A Procedure for Computing Optimal Stratum Boundaries and Sample Sizes for Multivariate Surveys

Karuna G. Reddy*, Mohammed G. M. Khan, Dinesh K. Rao

School of Computing, Information and Mathematical Sciences, The University of South Pacific, Suva, Fiji.

* Corresponding author. Tel.: +679 3232194; email: reddy_k@usp.ac.fj
Manuscript submitted January 13, 2016; accepted March 20, 2016.
doi: 10.17706/jsw.11.8.816-832

Abstract: In most surveys, the target variables (items of interest) commonly resemble right-skewed distributions where the Stratified Random Sampling technique is used as a method of sampling and estimation. The methodology of constructing strata is called stratification. Over a particular characteristic chosen as the stratification variable (such as gender, geographical region, ethnicity, or any natural criteria), the survey may fail to form homogeneous strata - this would impact the precision in the estimates of the target variables. Stratification can lead to substantial improvements in the precision of sample estimators, which not only depends on the sample size, but also on the heterogeneity among the units of the population. The principal reason for stratification in the design of sample surveys is to reduce the variance of sample estimates. Surveys normally have more than one target variable with several variables both available and desirable for stratification. Stratification in such multivariate situations has not been explored to a great deal like the univariate case and requires algorithms to determine efficient stratum boundaries. This paper takes into consideration multiple survey variables and attempts to present a computational procedure to construct optimal stratum boundaries (OSB) using Dynamic Programming (DP) technique. A numerical example to determine the OSB for two main variables under study is also presented.

Key words: Multivariate sample surveys, optimum stratum boundaries (OSB), mathematical programming problem (MPP), dynamic programming technique (DP).

1. Introduction

Stratified Sampling is an important sampling technique used in surveys in almost all fields, be it business, economic, social sciences or health. Sociology and other sciences arrange individuals into social strata using demographic and socio-economic factors to explore inequalities between groups. Geology classifies layers of soil into strata for analysis, and biology can consider strata in the context of layers of tissue. In stratified sampling, the sampling-frame is divided into non-overlapping groups or strata in such a way that the strata constructed are internally homogeneous with respect to the survey variable (y) under study that maximizes the precision of its estimate. Stratification by convenience manner (such as choosing strata using demographic, socio-economic fact*or*sors, or any natural characteristic) is not always a reasonable criterion as the strata so obtained may not be internally homogeneous with respect to the variable of interest. Thus, one has to create homogenous sub-populations by determining optimum strata boundaries (OSB) so that the variances within groups are minimised and hence the precision of overall population parameters are improved. The concept of optimal stratification extends the usual concepts of providing algorithms for estimating the number of strata, the determination of stratum boundaries, and the allocation of sample

units among the strata, in order to minimize the variance of estimates. This paper will look at the concept of determining OSB.

Determining OSB for a survey variable by using its frequency distribution f(y) is well documented in sampling literature. In order to achieve maximum precision in the OSB, the stratum variances σ_h^2 should be as small as possible for a given type of sample allocation. If distribution of the variable under study is known, the OSB is determined by cutting the range of this distribution at suitable points. This problem of determining OSB was first discussed by [1] who presents a set of minimal equations that are difficult to solve. Subsequently, attempts for determining approximate OSB have been made by several authors [2]–[7].

Several other numerical and computational methods [5], [7]–[29] have been developed for determining OSB when a single main variable is under study.

2. Related Works

Notably, all the methods mentioned above are univariate procedures that primarily deal with a single survey variable. The multivariate situation is more common practically, with several variables both available and desirable for stratification. Most real-world surveys are multipurpose - they have several variables, and many statistics compete for attention as the principal objectives of survey efforts. It is usually better to utilize several variables rather than just one - this often would be true even if the best one were known. The advantages of multivariate stratification can be appreciated with the fact that with increasing number of survey variables, there are reductions in the variance within strata and gains in precision of the estimates. However, construction of strata for the multivariate case is usually not as natural as it is in the univariate case.

Efforts made to take into consideration two or more variables have involved the generalization of univariate procedures to determine the boundaries. An analytic method for multivariate stratification was proposed by [30]. They used principal component analysis on the stratification variables and then formed the strata using only the first two components as stratification variables. Pla [31] extended this by applying approximate univariate methods, using the first principal component as the stratification variate. The method reduced the generalized and total variances, and outperforms the univariate or bivariate procedures for total and linear mean-vector variances and it found out that the relative gain is independent of the total sample size. This multivariate method of constructing OSB provides a step forward in solving multi-dimensional stratification problems. It has been established that the cumulative $\sqrt{f}$ rule proposed in [8] performs so well that for all practical purposes, it gives OSB in the bivariate case [32]. In the multi-dimensional case, applying the cumulative $\sqrt{f}$ rule to the first principal component also gives the maximum reduction in the variance within the strata.

Ghosh [33] extended the theory in [1] to a bivariate population (x,y) whereby he proposed the problem of optimum stratification with two characters under proportional allocation, assuming stratification variable to be identical to the estimation variable for a fixed number of strata. The generalized variance of the sample means was taken as a measure of precision for the bivariate characters. He proposed that a two-way rectilinear stratification would be optimum if the generalized variance of the sample means or the unbiased linear estimate is minimized. Sadasivan and Aggarwal [34] considered bivariate stratification under the Neyman-optimum allocation. They also extended the univariate method in [1] to the bivariate case by taking study variables as the basis for stratification, minimizing the generalized variance of the equations to get the OSB for multiple strata.

Samanta [35] considered optimum stratification for multiple variables under proportional allocation by

minimizing the generalized variance of estimators under the assumption that stratification variables are identical to survey variables. Procedures for more than two stratification variables were also suggested by [36] but could not be used when there were several stratification variables and relatively few units.

In an analytic approach to multivariate stratification, the cluster analysis method was used where the total within sum of squares of the stratification variables was minimized. This approach was introduced by [37] and was also treated by [38]. Thomsen [39] presented an approximation to the variance of the study variable under the assumption of a linear regression on two stratification variables where he demonstrated that under some conditions, one can expect a considerable reduction of the variance using two variables. Iachan [40] used a stratification method based on prior information - his method produced better results than simple random sampling.

Stratification was also investigated using cluster analysis in [37], [41], and [42]. Following their concepts, [43] proposed the concept of stratification as an optimization problem under the clustering approach which achieved more efficient stratification than other authors in this area. The optimization function to be minimized was formulated as the sum of the multivariate within-strata variances. The results in this were not considered as optimum since the constraints were not fulfilled. The problem of optimum stratification of two variates under the proportional and optimum allocation in the case of bivariate normal distribution was considered by [44] where two stratification attributes were taken into account - correlation coefficient between the characters and a sampling fraction. The theory of optimum bivariate stratification under the proportional allocation was developed by [45]. The problem of optimum bivariate stratification for two characters under the compromise allocation [46] was considered and a cumulative cube root rule for stratification was proposed.

Another approach to defining strata is via so-called *L–rot–*180 stratification geometry [47], [17]. Under this peculiar stratification geometry, in the bivariate case, strata, the elements of which are points on a plane, have a form of the capital L rotated through 180 degrees and can be generalized to more than two dimensions. Later, the *L–rot–*180 geometry was applied by [19], who presented stratification under the compromise allocation in a problem of stratifying a multivariate population orientated towards minimizing the maximum coefficient of variation of estimators studied.

In the multi-way stratification approach [48], the values of each stratifying variable are first categorized to define univariate strata and then the multi-way strata are simply obtained as the intersections of these univariate strata across all variables. They used cumulative $\sqrt{f}$ on one variate to generate a set of strata and then repeated the procedure on the other variables. Final strata were defined by the cross-combination of different numbers of cut-points on the two variates. The multi-way stratification shows large gains when the numbers of cut-points of the variates are similar, and when the correlations with the survey variables are high but a practical problem is that this number can be very large. A multi-way stratification approach was also proposed by [49] and [50] using the simple idea of linear programming. Large problems did become feasible, however, the main problem was that it was too huge computationally as the number of cells in the multi-way stratification increases, to the extent that it cannot be used in most realistic situations.

In this paper, a computational technique of determining OSB using a dynamic programming technique is proposed for multivariate survey variables. The proposed technique is illustrated with a numerical example that requires the construction of OSB for two study variables and can be extended to many survey variables. Sample sizes are also determined together with a comparative on the performance of the proposed method against other methods.

3. Preliminaries

In this section, the basic concepts on Stratified Random Sampling are presented together with a review of some of the foundation principles with regards to the construction of OSB.

Let the target population with p variables under study be stratified into L strata where the estimation of the means of these study variables $y_1, y_2, \ldots, y_p$ are of interest. If a simple random sample of size n_{jh} is to be drawn from h^{th} stratum with sample mean $\bar{y}_{jh}; (j = 1, 2, \ldots, p; h = 1, 2, \ldots, L)$, then the stratified sample mean, $\bar{y}_{jst}$, is given by

$$\bar{y}_{jst} = \sum_{h=1}^{L} W_{jh} \bar{y}_{jh} \tag{1}$$

where W_{jh} is the proportion of the population of the j^{th} variable contained in the h^{th} stratum. When the finite population correction factors are ignored, under the [51] allocation:

$$n_{jh} = n \frac{W_{jh}\sigma_{jh}}{\sum_{h=1}^{L} W_{jh}\sigma_{jh}} \tag{2}$$

the variance of $\bar{y}_{jst}$ is given by

$$Var(\bar{y}_{jst}) = \frac{\left(\sum_{h=1}^{L} W_{jh}\sigma_{jh}\right)^2}{n} \tag{3}$$

where σ_{jh}^2 is the stratum variance for the j^{th} main variable in the h^{th} stratum; $h = 1, 2, \ldots, L$ respectively and n is the preassigned total sample size.

4. The Proposed Scheme

In this section, a method of constructing OSB for each stratum is developed for multiple survey (or study) variables, which leads to substantial gains in the precision of the estimates. The determination of OSB based on the survey variable is not feasible in practice since the variable of interest is unavailable prior to conducting the survey, however, this method assumes that with prior surveys, the nature of distributions and initial values are known for the survey variables. The problem of finding the OSB is formulated as Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter under Neyman allocation. The MPP is then solved for OSB by developing a solution procedure of dynamic programming technique. A numerical example with simulated data sets of skewed populations that follow a 3-parameter (3P) Weibull and Gamma distributions.

Let $f(y_1), f(y_2), \ldots, f(y_p); a_j \le y_j \le b_j$ be the frequency functions of p main study variable, $y_1, y_2, \ldots, y_m$ on which compromise strata boundaries are to be constructed. If the population means of these study variables are estimated under Neyman allocation given in (2), then the problem of determining compromise strata boundaries is to cut up the range, $d = b - a$, at $(L-1)$ at intermediate points $a = y_0 \le y_1 \le y_2 \le, \ldots, \le y_{L-1} \le y_L = b$ such that (3) is minimum. The upper bound b is the maximum value of all upper bounds of the individual study variables ($b = max(b_1, b_2, \ldots, b_p)$) and the lower bound

a is the minimum value of lower bounds of all the study variables ($a = min(a_1, a_2, \ldots, a_p)$).

For a fixed sample size n, minimizing the expression of the right hand side of (3) is equivalent to minimizing

$$\sum_{j=1}^{p}\sum_{h=1}^{L} W_{jh}\sigma_{jh} = \sum_{h=1}^{L}\left(W_{1h}\sigma_{1h} + W_{2h}\sigma_{2h} + \ldots + W_{ph}\sigma_{ph}\right) \tag{4}$$

If $f(y_j)$ for the j^{th} study variables are known or made known after various statistical techniques and if these functions are integrable, W_{jh}, σ_{jh}^2 and μ_{jh} can be obtained as a function of the boundary points y_h and y_{h-1} by using the following expressions:

$$W_{jh} = \int_{y_{h-1}}^{y_h} f(y_j)dy_j; \tag{5}$$

$$\sigma_{jh}^2 = \frac{1}{W_{jh}}\int_{y_{h-1}}^{y_h} y_j^2 f(y_j)dy_j - \mu_{jh}^2 \tag{6}$$

where

$$\mu_{jh} = \frac{1}{W_{jh}}\int_{y_{h-1}}^{y_h} y_j f(y_j)dy_j \tag{7}$$

and (y_{h-1}, y_h) are the boundaries of h^{th} stratum.

Thus, the objective function in (4) could be expressed as a function of boundary points y_h and y_{h-1} only.

Further defining $l_h = y_h - y_{h-1}; h = 1, 2, \ldots, L$ where $l_h \geq 0$ denotes the range or width of the h^{th} stratum and the range of the distribution, $d = b - a$, is expressed as a function of stratum width as:

$$\sum_{h=1}^{L} l_h = \sum_{h=1}^{L}(y_h - y_{h-1}) = b - a = y_L - y_0 = d \tag{8}$$

The h^{th} stratification point $y_h; h = 1, 2, \ldots, L$ is then expressed as $y_h = y_{h-1} + l_h$ and from (8), the problem can be treated as an equivalent problem of determining optimum strata widths (OSW), $l_1, l_2, ., l_L$. Due to the special nature of functions, the problem may be treated as a function of l_h alone and can be expressed as:

$$\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{L}\phi_h(l_h), \\ &\text{subject to} \quad \sum_{h=1}^{L} l_h = d, \\ &\text{and} \quad l_h \geq 0;\ h = 1, 2, \ldots, L. \end{aligned} \tag{9}$$

5. The Solution Procedure

The MPP (9) is a multistage decision problem in which the objective function and the constraint are separable functions of l_h, which allows us to use a dynamic programming technique [27]. Dynamic programming determines the optimum solution of a multi-variable problem by decomposing it into stages, each stage compromising a single variable sub-problem. A dynamic programming model is basically a recursive equation based on Bellman's principle of optimality [52]. This recursive equation links the different stages of the problem in a manner which guarantees that each stage's optimal feasible solution is also optimal and feasible for the entire problem (see [53]; Chapter 10).

Consider the following sub-problem of (9) for first $k(< L)$ strata:

$$\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{k} \phi_h(l_h), \\ &\text{subject to} \quad \sum_{h=1}^{k} l_h = d_k, \\ &\text{and} \qquad\quad l_h \geq 0;\ h = 1, 2, \ldots, k \end{aligned} \tag{10}$$

where $d_k < d$ is the total width available for division into k strata or the state value at stage k. Note that $d_k = d$ for $k = L$.

The transformation functions are given by

$$\begin{aligned} d_k &= l_1 + l_2 + \ldots + l_k, \\ d_{k-1} &= l_1 + l_2 + \ldots + l_{k-1} = d_k - l_k, \\ d_{k-2} &= l_1 + l_2 + \ldots + l_{k-2} = d_{k-1} - l_{k-1}, \\ &\vdots \qquad\quad \vdots \\ d_2 &= l_1 + l_2 = d_3 - l_3, \\ d_1 &= l_1 = d_2 - l_2. \end{aligned} \tag{11}$$

Let $\Phi_k(d_k)$ denote the minimum value of the objective function of (10) that is, for $h = 1, 2, \ldots, k$ and $1 \leq k \leq L$,

$$\Phi_k(d_k) = \min\left[\sum_{h=1}^{k} \phi_h(l_h) \,\middle|\, \sum_{h=1}^{k} l_h = d_k, \text{and } l_h \geq 0\right]$$

With the above definition of $\Phi_k(d_k)$, the MPP (10) is equivalent to finding $\Phi_L(d)$ recursively by finding $\Phi_k(d_k)$ for $k = 1, 2, \ldots, L$ and $0 \leq d_k \leq d$. Hence, for $l_h \geq 0; h = 1, 2, \ldots, k$,

$$\Phi_k(d_k) = \min\left[\phi_k(l_k) + \sum_{h=1}^{k-1} \phi_h(l_h) \,\middle|\, \sum_{h=1}^{k-1} l_h = d_k - l_k\right]$$

For a fixed value of l_k; $0 \leq l_k \leq d_k$, and $l_h \geq 0; h = 1, 2, (k-1)$ and $1 \leq k \leq L$.

$$\Phi_k(d_k) = \phi_k(l_k) + \min\left[\sum_{h=1}^{k-1}\phi_h(l_h) \Big| \sum_{h=1}^{k-1} l_h = d_k - l_k\right]$$

Using the Bellman's principle of optimality, a forward recursive equation of the dynamic programming technique for $k \geq 2$ and minimizing on $0 \leq l_k \leq d_k$ could be written as:

$$\Phi_k(d_k) = \min\left[\phi_k(l_k) + \Phi_{k-1}(d_k - l_k)\right] \tag{11}$$

For the first stage ($k = 1$),

$$\Phi_1(d_1) = \phi_1(d_1) \Rightarrow l_1^* = d_1 \tag{12}$$

where $l_1^* = d_1$ is the optimum width of the first stratum. The relations (11) and (12) are solved recursively for each $k = 1, 2, \ldots, L$ and $0 \leq d_k \leq d$, and $\Phi_L(d)$ is obtained. From $\Phi_L(d)$ the optimum width of L^{th} stratum, l_L^*, is obtained. From $\Phi_{L-1}(d - l_L^*)$ the optimum width of $(L-1)^{th}$ stratum, l_{L-1}^*, is obtained and so on until l_1^* is obtained.

6. A Numerical Example

In this section, an example is presented using a bivariate data set where construction of OSB is being sought for the two variables (var1 and var2), both of size, $N = 725$. The two survey variables, having different distances, are first standardized to bring about consistency in the 'compromise' distance over which the OSB would be constructed. Their distributions are estimated using ***EasyFit*** software whereby the first survey variable follows approximately Weibull distribution while the second survey variable follows approximately Gamma distribution, both of which are 3P distributions (shape(r), scale(θ), and location(γ)). Since data have been standardized, they follow these skewed distributions which are families of the Normal Gaussian distribution ($\mu = 0, \sigma = 1$). The density functions of 3P Weibull and Gamma distributions are:

$$f(x; r, \theta, \gamma) = \frac{r}{\theta}\left(\frac{x-\gamma}{\theta}\right)^{r-1} e^{-\left(\frac{x-\gamma}{\theta}\right)^r} \tag{13}$$

$$f(x; r, \theta, \gamma) = \frac{(x-\gamma)^{r-1}}{\theta^r \Gamma(r)} e^{-\left(\frac{x-\gamma}{\theta}\right)} \tag{14}$$

where $r > 0$ is a shape parameter, $\theta > 0$ is the scale parameter and γ is the location parameter.

Using (4), (5), (6) and (7) the MPP is formulated as:

$$\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{L}\left\{\sqrt{W_{1h}^2\sigma_{1h}^2}+\sqrt{W_{2h}^2\sigma_{2h}^2}\right\}, \\ &\text{subject to} \quad \sum_{h=1}^{L} l_h = d \\ &\text{and} \quad l_h \geq 0;\ h=1,2,\ldots,L. \end{aligned} \tag{15}$$

where W_{1h} and W_{2h} are the h^{th} stratum weights and σ_{1h}^2 and σ_{2h}^2 are the h^{th} stratum variances of the of the two study variables. In simplified forms, $W_{1h}^2\sigma_{1h}^2$, which deals with 3P Weibull distribution, can be written as

$$\begin{aligned} &\theta_1^2\,\Gamma\left(\frac{2}{r_1}+1\right)\left[e^{-\left(\frac{x_{h-1}-\gamma_1}{\theta_1}\right)^{r_1}}-e^{-\left(\frac{x_{h-1}+l_h-\gamma_1}{\theta_1}\right)^{r_1}}\right] \\ &\times\left[Q\left(\frac{2}{r_1}+1,\left(\frac{x_{h-1}-\gamma_1}{\theta_1}\right)^{r}\right)-Q\left(\frac{2}{r_1}+1,\left(\frac{x_{h-1}+l_h-\gamma_1}{\theta_1}\right)^{r}\right)\right] \\ &-\left\{\theta_1\,\Gamma\left(\frac{1}{r_1}+1\right)\left[Q\left(\frac{1}{r_1}+1,\left(\frac{x_{h-1}-\gamma_1}{\theta_1}\right)^{r}\right)-Q\left(\frac{1}{r_1}+1,\left(\frac{x_{h-1}+l_h-\gamma_1}{\theta_1}\right)^{r}\right)\right]\right\}^2 \end{aligned} \tag{16}$$

and $W_{2h}^2\sigma_{2h}^2$, which deals with 3P Gamma distribution, can be written as

$$\begin{aligned} &\theta_2^2 r_2(r_2+1)\left[Q\left(r_2,\frac{x_{h-1}-\gamma_2}{\theta_2}\right)-Q\left(r_2,\frac{x_{h-1}+l_h-\gamma_2}{\theta_2}\right)\right] \\ &\times\left[Q\left(r_2+2,\frac{x_{h-1}-\gamma_2}{\theta_2}\right)-Q\left(r_2+2,\frac{x_{h-1}+l_h-\gamma_2}{\theta_2}\right)\right] \\ &-\left\{\theta_2 r_2\left[Q\left(r_2+1,\frac{x_{h-1}-\gamma_2}{\theta_2}\right)-Q\left(r_2+1,\frac{x_{h-1}+l_h-\gamma_2}{\theta_2}\right)\right]\right\}^2. \end{aligned} \tag{17}$$

The three parameters, minimum and maximum values for the two variables that follow Weibull and Gamma distributions respectively are given below.

	Weibull	Gamma
Shape (r)	6.1778	5.1196
Scale (θ)	5.7282	0.4509
Location (γ)	-5.3264	-2.3086
Minimum	-3.6734	-1.7763
Maximum	2.7939	3.7360

Substituting Equations (16) and (17) into MPP (15) and solving (via a computer program) it by using the DP technique over compromise distance *d* = max(var1, var2) - min(var1, var2) = 3.7360 - -3.6734 = 7.4094

with the initial value of $x_0 = 0$ gives the standardized OSB. To get the OSB for individual variables, it is first shifted appropriately (adding the minimum of that variable) since the initial value is 0. The OSB are then further transformed into real OSB by un-standardizing, which is, multiplying with the standard deviation and adding the mean. For the first survey variable, the OSB and the Objective function values (Variances) for up to 7 strata ($L = 1, 2, \ldots, 7$) are given in Table 1 below while the results for the second survey variable are given in Table 2.

Table 1. OSB and Variance for Variable 1

L	OSB $y_h^* = y_{h-1}^* + l_h^*$	Variance $\sum_{h=1}^{L} \phi_h(l_h)$
2	y_1^* = 7.6068	2.2880e-05
3	y_1^* = 7.0897 y_2^* = 8.3118	1.6682e-05
4	y_1^* = 6.8463 y_2^* = 7.6795 y_3^* = 8.7985	1.3067e-05
5	y_1^* = 6.7027 y_2^* = 7.3500 y_3^* = 8.1079 y_4^* = 9.1985	1.0749e-05
6	y_1^* = 6.6073 y_2^* = 7.1415 y_3^* = 7.7337 y_4^* = 8.4544 y_5^* = 9.5627	9.1441e-06
7	y_1^* = 6.5380 y_2^* = 6.9936 y_3^* = 7.4845 y_4^* = 8.0425 y_5^* = 8.7479 y_6^* = 9.9049	7.9503e-06

Table 2. OSB and Variance for Variable 2

L	OSB $y_h^* = y_{h-1}^* + l_h^*$	Variance $\sum_{h=1}^{L} \phi_h(l_h)$
2	y_1^* = 6.9342	7.0004e-02
	y_1^* = 5.0693	6.7069e-02

3	y_2^* = 9.4770	
4	y_1^* = 4.1913 y_2^* = 7.1963 y_3^* = 11.2323	6.1923e-02
5	y_1^* = 3.6733 y_2^* = 6.0079 y_3^* = 8.7413 y_4^* = 12.6748	5.6615e-02
6	y_1^* = 3.3294 y_2^* = 5.2559 y_3^* = 7.3919 y_4^* = 9.9913 y_5^* = 13.9884	5.1927e-02
7	y_1^* = 3.0793 y_2^* = 4.7227 y_3^* = 6.4931 y_4^* = 8.5057 y_5^* = 11.0499 y_6^* = 15.2225	4.7853e-02

In summary, the proposed method and its application to simulated data appear to work fine and the method is able to determine OSB that are quite efficient. Compromise stratum boundaries are first determined which are then used to determine the OSB for the individual study variables. With all the assumptions met, this method would work with as many main study variables that one wants to stratify in a survey.

7. Optimum Sample Sizes

The optimum sample sizes for individual strata can easily be computed once the OSB (y_h, y_{h-1}) have been determined via the proposed method. These sample sizes ($n_h; h = 1, 2, \ldots, L$) are obtained for a fixed total sample of size n under Neyman allocation given in Equation (2) where W_{jh} and σ_{jh} are the stratum weight and variance for the p main study variables and they are derived using Equations (5) - (7).

It is worth noting that the OSB, (y_h, y_{h-1}), are so obtained from the MPP (15) that n_h must satisfy the restriction of $1 \le n_h \le N_h$, where $N_h = NW_h$. The restriction $1 \le n_h$ is added to the formulation so that the h^{th} stratum must form with at least a unit and the restriction $n_h \le N_h$ is added to avoid over sampling.

The computed OSB, presented in Tables 1 and 2 are used to calculate the stratum sample sizes (n_h) for

$h = 1, 2, ., L$ with the total sample size arbitrarily fixed at n=200. For the two main variables under study, these are presented in Table 3.

Table 3. Sample Sizes for Both Variables

L	Variable 1	Variable 2
2	n_1 = 115	n_1 = 71
	n_2 = 85	n_2 = 129
3	n_1 = 75	n_1 = 49
	n_2 = 82	n_2 = 44
	n_3 = 43	n_3 = 107
4	n_1 = 55	n_1 = 38
	n_2 = 61	n_2 = 34
	n_3 = 61	n_3 = 35
	n_4 = 23	n_4 = 93
5	n_1 = 44	n_1 = 31
	n_2 = 48	n_2 = 28
	n_3 = 50	n_3 = 27
	n_4 = 47	n_4 = 32
	n_5 = 11	n_5 = 82
6	n_1 = 37	n_1 = 27
	n_2 = 40	n_2 = 25
	n_3 = 41	n_3 = 23
	n_4 = 42	n_4 = 23
	n_5 = 35	n_5 = 30
	n_6 = 5	n_6 = 72
7	n_1 = 31	n_1 = 24
	n_2 = 34	n_2 = 22
	n_3 = 35	n_3 = 20
	n_4 = 36	n_4 = 20
	n_5 = 35	n_5 = 22
	n_6 = 27	n_6 = 30
	n_7 = 2	n_7 = 62

8. Comparison with Other Methods

This section presents the comparison of the OSB and performance (variances) of the proposed method with other established methods. The following three univariate methods have been consistently used in literature and have been considered for comparison purposes:

1) Cum $\sqrt{f}$ method of Dalenius and Hodges [8].
2) Geometric method of Gunning and Horgan [18].
3) Lavallee and Hidiroglou [14] method with Kozak's [19] algorithm.

The 'stratification' package developed by [54] in the R statistical software is utilized to determine the OSB for the main study variables for the three methods above. The OSBs are then used to compute the variances of the estimated mean so that a comparative analysis could be carried out between the three established methods and the proposed method. Firstly, the OSBs are presented for the three methods in

Tables 4 and 5 and then secondly, the comparison of variances are given in Tables 6 and 7 for the two variables respectively.

Table 4. OSB for Variable 1 Using Other Methods

L	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	10.15	12.15	12.35
3	8.57	11.28	11.55
	12.03	13.23	12.75
4	7.87	10.64	11.35
	10.15	12.15	12.35
	13.1	13.66	13.05
5	7.48	10.2	9.25
	9.17	11.72	11.95
	11.24	12.8	12.75
	13.78	13.88	13.55
6	7.23	9.77	9.35
	8.57	11.07	12.05
	10.15	12.15	12.65
	12.03	13.01	13.05
	14.26	14.09	13.55
7	7.06	9.56	9.15
	8.16	10.85	11.75
	9.44	11.72	12.25
	10.92	12.58	12.75
	12.63	13.44	13.25
	14.61	14.52	15.9

Table 5. OSB for Variable 2 Using Other Methods

L	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	7.21	12.79	11.35
3	4.27	9.47	7.75
	12.18	16.77	11.65
4	3.29	8.14	5.65
	7.21	12.79	8.75
	15.82	18.76	11.45
5	2.81	6.81	5.45
	5.27	10.8	7.95
	9.88	14.78	9.9
	18.51	20.09	11.55
6	2.53	6.15	3.95
	4.27	9.47	5.75
	7.21	12.79	7.95
	12.18	16.77	9.9
	20.56	21.42	11.35
7	2.35	5.48	3.95
	3.68	8.8	5.65
	5.76	11.46	8.25
	9.03	14.78	9.9
	14.14	18.1	11.45
	22.15	22.08	12.85

Table 6. Comparison of Variances for Variable 1

L	Proposed DP	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	2.2879e-05	3.2725e-05	3.3216e-05	3.3223e-05
3	1.6681e-05	2.9255e-05	3.3144e-05	3.3176e-05
4	1.3066e-05	2.4816e-05	3.2985e-05	3.3153e-05
5	1.0748e-05	2.1086e-05	3.2755e-05	3.1490e-05
6	9.1432e-06	1.8164e-05	3.2351e-05	3.1703e-05
7	7.9495e-06	1.5909e-05	3.2056e-05	3.1251e-05

Table 7. Comparison of Variances for Variable 2

L	Proposed DP	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	0.070004	0.070096	0.070286	0.070285
3	0.067069	0.062613	0.070280	0.070200
4	0.061923	0.052260	0.070239	0.068694
5	0.056614	0.044952	0.069950	0.068236
6	0.051926	0.039040	0.069455	0.059479
7	0.047852	0.034224	0.068319	0.059450

The comparison of the relative efficiencies of the proposed method over the other methods are given in Tables 8 and 9 for the two variables respectively. The relative efficiency for i^{th} study variable is given by Equation (18) below where 'Other Method' indicates one of the other three methods being compared against.

$$RE_i = \frac{V_{i_{\text{ProposedMethod}}}}{V_{i_{\text{OtherMethod}}}} \times 100\% \tag{18}$$

9. Discussion

It is seen that the OSB can be constructed in the multivariate situation in a very efficient manner whereby the proposed method can be applied to as many survey variables as possible. Stratified random sampling

technique is used in this research to estimate the population parameter since it is an efficient and widely used sampling technique. Often, the two major difficulties surveyors encounter prior to drawing the sample while using stratified sampling are: (i) constructing the optimum stratum boundaries (OSB) within which the units are as homogeneous as much as possible and (ii) determination of the optimum size of the sample to be drawn from each stratum. Both the problems have been addressed by this paper for the multivariate situation. The OSB obtained by the proposed method could be used to compute the optimum sample sizes for each stratum so that the precision of the estimates of parameters of the study variables are maximized.

Table 8. Efficiencies for Variable 1

L	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	143.04	145.18	145.21
3	175.38	198.70	198.89
4	189.93	252.44	253.73
5	196.19	304.75	292.99
6	198.66	353.82	346.73
7	200.13	403.25	393.12

Table 9. Efficiencies for Variable 2

L	Geometric	Cum $\sqrt{f}$	L-H Kozak
2	100.13	100.40	100.40
3	93.36	104.79	104.67
4	84.40	113.43	110.93
5	79.40	123.56	120.53
6	75.18	133.76	114.55
7	71.52	142.77	124.24

Table 10: Average Relative Efficiencies

L	Geometric	Cum $\sqrt{f}$	Kozak
2	121.59	122.79	122.81
3	134.37	151.75	151.78
4	137.17	182.94	182.33
5	137.80	214.16	206.76
6	136.92	243.79	230.64
7	135.83	273.01	258.68

Using the proposed method, Tables 1 and 2 present the OSB and the objective function values (variances) for the two variables while Table 3 presents their respective sample sizes for $L = 1, 2, \ldots, 7$. The results reveal that with increasing number of strata (L), the variances decrease in an exponential manner. Comparing the OSB from the proposed method with the other three methods in Tables 4 and 5, it is seen that they are quite different from each other. A comparison of variances is carried out and presented in Tables 6 and 7 for both the variables. The variances of the estimate given by the proposed method for variable 1 are lower than the variances for all other established methods. The same could be said for variable 2 except that the variances under the proposed method are slightly greater than the Geometric method.

The relative efficiencies of the proposed method over other methods are presented in Tables 8 and 9 where one can see the substantial gains by the proposed method over all other methods in both variables except over Geometric method in variable 2. This is expected under the proposed method which basically works on the idea of compromise stratum boundaries whereby there could be gains in one variable and loss in another. However, when the two variables are combined, there would be an overall gain achieved by the proposed method in comparison with all other methods. Table 10 presents the average relative efficiencies (A.R.E.) of the proposed method over other methods, calculated by Equation (19) where $R.E._1$ is the relative efficiency for variable 1 while $R.E._2$ is the relative efficiency for variable 2.

$$A.R.E. = \frac{R.E._1 + R.E._2}{2} \tag{19}$$

From Table 10, it is noticeable that there are substantial gains in average relative efficiencies for the two study variables. For $L = 2,3,\ldots,7$, the gains over Geometric method on average range from about 122% to 138%, the gains over Cum $\sqrt{f}$ method range from about 123% to 273% while the gains over L-H Kozak's method range from about 123% to 259%.

10. Conclusion

In this paper, a scheme is proposed to construct the OSB for multiple survey or study variables. A numerical example using a simulated data set is presented to illustrate the computational details of the application of the proposed technique for two variables. Using the estimated frequency distributions of the standardized variables, the problem of creating optimum stratum boundaries is formulated into an MPP which results in a multi-stage decision problem to be solved on a compromise distance. The brute-force algorithm of the Dynamic Programming technique is implemented into a computer program to solve the MPP, which aims to minimize the total variance of all the study variables.

It is found out that the construction of strata for multiple survey variables, when the frequency distributions are known, is possible and it leads to substantial gains in average relative efficiencies, and hence, gains in the precision of the estimates. The advantage of the proposed method is that it does not require any initial approximate solution and it can be applied to any skewed population with whatever range. The proposed method is able to determine OSB simultaneously for multiple study variables with substantially improved average relative efficiency.

The optimum stratification based on the survey variables is not feasible in practice since they are unknown prior to conducting the survey. Thus, the proposed technique is useful in the sense that it does not require the data but requires the estimated parameters of the distributions for the multiple survey variables, which can be estimated from prior surveys. The proposed method obtains global optimum stratum boundaries and is slow in terms of computing efficiency but with improved computer processing power, it will surely be a thing of the past.

References

[1] Dalenius, T. (1950). The problem of optimum stratification-II. *Skand. Aktuartidskr*, *33*, 203-213.

[2] Dalenius, T., & Gurney, M. (1951). The problem of optimum stratification. *Scandinavian Actuarial Journal*, (1-2), 133-148.

[3] Mahalanobis, P. C. (1952). Some aspects of the design of sample surveys. *Sankhya*, *12*, 1-7.

[4] Hansen, M. H., & Hurwitz, W. N. (1953). On the theory of sampling from finite population. *Ann. Math. Statist*, *14*, 333-362.

[5] Sethi, V. K. (1963). A note on optimum stratification of population for estimating the population mean. *Aust. J. Statist.*, *5*, 20-33.

[6] Aoyama, H. (1954). A study of stratified random sampling. *Ann. Inst. Stat. Math.*, *6*, 1-36.

[7] Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Ann. Inst. Stat. Math.*, *30*, 1131-1134.

[8] Dalenius, T., & Hodges, J. L. (1959). Minimum variance stratification. *J. Amer. Statist. Assoc.*, *54*, 88-101.

[9] Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.*, *38*, 345-358.

[10] Serfling, R. J. (1968). Approximately optimum stratification. *Journal of Americal Statistical Association*, *63*, 1298-1309.

[11] Singh, R., & Sukhatme, B. V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*, *21(3)*, 515-528.
[12] Bühler, W., & Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, *22*, 161-175.
[13] Unnithan, V. K. G. (1978). The minimum variance boundry points of stratification. *Sankhya*, *40*, 60-72.
[14] Lavallée, P., & Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey* Methodology, *14*, 33-43.
[15] Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, *16(1)*, 15-29.
[16] Rivest, L. P. (2002). A generalization of Lavallée and Hidiroglou algorithm for stratification in business survey. *Survey Methodology*, *28*, 191-198.
[17] Lednicki, B., & Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, *6*, 287-306.
[18] Gunning, P., & Horgan, J. M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, *30(2)*, 159-166.
[19] Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, *6(5)*, 797-806.
[20] Horgan, J. M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, *74(1)*, 67-76.
[21] Kozak, M., & Verma, M. R. (2006). Geometric versus optimisation approach to stratification: A comparison of efficiency. *Survey Methodology*, *32(2)*, 157-163.
[22] Kozak, M., Verma, M. R., & Zieliński, A. (2007). Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition-New Series*, *8(2)*, 223-248.
[23] Keskintürk, T., & Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics and Data Analysis*, *52(1)*, 53-67.
[24] Khan, E. A., Khan, M. G. M., & Ahsan, M. J. (2002). Optimum stratification: A mathematical programming approach. *Culcutta Statistical Association Bulletin*, *52*, 205-208.
[25] Khan, M. G. M., Khan, E. A., & Ahsan, M. J. (2003). An optimal multivariate stratified sampling design using dynamic programming. *Australian and New Zealand Journal of Statistics*, *45(1)*, 107-113.
[26] Khan, M. G. M., Najmussehar, & Ahsan, M. J. (2005). Optimum stratification for exponential study variable under neyman allocation. *Journal of Indian Society of Agricultural Statistics*, *59(2)*, 146-150.
[27] Khan, M. G. M., Nand, N., & Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, *34(2)*, 205-214.
[28] Nand, N., & Khan, M. G. M. (2009). Optimum stratification for Cauchy and power type study variables. *Journal of Applied Statistical Science*, *16(4)*, 453-462.
[29] Khan, M. G. M., Ahmad, N., & Khan, S. (2009). Determining the optimum stratum boundaries using mathematical programming. *Journal of Mathematical Modelling and Algorithm*, *8(4)*, 409-423.
[30] Hagood, M. J., & Bernert, E. H. (1945). Component indexes as a basis for stratification in sampling. *Journal of the American Statistical Association*, *40*, 330-341.
[31] Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, 1409-1422.
[32] Anderson, D. W. (1976), Gains from multivariate stratification. Ph.D. thesis, Department of Biostatistics, The University of Michigan
[33] Ghosh, S. P. (1963). Optimum stratification with two characters. *Ann. Math. Statist.*, *34*, 866-72.
[34] Sadasivan, G., & Aggarwal, R. (1978). Optimum points of stratification in bivariate populations.

Sankhya, 41C, 92-96.
[35] Samanta, M. (1965). A note on the problem of optimum stratification of a bivariate population in stratified random sampling, *Annals of Institute of Statistical Mathematics*, *17*, 363—375.
[36] Schneeberger, H. (1970). Optimierung in der stichprobentheorie durch schichtung. *Statistische Hefte*, *11(4)*, 242—253.
[37] Golder, P. A., & Yeomans, K. A. (1973). The use of cluster analysis for stratification. *Appl. Statist.*, *22*, 213-219.
[38] Dahmstrom, P., & Hagnell, M. (1974). The formation of strata using cluster analysis. *Statist. Tidskr*, *12*, 477-486.
[39] Thomsen, I. (1977). On the effect of stratification when two stratifying variables are used. *JASA*, *72*, 149-153.
[40] Iachan, R. (1985). Optimum stratum boundaries for shellfish surveys. *Biometrics*, 1053-1062.
[41] Yeomans, K. A., & Golder, P. A. (1975). Further observations on the stratification of Birmingham wards by clustering: A riposte. *Applied Statistics*, *24*, 345-346.
[42] Heeler, R. M., & Day, G. S. (1975). A supplementary note on the use of cluster analysis for stratification. *Applied Statistics*, *24*, 342-344.
[43] Mulvey, J. M. (1983). Multivariate stratified sampling by optimization. *Management Science*, *29*, 715—724.
[44] Schneeberger, H., & Pollot, J. P. (1985). Optimum stratification with two variates. *Statist. Hefte*, *26*, 97-113.
[45] Rizvi, S. E. H., Gupta, J. P., & Singh, R. (2000). Approximately optimum stratification for two study variables using auxiliary information. *Journal of Indian Society of Agricultural Statistics*, *53(3)*, 287-298.
[46] Rizvi, S. E. H., Gupta, J. P., & Bhargava, M. (2002). Optimum stratification based on auxiliary variable for compromise allocation.
[47] Briggs, J., & Duoba, V. (2000). STRAT2D: Optimal bivariate stratification system. Statistics New Zealand.
[48] Kish, L., & Anderson, D. W. (1978). Multivariate and multipurpose stratification, *73*, 24-34.
[49] Sitter, R. R., & Skinner, C. J. (1994). Multi-way stratification by linear programming. Survey Methodology, *20*, 65-73.
[50] Lu, W., & Sitter, R. R. (2002). Multi-way stratification by linear programming made practical. *Survey Methodology*, *28*, 199-207.
[51] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*, 558-625.
[52] Bellman, R. E. (1957). *Dynamic Programming*. Princetown University Press, New Jersey.
[53] Taha, H. A. (2007), *Operations Research: An Introduction*, 8th edition, Pearson Education, Inc., New Jersey.
[54] Baillargeon, S., & Rivest, L. P. (2011). The construction of stratified designs in R with the package stratification. *Survey Methodology, 37(1)*, 53-65.

Karuna G. Reddy is a PhD candidate at the Faculty of Science and Technology at the University of South Pacific, Laucala Campus, Suva, Fiji. Mr. Reddy obtained his BSc (2005), PGDMA (2009) and MSc (2011) from the University of South Pacific, Suva, Fiji.

Mohammed M. G. M. Khan works as an associate professor of mathematics/statistics in the School of Computing, Information and Mathematical Sciences at the University of South Pacific, Laucala Campus, Suva, Fiji. Associte Professor Dr. Khan obtained his BSc (1985) from University of Calcutta, in Calcutta, India and his MSc (1989), MPhil (1992) and PhD (1996) from Aligarh Muslim University in Aligarh, India.

Dinesh K. Rao works as an assistant lecturer in mathematics/statistics in the School of Computing, Information and Mathematical Sciences at the University of South Pacific, Laucala Campus, Suva, Fiji. Mr. Rao obtained his BEd, PGDMA and MSc from the University of South Pacific, Suva, Fiji.