2013

Model-assisted optimal allocation for planned domains using composite estimation

Wilford Molefe
University of Botswana

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

19-13

Model-Assisted Optimal Allocation for Planned Domains Using Composite Estimation

Wiford B. Molefe and Robert Graham Clark

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

MODEL-ASSISTED OPTIMAL ALLOCATION FOR PLANNED DOMAINS USING COMPOSITE ESTIMATION

Wilford B. Molefe [1] and Robert Graham Clark [2]

ABSTRACT

This paper develops allocation methods for stratified sample surveys where small area estimates are a priority. Small areas are domains of interest with sample sizes too small to allow traditional direct estimation to be feasible. Composite estimation may then be used, to balance between using a grand mean estimate and an area-specific estimate for each small area. In this paper, we assume stratified sampling where small areas are strata. Similar to Longford (2006), we seek efficient allocations where the aim is to minimise a linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. Unlike Longford, we define mean-squared error in a model-assisted framework, allowing a more natural interpretation of results using an intra-class correlation parameter. This optimal allocation is only available analytically for a special case, and has the unappealing property that some strata may be allocated no sample. Some alternative allocations, including a power allocation with numerically optimized exponent, are found to perform nearly as well as the optimal allocation, but with better practical properties.

Key Words: small area estimation, sample design, sample size allocation, composite estimation, mean squared error, Taylor approximation.

[1]Department of Statistics, University of Botswana. E-mail: molefewb@mopipi.ub.bw

[2]National Institute for Applied Statistics Research Australia, University of Wollongong. Email: rclark@uow.edu.au

1. INTRODUCTION

Sample surveys have long been used as cost-effective means for data collection but it is also the case that general purpose surveys will often not achieve adequate precision for statistics for subpopulations of interest (often called domains or areas). Domains may be geographically based areas such as states. They may also be cross-classifications of a small geographic area and a specific demographic or social group. A domain or an area is considered large or major if the domain-specific sample is sufficiently large to yield *direct estimates* (using data from just that area) of adequate precision. On the other hand, a domain is regarded as *small* if the domain-specific sample is not large enough to produce a direct estimate with reliable precision. In the survey sampling literature, areas or domains with small sample are referred to as *small areas*. Small areas are also often referred to as *small domains*, *local areas*, *subdomain*, *small subgroup*, *subprovince* and *minor domain* (Rao, 2003).

Sampling designs and in particular sample sizes are chosen in practice so as to provide reliable estimates for aggregates of the small areas such as large geographical regions or broad demographic groups. Budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. Also, it is often the case that domains of interest are only specified after the survey has already been designed and carried out. In practice, it is not possible to anticipate and plan for all possible areas (or domains) and uses of survey data as "the client will always require more than is specified at the design stage" (Fuller, 1999).

The increased emphasis on small area estimation raises the question of how best to design samples when the precision of small area estimates is a priority. If small area data needs are to be served using survey data then there is a need to develop an

overall strategy that involves careful attention to satisfy these needs at the planning, sample design and estimation stages of the survey process (Singh et al., 1994). Singh et al. (1994) present an illustration of compromise sample size allocation to satisfy reliability requirements at the provincial level as well as sub-provincial level.

Marker (2001) concludes that it will never be possible to anticipate all survey uses, or to allocate sufficient sample sizes to all domains of interest, so indirect estimators will always be needed. But he found that it is possible to make design choices that will greatly improve the ability of national surveys to support direct estimation for many small areas and such choices could also improve the ability of surveys to be used to produce indirect estimates where they are needed.

In this paper, we suppose that small areas can be identified in advance, and that stratified sampling is used with H strata defined by the small areas, indexed by $h \in U^1$. The population of units, indexed by j, is denoted U, of size N. The population of N_h units in stratum h is U_h and the sample of n_h units selected by simple random sampling without replacement (SRSWOR) from stratum h is s_h. Let Y_j be the value of the characteristic of interest for the jth unit in the population. The small area population mean is $\bar{Y}_h$ and the national mean is $\bar{Y}$. The corresponding sample estimators are $\bar{y}_h$ and $\bar{y}$, respectively; $\bar{y}_h = n_h^{-1} \sum_{j \in s_h} y_j$ and $\bar{y} = \sum_{h \in U^1} P_h \bar{y}_h$, where $P_h = N_h/N$. Let the sampling variances be $v_h = var_p(\bar{y}_h)$ and $v = var_p(\bar{y})$.

Longford (2006) considers the problem of optimal sample sizes for small area estimation for this design. The approach is based on minimizing the weighted sum of the mean squared errors of the planned small area mean estimates and an overall estimate of the mean, with the weights specified to reflect the inferential priorities.

An analytical solution exists for the case where no weight is attached to estimating the overall mean but it has undesirable practical properties, and may sometimes result in zero sample sizes for some strata. When the overall mean is also important Longford does not find an exact or approximate analytical solution to the optimization problem. He suggests that the equation can be solved by numerical methods, such as the Newton method, but this may involve significant computation if there are a large number of small areas.

The aim of this paper is to find the best allocation to strata for a linear combination of small area composite estimates and an overall estimator of the mean similar to Longford. In Section 2 we reformulate the objective in model-assisted terms, and derive the model-assisted composite estimator. Section 3 is devoted to optimizing the design. In Subsection 3.1 we derive the optimal allocation for this objective when national estimation has no priority ($G = 0$) (similar in form to Longford but with different interpretation due to the explicit use of a model). Longford (2006) did not give an analytical solution for the case where both national (overall) and small area estimates are a priority ($G > 0$). A numerical algorithm was given but may be computationally intensive, and its iterative nature makes it harder to see what the method is doing.

In Subsections 3.2 and 3.3 we derive two different Taylor Series approximations to the optimum. Unfortunately the optimal allocations (both when $G = 0$ and when $G > 0$) have some undesirable properties, so in Subsection 3.4 we consider a power allocation with numerically optimized exponent and also suggest several ad-hoc sample allocations.

2. COMPOSITE ESTIMATION

Royall (1973), in a discussion of papers by Gonzalez (1973) and Ericksen (1973), suggested that a choice between direct and synthetic approaches need not be made but that '... a combination of the two is better than either taken alone'. A natural way to balance the potential bias of a synthetic estimator $\bar{y}$ for $\bar{Y}_h$ against the instability of a direct estimator $\bar{y}_h$, is to use a composite estimator $\tilde{y}_h^{\mathcal{C}}$.

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $\tilde{y}_h^{\mathcal{C}} = (1-\phi_h)\bar{y}_h + \phi_h\bar{y}$ of the sample mean $\bar{y}_h$ for the target area h and the overall sample mean $\bar{y}$ of the target variable. The (area-specific) coefficients ϕ_h and $1-\phi_h$ in this composition are set with the intent to minimise its mean squared error (MSE), see for example, Schaible (1978); Brock et al. (1980) and Rao (2003). The coefficients for which minimum MSE would be attained depend on some unknown parameters which have to be estimated.

The design-based MSE of the composite estimator is given by:

$$MSE_p\big(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\big) = (1-\phi_h)^2 v_h + \phi_h^2\{v + B_h^2\} + 2\phi_h(1-\phi_h)c_h$$

where c_h is the sampling covariance of $\bar{y}_h$ and $\bar{y}$, v_h is the sampling variance of the direct estimator $\bar{y}_h$, v is the sampling variance of the synthetic estimator $\bar{y}$ for $\bar{Y}_h$ and $B_h = \bar{Y}_h - \bar{Y}$ is the bias of using $\bar{y}$ to estimate $\bar{Y}_h$. Further,

$$MSE_p\big(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\big) \approx (1-\phi_h)^2 v_h + \phi_h^2 B_h^2 \tag{1}$$

because $c_h \ll v_h$ and $v \ll v_h$ when number of small areas is large.

The following model ξ will be assumed:

$$\left.\begin{aligned} &E_\xi[Y_j] = \mu \\ &var_\xi[Y_j] = \sigma^2 \\ &cov_\xi[Y_i, Y_j] = \rho\sigma^2 \ (i \neq j;\ i, j \in U_h) \\ &cov_\xi[Y_i, Y_j] = 0 \ (i \in U_h, j \in U_g, h \neq g) \end{aligned}\right\} \tag{2}$$

where i and j are units and h and g are small areas.

Under the model (2),

$$E_\xi[v_h] = E_\xi[var_p(\bar{y}_h)] = E_\xi\big[n_h^{-1}S_{hw}^2\big] = n_h^{-1}\sigma^2(1-\rho)$$

and

$$E_\xi[B_h^2] = E_\xi\big[\big(\bar{Y}_h - \bar{Y}\big)^2\big] \approx var_\xi[\bar{Y}_h] = var_\xi\left(N_h^{-1}\sum_{j\in U_h} Y_j\right) = \sigma^2 N_h^{-1}[1 + (N_h - 1)\rho].$$

While the areas may be small in terms of n_h, they are of reasonable size in terms of N_h, so that $E_\xi[B_h^2] \approx \sigma^2\rho$. Also,

$$E_\xi[v] = E_\xi[var_p(\bar{y})] = E_\xi\left(\sum_{h\in U^1} P_h^2 n_h^{-1} S_{hw}^2\right) = \sigma^2(1-\rho)\sum_{h\in U^1} P_h^2 n_h^{-1}.$$

Substituting for $E_\xi[v_h]$ and $E_\xi[B_h^2]$ into (1) we get the anticipated MSE or approximate model assisted mean squared error, denoted $AMSE_h$:

$$\begin{aligned} AMSE_h &= E_\xi MSE_p\big(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\big) \\ &= (1-\phi_h)^2 n_h^{-1}\sigma^2(1-\rho) + \phi_h^2\sigma^2\rho\left(1 + \frac{1-\rho}{\rho}\sum_{h\in U^1} P_h^2 n_h^{-1}\right) \qquad (3) \\ &\approx (1-\phi_h)^2 n_h^{-1}\sigma^2(1-\rho) + \phi_h^2\sigma^2\rho \end{aligned}$$

Optimizing with respect to ϕ_h we immediately obtain the optimal weight ϕ_h as:

$$\phi_{h(opt)} = (1-\rho)\big[1 + (n_h - 1)\rho\big]^{-1}. \tag{4}$$

We substitute the optimum weight (4) into (3) to obtain the approximate optimum anticipated MSE:

$$\begin{aligned} AMSE_h &= E_\xi MSE_p\big(\tilde{y}_h^{\mathcal{C}}\big[\phi_{h(opt)}\big]; \bar{Y}_h\big) \\ &\approx \Big(n_h\rho\big[1+(n_h-1)\rho\big]^{-1}\Big)^2 n_h^{-1}\sigma^2(1-\rho) + \Big((1-\rho)\big[1+(n_h-1)\rho\big]^{-1}\Big)^2\sigma^2\rho \\ &= \sigma^2\rho(1-\rho)\big[1+(n_h-1)\rho\big]^{-1}. \end{aligned}$$

3. OPTIMIZING THE DESIGN

3.1 Optimal Design When $G = 0$

Provision of precise survey estimates for domains of interest requires that samples of adequate sizes be allocated to the domains. Conflicts arise when equal precision is desired for domains with widely varying population sizes. If estimates are desired at the same level of precision for all domains, then an equal allocation may be the most efficient strategy. However, such an allocation can cause a serious loss of efficiency for national estimates.

One way of measuring the performance of designs for small area estimation is with a linear combination of the anticipated MSE's of the small area mean and overall mean estimates. Following Longford (2006), but using anticipated MSEs instead of design-based MSEs, we use

$$F = \sum_{h\in U^1} N_h^q AMSE_h + GN_+^{(q)} E_\xi v \tag{5}$$

where the weights N_h^q reflect the inferential priorities for areas h, with $0 \le q \le 2$, and $N_+^{(q)} = \sum_{h\in U^1} N_h^q$. The quantity G is a relative priority coefficient. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small area estimation corresponds to large values of G, since when G is very large the second

component in (5) is dominant in this case. If G is non-zero, it would typically be large because v would generally be much smaller than v_h, so that G has to be large if the last term of (5) is to have any influence on the outcome. The factor $N_+^{(q)}$ is introduced to appropriately scale for the effect of the absolute sizes of N_h^q and the number of areas on the relative priority G. Criteria (5) is similar to the criteria in Longford (2006), however unlike this paper we adopt the model-assisted approach which treats the design-based inference as the real goal of survey sampling, but employs models to help choose between valid randomization-based alternatives (Särndal et al., 1992).

The minimization is subject to a fixed sample size constraint. It would be straightforward to extend this to a fixed cost constraint with different cost coefficients in different strata.

When national estimation has no priority ($G = 0$), the solution for the number of units to be sampled from each strata is found by optimizing (5) subject to a fixed total sampling cost function. The stationary point for this optimization is

$$n_{h,opt.} = \frac{n\sqrt{N_h^q}}{\sum_{h\in U^1}\sqrt{N_h^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{N_h^q}}{H^{-1}\sum_{h\in U^1}\sqrt{N_h^q}} - 1\right) \tag{6}$$

The expression (6) is a stationary point. It is also the optimal design if it gives a feasible solution ($0 \le n_{h,opt.} \le N_h$ for all h); if not the optimal design must be obtained numerically. An approximate solution can be found by setting the non-feasible solutions to $n_{h,opt.} = 0$ when $n_{h,opt.} < 0$ or $n_{h,opt.} = N_h$ when $n_{h,opt.} > N_h$ and then reallocating the remaining small areas (Longford, 2006).

In practice it would almost always be appropriate to set $0 \le q \le 2$, with $q = 0$ corresponding to all areas being equally important regardless of size, and $q = 2$ being the best choice for national estimation. In many cases $q = 1$ would be a

sensible compromise.

The first term in (6) above is the optimal allocation for the direct estimator and corresponds to power allocation (Bankier, 1988). The second term will be positive for more populous areas (large N_h) and negative for less populous areas. Therefore, the allocation optimal for composite estimation has more dispersed subsample sizes $n_{h,opt.}$ than the allocation that is optimal for direct estimators.

3.2 First Taylor Series Approximation When $G > 0$

To incorporate priority for national estimation in optimizing design for small area estimation, we set the relative priority G to positive values. Unfortunately, this optimization has no simple closed form solution (Molefe, 2012). The solution can be expressed as a quartic equation. Analytic solutions can be found to quartic equations but finding the solution would be convoluted and difficult to interpret. Also, there are up to 4 real-valued solutions.

Another approach would be to find a Taylor series approximation based on ρ close to 0 and then minimize this with respect to n_h. Instead, we note that the optimal n_h depends on ρ; one could consider n_h to be a function of this quantity and write $n_h = n_h(\rho)$.The approximation will be

$$n_h \approx n_h(0) + n_h'(0)\rho + \frac{1}{2}n_h''(0)\rho^2$$

An explicit expression for $n_h(\rho)$ does not exist, so we cannot obtain $n_h'(0)$ and $n_h''(0)$ by direct differentiation of $n_h(\rho)$ with respect to ρ. Instead, our approach is to obtain these derivatives indirectly, by differentiating both sides of (7) and (8):

$$L = \sum_{h \in U^1} N_h^q \rho \big[1 + (n_h - 1)\rho\big]^{-1} + G N_+^{(q)} \sum_{h \in U^1} P_h^2 n_h^{-1} + \lambda \left(\sum_{h \in U^1} n_h - n \right)$$

Now

$$0 = \frac{\partial L}{\partial n_h} = -N_h^q \rho^2 \big[1 + (n_h - 1)\rho\big]^{-2} - G N_+^{(q)} P_h^2 n_h^{-2} + \lambda \tag{7}$$

$$0 = \frac{\partial L}{\partial \lambda} = \sum_{h \in U^1} n_h - n \tag{8}$$

We obtain the following results: $n_h(0) = nP_h$, $n_h'(0) = 0$, $n_h''(0) = n^3 P_h (G N_+^{(q)})^{-1} \Big\{ N_h^q - N^{-1} \sum_{h \in U^1} N_h^{q+1} \Big\}$ (See Molefe (2012) for details).

Hence when $\rho \approx 0$, an approximate stationary point for this optimization is:

$$n_h \approx nP_h \left(1 + \frac{1}{2} \rho^2 n^2 (G N_+^{(q)})^{-1} \left\{ N_h^q - N^{-1} \sum_{h \in U^1} N_h^{q+1} \right\} \right)$$

The approximate solution is a function of G, ρ and q. When G approaches ∞ the approximate solution for n_h tends to $n_h \approx nP_h$, which is proportional allocation. When G is large, priority is given to estimation of the national mean, hence this is as would be expected, since proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When $G = 0$ the approximate solution is undefined since division by zero is undefined. The approximate solution is therefore not suitable or appropriate when the only goal is small area estimation. When ρ approaches 0 the approximate solution is approximately equal to $n_h \approx nP_h$. When $\rho \approx 0$, units within a small area are somewhat similar to each other for the variable of interest but that the degree of similarity is very very low. When this happens it is natural for small areas to be represented in proportion to their population sizes.

When $q = 1$ or 2, it is not clear what the value of the approximate solution will be. The value of n_h depends on the magnitude and whether $N_h^q - N^{-1} \sum_{h \in U^1} N_h^{q+1}$ is positive or negative. We obtain large positive and negative values of n_h depending on the population size of the stratum. For relatively smaller strata, the result is

large negative values which would in practice be truncated at zero and the opposite is true for relatively large strata.

3.3 Second Taylor Series Approximation When $G > 0$

The approximate analytical optimal design based on $\rho \approx 0$ gave counter-intuitive results, particularly when G is small or zero. Hence we are now going to approximate n_h based on a different quantity based on both ρ and G rather than on ρ only, say, $n_h = n_h(\alpha)$ where $\alpha = f(\rho, G) = \rho(GN_+^{(q)})^{-1}N^q$. Our interest is the case where α is small. The problem is to minimize

$$F = \sum_{h \in U^1} N_h^q \rho[1 + (n_h - 1)\rho]^{-1} + GN_+^{(q)} \sum_{h \in U^1} P_h^2 n_h^{-1}$$

with respect to n_h subject to $\sum_{h \in U^1} n_h = n$. This is equivalent to minimizing

$$F = \alpha \sum_{h \in U^1} P_h^q [1 + (n_h - 1)\rho]^{-1} + \sum_{h \in U^1} P_h^2 n_h^{-1}$$

The partial derivatives of the corresponding Lagrangian function with respect to n_h and λ are, respectively,

$$0 = \frac{\partial L}{\partial n_h} = -\alpha P_h^q [1 + (n_h - 1)\rho]^{-2} - P_h^2 n_h^{-2} + \lambda \tag{9}$$

$$0 = \frac{\partial L}{\partial \lambda} = \sum_{d \in U^1} n_h - n \tag{10}$$

Equations (9) and (10) are easily solved when $\alpha = 0$. As with the first Taylor approximation, indirect differentiation is used because $n_h(\alpha)$ is not available explicitly.

Let n_h be the solution of (9) and (10) for any given value of α. We can then approximate $n_h(\alpha)$ by

$$n_h \approx n_h(0) + n_h'(0)\alpha$$

An approximate stationary point for this optimization problem when $\alpha \approx 0$ is

$$n_h \approx nP_h\left(1+\frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}N^q\left\{\frac{P_h^q}{\left[1+(nP_h-1)\rho\right]^2}-\sum_{h\in U^1}\frac{P_h^{q+1}}{\left[1+(nP_h-1)\rho\right]^2}\right\}\right) \tag{11}$$

(See Molefe, 2012, page 120 Theorem 3.7.1).

In the approximation in 3.2, which was based on $\rho \approx 0$, we obtained large positive or negative values of n_h when n was large. Here, as n approaches ∞ the approximate sample size is equal to:

$$\begin{aligned} n_h &\approx nP_h\left(1+\frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}N^q\left\{P_h^q\big(nP_h\rho\big)^{-2}-\sum_{h\in U^1}P_h^{q+1}\big(nP_h\rho\big)^{-2}\right\}\right) \\ &= nP_h\left(1+\frac{1}{2}(GN_+^{(q)})^{-1}N^q\left\{P_h^{q-2}-\sum_{h\in U^1}P_h^{q-1}\right\}\right) \end{aligned}$$

which seems more reasonable.

When $q=0$ and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GH)^{-1}N^0\left\{P_h^{-2}-\sum_{h\in U^1}P_h^{-1}\right\}\right)$$

where $H = N_+^{(0)} = \sum_{h\in U^1} N_h^0$.

When $q=1$ and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2G}\left\{P_h^{-1}-\sum_{h\in U^1}P_h^0\right\}\right)$$

When $q=2$ and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GN_+^{(2)})^{-1}N^2\left\{P_h^0-\sum_{h\in U^1}P_h^1\right\}\right) = nP_h$$

A priority exponent of $q=2$ implies proportional allocation, hence the result is as expected.

When G approaches ∞ the approximate sample size is equal to $n_h \approx nP_h$. This result is as expected since very large G implies more priority for national estimation.

Proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When G approaches 0, this corresponds to α approaching ∞, and the approximate solution is undefined. This means that the alternative approximate analytical optimal design for n_h breaks down as G approaches zero. Perhaps this is not surprising, as our approximation is based on small α not large α.

When ρ approaches 0 the approximate analytical design is equal to $n_h \approx nP_h$.

3.4 Power Allocation

Power allocation was considered by Bankier (1988). The within-stratum sample sizes are proportional to N_h^p given by

$$n_h = \frac{nN_h^p}{\sum_{h \in U^1} N_h^p} \tag{12}$$

for $h = 1, \ldots, H$, where $0 \leq p \leq 1$. A special case is the square root allocation when $p = \frac{1}{2}$. The exponent p is called the power of the allocation. Setting $p = 1$ results in the Neyman allocation and $p = 0$ results in equal allocation.

Bankier (1988) proposed choosing p based on perceived relative priorities. However, this was based on direct estimates being used in each stratum. We are interested in the case where composite estimation is to be used, and the objective is to obtain a low value for F in (5).

It appears it is intractable to derive analytically an optimal value of p for F given by (5). Instead we obtain the best power allocation by numerical optimization.

3.5 Ad-hoc Allocations

We also consider two sensible but ad-hoc allocations that include a design constructed as a mix of the optimal design when $G = 0$ and proportional allocation

(special case of power allocation when $p = 1$). The weighting of the two allocations is based on the relative priority coefficient G.

- Mixed Design

$$n_h = \frac{1}{G+1}\left\{\frac{n\sqrt{N_h^q}}{\sum_{h\in U^1}\sqrt{N_h^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{N_h^q}}{H^{-1}\sum_{h\in U^1}\sqrt{N_h^q}} - 1\right)\right\} + \frac{G}{G+1}nP_h$$

- Logarithmic Mixed design

$$n_h = \frac{1}{log(G+1)+1}\left\{\frac{n\sqrt{N_h^q}}{\sum_{h\in U^1}\sqrt{N_h^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{N_h^q}}{H^{-1}\sum_{h\in U^1}\sqrt{N_h^q}} - 1\right)\right\} + \frac{log(G+1)}{log(G+1)+1}nP_h$$

4. SAMPLE ALLOCATION

We use data on the 26 cantons of Switzeland (Longford, 2006); their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million. Throughout, we assume that $n = 10,000$, $\rho = 0.025$ and $\sigma^2 = 100$.

We use Longford's algorithm (Longford, 2006) to obtain a numerical optimum sample allocation using data on the 26 cantons of Switzerland.

To compare the efficiency of these designs, we consider the relative efficiency of the various designs, relative to a standard design, equal allocation, by computing ratios of the F values for a particular design to the value of a base design, equal allocation, for priority exponent $q = 1, 2$ and relative priority coefficients $G = \{0, 5, 10, 50, 100, 200\}$. A ratio less than one implies that a design is more efficient than the base design, whilst a ratio greater than one implies a design is less efficient than the base design.

In Table 1 we see that all the allocations are more efficient than equal allocation for all G. When $G = 0$, the area-only optimum (i.e. the optimal design based on $G = 0$), numerical optimum, optimum power allocation, the mixed design and logarithmic mixed allocations are the best designs, followed by proportional allocation, with equal allocation doing the worst.

When $G = 5$ optimum power allocation is the best design followed by the numerical optimum. The area-only optimum is also not performing badly. The second Taylor approximation is the worst design. In principle, the numerical optimum should have been the most efficient algorithm, but in some cases it was very slightly less efficient than other options, as this algorithm does not perfectly handle the fixing of values of n_h at the boundaries.

When $G = 10$, the second Taylor approximation performs worse than all the designs except the base design.The optimum power allocation is the best design in this case, slightly better than the numerical optimum. However, the area-only optimum becomes worse than other designs as G increases. The relative efficiencies of the mixed, logarithmic mixed, optimum power and proportional allocations are comparable to the numerical optimum and the second Taylor approximation.

When $G = 50$, the second Taylor approximation, optimum power and logarithmic mixed allocations are the best designs.

The other allocations perform as well as the numerical optimum when $G \geq 50$ with the exception of area-only optimum.

Table 1: Relative efficiency of stratified designs for $q = 1$

	Priority Coefficient (G)					
Designs	0	5	10	50	100	200
Equal allocation	1.000	1.000	1.000	1.000	1.000	1.000
Proportional allocation	0.887	0.766	0.701	0.562	0.529	0.510
Area-only optimum	0.786	0.718	0.682	0.604	0.585	0.575
Numerical optimum	0.788	0.715	0.670	0.560	0.529	0.510
First Taylor approximation	-	-	-	-	-	-
Second Taylor approximation	-	3.392	0.743	0.558	0.528	0.509
Optimum power allocation	0.786	0.712	0.668	0.558	0.528	0.509
Mixed allocation	0.786	0.743	0.691	0.562	0.529	0.509
Logarithmic mixed allocation	0.786	0.726	0.676	0.558	0.528	0.510

In Table 2 we observe that when $G = 0$ the area-only optimum, mixed and logarithmic mixed allocations are the best, marginally better than the numerical optimum. The relative efficiency for numerical optimum, proportional and optimum power allocation is the same. We also observe that the area-only optimum relative efficiency worsens as G increases.

In Tables 3 and 4 we present the percentiles of the sample sizes when $G = 0$ for $q = 1$ and 2. The sample allocation for mixed and logarithmic mixed allocations are the same in this case. We also observe that when $q = 2$ the area-only optimum, mixed and logarithmic allocations have $n_h = 0$ for some strata which is undesirable. The first Taylor and second Taylor approximations are not applicable when $G = 0$.

Table 2: Relative efficiency of stratified designs for $q = 2$

	Priority Coefficient (G)					
Designs	0	5	10	50	100	200
Equal allocation	1.000	1.000	1.000	1.000	1.000	1.000
Proportional allocation	0.493	0.492	0.491	0.489	0.488	0.488
Area-only optimum	0.488	0.493	0.496	0.502	0.503	0.504
Numerical optimum	0.493	0.492	0.491	0.489	0.488	0.488
First Taylor approximation	-	-	-	-	-	-
Second Taylor approximation	-	0.500	0.491	0.489	0.488	0.488
Optimum power allocation	0.493	0.492	0.491	0.489	0.488	0.488
Mixed allocation	0.488	0.491	0.490	0.489	0.488	0.488
Logarithmic mixed allocation	0.488	0.490	0.490	0.489	0.489	0.488

Table 3: Sample sizes of stratified designs for $G = 0$ and $q = 1$

	Percentiles of n_h				
Designs	Min	1st Quarter	Median	3rd Quarter	Max
Equal allocation	384.60	384.60	384.60	384.60	384.60
Proportional allocation	20.66	96.42	285.80	470.00	1693.00
Area-only optimum	72.00	200.50	372.50	489.20	963.00
Numerical optimum	100.50	217.20	373.50	479.40	910.10
First Taylor approximation	-	-	-	-	-
Second Taylor approximation	-	-		-	-
Optimum power allocation	84.82	200.10	366.20	483.50	987.70
Mixed allocation	71.74	200.20	372.40	489.00	963.40
Logarithmic mixed allocation	71.74	200.20	372.40	489.00	963.40

Table 4: Sample sizes of stratified designs for $G = 0$ and $q = 2$

	Percentiles of n_h				
Designs	Min	1st Quarter	Median	3rd Quarter	Max
Equal allocation	384.60	384.60	384.60	384.60	384.60
Proportional allocation	20.66	96.42	285.80	470.00	1693.00
Area-only optimum	0.00	67.50	275.00	478.20	1823.00
Numerical optimum	20.66	96.42	285.80	470.00	1693.00
First Taylor approximation	-	-	-	-	-
Second Taylor approximation	-	-	-	-	-
Optimum power allocation	20.67	96.43	285.80	470.00	1693.00
Mixed allocation	0.00	67.04	275.30	477.90	1823.00
Logarithmic mixed allocation	0.00	67.04	275.30	477.90	1823.00

In Tables 5 and 6 we present the percentiles of the sample sizes for $G = 10$ and $G = 100$, respectively, for $q = 1$ which can be interpreted to mean that larger small areas being are somewhat important. The sample sizes for the least populous cantons are boosted in relation to proportional allocation at the expense of relatively larger cantons when $G = 10$. When $G = 100$ we observe that the sample size allocation converges to proportional allocation.

Table 5: Sample sizes of stratified designs for $G = 10$ and $q = 1$

	Percentiles of n_h				
Designs	Min	1st Quarter	Median	3rd Quarter	Max
Equal allocation	384.60	384.60	384.60	384.60	384.60
Proportional allocation	20.66	96.42	285.80	470.00	1693.00
Area-only optimum	72.00	200.50	372.50	489.20	963.00
Numerical optimum	100.50	217.20	373.50	479.40	910.10
First Taylor approximation	-	-	-	-	-
Second Taylor approximation	62.49	265.20	421.30	483.50	684.30
Optimum power allocation	84.82	200.10	366.20	483.50	987.70
Mixed allocation	71.74	200.20	372.40	489.00	963.40
Logarithmic mixed allocation	71.74	200.20	372.40	489.00	963.40

Table 6: Sample sizes of stratified designs for $G = 100$ and $q = 1$

	Percentiles of n_h				
Designs	Min	1st Quarter	Median	3rd Quarter	Max
Equal allocation	384.60	384.60	384.60	384.60	384.60
Proportional allocation	20.66	96.42	285.80	470.00	1693.00
Area-only optimum	72.00	200.50	372.50	489.20	963.00
Numerical optimum	55.73	139.50	307.50	463.20	1477.00
First Taylor approximation	-	-	-	-	-
Second Taylor approximation	29.03	130.20	312.90	472.70	1491.00
Optimum power allocation	30.02	117.80	308.80	480.40	1498.00
Mixed allocation	21.66	98.45	287.50	470.40	1679.00
Logarithmic mixed allocation	31.02	117.50	303.40	473.90	1545.00

Figure 1 shows a graphical display of the sample size distribution by numerical optimization in R using Longford's algorithm for various values of G and q. When $G = 0$ and $q = 0$ each canton is allocated the same sample size of $n_h = 10,000/26 = 385$. When $q = 2$, 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, at the expense of reduced allocation to the most populous cantons.

As G increases we observe that the distances between the curves that connect the optimal sample size reduces, especially for smaller cantons. This shows that the priority exponent q has minimal impact on the sample allocation for very large G.

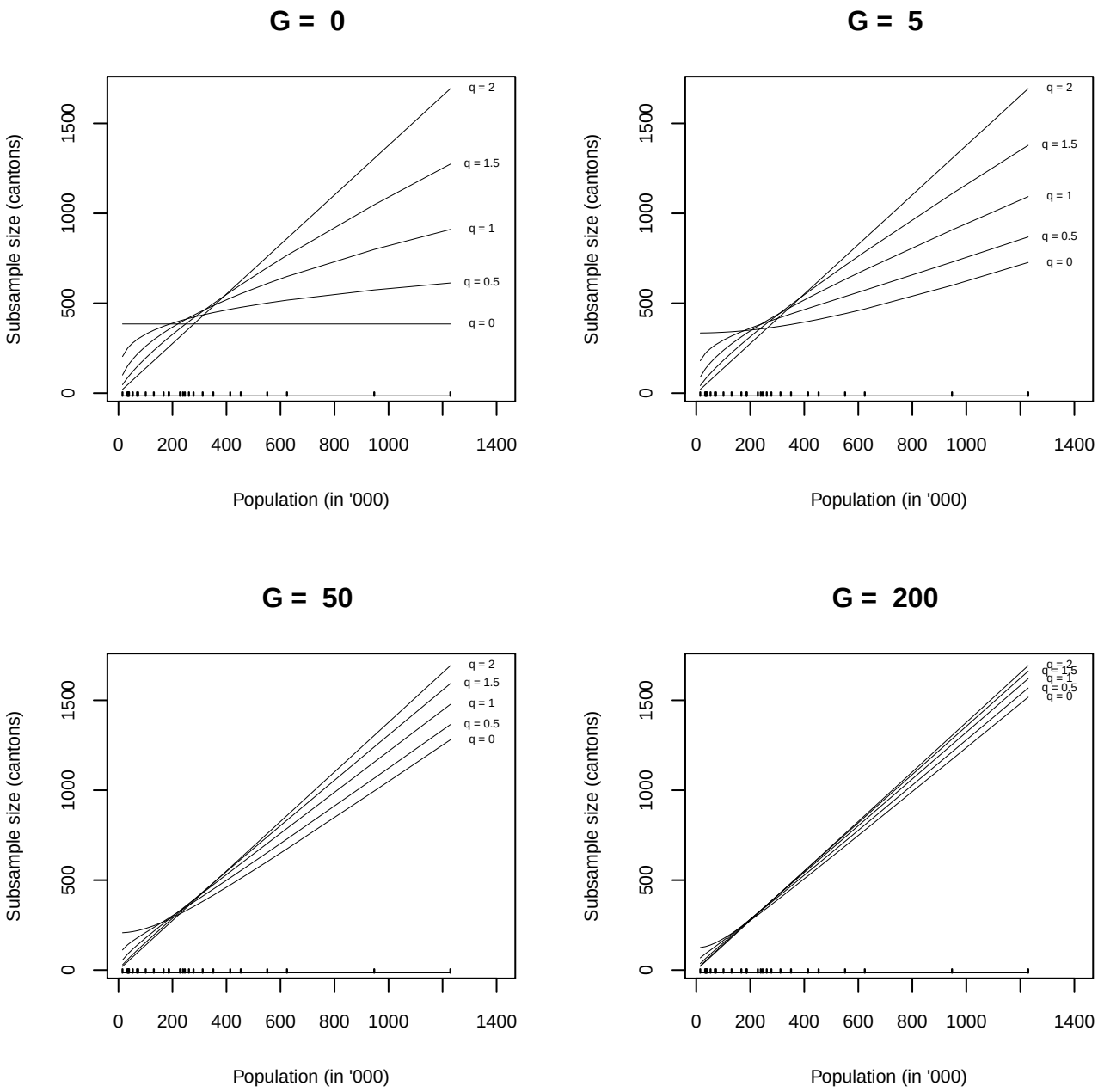

Figure 1: Numerical optimum sample distribution by relative priority coefficient

In Table 7 we show the numerical optimum value of p, the power of the allocation. For each relative priority coefficient G we observe that as q approaches 2, the optimum p approaches 1. When G is large, the optimum p is 1 even for small values of the priority exponent q. Also, when $G = 0$, the optimum $p \approx \frac{q}{2}$.

Table 7: Numerical optimum value of p

Priority Exponent	Priority Coefficient G					
q	0	5	10	50	100	200
0.00	0.000	0.192	0.293	0.597	0.730	0.839
0.25	0.138	0.300	0.392	0.677	0.794	0.882
0.50	0.277	0.416	0.500	0.756	0.852	0.917
0.75	0.417	0.537	0.612	0.828	0.899	0.945
1.00	0.557	0.658	0.721	0.887	0.936	0.966
1.25	0.698	0.776	0.823	0.934	0.963	0.981
1.50	0.837	0.886	0.912	0.970	0.983	0.991
1.75	0.975	0.984	0.988	0.996	0.998	0.999
2.00	1.000	1.000	1.000	1.000	1.000	1.000

5. CONCLUSIONS

The anticipated MSE is a sensible objective criteria for sample design, because the particular sample which will be selected is not available in advance of the survey. Hence a criteria which averages over all possible samples is appropriate. Särndal et al. (1992, Chapter 14) base their optimal designs on the anticipated variance, which similarly averages over both model realizations and sample selection, although they consider only approximately design-unbiased estimators.

An analytical solution for the stationary point exists when the only priority is small area estimation. However, there are difficulties in applying it because when the strata have disparate population sizes, the stationary point gives negative sample sizes, so that the optimum must be obtained numerically. The numerical optimum then has some strata with $n_h = 0$ which is also not desirable.

When priority is given to national estimation as well as to small area estimation so that $G > 0$, two approximate solutions were derived, based on $\rho \approx 0$, and $\alpha = f(\rho, G) = \rho(GN_+^{(q)})^{-1}N^q \approx 0$. Both had undesirable properties, giving very large positive and negative sample sizes in some cases.

An optimal power allocation, where $n_h \propto N_h^p$, and p is obtained numerically to minimize the objective function, has much better practical properties, is easier to calculate and would seem a more natural design by most survey statisticians. It is only slightly less efficient than the numerical optimum design. The mixed design and logarithmic mixed design are also found to be efficient and easy to calculate.

ACKNOWLEDGEMENTS: The authors wish to thank Professors Raymond Chambers and David Steel for their helpful suggestions to improve on this paper

REFERENCES

Bankier, M. D. (1988), "Power Allocations: Determining Sample Sizes for Subnational Areas," *The American Statistician*, 42, 174–177.

Brock, D. B., French, D. K., and Peyton, B. W. (1980), "Small Area Estimation: Empirical Evaluation of Several Estimators for Primary Sampling Units," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 766–771.

Ericksen, E. P. (1973), "Recent Developments in Estimation for Local Areas," in *Proceedings of the Section on Social Statistics, American Statistical Association*, pp. 37–41.

Fuller, W. A. (1999), "Environmental Surveys Over Time," *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331–345.

Gonzalez, M. E. (1973), "Use and Evaluation of Synthetic Estimates," in *Proceedings of the Section on Social Statistics, American Statistical Association*, pp. 33–36.

Longford, N. T. (2006), "Sample Size Calculation for Small-Area Estimation," *Survey Methodology*, 32, 87–96.

Marker, D. A. (2001), "Producing Small-Area Estimates from National Surveys: Methods of Minimizing use of Indirect Estimators," *Survey Methodology*, 27, 183–188.

Molefe, W. B. (2012), "Sample Design for Small Area Estimation," Ph.D. thesis, University of Wollongong, http://ro.uow.edu.au/theses/3495.

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

Royall, R. M. (1973), "Discussion of two Papers on Recent Developments in Estimation of Local Areas," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 43–44.

Särndal, C., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag.

Schaible, W. L. (1978), "Choosing Weight for Composite Estimators for Small Area Statistics," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 741–746.

Singh, M. P., Gambino, J., and Mantel, H. J. (1994), "Issues and Strategies for Small Area Data," *Survey Methodology*, 20, 3–22.