Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

On sample allocation for efficient domain estimation

by G. Hussain Choudhry, J.N.K. Rao and Michael A. Hidiroglou

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 23-29
Statistics Canada, Catalogue No. 12-001-X

On sample allocation for efficient domain estimation

G. Hussain Choudhry, J.N.K. Rao and Michael A. Hidiroglou [1]

Abstract

Sample allocation issues are studied in the context of estimating sub-population (stratum or domain) means as well as the aggregate population mean under stratified simple random sampling. A non-linear programming method is used to obtain "optimal" sample allocation to strata that minimizes the total sample size subject to specified tolerances on the coefficient of variation of the estimators of strata means and the population mean. The resulting total sample size is then used to determine sample allocations for the methods of Costa, Satorra and Ventura (2004) based on compromise allocation and Longford (2006) based on specified "inferential priorities". In addition, we study sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified. Performance of the three methods is studied using data from Statistics Canada's Monthly Retail Trade Survey (MRTS) of single establishments.

Key Words: Composite estimators; Compromise allocation; Direct estimators; Domain means; Non-linear programming.

1. Introduction

Stratified simple random sampling is widely used in business surveys and other establishment surveys employing list frames. The population mean $\bar{Y} = \sum_h W_h \bar{Y}_h$ is estimated by the weighted sample mean $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, where $W_h = N_h / N$ is the relative size of stratum h (= 1,..., L) and $\bar{Y}_h$ and $\bar{y}_h$ are the stratum population mean and sample mean respectively. The well-known Neyman sample allocation to strata is optimal for estimating the population mean in the sense of minimizing the variance of $\bar{y}_{st}$ subject to $\sum_h n_h = n$ where n is fixed or minimizing $\sum_h n_h$ subject to fixed variance of $\bar{y}_{st}$, where n_h is the stratum sample size. But the Neyman allocation may cause some strata to have large coefficients of variation (CV) of the means $\bar{y}_h$. On the other hand, equal sample allocation, $n_h = n/L$, is efficient for estimating strata means, but it may lead to a much larger CV of the estimator $\bar{y}_{st}$ compared to that of Neyman allocation.

Bankier (1988) proposed a "power allocation" as a compromise between Neyman allocation and equal allocation. Letting $C_h = S_h / \bar{Y}_h$ be the stratum CV, the power allocation is

$$n_h^B = n \frac{C_h X_h^q}{\sum_h C_h X_h^q}, \; h = 1, \ldots, L \qquad (1.1)$$

where X_h is some measure of size or importance of stratum h and q is a tuning constant. Power allocation (1.1) is obtained by minimizing $\sum_h \{X_h^q \, \mathrm{CV}(\bar{y}_h)\}^2$ subject to $\sum_h n_h = n$, where $\mathrm{CV}(\bar{y}_h)$ is the CV of the stratum sample mean $\bar{y}_h$. The choice $q = 1$ and $X_h = N_h \bar{Y}_h$ in (1.1) leads to Neyman allocation

$$n_h^N = n \frac{N_h S_h}{\sum_h N_h S_h}, \; h = 1, \ldots, L \qquad (1.2)$$

and $q = 0$ gives equal allocation if $C_h = C$ for all h, where S_h^2 is the stratum variance. Bankier (1988) viewed values of q between 0 and 1 as providing compromise allocations. He gave a numerical example to illustrate how q may be chosen in practice. The choice $X_h = N_h$ and $q = 1/2$ in (1.1) gives "square root allocation" $n_h = n\sqrt{N_h} / \sum_h \sqrt{N_h}$ if $C_h = C$. Power allocation (1.1) and some other allocations generally depend on the variable of interest *y* and hence in practice a proxy variable with known population values is used in place of *y*.

Costa *et al.* (2004) proposed a compromise allocation based on a convex combination of proportional allocation, $n_h = nW_h$, and equal allocation $n_h = n/L$, see section 2.1. Longford (2006) made a systematic study of allocation in stratified simple random sampling by introducing "inferential priorities" P_h for the strata h and G for the population. In particular, he assumed that $P_h = N_h^q$ for a specified q ($0 \le q \le 2$), see section 2.4. He also studied the case of small strata sample sizes n_h in which case composite estimators of strata means $\bar{Y}_h$ may be used.

The main purpose of our paper is to propose an "optimal" allocation method, based on non-linear programming (NLP), see section 2.3. It minimizes the total sample size $\sum_h n_h$ subject to specified tolerances on the CVs of the strata sample means $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$. The case of indirect (composite) estimators of strata means is studied in Section 3. In Section 4, we study optimal sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified.

1. G.H. Choudhry, Statistical Research and Innovation Division, Statistics Canada. E-mail: ghchoudhry@gmail.com; J.N.K. Rao, School of Mathematics and Statistics, Carleton University. E-mail: jrao@math.carleton.ca; M.A. Hidiroglou, Statistical Research and Innovation Division, Statistics Canada. E-mail: mike.hidiroglou@statcan.gc.ca.

The proposed method readily extends to multiple variables, but for simplicity we omit details. Using the optimal total sample size obtained from NLP, we make a numerical study of the performances of Costa *et al.* and Longford methods in terms of satisfying reliability requirements, Section 5.

2. Allocation for direct estimators

In this section, we consider direct estimators, $\bar{y}_h$, of strata population means, assuming stratified simple random sampling. The case of indirect estimators of strata means is studied in Section 3. Indirect strata estimators are used in the case of strata with small sample sizes n_h.

2.1 Costa *et al.* allocation

The sample allocation of Costa *et al.* (2004) is

$$n_h^C = k(nW_h) + (1-k)(n/L) \tag{2.1}$$

for a specified constant $k (0 \le k \le 1)$. This allocation reduces to equal allocation when $k = 0$ and to proportional allocation when $k = 1$. Formula (2.1) needs to be modified when $n / L > N_h$ for some h in a set of strata A. The modified allocation is

$$\tilde{n}_h^C = k(nW_h) + (1-k)\, n_h^0, \tag{2.2}$$

where $n_h^0 = N_h$ if $h \in A$ and $n_h^0 = (n - \sum_{h\in A} N_h) / (L - m)$ otherwise, where m is the number of strata in the set A. Note that when $k = 0$, (2.2) gives modified equal allocation. We study different choices of the constant k in the numerical study of Section 5, based on data from Statistics Canada's Monthly Retail Trade Survey (MRTS).

2.2 Longford allocation

Longford's (2006) method attempts to simultaneously control the reliability of the strata means $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$ by minimizing the objective function

$$\sum_{h=1}^{L} P_h V(\bar{y}_h) + (GP_+) V(\bar{y}_{st}) \tag{2.3}$$

with respect to the strata sample sizes n_h subject to $\sum_h n_h = n$, where $P_+ = \sum_h P_h$. The first component in (2.3) specifies relative importance, P_h, of each stratum h while the second component attaches relative importance to $\bar{y}_{st}$ through the weight G. Longford (2006) assumed that $P_h = N_h^q$ for some constant $q (0 \le q \le 2)$. The term P_+ in (2.3) offsets the effect of the sizes P_h and the number of strata on the weight G.

Under stratified simple random sampling, the sample allocation minimizing (2.3) is

$$n_h^L = n \frac{S_h \sqrt{P_h'}}{\sum_h S_h \sqrt{P_h'}},\ h = 1, ..., L \tag{2.4}$$

where $P_h' = P_h + GP_+ W_h^2$. If $q = 2$, then (2.4) does not depend on the value of G and it reduces to Neyman allocation, n_h^N, given by (1.2)

2.3 Nonlinear programming (NLP) allocation

We now turn to the NLP method of determining the strata sample sizes n_h subject to specified reliability requirements on both the strata sample means and the estimated population mean. Letting $\mathbf{f} = (f_1, ..., f_L)^T$ with $f_h = n_h / N_h$, we minimize the total sample size

$$g(\mathbf{f}) = \sum_{h=1}^{L} f_h N_h \tag{2.5}$$

with respect to f subject to

$$\mathrm{CV}(\bar{y}_h) \le \mathrm{CV}_{0h},\ h = 1, ..., L \tag{2.6}$$

$$\mathrm{CV}(\bar{y}_{st}) \le \mathrm{CV}_o \tag{2.7}$$

$$0 < f_h \le 1,\ h = 1, ..., L \tag{2.8}$$

where CV_{0h} and CV_0 are specified tolerances on the CV of the stratum sample mean $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$, respectively. Inequality signs are used in (2.6) and (2.7) because the resulting CVs for some strata h and/or for the aggregate may be smaller than the specified tolerances (Cochran 1977, page 122).

Letting $k_h = f_h^{-1}$, (2.5) becomes a separable convex function of the variables k_h,

$$\tilde{g}(\mathbf{k}) = \sum_{h=1}^{L} N_h k_h^{-1}. \tag{2.9}$$

We re-specify the constraints (2.6) and (2.7) in terms of relative variances so that the constraints are linear in the variables k_h. The relative variance (RV) of $\bar{y}_h$ is the square of its CV,

$$\mathrm{RV}(\bar{y}_h) = \frac{k_h - 1}{N_h} C_h^2. \tag{2.10}$$

Similarly, the relative variance of $\bar{y}_{st}$ is the square of its CV,

$$\mathrm{RV}(\bar{y}_{st}) = \bar{Y}^{-2} \sum_{h=1}^{L} W_h^2 \frac{k_h - 1}{N_h} S_h^2. \tag{2.11}$$

We used the SAS procedure NLP with the Newton-Raphson option to find the optimal k_h that would minimize (2.9) subject to

$$\mathrm{RV}(\bar{y}_h) \le \mathrm{RV}_{oh},\ h = 1, ..., L, \tag{2.12}$$

$$\mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_0, \tag{2.13}$$

$$k_h \ge 1,\ h = 1, ..., L. \tag{2.14}$$

$\mathrm{RV}(\bar{y}_h)$ and $\mathrm{RV}(\bar{y}_{st})$ are given by (2.10) and (2.11) where $\mathrm{RV}_{0h} = \mathrm{CV}_{0h}^2$ and $\mathrm{RV}_0 = \mathrm{CV}_0^2$. By expressing the constraints as linear constraints and the objective function as a separable convex function, we achieve faster convergence of the re-formulated NLP. Denoting the solution to NLP as $\mathbf{k}^0 = (k_1^0, ..., k_L^0)^T$, the corresponding vector of optimal strata sample sizes is given by $\mathbf{n}^0 = (n_1^0, ..., n_L^0)^T$, where $n_h^0 = N_h / k_h^0$. We can modify (2.14) to ensure that $n_h^0 \ge 2$ for all h which permits unbiased variance estimation.

The NLP method can be readily extended to multiple variables $y_1, ..., y_P$ by specifying tolerances on the CVs of strata means and the estimated population mean for each variable $(p = 1, ..., P)$. If the number of variables P is not small, then the resulting optimal total sample size $n^0 = \sum_h n_h^0$ may increase significantly relative to n^0 for a single variable. Huddleston, Claypool and Hocking (1970), Bethel (1989) and others studied NLP for optimal sample allocation in the case of estimating population means of multiple variables under stratified random sampling.

3. Allocation for composite estimators

Longford (2006) studied composite estimators of strata means of the form

$$\hat{\theta}_h = \alpha_h \bar{y}_h^S + (1 - \alpha_h)\bar{y}_h \tag{3.1}$$

where $\bar{y}_h^S$ is a synthetic estimator; here we take $\bar{y}_h^S = \bar{y}_{st}$. The MSE of $\hat{\theta}_h$ is

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_h) &= V(\hat{\theta}_h) + [B(\hat{\theta}_h)]^2 \\ &= \alpha_h^2 \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} + (1-\alpha_h)^2 W_h^2 \frac{S_h^2}{n_h} \\ &\quad + 2\alpha_h(1-\alpha_h) W_h \frac{S_h^2}{n_h} + \alpha_h^2 (\bar{Y}_h - \bar{Y})^2 \\ &\quad + \text{terms not depending on the } n_h. \end{aligned} \tag{3.2}$$

Longford (2006) showed that the optimal coefficient α_h in (3.1) minimizing (3.2) is approximately equal to $\alpha_h^* = S_h^2 (S_h^2 + n_h \Delta_h^2)^{-1}$, where $\Delta_h = \bar{Y}_h - \bar{Y}$. He then replaced Δ_h^2 in α_h^* by its average over the strata, denoted by $\sigma_B^2 = L^{-1} \sum_h (\bar{Y}_h - \bar{Y})^2$, leading to $\alpha_h^* \approx (1 + n_h \omega_h)^2$, where $\omega_h^2 = \sigma_B^2 / S_h^2$. The resulting MSE of $\hat{\theta}_h$ is approximated as

$$\mathrm{MSE}(\hat{\theta}_h) \approx \frac{\sigma_B^2}{1 + n_h \omega_h}. \tag{3.3}$$

Longford's allocation is obtained by minimizing the objective function

$$\sum_{h=1}^{L} P_h \mathrm{MSE}(\hat{\theta}_h) + (\mathrm{GP}_+) V(\bar{y}_h) \tag{3.4}$$

with respect to the n_h. The resulting solution satisfies

$$\frac{P_h \sigma_B^2 \omega_h}{(1 + n_h \omega_h)^2} + (\mathrm{GP}_+) W_h^2 \frac{S_h^2}{n_h} = \text{const.}, \ h = 1, ..., L. \tag{3.5}$$

Longford used an iterative method to obtain the solution to (3.5) since it does not have a closed-form solution.

Our NLP procedure minimizes $g(\mathbf{f})$ given by (2.5) subject to

$$\mathrm{RMSE}(\hat{\theta}_h) \le \mathrm{RMSE}_{0h}, h = 1, ..., L;\ \mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_o \tag{3.6}$$

and (2.8), where $\mathrm{RMSE}(\hat{\theta}_h) = \mathrm{MSE}(\hat{\theta}_h) / \bar{Y}_h^2$ and RMSE_{0h} is a specified tolerance. The approximation (3) to $\mathrm{MSE}(\hat{\theta}_h)$ is used in (3.6).

4. Allocation for domain estimation

Suppose that the population U is partitioned into domains ${}_dU$ $(d = 1, ..., D)$ that cut across the strata. Also, suppose that the estimators of domain means need to satisfy specified relative variance tolerances, ${}_d\mathrm{RV}_0$, $d = 1, ..., D$. We find the optimal additional strata sample sizes that are needed to satisfy the domain tolerances, using the NLP method.

An estimator of domain mean ${}_d\bar{Y} = {}_dN^{-1} \sum_{k \in U_d} y_k$ is the ratio estimator

$${}_d\hat{\bar{Y}} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k \, y_k}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k}, \tag{4.1}$$

where ${}_d\delta_k = 1$ if $k \in {}_dU$ and ${}_d\delta_k = 0$ otherwise, s_h is the sample from stratum h and ${}_dN$ is the size of domain d. The relative variance of the ratio estimator (4.1) is $\mathrm{RV}({}_d\hat{\bar{Y}}) = V({}_d\hat{\bar{Y}}) / {}_d\bar{Y}^2$, where the variance $V({}_d\hat{\bar{Y}})$ is obtained by the usual linearization formula for a ratio estimator.

Let $\tilde{n}_h$ denote the revised total sample size from stratum h so that the sample increase from stratum h is $\tilde{n}_h - n_h^0$. Let $\tilde{f}_h = \tilde{n}_h / N_h$ be the corresponding sampling fraction. We obtain the optimal $\tilde{\mathbf{n}} = (\tilde{n}_1, ..., \tilde{n}_L)^T$ by minimizing the sample increase

$$g(\tilde{\mathbf{f}}) - \sum_{h=1}^{L} n_h^0 N_h = \sum_{h=1}^{L} (\tilde{f}_h - f_h^0) N_h \tag{4.2}$$

with respect to $\tilde{\mathbf{f}} = (\tilde{f}_1, ..., \tilde{f}_L)^T$ subject to

$$f_h^0 \le \tilde{f}_h \le 1, \quad h = 1, ..., L \tag{4.3}$$

$$\mathrm{RV}(\hat{\bar{Y}}_d) \le {}_d\mathrm{RV}_0, \quad d = 1, ..., D. \tag{4.4}$$

As before, we reformulate the problem by expressing (4.2), (4.3) and (4.4) in terms of $\tilde{\mathbf{k}} = (\tilde{k}_1, ..., \tilde{k}_L)^T$, where $\tilde{k}_h = \tilde{f}_h^{-1}$. This leads to minimization of the separable convex function

$$g^*(\tilde{\mathbf{k}}) = \sum_{h=1}^{L} N_h \tilde{k}_h^{-1} \tag{4.5}$$

with respect to $\tilde{\mathbf{k}}$ subject to the linear constraints

$$1 \le \tilde{k}_h \le k_h^0, \quad h = 1, ..., L \tag{4.6}$$

and

$$\mathrm{RV}(\hat{\bar{Y}}_d) = {}_d\bar{Y}^{-2} \sum_{h=1}^{L} \left(\frac{N_h}{{}_dN} \right)^2 \frac{\tilde{k}_h - 1}{N_h} {}_dS_{eh}^2 \le {}_d\mathrm{RV}_0, \quad d = 1, ..., D \tag{4.7}$$

where ${}_d\mathrm{RV}_0$ is the specified tolerance, ${}_dS_{eh}^2$ denotes the stratum variance of the residuals ${}_de_k = {}_d\delta_k(y_k - {}_d\bar{Y})$ for $k \in U_h$ and U_h denotes the stratum population. Denote the resulting optimal $\tilde{k}_h$ and $\tilde{n}_h$ as $\tilde{k}_h^0$ and $\tilde{n}_h^0$ respectively, so that the optimal sample increase in stratum h is $\tilde{n}_h^0 - n_h^0$.

It can be shown that the minimization of total sample size subject to all the constraints $\mathrm{RV}(\bar{y}_h) \le \mathrm{RV}_{0h}$, $h = 1, ..., L$, $\mathrm{RV}(\hat{\bar{Y}}_d) \le {}_d\mathrm{RV}_0$, $d = 1, ..., D$, $\mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_0$, and $0 < f_h \le 1$, $h = 1, ..., L$ will lead to the same optimal solution, $\tilde{n}^0 = (\tilde{n}_1^0, ..., \tilde{n}_L^0)^T$. However, domain reliability requirements may often be specified after determining n^0.

5. Empirical results

In this section, we study the relative performance of different sample allocation methods, using data from the MRTS. Section 5.1 and 5.2 report our results for direct estimators and composite estimators of strata means, respectively. Results for the domain means are given in section 5.3.

5.1 Strata means: Direct estimators

For the empirical study, we used a subset of the MRTS population values restricted to single establishments. Strata sizes, N_h, strata population means, $\bar{Y}_h$, strata standard deviations, S_h, and strata CVs, $C_h = S_h / \bar{Y}_h$, are given in Table 1 for the ten provinces in Canada (treated as strata). For the NLP allocation, we have taken the CV tolerances as $\mathrm{CV}_{0h} = 15\%$ for the strata means $\bar{y}_h$ and $\mathrm{CV}_0 = 6\%$ for the weighted sample mean $\bar{y}_{st}$, denoted Canada (CA).

The NLP allocation satisfying the specified CV tolerances resulted in a minimum overall sample size $n^0 =$ 3,446. Table 2 reports the sample allocation n_h^0 and the associated $\mathrm{CV}(\bar{y}_h)$ and $\mathrm{CV}(\bar{y}_{st})$ for the NLP allocation. It shows that the NLP allocation respects the specified tolerance $\mathrm{CV}_0 = 6\%$, gives CVs smaller than the specified tolerance $\mathrm{CV}_{0h} = 15\%$ for two of the larger provinces (QC: 11.4% and ON: 11.0%) and attains a 15% CV for the remaining provinces.

Table 1
Population values for the MRTS

Provinces	N_h	$\bar{Y}_h$	S_h	C_h
Newfoundland (NL)	909	963	1,943	2.02
Price-Edward-Island (PE)	280	712	1,375	1.93
New-Brunswick (NB)	1,333	1,368	3,200	2.34
Nova-Scotia (NS)	1,153	1,568	4,302	2.74
Quebec (QC)	11,135	2,006	4,729	2.36
Ontario (ON)	21,531	1,722	6,297	3.66
Manitoba (MN)	1,700	1,295	2,973	2.30
Saskatchewan (SK)	1,743	1,212	3,019	2.49
Alberta (AL)	5,292	1,698	5,358	3.16
British Columbia (BC)	7,803	1,291	4,013	3.11
Canada (CA)	52,879	1,654	-	-

Table 2
Equal, proportional, square root and NLP allocations and associated CVs (%)

Province	Equal		Proportional		Square-Root		NLP	
	n_h	CV_h	n_h	CV_h	n_h	CV_h	n_h	CV_h
NL	352	8.4	59	25.4	169	14.0	151	15.0
PE	280	0.0	18	44.1	94	16.2	104	15.0
NB	352	10.7	87	24.2	205	15.0	206	15.0
NS	352	12.2	75	30.6	191	18.1	259	15.0
QC	352	12.4	726	8.5	593	9.4	410	11.4
ON	352	19.3	1,403	9.4	824	12.5	1,056	11.0
MN	352	10.9	111	21.1	232	14.0	206	15.0
SK	352	11.9	114	22.6	234	15.2	238	15.0
AL	352	16.3	345	16.4	408	15.0	409	15.0
BC	352	16.2	508	13.3	496	13.5	407	15.0
CA	3,446	9.1	3,446	5.2	3,446	6.3	3,446	6.0

Using the optimal overall sample size 3,446, we calculated the sample allocations n_h and the associated $\mathrm{CV}(\bar{y}_h)$ and $\mathrm{CV}(\bar{y}_{st})$ for the modified equal allocation, proportional allocation and square-root allocation, reported in Table 2. It is clear from Table 2 that the modified equal allocation is not suitable in terms of satisfying specified CV tolerances because it leads to $\mathrm{CV}(\bar{y}_{st}) = 9.1\%$ which is significantly larger than the specified $\mathrm{CV}_0 = 6\%$. Also, under the modified equal allocation, $\mathrm{CV}(\bar{y}_h)$ equals 19.3%, 16.3% and 16.2% for the larger provinces ON, AL and BC respecttively. Note that for the smallest province PE Table 2 gives $\mathrm{CV}(\bar{y}_h) = 0\%$ for the modified equal allocation because for PE it gives $n_h = N_h$.

Turning to proportional allocation, Table 2 reports $CV(\bar{y}_{st}) = 5.2\%$ but it leads to considerably larger strata CVs relative to the specified 15% for seven of the provinces, ranging from 16.4% to 44.1%. On the other hand, Table 2 shows that square-root allocation offers a reasonable compromise in terms of desired CV tolerances. We have $CV(\bar{y}_{st}) = 6.3\%$ and $CV(\bar{y}_h) \leq 15\%$ for seven of the provinces and the three provinces with CVs greater than 15% are SK with 15.2%, PE with 16.2% and NS with 18.1%.

Table 3 reports the results for the Costa *et al.* allocation (2.1) with $k = 0.25$, 0.50 and 0.75, using $n = 3{,}446$ obtained from NLP. We observe from Table 2 that the choice $k = 0.25$, which assigns more weight to equal allocation, is not satisfactory for the estimation of the population (Canada) mean: $CV(\bar{y}_{st}) = 7.2\%$, but performs well for strata means, except AL with $CV(\bar{y}_h) = 16.3\%$. On the other hand, the choice $k = 0.75$, which assigns more weight to proportional allocation, performs poorly in estimating provincial means, with $CV(\bar{y}_h)$ ranging from 16.2% to 21.4% for seven of the provinces, although $CV(\bar{y}_{st})$ is smaller than the desired tolerance, 6%. The compromise choice $k = 0.50$ performs quite well, leading to $CV(\bar{y}_{st}) =$ 6.4% and $CV(\bar{y}_h)$ around 15% or less except for two provinces (NS and AL) with CVs of 17.0% and 16.5% respectively. Performance of the Costa *et al.* method with $k = 0.50$ and square-root allocation are somewhat similar, and both allocations do not depend on the variable of interest, y, unlike the Longford and NLP allocations.

We next turn to Longford's allocation (2.4) which depends on q and G. Table 4 provides results for $q = 0$, 0.5, 1.0, 1.5 and $G = 0, 10, 100$, using $n = 3{,}446$ obtained from NLP. For $q = 2.0$, Longford's allocation does not depend on G and in fact it reduces to the Neyman allocation (1.2) which minimizes $CV(\bar{y}_{st})$ for fixed n but leads to highly inflated $CV(\bar{y}_h)$, ranging from 16% to 85% for seven provinces. We see from this table that $CV(\bar{y}_h)$, for a given q, increases with G rapidly while $CV(\bar{y}_{st})$ decreases slowly as G increases and in fact is virtually a constant ($\approx 5.1\%$) for $G > 100$ (values not reported here). Also, $CV(\bar{y}_h)$ for a given G, increases rapidly as q increases while $CV(\bar{y}_{st})$ decreases. Langford's allocation, for $q \geq 0.5$ and/or $G \geq 10$, leads to significantly larger $CV(\bar{y}_h)$ than the specified tolerance $CV_{0h} = 15\%$ for several provinces, even though $CV(\bar{y}_{st})$ respects the specified tolerance of 6%. On the other hand, for $q = 0$ and $G = 0$, $CV(\bar{y}_h)$ is below the specified tolerance except for BC with 15.7%, but $CV(\bar{y}_{st}) = 7.3\%$ significantly exceeds the specified tolerance. For $q = 1.0$ and $q = 1.5$, $CV(\bar{y}_{st})$ stays below 6% when $G = 0$, but $CV(\bar{y}_h)$ exceeds 15% for six provinces, ranging from 17.7% to 34.0% for $q = 1.0$ and 22.0% to 54.6% for $q = 1.5$. On the whole, Table 4 suggests that no suitable combination of q and G can be found that ensures that all the specified reliability requirements are satisfied even approximately.

Table 3
Costa *et al.*'s allocation and associated CVs (%) for k = 0.25, 0.50 and 0.75

Province	k = 0.25		k = 0.50		k = 0.75	
	n_d	CV_d	n_d	CV_d	n_d	CV_d
NL	278	10.1	205	12.4	132	16.2
PE	214	6.4	149	10.8	83	17.8
NB	286	12.3	219	14.5	153	17.8
NS	282	14.2	213	17.0	144	21.4
QC	446	10.9	539	9.9	633	9.1
ON	615	14.5	878	12.1	1,140	10.5
MN	292	12.2	231	14.0	171	16.6
SK	292	13.3	733	15.2	174	17.9
AL	350	16.3	349	16.3	347	16.4
BC	391	15.3	430	14.6	469	13.9
CA	3,446	7.2	3,446	6.2	3,446	5.6

Table 4
CVs (%) for Longford's allocation with q = 0, 0.5, 1.0, and 1.5

Province	q = 0			q = 0.5			q = 1.0			q = 1.5		
	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100	G = 0	G = 10	G = 100
NL	13.5	19.3	29.7	17.2	23.0	33.4	22.7	29.0	38.3	30.4	36.2	40.6
PE	12.7	20.4	34.6	21.4	29.6	48.5	34.0	45.4	67.3	54.6	67.3	85.6
NB	12.0	17.1	25.0	14.5	19.4	26.8	18.3	23.1	29.0	23.5	27.6	30.3
NS	11.1	16.7	25.5	14.2	19.5	27.9	18.7	24.1	30.9	24.9	29.4	32.8
QC	11.0	9.8	9.1	9.9	9.4	9.0	9.2	9.0	8.9	8.9	8.9	8.8
ON	14.9	9.8	8.7	12.3	9.5	8.6	10.5	9.1	8.5	9.3	8.7	8.5
MN	12.7	17.6	24.3	14.7	19.1	25.2	17.7	21.9	26.5	22.0	25.4	27.5
SK	13.6	18.9	25.9	15.7	20.5	26.9	19.0	23.5	28.3	23.5	27.0	29.4
AL	13.5	15.7	16.1	13.3	15.2	15.9	13.6	15.2	15.9	14.6	15.5	15.9
BC	15.7	16.1	15.4	14.7	15.4	15.3	14.3	15.0	15.1	14.5	15.0	15.1
CA	7.3	5.5	5.1	6.2	5.3	5.1	5.5	5.2	5.1	5.2	5.1	5.1

5.2 Strata means: Composite estimators

We now report some results for the composite estimators, $\hat{\theta}_h$, of strata means. We obtained the optimal total sample size as $n = 3{,}368$ using NLP and the reliability requirements (3.6). This value is slightly smaller than the optimal $n^0 =$ 3,446 for the direct estimators. For the Longford allocation, we used $n = 3{,}368$ and calculated the sample allocation and associated CVs of the composite estimators $\hat{\theta}_h$ and the weighted mean $\bar{y}_{st}$ for specified q and G, constraining n_h to be at least two. For the MRTS data we have used, the first term of (3.5) is small relative to the second term. As a result, the sample allocation is flat across G - values for a given q which means that the CVs for the Longford allocation did not vary significantly with G.

Therefore, we have reported results in Table 5 only for $G = 0$ and $q = 0, 0.5, 1.0$ and 1.5. We note from Table 5 that $\mathrm{CV}(\hat{\theta}_h)$ decreases with q for the two largest provinces (QC and ON) because the sample shifts from the smaller provinces to these two provinces as q increases. Also, $\mathrm{CV}(\hat{\theta}_h)$ initially decreases for AL and BC but it starts increasing when q is large because the sample starts shifting to QC and ON from these provinces as well. Further, $\mathrm{CV}(\hat{\theta}_h)$ increases for all other provinces with q except for NS for which it starts decreasing for large q because of larger synthetic component and very negligible bias. In particular, $\mathrm{CV}(\hat{\theta}_h)$ increases rapidly for NL and PE because of very large bias.

Table 5
CVs (%) for the composite estimators using Longford's allocation: $G = 0$ and $q = 0, 0.5, 1.0$ and 1.5

Province	$q = 0$	$q = 0.5$	$q = 1.0$	$q = 1.5$
NL	12.7	17.0	24.2	37.3
PE	12.4	23.8	46.0	112.2
NB	10.4	12.8	16.1	20.4
NS	9.4	11.9	14.5	11.7
QC	10.3	9.0	8.3	8.0
ON	13.9	11.1	9.3	8.2
MN	11.2	13.1	16.0	20.3
SK	12.4	14.6	17.9	23.2
AL	11.4	11.2	11.5	12.2
BC	14.4	13.3	12.9	13.1
CA	8.0	6.3	5.4	5.6

On the other hand, $\mathrm{CV}(\bar{y}_{st})$ decreases initially with q but starts increasing when q is large because most of the sample gets allocated to QC and ON and very little sample is assigned to the smaller provinces. It appears from Table 5 that the Longford allocation performs reasonably well only for $q = 0$ and $G = 0$, giving $\mathrm{CV}(\hat{\theta}_h)$ less than 15% for all provinces at the expense of $\mathrm{CV}(\bar{y}_{st}) = 8.0\%$.

5.3 Domain means

Establishments on the Canadian Business Register are classified by industry using the North American Industry Classification System (NAICS). NAICS is principally a classification system for establishments and for the compilation of production statistics. The industry associated with each establishment on the Canadian Business Register is coded to six digits using the North American Industry Classification System. There are 67 six digit codes associated with the Retail sector. These six digit codes are regrouped into 19 trade groups (TG) for publication purposes.

We took the trade groups as domains that cut across provinces (strata). The trade group with the smallest number of establishments is TG 110 (beer, wine and liquor stores) with 307 establishments and the TG with the largest number of establishments is TG 100 (convenience and specialty food stores) with 7,752 establishments. Establishments were coded to all the 19 trade groups in all but one province: in PE, establishments were coded to only 16 trade groups.

We applied NLP based on (4.5), (4.6) and (4.7), and obtained the optimal total sample size increase to meet specified reliability requirements on the domain estimators ${}_d\hat{\bar{Y}}$. We found that no increase in the total sample size is needed if the tolerance on $\mathrm{CV}({}_d\hat{\bar{Y}})$ is less than or equal to 30% for each domain. If the tolerance is reduced to 25%, then the optimal total sample size increase is 622 and the total sample size after the increase is 4,068. If the tolerance is further reduced to 20%, then the optimal total sample size increase is 2,100 and the total sample size after the increase is 5,546, which is considerably larger than the original 3,446. Note that as the total sample size is increased, CVs of strata means $\bar{y}_h$ and the weighted sample mean $\bar{y}_{st}$ decrease.

6. Summary and concluding remarks

We have proposed a non-linear programming (NLP) method of sample allocation to strata under stratified random sampling. It minimizes the total sample size subject to specified tolerances on the coefficient of variation of estimators of strata means and the population mean. We considered both direct estimators of strata means and composite estimators of strata means. The case of domains cutting across strata is also studied. Difficulties with alternative methods in satisfying specified reliability requirements are demonstrated using data from the Statistics Canada Monthly Retail Trade Survey of single establishments. We also noted that NLP can be readily extended to handle reliability requirements for multiple variables. Compromise allocations that perform reasonably well in terms of reliability requirements are also noted.

Acknowledgements

The authors thank two referees and an associate editor for constructive comments and suggestions.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 47-57.

Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York : John Wiley & Sons, Inc.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT*, 28, 69-86.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimum allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

North American Industry Classification System, Version 1.4 (2008). Catalogue 12F0074XCB. Statistics Canada.