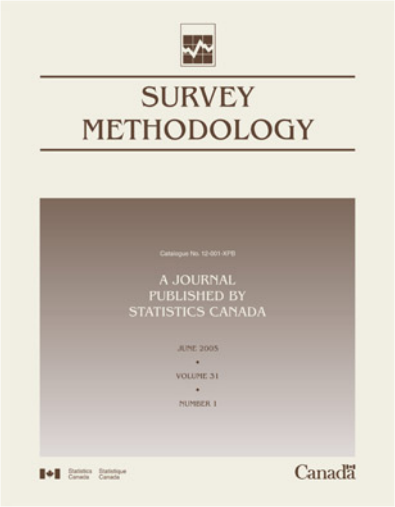
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (n^{o} 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 32, No. 1, pp. 87-96
Statistics Canada, Catalogue No. 12-001

Sample Size Calculation for Small-Area Estimation

Nicholas Tibor Longford [1]

Abstract

We describe a general approach to setting the sampling design in surveys that are planned for making inferences about small areas (sub-domains). The approach requires a specification of the inferential priorities for the areas. Sample size allocation schemes are derived first for the direct estimator and then for composite and empirical Bayes estimators. The methods are illustrated on an example of planning a survey of the population of Switzerland and estimating the mean or proportion of a variable for each of its 26 cantons.

Key Words: Efficiency; Inferential priority; Sample size allocation; Small-area estimation.

1. Introduction

Sampling design is a key device for efficient estimation and other forms of inference about a large population when the resources available do not permit collecting the relevant information from every member of the population. In this context, efficiency is interpreted as the optimal combination of a sampling design and an estimator of a population quantity θ. By optimum we understand minimum mean squared error, although the development presented in this paper can be adapted for other criteria. The pool of the possible sampling designs is delimited by the resources, and these are usually expressed in terms of a fixed sample size. This is not always appropriate because the designs may not entail identical average costs per subject. However, within a limited range of designs, this issue can be ignored.

The problem of setting the sampling design for the purpose of efficient estimation of a single quantity is well understood, and solutions are available for many commonly encountered settings. Most of them involve a univariate constrained optimisation problem. Setting the sampling design for estimating several quantities represents a quantum leap in complexity, because the problem involves several factors, typically one for each quantity. It is essential to optimise the design simultaneously for all the factors, because the goals of efficient inference about the target quantities may be in conflict. For example, in small-area estimation, a more generous allocation of the sample size to one area has to be compensated by a less generous allocation to one or several other areas.

Small-area statistics have become an important research topic in survey methods in the last few decades (Fay and Herriot 1979; Platek, Rao, Särndal and Singh 1987; Ghosh and Rao (1994), Longford 1999; and Rao 2003), stimulated by increasing interest of government agencies, the advertising and marketing industry and the financial and insurance sector. At present, many large-scale surveys are designed for estimating national quantities but, sometimes almost as an afterthought, are used for inferences about small areas. This would be appropriate if the sampling designs optimal for small-area and national inferences were similar. We illustrate in this paper that this is not the case and that sampling design can be effectively targeted for small-area estimation, taking into account the goal of efficient estimation of national quantities. To avoid the trivial case, we assume that the areas have unequal population sizes. We apply the methods to the problem of planning inferences about the 26 cantons of Switzerland; their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million.

Literature on the subject of planning surveys for small-area estimation is rather sparse. An important contribution is Singh, Gambino and Mantel (1994). In one of the approaches they discuss, the planned sample size for the Canadian Labor Force Survey is split into two parts. One part is allocated optimally for the purpose of national (domain) estimation and the remainder optimally for small-area estimation. For the latter goal, equal subsample sizes are allocated to each area when the areas have equal within-area variances, the finite population correction can be ignored and the areas have equal survey costs per subject, but also when the targets of inference are area-level means. When the targets are population totals, equal allocation to the areas is not efficient, because it handicaps estimation for more populous areas. Even when proportions or rates (percentages) are estimated, the within-area variances depend on the population proportion, although the dependence is weak when all the proportions are distant from zero and unity. For more recent developments in sampling design for small-area estimation, see Marker (2001).

The next section describes the proposed approach based on minimising the weighted sum of the sampling variances (mean squared errors) of the planned estimators, with the

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

weights specified to reflect the inferential priorities. It is applied first to direct estimation of the area-level quantities. Then it is extended to incorporate the goal of national estimation, and, finally, to composite estimation in section 3. The concluding section 4 contains a discussion.

The remainder of this section introduces the notation used in the rest of the paper. We assume that area-level population quantities θ_d, $d = 1, ..., D$, are estimated by $\hat{\theta}_d$ with respective mean squared errors (MSE) v_d that are functions of the within-area subsample sizes $n_d; v_d = v_d(n_d)$. The overall sample size is denoted by n, and is assumed to be fixed. The population sizes are denoted by N (overall) and N_d (for area d). For brevity, we denote $\mathbf{n} = (n_1, ..., n_D)^{\mathsf{T}}$. Most population quantities θ are functions of a single variable, such as its mean, total, and the like. The variable may be recorded in the survey directly, or constructed from one or several such variables. Although our development is not restricted to such quantities, the motivation is more straightforward with them. An estimator of θ_d is said to be *direct* if it is a function of only the variable concerned on subjects in area d.

We assume that each direct estimator considered is unbiased. This is not particularly restrictive, as most direct estimators are naive estimators or are closely related to them. We assume that the sample sizes for the areas are under the control of the survey designer. This is the case in stratified sampling designs in which the strata coincide with the areas. In section 4, we discuss sampling designs in which such control cannot be exercised; they are particularly relevant for divisions of the country into many (hundreds of) areas.

2. Optimal Design for Direct Estimation

We resolve the conflict between the goals of efficient estimation of the area-level quantities θ_d by choosing the area-level sampling design that minimises the weighted sum of the sampling variances (MSEs),

$$\min_{\mathbf{n}} \sum_{d=1}^{D} P_d v_d, \tag{1}$$

subject to the constraint of fixed overall sample size $n = \mathbf{n}^{\mathsf{T}} \mathbf{1}_D; \mathbf{1}_D$ is the vector of unities of length D. The coefficients P_d are called *inferential priorities*. Greater value of P_d (in relation to the values $P_{d'}$, $d' \neq d$) implies a greater urgency to reduce v_d, because the contribution of area d to the sum in (1) in magnified more than for the other areas.

The optimisation problem in (1) is solved by the method of Lagrange multipliers, or simply by substituting $n_1 = n - n_2 - ... - n_D$, so that the problem then involves $D - 1$ functionally unrelated variables. The solution satisfies the condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

An analytical expression for the optimal subsample sizes n_d cannot be obtained in general, but when $v_d = \sigma_d^2 / n_d$, as in simple random sampling within areas, the solution is proportional to $\sigma_d \sqrt{P_d}$, that is,

$$n_d^{\dagger} = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + ... + \sigma_D \sqrt{P_D}}.$$

When the within-area variances σ_d^2 coincide, $\sigma_1^2 = ... = \sigma_D^2 = \sigma^2$, this simplifies further; the optimal sample sizes are proportional to $\sqrt{P_d}$ and do not depend on σ^2.

In most contexts, it is difficult to elicit a suitable set of priorities P_d, and so it is more constructive to propose a convenient parametric class of priorities $\mathbf{P} = (P_1, ..., P_D)^{\mathsf{T}}$ and illustrate their impact on the sample size allocation. We propose the priorities $P_d = N_d^q$ for $0 \leq q \leq 2$. For $q = 0$, inference is equally important for every area. With increasing q, relatively greater importance is ascribed to more populous areas. When $v_d = \sigma^2 / n_d$, the optimal sample size allocation for $q = 2$, $n_d^{\dagger} = n N_d / N$, is proportional to the population sizes in the areas, and so the same sampling design is optimal for national and area-level inferences. For $q > 2$ the sample size allocation is even more generous to the most populous areas, at the expense of less populous areas. As this is counterintuitive in the context of small-area estimation, the choice of an exponent $q > 2$ is probably never appropriate. A negative priority exponent q would be suitable for a survey that aims to focus on the least populous areas. Of course, such a design is very inefficient for estimating the national quantity θ, especially when the areas have widely dispersed population sizes.

The inferential priorities P_d may be functions of quantities other than N_d. For example, the sizes of certain subpopulations of focal interest, such as an ethnic minority in the area, may be used instead of N_d, P_d may be defined differently in the country's regions, or the formula for them may be overriden for one or a few areas.

In some publications of survey analyses, an estimate is reported only when it is based on a sufficiently large sample size or its coefficient of variation (the ratio of the estimated standard error and the estimate) is smaller than a specified threshold. If a 'penalty' for not reporting a quantity is specified, it can be incorporated in the definition of the inferential priorities. The difficulty that may arise is that the objective function in (1) is discontinuous and the standard approaches to its optimisation are no longer applicable. The penalty has to be set with care. If it is too low it is ineffective; if it is set too high the solution will prefer reporting estimates for as many areas as possible, but each with sample size or precision that narrowly exceeds the set threshold. See Marker (2001) for an alternative approach to this problem.

Figure 1 illustrates the impact of the priority exponent q on the sample size allocation for a survey planned in Switzerland, with the aim of estimating the population means of a variable in its 26 cantons, assuming a common within-canton variance σ^2. The planned overall sample size is $n = 10{,}000$. The curves in either panel connect the optimal sample sizes for each exponent q; they are drawn on the linear scale (on the left) and on the log scale (on the right). The population sizes are marked on the horizontal bar at the bottom of each plot. On the log scale, the curves are linear. The log scale is useful also because the population sizes of the cantons are more evenly distributed on it.

For $q = 0$, each canton is allocated the same sample size, $10{,}000/26 = 385$, and for $q = 2$ the allocation is proportional to the canton's population size. For intermediate values of q, sample sizes of the least populous cantons are boosted in relation to proportional allocation $(q = 2)$, at the expense of reduced allocation to the most populous cantons. The subsample sizes depend very little on q for cantons with population of about 250,000, approximately 3% of the national population size.

2.1 The Priority for National Estimation

As the canton-level subsample sizes differ from the proportional allocation for priority exponent $q < 2$, optimal canton-level estimation is accompanied by a loss of efficiency of the national estimator. Consider the stratified estimator

$$\hat{\theta} = \frac{1}{N} \sum_{d=1}^{D} N_d \hat{\theta}_d$$

of the national mean θ of a variable, where $\hat{\theta}_d$ are unbiased estimators of the within-canton means of the same variable. Assuming stratified sampling with simple random sampling within strata (cantons), with $\hat{\theta}_d$ set to the within-stratum sample means,

$$\operatorname{var}(\hat{\theta}) = \frac{1}{N^2} \sum_{d=1}^{D} \frac{N_d^2}{n_d} (1 - f_d) \sigma_d^2,$$

where $f_d = n_d / N_d$ is the finite population correction.

Figure 2 displays the function that relates the standard error $\sqrt{\operatorname{var}(\hat{\theta})}$ to the priority exponent q, calculated assuming $\sigma^2 = 100$. The standard error is a decreasing function of q; it decreases more steeply at $q = 0$ than at $q = 2$, where it is quite flat. For $q = 2$, the goals of canton-level and national estimation are in accord, and $\sqrt{\operatorname{var}(\hat{\theta})} = 0.100$. For $q = 0$, $\sqrt{\operatorname{var}(\hat{\theta})} = 0.143$; in this setting, optimality of the small-area estimation exerts a considerable toll on national estimation, equivalent to halving the sample size $(0.143/0.100 \doteq \sqrt{2})$. For negative q, the toll is even greater.

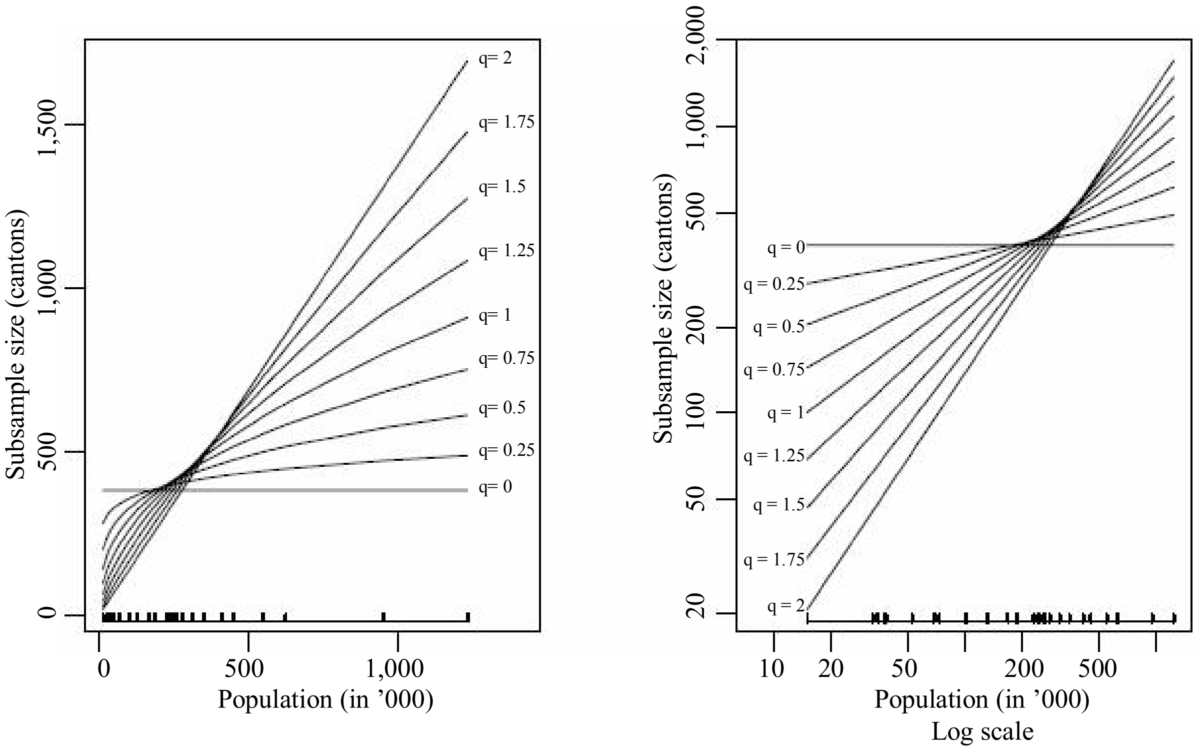


Figure 1. The sample size allocation to the Swiss cantons for a range of priority exponents q. The population sizes of the cantons are marked on the horizontal bar at the bottom of each plot.

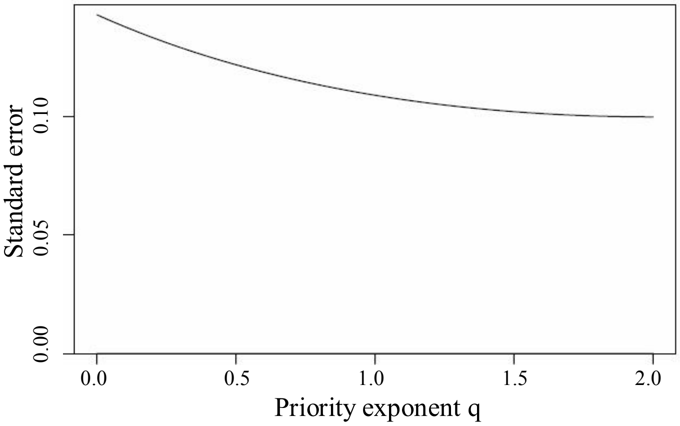


Figure 2. The standard error of the national estimator $\hat{\theta}$ of the mean of a variable, as a function of the exponent q for priorities of the canton-level estimation.

Thus, the need for efficiency of the national estimator can be addressed by increasing the priority exponent. For example, the parties with rival inferential interests may negotiate about how much loss in efficiency of $\hat{\theta}$ can be afforded, and the priority exponent would then be set to match this loss. Alternatively, this loss may be considered by applying the optimal design for area-level estimation. If it is regarded as excessive, q is increased until a balance is struck between the losses of efficiency for national and small-area estimation.

An unsatisfactory feature of these approaches is that they compromise the original purpose of the priorities $\mathbf{P}-$ to reflect the relative importance of the inferences about the distinct small areas. This drawback is addressed by associating $\hat{\theta}$ with a priority, denoted by $G,$ relative to small-area estimation, and considering optimal estimation of the set of D area-level targets θ_d together with the national target $\theta.$ Thus, we minimise the objective function

$$\sum_{d=1}^{D} P_d v_d(n_d) + GP_+ v(\mathbf{n}),$$

where $v = \operatorname{var}(\hat{\theta})$ and $P_+ = \mathbf{P}^{\mathsf{T}}\mathbf{1}_D.$ The factor P_+ is introduced to ameliorate the effect of the absolute sizes of P_d and the number of areas on the relative priority $G.$ The priorities P_d can be interpreted only by their relative sizes, as, for any constant $c > 0,$ P_d and cP_d correspond to identical sets of priorities for small-area estimation in (1).

When the sampling design within each area is simple random and $\hat{\theta}$ is the standard stratified estimator, the minimum is attained when

$$\sigma_d^2 \frac{P_d'}{n_d^2} = \text{const},$$

where $P_d' = P_d + GP_+ N_d^2 / N^2.$ The optimal sample sizes for the areas are

$$n_d^* = n \frac{\sigma_d \sqrt{P_d'}}{\sigma_1 \sqrt{P_1'} + \ldots + \sigma_D \sqrt{P_D'}}.$$

This corresponds to an adjustment of the priorities P_d by $GP_+ N_d^2 / N^2.$ Note that this adjustment is neither additive nor multiplicative. The priority is boosted more for the more populous areas. As a consequence, the area-level subsample sizes are dispersed more when the relative priority for national estimation is incorporated and the area-level priorities are unchanged. The finite population correction has no impact on n_d^* because it reduces each sampling variance v_d and v by a quantity that does not depend on $\mathbf{n}$.

The priority G can be set by insisting that the loss of efficiency in estimating the national quantity θ does not exceed a given percentage or that at most a few (or none) of the absolute differences $|P_d' - P_d|$ or log-ratios $|\log(P_d' / P_d)|$ are very large. However, the analytical problem is simple to solve, so the survey management can be presented by the sampling designs that are optimal for a range of values of $G.$

The dependence of the subsample size on the exponent q and relative priority G is plotted in Figure 3 for the least and most populous cantons, Appenzell-Innerrhoden and Zürich, in the respective panels A and C. Panels B and D plot the same curves as A and C, respectively, on the log scale. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small-area estimation to very large values of $G.$ Throughout, we assume that $n = 10{,}000$ and $\sigma^2 = 100,$ common to all cantons.

For each exponent $q < 2,$ the sample-size curve $n_d(G)$ decreases for the less populous and increases for the more populous cantons toward the proportional representation $n_d = nN_d / N,$ which corresponds to $q = 2.$ On the linear scale, the increase is quite rapid for Zürich for small q and $G,$ whereas the reduction for Appenzell-Innerrhoden is more gradual. As the relative priority G is reduced, the excess sample size is re-distributed from Zürich (and a few other populous cantons) to several less populous cantons.

Figure 4 plots the 'national' standard error $\sqrt{\operatorname{var}(\hat{\theta})}$ under the optimal sample allocation for an array of values of q and $G.$ The diagram shows that the standard error of $\hat{\theta}$ is reduced radically by a small increase of G in the vicinity of $G = 0,$ whereas for larger values of G it is affected only slightly. For each $G,$ higher priority exponent q is associated with higher precision of $\hat{\theta}.$

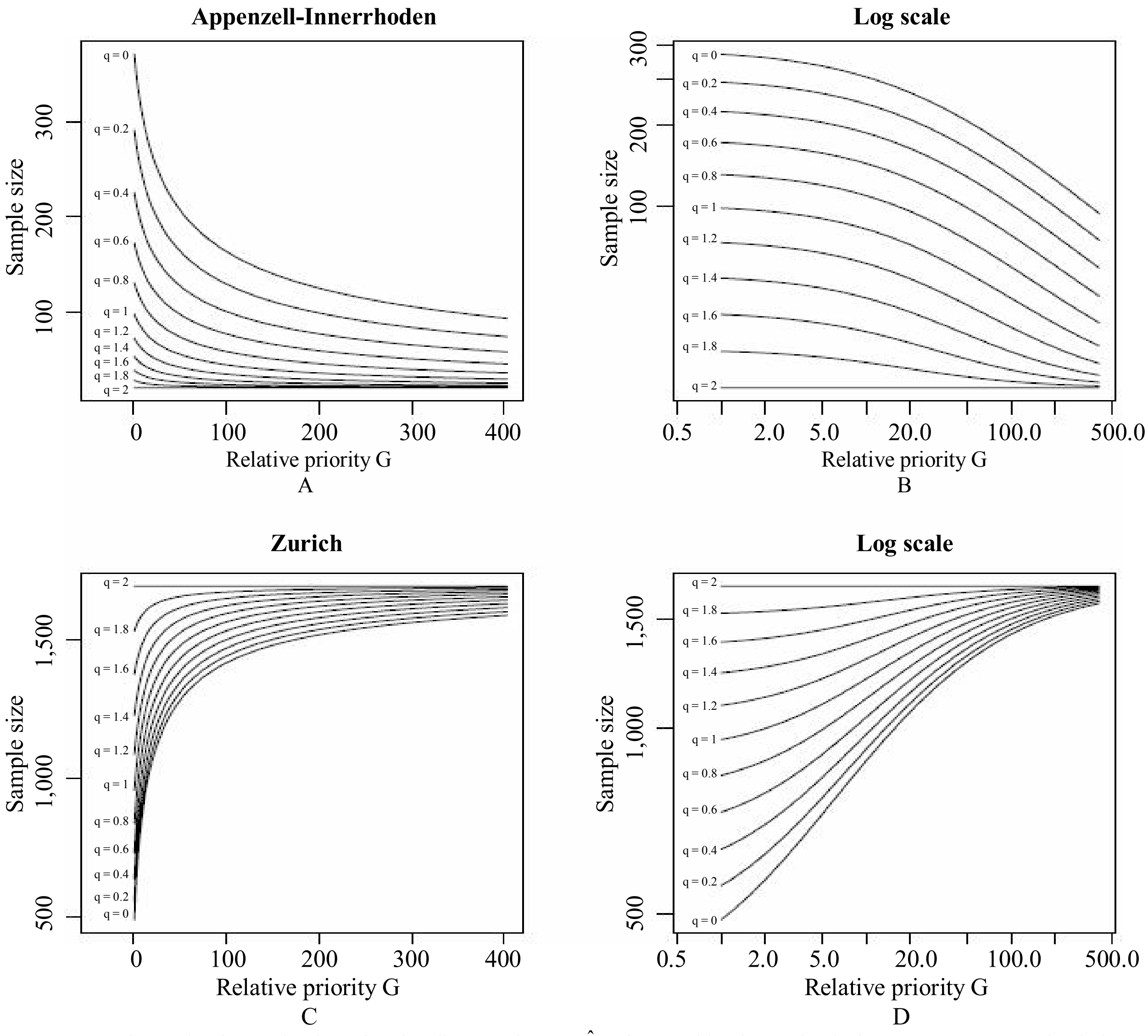


Figure 3. The optimal sample sizes for the direct estimator $\hat{\theta}_d$ for combinations of priority exponents q and relative priorities G for the least and most populous cantons.

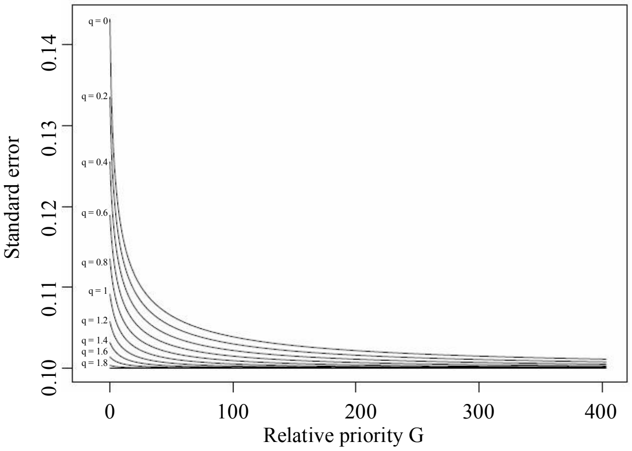


Figure 4. The standard error of the national estimator for the allocation that is optimal under an array of priorities given by q and G.

3. Composite Estimation

The resources available for the conduct of a survey are used most effectively by the optimal combination of a sampling design and estimator(s), and so the sampling design and (the selection of) the estimator should be, in ideal circumstances, optimised simultaneously. This problem is difficult to solve formally in most settings, although some estimators are more efficient than their competitors in a wide range of designs. Composite estimators (Longford 1999, 2004) are one such class. They are convex combinations of the direct small-area and national estimators,

$$\tilde{\theta}_d = (1 - b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}, \qquad (2)$$

with area-specific coefficients b_d that are estimates of the optimum. The composition $\tilde{\theta}_d$ exploits the similarity of the areas; it is particularly effective when the areas have a small between-area variance $\sigma_B^2 = D^{-1}\sum_d (\theta_d - \overline{\theta})^2$, where $\overline{\theta} = D^{-1}\sum_d \theta_d$. This variance is defined over the D population quantities θ_d and is unaffected by the sampling design. In practice, σ_B^2 has to be estimated. When planning a survey, estimates from other surveys of the same or a related population have to be used, and the uncertainty about σ_B^2 addressed. This can be done by sensitivity analysis, exploring the optimal designs for a range of plausible values of σ_B^2.

If the deviations $\Delta_d = \theta_d - \overline{\theta}$ were known the optimal coefficient b_d in (2) would be, approximately, $b_d^* = \sigma_d^2 / (\sigma_d^2 + n_d \Delta_d^2)$. As Δ_d is not known (otherwise θ_d would be estimated with high precision by $\overline{\theta} + \Delta_d$), we replace Δ_d^2 by its average over the areas, equal to σ_B^2, yielding the coefficient $b_d = 1/(1 + n_d \omega_d)$, where $\omega_d = \sigma_B^2 / \sigma_d^2$ is the variance ratio. The variance σ_B^2 also has to be estimated, but when there are many areas it is estimated with precision much higher than most Δ_d^2 are.

If the coefficients b_d are estimated with sufficient precision the composite estimator $\tilde{\theta}_d$ is more efficient than the two constituent estimators $\hat{\theta}_d$ and $\hat{\theta}$. Ignoring the uncertainty about the within- and between-area variances, as well as the national mean $\overline{\theta}$ and the correlation between the national and area-level (direct) estimators, the average MSE of $\tilde{\theta}_d$ is

$$\mathrm{aMSE}(\tilde{\theta}_d) = \frac{\sigma_B^2}{1 + n_d\,\omega_d}, \tag{3}$$

where 'aMSE' denotes the MSE in which Δ_d^2 is replaced by σ_B^2, its average over the areas. The aMSE in (3) is also an approximation to the conditional variance of the EBLUP estimator of the area-level mean based on the two-level (empirical Bayes) model (Longford 1993, Goldstein 1995, Marker 1999, and Rao 2003). See Ghosh and Rao (1994) for an authoritative review of application of these models to small-area estimation.

For the composite estimators of the area-level means, we search for the sample allocation that minimises the objective function

$$\sum_{d=1}^{D} P_d \,\mathrm{aMSE}\,(\tilde{\theta}_d) + GP_+ v.$$

The solution satisfies the condition

$$\frac{N_d^q \sigma_B^2 \omega_d}{(1 + n_d \omega_d)^2} + GP_+ \frac{N_d^2}{N^2} \frac{\sigma_d^2}{n_d^2} = \text{const.} \tag{4}$$

This equation does not have a convenient closed-form solution, but iterative schemes can be applied to solve it. The value of n_1 determines the remaining sample sizes n_d, and so optimisation corresponds to a one-dimensional search. If the provisional sample sizes $\mathbf{n}$ based on a set value of n_1 are too large, $\mathbf{n}^\top \mathbf{1}_D > n$, n_1 is reduced and the other sample sizes n_d are calculated by solving (4). Note that the solution depends on the variances σ_d^2 and σ_B^2. The problem is simplified somewhat when the areas have a common variance $\sigma^2 = \sigma_1^2 = ... = \sigma_D^2$. Then the solution of (4) depends on the variances only through the ratio $\omega = \sigma_B^2 / \sigma^2$ because σ^2 is a multiplicative factor and has no impact on the optimisation.

By way of an example, suppose $q = 1$ and $G = 10$ in planning a survey of the population of Switzerland with $n = 10{,}000$, and $\omega = 0.10$ is assumed. As the initial solution, we use the allocation optimal for direct estimation with the same values of q and G. One iteration updates the sample size for each canton and, within it, the updating for all but the arbitrarily selected reference canton $d = 1$ is also iterative. The reference canton's provisional subsample size determines the current value of the constant on the right-hand side of (4). Then equation (4) is solved, iteratively, for each canton $d = 2, ..., D$, using the Newton method. In the application, the number of these iterations was in single digits for each canton. Finally, the subsample size for the reference canton is adjusted by the $1/D$-multiple of the difference between the current total of the subsample sizes and the target total n. The updating of the cantons is itself iterated, but only a few iterations are required to achieve convergence; for example, all the changes in the subsample sizes were smaller than 1.0 after three iterations, and smaller than 0.01 after eight iterations. The convergence is fast because the starting solution is close to the optimum; the largest difference between the two subsample sizes is for Zürich, 20.0 (from 1199.5 at the start to 1219.5 after eight iterations). For Appenzell-Innerrhoden, the sample size is reduced from 81.6 to 73.4. Change by less than unity takes place for five cantons with population sizes in the range 228,000–278,000. Note that the subsample sizes would in practice be rounded, and possibly adjusted further to conform with various survey management constraints.

No priority for national estimation

If national estimation has no priority, $G = 0$, equation (4) has the explicit solution

$$n_d^* = \frac{n\omega + D}{\omega} \frac{N_d^{q/2}}{U^{(q)}} - \frac{1}{\omega},$$

where $U^{(q)} = N_1^{q/2} + ... + N_D^{q/2}$. This allocation is related to the allocation $n_d^{\dagger}$, $d = 1, ..., D$, that is optimal for direct estimation of θ_d by the identity

$$n_d^* = n_d^{\dagger} + \frac{1}{\omega}\left(\frac{DN_d^{q/2}}{U^{(q)}} - 1\right).$$

Hence, when $q > 0$, the allocation optimal for composite estimation is more dispersed than for direct estimation. The break-even population size is $N_{\mathrm{T}} = (U^{(q)}/D)^{2/q}$; areas with population sizes $N_d < N_{\mathrm{T}}$ have smaller subsample sizes for composite than for direct estimation, and areas with greater population sizes have greater subsample sizes. (For $q = 0$, $n_d^* \equiv n/D$). The amount of extra dispersion is inversely proportional to ω.

For $\omega = 0$, the equations for the optimal sampling design lead to a singularity. In this case, each θ_d is estimated efficiently by the national estimator $\hat{\theta}$, and so the design optimal for composite estimation coincides with the design that is optimal for the national estimator $(n_d^* = nN_d/N)$. For $q > 0$, the optimal allocation yields negative sample sizes n_d^* when

$$N_d < \left\{\frac{U^{(q)}}{n\omega + D}\right\}^{2/q}. \qquad (5)$$

This (formal) solution is not meaningful. A negative solution should come as no surprise because the aMSE in (3) is an analytical function for $n_d > -1/\omega_d$. For small $\omega > 0$, the aMSE is a shallow decreasing function of the sample size n_d. A negative n_d^* indicates that a (small) canton is not worth sampling because of its low inferential priority P_d. Although additional sample size for a more populous canton d' may yield a smaller reduction of aMSE than it would for a small canton d, its impact is magnified by the larger priority $P_{d'}$.

Positive priority for the national mean

The aMSE in (3) ignores the uncertainty about the national mean θ, and this becomes acute when one of the cantons is not represented in the sample. This deficiency of (3) can be compensated for by setting the relative priority G to a positive value.

Figure 5 summarises the impact of the relative priority G and the priority exponent q on the optimal sample sizes of the least and most populous cantons, together with canton Thurgau which has the 13th (median) largest population size, 228,000. Each setting of q, indicated in the title, and G, using different line types, is represented for a canton by a graph of the optimal sample size as a function of the variance ratio ω. The limit of this function for $\omega \to +\infty$, equal to the sample size optimal for direct estimation, is marked by a bar at the right-hand margin of the panel. For $\omega = 0$, the sampling design optimal for estimation of the national mean θ is obtained. Panels A and B at the top are for the overall sample size $n = 10,000$ and panels C and D for $n = 1,000$.

The diagram shows that the optimal sample sizes are nearly constant in the range $\omega \in (\omega^*, +\infty)$; ω^* increases with q, G and $1/n$. This is a consequence of the relatively large sample size n, which ensures that the subsamples of most cantons are too large for any substantial borrowing of strength across the cantons to take place, unless the cantons are very similar $(\omega < \omega^*)$. Most shrinkage coefficients $b_d = 1/(1 + n_d\,\omega)$ are very small. When $n = 10,000$ is planned, for small values of ω, the optimal sample size increases steeply for the least populous canton and drops precipitously for the most populous canton. Dispersion of the optimal sample sizes increases with q and G, converging to the optimal allocation for estimating the national mean θ, which corresponds to $\omega = 0$. In contrast, the optimal sample sizes are discontinuous at $\omega = 0$ when $G = 0$; the solutions diverge to $-\infty$ for the least populous cantons.

In panels C and D, for $n = 1,000$, the dependence of the sample sizes on ω persists over a wider range of ω because there is a greater scope for borrowing strength across the cantons with the smaller sample sizes. The optimal sample sizes are not monotone functions of ω; for the least populous cantons there is a dip at small values of ω. The dip is more pronounced for small G and large q, that is, when the disparities of the cantons' priorities are greater and inference about the national mean is relatively unimportant. This phenomenon, somewhat exaggerated by the log-scale of the vertical axis, is similar to the case discussed for $G = 0$. Because of the disparity in the priorities P_d, a small reduction of aMSE for a more populous canton is preferred to a greater reduction for a less populous canton. The dip is present also when $n = 10,000$, but it is so shallow and narrow as to be invisible with the resolution of the graph. Note that the horizontal axes in panels C and D have three times wider range of values of ω than in panels A and B.

In the context of the planned survey, it was agreed that ω is unlikely to be smaller than 0.05. Therefore, the sample size calculations could be based on the direct estimator.

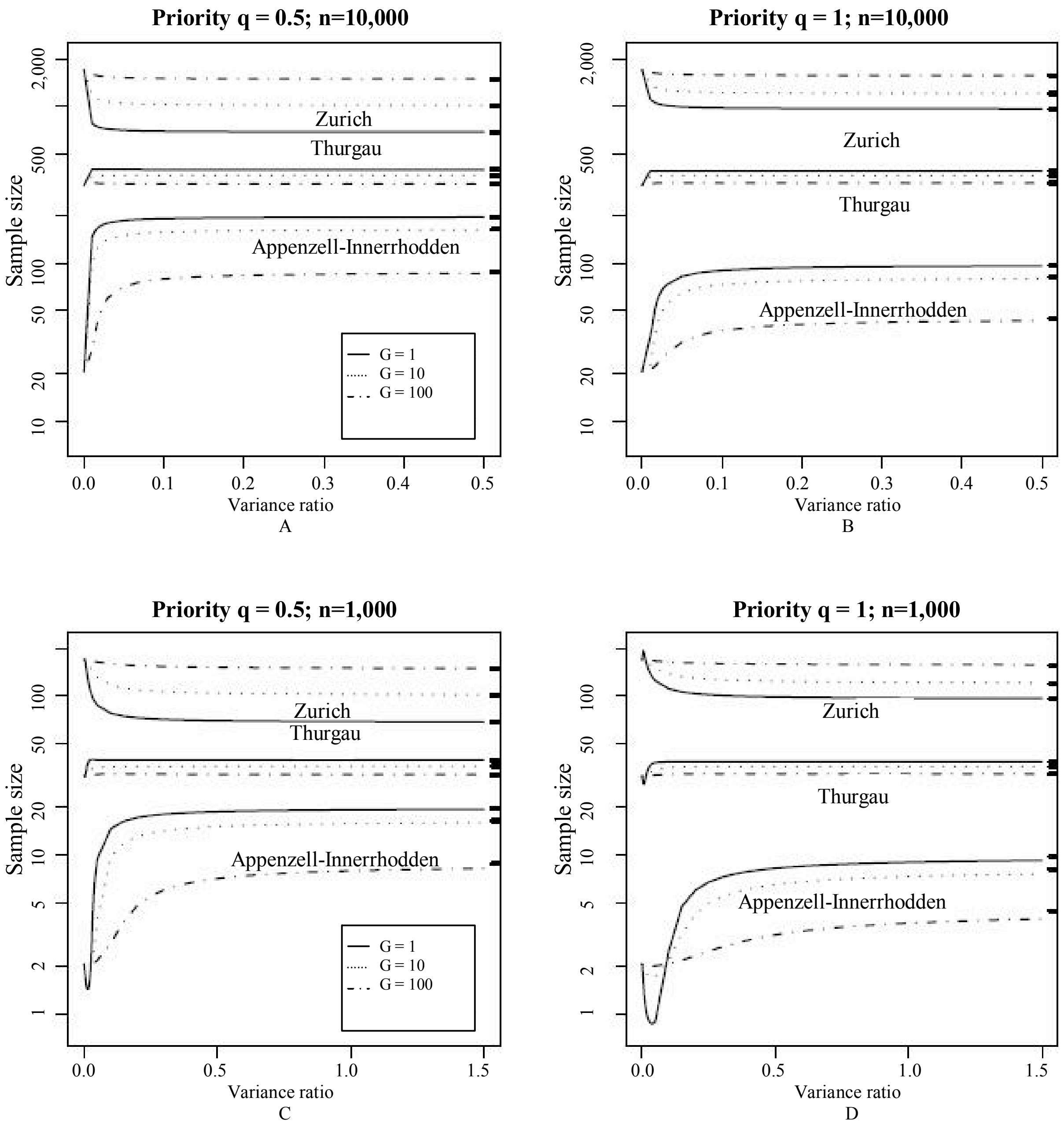


Figure 5. The sample sizes optimal for composite estimation of the population means for three cantons for a range of variance ratios ω, priority exponents $q = 0.5$ and $q = 1.0$ and relative priorities $G = 1$, 10 and 100. The overall sample sizes are 10,000 (panels A and B) and 1,000 (panels C and D).

4. Discussion

The method described in this paper identifies the optimal design for the artificial setting of stratified sampling with simple random sampling within homoscedastic strata. Specifying the priorities for small-area and national estimation is a key element of the method. In practice, the priorities may be difficult to agree on, and some of the assumptions made may be problematic, the assumptions of equal within-stratum variances and simple random sampling in particular. The method can be extended to more complex estimators, but then the values of further parameters are required. A more constructive approach regards the optimal sampling design for the simplified setting as an approximation to the sampling design that is optimal for the more realistic setting. Even if the optimal sampling design were identified, it could not be implemented literally, because of imperfections in the sampling frame and (possibly) informative and unevenly distributed nonresponse. However, the approach can be applied, in principle, to any small-area estimator that has an analytical expression for the exact or approximate MSE. This includes all estimators based on empirical Bayes models, to which the composite estimator is closely related. Sampling weights can be incorporated in sample size calculation if they, or their within-area distributions, are known, subject to some approximation, in advance. Sample size calculation for a single (national) quantity entails the same problem.

Although the numerical solution of the problem for composite estimation with a positive priority G is simple and involves no convergence problems, it is advantageous to have an analytical solution, so that a range of scenarios can be explored. The proximity of the solutions for the direct and composite estimation suggests that the allocation optimal for direct estimation may be close to optimum also for composite estimation with realistic values of ω, say, $\omega > 0.05$.

Various management and organisational constraints are another obstacle to the literal implementation of an analytically derived sampling design. In household surveys, it is often preferable to assign an (almost) full quota of addresses to each interviewer, and so sample sizes that are multiples of the quota are preferred. These and numerous other constraints can be incorporated in the optimization problem, although they are often difficult to quantify or the designer may not be aware of them because of imperfect communication. Improvisation, after obtaining the sampling design that is optimal for a simpler setting, may be more practical. Also, priorities, or expert opinion about them, may change over time, even while the survey is being conducted and analysed. Estimates that are associated with standard errors or coefficients of variation greater than a specified threshold are often excluded from analysis reports. Intention to do this can be reflected in sample size calculation by regarding $\hat{\theta}$ as the estimator of θ_d, that is, by setting the associated MSE to the corresponding aMSE $\sigma_{\mathrm{B}}^2 + \mathrm{var}(\hat{\theta})$ or to another (large) constant.

Although we propose a particular class of priorities for the small areas, no conceptual difficulties arise when another class is used instead. It may depend on several population quantities, not only the population size. In principle, the priorities can also be set for the areas individually, although this is practical only when the number of areas is small. The formula-based and individually set priorities can be combined by adjusting the priorities, such as $P_d = N_d^q$, for a few areas to reflect their exceptional role in the analysis.

Sensitivity analysis, exploring how the sampling design is changed as a result of altered input, is essential for understanding the impact of uncertainty about the estimated parameters (the variance ratio ω in particular) and the arbitrariness, however limited, in how the priorities are set. For this, an analytically simple solution that can be executed many times, for a range of settings, is preferred to a more complex solution, the properties of which are more difficult to explore.

Multivariate composite estimators exploit the similarity not only across areas, but also across (auxiliary) variables, time, subpopulations, and the like (Longford 1999 and 2005). The aMSEs of these estimators depend on the scaled variance matrix $\mathbf{\Omega}$, the multivariate counterpart of ω. Sample size calculation for this method is difficult to implement directly because both variances and covariances in $\mathbf{\Omega}$ are essential to the efficiency of the estimators. A more constructive approach matches the matrix $\mathbf{\Omega}$ with a ratio ω that can be interpreted as the similarity of the areas after adjusting for the auxiliary information, as in empirical Bayes methods.

When control over the sample sizes allocated to the areas is not possible sample size calculation is still meaningful as a guide for how the sample sizes should be allocated *on average*. In general, a unit reduction of the sample size is associated with greater loss of precision than a unit increase. Therefore, designs in which the sampling (replication) variance of the subsample sizes n_d (d fixed) is smaller are better suited for small-area estimation. In designs with large clusters, such variances are large because, at an extreme, an area may not be represented in the survey in some replications and may be over-represented several times in others. Using smaller clusters is in general preferable for small-area estimation if this does not inflate the survey costs and a fixed overall sample size can be maintained.

Acknowledgements

I am grateful to the Deputy Editor and referees for suggesting several improvements but mainly for leading me to discover an error in an earlier version of the manuscript. Discussions with the Polish team in the EURAREA project are acknowledged.

References

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

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

Goldstein, H. (1995). *Multilevel Statistical Models.* Second Edition. Edward Arnold, London, UK.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

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. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society*, Series A, 167, 341-373.

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

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.

Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Eds.) (1987). S*mall Area Statistics*. New York: John Wiley & Sons.

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

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.