisfied.

The direct and composite estimators are uncompetitive in all the settings discussed.

7 Discussion

The policy-related estimator developed in Sections 3 and 5 and its assessment by simulations indicate that there is no single small-area estimator that is preferable to all others, because different estimators are optimal for different policies, or criteria. Shen and Louis (1998) highlight a related problem, that a nonlinear transformation of an efficient small-area estimator is not efficient for the same transformation of the target(s). They draw a similar conclusion about a nonlinear summary, such as the standard deviation (of θ_m, $m = 1, \ldots, M$), estimated naively from a set of efficient estimators $\tilde{\theta}_m$. Evaluation of small-area estimators has so far almost

exclusively focussed on the MSE criterion. We argue that this criterion should not be taken for granted and alternatives that reflect the policy objectives served by the analysis be carefully considered. Elicitation of the policy, or purpose, imposes an additional burden on the analyst and the client (the policy maker), but its outcome, a range of loss functions, enables them to tailor the analysis closely to the needs, priorities and the perspective of the client.

The simulations confirm that composite (empirical Bayes) 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 made several compromises to maintain tractability. First, we imposed the equilibrium condition, which has the flavour of unbiasedness, and then we minimised the (symmetric) averaged MSE instead of the expectation of the specified loss function. However, the gains made over the established estimators are substantial in a range of settings studied by simulations, some of them not reported here.

In the simulations, we focussed on the setting with a minority of 'positives', districts that require intervention, and assumed higher loss for false negatives than for false positives. In practice, it is unlikely that an intervention would be applied to a majority of the districts and at the same time a failure to identify a district that requires intervention would have more serious consequences than the inappropriate application of the intervention. Nevertheless, our results can be extended to such a setting.

No simulations can be conclusive for all plausible scenarios. Our simulations, conducted in R (R Development Core Team, 2009), can be easily adapted to other settings. The principal difficulty is in specifying 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 140 seconds, and one set of 2000 (bivariate) replications in Section 5.1 or 6 about 400 seconds of CPU time, so a wide range of alternative scenarios and loss structures can be explored in real time. We have found that the results are quite robust with respect to the details of how the loss functions are defined, although all these details are very distant from the mean squared loss used conventionally to assess efficiency. The direct and empirical Bayes (composite) estimators have a higher expected weighted-total loss (as well as unweighted loss) than the policy-related estimator in all the simulated scenarios, many of them not described here.

These results can be broadly interpreted as a failure of an analysis conducted in stages (stage 1 — estimation; stage 2 — assignment of action). Other examples of such failure are summarising estimated quantities (stage 1 — estimation; stage 2 — summary of the estimates), when the summary is a non-linear function, and search for a model followed by applying the estimator based on the selected model. The EM algorithm (Dempster, Laird and Rubin, 1977) explains this failure in its generality as follows. The second stage (the M-step in the EM algorithm) has to use the linear sufficient statistics in the missing data; using efficient estimates of the missing values is suboptimal. The two stages in our case are estimation and selective application of intervention (action A) based on the estimates.

We have treated the districts as isolated units and assumed that there is no interference among them. In practice, labour force as well as employers respond to government's anticipated or applied 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.

Independence of national statistical institutes, discussed extensively in the recent years (Royal Statistical Society, 2005), is often interpreted as a separation of the tasks of survey design and analysis (conducted by the institute) and interpretation and action (done by the sponsor or the client), and noninterference of the parties in their respective remits. Our development suggests that this division may lead to poor practice, because there is no single criterion for good quality of an estimator and the details of the intended policy have to inform the construction of the estimator. Thus, a single data source (a survey) may yield two different sets of high-quality estimates of the same set of targets, for two distinct purposes (clients). This is not a problem with the estimators of most national quantities, because they have sampling variation that can for most purposes be ignored. However, small-area estimators usually have non-trivial sampling variation. Borrowing of strength by empirical Bayes and related methods reduces it somewhat, but not always to the level at which it could be ignored. In fact, we have found that such shrinkage is detrimental for the purpose and, in some cases, it has to be applied in a different direction, and toward a different focus.

References

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* McGraw Hill, New York.

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

Efron, B., and Morris, C.N. (1972). Limiting the risk of Bayes and emprirical 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* Ser. B **69**, 221–238.

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

Lindley, D.V. (1992). Is our view of Bayesian statistics too narrow? In Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. (Eds.) *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meting.* Clarendon Press, Oxford, UK; pp. 1–15.

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

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

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

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.

Prasad, N., and Rao, J. N. K. (1990). The estimation of mean-squared errors of small-area

estimators. *Journal of the American Statistical Association* **80**, 163–171.

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

Robbins, H. (1955). An empirical Bayes approach to statistics. In Neyman, J. (Ed.) *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 157–164. University of California Press, Berkeley, CA.

R Development Core Team. (2009). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Royal Statistical Society (2005). Vision for National Statistics. `http://www.rss.org.uk/main.asp?page=2616`. Retrieved on 11th Jan. 2011.

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* Ser. B **69**, 238–257.

CEPS
INSTEAD

3, avenue de la Fonte
L-4364 Esch-sur-Alzette
Tél.: +352 58.58.55-801
www.ceps.lu