R&D Report 2002:3
Research – Methods – Development

Statistiska centralbyrån Statistics Sweden

On the Choice of Sampling Design in Business Surveys with Several Important Study Variables

by

Anders Holmberg

Patrik Flisberg

Mikael Rönnqvist

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 2002:3. On the choice of sampling design in business surveys with several important study variables / Anders Holmberg m.fl.
Digitalt skapad fil, anpassad efter de digitaliserade delarna i serien. Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-2002-X101OP0203

R&D Report **2002:3**
Research – Methods – Development

On the Choice of Sampling Design in Business Surveys with Several Important Study Variables

by

Anders Holmberg
Patrik Flisberg
Mikael Rönnqvist

R&D Report 2002:3

Research - Methods - Development

On the Choice of Sampling Design in Business Surveys with Several Important Study Variables

---

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

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

© 2002, Statistiska centralbyrån

ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 2002 11 MILJOMÄRKT Trycksak 341590

On the Choice of Sampling Design in Business Surveys with Several Important Study Variables

Anders Holmberg, Patrik Flisberg and Mikael Rönnqvist

ABSTRACT

The typical business survey has several study variables and several target parameters. To improve the precision of the estimators, it normally also involves several auxiliary variables that can be used in the design as well as in the estimators. Different auxiliary variables may have varying strength for different target parameters, and the design that is best for one parameter may not be best for another. With multiple target parameters and multiple requirements on precision, the practising statistician then must select a compromise design. In this paper, we present methods to obtain good compromise designs. Our approach yields unequal first order inclusion probabilities, which can be applied with both fixed size and random size sampling schemes and it offers a flexible use of auxiliary variables in the design. An example from a real Swedish business population is given.

Aim with this report

This report is a follow up of an earlier R&D report in this series (R&D report 2002:1.) Besides extended theory, it contains an application on Swedish Business data for an approach that was only outlined in the earlier work. This paper is a part of a joint work between personnel at the R&D department at Statistics Sweden and the Division of Optimization at Linköping Institute of Technology.

Acknowledgements

The authors acknowledge the work of Lennart Nordberg, Statistics Sweden and Bengt Swensson, Örebro University; Lennart Nordberg for his general advice and his assistance in providing register data for the Swedish manufacturing industry, and Bengt Swensson for his careful comments and suggestions that have led to improvements of the report.

Contents

On the Choice of Optimal Sampling Design in Business Surveys with Several Important Study Variables

Anders Holmberg*, Patrik Flisberg and Mikael Rönnqvist†

October 28, 2002

Abstract

The typical business survey has several study variables and several target parameters. To improve the precision of the estimators, it normally also involves several auxiliary variables that can be used in the design as well as in the estimators. Different auxiliary variables may have varying strength for different target parameters, and the design that is best for one parameter may not be best for another. With multiple target parameters and multiple requirements on precision, the practising statistician then must select a compromise design. In this paper, we present methods to obtain good compromise designs. Our approach yields unequal first order inclusion probabilities, which can be applied with both fixed size and random size sampling schemes and it offers a flexible use of auxiliary variables in the design. An example from a real Swedish business population is given.

MCS 2000 *subject classifications:* 62D05

Key Words and Phrases: Multiparameter surveys, Optimal sampling designs, Unequal Probability Sampling, Non-linear programming.

1 Introduction

A typical sample survey is concerned with estimating a large number of parameters of a finite population. 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) i.e., such that the final choice results in 'small' mean square error for each estimator $\hat{\theta}_i$. If, as frequently is the case in large survey organizations, there is auxiliary information available, the statistician can use this information to his advantage and thereby obtain a highly efficient strategy. The auxiliary information may be used in the choice of design as well as in the choice of estimator. However, whereas the auxiliary information for the estimators may be individually selected and different for different estimators, the auxiliary information used in the design is shared, and will affect all the parameter estimates. Before implementing a design that uses auxiliary information, the statistician must therefore closely examine its effects on all his or her key estimators. The design that is optimal for one target parameter may be far from optimal for another. Therefore, with a fixed amount of resources, it can sometimes be hard to find a design that meets the desired levels of precision for all important target parameters simultaneously.

*Statistics Sweden, Department of Research and Development, SE-701 89 Örebro, Sweden. e-mail: Anders.Holmberg@scb.se

†Linköping Institute of Technology, Division of Optimization SE-581 83 Linköping, Sweden

A well-known example of this problem appears for multi-item optimal allocation in stratified sampling. The solution is one in nonlinear programming (Bethel (1989), Cochran (1977) section 5A3-5A4). In this paper, we are concerned with a similar problem. However, instead of determining the best sample allocation between strata, we aim to determine the first order inclusion probabilities π_k $(k = 1, \ldots, N)$ so as to get a design which can qualify as an 'optimal' compromise in meeting desired levels of precision for multiple targets; i.e. we aim for a design where the estimator variances deviate as little as possible from the estimator variances resulting from single variable 'optimal' designs. Our approach is general in the sense that, when the π_k:s have been determined, they can directly be applied to both fixed size sampling schemes and random size sampling schemes. It covers regression estimators such as the family of (generalized regression) GREG estimators and the so called 'optimal regression estimator' (see Rao (1994) and Montanari (1998).) The approach also allows different auxiliary variables to be considered in combination with the different study variables in the design, which means a high degree of flexibility in the use of the auxiliary variables. Related work in this area having similar objectives have been made by Sigman and Monsour (1995), Saavedra (1999), Kott and Bailey (2000) and Holmberg (2002). Sigman and Monsour (1995) sketched a procedure using nonlinear programming that is similar to the one described here for Poisson πps sampling and the Horvitz-Thompson estimator. Saavedra (1999) applied these ideas using the algorithm by Chromy (1987) to determine probabilities to be used for Pareto πps sampling in a price and volume petroleum product survey. Kott and Bailey (2000) describe the theory and practice of a method they call *Maximal Brewer Selection,* which is used at the U.S. National Agricultural Statistics Service (NASS). Some of the theory used in this paper is also used in the paper by Kott and Bailey and in Holmberg (2002).

2 The problem and some theory

Assume that we are planning a survey of a finite population $U = \{1, \ldots, k, \ldots, N\}$, and assume that the key parameters we want to estimate are the population totals of the unknown study variables $y_1, \ldots, y_q, \ldots, y_Q$, i.e. $\mathbf{t} = \left(t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q}\right)'$, where $t_{y_q} = \sum_{k \in U} y_{qk} = \sum_U y_{qk}$. A without replacement sampling design $p(\cdot)$, with first-order inclusion probabilities π_k $(k = 1, \ldots, N)$ and second-order inclusion probabilities π_{kl} $(k, l = 1, \ldots, N)$, will be used to select a random set sample $s \subseteq U$ of size n_s, and each of the population totals are to be estimated with estimators $\hat{\mathbf{t}} = \left(\hat{t}_{y_1}, \ldots, \hat{t}_{y_q}, \ldots, \hat{t}_{y_Q}\right)'$. The design is to be determined to minimize a function f of all Q estimator variances and fulfill specified restrictions, v_q, made on functions g of each estimator variance (or variance approximation), i.e. minimize $f(g(V(\hat{t}_{y_1})), g(V(\hat{t}_{y_2})), \ldots, g(V(\hat{t}_{y_Q})))$ with restrictions on

$$g(V(\hat{t}_{y_q})) \leq v_q, \qquad (q = 1, \ldots, Q) \tag{1}$$

At the planning stage P auxiliary variables are available; 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. For $q = 1, \ldots, Q$, let $\mathbf{x}_q' = \left(x_{1_q}, \ldots, x_{j_q}, \ldots, x_{J_q}\right)$ be a suitable set of J_q (positive) auxiliary variables formed from $u_1, \ldots, u_p, \ldots, u_P$. Each population total can then be estimated by a 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, $\mathbf{t}_{x_q} = \left(t_{x_{1_q}}, \ldots, t_{x_{j_q}}, \ldots, t_{x_{J_q}}\right)'$ is 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. (Details of GREG estimation are given in Särndal, Swensson and Wretman (1992) sections 6.4-6.7.)

Suppose that the regression model ξ_q underlying the GREG estimator $\hat{t}_{y_q r}$ is

$$y_k = \mathbf{x}'_{qk}\boldsymbol{\beta}_q + \varepsilon_{qk} \tag{4}$$

with

$$\begin{cases} E_{\xi_q}(\varepsilon_{qk}) = 0 \\ V_{\xi_q}(\varepsilon_{qk}) = \sigma^2_{qk} \\ E_{\xi_q}(\varepsilon_{qk}\varepsilon_{ql}) = 0; \quad k \neq l \end{cases} \tag{5}$$

where $\boldsymbol{\beta}_q = \left(\beta_{1_q}, \ldots, \beta_{j_q}, \ldots, \beta_{J_q}\right)'$ are model parameters and the values $\sigma^2_{q1}, ..., \sigma^2_{qN}$ are considered known up to a constant multiplier. The anticipated variance (see Isaki and Fuller (1982))

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

is the variance of $\hat{t}_{y_q r} - t_{y_q}$ under the model and under the design. Here, an approximation to the anticipated variance denoted $ANV(\hat{t}_{y_q r})$ is given by

$$ANV_q(\hat{t}_{y_q r}) = \sum_U (\pi_k^{-1} - 1)\sigma^2_{qk} \tag{6}$$

For a single parameter, Result 12.2.1. in Särndal et al. states that for a sampling design such that $E_p(n_s) = n$, an 'optimal' design, i.e. a design that minimizes $ANV_q(\hat{t}_{y_q r})$, is such that the first-order inclusion probabilities are given by

$$\pi_k = \pi_{q(opt)k} = n\sigma_{qk} / \sum_U \sigma_{qk} \tag{7}$$

Hence, in the single parameter case we obtain an 'optimal' design by choosing $\pi_k \propto \sigma_{qk}$,i.e. a probability proportional -to-size (πps) design with σ_{qk} as a measure of size. It is sometimes assumed that a regression model as (4)-(5) is such that the heteroscedasticity is given by $\sigma^2_{qk} = \sigma^2_q u^{\gamma_q}_{qk}$ (where σ^2_q is a possibly unknown constant and $0 \leq \gamma_q \leq 2$.) If that is the case, then when $\gamma_q = 2$, an 'optimal' design is πps sampling design with $\pi_{q(opt)k} = nu_{qk} / \sum_U u_{qk}$.

Concerning the 'optimal' properties of πps designs with σ_{qk} as a measure of size for the single parameter case, others than Särndal et al. have shown similar results; some important references are Godambe (1955), Hájek (1959), Brewer (1963), Cassel, Särndal and Wretman (1976), Rao and Bellhouse (1978), Wright (1983) and Rosén (2000).

However, when $Q \geq 2$ and σ^2_{qk} differs (different u_q or the same u_q but different γ_q) single variable results are insufficient since they would suggest different designs for different study variables. Consequently, a compromise design has to be selected. The compromise should be a design that makes the 'best' use of the auxiliary information, considers all our important target parameters and gives results which do not differ too much from the individually optimal designs. Moreover, in contrast to the single parameter case, there is no single well-defined meaning of 'optimality' in the multiparameter case. Holmberg (2002) discussed three different approaches to achieve good compromise designs. They all have different minimization criteria, i.e. minimizing different overall

measures of variability, and they result in designs that in a sense are 'optimal', given the chosen criteria. Nevertheless, although those approaches yield designs with properties of overall 'optimality', that paper only outlines how 'optimal' solutions with individual restrictions such as (1), could be handled with nonlinear programming.

Restrictions or specified tolerance limits, v_q, for the variability of the most important survey estimates are commonly set. If we follow the situation outlined above, we can put restrictions on the individual $ANV_q(\hat{t}_{y_q r})$, e.g. $ANV_q(\hat{t}_{y_q r}) \leq v_q$, $(q = 1, \ldots, Q)$. However, this is seldom practical in multiparameter situations due to problems with scaling. Instead, we propose that restrictions are set with a dimensionless measure that relates to the individual minimum $ANV_q(\hat{t}_{y_q r})$, obtainable by inserting expression (7) in (6). Hence, for $q = 1, \ldots, Q$, we use

$$\frac{ANV_q(\hat{t}_{y_q r})_{p_i}}{ANV_{q\,\min}(\hat{t}_{y_q r})} \leq v_q \tag{8}$$

where $ANV_{q\,\min}(\hat{t}_{y_q r}) = \sum_U (\pi_{q(opt)k} - 1)\sigma_{qk}^2$ and $ANV_q(\hat{t}_{y_q r})_{p_i}$ is the anticipated variance of $\hat{t}_{y_q r}$ under a design $p_i(\cdot)$ with $\pi_k = \pi_{p_i k}$. (We use the index i to distinguish between various designs that are considered at the planning stage of a survey.) Restrictions can also be set on approximations to the anticipated coefficients of variation or anticipated relative variances (see Kott and Bailey (2000) and Holmberg (2002).

A good compromise design would be a design that fulfills the Q restrictions in equation (8) and minimizes some objective function that considers the overall precision of all $\hat{t}_{y_q r}$. The arithmetic mean (possibly weighted) of the relative ratios $ANV_q(\hat{t}_{y_q r})_{p_i}/ANV_{q\,\min}(\hat{t}_{y_q r})$, which we call the Anticipated Overall Relative Efficiency Loss (ANOREL) is one such objective function i.e.,

$$ANOREL = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})_{p_i}}{ANV_{q\,\min}(\hat{t}_{y_q r})} \tag{9}$$

where H_q $(\sum_{q=1}^{Q} H_q = 1)$ are importance weights attached to the Q ratios. ($H_q = 1/Q$ for every $q = 1, \ldots, Q$ when all ratios are considered equally important.)

Remark 1 *If no upper restrictions v_q are set on the individual ratios (8), the minimum of ANOREL for a given sample size is obtained if $\pi_k \propto \sqrt{\sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2}}$. (For a proof see Holmberg (2002).)*

Remark 2 *The geometric mean of $ANV_q(\hat{t}_{y_q r})_{p_i}/ANV_{q\,\min}(\hat{t}_{y_q r})$ is another possible objective function. Then, a logaritmic transformation would mean that focus would be both on anticipated standard errors and anticipated variances (Kott (2002).)*

In the following sections we will illustrate our method to determine the inclusion probabilities of a compromise design that minimizes functions such as ANOREL under restrictions. First we will give a description of the optimization models and thereafter we will give a numerical example using an authentic Swedish business population.

3 Optimization Models

The minimization of (9) under restrictions is a non-linear optimization problem where the $\pi_k : s$ are the variables that are to be determined. To avoid Greek symbols let us in this optimization part denote these by z_k $(k = 1, \ldots, N)$. The optimization problem that is to be solved can then be written as follows: Minimize the objective function

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

subject to the $2N + Q + 1$ constraints

$$\begin{aligned} 0 &< z_k \leq 1 \qquad k = 1, \ldots, N \\ g_0(\mathbf{z}) &= \sum\nolimits_U z_k - n = 0 \\ g_q(\mathbf{z}) &= \sum\nolimits_U (z_k^{-1} - 1) \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2} \leq v_q \qquad q = 1, \ldots, Q \end{aligned}$$

The number of restrictions as well as the mix between linear and nonlinear restrictions makes it hard to find useful analytical solutions. However, if the restrictions are not set too strictly, it is possible to find numerical solutions. A numerical solution is sufficient for our purpose to find inclusion probabilities for a good compromise design.

3.1 Convexity

An interesting question is if the problem above is a *convex problem.* Such a problem has the property that every locally optimal solution is also a global optimal solution. This is interesting as most solution methods for non-linear optimization problems are such that they find locally optimal solutions. There are a few methods which finds global solutions but they are in practice working only for much smaller problems than the ones we consider in this paper. Regarding convexity we refer to, for example, Fletcher (1991), Chapter 9. There are two properties that guarantee that a problem is a convex problem The first is that the feasible region is a convex set. The second is that the objective function is convex (in the case of minimization). For our problem it is easy to verify that the feasible region in fact is a convex set. Furthermore, the objective function is a convex function.

3.2 Model

We can rewrite the optimization model by using the following coefficients

$$\begin{aligned} a_k &= \sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2} \\ b_{qk} &= \frac{\sigma_{qk}^2}{\sum_U (\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2} \end{aligned}$$

The model can now be formulated as

$$[P1] \quad \min \quad f(\mathbf{z}) = \sum\nolimits_U a_k (z_k^{-1} - 1) \tag{10}$$

subject to

$$\begin{aligned} \textstyle\sum_U z_k &= n && (11)\\ \textstyle\sum_U b_{qk}(z_k^{-1}-1) &\leq v_q, \quad q=1,\ldots,Q && (12)\\ \epsilon \leq z_k &\leq 1, \quad k \in U && (13) \end{aligned}$$

The value of ϵ in constraint (13) is chosen arbitrarily small.

3.3 A practical model

For general values of the right hand side values of constraints (12), i.e. coefficients v_q, it is difficult to know if the problem has a feasible solution. For practical purposes we therefore introduce slack variables, s_q, in these constraints that are penalized with a factor δ. The new formulation will be

$$[P2] \quad \min \quad f(\mathbf{z}) = \textstyle\sum_U a_k(z_k^{-1}-1) + \delta \sum_{q=1}^{Q} s_q \qquad (14)$$

subject to

$$\begin{aligned} \textstyle\sum_U z_k &= n && (15)\\ \textstyle\sum_U b_{qk}(z_k^{-1}-1) - s_q &\leq v_q, \quad q=1,\ldots,Q && (16)\\ \epsilon \leq z_k &\leq 1, \quad k \in U && (17) \end{aligned}$$

The penalty parameter δ is chosen large as compared to the coefficients a_k. It is possible to show that the optimal solution (if one exists) to [P1] is equivalent to the optimal solution to [P2] if δ is chosen large enough, see e.g. Fiacco and McCormick (1990). The reason for introducing the penalized slack variables is as follows. If there exist a feasible solution to [P1] then there is no need of the slack variables; they will all receive a value of zero to avoid a possibly very large contribution to the objective function due to the large value of the penalty parameter δ. On the other hand, if no feasible solution exists to [P1] then the slack variables in formulation [P2] will be used. We may note that there is no intrinsic meaning of the contribution to the objective function from the slack variables as they are only used as a technical device to ensure feasible solutions. The only, but important, difference between [P2] and [P1] is that the former formulation always has a feasible solution. The information from a solution based on nonzero slack variables is very useful in practice. Firstly, there is an indication of which constraint that causes the unfeasability of [P1]. Secondly, it provides values that give information of when the (original) problem will become feasible.

3.4 Solution method

There are several methods available to solve the problem. We use the subroutine package NPSOL Version 4.0 (Gill, Murray, Sanders and Wright (1986)). The method is a Sequential Quadratic Programming (SQP) method. A description can be found in e.g. Fletcher (1991), Chapter 12. The basic structure of the SQP method is to generate a sequence of points $\{\mathbf{z}^{(l)}\}$ that converges to a local optimal solution. In each iteration l a new point $\mathbf{z}^{(l+1)}$ is generated as

$$\mathbf{z}^{(l+1)} = \mathbf{z}^{(l)} + r^{(l)}\mathbf{d}^{(l)}$$

where $\mathbf{d}^{(l)}$ is a search-direction obtained by solving a quadratic programming subproblem and r a step-length. The value of the step-length is computed by performing a linesearch in the direction

$\mathbf{d}^{(l)}$ of a merit function. There exist a range of merit functions, and the one used in NPSOL is an augmented Lagrangian function. Advantages with NPSOL is that it treats the lower and upper bounds on the variables implicitly and that linear constraints are efficiently utilized. The method is very robust and efficient. It is used in many commercial systems to solve hard nonlinear problems.

3.5 A comment on other objective functions

In section 2 we mentioned that other objective functions f than the one based on the ANOREL measure (equation (9)) can be of interest, e.g. approximations to anticipated *relative* variances or just anticipated variances (see Holmberg (2002).) For a weighted arithmetic mean we can use the models as described above with some minor adjustments. If the objective function is based on approximations to anticipated *relative* variances, i.e. minimizing $\sum_{q=1}^{Q} H_q ANV_q(\hat{t}_{y_q r})/t_q^2$, we use $a_k = \sum_{q=1}^{Q} H_q \sigma_{qk}^2/\dot{t}_q^2$, where $\dot{t}_q$ is a planning value (a guesstimate) of t_q. (If $\dot{t}_q = 1$ the objective function would be based on anticipated variances only.) Of course, the values v_q are adjusted to suit the chosen objective function but otherwise the constraints are kept the same.

If instead an objective function with a geometric mean is preferred, the problem is reformulated and the properties of the solution change.

3.5.1 Logarithm

If we use the weighted geometric mean of the ratios $ANV_q(\hat{t}_{y_q r})/ANV_{q\min}(\hat{t}_{y_q r})$

$$G = \prod_{q=1}^{Q} \left(\frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} \right)^{H_q}$$

as an objective function, i.e. minimizing $\ln G = \sum_{q=1}^{Q} H_q(\ln ANV_q(\hat{t}_{y_q r}) - \ln ANV_{q\min}(\hat{t}_{y_q r}))$, then the problem ((14)-(17)) is reformulated (using $a_{qk} = \sigma_{qk}^2 / \sum_{k\in U}(\pi_{q(opt)k}^{-1} - 1)\sigma_{qk}^2$) as

$$[P3] \quad \min \quad f(\mathbf{z}) = \sum_{q=1}^{Q} H_q \ln(\sum_U a_{qk}(z_k^{-1} - 1)) + \delta \sum_{q=1}^{Q} s_q \tag{18}$$

subject to

$$\sum_U z_k = n \tag{19}$$

$$\ln(\sum_U b_{qk}(z_k^{-1} - 1)) - s_q \leq v_q, \quad q = 1, \ldots, Q \tag{20}$$

$$\epsilon \leq z_k \leq 1, \quad k \in U \tag{21}$$

The values of v_q are, of course, chosen differently as compared to problem [P2]. In this model, it is not possible to show that the problem is convex. Therefore, we only can guarantee that a local optimal solution is found.

4 An application to a business population

4.1 Planning stage calculations

We illustrate with an application on a Swedish business population, (Manufacturers of food products and beverages, $N = 749$.) The sampling design in business surveys is often an unequal probability sampling design, and auxiliary variables are often available both for the design and estimation stages. In our case, we aim to estimate the (yearly) population totals of *Number of Employees* (t_{y_1}), *Turnover* (t_{y_2}), and *Personnel Expenses* (t_{y_3}). All three are considered equally important, we suppose our budget allows $E_p(n_s) = 112$ and we prefer a design p_i where no $ANV_q(\hat{t}_{y_q r})_{p_i}$ exceeds $ANV_{q\min}(\hat{t}_{y_q r})$ $(q = 1, 2, 3)$ by more than 6%. For every $k \in U$ we have the values from the previous year to use as auxiliary information. Previous experience and monitoring of this population have indicated that the pairwise correlations are high over a two year period. If we only had one target parameter, e.g. total *Number of Employees,* it would therefore be efficient to select a design using previous year's *Number of Employees* (u_1) as auxiliary variable. However, to estimate the total *Turnover,* it would be more efficient to use a design with previous year's *Turnover* (u_2) as auxiliary variable, and previous year's *Personnel Expenses* (u_3) would be best suited as auxiliary for t_{y_3}. Here, we consider the auxiliary vector $\mathbf{x}'_{qk} = (1, u_{1k}, u_{2k}, u_{3k})$ and $c_{qk} = 1$ for every $\hat{t}_{y_q r}$ $(q = 1, 2, 3)$. The planning values of σ^2_{qk} are $\tilde{\sigma}^2_{1k} = u_{1k}$, $\tilde{\sigma}^2_{2k} = u_{2k}$ and $\tilde{\sigma}^2_{3k} = u_{3k}$.

By applying equation (7) we have three alternative designs p_1, p_2, and p_3 (one for each study variable) where $\pi_k = \pi_{q(opt)k} \propto \tilde{\sigma}_{qk}$ $(q = 1, 2, 3)$. A fourth alternative, p_4, is a compromise design where $\pi_k \propto \sqrt{\sum_{q=1}^{Q} H_q \frac{\tilde{\sigma}^2_{qk}}{\sum_U (\pi^{-1}_{q(opt)k} - 1)\tilde{\sigma}^2_{qk}}}$, i.e. a design which minimizes (9) without restrictions on $ANV_q(\hat{t}_{y_q r})/ANV_{q\min}(\hat{t}_{y_q r})$ (see remark 1.) If p_4 is a preferred design we are through.

By calculations based on the planning values $\tilde{\sigma}^2_{qk}$, we get the cells of table 1 which contain the relative values $\tilde{R}_{p_i q} = \left[ANV_q(\hat{t}_{y_q r})_{p_i}/ANV_{q\min}(\hat{t}_{y_q r})\right] - 1$ for every combination of design and key parameter. The mean of $\tilde{R}_{p_i q}$ for each (planning stage) design $m(\tilde{R}_{p_i q})$ is the given in the right margin. (The values of $ANV_{q\min}(\hat{t}_{y_q r})$ were 77543, 89730 and 20262 for $(q = 1, 2, 3)$.)

Table 1: Planning stage relative efficiency losses, $100 \cdot \tilde{R}_{p_i,q}$, for four alternative sampling designs, $(E_p(n_s) = 112)$, when estimating three population totals of our business population. (Boldface numbers show the largest efficiency loss for each design.)

	Parameters			
Design approach	t_{y_1}	t_{y_2}	t_{y_3}	$m(\tilde{R}_{p_i,q})$ (%)
p_1: 'Optimal' for t_{y_1}	0	**24.3**	3.5	9.3
p_2: 'Optimal' for t_{y_2}	**24.5**	0	19.1	14.5
p_3: 'Optimal' for t_{y_3}	3.3	**16.4**	0	6.5
p_4: 'Optimal' for $m(\tilde{R}_{p_i,q})$	4.0	**7.2**	1.9	4.4

From table 1 we see that designs p_1 and p_3 have similar properties. With those we can expect good precision for t_{y_1} and t_{y_3} but poor for t_{y2}. With p_2 it is the other way around. The compromise design p_4 is the best as a whole with the lowest mean relative efficiency loss of $\tilde{R}_{p_4,q} = 4.4\%$.

However, design p_4 is not completely satisfactory, since the predicted efficiency loss of 7.2% for t_{y2} exceeds 6%.

Hence, we solve [P2] subject to $v_q = 6\%$ $(q = 1, 2, 3)$ (inserting $\tilde{\sigma}^2_{qk}$ for σ^2_{qk} and $n = 112$) and get a feasible solution of π_k values (a fifth alternative, p_5) such that when they are inserted into $ANV_q(\hat{t}_{y_q r})$, the ratios $ANV_q(\hat{t}_{y_q r})_{p_5}/ANV_{q\min}(\hat{t}_{y_q r})$ $(q = 1, 2, 3)$ (corresponding to the cells in table 1) become 4.9%, 6.0% and 2.4%. (It just takes one minute to solve P2 with an ordinary desktop with a Pentium 500MHz processor.) With solution design p_5, we expect an increased precision for estimating t_{y2} compared to design p_4, but we also sacrifice expected precison for the estimators of t_{y1} and t_{y3}. In this case, however, this sacrifice is small. The value of the mean efficiency loss (i.e. $\tilde{R}_{p_5,q}$) is slightly larger than $\tilde{R}_{p_4,q}$ but also rounded off to 4.4%.

4.2 Variance comparisons

In the preceeding section we compared alternative sampling designs using the auxiliary information only. In a real situation we would have used the knowledge from the planning stage calculations and applied a suitable sampling scheme (for example the Poisson πps or Pareto πps (see Rosén (1997)) to implement design p_5. However, it is not certain that p_5 is the best of our design when it comes to practice. The model assumptions made at the planning stage will deviate more or less from factual conditions, and therefore it is of interest to compare our alternative designs by calculating estimator variances.

Suppose we use the Poisson sampling scheme with π_k according to our designs $p_1, \ldots, p_5$ and suppose we use the GREG estimator, $\hat{t}_{y_q r}$, as mentioned earlier, with the auxiliary vector $\mathbf{x}'_{qk} = (1, u_{1k}, u_{2k}, u_{3k})$ and $c_{qk} = 1$ for every $q = 1, 2, 3$. Then the Taylor expansion variance of $\hat{t}_{y_q r}$ is

$$V_{T(PO)}(\hat{t}_{y_q r}) = \sum\nolimits_U \left(\pi_k^{-1} - 1\right) E^2_{qk}$$

where $E_{qk} = y_k - \mathbf{x}'_{qk}\mathbf{B}_q$ $(k = 1, \ldots, N)$ are population fit residuals, with $\mathbf{B}_q = \left(\sum_U \mathbf{x}_{qk}\mathbf{x}'_{qk}\right)^{-1} \sum_U y_{qk}\mathbf{x}'_{qk}$, a finite population regression coefficient.

For every parameter t_{y_q} and every design p_i, $V_{T(PO)q}(\hat{t}_{y_q r})_{p_i}$ $(i = 1, \ldots, 6)$ is calculated (p_6 is Bernoulli sampling with $\pi_k = n/N$, which is used here as a benchmark design.) One of our designs will give us the smallest estimator variance, i.e. $V'_q(\hat{t}_{y_q r}) = \min_{i=1}^{6} V_{T(PO)q}(\hat{t}_{y_q r})_{p_i}$ for every q. Given E_{qk} and given the various design alternatives at the planning stage, $V'_q(\hat{t}_{y_q r})$ represents the 'best' obtainable result for every parameter separately. Dividing $V_{T(PO)q}(\hat{t}_{y_q r})_{p_i}$ by $V'_q(\hat{t}_{y_q r})$ parameter by parameter we get a measure comparable to the one used in table 1 and we can study the effect of different choices of design.

Table 2 illustrates the losses in efficiency due to design choice, in terms of relative estimator variances. (The values of $V'_q(\hat{t}_{y_q r})$ are 237202, 472733 and 15922 for $q = 1, 2, 3$ respectively.)

The general pattern of table 2 is similar to that of p_1 to p_4 in table 1. The differences are that p_3 is the best choice for two of the three parameters while p_1 never is the best. Both are poor when the target parameter is total *Turnover* t_{y_2}. Despite that all auxiliary variables are used in the estimators, strategies using the Bernoulli design (p_6) are very poor, relatively speaking. This indicates that using auxiliary information in the design as well as the estimator pays off here. If we study the mean efficiency loss, the designs p_4 and p_5 are the best choices (6.8% and 6.4% mean efficiency loss respectively.) Hence, from a multiparameter perspective the choice of p_5 is good. Our aim of at most 6% loss for each parameter is not met (which is not to be fully expected) but it still is the best compromise design.

Table 2: Estimated relative efficiency losses, $100(\frac{V_{T_{(PO)q}}(\hat{t}_{yqr})_{p_i}}{V_q'(\hat{t}_{yqr})} - 1)$, for six alternative Poisson sampling designs, $(E_p(n_s) = 112)$, when estimating three population totals. (Boldface numbers show the largest efficiency loss for each design.)

	Parameters			
Design	t_{y1}	t_{y2}	t_{y3}	*Mean efficiency loss (%)*
p_1	11.2	**41.3**	3.2	18.6
p_2	**19.7**	0	12.4	10.7
p_3	0	**34.2**	0	11.4
p_4	5.0	**15.2**	0.3	6.8
p_5	5.6	**12.9**	0.6	6.4
p_6	171.3	212.4	**217.9**	200.5

The same conclusion is drawn from similar studies on other business populations in various branches of the Swedish manufacutring industry. For sake of space we choose not to present those studies here.

5 Summary

In large survey organizations that repeatedly do business surveys it is common to have access to strong auxiliary information. Often this information is used both in the design as well as in the estimators. However, when the auxiliary information is used in the design it is often done so in a simplified fashion, which does not take into consideration the fact that the different study variables have different relations to the auxiliary variables. The method we present here gives the survey statistician a more flexible and efficient way to use his or her available auxiliary information in the design. Our method involves non-linear programming at the planning stage of the survey to determine the inclusion probabilities that for several study variables minimizes various functions of anticipated estimator variances. The minimization is done subject to constraints on the functions of the individual estimator variances. The statistician can choose these constraints in order to find the design that simultaneously fulfills the precision requirements for several parameters of the survey.

We illustrated our method with an example from an authentic Swedish business population. The example with various design alternatives for this population and the estimator variance calculations later illustrate that, although we use strong auxiliary information in the estimators, it matters how we use the auxiliary information in the design. The numerical calculations require very little time even on an ordinary desktop. In the planning process, the statistician can easily try different solutions with different sets of constraints to find a satisfactory design. The tool we developed also computes design diagnostics such as table 1. This gives easy access to valuable information for the final design choice.

6 References

Bethel, J., 1989. Sample Allocation in Multivariate Surveys, *Survey Methodology*, **15**, 47-57.

Cochran, W.G. 1977. *Sampling Techniques.* Wiley, New York

Chromy, J. 1987. Design Optimization with Multiple Objectives, *Proceedings of the Section on Survey Research Methods, American Statistical Association 1987* 194-199.

Fiacco, A.V. and McCormick, G.P., 1990. *Nonlinear Programming - Sequential Unconstrained Minimization Techniques*, Classics in applied mathematics, SIAM.

Fletcher, R., 1991. *Practical Methods of Optimization*, 2:nd edition, John Wiley & Sons, Chichester.

Gill, P.E., Murray, W., Saunders, M.A. and Wright, M.H., 1986. *User's guide for NPSOL (Version 4.0): A Fortran Package for Nonlinear Programming,* Technical Report SOL 86-2, Systems Optimization Laboratory, Department of Operations Research, Stanford University, California 94305, January 1986.

Hájek, R., 1959. Optimum strategy and other problems in probability sampling, *Časopis Pěst. Mat.*, ***84***, 387-423.

Holmberg, A., 2002. A Multiparameter Perspective on the Choice of Sampling Design in Surveys,.Paper presented at the Baltic-Nordic Conference on Survey Sampling, August 17-23, Ammarnäs Sweden. (To appear in *Statistics in Transition*).

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

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

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

Kott, P.S., 2002. Personal Communication

Kott, P.S. and Bailey, J.T., 2000. The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling, *Proceedings of the second International Conference on Establishment Surveys,* June 17-21, 2000, Buffalo 269-279.

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. and Bellhouse, D.R., 1978. Optimal estimation of a finite mean under generalized random permutation models. *Journal of Statistical Planning and Inference,* **2,** 125-141.

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

Saavedra, P.J., 1999. Application of the Chromy Algorithm with Pareto Sampling, *Proceedings of the Section on Survey Research Methods, American Statistical Association 1999* 355-358.

Sigman, R.S., and Monsour, N.J., 1995. Selecting Samples from List Frames of Businesses, in Cox, B.G., Binder, D.A., Chinnappa, N., Christianson, A., Colledge, M.J., and Kott, P.S. (eds) *Business Survey Methods*, New York: Wiley, 153-169.

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

2002:2 Model-based calibration for survey estimation, with an example from expenditure analysis Surveys *(Claes Cassel, Peter Lundquist and Jan Selén)*

2002:3 On the Choice of Sampling Design in Business Surveys with Several Important Study Variables *(Anders Holmberg, Patrik Flisberg and Mikael Rönnqvist)*

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).