Yugoslav Journal of Operations Research
33 (2023), Number 3, 499-508
DOI: https://doi.org/10.2298/YJOR221015003J

CAT SWARM OPTIMIZATION FOR THE DETERMINATION OF STRATA BOUNDARIES

Raghad M. JASIM
Department of Veterinary Public Health, College of Veterinary Medicine, University of Mosul, Mosul, Iraq
raghad.m.j81@uomosul.edu.iq

Received: October 2022 / Accepted: February 2023

Abstract: A stratified random sampling method is preferred for selecting varied populations with outliers. As opposed to plain random sampling, stratified sampling increases statistical precision by reducing estimator variance. Before reducing the estimator's variance, stratum boundary identification and data apportionment must be solved. In this study, a Neyman allocation strategy is used to address the stratum boundary determination issue in mixed populations. In addition to evaluating CSO on two groups of people, a comparison study was conducted using Kozak, GA, PSO, and Delanius and Hodge's approaches. Compared to previous algorithms, the numerical results indicate that the proposed technique can select the best-stratified boundaries for various standard populations and test functions.

Keywords: Cat swarm optimization, stratified random sampling, Neyman allocation, optimal strata boundaries.

MSC: 62B86,62D05,62F07.

1. INTRODUCTION

Sampling studies commonly partition a population into elements before allocating a sample to obtain the most useful information. This process is known as stratification. Being based on one or more attributes, it can be used for accomplishing a variety of purposes, such as ensuring the availability of information regarding various geographic areas within a country or boosting population estimation precision [1, 2, 3]. Stratification is often used for improving precision in determining how much of the sample should be drawn from each stratum after deciding on a non-uniform allocation scheme, i.e., the number of samples collected from each stratum is inversely related to the size of the stratum. Therefore, it is vital to perform an adequate stratification of the allocation system while reviewing it [4, 5, 6].

Numerous numerical and computational methodologies have been developed for stratified sampling. Some are only appropriate for severely skewed populations, whereas

others are appropriate for all. The cumulative square root of frequency approach (CUMF) was developed by Dalenius and Hodges [7], which was a straightforward and early method. For highly skewed populations, studies have suggested [8] algorithm; Gunning and Horgan's [9] geometric method, while Keskinturk and Er's [10] Genetic Algorithm (GA) method was implemented in many scenarios expanding to non-skewed arrays. Ali and Al-Kassab [11] introduced particle swarm optimization (PSO) to find optimal strata boundaries. Yıldız et al [12] introduced the sine-cosine optimization algorithm (SCO) used to solve the shape optimization of a vehicle clutch lever, and the results demonstrate the algorithm's ability to optimize automobile components in the industry. In the same year, Panagant et al [13], proposed a new surrogate-assisted shape optimization metaheuristic. To solve the shape optimization of a vehicle bracket, a seagull optimization algorithm (SOA) is used. Yıldız et al [14] applied three structural optimization methods using size, shape, and topology optimization together. In the same year, Yıldız and Mehmet [15] developed a new Hybrid Taguchi-salp swarm algorithm (HTSSA) to speed up the optimization processes of structural design problems in industry and to approach a globally optimum solution. In [16], the Chaotic Lévy flight distribution (CLFD) algorithm was proposed to address physical world engineering optimization problems that incorporate chaotic maps in the elementary Lévy flight distribution (LFD).

The purpose of this paper is to introduce a CSO method for identifying stratum boundaries. To evaluate the CSO algorithm's efficiency, we compare it to Kozak's [17] GA, Dalenius', and Hodges' [7] algorithms. The rest of this work is outlined as follows. Stratified random sampling is discussed next. We summarize the background and prior activities of the CSO in Section 3. In addition, we examine a model for ideal stratum borders developed by the CSO. Kozak's [17] GA, Dalenius', and Hodges' [7] approaches are compared in Section 4 to determine the efficiency of the proposed CSO. A discussion of births outside of Iraq is presented in Section 5. Lastly, recommendations are suggested for future research in Section 6. Finally, a Neyman allocation strategy is used in this study to address the stratum boundary determination issue in mixed populations, and the numerical results show that the proposed technique can select the best-stratified boundaries for various standard populations and test functions.

2. STRATIFIED RANDOM SAMPLING

Among the options which are equal, and proportionate, is Neyman's allocation [18, 19]. As each stratum has the same sample size, the equal allocation method is the easiest. A proportional allocation approach ensures that stratum and sample sizes are consistent. When the variations within a stratum are comparable, these two strategies are efficient and effective. In highly-deviated scenarios, the Neyman technique is implemented if the strata variances differ. The objective is to collect fewer components from strata with low internal variability and increased samples from highly-variable internal layers.

We use Neyman's allocation technique to distribute sample size, and matching costs are hypothesized across strata. This document shows each character as its value. *Y* stands for stratification variable, *N* stands for population size, *n* for sample size, *L* represents strata number, and *Nh* and *nh* are the quantity and size of stratum *h* items, respectively ($h = 1,..., L$). As for, *Yh*: it means the value of stratum h elements. Stratified sampling estimated the mean's value. A population of *N* units are subdivided into *L* groups with

*N*1, *N*2,..., *Ni*,..., and *NH* units [6]. These are termed strata. which do not interfere while the population is exhausted. Consequently,

$$N1 + N2 + N3 + \cdots . + Nh + \cdots . + NL = N. \quad (1)$$

When following stratification, which is determined based on demographic characteristics; Sample units are selected individually based on criteria within each stratum. The sample sizes are indicated by n_1, n_2,.., n_h,....., and n_L. In this case, the sample size is denoted by the letter *n*. As a result,

$$(h = 1)Ln\ h = n. \quad (2)$$

The h is the mean of the stratum and IT is denoted by the following equation:

$$\mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} \quad (3)$$

This equation represents the sample's mean acquired from the stratum.

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \quad (4)$$

The stratum *h* variance is signified as shown below:

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{hi} - \mu_h)^2 \quad (5)$$

As for stratum *h* samples' variance, labeled as

$$S_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 \quad (6)$$

Lastly, the weight of stratum *h* is represented by W_h as follows:

$$W_h = \frac{N_h}{N} \quad (7)$$

And can be alternatively acquired from the mean shown by μ as follows:

$$\mu = \sum_{h=1}^{L} W_h\, \mu_h \quad (8)$$

Another multiplication yields a stratified mean denoted by $\bar{y}_{st}$ as follows:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h\, \bar{y}_h \quad (9)$$

Further, the variance of the stratified sampling mean is given below:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} \quad (10)$$

The computation of overall sample size *n* is governed using Neyman's method as shown below:

$$n_h = n\frac{W_h\sigma_h}{\sum_{h=1}^{L} W_h\sigma_h} \quad (11)$$

Equation (11) is associated with Neyman's allocation [19]. By substituting n_h in (11) by (10), we have the following equation:

$$V_{Ney}(\bar{y}_{st}) = \frac{1}{n}(\sum_{h=1}^{L} W_h \sigma_h)^2. \quad (12)$$

3. CSO FOR OPTIMAL STRATA BOUNDARIES

To solve this problem, Dalenius and Hodges [7] established the use of Metaheuristics called "Cat Swarm Optimization" that are based on cat behavior. The first is the ability to hunt, which is present in all felines, though in the case of the domestic cat, this ability was reduced, as was the considerable curiosity about moving objects [20]. Despite this, the cat spends most of its time resting, its second behavior, but these animals maintain an elevated level of alertness even when resting. CSO, likely, has these two states, which are referred to as Tracing Mode and Seeking Mode, respectively.

A. Input Information:

The software used to determine stratum boundaries using Neyman allocation considers these parameters:

- The number of strata *L*
- *D* is the population under study, or the population function *F*(*x*) in the range [0, 1].

B. The Fitness Function:

Fitness is an essential part of the CSO model. Within the CSO population, cats are assigned fitness values, and the value they fulfilled determines where they travel within the solution space. The Neyman allocation's variance in stratified sampling, represented by Eq. (12), is the fitness value in this work, which must be reduced during the iteration process.

C. Cat Structure:

Fitness is an essential part of the CSO model. Within the CSO population, cats are assigned fitness values, and the value they fulfilled determines where they travel within the solution space. The variance of Neyman allocation in stratified sampling, represented by Equation (12), is the fitness value in this work, which must be reduced during the iteration process.

4	8	14	18	27	30

Figure 1: Cat Structure

D. Initial Population Creation Performance and quality of the algorithm:

Fitness is heavily influenced by the particle counts reflecting the datum size and how the population is formed. In addition, the individual particle should possess equal gene quantities as strata L since the final gene marks the maximum border of the last stratum. To navigate the search space more effectively, the largest diversity of particles is ideal.

E. Tracing Mode

This condition resembles a cat identifying and pursuing its prey, with the animal moving at the speed of each dimension. This technique is divided into three stages:

1. The velocities of each dimension $V_{k,d}$ are modified according to the following function

$$V_{k,d} = round(V_{k,d} + c * r(X_{best,d} - X_{k,d}) \quad (13)$$

Where, X_{best} is the position of the best cat, which has the best fitness value; while *K* is the position of the cat, *C* is a constant and *r* is a random value between 1 and 0.

2. It is checked that the speeds do not exceed the established limit, if it occurs, it is corrected, placing it at the upper limit.
3. Update the cat position, using the following function.

$$X_{k,d} = X_{k,d} + V_{k,d} \quad (14)$$

F. Seeking Mode

- The Metaheuristic regularly slips into this condition in real life, and it replicates the behavior of a cat that is attentive and searching around for its next place to move when resting. Four key elements are used in this sub-model: "Seeking Memory Pool" (SMP), which is used for defining the search size for each cat.
- The search range of the selected dimension (SR(SRD) declares the mutative relationship for the selected dimensions.
- The number of dimensions to change (CDC).
- Its position consideration (SPC) is a Boolean that indicates if the current position of the cat is a candidate point.

When a cat engages in this activity, it follows a five-step process.

1. Make J duplicates of the location of the cat, where *J* equals SMP. Let *J* = *SMP* – 1 if the value of SPC is true, indicating that it is a candidate cat.
2. Increase or reduce the proportion of the SRD at random for each copy, according to the CDC.
3. Calculate the fitness values of all candidate points together.
4. Calculate the probability of selection for each candidate point in the equation if all the fitness values are not precisely the same (15). Set this value to 1 if not otherwise specified.

$$P_i = \frac{|FS_i - FS_{max}|}{FS_{max} - FS_{min}} \quad (15)$$

5. Move the cat's position to a candidate point randomly.

G. Stopping criteria

Termination criteria might include things like the number of iterations necessary to attain the maximum and the presence of a minor improvement in the goal task. The general structure of the algorithm looks for the best answers by conducting the following steps:

(1) Determine the upper and lower limits for the solution sets.
(2) At random, generate *N* cats (solution sets) and distribute them in *M*-dimensional space, each with a random velocity value less than a pre-defined maximum velocity value.
(3) Sort the cats into searching and tracing modes at random using *MR*. *MR* is a mixing ratio that has a value between 0 and 1. If *N* is equal to 10 and *MR* is set to 0.2, for example, 8 cats will be randomly selected to go through searching mode while the other two cats will go through tracing mode.

(4) Using the domain-specific fitness function, calculate each cat's fitness value. The best cat is then selected and remembered.
(5) After then, the cats alternate between searching and tracing modes.
(6) For the following iteration, randomly allocate the cats into searching or tracing modes based on *MR* after they have gone through seeking or tracing mode.
(7) Check the termination condition; if it is met, the program should be terminated; otherwise, repeat Steps 4–6.

4. NUMERICAL RESULTS

We conducted stratification sampling experiments on two populations: data and functions, to optimize strata boundaries based on Neyman allocation variance. MATLAB 9.0.0 (R2016a) was used for all analyses. Testing the CSO algorithm is to achieve the stratified boundaries of data.

A. Testing PSO algorithm looking for the stratified boundaries of data:

We stratify a variety of populations with variable skewness, kurtosis, mean, standard deviation, and size attributes. The stratification is performed using populations from the R stratification [21]; GA4 Stratification packages [22]. Each population is divided into three, four, five, and six strata. We selected 100 samples, and then we used Kozak's, GA An PSO techniques to select the bounds with random initial boundaries.

Pop 1: Population of Irish company debtors.

Pop 2: The employee count in 284 Swedish municipalities was surveyed in 1984 (ME84).

Pop3: (MRTS) uses data from the Statistics of Monthly Retail Trade Survey, in Canada.

Pop4: Pop4 was the name given to the population of thousands of people in 284 Swedish municipalities in 1975. (P75).

Pop5: Values of real estate in million kronor, based on a survey of 284 Swedish towns in 1984 (REV84).

Pop6: Commercial US banks have resources worth millions of dollars (US banks).

Pop7: 1940 US cities' population of (US cities).

Pop8: In 1952-1953, the total of four-year college students in the United States was estimated to be over 8 million. (US colleges). The variance of the estimate provided by Eq. 12 is used to compare the efficiency of three techniques. To instrument our suggested technique on a PC, we used the MATLAB programming language (CPU 3.00 GHz, 3GB RAM). Table 1 lists the CSO parameter values employed to stratify. Table 2 summarises the deviation of the estimators derived using the CSO, GA, PSO, and Kozak's approaches [17].

Table 1: CSO Parameters

CSO parameters	H =2,3	H=5,6
Swarm size	100	100
Maximum iteration	100	200
C1	2	2.5
C2	2	1.5

Table 2: Variations of stratification estimators according to CSO, PSO, GA, and Kozak methods

Pop	H	CSO	PSO	GA	Kozak
Pop1	3	2467.5	2467.5	2469.5	4090.9
	4	1359.2	1359.2	1369.2	2291.7
	5	822.5	822.5	831.2	1269.5
	6	572.4	572.5	588.98	605.58
Pop2	3	6506.1	6506.1	36797	36797
	4	3115.87	3116	34787	34787
	5	2117.1	2117.1	35614	35614
	6	1555	1555.1	24207	35577
Pop3	3	591721	591721	593160	1039100
	4	310783	310783	311190	311190
	5	204832	204833	207070	207000
	6	148921	148336	150780	150750
Pop4	3	1.81689	1.81689	5.3956	5.3956
	4	0.90768	0.90768	4.9031	4.9031
	5	0.59364	0.59366	4.0269	5.4344
	6	0.4246	0.4321	3.9140	5.3821
Pop5	3	18231	18231	47733	47733
	4	9296	9296	46545	46545
	5	5590	5590	34483	137400
	6	3815	3816	31654	42403
Pop6	3	33.519	33.519	36.850	36.850
	4	17.327	17.339	27.331	27.331
	5	11.007	11.008	20.370	20.370
	6	6.7476	6.7485	18.435	18.448
Pop7	3	0.89195	0.89195	0.917173	0.917173
	4	0.47227	0.47276	0.473657	0.873657
	5	0.264201	0.264204	0.266574	0.569189
	6	0.19422	0.196972	0.199325	0.274273
Pop8	3	2451.4	2451.4	2469.7	2469.7
	4	1500.4	1500.4	1539	1539
	5	928.9	928.9	1020.9	2763.70
	6	603.1	603.2	892.40	892.33

B. CSO method is being tested to discover stratified boundaries for function populations.

Three benchmark functions are used for evaluating the suggested CSO. For example, Delanius and Hodges [7] are likewise utilize these computations. The test functions are detailed in Table 3.

Table 3: Benchmark functions (f1-f3)

Function	Range
$f_1(x) = xe^{-x}$	$0 \leq x < \infty$
$f_2(x) = e^{-x}$	$0 \leq x < \infty$
$f_3(x) = 2\,(1-x)$	$0 \leq x \leq 1$

The comparative assessment of these two approaches encompassing three benchmark computations across 4 strata is presented in Table 4.

Table 4. Results comparing 3 benchmark functions

Fun.	H	D and H $V_{Ney}(\bar{y}_{st})$	PSO $V_{Ney}(\bar{y}_{st})$	CSO $V_{Ney}(\bar{y}_{st})$
f_1	2	0.2855	0.2835	0.2835
	3	0.1339	0.1321	0.1321
	4	0.0774	0.0761	0.0761
	5	0.0503	0.0494	0.0495
f_2	2	0.6389	0.6177	0.6177
	3	0.3069	0.2964	0.2964
	4	0.1817	0.1732	0.1732
	5	0.1192	0.1133	0.1132
f_3	2	0.0152	0.015	0.2835
	3	0.0069	0.0069	0.1321
	4	0.0039	0.0039	0.0761
	5	0.0495	0.00253636	0.00253634

5. PRACTICAL APPLICATION OF BIRTHS OUTSIDE IRAQ

Iraqis living abroad are Iraqi nationals or persons of Iraqi ancestry who live in countries other than Iraq as immigrants or refugees. According to the United Nations, the situation of Iraqis living abroad is one of the world's most important humanitarian problems, which began after the US occupation of Iraq in 2003 and continues to this day. More than four million Iraqis live outside of Iraq, and according to official statistics, these people require consular services such as registering births and deaths, necessitating an increase in the number of embassies and consulates staff by the Iraqi Ministry of Foreign Affairs to improve service equations simultaneously with a dense population of Iraqis abroad.

Outside of Iraq, the birth community consists of 97 nations, with a total of 31794 births in those 97 countries. 328 and 995 are the mean and standard deviation, respectively. Table 5 shows the results of applying the CSO algorithm to discover the optimal stratum borders for dividing the population of births into homogenous strata.

Table 5: Shows optimal stratum borders for dividing the population of births into homogenous strata

H	$V_{Ney}(\bar{y}_{st})$	Boundaries	*N*
2	**1.0071e+005**	1-904	91
		904-7890	8
3	**4.9317e+004**	1-120	69
		120-904	22
		904-7890	8
4	**2.7985e+004**	1-120	69
		120-904	22
		904-2955	6
		2955-7890	3
5	**1.8860e+004**	1-92	67
		92-554	23
		554-1877	5
		1877-2955	4
		2955-7890	3
6	**1.4367e+004**	1-50	59
		50-253	21
		253-554	10
		554-1877	5
		1877-2955	4
		2955-7890	3

6. CONCLUSIONS

A stratified sampling approach provides more precision than other sampling methods for diverse populations. Several test scenarios are used in this research to evaluate the performance of a CSO method for stratified borders with Neyman allocation. Optimal strata boundaries have been determined by the CSO algorithm using numerical results. Surprisingly, it outperforms other approaches like Kozak, GA, PSO, Delanius, and Hodges. Although the PSO technique's results are similar to those of the CSO algorithm, the suggested algorithm's performance improves as the number of strata increases. CSO can be used to stratify diverse populations efficiently. In the future, CSO could be utilized in studies where the strata quantification, sample boundaries, and cost are all variable.

Funding. This research received no external funding.

REFERENCES

[1] S. Y. Al-Saffawi, *Applying Statistical Techniques Analysis and Estimating a certain Agricultural Production*, Master Thesis, University of Baghdad, Iraq, 1976.

[2] A. Al-Hasow and M.M.T. Al-Kassab, *Method to find Stratum Boundaries Using Neyman Allocation*, Master Thesis, University of Mosul, Iraq, 1996.

[3] M. M. T. Al-Kassab and H. Al-Taay, "Approximately Optimal Stratification Using Neyman Allocation", *Journal of Tanmiyat Al-Rafidain*, 1994.

[4] T. Bäck, *Evolutionary algorithms in theory and practice*, New York: Oxford Univ. Press. 1996.

[5] A. Korel, *Software Test Data Generation*, IEEE, Computer Society and Association for Computing Machinery, 1990.

[6] W. G. Cochran, *Sampling Techniques, 3rd ed*, John Wiley & Sons, Inc. USA, 1977.

[7] T. Dalenius and J. L Hodges, "Minimum Variance Stratification", *Journal of the American Statistical Association*, vol. 54, no. 285, pp. 88-101, 1959.

[8] P. Lavallée and M. Hidiroglou, "On the Stratification of Skewed Populations", *Survey Methodology*, vol. 14, no. 1, pp. 33-43, 1988.

[9] P. Gunning and J. M. Horgan, "A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations", *Survey Methodology*, vol. 30, no. 2, pp. 159-166, 2004.

[10] T. Keskintürk and Ş. Er, "A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling", *Computational Statistics &Data Analysis*, vol. 52, no. 1, pp. 53-67, 2007.

[11] A. A. Ali and M. M. AL-Kassab, "Using Particle Swarm Optimization to Determine the Optimal Strata Boundaries", *Journal of Advances in Mathematics*, vol. 11, no. 1, 3890–3901, 2015.

[12] A. B. S, Yıldız, N. Pholdee, S. Bureerat, A. R. Yıldız, and S. M. Sait, "Sine-cosine optimization algorithm for the conceptual design of automobile components", *Materials Testing*, vol. 62, no. 7, pp. 744 748, 2020.

[13] N. Panagant, N. Pholdee, S.Bureerat, K. Kaen, A. R. Yıldız, and S. M. Sait, "Seagull optimization algorithm for solving real-world design optimization problems". *Materials Testing*, vol. 62, no. 6, pp. 640-644, 2020.

[14] B. S.,Yıldız, N., Pholdee, S., Bureerat, M. U., Erdaş, A. R., Yıldız, and S. M. Sait, "Comparision of the political optimization algorithm, the Archimedes optimization algorithm, and the Levy flight algorithm for design optimization in industry", *Materials Testing*, vol. 63, no. 4, pp. 356-359, 2021.

[15] A. R. Yıldız and M. U. Erdaş, "A new Hybrid Taguchi-salp swarm optimization algorithm for the robust design of real-world engineering problems", *Materials Testing*, vol. 63, no. 2, pp. 157-162, 2021.

[16] B. S. Yıldız, S. Kumar, N. Pholdee, S. Bureerat, S. M. Sait, and A. R. Yildiz, "A new chaotic Lévy flight distribution optimization algorithm for solving constrained engineering problems", *Expert Systems*, vol. 39, no. 8, e12992, 2022.

[17] M. Kozak, "Optimal Stratification Using Random Search Method in Agricultural Surveys", *Statistics in Transition*, vol. 6, no. 5, pp.797-806, 2004.

[18] T. H. N. Daghistani, *An Approximately Optimal Stratification Using Proportional Allocation*, Master Thesis, University of Mosul, Iraq,1995.

[19] J. Neyman, "On the Two Different Aspects of the Representative Methods: The Method of Stratified Sampling and the Method of Purposive Selection", *Journal of the Royal Statistical Society*, vol. 97, no. 4, pp. 558-625, 1934.

[20] R. C. Eberhart, Y. Shi, J. Kennedy, *Swarm Intelligence*, New York: Morgan Kaufmann, 2001

[21] R: stratification. http://CRAN.R-project.org/package=stratification

[22] R: GA4Stratification.http://CRAN.R-project.org/package=GA4stratification