STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 65–82

POLICY-ORIENTED INFERENCE AND THE ANALYST-CLIENT COOPERATION. AN EXAMPLE FROM SMALL-AREA STATISTICS

Nicholas T. Longford[1]

ABSTRACT

We show on an application to small-area statistics that efficient estimation is not always conducive to good policy decisions because the established inferential procedures have no capacity to incorporate the priorities and preferences of the policy makers and the related consequences of incorrect decisions. A method that addresses these deficiencies is described. We argue that elicitation of the perspectives of the client (sponsor) and their quantification are essential elements of the analysis because different estimators (decisions) are appropriate for different perspectives. An example of planning an intervention in a developing country's districts with high rate of illiteracy is described. The example exposes the deficiencies of the general concept of efficiency and shows that the criterion for the quality of an estimator has to be formulated specifically for the problem at hand. In the problem, the established small-area estimators perform poorly because the minimum mean squared error is an inappropriate criterion.

Key words: composition, empirical Bayes, expected loss, borrowing strength, exploiting similarity, shrinkage, small-area estimation.

1. Introduction

Survey methods have in the recent decades been greatly stimulated by the big-budget departments of national governments, such as social security, health care, education and employment, owing to their greater appreciation of the role of statistical information and inference in policy making. Developments in small-area estimation have responded to the demand for greater detail about the (administrative) divisions of a country, such as regions and districts. In a typical setting, a national survey is conducted, collecting information about the key variables, such as employment status, and an established set of background variables (age, sex, educational level, marital status, and the like), and estimates related to a key variable

[1] SNTL and Universitat Pompeu Fabra, Barcelona, Spain. E-mail: sntlnick@sntl.co.uk

are sought for each district. The districts are of varying sizes, and some of them are represented in the survey by small samples that on their own are not sufficient for estimating their key characteristics, usually percentages (e.g., the unemployment rate), with any appreciable precision.

The main advance in small-area estimation is the exploitation of similarity, also referred to as borrowing strength, in any aspect for which (auxiliary) data is available. Obvious examples of such data (and information) are values of the target variable observed in the other districts of the country, the background variables recorded in the survey, and information obtained from censuses and surveys. Somewhat less appreciated is the potential of variables prima facie related to the target variable and information from the previous years of the surveys in the same programme. For estimating subpopulation characteristics (e.g., for minorities), other subpopulations (or the complementary subpopulation) are often very effective auxiliaries. See Longford (2005) for examples. In particular, insisting on having the values of auxiliary variables for the entire population is extremely restrictive (Elbers, Lanjouw and Lanjouw, 2003).

In small-area estimation, as in other survey inference, efficient estimation is generally regarded as superior. A lot of the theory is concerned with deriving estimators that are efficient, or nearly so, sometimes in uncongenial circumstances, using models known not to be valid. Robust estimation is therefore regarded as invaluable. Estimation of standard errors of these estimators is also an important preoccupation. This research implies, often without a clinical statement to that effect, that the estimates obtained are best suited for a policy related to the target variable and that it will be well informed by efficient small-area estimators. The statistical analysis is concluded by the presentation of the estimates and the associated standard errors, and the remainder of the analysis is left to be dealt with by the policy maker. We show by example that this is a poor strategy and argue that the analytical skills of a statistician and the insight and other qualities of the policy maker have to be integrated much more closely.

The core of the problem is that the consequences of the (estimation) errors made have to be taken into account because they are in substantial discord with the default assumption of the (symmetric) quadratic loss. These consequences are difficult to elicit from experts and to quantify them, but that is hardly an excuse for applying methods that assume a particular structure of such consequences, especially when the assumed structure is in obvious discord with the client's perception. The analyst may not be aware of the consequences and of their relevance, and therefore would

not communicate the (default) assumptions to the client. It is essential to sensitise the client to this issue because the solution to the problem is statistical, and there is a danger that when the client becomes aware of the issue he or she will seek a second-rate ad hoc solution not integrated with the original (incomplete) analysis.

In the next section, we outline the policy planned in an application and describe a survey that is intended to inform the policy. The policy is related to combating illiteracy in the districts of a developing country. Illiteracy is regarded as a major barrier to gaining employment and to economic development in general, but also to the spread of government information and, admittedly, of the governing party's political propaganda as well. In Section 3, we review the established methods for small-area estimation and highlight their deficiency which is then addressed in Section 4 by the proposed estimator, called *policy-related*. In Section 5, we compare by simulations the implementations of the policy using three estimators:

- direct estimators that no one would recommend (for small districts);
- composite (empirical Bayes) estimators that are generally regarded as superior;
- policy-related estimators constructed with the intent to minimise the total expected loss.

We show that for a wide range of perspectives and priorities of the decision maker the policy-related estimator is far superior to the other two estimators. The concluding section discusses the implications of this result on how small-area analysis should be conducted, and extends them to some general principles, including how official statistical institutes should operate.

Direct estimators use no auxiliary information; they are based only on the data for the target variable and the region concerned. Often they are standard survey-methods estimators (see, e.g., Särndal, Bengtsson and Wretman, 1992) restricted to the region. Empirical Bayes estimators exploit auxiliary information by means of a two-level regression model (Goldstein, 2002; Rao, 2003) and composite estimators (Longford, 1999 and 2005) combine direct and auxiliary estimators (or exact quantities) without a reference to a model. The policy-related estimator is developed in Longford (2013, Chapter 7, and 2015), where some technical details omitted from this article can be found. The method exploits auxiliary information, but requires also input about the consequences (ramifications) of the errors that may be committed. These errors are:

- *false negative*: failure to apply an intervention when it should have been applied;
- *false positive*: applying the intervention when it is not necessary.

Our interest is in settings in which these consequences are uneven. In the example of combating endemic illiteracy, a false negative has much more serious consequences (greater losses) than a false positive. Empirical Bayes methods and, more generally, methods that aim to minimize the mean squared error (MSE), are oblivious to such consequences.

Similar issues arise in medical screening (Longford, 2013, Chapter 6), where a false positive (incorrect labelling as diseased) is regarded as less serious an error than a false negative (failure to discover the disease), in (production) quality control, where a false claim of satisfying a standard is much more costly (in the long term) than the pursuit of further (unessential) improvements when the standard has already been achieved, and in the operation of warning systems (for epidemics, natural disasters, military or terrorist attacks, and the like), where false warnings may erode the credibility of the system, but a failure to warn is an unmitigated disaster.

2. Illiteracy in the districts of a country

We consider a developing country with an adult population (aged 16 or over) approaching 40 million and adult illiteracy rate of about 18%. The country has 72 districts, of population sizes (numbers of adults) varying from 50 000 to 1.8 million (the capital). Illiteracy tends to be more prevalent in rural districts. Its rates are smaller in the most populous districts which are mostly urban (large cities and their environs). However, some less populous districts are formed by single (smaller) towns and cities, and there the illiteracy rates tend to be smaller. Some of these towns are satellites of larger cities.

The Ministry of Education has appropriated funds for conducting a survey to study the illiteracy rate in the country. The results would then be used for implementing a particular policy aimed at the districts with illiteracy rates higher than 25%. The survey has several sponsors and subscribers with purposes different from small-area estimation, and so a compromise stratified sampling design is implemented, with the districts as the strata. Some clustering is applied within the strata, which is of marginal interest for our purposes, and its details are omitted. It is impossible to ensure that each stratum (district) would have a sample size sufficient

for reliable direct estimation of its illiteracy rate.

Denote by θ_d the rate of illiteracy in district $d = 1, \ldots, D$, and by $\hat{\theta}_d$ and $\tilde{\theta}_d$ the respective direct and composite estimators of θ_d. The composite estimator is defined as the convex combination of the direct and (overall) national estimator,

$$(1 - b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}\,, \tag{1}$$

where the coefficient $b_d = 1/(1+n_d\,\omega)$ involves the ratio $\omega = \sigma_{\mathrm{B}}^2/\sigma^2$ of the between- and within-stratum variances and n_d is the sample size of the stratum. When the sampling weights are not constant within strata, n_d has to be replaced by the effective sample size. The variance σ^2 is estimated by pooling the within-district estimates of the variance and σ_{B}^2 is estimated by moment matching applied to the sum-of-squares statistic $\sum_d (\hat{\theta}_d - \hat{\theta})^2$.

As part of the policy, an intervention is designed and planned to be applied in districts in which $\theta_d > 0.25$. We assume that it will be applied to districts with $\hat{\theta}_d > 0.25$ or $\tilde{\theta}_d > 0.25$, or to districts that satisfy this inequality for a different estimator. At the beginning of the author's involvement, all parties involved agreed that the composite estimator $\tilde{\theta}_d$ should be used. The Research Department of the Ministry agreed that a simulation study would be conducted, mainly to assess the potential problems with districts that have extreme rates of illiteracy. The related methodological issue is discussed in Longford (2007), and it concerns mainly estimation of the standard errors and the claim that empirical Bayes and composite estimators are superior to direct estimators for every district. Also, the funding for the survey and the intervention come from the same budget, and so its split for the survey (data collection and analysis) and policy implementation was negotiated extensively, until it was agreed that a simulation study would inform this issue.

The simulation study, and the detailed discussion of its set-up, including the information on which it would be based, as well as the arguments about how to evaluate the results, brought to the fore the purpose of the survey, namely, allocation of funds to the districts. Here the established criterion of minimum MSE turned out to be irrelevant because in the Ministry's perspective, the evaluation should focus on the two types of error (false negative and false positive) in classifying the districts as

- deserving the intervention ($\theta_d > 0.25$), and
- not deserving the intervention ($\theta_d < 0.25$).

At first, this might suggest that hypothesis testing (HT) should be applied. However, this was also dismissed after a set of simulations when it became clear that HT is oblivious to the consequences of the two kinds of error. The following example disqualifies HT from almost any problem in which we have to decide how to proceed; as if the (null) hypothesis were valid, or not. Suppose incorrect omission of a district from intervention is five times as serious an error as incorrect inclusion. By a hypothesis test, with the conventional level of significance of $\alpha = 0.05$, we would come to a particular decision. Next, suppose incorrect omission is 50 times more serious than incorrect inclusion. With conventional statistical tools, and conventional operational mindset, we would apply the same hypothesis test, and come to the same decision. No procedure that always comes to the same conclusion in these two settings could possibly be appropriate. A decision (choice between two complementary options) is poorly founded if it is not influenced by the consequences of the two kinds of incorrect choices, unless the decision is always correct (or always incorrect) and entails no uncertainty. In the general problem of estimation, a similar dismissing argument is easy to formulate. The consequences of the errors that are committed in estimation have to be an important factor in how an estimator is constructed. Ignoring them is a licence for making poor decisions and, ultimately, rendering the statistical analysis irrelevant.

Eliciting information about the consequences of errors is, unfortunately, not a mainstream statistical activity. In practice, it can be surprisingly contentious, because many clients believe that 'it is all in the data', and the analyst's task is to process the data and deliver an unambiguous verdict. Also, a client may suspect that the analyst wants to elicit the client's perspective, priorities and goals, merely to fix up the results so as to superficially please the client, or to obtain confidential information that would later be disclosed to the client's detriment. In the political and civil-service sphere, the value of information is well appreciated, but often leads to the practice of divulging it on a strictly need-to-know basis. This involves liberally placed controls and requirements for approvals (hurdles) across the layers of management that hinder and sometimes entirely disable the process of informing the statistical analysis. A change in the perspectives and priorities may appear as an embarrassment to the client. However, it is a responsible act when it responds appropriately to new information and circumstances. Elicitation is more commonly considered for prior distributions in Bayesian estimation (Garthwaite, Kadane and O'Hagan, 2005). The same principles apply to elicitation in our context, although much less of the experience is recorded in the literature.

An important element of elicitation is to put the clients (or experts) at ease by not rushing them to any quick decisions (or an uneasy consensus), and explaining that they are not expected to have answers ready at a moment's notice. In the example of combating illiteracy, the key question relates to the so-called *penalty ratio R* which quantifies how many times more costly is an error of one kind than of the other. There is no need to conclude with a single value R. It is more constructive to set (or declare) a plausible range of values of R. The key property of such a range is that any value outside it can be ruled out—that the client is satisfied that such a value is not realistic. Of course, it is advantageous to have as narrow a plausible range as possible, but its plausibility is an imperative. The wider the range, the greater the threat of an inferential *impasse*, when both decisions are plausible; one for certain plausible values of R and the other for the complement.

The rationale for reflecting in inferential statements different perspectives and priorities is hinted by Shen and Louis (1998). They coined the term 'triple-goal estimator' to highlight the need for different estimators for three distinct purposes in small-area estimation: estimating each district's population quantity, ranking the districts according to this quantity, and estimating the district-level distribution of these quantities. A compact summary of their conclusion is that efficiency and unbiasedness (of an estimator) are *fragile* properties. Fragility refers to the fact that these properties are not maintained by non-linear transformations, and even less so by some discontinuous ones, such as assessing whether a parameter is greater or smaller than a set threshold. We can paraphrase these conclusions by saying that optimality of an estimator is conditioned on the scale used for the error in estimation. By the same token, an estimator with minimum MSE may be far from optimal with a distinctly asymmetric loss function. Loss function is as important an input and has a similar nature as (informative) prior distribution in Bayesian analysis, where it is taken for granted that different priors lead to different posteriors and conclusions based on them. On this account, we have to dismiss the idea that the results of a respectable analysis, even in a frequentist paradigm, have to be 'objective' and applicable to a wide range of perspectives. Instead, they have to be client-specific, that is, responsive and sensitive to the client's perspective and value judgements.

3. Shrinkage estimation

Empirical Bayes estimators (Robbins, 1995; Carlin and Louis, 2000) are a general example of shrinkage estimators. Composite estimators (Longford, 1999) are also

shrinkage estimators; in fact, they apply shrinkage directly, without the intermediation of a model, and thus absolve the analyst from the responsibilities related to the validity of the model, including the distributional assumptions. In small-area estimation, this feature is important because the analysts rarely have the freedom to choose what auxiliary information will be used; they have to operate with what is available and can neither present an excuse nor apportion the blame to anybody when the model is assessed not to fit well.

Shrinkage estimators pull a direct estimator $\hat{\theta}_d$ for district d toward a (national) focus $\hat{\theta}$; see equation (1). The amount of shrinkage (the strength of the pull) depends on the balance of the sampling variance of $\hat{\theta}_d$, equal to σ^2/n_d in the simplest setting, and the district level variance $\sigma_B^2 = \mathrm{var}_d(\theta_d)$. The latter variance is defined for the districts and is not related to sampling; it is a population quantity. If σ_B^2 is very small, then the shrinkage is strong because the districts are very similar and the national estimator $\hat{\theta}$ is very useful for every district. In contrast, if σ_B^2 is very large, much less shrinkage takes place because θ may be far away from θ_d, and therefore $\hat{\theta}$ a poor estimator of θ_d; this is the case for a substantial fraction of the districts. Further, more shrinkage takes place for small districts (districts with large sampling variance of $\hat{\theta}_d$) and less for districts with more precise estimators $\hat{\theta}_d$.

Although this interpretation applies only to the simplest form of (univariate) shrinkage estimation, it offers some insights as to why it may be poorly suited for the Ministry's task. If a false positive has less serious consequences than a false negative then we should focus on the deserving districts, for which only errors of the former type are possible. In our example, these districts are in a minority because the threshold of $T = 25\%$ is far greater than the national rate of about 18%. Direct estimation for the deserving districts will result in an error if the estimation error is negative and $\hat{\theta}_d < 0.25 < \theta_d$. Positive or small negative estimation errors have no consequences because then both $\hat{\theta}_d$ and θ_d exceed the threshold. The likelihood that $\hat{\theta}_d$ is close to the national rate, $\theta \doteq 0.18$, is quite small for most deserving districts, because an error smaller than (more negative than) -0.07 is quite rare.

Shrinkage applied to the direct estimator in (1) pulls it toward the national rate, and therefore reduces it for nearly all deserving districts. As an aside, note that this implies that empirical Bayes estimation for a deserving district is biased, contrary to the acronym EBLUP (empirical Bayes linear *unbiased* predictor) used in the context of hierarchical linear models. There may be deserving districts with $\hat{\theta}_d > T > \tilde{\theta}_d$, for which direct estimation would lead to the appropriate decision (intervention), but the shrinkage estimation would yield the inappropriate decision. The opposite,

$\hat{\theta}_d < T < \tilde{\theta}_d$, is less likely to happen, because the focus of shrinkage is $\hat{\theta}$, an efficient estimator of θ, and θ is much smaller than T. Thus, efficient estimation contradicts good policy implementation. This is in accord with a clinical proposal described in Longford (2013, Chapter 7), in which shrinkage is applied, but with a different focus, and to an extent different from the empirical Bayes shrinkage.

4. Policy-related estimation

Before describing the proposed estimator, we give a minimum background to decision theory, which motivates it. Suppose our target is a quantity (parameter) ν, and let $\hat{\nu}$ be an estimator of ν. The estimator is unlikely to be without error; $\Delta\nu = \hat{\nu} - \nu \neq 0$. The conventional criterion for 'good' (efficient) estimation, the MSE, assigns the cost of $\Delta\nu^2 = (\hat{\nu} - \nu)^2$, and the efficient estimator is defined as the one that minimises the expectation $\mathrm{E}(\Delta\nu^2)$.

Suppose the cost is not symmetric. A simple adaptation of MSE is that the cost is $(\hat{\nu} - \nu)^2$ if $\hat{\nu} > \nu$, but it is $R(\hat{\nu} - \nu)^2$ if $\hat{\nu} < \nu$. That is, given a fixed absolute error $|\Delta\nu|$, understatement is R times more costly than overstatement. The penalty ratio R is positive, but may be smaller than unity. No generality is lost by having a factor (R) with only one of the squared errors, because multiplying both error functions by the same constant does not alter the nature of the costs; only their relative size matters, and it is very convenient that their ratio is a constant (R). An estimator may be optimal for $R = 1$, that is, for MSE as the criterion, but then it is not optimal for $R = 10$, nor for $R = 0.1$.

Further, suppose we are not interested in the value of ν as such, but merely want to establish whether ν is greater or smaller than a threshold T. In this setting, estimation is associated with no error if $\hat{\nu}$ and ν are on the same side of T—if both $\hat{\nu}$ and ν are greater than T, or both are smaller. Suppose one unit of loss is incurred if $\nu < T$ but $\hat{\nu} > T$ (incorrect inclusion, in the context of our study) and R units are lost if $\nu > T$ but $\hat{\nu} < T$ (incorrect omission).

This setting resembles HT, but if we wanted to apply it we would not know which case ($T \leq 0.25$ or $T \geq 0.25$) to declare as the hypothesis and which as the alternative. Even if we resolved this issue, we would not know how to act when the hypothesis is not rejected because in that case the hypothesis has not been accepted, merely we would have failed to find evidence against it. On the one hand, we are well aware of this; on the other hand, we liberally abuse this wisdom because the correct conclusion that we are ignorant about the relation of ν and T is unacceptable.

A hypothesis test can provide evidence for its alternative, by concluding that there is a probabilistic contradiction with the hypothesis. But in the absence of such a contradiction, it does not provide any evidence for the hypothesis. Continuing the analysis, a business agenda, or any other plan as if the hypothesis were valid, where there is no support for it, is a common logical inconsistency that does no favours to the image of any scientist.

In the decision-theoretical approach (Lindley, 1985; DeGroot, 2004), we evaluate the expected loss with the two options we have, to conclude A, that $\nu < T$, or B, that $\nu > T$, and choose the option that is associated with smaller expected loss. The evaluations are somewhat more complex than in HT, but they are a small price to pay for better allocation of our own money (assuming that we all are taxpayers) or, in general, for tailoring the solution closer to the clients' perspective, priorities and goals.

The policy-related estimator is developed from the following considerations. Let

$$\tilde{\theta}_d^* = (1 - b_d)\,\hat{\theta}_d + b_d F_d\,, \tag{2}$$

where b_d is the shrinkage coefficient and F_d is the focus of shrinkage, set separately for each district. In empirical Bayes and composite estimation, $F_d \equiv \hat{\theta}$; see (1). In our problem, one might contemplate $F_d \equiv T$. Both proposals lead to poor solutions. The coefficients b_d and F_d are determined by two conditions:

- *equilibrium at* T: the choice between options A and B is immaterial for a district with $\theta_d = T$;
- minimum MSE.

The solution is derived in the Appendix.

The second condition is somewhat out of line with our general arguments, and is included to obtain a unique solution that is tractable. Thus, our proposal is not optimal; we have only empirical evidence that it is far superior to both direct and composite estimation. Further, decisions based on composite estimation are poorer than on direct estimation.

5. Simulations

For a simulation study, we form a computer version of the country, with its districts and within-district rates of illiteracy. We define a sampling design: stratified simple random sampling with the districts as the strata and sampling fractions slightly

greater in the smallest districts than in the largest, with some variation to make the setting more realistic. The overall sample size is 20 000, and the average within-district sample sizes are in the range 21 – 1030. In the sampling design, these sample sizes are not fixed and involve some moderate randomness. The population rates of illiteracy within the districts are in the range 2 – 29%, set by a random process in which some prior information is used. These rates are fixed across replications. The rates are smaller in the largest and some of the smallest districts, and are highest for a few mid-size districts. The population is fixed (not altered) in the replications of a simulation study; the sampling process is the sole source of randomness. However, we conduct a large number of simulation studies, with different populations that are plausible in the considered setting, to check that the findings are replicated across studies.

A replication of the simulation comprises drawing a sample from the (fixed, artificially generated) population, and applying the three estimators, direct (D), composite (C) and policy-related (P). The errors of the two kinds are then summarised for each estimator (applied in 72 districts) and the losses added up. The population size of a district is reflected in these summaries by multiplying the loss, when an error in classifying the district is committed, by its population size (in millions). From $M = 1000$ replications, we obtain 1000 triplets of losses and compare their averages. This exercise is repeated for several values of R, which influence only the policy-related estimator; the direct and composite estimators do not depend on R. However, R is a factor in evaluating the average loss even for the direct and composite estimators. More detail can be obtained by evaluating the average losses within the two groups of districts: those deserving the intervention (15 districts) and the complement (57 districts).

Table 1 presents the results of one set of 1000 replications. For each replication, we record the numbers of districts with errors of the two kinds, the total population in them and the total of the losses scaled by the district-level population sizes. These summaries are evaluated for the three estimators and several penalty ratios R, indicated at the left-hand margin. For example, for $R = 10$, the policy-related estimator (P) generates false negatives (F–) for 1.484 districts on average (out of 15) and false positives (F+) for 12.346 districts (out of 57), that is, 13.830 in total. Their respective populations (numbers of adults) are 0.278 and 4.419 million on average, 4.697 million in total. Judging by these two totals, the policy-related estimator is inferior to both the direct and composite estimators, which have errors in approximately $4.4 + 5.3 = 9.7$ and $7.4 + 2.4 = 9.8$ districts, respectively, both involving

only 2.9 million adults. However, on the criterion of expected loss, the direct and composite estimators are far inferior to the policy-related estimator. The former two have expected losses 0.495 and 0.711, respectively, whereas the expected loss of the policy-related estimator is only 0.283. The policy-related estimator has greater expected loss on false positives, but the other estimators have excessive expected losses on false negatives.

Table 1: The average number of districts, population and the loss associated with incorrect decisions; 1000 replications.

		Districts		Population		Loss		
R	Estimator	F–	F+	F–	F+	F–	F+	Total
1	P	4.380	5.312	1.035	1.923	0.040	0.057	0.097
	D	4.395	5.295	1.029	1.899	0.043	0.062	0.105
	C	7.387	2.447	1.649	1.247	0.068	0.029	0.097
5	P	1.926	10.072	0.418	3.581	0.077	0.139	0.216
	D	4.395	5.295	1.029	1.899	0.217	0.062	0.279
	C	7.387	2.447	1.649	1.247	0.341	0.029	0.370
10	P	1.484	12.346	0.278	4.419	0.097	0.186	0.283
	D	4.395	5.295	1.029	1.899	0.433	0.062	0.495
	C	7.387	2.447	1.649	1.247	0.682	0.029	0.711
25	P	0.704	15.162	0.138	5.532	0.123	0.256	0.379
	D	4.395	5.295	1.029	1.899	1.083	0.062	1.145
	C	7.387	2.447	1.649	1.247	1.705	0.029	1.734
50	P	0.606	17.314	0.102	6.414	0.162	0.317	0.479
	D	4.395	5.295	1.029	1.899	2.165	0.062	2.227
	C	7.387	2.447	1.649	1.247	3.410	0.029	3.439

Notes: The estimators are: P: policy-related; D: direct; C: composite. The components of loss are: F–: false negative; F+: false positive.

For greater penalty ratios R, the expected loss of the policy-related estimator is greater, but for the other two estimators it increases at a much faster rate. For $R = 50$, their expected losses are 0.47, 2.23 and 3.44. For $R = 1$, the policy-related and composite estimators have the same expected loss, 0.097, and the direct estimator has a slightly higher expected loss, 0.105. However, even for slightly higher R, the policy-related estimator has the smallest expected loss, followed by the direct and

composite estimators. Note that $R = 1$ is not equivalent to MSE, because no loss is incurred when the appropriate decision is made, even when the estimate differs from the target substantially, so long as it is in the direction in which the decision is not altered.

In the implementation in the statistical language for computing and graphics R (R Core Development Team, 2012), such a simulation takes about 10 seconds, and so a wide variety of settings can be explored. In some situations, fewer than 1000 replications would suffice for comparisons with a high level of certainty, but the saving in computing is negligible. The key to efficient processing of the results is a partially automated assessment of the results and their compact tabular and graphical presentation.

In particular, the simulations can be re-run with different sample sizes, to explore whether the Ministry's funds could be allocated to the conduct of the survey and the implementation of the programme more effectively. A greater sample size requires higher expenditure, but the allocation of the remainder is then associated with smaller expected losses. It turns out that a sample size around $n = 35\,000$ would lead to a near-optimal split of the resources, although the calculation involved is based on some further assumptions which make some of the existing assumptions more onerous.

The issue of possibly insufficient funds can be explored similarly. Although it uncovers some weaknesses of the policy-related estimator, it remains far superior to its established competitors. The problem is that by erring on the side of inappropriate inclusion in the programme, the expenditure on its implementation increases, and the awards to the selected districts have to be reduced.

The composite and policy-related estimators have their multivariate versions in which auxiliary information from other variables is exploited. In further simulations, we generated such information as the same survey from the previous year. The bivariate composite estimator has substantially smaller expected loss than the univariate composite estimator, but it is still inferior in some settings to the direct estimator. The improvement of the policy-related estimator, based on a multivariate version of the shrinkage in equation (2), is somewhat more modest, but it remains far superior to the direct and composite estimators, except for R very close to unity, such as $R = 1.05$. Even when we get the value of R wrong, say by 50% in either direction, that is, we conduct the estimation with R, but evaluate the losses with $1.5R$ or $R/1.5$, the policy-related estimator is superior to the direct estimator, although the expected losses are appreciably greater than they would be with the assumed

value of R. Thus, a modicum of uncertainty about R is acceptable, but there are obvious rewards for its more precise specification by a narrower plausible range.

In summary, Table 1 shows that no single estimator or method is superior by all criteria, so it is essential to specify the criterion that best describes the perspective and priorities of the client. The penalty ratio R is a key quantity in this regard. The policy-related estimation allows some leeway, and it suffices to declare a range of plausible values of R. In the application, this range was set to $(20, 30)$, and the survey design was based on $R = 25$. In more detailed exploration of the results, we may find strong points of the composite estimator but on the principal criterion of minimum expected loss it is a total failure; it is inferior even to the direct estimator.

6. Analysts and clients

National statistical institutes and their principal clients, the national government departments and agencies, have over the recent decades negotiated a code of conduct that would ensure the efficient functioning of the institutes without the client exercising any undue influence on the outcomes of the assigned tasks. Transparency and unbiasedness of the institutes, sometimes interpreted as absence of any political influence, are highly prized by the public and are generally regarded as essential elements of the proper use of statistics by government agencies. This mode of operation is appropriate for the production of inferential statements that are effective without any uncertainty, such as national unemployment rates (based on national Labour Force Surveys), consumer price indices, and the like.

Our example of planning an intervention in the districts, for which the uncertainty about key quantities is nontrivial and has to be reckoned with, indicates that such a detached mode of cooperation between the client and the analyst is not conducive to good practice of statistics. Much closer cooperation and integration of the two sets of activities, analytical and decision making, is required. The process of elicitation implies such an integration, even though it is contrary to the current trends in which transparent detachment of the two parties is paramount. The analyst has to be privy to the details of how the estimates or other inferential statements are going to be applied, and under what conditions, to actually choose an appropriate estimator and, more generally, a format of the inferential statement and the assessment of its quality (the expected loss).

The argument that the client could deal with the decision-theoretical aspects of the inferential task without the analyst's assistance does not hold water. These

evaluations entail nontrivial optimisation that is firmly in the remit of computational statistics, and for which the national institutes are, or should be, equipped much better than the client. These evaluations cannot be replaced by postprocessing the results obtained by established methods and presented in a standard format.

We see the resolution of this problem in altering the professional ethos in statistics and bringing it much closer to the standard of the (corporate) legal profession. Their standard of 'representing the best interests of the client' should be translated to the statistical profession as

> representing the best interests as regards data and information, their collection and all intermediate processes leading to decisions based on them.

Transparency is not disregarded in this process, because the loss functions, the quantified versions of the government priorities and perspectives, can and should be declared openly, as a matter of course by a transparent government. Uncertainty about them is not necessarily a sign of poor management or lack of control over the processes in the remit of the client (the government). On the contrary, it may be an indication of its integrity. Denial of such uncertainty is a sign of poor understanding of its relevance in statistical inference, sometimes combined with misplaced concerns about the false image of the government as an omniscient body.

Acknowledgements

This paper was presented at the First South European Survey Methodology Conference in Barcelona, Spain, in December 2013. By agreement with the author, the identity of the client (country and its government) cannot be disclosed. The person representing the client has approved the text of this paper. His and his colleagues' cooperation and understanding for the authors' academic perspective are appreciated. Constructive comments of the journal referees are acknowledged.

REFERENCES

CARLIN, B. P., LOUIS, T. A. (2009). *Bayes and Empirical Bayes Methods for Data Analysis*, 3rd ed., Boca Raton, FL: Chapman and Hall/CRC.

DEGROOT, M. H. (2004). *Optimal Statistical Decisions*, New York: McGraw-Hill.

ELBERS, C., LANJOUW, J. O., LANJOUW, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355–364.

GARTHWAITE, P. H., KADANE, J. B., O'HAGAN, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680–701.

GOLDSTEIN, H. (2002). *Multilevel Statistical Models*, 3rd ed., London: E. Arnold.

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

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

LONGFORD, N. T. (2005). *Missing Data and Small-Area Estimation*, New York: Springer-Verlag.

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

LONGFORD, N. T. (2013). *Statistical Decision Theory*, New York: Springer-Verlag.

LONGFORD, N. T. (2015). Policy-related small-area estimation. *South African Journal of Statistics* 49, 105–119. Also available as Working Paper 1427, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona, Spain, http://www.econ.upf.edu/en/research/onepaper.php?id=1427.

R CORE DEVELOPMENT TEAM (2012). R: *A language for statistical computing and graphics*, Vienna, Austria: Foundation for Statistical Computing.

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

ROBBINS, H. (1995). An empirical Bayes approach to statistics. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 157–164.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

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.

Appendix

This appendix derives the policy-related estimator. Suppose $\hat{\theta}_d$ is normally distributed with expectation γ_d and variance v_d^2; $\hat{\theta}_d$ may be biased for θ_d. Denote $\tilde{z} = (\gamma_d - T)/v_d$ and $\tilde{z}^\dagger = (\gamma_d - \theta_d)/v_d$. We use ϕ and Φ for the density and the distribution function of the standard normal distribution.

For the piecewise constant loss, the expected losses for the false positives (when $\theta_d < T$) and false negatives (when $\theta_d > T$) are

$$\begin{aligned} Q_+ &= \frac{1}{v_d}\int_T^{+\infty} \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y = \Phi(\tilde{z}_d) \\ Q_- &= \frac{R}{v_d}\int_{-\infty}^{T} \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y = R\left\{1-\Phi(\tilde{z}_d)\right\}, \end{aligned}$$

respectively. The corresponding identities for the piecewise quadratic loss function are

$$\begin{aligned} Q_+ &= \frac{1}{v_d}\int_T^{+\infty} (y-\theta_d)^2 \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y \\ &= v_d^2\left\{\left(1+\tilde{z}_d^{\dagger 2}\right)\Phi(\tilde{z}_d) + \left(2\tilde{z}_d^\dagger - \tilde{z}_d\right)\phi(\tilde{z}_d)\right\} \\ Q_- &= \frac{R}{v_d}\int_{-\infty}^{T} (y-\theta_d)^2 \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y \\ &= Rv_d^2\left[\left(1+\tilde{z}_d^{\dagger 2}\right)\left\{1-\Phi(\tilde{z}_d)\right\} - \left(2\tilde{z}_d^\dagger - \tilde{z}_d\right)\phi(\tilde{z}_d)\right\}. \end{aligned}$$

These identities are obtained by integration by parts. Under the condition of equilibrium at T, $\tilde{z}_d = \tilde{z}_d^\dagger$, and we have the equations

$$\begin{aligned} \Phi(\tilde{z}_d) &= \frac{R}{R+1} \\ (R+1)\left\{\left(1+\tilde{z}_d^2\right)\Phi(\tilde{z}_d) + \tilde{z}_d\,\phi(\tilde{z}_d)\right\} &= R\left(1+\tilde{z}_d^2\right) \end{aligned}$$

for the respective constant and quadratic loss. The former equation has an explicit solution for $\tilde{z}_d$, while the latter is solved by the Newton method. Denote the solution of the relevant equation by $\tilde{z}_d^*$. The results in Table 1 are based on the quadratic loss. The MSE of $\tilde{\theta}_d$ is

$$\mathrm{MSE}\left(\tilde{\theta}_d;\theta_d\right) = (1-b_d)^2 v_d^2 + b_d^2 (F_d - \theta_d)^2.$$

We replace the square $(F_d - \theta_d)^2$ by its average over the districts d, to eliminate the target θ_d. We refer to this operation as averaging. It results in the identity

$$\text{aMSE}\left(\tilde{\theta}_d; \theta_d\right) = (1-b_d)^2 v_d^2 + b_d^2 \left\{\sigma_B^2 + (F_d - \theta)^2\right\}.$$

Equilibrium at $\theta_d = T$ is satisfied when

$$F_d = T + \frac{|1-b_d|}{b_d} \tilde{z}_d^* v_d;$$

then the estimator is

$$\tilde{\theta}_d = (1-b_d)\,\hat{\theta}_d + b_d T + \tilde{z}_d^* |1-b_d|\, v_d$$

and its aMSE is

$$\begin{aligned}\text{aMSE}\left(\tilde{\theta}_d; \theta_d\right) &= (1-b_d)^2 \left(1 + \tilde{z}_d^{*2}\right) v_d^2 + b_d^2 \left\{\sigma_B^2 + (T-\theta)^2\right\} \\ &\quad + 2 b_d |1-b_d|\,(T-\theta)\tilde{z}_d^* v_d\,.\end{aligned}$$

This quadratic function of b_d attains an extreme when

$$b_d^* = \frac{v_d^2\left(1+\tilde{z}_d^{*2}\right) - \text{sign}\left(1-b_d^*\right)\,(T-\theta)\tilde{z}_d^* v_d}{v_d^2 + \sigma_B^2 + \left\{\tilde{z}_d^* v_d - \text{sign}\left(1-b_d^*\right)\,(T-\theta)\right\}^2},$$

and it can be shown that one of the two solutions is the unique minimum. For further details, see Longford (2015).