Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

On the Choice of Strategy in Unequal Probability Sampling

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

KEY WORDS **Probability proportional-to-size, Optimal strategy, GREG-estimation, Order Sampling, Model-based stratification.**

Abstract

There have been recent advances in the area of probability proportional to size sampling. Relatively new methods are PoMix sampling and Pareto sampling. Both are advocated as potential standard methods for conducting unequal probability sampling without replacement, with inclusion probabilities proportional to a known size measure. This paper is a comparison of these two designs and model based stratified sampling. From both an efficiency and practical point of view, we discuss various strategies (design and estimator combinations) that a survey statistician can choose from.

1. Introduction

In sampling from finite populations the topic of how to use auxiliary information is frequently discussed. Here our discussion is restricted to auxiliary information used in connection with sample selection and estimation. Roughly, we may divide the ways survey theory deals with auxiliary variables into two different categories: (i) Use of auxiliary variables at the *sampling design stage*, (e.g. for stratification or probability proportional-to-size sampling). (ii) Use of auxiliary variables at the *estimation stage*, for example through Generalized regression estimation (GREG), calibration and imputation methods. This paper treats the recent increased focus on probability proportional-to-size without replacement sampling, (πps sampling), and we aim to highlight some issues concerning the interaction of using auxiliary information at the sampling stage and at the estimation stage.

The question the survey statistician seeks to answer is: Which *strategy*, (combination of design and estimator), should be used to obtain as precise estimates as possible? If we ignore non-sampling errors, this question is reduced to finding the strategy that minimizes sampling errors, which in turn, (if unbiased point estimators are considered), means finding the strategy that yields the smallest point estimator variance. However, this search for efficiency and optimality must be balanced by considerations that make the chosen strategy simple to implement, e.g. simple sample selection schemes and computing algorithms.

Constructing a πps sampling scheme with desirable properties is one way of using the strength of an auxiliary variable to find an efficient sampling design. Relatively new πps designs are *PoMix* designs proposed by Kröger, Särndal and Teikari (1999, 2000), and *order sampling* designs like Pareto πps and sequential Poisson πps (see Rosén (1997), Saavedra (1995) and Ohlsson (1995) respectively). From a strategy perspective a comparison between the results given in these papers is needed. Furthermore, in situations when the survey statistician considers using these schemes, he/she should also pay attention to other alternative strategies. One such alternative that we choose to consider here is *model-based stratified simple random sampling* (mb-STSI) proposed by Wright (1983).

1.1 Background and notation

The setup is a situation where direct element sampling from a finite population, $U = \{1, \ldots, k, \ldots, N\}$, is possible. Our objective is to estimate a population total, $t_y = \sum_{k \in U} y_k$, of a study variable y. We have access to Q auxiliary variables, whose values u_{qk}, $(q = 1, \ldots, Q)$, are known for every element k in the population, and we want to make the best possible use of this information.

For every element in the population we have an auxiliary vector $\mathbf{u}_k = (u_{1k}, \ldots, u_{Qk})^{'}$, and with a suitable function $h(\cdot)$, we can create a strictly positive size variable $z_k = h(\mathbf{u}_k)$, $(k = 1, \ldots, N)$, and use it in a πps sampling design denoted $\pi ps(z)$. Hence, we use z to compute first order inclusion probabilities $\pi_k = nz_k/t_z$, (henceforth $\pi_k \leq 1$ is assumed for every $k \in U$), and select a (set) sample $s \subseteq U$ of size n_s, where $E(n_s) = n$.

To use the auxiliary variables in the estimation stage we can use GREG estimation. The Q auxiliary variables are then used to form another set of variables, $x_1, \ldots, x_j, \ldots, x_J$, i.e. for $k = 1, \ldots, N$, we have a known vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})^{'}$.

The GREG estimator for t_y, can be defined as

$$\hat{t}_{greg} = \hat{t}_\pi + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^{'}\hat{\mathbf{B}} \quad (1)$$

where $\hat{t}_\pi = \sum_{k \in s} y_k/\pi_k$ is the well-known π or Horvitz-Thompson estimator, $\mathbf{t}_x = \sum_U \mathbf{x}_k$, $\hat{\mathbf{t}}_{x\pi} = (\hat{t}_{x_1\pi}, \ldots, \hat{t}_{x_J\pi})^{'}$ and

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k^{'}}{c_k \pi_k} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k y_k}{c_k \pi_k} \quad (2)$$

where c_k is suitably chosen constant.

1.2 A brief review of theoretical results on "optimal" choice of strategy

Although a πps sampling design combined with the π estimator, or a non-πps sampling design combined with GREG estimation can be a good strategy, a natural alternative when we seek an efficient use of the auxiliary information, would be to combine πps sampling with GREG estimation. Theoretical results that support this exist, e.g. see (a) Result 12.2.1 in Särndal, Swensson and Wretman (1992) and (b) Theorem 1 or Theorem 4.1 respectively in Cassel et. al. (1976, 1977), (for a modification of the latter see also Theorem 2.1 in Rosén (2000).) It is interesting to note that, although these two results are derived from different angles of approach, the practical advice they carry is similar. Both results address the strategy issue in terms of an underlying superpopulation model, ξ, which we can write as

$$y_k = \mathbf{x}_k^{'}\boldsymbol{\beta} + \varepsilon_k$$

with

$$\begin{cases} E_\xi(\varepsilon_k) = 0 \\ V_\xi(\varepsilon_k) = \sigma_k^2 \\ E_\xi(\varepsilon_k \varepsilon_l) = 0; \qquad k \neq l \end{cases}$$

where, (for $k = 1, \ldots, N$) σ_k^2 are, (often through a function of $\mathbf{u}_k$), known constants.

Both result (a) and result (b) above advocate a strategy with GREG estimation combined with a $\pi ps(\sigma)$ design, i.e. $\pi_k \propto \sigma_k$, $\pi_k = n\sigma_k / \sum_U \sigma_k$. Rosén (2000) seems to suggest using the GREG estimator (1) and (2) with $c_k = \sigma_k^2$,while Särndal et. al. (1992) do not impose the restriction $c_k = \sigma_k^2$.

Although these results should be considered when planning a survey, their potential benefit are based on our belief about the structure of the population for one single study variable! Therefore, they are of limited use, unless we study or have knowledge of the likely side effects on efficiency that a certain choice has for all our important study variables. The choice of size measure z, is especially delicate. It is only made once and this choice affects the efficiency of all the estimators. For example, if $\sigma_k^2 = z_k$, ($k = 1, \ldots, N$), for one important variable, then our 'optimal' strategy would be a $\pi ps(\sqrt{z})$ design combined with $\hat{t}_{greg}$ with $c_k = z_k$, but for other variables where $\sigma_k^2 \neq z_k$, selecting the $\pi ps(\sqrt{z})$ design would not be 'optimal'. Although we can use different c_k-weighting in $\hat{t}_{greg}$ for those variables, using the $\pi ps(\sqrt{z})$ in the sampling stage might severely jeopardize the efficiency of the estimates. A way to study the effect of choosing a $\pi ps(\sigma)$ design with non-optimal σ_k^2 is presented in Holmberg and Swensson (2001).

Besides the fact that for some variables there will be unavoidable losses in efficiency, due to 'non-optimal' inclusion probabilities, other deviations between an assumed model and reality also affect how well a chosen strategy works. The practising survey statistician must also carefully choose the auxiliary vector $\mathbf{x}_k$ and at the same time keep estimation as simple as possible, (the latter primarily to make production run smoothly). Based on a common design, we ideally seek one simple and robust strategy for every study variable, where only moderately large losses in efficiency are made if model assumptions are poorly fulfilled. In the following we present parts of an ongoing study to compare aspects of different strategy choices.

2. πps sampling designs

Many sample selection schemes which implement πps sampling designs have been proposed over the years. However, if we exclude random size designs, it has turned out to be hard to devise a scheme for arbitrary sample size n that has a number of desirable properties, e.g. (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 $V_p(\hat{t}_\pi)$ and $V_p(\hat{t}_{greg})$. 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.

We will give a very brief account of, and a few comments on, *Pareto πps sampling,* and *Poisson mixture (PoMix) sampling.* These designs will be compared to Wright's *model-based stratified simple random sampling* as it is outlined in chapter 12 of Särndal et al (1992). All these designs have simple algorithms for sample selection, there are suggested solutions for variance estimation and they may be alternatives for the practitioner when the use of the PRN technique is desirable. With regard to desirable property (b) above, this is only approximately fulfilled. Since mb-STSI is well described in the references, this section focuses on the link between Pareto πps and fixed size PoMix based on a Pareto πps sampling scheme. Thus, although applicable to a wide range of πps sampling schemes, here we will present PoMix merely as a transformation of the target inclusion probabilities of Pareto πps.

2.1 Pareto πps

The Pareto πps design has the following sample selection scheme: (i) For every element $k \in U$, compute target inclusion probabilities $\lambda_k = nz_k/t_z$. (ii) Generate N independent standard uniform random variables $\delta_1, \delta_2, \ldots, \delta_N$ and compute ranking variables $Q_k = \delta_k(1-\lambda_k)/(1-\delta_k)\lambda_k \quad (k = 1, \ldots, N)$. (iii) The elements with the n smallest Q_k then constitute the sample s.

To estimate the population total t_y, Rosén considers $\hat{t}_\lambda = \sum_s \frac{y_k}{\lambda_k} = \sum_s \frac{y_k}{\pi_k} \cdot \frac{\pi_k}{\lambda_k}$ which is very close to the π estimator, if λ_k are very close to π_k for $(k = 1, \ldots, N)$. The variance of $\hat{t}_\lambda$ is, unless n and N are very small, and/or the pattern of λ_k is unfavorable, well approximated by

$$AV\left(\hat{t}_\lambda\right) = \frac{N}{N-1}\sum\nolimits_U \lambda_k(1-\lambda_k) \times \left(\frac{y_k}{\lambda_k} - \frac{\sum_U y_k(1-\lambda_k)}{\sum_U \lambda_k(1-\lambda_k)}\right)^2 \quad (3)$$

Pareto πps has several advantages: (i) The sampling scheme is very simple. (ii) It enables control of the sample size. (iii) For order sampling designs with fixed distribution shape, Rosén (1997) shows that Pareto πps is optimal, in the sense of having the smallest asymptotic variance of $\hat{t}_\lambda$. (iv) A variance estimator is suggested by Rosén

$$\hat{V}(\hat{t}_\lambda) = \frac{n}{n-1}\sum\nolimits_s (1-\lambda_k) \times \left(\frac{y_k}{\lambda_k} - \frac{\sum_s y_k(1-\lambda_k)/\lambda_k}{\sum_s(1-\lambda_k)}\right)^2 \quad (4)$$

If N and n are reasonably large, the λ_k are close to the true π_k, and the bias of $\hat{t}_\lambda$, the approximation error of (3) and the bias of the variance estimator (4) are negligible. Aires and Rosén (2000) provide guidelines for when we can expect the bias of $\hat{t}_\lambda$ to be small for different size and shape patterns of the size measures. Hence, assuming that the λ_k are close approximations of π_k, the Pareto πps has several advantages. Still, only limited studies exist for a strategy combining Pareto sampling with the GREG estimator, (see Rosén (2000) and Holmberg (2000)). We therefore need ideas for GREG estimation under the Pareto πps.

2.2 GREG Estimation under Pareto πps

Since the exact inclusion probabilities for Pareto πps normally are unknown, we cannot derive the GREG estimator exactly. However we can form a related 'quasi GREG-estimator' by replacing the π_k by λ_k, i.e.

$$\hat{t}_{qgreg} = \hat{t}_\lambda + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\lambda})'\hat{\mathbf{B}}_\lambda = \sum\nolimits_s \lambda_k^{-1} g_k y_k$$

with $g_k = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\lambda})'\hat{\mathbf{T}}^{-1}\mathbf{x}_k/c_k$ where $\hat{\mathbf{T}}^{-1} = \left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k'}{c_k\lambda_k}\right)^{-1}$. (If λ_k is a close approximation to π_k for every $k \in U$, $\hat{t}_{qgreg}$ will be close to $\hat{t}_{greg}$.)

We also require that there is a simple and adequate way of estimating the variance $V_p(\hat{t}_{qgreg})$. Combining Result 6.6.1 in Särndal et al. and equation (4) suggests the variance estimator

$$\hat{V}(\hat{t}_{qgreg}) = \frac{n}{n-1}\sum\nolimits_s (1-\lambda_k) \times \left(\frac{g_k e_k}{\lambda_k} - \frac{\sum_s g_k e_k(1-\lambda_k)/\lambda_k}{\sum_s(1-\lambda_k)}\right)^2 \quad (5)$$

where $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}_\lambda$. It should be noted, that the above suggestions are based on several approximations. The effect of these approximations are not fully explored. However, the studies shown here, in Rosén (2000) and Holmberg (2000), indicate that this approach is practically feasible.

2.3 PoMix sampling

At the sampling stage the surveyor has full disposal over the size measures z. For a possible gain in precision of the point estimators the size measures and thereby the inclusion probabilities may be transformed. When considering the Pareto πps design, there are two reasons why it is wise to take a closer look at the size measures. Firstly, as pointed out by Aires and Rosén, if the shape pattern of the size measures does not meet certain criteria, we can get a bias which is unacceptably large. Secondly, by modifying the size measures we can in certain situations also get a more efficient design. We give a tentative explanation for this second statement linked to the distribution of z_k (or λ_k). The finite populations for which πps sampling is suggested are often skewed to the right with respect to z, (e.g. business populations where z is some measure of company size.). It is well known that elements with very small inclusion probabilities can cause problems. We risk abnormal estimates and a high sample to sample variability for our estimators, since λ_k^{-1} of these small elements, are large, this will especially be the case if there are many such elements in the population and if the design is such, that the probability is large that at least some of them will be included in the sample. A method which still keeps the advantages of the πps design, but where we automatically avoid the risk of elements with abnormal weights is desirable. Although the overall effect on precision is uncertain, time can be saved, because elements with this property often have to be singled out and treated in some special way as extremes. Slightly smoothing the size measures before selecting a Pareto πps sample can be a way to achieve this. Such smoothing is offered by the Pomix approach.

A fixed size PoMix transformation based on the Pareto πps sampling scheme is attained in the following way: For a given n, and for every $k \in U$, make the linear transformation $\lambda_k^{'} = \omega + \lambda_k(1-\omega/f)$ of the target inclusion probabilities in the Pareto πps scheme of section 2.1. Here ω is a parameter in the interval $(0 \leq \omega \leq f)$, (called Bernoulli width in Kröger et al.) and $f = n/N$. As before, λ_k $(0 < \lambda_k < 1)$ is the original Pareto πps target inclusion probabilities.

It is easy to see that when $\omega = 0$ we have the 'original' Pareto πps design, and that the design reduces to simple random sampling when $\omega = f$. Consequently, as ω moves from 0 to f, the variance and the range of $\lambda_k^{'}$, will decrease. In fact, since λ_k is the original target inclusion probabilities of the Pareto πps, it follows from choosing $\omega > 0$ and applying the proposed transformation that:

$$\begin{aligned} \lambda_k^{'} &\geq \lambda_k \, for \, \{k \in U : \lambda_k \leq \textstyle\sum_U \lambda_k/N = f\} \\ \lambda_k^{'} &< \lambda_k \, for \, \{k \in U : \lambda_k > \textstyle\sum_U \lambda_k/N = f\} \end{aligned}$$

Thus, the elements in Pareto πps with a target inclusion probability less than or equal to the sampling fraction, will have an increased modified target inclusion probability, and the elements with a target inclusion probability larger than the sampling fraction will have a decreased modified target inclusion probability.

If the finite population is skewed to the right with respect to λ_k, the proportion of elements with an increased target inclusion probability due to transformation will be larger than the number of elements with a decreased ditto. If the shape pattern of the size measures is such that we suspect the bias of $\hat{t}_\lambda$ (and $\hat{t}_{qgreg}$) to be non-negligible, (see Aires and Rosén), this modification can automatically, if not adjust, so at least reduce this bias by changing the shape of the target inclusion probabilities. Furthermore by increasing the smallest λ_k, the risk of abnormal estimates due to extreme weights become less and the estimator variance could decrease.

A perspective which deserves attention, not regarded in Kröger et. al, concerns the properties of Pomix sampling when $\lambda_k \propto \sigma_k$, (i.e. according to the theoretical results for 'optimal' strategies mentioned in section (1.2).)

At the estimation stage we can for this fixed size PoMix design, apply the same principles for GREG estimation as described in section 2.2. In PoMix sampling you have to decide what value of ω to choose. The survey statistician who considers PoMix has to make this choice on grounds of rough guidelines and practical experience. In the comparisons of the next section we will use the value $\omega = 0.3f$ tentatively suggested by Kröger et. al.

3. A simple comparison of strategies

A small numerical example illustrates some issues when a strategy involving a πps design is to be chosen. We have a finite population, Holmberg and Swensson (2001), of $N = 1000$ elements, two study variables y_1 and y_2, and one auxiliary variable highly correlated with the y-variables. The latter is used both as size variable z, and as regressor x in the GREG estimator. The relationship is such that it reasonable to believe, that for y_1 as well as y_2 the finite population scatterplot is well described by a model $E_\xi(y_k) = \beta x_k$, $V_\xi(\varepsilon_k) = V_\xi(y_k) = \sigma_k^2 = cx_k^\gamma = cz_k^\gamma$, and $E_\xi(\varepsilon_k \varepsilon_l) = 0$, where γ is a constant reflecting the degree of heteroscedasticity.

In πps sampling, statisticians often use (due to tradition, belief, or analytical neglect of previous surveys) designs with inclusion probabilities proportional to $\sqrt{z}$ or z. But, if we strive for a $\pi ps(\sigma)$ design, recommended by the results mentioned in section (1.2), we then implicitly assume $\gamma = 1$ and $\gamma = 2$. For the designs described and the design mb-STSI, we will compare this practice, with the results we get when exploiting better knowledge of γ. As estimators we use $\hat{t}_\lambda$, variants of $\hat{t}_{qgreg}$ in the GREG family for the Pareto πps based schemes and their counterparts $\hat{t}_\pi$ and $\hat{t}_{greg}$ for mb-STSI. For $\hat{t}_{qgreg}$ and $\hat{t}_{greg}$ the following constants c_k are used: $(\hat{t}_1 : c_k = x_k^\gamma)$, $(\hat{t}_2 : c_k = \lambda_k/(1-\lambda_k))$, $(\hat{t}_3 : c_k = x_k)$ and $(\hat{t}_4 : c_k = 1)$. If we knew the true value of γ, $\hat{t}_1$ would, in combination with a $\pi ps(z_k^{\gamma/2})$ design, be an efficient strategy choice according the theoretical results; $\hat{t}_2$ is used by Kröger et al. (1999); $\hat{t}_3$ simplifies to the well known ratio estimator with $g_k = t_x/\hat{t}_x$ and $\hat{t}_4$ is a simple regression estimator. In our population the ML estimate of γ according to the method suggested by Harvey (1976) is $\hat{\gamma}_1 = 1.09 \approx 1$ for y_1, and $\hat{\gamma}_2 = 1.45 \approx 1.5$ for y_2. Depending on which variable y_1 or y_2 that is considered to be the most important, reasonable approaches for the survey statistician who seeks an efficient strategy, would then be to apply a $\pi ps(z_k^{1/2})$ or a $\pi ps(z_k^{3/4})$ design. As benchmark designs we consider $\pi ps(z_k^1)$ (assuming $\gamma = 2$) and simple random sampling, (SRS). This gives in total twelve different designs that, together with the studied estimators add up to 60 different strategies. In other words, for each of three different sampling schemes, (original Pareto πps $(\omega = 0)$, Pareto πps with PoMix transformation ($\omega = 0.3f$), and mb-STSI with 10 strata), we study four different transformations of the auxiliary variable as a size variable, five estimators for the Pareto πps based sampling schemes each, and five for mb-STSI.).

Table 1 shows the estimated estimator variances, $S_p^2(\hat{t}_i)$ of the strategies, (where p denotes the designs under consideration and i the estimator), relative to the estimator variance for SRS combined with the π-estimator, $S_{SRS}^2(\hat{t}_\pi)$. The estimates, $S_p^2(\hat{t}_i)$, are based on 10000 independent samples of size $n = 100$ for each design, and the table cells contain the ratio $R = S_{SRS}^2(\hat{t}_\pi)/S_p^2(\hat{t})$. $S_{SRS}^2(\hat{t}_\pi)$ is $5.51E6$ for y_1 and $3.87E8$ for y_2 and the higher the value of R the higher is the strategy effect compared to SRS and the π-estimator. The ratio of the cell values can be used to compare two different strategies. For example by comparing Pareto $\pi ps(z_k^{1/2})$ and $\hat{t}_1$ with mb-STSI $\pi ps(z_k^{1/2})$ and $\hat{t}_1$, we see that the estimator variance of the Pareto $\pi ps(z_k^{1/2})$ strategy is 4.7%, $(100 * 163.6/156.2)$ higher for y_1 and 19.9%, $(100 * 61.4/51.2)$, higher for y_2.

Not surprisingly we note, that the GREG estimators are superior to the π or (λ) estimator for all designs, and for both y_1 and y_2, (Pareto $\pi ps(z_k)$ where the point estimator expressions are the same is the exception.) The effect of using the auxiliary variable in the design can be studied by comparing the strategies based on $\pi ps(1)$ with the other designs using $\pi ps(z_k^{\gamma/2})$. Hence the table reveals that a strategy which uses the auxiliary information in both the design as well as in the estimation is preferable. The survey statistician is then left with 32 possible strategy choices, $\hat{t}_1 - \hat{t}_4$ and one of the $\pi ps(z_k^{\gamma/2})$ alternatives. The choice of GREG estimator seems to be of minor importance so we concentrate our discussion on the choice of design. This choice depends on which of the variables y_1 and y_2 that is most important and the knowledge of the γ values. Presuming that y_1 is the key variable of a survey and that information saying $\gamma_1 \doteq 1$ exists, then according to the theoretical results we should choose one of the $\pi ps(z_k^{1/2})$ designs. When comparing the Pareto, PoMix and mb-STSI $\pi ps(z_k^{1/2})$ we see that mb-STSI is more efficient than Pareto $\pi ps(z_k^{1/2})$ (approximately 4% smaller estimator variance for all the GREG estimators). Pareto $\pi ps(z_k^{1/2})$ is in turn more efficient than PoMix $\pi ps(z_k^{1/2})$. On the other hand, if y_2 is the key variable and with the guess that $\gamma_2 \doteq 1.5$, we should according to the theoretical results choose a $\pi ps(z_k^{3/4})$ design. Then again mb-STSI would be the design resulting in the smallest GREG estimator variance. Over the studied GREG estimators, the estimator variances for the Pareto and PoMix designs are similar and roughly 18%, $(100 * 65/55)$, larger than $S_{mbSTSI}^2(\hat{t}_i)$.

However, if we regard estimates of both variables equally important, the choice of sampling design is more delicate. In this population it turns out that if we sum our measure R, over both variables y_1 and y_2 in each cell, the strategy which gives the most efficient estimates is the mb-STSI $\pi ps(z_k^{3/4})$ design combined with $\hat{t}_4$, (boldfaced in the table). This strategy is not the most efficient for estimating t_{y_1} but the relative loss is very small.

From the table we also note that by using a 'standard' $\pi ps(z_k)$ design, we would loose in efficiency, either for estimates of t_{y_1} or for estimates of t_{y2}. The results for PoMix however, seem to be similar to those of PoMix $\pi ps(z_k^{3/4})$. This latter observation was the only notable difference between the Monte Carlo results shown here, and the results of a Monte Carlo study with a sample size of $n = 50$, in which case the PoMix $\pi ps(z_k)$ showed higher estimator variances than PoMix $\pi ps(z_k^{3/4})$.

4. Summary and conclusions

This study highlights that we have practically working methods (i.e. simple to implement, allow PRN, and have nice variance estimation properties) to combine GREG estimation and πps designs, and that considerable gains in efficiency can be made by doing so. The efficiency gain is supported by a Monte Carlo study and a review of two theoretical results. Recent developments in πps sampling propose the use of the Pareto πps or fixed-size PoMix designs. However, from an efficiency point of view, the mb-STSI proposed by Wright seems to work at least as well. Also adding that Pareto πps can be sensitive to high skewness and extremely small values of π_k, and that PoMix is sensitive for the choice of the Bernoulli width parameter, ω, the relative attractiveness of mb-STSI increases.

We also want to stress that, in practical work, more attention should be payed to σ_k and thereby the use of $\pi ps(\sigma_k)$ design rather than just using a plain untransformed size measure z_k to determine inclusion probabilities. In the light of the theoretical results, we also note that a $\pi ps(\sigma_k)$ design makes the PoMix design (or PoMix modification of Pareto πps) unnecessary from an efficiency perspective. However, since there are usually several study variables involved and only one σ_k can be used in the design, and since there usually are deviations from a presumed model, PoMix might be useful. As our Monte Carlo studies show, it might be more efficient

Table 1: Strategy effects measured as the ratio $S^2_{SRS}(\hat{t}_\pi)/S^2_p(\hat{t}_i)$.

Designs		Estimators				
		$\hat{t}_{\lambda y_1}, \hat{t}_{\lambda y_2}$	$\hat{t}_{1y_1}, \hat{t}_{1y_2}$	$\hat{t}_{2y_1}, \hat{t}_{2y_2}$	$\hat{t}_{3y_1}, \hat{t}_{3y_2}$	$\hat{t}_{4y_1}, \hat{t}_{4y_2}$
$\pi ps(1)$	Pareto	1, 1	110.7, 37.6	110.7, 37.6	111.1, 31.7	110.7, 37.6
	PoMix	1, 1	110.7, 37.6	110.7, 37.6	111.1, 31.7	110.7, 37.6
	mb-STSI	6.3, 5.7	109.9, 40.6	109.9, 40.6	109.5, 39.3	109.9, 40.6
$\pi ps(z_k^{1/2})$	Pareto	5.7, 4.8	156.2, 51.2	156.7, 54.0	156.2, 51.2	156.9, 56.6
	PoMix	3.3, 3.0	153.4, 48.3	153.9, 52.4	153.4, 48.3	153.9, 54.4
	mb-STSI	24.8, 19.5	163.6, 61.4	163.6, 61.9	163.6, 61.4	163.6, 62.4
$\pi ps(z_k^{3/4})$	Pareto	23.8, 14.1	142.0, 50.7	146.2, 53.5	145.9, 53.1	146.6, 55.8
	PoMix	6.7, 5.7	162.3, 49.4	165.1, 55.7	164.7, 53.4	165.5, 58.5
	mb-STSI	33.3, 23.9	163.4, 65.2	164.4, 65.6	164.3, 65.6	**164.5, 65.7**
$\pi ps(z_k)$	Pareto	106.9, 47.5	106.9, 47.5	106.9, 47.5	106.9, 47.5	106.9, 47.5
	PoMix	14.7, 10.5	140.8, 49.9	163.8, 56.4	163.6, 55.8	164.2, 59.2
	mb-STSI	25.5, 23.9	137.2, 60.4	152.0, 63.0	151.7, 63.0	151.5, 63.0

than Pareto πps and it can automatically take care of some unfavorable shape patterns that might lead to bias in Pareto πps. In conclusion, more work is needed to study which GREG plus πps strategy that works best in an overall sense. This paper suggests that mb-STSI is a serious competitor to the newer πps designs. Work to find out which strategy is least sensitive to deviations from model assumptions is currently in progress.

References

Aires, N and Rosén, B. (2000). On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling. In: *Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs.* Doctoral thesis Chalmers University, Göteborg Sweden.

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

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).

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

Holmberg, A. (2000) Developments in unequal probability sampling, a comparison of designs. Baltic-Nordic Workshop on Survey Sampling Theory and Methodology June 18-22, 2000 in Pärnu, Estonia.

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

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

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

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