Survey Methodology, December 1996
Vol. 22, No. 2, pp. 199-204
Statistics Canada

Optimal Sample Redesign Under GREG in Skewed Populations With Application

GURUPDESH S. PANDHER [1]

ABSTRACT

Within a survey re-engineering context, the combined methodology developed in the paper addresses the problem of finding the minimal sample size for the generalized regression estimator in skewed survey populations (*e.g.*, business, institutional, agriculture populations). Three components necessary in identifying an efficient sample redesign strategy involve i) constructing an efficient partitioning between the "take-all" and "sampled" groups, ii) identifying an efficient sample selection scheme, and iii) finding the minimal sample size required to meet the desired precision constraint(s). A scheme named the "Transfer Algorithm" is devised to address the first issue (Pandher 1995) and is integrated with the other two components to arrive at a combined iterative procedure that converges to a globally minimal sample size and population partitioning under the imposed precision constraint. An equivalence result is obtained allowing the solution to the proposed algorithm to be alternatively determined in terms of simple quantities computable directly from the population auxiliary data. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario are reported. A 52% reduction in the total sample size is achieved for the regression estimator of the total at a minimum coefficient of variation of 2%.

KEY WORDS: Minimal sample size; Optimal sample selection; Precision constraint; Sampled group; Take-all group.

1. INTRODUCTION

In many survey situations additional information is available on all population units before the survey is undertaken. This auxiliary information is frequently useful in devising a more efficient sample design and estimation strategy. In a survey redesign context, the most optimal strategy holds the promise of offering the largest reduction in survey costs by requiring the lowest sample size necessary to meet the desired precision constraint on the estimates. In repeat surveys of skewed populations, an efficient sample design and estimation strategy may be realized by exploiting a) the correlation structure between the size-based auxiliary information x (*e.g.*, population of municipality, employees in a firm, farm acreage) and the survey variables y (*e.g.*, municipality expenditures, value of shipments, farm yield) and b) the variance relationship between the survey variable and the auxiliary size information.

In this paper, a comprehensive sample redesign methodology is developed for skewed populations with the ultimate objective of bringing about maximal reductions in the current sample size while ensuring a desired level of precision for the generalized regression estimator of the total. This work was motivated by the redesign of the Local Government Finance Survey (LGFS) conducted by Statistics Canada's Public Institutions Division. Financial information (*e.g.*, revenues, expenditures, debt, *etc.*) obtained from local government units is used in the estimation and publication of financial statistics on a provincial and national basis. Although the work presented in this paper is motivated by a concrete application, the sample design methodology devised applies generally to all surveys based on skewed populations (*e.g.*, agricultural, business, and institutional surveys).

In identifying an efficient new sample design, the overall methodology addresses and integrates the solution to three problems:

1) Creation of the "Take-all" and "Sampled Groups"

Since the variability of the survey response y_k tends to increase with the size of the unit x_k, it is common in skewed populations to sample the largest x-valued units with certainty in order to improve the efficiency of the population estimators. The demarcation of the population into the non-overlapping "take-all" $U_a = \{1, ..., N_a\}$ and "sampled" groups $U_b = \{1, ..., N_b\}$ is obtained through a new scheme named the "Transfer Algorithm".

2) Choosing an Efficient Sample Selection Scheme

Let $p(s; \lambda) = (p_a(s_a), p_b(s_b; \lambda))$ represent the complete sample design where the sample design parameter λ determines the type of sample selection implemented in the sampled group U_b. The sample inclusion probabilities due to $p_b(s_b; \lambda)$ may be expressed as $\pi_k(\lambda) = n_b(x_k^{\lambda/2} / \sum_{U_b} x_j^{\lambda/2}), k \in U_b$. Note that the parameter λ defines a broad class of sample designs with SRS $(\lambda = 0)$ and pps $(\lambda = 2)$ as particular cases. Design optimality results (Godambe and Joshi 1965) allow the identification of the most optimal value for the sample design parameter λ.

[1] Gurupdesh S. Pandher, Survey Analysis and Methods Development Section, Household Survey Methods Division, Methodology Branch, Statistics Canada, 16th Floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

3) Minimal Sample Size Determination

The third component of the overall methodology is aimed at finding the minimal sample size required to meet the imposed precision constraints for the estimator.

The combined procedure devised integrates these components to allow a new globally minimal sample size and optimal population partitioning to be determined under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Firstly, the "Transfer Algorithm" is proposed which finds an optimal population allocation between the take-all and sampled population groups in the sense of minimizing the variance of the generalized regression estimator (GREG) of the total. Desirable mathematical properties of this algorithm such as existence and optimality of solution along with an equivalence result were established in Pandher (1995). The equivalence result allows the solution to be determined in terms of simple quantities computable directly from the population auxiliary data.

The Transfer Algorithm in then synthesized iteratively with the sample size determination step to find the minimal sample size needed to satisfy the imposed precision constraints through an iterative procedure. The combined methodology produces a sequence of sample sizes and population partitionings which converge to a globally optimal solution where further reductions in the sample size are not possible given the imposed precision constraint. An application of the procedure is given for Ontario using provincial data from the Local Government Finance Survey.

Lavallée and Hidiroglou (1988), Hidiroglou and Srinath (1993) (subsequently denoted as L&H and H&S, respectively), and Glasser (1962) have proposed alternative methodologies for constructing the take-all and sampled groups within the context of stratified SRS design. The proposed approach differs from other methods in three respects. Firstly, the population demarcation is obtained under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Secondly, the criterion for constructing the population demarcation is based on minimizing the variance of the GREG estimator of the total under the desired sample selection strategy (Glasser and L&H base their allocation on minimizing the within-stratum sum-of-squares *x*; H&S use the total regression sum-of-squares under a regression model with a compulsory intercept assuming SRS). Thirdly, the proposed methodology explicitly captures the size-induced heteroscedasticity present in skewed survey populations which has been ignored in other frameworks.

Lastly, it is useful to qualify the sense in which the term "optimal" is used. Since, the redesign uses auxiliary information from a previous cycle of the survey to estimate the design parameters, there is a level of sub-optimality introduced in the redesign methodology by this lag. But as a practical matter, using the data from the most recent survey is the best that can be done. Once the design parameters have been estimated or are known however, the cut-offs and sample sizes required to achieve the desired precision yield the lowest anticipated design variance given that the estimates are true (or close to it). It is therefore, in this sense that the word "optimal" is used.

2. SURVEY FRAMEWORK

The model assisted survey framework is adopted for the skewed population whose auxiliary and survey characteristics are denoted by $C_U = \{(x_1, y_1), ..., (x_N, y_N)\}$. In this framework, underlying the class of generalized regression estimators for the population total are regression models (Särndal 1992, p. 255) exploiting the correlation between the survey variables y and the auxiliary covariates x. Different model assumptions on the deterministic and stochastic components of the underlying model lead to different regression estimators for the population total. For example, a ratio-form heteroscedastic model

$$y_k = \beta x_k + \epsilon_k, \tag{2.1}$$

with the error $\epsilon_k \sim (0, \sigma_k^2)$ and the variance structure given by $\sigma_k^2 = c\, x_k^{\gamma}$ (γ is the heteroscedasticity parameter) leads to the following GREG estimator:

$$\hat{t}_{Rb} = \sum_{U_b} x_k \hat{B} + \sum_{s_b} \frac{(y_k - x_k \hat{B})}{\pi_k} \tag{2.2}$$

where $\hat{B} = (\sum_s y_k/\pi_k)/(\sum_s x_k/\pi_k)$ is the sample-based probability weighted estimate of the population regression parameter B.

Given this estimation framework, the total across both groups $t = t_a + t_b$ is estimated by $\hat{t} = t_a + \hat{t}_{Rb}$ where $\hat{t}_a = t_a = \sum_{U_a} y_k$ since all units are sampled in the take-all group and $\hat{t}_{Rb}$ is the GREG estimator under the relevant model. The anticipated variance of $\hat{t}_{Rb}$ (defined as the variance with respect to both the design and the model, denoted p and ξ, respectively) is expressible as

$$V(\hat{t}_{Rb}) \equiv \mathrm{E}_{\xi} V_p(\hat{t}_{Rb}) \doteq \sum_{k \in U_b} \left(\frac{1}{\pi_k} - 1 \right) \sigma_k^2. \tag{2.3}$$

Furthermore, if σ_k^2 depends on the auxiliary measure x_k according to the formulation $\sigma_k^2 = c x_k^{\gamma}$ (2.4), then design optimality (Godambe and Joshi 1965) implies that the optimal sample inclusion probabilities are $\pi_k^*(\gamma) \propto x_k^{\gamma/2}$, $k \in U_b$. Therefore, the sample design $p_b^*(s_b; \lambda = \gamma)$ in the sampled sub-population, defining the first order inclusion probabilities $\pi_k^*(\gamma) = n(x_k^{\gamma/2} / \sum_U x_j^{\gamma/2}), k \in U_b$, minimizes the anticipated variance $V(\hat{t}_{Rb})$.

In the model assisted framework used in this paper, the auxiliary measure x_k is assumed to be a scalar. As noted by a referee, the more general case where x_k is a vector could be handled by fitting the appropriate parametric relationship $\sigma_k^2 = f(x_{k1}, ..., x_{kq})$ and using the estimated $\hat{\sigma}_k$ in lieu of x_k in defining the inclusion probabilities. The approach for the multivariate x_k seems intuitively sound and is mentioned here for completeness but requires further study and investigation.

Three methods for estimating the heteroscedasticity parameter γ from past survey data called the "Least Squares Method", the "Maximum Likelihood Method", and the "Graphical Method" are described in Appendix A of Pandher (1995).

3. TRANSFER ALGORITHM

In this section, an iterative scheme named the "Transfer Algorithm" is proposed to determine the optimal demarcation between the take-all and sampled sub-populations under the sample design $p(s;\lambda)$. The criterion for this construction is based on finding a population partitioning minimizing the estimated anticipated variance of $\hat{t}_{Rb}$. An equivalence result from Pandher (1995) is used to find an alternative and simpler method of solution based entirely on quantities defined on the auxiliary population data.

The proposed scheme for constructing the take-all and sampled sub-populations, U_a and U_b, respectively, is based on the following idea. Initially, place all population units in the sampled group, labelling it $U_b^{(0)}$ (the superscript l represents the iteration cycle). Hence, the take-all group is an empty set $U_a^{(0)}=\{\varnothing\}$. The resulting population and sample size allocation at $l=0$ is given by $N_a^{(0)}=0$, $n_a^{(0)}=0$, $N_b^{(0)}=N$, and $n_b^{(0)}=n_0$ where n_0 is the current sample size.

In a repeat survey setting, the variances σ_k^2 in (2.3) can be empirically modelled using the relation $\sigma_k^2 = c\,x_k^{\gamma}$ (2.4) where γ and c are estimated from previous sample data as mentioned before. Using the estimated version of (2.4) in (2.3) yields the following estimator for $V^{(l)}(\hat{t}_{Rb};\cdot)$:

$$\hat{V}^{(l)}(\hat{t}_R;\lambda,N_b^{(l)},n_b^{(l)}) = \sum_{k\in U_b^{(l)}} \left(\frac{1}{\pi_k(\lambda)} - 1\right) \hat{c}x_k^{\hat{\gamma}} \qquad (3.1)$$

where the largest l x-valued units have been removed from $U_b^{(0)}$. Note that λ is used here to parameterize the sample design to allow greater generality when $\lambda \neq \gamma$.

In the iterative algorithm, we start initially with all population units placed in $U_b^{(0)}$. Then for each iteration l, $0 \le l < n$, the largest $l+1$ x-valued unit $x_{(N-l-1)}$ is transferred from $U_b^{(l)}$ to $U_a^{(l)}$ and the difference

$$\Delta(l) = \hat{V}^{(l+1)}(\hat{t}_{Rb};\lambda,N-l-1,n-l-1) - \hat{V}^{(l)}(\hat{t}_{Rb};\lambda,N-l,n-l) \qquad (3.2)$$

is computed. Negative values of $\Delta(l)$ mean that the transfer of the unit corresponding to the ordered value $x_{(N-l-1)}$ lead to a decrease in the variance. Moreover, such transfers continue to result in a reduction in the variance of $\hat{t}_{Rb}$ as long as $\Delta(l)<0$. In general, for any iteration l, the relationship between the population and sample size allocations is described by the following relations: $N_b^{(l)}=N-l$, $n_b^{(l)}=n-l$, and $N_a^{(l)}=n_a^{(l)}=l$. These relations hold because the overall population and sample sizes must remain constant ($N=N_a^{(l)}+N_b^{(l)}$ and $n=n_a^{(l)}+n_b^{(l)}$) for all iterations.

The solution is also constrained by the condition $\pi_k(\lambda)<1,\, k\in U_b^*(l^*)$. Let $l^*(\lambda)$, $0 \le l^* < n$, represent the solution to the Transfer Algorithm. Given the discussion above, the solution to the Transfer Algorithm under the sample design $p(s;\lambda)$ may be formulated as

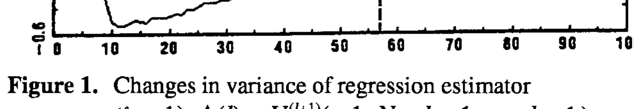
$$l^*(\lambda) = \min_{l}\{l\colon [\pi_{(N-l)}(\lambda)<1] \quad \text{and}$$
$$\hat{\Delta}(l) = [\hat{V}^{(l+1)}(\hat{t}_{Rb};\lambda) - \hat{V}^{(l)}(\hat{t}_{Rb};\lambda)] \ge 0, 0 \le l < n\}. \quad (3.3)$$

The optimal population allocation to the take-all group $U_a^*(l^*)$ is then given by the population units coinciding with the l^* ordered units transferred to the take-all auxiliary vector $X_a^* = (x_{(N-l^*)}, x_{(N-l^*+1)}, \ldots, x_{(N)})$; correspondingly the sampled group $U_b^*(l^*)$ consists of the units corresponding to $X_b^* = (x_{(1)}, x_{(2)}, \ldots, x_{(N-l^*-1)})$.

Transferring a unit from $U_b^{(l)}$ to $U_a^{(l)}$ causes two opposite effects on the variance $V^{(l)}(\hat{t}_{Rb};\cdot)$. The reduction in the population size ($N_b^{(l+1)} = N_b^{(l)} - 1$) has the impact of decreasing the variance, while the equivalent reduction in the sample size ($n_b^{(l+1)} = n_b^{(l)} - 1$) has the reverse effect of increasing $V^{(l)}(\hat{t}_{Rb};\cdot)$. Somewhere in this process, a critical value l^*, $0 \le l^* < n$, exists which gives the optimal breakdown $\{U_a^*(l^*), U_b^*(l^*)\}$. Moreover, in Theorem 3 of Pandher (1995), it is shown that as long as the conditions $(x_{(N-l)}^{\lambda/2} - x_{(N-l-1)}^{\lambda/2}) \ge 0$ and $(x_{(N-l)}^{\gamma-\lambda/2} - x_{(N-l-1)}^{\gamma-\lambda/2}) \ge 0$, $0 \le l < n$, hold, a solution to the Transfer Algorithm exists and that the system remains stable (optimal) upon reaching l^*. Stability further implies that the solution is optimal since the conditions leading to the solution do not change in the range $l^* \le l < n$. These two properties may be more precisely defined as follows:

Existence: $\exists\, l^*, 0 \le l^* < n$, such that $V^{(l^*+1)} - V^{(l^*)} \ge 0$
and $\pi_{(N-l^*)}^{(l^*)} < 1$.
Stability: If $V^{(l^*+1)} - V^{(l^*)} \ge 0$, then $V^{(l+1)} - V^{(l)} \ge 0$
and $\pi_{(N-l)}^{(l)} < 1$ for $0 \le l^* < l < n$.

An example of the application of the Transfer Algorithm to the LGF survey population of local municipalities in Ontario (with $N=793$, $n=108$, $\gamma=2$, and $\lambda=1$) is given in Figure 1. The curves are plotted for $l > 8$ because in the interval $0 < l \le 8$, the first condition of (3.3), namely $[\pi_{(N-l)}(\lambda)<1]$, is not satisfied. The minimum value of $\hat{V}^{(l)}(\hat{t}_{Rb})$ is achieved at $l^* = 57$ where $\Delta(l^*) = \hat{V}^{(\hat{l}^*+1)} - \hat{V}^{(\hat{l}^*)} \ge 0$.



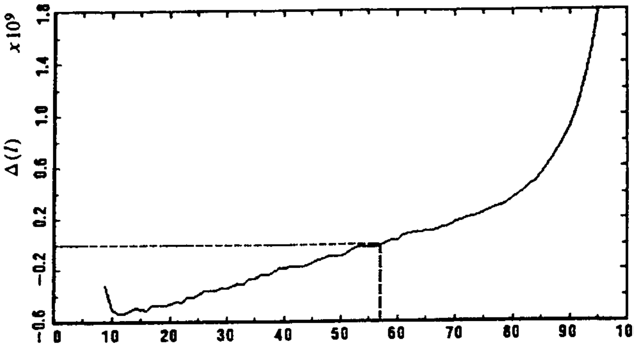


Figure 1. Changes in variance of regression estimator ($\lambda = 1$): $\Delta(l) = V^{(l+1)}(t; 1, N-l-1, n-l-1) - V^{(l)}(t; 1, N-l, n-l)$

Theorem 2 from the complete paper is an important result which allows the solution to the Transfer Algorithm to be equivalently expressed in terms of simpler quantities based on the auxiliary data. A brief sketch of the development of this theorem is given in the Appendix.

Theorem 2. Equivalent Solution to the Transfer Algorithm

The solution $l^*(\lambda)$ to the Transfer Algorithm stated in (3.3) in terms of $V^{(l)} - V^{(l-1)}$ and $\pi^{(l)}_{(N-l)}(\lambda)$ may also be equivalently expressed as

$$l^*(\lambda) = \begin{Bmatrix} \min\limits_{l} \{l : n - l \le R(l; \gamma - \lambda/2), 0 \le l < n\}, 0 \le \lambda < \gamma \\ \min\limits_{l} \{l : n - l \le R(l; \gamma/2), 0 \le l < n\}, \lambda = \gamma \\ \min\limits_{l} \{l : n - l \le R(l; \lambda/2), 0 \le l < n\}, \gamma < \lambda \le 2\gamma \end{Bmatrix}$$

where $R(l; \gamma - \lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} / x_{(N-l)}^{\gamma-\lambda/2}$ and $R(l; \lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2} / x_{(N-l)}^{\lambda/2}$ define the critical values.

This use of this theorem to find the optimal population allocation is illustrated graphically in Figure 2 (Ontario data). In this case, $0 \le \lambda < \gamma$, and the solution is determined by the behaviour of the functions $R(l;\ \gamma - \lambda/2)$ (the lower curve in the graph) and $n - l$. The same solution $l^* = 57$ is obtained as before.

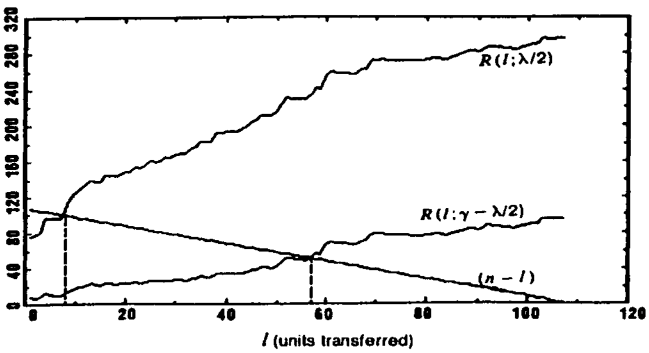


Figure 2. Use of $R(l;\ \gamma - \lambda/2)$, $R(l;\lambda/2)$, and $(n - l)$ to construct optimal take-all/sampled groups (Ontario)

4. SAMPLE SIZE DETERMINATION AND COMBINED ITERATIVE PROCEDURE

Given a sample design $p(s, \lambda)$, $0 \le \lambda \le 2\gamma$, with sample size n, the Transfer Algorithm yields an optimal construction of the take-all and sampled sub-populations, $U_a^*(l^*)$ and $U_b^*(l^*)$, respectively. Next, an expression for finding the minimal sample size is obtained which meets the imposed precision constraint – expressed in terms of the coefficient of variation $\mathrm{CV}_{\min}$. The sample determination step is then integrated with the Transfer Algorithm to develop a combined procedure which allows the survey designer to find the globally minimal sample size and optimal population partitioning.

4.1 Expression for New Sample Size

Let q represent the iteration cycle for the combined procedure and $n_q^* = n_{aq}^* + n_{bq}^*$ denote the total minimal sample size required to satisfy the precision constraint. Given the sample design $p_q(s, \lambda, l_q^*(\lambda, n_q))$, current sample size n_q, and the population partitioning $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$, the precision constraint for $\hat{t}_R = t_a + \hat{t}_{Rb}$ may be stated formally as

$$\mathrm{CV}_{\min} \ge \frac{\hat{V}_q^{1/2}(\hat{t}_{Rb}; \lambda, N - l_q^*, n_q - l_q^*)}{\hat{t}_R}. \tag{4.1}$$

Solving this inequality for n_{bq}^* gives the following expression for the minimal sample size needed in the sampled group $U_{bq}^*(l_q^*)$ to meet the precision constraint:

$$n_{bq}^* = n_q^* - l_q^*(n_q) = \frac{X(l_q^*, \lambda/2)\, X(l_q^*, \hat{\gamma} - \lambda/2)\, \hat{c}}{\hat{t}_R^{\,2}\mathrm{CV}_{\min} + X(l_q^*, \hat{\gamma})\, \hat{c}} \tag{4.2}$$

where $X(l_q^*, \lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\lambda/2}$, $X(l_q^*, \hat{\gamma} - \lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\hat{\gamma}-\lambda/2}$, and $\hat{t}_R$ may be estimated from past survey data corresponding to the period of the auxiliary information. The total new minimal sample size required to meet the precision constraint is then given by

$$n_q^* = n_{aq}^* + n_{bq}^* = l_q^*(n_q) + n_{bq}^*. \tag{4.3}$$

4.2 Combined Sample Redesign Methodology

Next, note that the solution to the Transfer Algorithm l^* depends on the current total sample size: $l_q^*(\lambda) \equiv l_q^*(\lambda, n_q)$. Once the new minimal sample size n_q^* is determined, the existing partitioning $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$ which was optimal at n_q is no longer optimal at the new minimal sample size n_q^* because $l^*(\lambda, n_q^*) \neq l^*(\lambda, n_q)$ if $n_q^* \neq n_q$. Therefore, letting $n_{q+1} = n_q^*$, a new population partitioning from the Transfer Algorithm based on $l_{q+1}^*(\lambda, n_{q+1})$, given by $\{U_{a,q+1}^*(l_{q+1}^*), U_{b,q+1}^*(l_{q+1}^*)\}$, is required to optimize the construction of the take-all and sampled sub-populations. Next, applying (4.2) over $U_{b,q+1}^*(l_{q+1}^*)$ gives a new minimal sample size $n_{q+1}^* = l_{q+1}^*(n_{q+1}) + n_{b,q+1}^*$ required to achieve the desired precision $\mathrm{CV}_{\min}$. Proceeding in this fashion, the combined scheme produces a sequence of population partitionings, sample sizes, and sample allocations

$$\begin{aligned} &(l^*(\lambda, n_q), (n_{aq} = l_q^*, n_{bq} = n_q - l_q^*), \\ &\quad (N_{aq}^* = l_q^*, N_{bq}^* = N - l_q^*), (n_{aq}^* = l_q^*, n_{bq}^*)),\ q = 0, 1, \ldots \end{aligned} \tag{4.4}$$

with $n_{q+1} = n_q^* = n_{aq}^* + n_{bq}^*$ and the initial value n_0 (current survey sample size). The combined procedure is repeated until further reductions in the minimal sample size can no longer be achieved. This leads to the stopping rule

$$q^* = \min_q \{q : n_{q+1}^* - n_q^* \ge 0\}. \tag{4.5}$$

The optimality of the combined procedure can be established using Theorem 2 and is omitted here due to space (see Pandher 1995). The main result is that the combined procedure converges to a globally optimal solution along the path defined by (4.4) to a point where further reductions in the sample size are not possible (by reconstructing U_a^* and U_b^*) given the imposed precision constraint.

5. APPLICATION

The combined sample design procedure described above is now applied to the redesign of the Local Government Finance Survey in the province of Ontario. The survey response y in this application is the actual expenditures reported for sampled local government units for Ontario in 1989. The actual estimates are prepared 30 months after the end of the survey year from financial statements submitted by the local government units to the Department of Municipal Affairs (provincial). Population counts for the local government units from the nearest census (1991) are used as the auxiliary variable x. The population of local-level municipalities for Ontario consists of a total of 793 units of which a sample of 108 units is currently taken.

The results of applying the combined methodology to Ontario LGFS data are reported in Table 1. The level of desired precision $\mathrm{CV}_{\min}$ was set at 2% for the total regression estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$. Using the methods of Pandher (1995), the best value for the heteroscedasticity parameter γ in Ontario was determined to be $\hat{\gamma} = 2$; the corresponding proportionality constant was estimated to be $\hat{c} = .0825$. The near optimal sample design defined by $\lambda = \hat{\gamma}$ $(p(s;\hat{\gamma}))$ was used.

Table 1
Application of Combined Methodology to LGF Survey Data (Ontario, 1989)

Iteration (q)	n_q	$l_q(\lambda, n_q)$	n_{aq}^*	n_{bq}^*	n_q^*
0	108	39	39	18	57
1	57	16	16	34	50
2	50	12	12	38	50

For Ontario the combined scheme stopped at iteration $q^* = 2$. The globally optimal population partitioning between the take-all and sampled groups is $N_a^* = 16$ and $N_b^* = 777$. The new minimal total sample size is $n^* = 50$ with allocations $n_a^* = 16$ and $n_b^* = 34$. A total sample size reduction of $n_0 - n_2^* = 108 - 50 = 58$ is achieved at the desired CV of 2% for the regression estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$.

6. CONCLUDING REMARKS

This paper provides a comprehensive methodology for identifying and implementing an efficient sample design for recurrent surveys of skewed populations. The combined procedure integrates the solution to the following three problems: i) identifying an efficient sample selection scheme, ii) constructing an efficient demarcation between the take-all and sampled population groups at a given sample size, and iii) determining the minimal sample size required to meet the precision constraint(s).

The equivalence result to the Transfer Algorithm (Pandher 1995) was used to create the take-all and sampled groups. The first two components were then combined with a sample size determination step through an iterative procedure. Under the stoping rule, the combined iterative procedure converges to a globally minimal sample size and optimal population partitioning. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario were reported. A 52% reduction in the total sample size was achieved for the regression estimator of the total $(\hat{t}_R = t_a + \hat{t}_{Rb})$ at the desired precision of CV = 2%.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support of Public Institutions Division for sponsoring this work and thank M.P. Singh, H. Mantel, M.S. Kovacevic, S. Wu and the referees for their valuable comments on earlier drafts of the paper.

APPENDIX

A brief sketch of the development behind Theorem 2 (Equivalence Result) is given here; for technical details see Pandher (1995). The same paper also establishes the desirable mathematical properties of the Transfer Algorithm such as existence and optimality of solution as well as the optimality of the combined procedure.

Using the expression for the variance of $V^{(l)}(\hat{t}_{Rb};\cdot)$ given in (3.1), the difference $V^{(l+1)} - V^{(l)}$ may be expressed as

$$\hat{V}^{(l+1)} - \hat{V}^{(l)} = c\,\frac{A(l)\;B(l)}{(n-l)\;(n-l-1)} \tag{A.1}$$

where

$$A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\,x_{(N-l)}^{\lambda/2}$$

and

$$B(l) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} - (n-l)\;x_{(N-l)}^{\gamma-\lambda/2}.$$

The condition $B(l) < 0$ may also be expressed as $n - l > R(l;\gamma - \lambda/2)$ where $R(l;\alpha) = \sum_{k=1}^{N-l} x_{(k)}^{\alpha} / x_{(N-l)}^{\alpha}$. Similarly, the condition $A(l) > 0$ corresponds to $n - l < R(l;\lambda/2)$. All possible states of the system defined by the Transfer Algorithm are summarized in Table A.1.

Table A.1
Outcomes for $V^{(l+1)} - V^{(l)} < 0$ and $V^{(l+1)} - V^{(l)} \geq 0$ in Terms of $n^{(l)} = n - l$

Behaviour of A and B	$V^{(l+1)} - V^{(l)} < 0$ Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) < 0$	$R(l; \gamma - \lambda/2) < n - l < R(l; \lambda/2)$ (T.1)
$A(l) < 0$ $B(l) > 0$	$R(l; \lambda/2) < n - l < R(l; \gamma - \lambda/2)$ (T.3)
	$V^{(l+1)} - V^{(l)} \geq 0$ Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) \geq 0$	$n - l \leq \min\{R(l; \lambda/2),\ R(l; \gamma - \lambda/2)\}$ (T.2)
$A(l) \leq 0$ $B(l) \leq 0$	$n - l \geq \max\{R(l; \lambda/2),\ R(l; \gamma - \lambda/2)\}$ (T.4)

The first column describes the behaviour of $A(l)$ and $B(l)$ leading to the outcome $V^{(l+1)} - V^{(l)} < 0$ and $V^{(l+1)} - V^{(l)} \geq 0$, respectively. The second column describes the equivalent condition in terms of $n^{(l)} = n - l$, $R(l; \gamma - \lambda/2)$, and $R(l; \lambda/2)$ corresponding to $V^{(l)} - V^{(l-1)} < 0$ and $V^{(l)} - V^{(l-1)} \geq 0$, respectively. An important condition required for the solution to the Transfer Algorithm $l^*(\lambda)$ is that $\pi_{(N-l)}(\lambda) < 1$ hold. It is easy to verify that $\pi_{(N-l)}(\lambda) < 1 \Rightarrow A(l) > 0$. In terms of the description for the Transfer Algorithm given in Table A.1, this condition means that the solution can occur only when both $A(l) > 0$ and $B(l) \geq 0$ or, equivalently, when $n - l$ satisfies condition (T.2).

Table A.1 completely enumerates all possible states of the system defined by the Transfer Algorithm. The correspondence between the internal cell quantities (computable directly from the auxiliary data and estimated parameters) and the margins $(A(l), B(l), V^{(l+1)} - V^{(l)})$ represents a tautology which leads directly to Theorem 2 (Equivalence Result). The behaviour of the system described in the table also depends on the sample design $p(s; \lambda)$ employed. The three relevant cases are:

a) $0 \leq \lambda < \gamma \Rightarrow [R(l; \gamma - \lambda/2) < R(l; \lambda/2)]$,

b) $\lambda = \gamma \Rightarrow [R(l; \gamma - \lambda/2) = R(l; \lambda/2)]$, and

c) $\gamma < \lambda \Rightarrow [R(l; \gamma - \lambda/2) > R(l; \lambda/2)]$.

In case a) the system starts ($l = 0$) in state (T.4), moves to (T.1) and then finally rests in state (T.2); state (T.3) is infeasible here. The solution to the Transfer Algorithm $l^*(\lambda)$ is given by the smallest l leading the system to move into state (T.2). In case b), the system starts in state (T.4) and moves to (T.2); (T.1) and (T.3) do not apply. Finally, in case c), the transition path is from (T.4) to (T.3) to (T.2); here (T.1) is invalid.

REFERENCES

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute,* 30, 28-32.

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Annals of Mathematical Statistics,* 36, 1702-1722.

HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics,* 11, 397-405.

LAVALLÉE, P., and HIDIROGLOU, M.A. (1988). On the stratification of skewed populations. *Survey Methodology,* 14, 33-43.

PANDHER, G.S. (1995). Surveys of skewed populations: optimal sample redesign under the generalized regression estimator with applications to the Local Government Finance Survey. Methodology Branch Working Paper: HSMD-95-006, Statistics Canada.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.