R&D Report 2002:1
Research – Methods – Development

Statistiska centralbyrån Statistics Sweden

On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys

Anders Holmberg

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 2002:1. On the choice of sampling design under GREG estimation in multiparameter surveys / Anders Holmberg.
Digitalt skapad fil, anpassad efter de digitaliserade delarna i serien. Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-2002-X101OP0201

R&D Report **2002:1**
Research – Methods – Development

On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys

Anders Holmberg

R&D Report 2002:1

Research - Methods - Development

On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys

---

Från trycket Februari 2002

Producent Statistiska centralbyrån, *Statistics Sweden,* metodenheten
Box 24300, *SE*-104 51 STOCKHOLM

Förfrågningar Anders Holmberg
anders.holmberg@scb.se
telefon 019- 17 69 05

ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 2002

On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys

Anders Holmberg

ABSTRACT

At the design and estimation stages of a survey, large survey organizations often use auxiliary information. Technological advances in data capture and a better accessibility of registers open up for an increased and more efficient use of such information. This paper addresses issues of how to use auxiliary information efficiently in sampling from finite populations. Previous results regarding the choice of optimal design are extended to the case of several study variables. We suggest approaches to achieve a high overall efficiency, and compare these approaches with single-variable routines, often used by practising survey statisticians

---

Aim with this report

The Pareto πps sampling scheme is already in use at various of Statistics Sweden's surveys. In an earlier report in this series by Rosén (R&D report 2000:5) the combination of Pareto πps and GREG estimation is put forward as a sampling strategy with particularly good properties. In this report, we also end up in studying strategies that combine πps sampling and GREG estimation. However, in this report we address the problem of choosing an efficient design from a multiparameter perspective. Compared to "optimal" designs in the single-parameter case, a multiparameter perspective requires some sort of compromise approach. We present three approaches to achieve an (overall) efficient compromise design for a multiparameter survey. Moreover, we stress the importance of good planning routines in multiparameter surveys, and in a numerical example we outline how the survey statistician diagnostically can compare different design alternatives considered at the planning stage.

Contents

On the choice of sampling design under GREG estimation in multiparameter surveys.

Anders Holmberg
R&D Department, Statistics Sweden, SE-701 89 Örebro, Sweden.

Abstract

At the design and estimation stages of a survey, large survey organizations often use auxiliary information. Technological advances in data capture and a better accessibility of registers open up for an increased and more efficient use of such information. This paper addresses issues of how to use auxiliary information efficiently in sampling from finite populations. Previous results regarding the choice of optimal design are extended to the case of several study variables. We suggest approaches to achieve a high overall efficiency, and compare these approaches with single-variable routines, often used by practising survey statisticians.

Key Words and Phrases: Auxiliary information, GREG Estimator, Optimal designs, Survey planning.

1 Introduction

A sample survey is in general taken with the purpose of estimating a large set of parameters $\theta_1, \theta_2, \ldots$ (totals, means, ratios, medians, Gini coefficients etc.) of a finite population $U = \{1, \ldots, k, \ldots, N\}$. The most important of these parameters, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_i, \ldots, \theta_I)'$, form the basis for planning the survey. The task of the survey statistician is then to find an efficient combination of sampling design $p(\cdot)$, and estimator vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_i, \ldots, \hat{\theta}_I)'$, (efficient strategy $\Omega_{p,\hat{\theta}} = \left[p(\cdot),\ \hat{\boldsymbol{\theta}}\right]$), i.e., such that the final choice results in 'small' mean square error for each estimator $\hat{\theta}_i$.

When there is only one parameter θ to estimate, say a population total $t = \sum_{k \in U} y_k = \sum_U y_k$, the statistician might start by choosing a suitable member $\hat{\theta}$ of a specific class of estimators known to have good properties, e.g. GREG (generalized regression) estimators, followed by an attempt to find a sampling design that minimizes the mean square error of $\hat{\theta}$, or variance if $\hat{\theta}$ is unbiased.

There is no simple straightforward generalization of this single parameter approach to the multiparameter case. However, in this paper we will discuss a couple of approaches that might be useful to achieve high overall efficiency.

The paper has the following structure. In the following preliminaries we give the background to the problem, introduce our basic notations and specify the survey situation that is considered. In section 2 we treat the single parameter case and give an overview of results

on the choice of 'optimal' designs. It lays the basis of section 3, which contains results of our approaches concerning the multiparameter case. Both section 2 and 3 naturally leads to treatment of without replacement probability proportional-to-size sampling, (πps sampling), combined with GREG estimation. Since the survey statistician needs feasible methods to implement theory, and since relatively recent progress in the area of πps sampling has been made, a short overview is provided in section 4. Comparisons between our suggested approaches are made in section 5. In section 6, a further extension of the multiparameter approaches is outlined, and, finally, our conclusions and recommendations are given in section 7.

1.1 Background and notation

Many results in survey theory address the problem of finding an efficient strategy. Those results rely on the availability of auxiliary information (e.g. the literature on optimum allocation in stratified sampling, on πps sampling designs or on estimators using auxiliary variables.) This paper assumes that there are P auxiliary variables accessible at the planning stage. They are denoted $u_1, \ldots, u_p, \ldots, u_P$, and their values u_{pk}, $(p = 1, \ldots, P)$, are known for every element k in the population. The vector of the most important parameters, $\boldsymbol{\theta}$, often consists of functions of the population totals of the unknown study variables, $y_1, \ldots, y_q, \ldots, y_Q$, i.e. $\boldsymbol{\theta} = (f_1(\mathbf{t}), f_2(\mathbf{t}), \ldots, f_i(\mathbf{t}), \ldots, f_I(\mathbf{t}))'$ where $\mathbf{t} = (t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q})'$ Henceforth, we will consider the case where $\boldsymbol{\theta} = \mathbf{t}$ with a corresponding estimator vector $\hat{\boldsymbol{\theta}} = \hat{\mathbf{t}}$ (i.e. $I = Q$.)

When we plan our survey strategy, we try to use the auxiliary information in the choice of *design* as well as in the choice of *estimator,* in such a way that the sampling error of $\hat{\mathbf{t}}$ becomes as small as possible. Statistical models can be used as an aid in this planning process. Hence, we assume that the statistician has useful a priori knowledge about the relations between the study variables y_q and the auxiliary variables.

We presume the structure of these relations makes it relevant to formulate linear models, ξ_q, $(y_{qk} = \mathbf{x}'_{qk}\boldsymbol{\beta}_q + \varepsilon_{qk})$ for the study variables, with $E_{\xi_q}(\varepsilon_{qk}) = 0$, $V_{\xi_q}(\varepsilon_{qk}) = \sigma^2_{qk}$ and $E_{\xi_q}(\varepsilon_{qk}\varepsilon_{ql}) = 0 \quad (k \neq l)$, i.e.,

$$\begin{aligned} E_{\xi_q}(y_{qk}) &= \mathbf{x}'_{qk}\boldsymbol{\beta}_q \qquad (k = 1, \ldots, N) \\ V_{\xi_q}(y_{qk}) &= \sigma^2_{qk} \qquad (k = 1, \ldots, N) \end{aligned} \tag{1}$$

where $\mathbf{x}'_{qk} = (x_{1_qk}, \ldots, x_{j_qk}, \ldots, x_{J_qk})$ is a suitable set of J_q (positive) auxiliary variables formed from $u_1, \ldots, u_p, \ldots, u_P$, $\boldsymbol{\beta}_q = \left(\beta_{1_q}, \ldots, \beta_{j_q}, \ldots, \beta_{J_q}\right)'$ are model parameters. The values $\sigma^2_{q1}, \ldots, \sigma^2_{qN}$ are considered known, although knowledge up to a constant multiplier sometimes is sufficient.

To select a set sample $s \subseteq U$ of size n_s, a without replacement sampling design $p(\cdot)$ will be used. We denote the first-order inclusion probabilities by π_k $(k = 1, \ldots, N)$ and the second-order inclusion probabilities by π_{kl} $(k, l = 1, \ldots, N)$. (Whenever necessary, a dot, i.e. $\dot{\pi}$, is used to emphasize inclusion probabilities that depend on assumed or approximated numerical values of σ^2_{qk}.)

The population total of y_q, $t_{y_q} = \sum_U y_{qk}$, can be estimated by the GREG estimator, which is defined as

$$\hat{t}_{y_q r} = \hat{t}_{y_q \pi} + (\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q \pi})' \hat{\mathbf{B}}_q \tag{2}$$

Here, $\hat{t}_{y_q \pi} = \sum_{k \in s} y_{qk}/\pi_k = \sum_s y_{qk}/\pi_k$ is the well known Horvitz-Thompson or π estimator, $\mathbf{t}_{x_q} = \left(t_{x_{1q}}, \ldots, t_{x_{Jq}}, \ldots, t_{x_{J_q}}\right)'$, i.e. a J_q-dimensional vector of x_q totals, $\hat{\mathbf{t}}_{x_q \pi}$ is a vector of the corresponding π estimators and

$$\hat{\mathbf{B}}_q = \left(\sum_s \frac{\mathbf{x}_{qk} \mathbf{x}_{qk}'}{c_{qk} \pi_k} \right)^{-1} \sum_s \frac{\mathbf{x}_{qk} y_{qk}}{c_{qk} \pi_k} \tag{3}$$

is an estimated vector of regression coefficients, where c_{qk} is a suitable constant.

Moreover, the Taylor expansion variance of $\hat{t}_{y_q r}$ is given by

$$V_T(\hat{t}_{y_q r}) = \sum_{k \in U} \sum_{l \in U} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_k \pi_l} E_{qk} E_{ql} \tag{4}$$

where $E_{qk} = y_k - \mathbf{x}_{qk}' \mathbf{B}_q$ $(k = 1, \ldots, N)$ are population fit residuals, with $\mathbf{B}_q = \left(\sum_U \mathbf{x}_{qk} \mathbf{x}_{qk}'/c_{qk}\right)^{-1} \sum_U y_{qk} \mathbf{x}_{qk}'/c_{qk}$ a finite population regression coefficient. (Details of GREG estimation are given in Särndal, Swensson and Wretman (1992) sections 6.4-6.7.)

Henceforth, the results in this paper concern the family of GREG estimators which includes several well known estimators used in practice, (e.g. the post-stratified estimator and the common ratio estimator.) A GREG estimator is approximately unbiased even with poorly fitted models, but with a strong relationship between y and $\mathbf{x}$, and a fair knowledge of that relationship, the GREG estimator will outperform the π estimator as far as efficiency is concerned. However, other types of estimators could also be mentioned when we discuss strategy selection using auxiliary information.

An estimator related to the GREG estimator is the 'optimal regression estimator', $\hat{t}_{y_q or}$, (Rao (1992, 1994, 1997), Cassady and Valiant (1993) and Montanari (1987, 1998)),

$$\hat{t}_{y_q or} = \hat{t}_{y_q \pi} + (\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q \pi})' \tilde{\mathbf{B}}_q$$

where $\tilde{\mathbf{B}}_q = \left[\hat{V}(\hat{\mathbf{t}}_{x_q \pi})\right]^{-1} \hat{C}(\hat{\mathbf{t}}_{x_q \pi}, \hat{t}_{y_q \pi})$, with $\hat{V}(\hat{\mathbf{t}}_{x_q \pi})$ and $\hat{C}(\hat{\mathbf{t}}_{x_q \pi}, \hat{t}_{y_q \pi})$ being unbiased estimators of $V(\hat{\mathbf{t}}_{x_q \pi})$ and $C(\hat{\mathbf{t}}_{x_q \pi}, \hat{t}_{y_q \pi})$ with dimensions $J_q \times J_q$ and $J_q \times 1$ respectively. (Expressions for the well known Horvitz-Thompson or alternatively the Sen-Yates-Grundy variants of $\hat{V}(\hat{\mathbf{t}}_{x_q \pi})$ and $\hat{C}(\hat{\mathbf{t}}_{x_q \pi}, \hat{t}_{y_q \pi})$ can be found in Särndal et al. pp 44-45, 170.) To use the variance and covariance estimators in point estimation can be very impractical and sometimes it leads to trouble. Nevertheless, Montanari (1998) discusses some situations when $\hat{t}_{y_q or}$ can be preferred to the GREG estimator.

Another more useful and wider family of estimators, are the calibration estimators described in DeVille and Särndal (1992), Lundström and Särndal (1999) and Estevao and Särndal (2000). They are asymptotically equivalent to the GREG estimator, and they are appealing to practitioners in attempts to reduce non-response bias.

Nonetheless, in the following neither calibration estimators nor $\hat{t}_{y_{qor}}$, will be discussed. This restricts our strategy concept to strategies where the estimator is a member of the GREG estimator family, (different GREG estimators for different parameters are allowed.) Since the above estimators are related to the GREG estimator, we believe that this is a mild restriction. In the planning stage of a multiparameter survey, the choice of design, which affects all parameter estimates, is likely to be more important than the choice between related estimators. Therefore, if we can find an 'optimal' *design* for an efficient estimator such as a GREG estimator, such a design is apt to work well combined with the other estimators as well. (As to the choice of a *specific* GREG estimator, it is henceforth a tacit understanding that when a model like (1) is explicitly presented; the model is satisfactory, and the information given is good enough for the statistician to make a suitable GREG estimator choice.)

Note that we consider design-based inference, but we try to make as efficient use of supplementary information as possible through models. Our approach will be model assisted, and we use models to assist in our design selection and in choosing estimators for $\mathbf{t}$. In the next section we reproduce results on finding an 'optimal' design for a GREG estimator in the single parameter case.

2 Selecting an optimal design in the single parameter case.

If we consider GREG estimation, the question of which strategy to choose can be limited to the issue of finding a design that minimizes $V(\hat{t}_{y_{qr}})$. However, direct minimization of $V(\hat{t}_{y_{qr}})$ is impossible, but given a model ξ_q, and utilizing the statistical properties of the model errors ε_{qk}, we can try to minimize the anticipated variance, (i.e. the variance over both the model ξ_q and the design p, e.g. see Isaki and Fuller (1982)).

$$E_{\xi_q} E_p[(\hat{t}_{y_{qr}} - t_{y_q})^2] - [E_{\xi_q} E_p(\hat{t}_{y_{qr}} - t_{y_q})]^2$$

If the model ξ_q is well specified, then an approximation to the anticipated variance can be written as (see Särndal et al. pp 450-451),

$$ANV_q(\hat{t}_{y_{qr}}) = \sum\nolimits_U (\pi_k^{-1} - 1)\sigma_{qk}^2 \tag{5}$$

For a sampling design $p(\cdot)$ such that $E_p(n_s) = n$, Result 12.1.1. in Särndal et al. show that a near 'optimal' design, i.e. a design that minimizes (5), is such that the first-order inclusion probabilities for $k = 1, \ldots, N$ are given by

$$\pi_k = \pi_{q(opt)k} = n\sigma_{qk} / \sum\nolimits_U \sigma_{qk} \tag{6}$$

and the minimum of $ANV_q(\hat{t}_{y_{qr}})$ is

$$\begin{aligned} ANV_{q\min}(\hat{t}_{y_{qr}}) &= \sum\nolimits_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2 \\ &= \frac{1}{n}\left(\sum\nolimits_U \sigma_{qk}\right)^2 - \sum\nolimits_U \sigma_{qk}^2 \end{aligned} \tag{7}$$

Hence, we obtain an 'optimal' design by choosing $\pi_k \propto \sigma_{qk}$, i.e. the statistician should use a πps design with σ_{qk} as size measures. For example, if $\sigma_{qk}^2 = \sigma_q^2 u_{qk}^{\gamma_q}$, where σ_q^2 is a constant (possibly unknown) and $u_q \in \{u_1, \ldots, u_p, \ldots, u_P\}$ is one of the auxiliary variables, this result suggests that by using a well chosen GREG estimator, and choosing a design where $\pi_k \propto u_{qk}^{\gamma_q/2}$, we obtain a near 'optimal' strategy for estimating t_{y_q}.

We will return to the issues of implementing a πps design in section 4, but from here on, designs where π_k are proportional to some known size measure, z, are referred to as $\pi ps(z)$ designs. For the moment we conclude that from a theoretical view, the combination of GREG estimation and a $\pi ps(\sigma)$ design is near 'optimal' with respect to minimizing an approximation to the anticipated variance. In the single parameter case, similar conclusions can be made, (for fixed size designs), from the results in Cassel, Särndal and Wretman (1976, 1977), (Theorem 1 or Theorem 4.1 respectively) and Theorem 2.1 in Rosén (2000a).

The results above can be helpful for survey planning but the limitations are obvious. As for many other optimality results, focus is on a single study variable, which is insufficient for a practising survey statistician having to deal with several parameters of importance. (For thorough reviews on optimization problems in choosing sampling designs for surveys, see Rao (1979) and Bellhouse (1984).)

3 Selecting a 'best' overall design in the multiparameter case

More realistically, suppose that we want to estimate $\mathbf{t} = \left(t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q}\right)'$, where $Q \geq 2$, and let the relative importance of the parameters be reflected by a set of importance weights H_q, $(\sum_{q=1}^{Q} H_q = 1)$. Moreover, suppose that a good choice of GREG estimator can be made for each population total t_{y_q}. Then, the statistician's task is to find a design that in some sense could be considered optimal for all parameters.

However, one problem now is that there is no - in contrast to the single parameter case - single well-defined meaning of 'optimality'. A design that is optimal or close to optimal in the single parameter case, might not be the best choice in an overall multiparameter sense. For example, suppose a diligent statistician with a specified amount of auxiliary information, time and skill, should use the 'optimal' design result above to seek what he a priori believes to be the theoretically 'optimal' design for every parameter, t_{y_q}, $(q = 1, \ldots, Q)$, separately. Most likely, he then would find design solutions that differ, and he can only choose one of them. The statistician in such a situation is forced to seek a compromise design, which he believes works reasonably well for all parameters to be estimated.

Here, three different compromise approaches that can be used to plan the selection of a design in the multiparameter case are discussed. They all have different minimization criteria, and all are in a sense extensions of the result from the previous section. The first two approaches (A and B) are appealing at a first glance, but they have some built in scaling problems that are avoided in approach C. Since the proofs are carried out in a similar way we will only provide details for the third approach (approach C).

3.1 Approach A: Minimizing a weighted sum of variances

In the multiparameter situation, a straightforward criterion for selecting the best overall design for estimating $\mathbf{t}$ is to minimize the trace of $V(\hat{\mathbf{t}}_r)$, i.e. minimizing the sum $\sum_{q=1}^{Q} V(\hat{t}_{y_q r})$, or if we like to attach importance weights H_q, minimizing $\sum_{q=1}^{Q} H_q V(\hat{t}_{y_q r})$. Since $V(\hat{t}_{y_q r})$, $(q = 1, \ldots, Q)$, is unknown, we could instead look for the design that, under the restriction $\sum_U \pi_k = n$ and assumed models ξ_q (see eq. (1)) for $(q = 1, \ldots, Q)$, minimizes a weighted sum of approximated anticipated variances, i.e. a design that minimizes $SANV(\hat{\mathbf{t}}_r) = \sum_{q=1}^{Q} H_q ANV_q(\hat{t}_{y_q r})$. Rewriting this weighted sum as $SANV(\hat{\mathbf{t}}_r) = \sum_U (\pi_k^{-1} - 1) \sum_{q=1}^{Q} H_q \sigma_{qk}^2$, and using the Cauchy-Schwarz inequality directly yields Result 3.1

Result 3.1. *A sampling design $p(\cdot)$ with the expected sample size $E_p(n_s) = n$ that minimizes $SANV(\hat{\mathbf{t}}_r)$, is such that the first order inclusion probabilities for $k = 1, \ldots, N$ are determined by*

$$\pi_k = \pi_{(A)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2}}. \tag{8}$$

Clearly, a design with $\pi_k = \pi_{(A)k}$ is a compromise that considers all the involved parameters and their importance. However, for $q = 1, \ldots, Q$, $ANV_{qA}(\hat{t}_{y_q r})$ will differ from $ANV_{q\min}(\hat{t}_{y_q r})$, since, in general, the ratios in the inequality below will be larger than 1.

$$\frac{ANV_{qA}(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} = \frac{\sum_U (\pi_{(A)k}^{-1} - 1)\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2} \geq 1 \qquad q = 1, \ldots, Q$$

The sizes of the Q ratios between $ANV_{qA}(\hat{t}_{y_q r})$ and $ANV_{q\min}(\hat{t}_{y_q r})$ will depend on H_q, and σ_{qk}^2. In addition, since $V_{\xi_q}(y_{qk}) = \sigma_{qk}^2$ for $q = 1, \ldots, Q$, the resulting design will have properties that are dependent on the measurement scales of the variables involved in the Q models.

3.2 Approach B: Minimizing a weighted sum of relative variances

Another measure often used by statisticians in survey planning is the coefficient of variation of the estimators. In approach B we use a relative variance, $RV_q(\hat{t}_{y_q r}) = V_q(\hat{t}_{y_q r})/t_q^2$. Suppose that we want to minimize $\sum_{q=1}^{Q} H_q RV_q(\hat{t}_{y_q r})$. Again, $V_q(\hat{t}_{y_q r})$ is unattainable but we can use $ANV_q(\hat{t}_{y_q r})$. Our approximated relative variance measure then becomes $ANRV_q(\hat{t}_{y_q r}) = ANV_q(\hat{t}_{y_q r})/t_q^2$, and we seek the design that minimizes, $SANRV(\hat{\mathbf{t}}_r) = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})}{t_q^2}$.

Result 3.2. *A sampling design $p(\cdot)$ with the expected sample size $E_p(n_s) = n$ that minimizes $SANRV(\hat{\mathbf{t}}_r)$, is such that the first order inclusion probabilities for $k = 1, \ldots, N$ are determined by*

$$\pi_k = \pi_{(B)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2/t_q^2}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2/t_q^2}} \tag{9}$$

Since we are minimizing a sum of relative measures, this approach is less sensitive to different scales of the involved variables than approach A. However, if $\sigma_{qk}^2 = \sigma_q^2 f_q(u_{qk})$, the

constant multiplier σ_q^2 does not cancel out and it may differ in size between the terms in $\sum_{q=1}^{Q} H_q \sigma_{qk}^2 / t_q^2$. Moreover, the requirements on the auxiliary information with this approach are extremely high. In addition to σ_{qk}^2, the approach involves knowledge of the parameters, t_q $(q = 1, \ldots, Q)$, that we want to estimate! In practise, t_q^2, (as well as σ_{qk}^2) must be substituted by planning values $\dot{t}_q^2$, and poor guesses of t_q^2 can have large effects on the design properties.

Instead, we propose a less scale dependent approach, which also relate to how much we lose in precision compared to the single parameter 'optimal' designs, given in section 2.

3.3 Approach C: Minimizing a weighted sum of relative efficiency losses.

Variance ratios are often used to compare the efficiency of strategies. This principle is used in approach C. As a background to motivate our measure and minimization criterion, let $V(\hat{t})_{\Omega_{opt}}$ denote the estimator variance of an optimal strategy (i.e. the strategy gives the smallest possible sampling error when estimating t), and let $V(\hat{t})_{\Omega_{p,\hat{t}}}$ be the estimator variance of $\hat{t}$ for another strategy $\Omega_{p,\hat{t}}$. The relative loss in efficiency (i.e. variance increase) for one strategy compared to the optimal strategy can then be defined as $REL = (V(\hat{t})_{\Omega_{p,\hat{t}}} - V(\hat{t})_{\Omega_{opt}})/V(\hat{t})_{\Omega_{opt}}$. Then, with Q y-totals to estimate, the overall (total) relative loss in efficiency is

$$OREL = \sum_{q=1}^{Q} \frac{V(\hat{t}_{y_q})_{\Omega_{p,\hat{t}_{yq}}} - V(\hat{t}_{y_q})_{\Omega_{qopt}}}{V(\hat{t}_{y_q})_{\Omega_{qopt}}} \tag{10}$$

With $\hat{\mathbf{t}} = \hat{\mathbf{t}}_r$ we realize from section 2, that, by using a model ξ_q, (5) and (6), we can theoretically derive an 'optimal' design, with $\pi_{q(opt)k}$, $(k = 1, \ldots, N)$, for every q, $(q = 1, \ldots, Q)$. We can also calculate $ANV_{q\min}(\hat{t}_{y_qr})$ (see equation (7)) for every $\hat{t}_{y_qr}$, $(q = 1, \ldots, Q)$.

By letting $ANV_{q\min}(\hat{t}_{y_qr})$ take the place of $V(\hat{t}_{y_q})_{\Omega_{qopt}}$ in (10) we can formulate approach C as finding the design that minimizes an approximation to (a weighted) *Anticipated Overall Relative Efficiency Loss*, $(ANOREL)$, here defined for GREG estimators as

$$ANOREL = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_qr}) - ANV_{q\min}(\hat{t}_{y_qr})}{ANV_{q\min}(\hat{t}_{y_qr})} \tag{11}$$

Result 3.3. *Let $p(\cdot)$ be a sampling design with the expected sample size $E_p(n_s) = n$. Suppose that $\mathbf{t} = (t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q})'$ is estimated by $\hat{\mathbf{t}}_r = (\hat{t}_{y_1r}, \ldots, \hat{t}_{y_qr}, \ldots, \hat{t}_{y_Qr})'$ as defined by (2). The design for which the anticipated overall relative efficiency loss (11) is minimized, is such that the first order inclusion probabilities for $k = 1, \ldots, N$ are determined by*

$$\pi_k = \pi_{(C)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U \left(\frac{1}{\pi_{q(opt)k}} - 1\right)\sigma_{qk}^2}}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U \left(\frac{1}{\pi_{q(opt)k}} - 1\right)\sigma_{qk}^2}}} \tag{12}$$

where $\pi_{q(opt)k}$ is given by (6). The minimum value of ANOREL is then

$$\begin{aligned} ANOREL_{\min} &= \sum_{q=1}^{Q} H_q \frac{ANV_{qC}(\hat{t}_{y_{qr}})}{ANV_{q\min}(\hat{t}_{y_{qr}})} - \sum_{q=1}^{Q} H_q \\ &= n\sum_{q=1}^{Q} H_q \frac{\sum_U \pi_{(C)k}^{-1}\sigma_{qk}^2 - \sum_U \sigma_{qk}^2}{\left(\sum_U \sigma_{qk}\right)^2 - \sum_U \sigma_{qk}^2} - 1 \end{aligned} \tag{13}$$

where $ANV_{qC}(\hat{t}_{y_{qr}}) = \sum_U(\pi_{(C)k}^{-1} - 1)\sigma_{qk}^2$ and $ANV_{q\min}(\hat{t}_{y_{qr}})$ is given by (7).

Proof. Let π_k (where $\pi_k \leq 1$ *for* $k = 1, \ldots, N$) denote the first-order inclusion probabilities of a design $p(\cdot)$ where $E_p(n_s) = \sum_U \pi_k = n$. Minimizing (11) is equivalent to minimizing

$$\begin{aligned} &\sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_{qr}})}{ANV_{q\min}(\hat{t}_{y_{qr}})} \\ &= \sum_{q=1}^{Q} H_q \sum_U (\pi_k^{-1} - 1)\frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2} \end{aligned} \tag{14}$$

Simplifying by setting $w_{qk} = \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2}$ and changing the order of summation we get,

$$\sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_{qr}})}{ANV_{q\min}(\hat{t}_{y_{qr}})} = \sum_U \left(\frac{1}{\pi_k} - 1\right) \sum_{q=1}^{Q} H_q w_{qk} \tag{15}$$

Evaluating the right side of (15) and applying the Cauchy-Schwarz inequality gives

$$\left(\sum_U \pi_k\right) \sum_U \frac{\sum_{q=1}^{Q} H_q w_{qk}}{\pi_k} \geq \left(\sum_U \left(\sum_{q=1}^{Q} H_q w_{qk}\right)^{1/2}\right)^2$$

where equality holds if and only if $\pi_k = \pi_{(C)k} \propto \sqrt{\sum_{q=1}^{Q} H_q w_{qk}}$. Equation (13) is obtained by inserting $\pi_{(C)k}$ for π_k in the numerator and $\pi_{q(opt)k}$ for π_k in the denominator of (11) and evaluating, (using (7) and $\sum_{q=1}^{Q} H_q = 1$.) ■

Remark 1 *Proofs of results 3.1 and 3.2 are derived in a similar manner.*

A simple example illustrates how result 3.3 can be used in practice. For simplicity we now consider the case when all the parameters have equal importance weights, i.e. $H_q = 1/Q$, for $(q = 1, \ldots, Q)$.

Example 2 *In the planning stages, Result 3.3 can be used to select a design which gives us a strategy $\Omega_{p,\hat{\mathbf{t}}}$ that will give a low overall relative efficiency loss. Again, suppose we want to estimate $\mathbf{t} = \left(t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q}\right)'$, and that we have auxiliary information to formulate models ξ_q, $(q = 1, \ldots, Q)$ that are good descriptions of the (y_q, u_q) scatterplots.*

$$\begin{aligned} E_{\xi_q}(y_{qk}) &= \beta_{0q} + \beta_{1q} u_{qk} \qquad (k = 1, \ldots, N) \\ V_{\xi_q}(y_{qk}) &= \sigma^2_{qk} = \sigma^2_q u_{qk}^{\gamma_q} \qquad (k = 1, \ldots, N). \end{aligned}$$

Furthermore, with 'guesstimates', $\tilde{\gamma}_q$, of γ_q, perhaps taken from previous surveys or subject knowledge, we can for $q = 1, \ldots, Q$ and $k = 1, \ldots, N$ easily calculate $\dot{\pi}_{q(opt)k} = n u_{qk}^{\tilde{\gamma}_q/2} / \sum_U u_{qk}^{\tilde{\gamma}_q/2}$. Result 3.3 then implies that by using GREG estimators $\hat{\mathbf{t}}_r = \left(\hat{t}_{y_1 r}, \ldots, \hat{t}_{y_q r}, \ldots, \hat{t}_{y_Q r}\right)'$ as defined in (2) and choosing a design such that the first-order inclusion probabilities are determined by,

$$\pi_k = \dot{\pi}_{(C)k} = \frac{n \sqrt{\sum_{q=1}^{Q} \frac{u_{qk}^{\tilde{\gamma}_q}}{\sum_U \left(\frac{1}{\dot{\pi}_{q(opt)k}} - 1\right) u_{qk}^{\tilde{\gamma}_q}}}}{\sum_U \sqrt{\sum_{q=1}^{Q} \frac{u_{qk}^{\tilde{\gamma}_q}}{\sum_U \left(\frac{1}{\dot{\pi}_{q(opt)k}} - 1\right) u_{qk}^{\tilde{\gamma}_q}}}}$$

we expect to get a 'small' overall relative efficiency loss.

Approach C has an advantage over A and B. Since the measure that is minimized is based on ratios, the scaling effects of the involved variables are neutralized. This can be observed in the example above, where the constant factors σ^2_q cancel out.

The reasoning in this section can also be applied for studying the effect of different sample sizes. For example, calculating $ANV_{q\min}(\hat{t}_{y_q r})$ for $q = 1, \ldots Q$ with a given n, gives us the opportunity to study the sample size, n^*, that is needed for (13) to meet certain constraint criteria.

Example 3 *Suppose all parameters are equally important (i.e. $H_q = 1/Q$ for $q = 1, \ldots, Q$), and let n^* be the sample size needed so that, on average, $ANV_q(\hat{t}_{y_q r})$ does not exceed $ANV_{q\min}(\hat{t}_{y_q r})$ by more than $100 \times c$ %. Furthermore, suppose we begin with a starting value of n and calculate $ANV_{q\min}(\hat{t}_{y_q r})$ for every $q = 1, \ldots, Q$. If the constraint c is not too strict, we then can calculate the smallest value of n^* that satisfies the inequality,*

$$\begin{aligned} \frac{1}{Q} \sum_{q=1}^{Q} \frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} - 1 &\leq c \\ \sum_{q=1}^{Q} \frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} &\leq (1 + c)Q. \end{aligned} \tag{16}$$

Writing $a_{qk} = \sum_{q=1}^{Q} w_{qk} = \sum_{q=1}^{Q} \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2}$ *and* $\pi_{(C)k} = \frac{n^* \sqrt{a_{qk}}}{\sum_U \sqrt{a_{qk}}}$ $(k = 1, \ldots, N)$, *and using (15), the left hand side of (16) can be written* $\frac{1}{n^*} \left(\sum_U \sqrt{a_{qk}}\right)^2 - \sum_U a_{qk}$, *and after some algebra we get,*

$$n^* \geq \frac{\left(\sum_U \sqrt{a_{qk}}\right)^2}{(1+c)Q + \sum_U a_{qk}}.$$

Hence, it is possible determine the sample size that is needed to meet specifications made on $ANOREL_{\min}$ *in equation (13). (When further elaborating approach C (see section 6) such calculations might be helpful.)*

Equations (6), (8), (9) and (12) all suggest that the inclusion probabilities should be chosen proportional to some function of $V_{\xi_q}(y_{qk}) = \sigma_{qk}^2$. To apply this in practice, we need a good sampling scheme to implement the suggested designs, and we need to have a good idea of the values of σ_{qk}^2. In reality σ_{qk}^2 is unattainable, but it is often fruitful (as in Example 2) to use a model where σ_{qk}^2 is a function of an auxiliary variable u_q. Subject knowledge, guesses, or previous survey estimates can be used as planning values of σ_{qk}^2. The next section gives an overview of πps sampling designs, and to connect to the previous sections, we give recent references where πps sampling is combined with GREG estimation.

4 Implementing a πps design

The designs discussed in the previous section all suggest that unequal probability sampling should be used (π_k should be chosen proportional to some measure z.) Hence, we need a sampling scheme that implements $\pi ps(z)$ designs. The use of $\pi ps(z)$ designs has a long history in survey sampling, and it is one way of using auxiliary information in the design stage. However, much of the discussion in the literature focuses on strategies where a $\pi ps(z)$ design is combined with the π estimator. The reason for this is tradition. When a study variable y_q is strictly proportional to z, and $\pi_k = nz_k / \sum_U z_k$, we have $\hat{t}_{y_q\pi} = \frac{n_s}{n} t_{y_q}$. This means that with a random size $\pi ps(z)$ design, $\hat{t}_{y_q\pi}$ will vary due to the variation in sample size only, and for a fixed size design, $\hat{t}_{y_q\pi}$ has no variation at all.

Nevertheless, given a $\pi ps(z)$ design, there is no need to restrict ourselves to the π estimator. Since we have auxiliary information at hand we can use the GREG estimator. Furthermore, with GREG estimation our loss in efficiency by using a random size design, instead of a fixed size design, is likely to be small (if $E_p(n_s)$ is not too small.) Random size designs like the traditional Poisson πps sampling, or the recently proposed PoMix sampling, are described in Kröger, Särndal and Teikari (1999).

However, statisticians often prefer fixed size designs. They enable control over the sample size and the statistician avoids the task of explaining to clients that the initial sample size, and thereby the cost of the survey, is a random component. Over the years, these circumstances have led to an extensive effort to find a selection scheme for implementing a fixed size $\pi ps(z)$ sampling design. Although many sample selection schemes have been proposed, it has turned out to be hard to devise a fixed size scheme for an arbitrary sample size n that has a number

of desirable properties, such as (a) the actual selection of the sample is relatively simple, (b) all first-order inclusion probabilities are strictly proportional to the size variable, (c) the design admits (at least approximately) unbiased estimation of the design variances $V(\hat{t}_{y_q\pi})$ and $V(\hat{t}_{y_qr})$. If we also want to use the technique of permanent random numbers (*PRN*) in the sample selection, which is desirable in large survey organizations, it will be even harder.

Nevertheless, for statisticians preferring fixed size πps sampling, some sampling schemes fulfill most of the requirements above. Relatively new fixed size πps designs are *order sampling* designs, as Pareto πps and sequential Poisson πps (see Rosén (1997), Saavedra (1995) and Ohlsson (1995) respectively). Fixed size *PoMix* proposed by Kröger, Särndal and Teikari (2000) is another alternative, and comparisons have shown (see Holmberg (2001) and Holmberg and Swensson (2001)), that also *model-based stratified simple random sampling* (mb-STSI) proposed by Wright (1983) is a method that should be considered.

However, which sampling scheme to use is not the issue here, and depending on the situation there are pros and cons for all of them. Here, we merely state that alternatives that approximately fulfill the requirements above exist. Rosén (1997, 2000a, 2000b) and the references therein provide details on the Pareto πps, which is used in the Swedish Crop Yield Survey and Swedish Market Tendency Survey. PoMix sampling is described in the references by Kröger et al, and Holmberg, and mb-STSI can be studied in Wright as well as in Särndal et. al chapter 12.

5 A numerical comparison of the multiparameter approaches.

In this section, we will give an example on the use of the above multiparameter approaches, and how it is possible to use information collected at the planning stage for further elaboration of and for support in the choice of sampling design.

One of the purposes of this paper is to suggest approaches that might be useful to achieve a high overall efficiency. In a factual survey situation, the success in achieving this depends highly on the quality and structure of the auxiliary information and how the statistician uses the auxiliary information. The usefulness of the approaches in previous sections will therefore vary from case to case. However, we can mimic a real survey situation to give an idea on how they can be applied and how they might work in practice.

In the next section's example, we use a real and easy accessible finite population, (the population of Swedish municipalities MU281 available in appendix B in Särndal et al.) We place ourselves at the planning stage of a multiparameter survey from this population, where we are supplied with auxiliary variables, and where we have certain more or less valid beliefs and guesses about the relationships between our study variables and these auxiliary variables. The chosen relationships are not necessarily the best for this specific survey population, yet chosen to mimic a realistic starting point for a statistician planning a survey with auxiliary information, and to mimic a situation where the involved study variables are thought to have varying relations with the available auxiliary variables. (The latter is common in farm surveys, where some auxiliary variables that work well for farms specialized on crop might be poor for farms specialized on animals and vice versa.) Hence, the relationships we use

suggest a point estimator (a GREG estimator) for each parameter, and they give us ideas of alternative sampling designs where the auxiliary information can be utilized.

Altogether, our planning stage conditions give us a variety of alternatives for selecting a sampling design. The statistician subjectively determines many of these conditions, (i.e. through his beliefs about the relationships between the auxiliary variables and the study variables, and his choice of important parameters.) Still, given these conditions, we can compare the design alternatives, as is done below. In the end, the results of such comparisons might lead to a design decision that is good in meeting the overall demands of the survey.

5.1 Planning a multiparameter survey to achieve overall efficiency: An example

In our example we use all the quantitative variables in the MU281 population. As auxiliary variables we have $u_1 = P75$ (1975 population) and $u_2 = S82$ (total number of seats in the municipal council 1982) (and a constant $u_{3_k} = 1$ for every $k \in U$.) The six important study variables $y_1, \ldots, y_6$ are:

y_1	=	P85	(1985 population)
y_2	=	RMT85	(Revenues from the 1985 municipal taxation)
y_3	=	ME84	(Number of municipal employees 1984)
y_4	=	REV84	(Real estate values according to 1984 assessment)
y_5	=	CS82	(Number of conservative seats in municipal council 1982)
y_6	=	SS82	(Number of Social Democratic seats in municipal council 1982)

We plan to use GREG estimators $\hat{\mathbf{t}}_r = (\hat{t}_{y_1 r}, \ldots, \hat{t}_{y_6 r})'$ to estimate $\mathbf{t} = (t_{y_1}, \ldots, t_{y_6})'$, and the parameters are rated as equally important, (i.e. $H_q = 1/6$ for $q = 1, \ldots, 6$.) To assist in the planning of the sampling design, we have some a priori ideas of the relations between the study variables and the auxiliary variables. These are described in table 1, i.e. we believe it is reasonable to apply linear models ξ_q, $(q = 1, \ldots, 6)$, according to (1), where σ^2_{qk} are substituted with 'guesstimates', $\tilde{\sigma}^2_{qk}$. In this example, the $\tilde{\sigma}^2_{qk}$:s are different functions

Table 1: Planning stage assumptions for the relations between the auxiliary variables and the study variables.

q	1	2	3	4	5	6
y_q	$P85$	$RMT85$	$ME84$	$REV84$	$CS82$	$SS82$
$\mathbf{x}'_\mathbf{q}$	$(1, P75)$	$(1, P75, S82)$	$(1, P75)$	$(1, P75)$	$(1, P75, S82)$	$(1, S82)$
$\tilde{\sigma}^2_{qk}$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$S82^{\tilde{\gamma}_q}$	1
$\tilde{\gamma}_q$	1.4	2	2	0.4	1.2	0

of different auxiliary variables, and the constant factors, $\tilde{\sigma}^2_q$ (discussed in example 2), are assumed to be 1 for every q. (Knowledge of such factors is important when approach A and

approach B are applied, but for approach C and the single parameter approach of section 2, it is not necessary.)

Note that the purpose of the information given in table 1 is not to illustrate some true or even necessarily good model relations for this population. The purpose of the chosen model relations is to reflect what might be a realistic planning stage situation. By this, we mean a situation where the statistician, for each parameter individually, believes that the survey can benefit substantially from using the auxiliary information in the design as well as in the estimator. Furthermore, the six model relations in table 1 illustrate flexible differences in the planned way to use the auxiliary variables, especially with respect to $\tilde{\sigma}_{qk}^2$. For $q = 1, 2, 3, 4$, $\tilde{\sigma}_{qk}^2$ is a function of the auxiliary variable $P75$, for $q = 5$ it is a function of $S82$, while for $q = 6$ it is constant. Obviously, the mixture of different functions for $\tilde{\sigma}_{qk}^2$ will influence the properties of a compromise design. To further clarify, we translate some information in table 1 to more 'well known' cases. For example, if we insert the values of $\tilde{\sigma}_{qk}$ into equation (6) of section 2, the planning relations of table 1 imply: (i) When it comes to estimating t_{y3}, then $\tilde{\sigma}_{3k} = u_{1k} = P75_k$, and the planning strategy the statistician believes in is a $\pi ps(P75)$ sampling design combined with a GREG estimator using $\mathbf{x}_q' = (1, P75)$. (ii) To estimate t_{y6} he suggests a design with equal inclusion probabilities, (i.e. $\dot{\pi}_{6(opt)k} = n/N$ for every $k \in U$), combined with a GREG estimator where $\mathbf{x}_q' = (1, S82)$. A similar kind of reasoning can be used to understand the implications of other planning stage assumptions given in table 1.

We can use the setup from table 1 to create a diagnostic table for the alternative planning stage designs. By calculating the $\tilde{\sigma}_{qk}^2$ values and applying them to equations (6), (8) and (12), we can determine $\dot{\pi}_{q(opt)k}$, $\dot{\pi}_{(A)k}$ and $\dot{\pi}_{(C)k}$ for $k = 1, \ldots, 281$ and $q = 1, \ldots, 6$. Then, for the different design alternatives (the different sets of $\dot{\pi}$:s), planning values, $ANV_q^*(\hat{t}_{y_q r})$ and $ANV_{q\min}^*(\hat{t}_{y_q r})$ can be computed from equations (5) and (7). These values can then be studied and used to make a prediction of how the different design alternatives might affect univariate and overall precision of the survey. If the information collected from such a prediction also carries over to the implementation of the survey, it is valuable for the final design choice.

Table 2 illustrates a planning stage comparison between the designs considered for a MU281 survey, with $E_p(n_s) = 40$. From the information given in table 1, we have computed $ANV_{q\min}^*(\hat{t}_{y_q r})$ for the six designs considered from the single parameter approach, (here denoted p_i $i = 1, \ldots, 6$.) Then, predicted relative efficiency losses, i.e.

$$PREL_{p_i,q} = 100\left(\frac{ANV_q^*(\hat{t}_{y_q r})_{p_i}}{ANV_{q\min}^*(\hat{t}_{y_q r})} - 1\right),$$

have been computed for $p_1 - p_6$, as well as for the compromise designs, (p_7 and p_8), that follows from approaches A and C of section 3. For each design alternative, the predictions of the overall, (total), efficiency loss are summarized by the rowmeans, given in the last column. The rowmeans can be interpreted as planning stage predictions of $ANOREL$, i.e. $100 \cdot ANOREL_{p_i}^* = \sum_{q=1}^{6} PREL_{p_i,q}/6$.

Table 2: Planning stage relative efficiency losses, $100(\frac{ANV_q^*(\hat{t}_{yqr})_{p_i}}{ANV^*_{q\min}(\hat{t}_{yqr})} - 1)$, for eight alternative sampling designs, $(E_p(n_s) = 40)$, when estimating six population totals of MU281. (Boldface numbers show the largest efficiency loss for each design.)

Design approach	Parameters t_{y1}	t_{y2}	t_{y3}	t_{y4}	t_{y5}	t_{y6}	$ANOREL^*_{p_i}$ (%)
p_1: 'Optimal' for t_{y1}	0	8.6	8.6	20.1	25.8	**39.4**	17.1
p_2: 'Optimal' for t_{y2}	8.0	0	0	57.9	68.5	**93.4**	37.9
p_3: 'Optimal' for t_{y3}	8.0	0	0	57.9	68.5	**93.4**	37.9
p_4: 'Optimal' for t_{y4}	23.3	**72.2**	**72.2**	0	0.7	2.8	28.5
p_5: 'Optimal' for t_{y5}	29.2	**83.5**	**83.5**	0.7	0	1.8	33.1
p_6: 'Optimal' for t_{y6}	49.1	**125.7**	**125.7**	3.0	1.9	0	50.9
p_7: Approach A	3.6	1.1	1.1	38.9	45.7	**63.7**	25.6
p_8: Approach C	2.9	17.4	17.4	8.9	11.6	**19.6**	13.0
$ANV^*_{q\min}(\hat{t}_{yqr})$	119192	845420	845420	5428.4	170311	1693.0	

Not surprisingly, table 2 indicates that the smallest $ANOREL^*_{p_i}$, (13.0 %), is obtained for the design following as a result of applying approach C. The design p_1, with π_k 'optimally' chosen for estimating t_{y1} (i.e. with $\pi_k = \dot{\pi}_{1(opt)k} \propto z_k = u_{1k}^{0.7}$), has the second smallest (17.1 %). For design p_1, small (<10%) relative efficiency losses are predicted when t_{y1}, t_{y2}, t_{y3} are to be estimated, but large (>20%) for t_{y4}, t_{y5}, t_{y6}. The designs p_2 and p_3 (which by the way are identical), and the design following from approach A, also predict small losses for t_{y1}, t_{y2}, t_{y3} and large for estimating t_{y4}, t_{y5}, t_{y6}. For the designs p_4, p_5 and p_6, the pattern is reversed. From a multiparameter perspective none of the designs $p_1 - p_7$ seem to be satisfactory as compromise designs.

Remark 4 *Concerning approach B, our data does not permit a fair comparison with the other approaches, and we suspect that in most practical situations, the information needed at the planning stages is too demanding. However, sometimes a planning value for* $\mathbf{B}_q$, *say* $\dot{\mathbf{B}}_q$, *might be available, and then*

$$\dot{t}^2_{y_q} = \left(\sum\nolimits_U \mathbf{x}'_{qk}\dot{\mathbf{B}}_q\right)^2 \tag{17}$$

could be used as planning values for $t^2_{y_q}$.

5.2 Design comparisons on population data

The results in table 2 give rough guidelines of the properties of the considered designs. For any real finite population, the model assumptions made at the planning stage will deviate more or less from factual conditions. Therefore, actual calculation of estimator variances from our population, will give valuable information on what would have happened, if we had implemented the planning stage ideas of table 1. It will also give indications on to what extent the predicted design properties, as those of table 2, are transferable and valid for the actual sample survey.

For all our six parameters $t_{y_1} - t_{y_6}$, we consider GREG estimation, $\hat{t}_{y_1 r} - \hat{t}_{y_6 r}$ (see equation (2)), using $\mathbf{x}'_{qk}$ and $c_{qk} = \tilde{\sigma}^2_{qk}$ from table 1. A simple way to compare the alternative designs of table 2, is to calculate the estimator variances under Poisson sampling. For Poisson sampling, the Taylor expansion variance of equation (4) is $V_{T_{(PO)}}(\hat{t}_{y_q r}) = \sum_U \left(\pi_k^{-1} - 1\right) E_k^2$, and we calculated $V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ for $q = 1, \ldots 6$ and all considered designs p_i, $(i = 1, \ldots 8)$. Table 3 is based on the results from those variance calculations.

Table 3: Estimated relative efficiency losses, $100(\frac{V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}}{V_q^*(\hat{t}_{y_q r})} - 1)$, for eight alternative Poisson sampling designs, $(E_p(n_s) = 40)$, when estimating six population totals of MU281. (Boldface numbers show the largest efficiency loss for each design.)

	Parameters						
Design approach	t_{y_1}	t_{y_2}	t_{y_3}	t_{y_4}	t_{y_5}	t_{y_6}	$\widehat{\text{OREL}}$ (%)
p_1	0	6.9	6.7	21.2	18.7	**37.8**	15.2
p_2	8.9	0	0	59.2	54.3	**90.1**	35.4
p_3	8.9	0	0	59.2	54.3	**90.1**	35.4
p_4	19.6	**58.7**	56.3	0	0	2.7	22.9
p_5	26.8	**65.8**	62.3	1.8	2.1	1.1	26.6
p_6	42.4	**101.1**	96.2	2.3	2.9	0	40.8
p_7:Approach A	4.7	1.7	1.6	41.4	36.1	**60.7**	24.4
p_8:Approach C	3.8	17.9	17.9	10.1	9.6	**19.1**	13.1
$V_q^*(\hat{t}_{y_q r})$	6810	895719	6.36E07	1.75E09	20156	37271	

To simplify comparisons, table 3 has the same structure as table 2. Thus, we show results for all the eight alternative sampling designs considered at the planning stage. For each parameter we determine the smallest estimator variance V^* over all the considered designs, i.e. $V_q^*(\hat{t}_{y_q r}) = \min_{i=1}^{8} V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ is calculated for $q = 1, \ldots, 6$. Given the information we used at the planning stage, the values of $V_q^*(\hat{t}_{y_q r})$, will then represent the 'best' result we might obtain for every parameter separately. By dividing every $V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ with $V_q^*(\hat{t}_{y_q r})$, parameter by parameter, we then get a measure comparable to the predicted relative efficiency losses $PREL_{p,q}$ shown in table 2.

The general pattern of the relative efficiency losses in table 3 is the same as in table 2. Hence, given the selected GREG estimators, table 2 gives a good image of the relative efficiency losses for the considered designs under Poisson sampling. Parameter by parameter, the design based on approach C is never the best alternative, but in an overall sense, it is the most efficient, with (13.1%) mean efficiency loss. Therefore it can be argued that it also is the best compromise design, followed by design p_1, just as in table 2. One notable difference between the planning values of table 2 and the population values in table 3, is that design p_4 is slightly better than design p_5 for estimating t_{y5}. This is likely to happen in reality as well, the planning values are just guesses, more or less accurate, and $ANV_q^*(\hat{t}_{y_q r})$ and $V_{T_q}(\hat{t}_{y_q r})$ are different.

Remark 5 *Table 3 is based on Poisson sampling estimator variances. Similar tables with similar results can be computed for the fixed size designs proposed in section 4. However, since there will be more approximations involved, the results might be somewhat less reliable.*

Remark 6 *The designs p_2 (or p_3) and p_6 in tables 2-3 are in a sense traditional textbook designs. In p_6, all inclusion probabilities are equal as in simple random sampling or (as for the data in table 3) Bernoulli sampling. p_2 is traditional since the inclusion probabilities are proportional to the most useful auxiliary variable P75, i.e. ($\pi_k \propto u_{1k}$). An untransformed auxiliary variable is often used as size measure in πps designs. It should be noted from tables 2-3 that these latter designs are the worst in overall efficiency terms. Therefore, from a multiparameter perspective, they cannot be recommended in situations similar to the one described.*

If the effects of using a design such as p_8 not are entirely satisfactory, then a more elaborate compromise design is possible by applying approach D, given in the next section.

6 Approach D: Minimizing the weighted sum of relative efficiency losses under restrictions.

This section outlines yet another multiparameter alternative. Here, we choose to extend approach C, although similar reasoning can be applied for approaches A and B as well.

The statistician may consider this approach from start, but for two reasons we suggest that the implications from applying approach C (which gives a minimum ANOREL) are examined first. Firstly, realistic restrictions are more easily found after studying results from approach C, and secondly, approach C will provide useful numerical knowledge for the calculations in approach D.

If we once again study table 2, and if we regard approach C as the approach most suitable for our goals, we observe that, although approach C implies the smallest overall efficiency loss, these losses vary between the parameter estimates. For example in table 2 we have more than 15% efficiency losses for the parameters t_{y2}, t_{y3} and t_{y6}, while the others have smaller values.

We might want a design where the sum, ($ANOREL_{p_8}^*$), still is small, but where certain restrictions on the individual variances are fulfilled, so that the efficiency loss compared to

the 'optimal' variance does not exceed, say c_q. Hence, we can formulate the non-linear optimization problem below, where the specified restrictions now make the importance weights H_q redundant.

Problem 7 *Minimize the objective function*

$$f(\pi) = \sum_{q=1}^{Q} \sum\nolimits_U (\pi_k^{-1} - 1) \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2}$$

where $\sigma_{qk}^2 \geq 0$ *and* $\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2 > 0$, *under the* $2N + Q + 1$ *restrictions*

$$0 < \pi_k \leq 1 \qquad k = 1, \ldots, N$$

$$g_0(\pi) = \sum\nolimits_U \pi_k - n = 0$$

$$g_q(\pi) = \sum\nolimits_U (\pi_k^{-1} - 1) \frac{\sigma_{qk}^2}{ANV_q} \leq c_q \qquad q = 1, \ldots, Q$$

The number of restrictions and the mixing between linear and non-linear restrictions, as well as equality and inequality restrictions complicate the problem. Therefore, it is hard to find useful analytical solutions. However, the Karush-Kuhn-Tucker conditions apply, and if the Q $g_q(\pi)$-restrictions are not too strictly set, solutions can be obtained with non-linear programming algorithms. Hence, with some effort the flexibility in choosing a compromise design can be increased.

7 Conclusions

Planning a multipurpose survey with several important parameters is not a straightforward task. In this paper, we have presented some potentially useful approaches when auxiliary information is available.

By adapting a multiparameter perspective already at the planning stage, we illustrate that compared to a single parameter approach, significant improvements on the overall precision of a survey is possible. If we plan for the possibility of using the auxiliary information in the sampling stage as well as in the estimation stage, we can, by conditioning on an efficient estimator such as the GREG estimator and focusing on the design choice, construct diagnostic tables such as the one exemplified by table 2 of section 5. This can give us valuable information to compare the properties of the designs alternatives we consider, and help us to choose the compromise design that best fulfills our goals. Since the final survey plan depends on the overall objectives, we cannot give absolute recommendations on the planning of a multiparameter survey. However, approach C of this paper seems to be a useful approach. The multiparameter perspective used in that approach takes into consideration 'optimal' results for the single parameter case, and by minimizing a relative measure it seems as if the approach has an overall robustness. That is, the loss in efficiency compared to best

possible single parameter solutions will not be extremely high for any parameter estimate. Furthermore, if the solution from applying approach C not is satisfactory, a certain amount of flexibility in the design choice is possible. Under certain regularity conditions, non-linear programming can be used to construct a design that fulfills certain precision requirements.

A detailed discussion of the sampling schemes that can be used to implement a πps sample is beyond the scope of this text. However, there has been progress in that area in recent years, especially concerning fixed size sampling designs. Detailed information is provided in the references on πps sampling in section 4.

Finally, a crucial issue for applying the results in this paper is to have good planning values of σ_{qk}^2. In practice, statisticians often seem to use the approximation $\sigma_{qk}^2 \doteq u_{qk}^{\tilde{\gamma}_q}$, where $\tilde{\gamma}_q$ is an estimated or guessed value of a parameter γ_q (Harvey (1976) describes how ML-estimates of γ_q can be obtained, and in finite populations, γ_q often lies in the interval $(0, 2)$ or according to Brewer (1963) in the narrower interval $(1, 2)$.) Results from Rosén (2000a) and Holmberg & Swensson indicate, that the positive effects of having good $\tilde{\gamma}_q$ values (or $\tilde{\sigma}_{qk}^2$ values) for the design, can be substantial. In addition, from the example in the present paper we also note that unreflectively chosen values, e.g. choosing a design implicitly based on $\tilde{\gamma}_q = 2$, can have a large negative effect in a multiparameter perspective (see designs p_2 and p_3 of tables 2-3.) Hence, more attention (than we believe is the case today) should be paid on finding good planning values for σ_{qk}^2. Especially in surveys repeated over time, such attention could be a relatively cheap way to improve the survey quality.

8 References

Bellhouse, D. R., (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, **12**, pp 53-65.

Brewer, K. R. W., (1963). Ratio Estimation and Finite population: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93-105.

Cassady, R. J. and Valiant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology,* **19,** 183-192.

Cassel, C. M., Särndal, C-E. and Wretman, J., (1976). Some results on generalized difference estimators and generalized regression estimators for finite populations. *Biometrika* **63**, 615-620.

Cassel, C. M., Särndal, C-E. and Wretman, J., (1977). Foundations of Inference in Survey Sampling. Wiley & Sons, New York

DeVille, J.-C. and Särndal, C-E., (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Estevao, V. and Särndal, C.E., (2000). A Functional Form Approach to Calibration. *Journal of Official Statistics*, **16**, No. 4, 379–399.

Harvey, A.C., (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrika*, **44**, No. 3, 461-465

Holmberg, A., (2001). On the Choice of Strategy in Unequal Probability Sampling. (To appear in American Statistical Association 2001 Proceedings of the Section on Survey Research Methods.)

Holmberg, A. and Swensson, B., (2001). On Pareto πps sampling: Reflections on unequal probability sampling strategies. Theory of Stochastic Processes, **7(23),** No. 1-2 (2001) 142-155.

Kröger, H., Särndal, C-E. and Teikari, P., (1999). Poisson Mixture Sampling: A family of designs for Coordinated Selection Using Permanent Random Numbers, *Survey Methodology,* **25**, No 1, 3-11.

Kröger, H., Särndal, C-E. and Teikari, P., (2000). Poisson Mixture Sampling Combined with Order Sampling: a Novel use of the Permanent Random Number Technique. Manuscript submitted for publication (date 00/08/30).

Lundström, S. and Särndal, C-E., (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, **15**, 305–327.

Montanari, G.E., (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review,* **55,** 191-202.

Montanari, G.E., (1998). On Regression Estimation of Finite Population Means. *Survey Methodology,* **24,** No 1, 69-77.

Isaki, C. T. and Fuller, W. A., (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89 96.

Ohlsson, E., (1995). Sequential Poisson Sampling. Research Report from Institute of Actuarial Mathematics and Mathematical Statistics at Stockholm University.

Rao, J. N. K., (1979). Optimization in the design of sample surveys. In: J.S. Rustagi (ed.), *Optimization methods in Statistics: Proceedings of an International Conference.* New York, Academic Press, pp 419-434

Rao, J. N. K., (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. *Proc. Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden.

Rao, J. N. K., (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. *Journal of Official Statistics*, **10**, 153-165.

Rao, J. N. K., (1997). Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics*, **25**, 1-21.

Rosén, B., (1997). On sampling with Probability Proportional to Size, *Journal of Statistical Planning and Inference*, **62**, 159-191.

Rosén, B., (2000a). Generalized Regression Estimation and Pareto πps, R & D report 2000:5 Statistics Sweden.

Rosén, B., (2000b). A User's Guide to Pareto πps Sampling, R & D report 2000:6 Statistics Sweden.

Saavedra, P., (1995). Fixed Sample Size PPS Approximations with a Permanent Random Number. *Proceedings of the section on Survey research Methods Joint Statistical Meetings,* American Statistical Association, *697-700.*

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

Wright, R. L., (1983). Finite Population Sampling with Multivariate Auxiliary Information, *Journal of the American Statistical Association*, **78**, 879-884.

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och Metodenheten gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published during 1999 and onwards:

1999:1 Täckningsproblem i Registret över totalbefolkning RTB. Skattning av övertäckning med en indirekt metod *(Jan Qvist)*

1999:2 Bortfallsbarometer nr 14 *(Per Nilsson, Antti Ahtiainen, Mats Bergdahl, Tomas Garås, Jan Qvist och Charlotte Strömstedt)*

1999:3 Att mäta statistikens kvalitet *(Claes Andersson, Håkan L. Lindström och Thomas Polfeldt)*

2000:1 Kalibrering av vikter – beskrivning av tekniken och de SCB-fall den prövats i *(Sixten Lundström et al)*

2000:2 On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling *(Nibia Aires and Bengt Rosén)*

2000:3 Bortfallsbarometer nr 15 *(Per Nilsson, Ann-Louise Engstrand, Sara Tångdahl, Stefan Berg, Tomas Garås och Arne Holmqvist)*

2000:4 Bortfallsanalys av SCB-undersökningarna HINK och ULF *(Jan Qvist)*

2000:5 Generalized Regression Estimation and Pareto πps *(Bengt Rosén)*

2000:6 A User's Guide to Pareto πps Sampling *(Bengt Rosén)*

2001:1 Det statistiska registersystemet. Utvecklingsmöjligheter och förslag *(SCB, Registerprojektet)*

2001:2 Order πps Inclusion Probabilities Are Asymptotically correct *(Bengt Rosén)*

2002:1 On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys *(Anders Holmberg)*

CATARINA ELFFORS
MR/BY-S

ISSN 0283-8680

Tidigare utgivna **R&D Reports** kan beställas genom Katarina Klingberg, SCB, MET, Box 24 300, 104 51 STOCKHOLM (telefon 08-506 942 82, fax 08-506 945 99, e-post katarina.klingberg@scb.se).
R&D Reports from 1988-1998 can - in case they are still in stock - be ordered from Statistics Sweden, attn. Katarina Klingberg, MET, Box 24 300, SE-104 51 STOCKHOLM (telephone +46 8 506 942 82, fax +46 8 506 945 99, e-mail katarina.klingberg@scb.se).

SCB-Tryck, Örebro 2002.02