International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 04 | July-2015 www.irjet.net p-ISSN: 2395-0072

Approaching an Optimal Sample Using Variable Neighborhood Search Algorithm

Priyaranjan Dash

[1] *Department of Statistics, Tripura University, India.*

Abstract - ***In almost all random sampling schemes, we adopt different sampling designs with an objective of obtaining a better representative sample (optimal sample) for the population. Application of different randomization techniques were adopted for providing a supportive basis for this. Now the question arises, whether the final sample selected, on which all our efforts are utilized, from the population is an optimal sample or not? No where we are checking about the optimality of this sample, i.e., whether this sample is the best one or there exists any other sample which is more optimal than the selected one satisfying all the constraints. In all these procedures, we only assume but, nowhere we are establishing a guarantee about the achievement of such a representative sample. The present paper emphasizes on achieving an optimal sample by using variable neighborhood search (VNS) technique.***

Key Words: *Optimal Sample, Variable Neighborhood Search, Metaheuristics.*

AMS Subject Classification: 62D05

1. INTRODUCTION

In any sample survey, we first develop a frame, which emphasizes on specifying the sampled population identical with the target population lacking any kind of ambiguity there on. A sample plays a role of centripetal force in sampling theory literature. An optimum sample is always desirable and fetches attention at all phases because a poor sample ruins the entire effort of the survey whatever attention may be put to other aspects. We put our entire effort in sampling theory to develop methods of sample selection i.e. to get an optimum sample and to draw inferences on the principles of specified precision and minimum cost. In this connection, two rivalry methods of selection of a sample came into existence: (1) random selection and the other one is (2) purposive (non-random) selection. Jensen (1926), Gini and Galvani (1929) advocated about these methods of selection. But, all of these based on the hope that the sample we get is a representative one. Since, our desire lies on getting an optimum sample (as a proper subset of the target population) whose characteristics $\hat{\Phi}_y = \tilde{y}(y_1, y_2, \cdots \cdots, y_n)$ under study are almost similar with the population characteristic $\Phi_y = Y(y_1, y_2, \cdots \cdots, y_N)$, when we have a sample of size n from the population of size N to infer about the variable y. Unfortunately, an optimum sample does not exist and even if it exists, it is very difficult, even not possible to identify it. In this regard Godambe(1955), Hege (1965), Hanurav (1966) had given significant contributions.
The above idea encourages to design a sampling scheme, which will guide us at each step of selection of the units for moving towards optimality. However, we have to keep in view about the cost incurred for selecting the sample.

2. AN OVERVIEW OF VNS ALGORITHM

Variable neighborhood search (VNS) is a metaheuristics best suitable for high dimensional optimization problems involving real life situations outperforming other recent local search methods. It is based on the three basic components: generation, improvement and shaking. In the generation or initialization component, we start with an initial point. In improvement component we step towards an improved solution by utilizing the available information and at the last, an evolutionary algorithm specialized in providing local diversity as shaking component. Mladenovic and Hansen (1997) used the variable neighborhood approach for solving the vehicle routing problems. Variable neighborhood search exploits the idea of systematic change of neighborhood within a possibly randomized local search algorithm yields a simple and effective metaheuristic for combinatorial and global optimization (Hansen and Mladenovic (1999,2001). Contrary to the other metaheuristics based on local search methods, VNS does not follow a trajectory but explores increasingly distant neighborhoods of the current solution, and jumps from this solution to a new one, if and only if an improvement has been made. In this way, favorable characteristics of the current solution (e.g., many variables are already at their optimal value), will often be kept and used to obtain promising neighboring solutions. Moreover, a local

search routine is applied repeatedly to get from these neighboring solutions to local optima. A basic VNS algorithm is stated as follows:

Generation: Choose an initial solution x and a set of neighborhood structure $\mathcal{N}_k, k = 1,2, \dots \dots, k_{max}$ to be used in the search and a *stopping condition*.

Improvement: Repeat the following steps until a stopping condition is met:

1) Set $k \leftarrow 1$.
2) Repeat the following steps until $k = k_{max}$.
   a. ***Shaking:*** Generate a point x' at random from the k −th neighborhood of x $(x' \in \mathcal{N}_k(x))$. ($x'$ is chosen at random to avoid cycling);
   b. ***Local Search:*** Apply some local search method with x' as initial solution; denote with x'' the so obtained local optimum;
   c. ***Move or Not:*** If this local optimum is better than the incumbent, move there $(x \leftarrow x'')$ and continue the search with $\mathcal{N}_1(k \leftarrow 1)$; otherwise set $k \leftarrow k + 1$.

3. AN OPTIMAL SAMPLE USING VNS ALGORITHM

Let x be the auxiliary variable closely related to the study variable y. Define the corresponding parametric function of interest for x as $\Phi_x = X(x_1, x_2, \cdots \cdots, x_N)$. An initial sample of size n selected using the design $\mathcal{P}$ from N units gives the sample observation vector for auxiliary variable only (need not observe the study variable) $\alpha^{(0)} = \{\alpha_1, \alpha_2, \cdots \cdots, \alpha_n\}$.

Another sample of size k $(\leq n)$ also has been selected randomly from remaining $(N - n)$ population units gives the sample observation vector $\beta = \{\beta_1, \beta_2, \cdots, \beta_k\}$ with the relation

$$\beta^{(j)} = \beta^{(j-1)} - \{\beta_j\}; \beta^{(0)} = \beta, j = 1,2, \cdots \cdots, k.$$

We repeat the following steps until a stopping condition is met.

Step 1: Start with $j = 1$. Select the j^{th} unit from $\beta^{(j-1)}$.

Step 2: Choose the initial sample $\alpha^{(0)}$ and calculate the value of $\hat{\Phi}_x$ and the corresponding value of the objective function $V(\alpha^{(0)}) = \|\hat{\Phi}_x - \Phi_x\|$ (say).

Step 3: We generate set of n samples, each of size $(n - 1)$, as $\alpha_{(i)}^{(0)} = \alpha^{(0)} - \{\alpha_i\}, i = 1,2, \cdots \cdots, n.$

Step 4: The neighborhood of $\alpha^{(0)}$ is constructed as

$$\mathcal{N}(\alpha^{(0)}) = \left\{\alpha_{(i)}^{(0)}, \beta_j : i = 1,2, \cdots \cdots, n\right\}$$

Step 5: Calculate the value of the objective function for all points in the neighborhood and find its minimum as $\alpha^{(1)\prime} = argmin\left\{V(\alpha^{(0)'}): \alpha^{(0)'} \in \mathcal{N}(\alpha^{(0)})\right\}$.

Choose the next improved sample to be $\alpha^{(1)}$ better than $\alpha^{(0)}$, where

$$\alpha^{(1)} = \begin{cases} \alpha^{(1)\prime}, & if\ V(\alpha^{(1)\prime}) < V(\alpha^{(0)}) \\ \alpha^{(0)}, & otherwise. \end{cases}$$

Step 6: *(Moving towards optimality)* Considering the next improved sample to be $\alpha^{(1)}$, we replace $\alpha^{(0)}$ by $\alpha^{(1)}$ and j by $j + 1$ and start again from **Step 1**.

Step 7: *(Stopping Condition)* Continue repeating the above steps until all k units are examined one by one or $\beta^{(k)} = \phi$.

4. EMPIRICAL ILLUSTRATION

The following table (Cochran (2011), p. 151-152.) gives the number of inhabitants (in 1000's) of 15 cities of United States in the years 1920 and 1930.

Table 1: Sizes of 15 Large US Cities (in 1000's) in 1920 (x_i) and 1930 (y_i)

Sl.No.	1	2	3	4	5
x_i	76	138	67	29	381
y_i	80	143	67	50	464
Sl.No.	6	7	8	9	10
x_i	23	37	120	61	387
y_i	48	63	115	69	459
Sl.No.	11	12	13	14	15
x_i	93	172	78	66	60
y_i	104	183	106	86	57

In order to estimate the total number of inhabitants $Y = \sum y$ in these cities in the year 1930, we select an initial sample of 4 cities using SRSWOR scheme. Let the selected cities are 3, 6, 7 and 12. So, we have $\alpha^{(0)} = \{\alpha_1 = 67, \alpha_2 = 23, \alpha_3 = 37, \alpha_4 = 172\}$.

Again we select another sample of size 3 from the remaining $15 - 4 = 11$ cities as 1, 5 and 8. Thus, $\beta = \{\beta_1 = 76, \beta_2 = 381, \beta_3 = 120\}$

Step 1: Start with $j = 1$. Select the 1^{st} unit from β as $\beta_1 = 76$.

Step 2: Choose the initial sample $\alpha^{(0)} = \{67, 23, 37, 172\}$ and calculate the value of

$$V(\alpha^{(0)}) = \|\hat{\Phi}_x - \Phi_x\|$$

$$= (N\bar{x} - X)^2 = (1121.25 - 1788)^2 = 444555.6.$$

Step 3: We generate set of 4 samples, each of size 3 as $\alpha_{(1)}^{(0)} = \{23,37,172\}, \alpha_{(2)}^{(0)} = \{67,37,172\}$, $\alpha_{(3)}^{(0)} = \{67,23,172\}$ and $\alpha_{(4)}^{(0)} = \{67,23,37\}$.

Step 4: The neighborhood of $\alpha^{(0)}$ is constructed as

$$\mathcal{N}(\alpha^{(0)}) = \left\{\alpha_{(i)}^{(0)}, \beta_j : i = 1,2,\cdots\cdots,n\right\} = \begin{Bmatrix}\{23,37,172,76\},\{67,37,172,76\},\\ \{67,23,172,76\},\{67,23,37,76\}\end{Bmatrix}$$

Step 5: Here,

$\alpha^{(1)\prime} = argmin\{V(\alpha^{(0)\prime}): \alpha^{(0)\prime} \in \mathcal{N}(\alpha^{(0)})\} = \{67,37,172,76\}$.
and
$V(\alpha^{(1)\prime}) = 219024 < V(\alpha^{(0)})$.
So, the corresponding units {3rd; 7th; 12th; 1st} gives a better representation of the population than the initial sample.

Step 6: *(Moving towards optimality)* We replace $\alpha^{(0)}$ by $\alpha^{(1)\prime} = \{67,37,172,76\}$.
Step 7: *(Stopping Condition)* Again, proceeding in the previous manner, after two such iterations, we can get $\beta^{(3)} = \phi$ and the sample units {3rd; 7th; 12th; 1st} is the optimum sample as it has the smallest argument.
Here, we get the optimum sample as $s = \{u_3, u_7, u_{12}, u_1\}$. Now, we can only study these units for getting y values. The following table gives the values of x and y for this optimum sample.

Table 2: Sample values for x and y

Sample Units	u_1	u_3	u_7	u_{12}
y values:	80	67	63	183
x values:	76	67	37	172

If an equivalent two phase sample is selected from this population with $n + k$ units to estimate the unknown population mean of auxiliary variable X and a second phase sample of size n units out of $n + k$ units, then in the present example (with $n = 4; k = 3$), observed sample values for x are 67, 23, 37, 172, 76, 381, 120. The following table gives the estimated standard errors of different estimators and relative gain in efficiency for estimating population total (Y) in adopting proposed optimal sample to the usual (initial) sample using different estimators under SRSWOR scheme.

Table 3: S.E.s of Different Estimators and their Relative Gain in Efficiency in Estimating Y

Sampling Scheme	Different Estimators	Sample Type	Estimate of Variance / MSE	Est. of Relative Gain in Efficiency of (a) to (b)
SRSWOR	$N\bar{y}$	a	178470.4	39.91885
		b	249713.7	
Ratio Estimator	$\bar{y}_r = \bar{x}^{\bar{y}}/_{\bar{x}}$	a	11669.7	66.68025
		b	19451.18	
Regression Estimator	$\bar{y}_{reg} = N[\bar{y} + b_{yx}(\bar{X} - \bar{x})]$	a	10256.91	10.86615
		b	11371.44	
Ratio Estimator in Double Sampling	$\bar{x}'\,{}^{\bar{y}}/_{\bar{x}}$	a	80989.51	42.17238
		b	115144.7	
Regression Estimator in Double Sampling	$N[\bar{y} + b_{yx}(\bar{x}' - \bar{x})]$	a	80163.82	37.74642
		b	110422.8	

a. Optimum sample b. Traditional sample

5. CONCLUSIONS

In all traditional sample survey literature, we are emphasizing on improving the sampling design or the estimators there on by efficiently utilizing the auxiliary information but neglecting the representativeness of the selected sample. The present paper utilizes the readily available auxiliary information in order to get an improved sample, viewed by a better representation of the population, to estimate the parameters of interest. The proposed procedure provides, by sacrificing a little cost to study the auxiliary variable, a safeguard for arriving at a better representative sample employing variable neighborhood search (VNS) technique. It does not require any kind of abstract knowledge about the population values like population correlation coefficient (ρ) between y and x as in case of ratio and regression methods of estimation. The optimality of the final selected sample is established by the relative gain in efficiency to the traditional sample, on the basis of a numerical study, shown in Table 3. Therefore, the proposed VNS algorithm for selecting an optimal sample strongly advocates about its better representativeness.

REFERENCES

[1] W.G. Cochran. Sampling Techniques. Wiley India Pvt. Ltd., New Delhi, 3rd edition, 2011.
[2] D. Freedman. A remark on the difference between sampling with and without replacement. Journal of the American Statistical Association, 72(359):681, 1977.
[3] C. Gini and L. Galvani. Di una applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione (10 dicembri, 1921). Annali di Statistica, 6, 4:1{107}, 1929.
[4] V.P. Godambe. A uni_ed theory of sampling from finite populations. Journal of the Royal Statistical Society. Series B (Methodological), 17(2):269-278, 1955.
[5] P. Hansen and N. Mladenovi_c. An introduction to variable neighborhood search in: Metaheuris-tics, Advances and Trends in Local Search Paradigms for Optimization. S. Voss et al., eds, Kluwer, Dordrecht, 1999.
[6] P. Hansen and N. Mladenovi_c. Variable neighborhood search: Principles and applications. European Journal of Operational Research, 130:449-467, 2001.
[7] T.V. Hanurav. Some aspects of unified sampling theory. Sankhya, 28:175-204, 1966.
[8] V.S. Hege. Sampling designs which admit uniformly minimum variance unbiased estimators. Calcutta Statistical Association Bulletin, 14:160-162, 1965.
[9] Jensen. Report on representative method in statistics. Bulletin of the International Statistical Institute, 22, Liv.1:381-439, 1926.
[10] N. Mladenovic and P. Hansen. Variable neighborhood search. Computers & Operations Research, 24:1097-1100, 1997.

BIOGRAPHY

The author is working in Department of Statistics, Tripura University (A Central University). He has more than 15 years of teaching and research experience. and has several publications.

