International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 3 Issue 4 || April. 2015 || PP-01-10

A note on the Efficiency of Geographic Stratification

Akeem O. Kareem[1], Gafar M. Oyeyemi[2] and Adijat B. Aiyelabegan[3]
[1]*Institute for Security Studies, P. M. B. 493, Abuja, Nigeria.*
[2]*Department of Statistics, University of Ilorin, Ilorin, Nigeria.*
[3]*Department of Mathematics & Statistics, Kwara State Polytechnic, Ilorin, Nigeria.*

ABSTRACT: *Often times, administrative convenience or the need for estimates for the domain of study dictates the use of Stratified Sampling using Geographic Stratification (GS). This method of strata boundary determination is ill adapted in practice due to its less amenability to mathematical approach. Despite its poor performance in terms of Precision, empirical investigationusing four sets of real life data with varying degrees of skewness shows that GS yields minimum Mean Square Error (MSE) value when compared with popular strata construction methods like cumulativesquare root of frequency method Dalenius and Hodges, DHR (1959) and Geometric Stratification of Gunning and Horgan, GMS, (2004) using optimum allocation.*

KEY WORDS: *Geographic; Stratification; Mean Square Error and Efficiency.*

I. INTRODUCTION

Stratification is a common technique employed in sample survey not only for its improved precision and provision of samples that are representative of the population, but because of administrative convenience and when estimates are required for a subpopulation. It is also very important when dealing with skewed population.

Literature adequately reports stratified sampling in terms of procedures and methods of estimation: Murthy (1967); Sukhatme and Sukhatme (1984); Mendehall et al. (1971); Cochran (1977); Raj and Chandhok (1998) and Okafor (2002). In order to reduce the amount of variations contained within the samples and obtain more precise estimates, we employ stratified sampling technique which was described by Mendehall et al. (1971) as a sampling procedure which separates the population units into non-overlapping groups called strata and thereafter select sample independently from each stratum. Okafor (2002) on his part, referred to it as the procedure of drawing independent samples after grouping the whole units in the population into homogenous distinct strata as stratified sampling. When a simple random sampling is used to select sample in each stratum, the procedure is called stratified random sampling. Thus, classes of stratified sampling derived their names from the sampling schemes employed in drawing the samples from the strata. When systematic sampling is employed, we have stratified systematic sampling. However, we shall be satisfied with the definition of Cochran (1977) which states that "In stratified sampling, the population of N units is first divided into L subpopulations of $N_1, N_2, \ldots, N_L$ units, respectively. These subpopulations are non-overlapping and together they comprise the whole of the population, such that $N_1 + N_2 + \ldots + N_L = N$. i.e. $\sum_{h=1}^{L} N_h = N$. The subpopulations are called strata. If simple random samples of sizes $n_1, n_{2,\ldots},n_L$ respectively are taken from each stratum, such that $\sum_{h=1}^{L} n_h = n$, the whole procedure is described as stratified random sampling.

Horgan (2006) stated that stratification technique is often employed majorly to maximize the precision of some estimator $\hat{\theta}$ or equivalently to minimize the Mean Square Error of $\hat{\theta}$, [MSE($\hat{\theta}$)]. Maximization of the precision dominated literature in the appraisal of the best method for strata boundary determination thus, the method yielding the least variance (but which fails to account for theassociated bias) is adjudged the best, which makes the inherent beauty of GS as a procedure yielding minimum MSE value not to be unveiled.

This study seeks to identify the most efficient method of stratification using the MSE criterionwhich incorporatesthe variance and the bias rather than the minimum variance approach commonly in use.

Murthy (1967), Cochran (1977), Raj and Chandkok (1998) and Okafor (2002) have all itemized reasons why stratification technique is commonly used to include:

i. the need for estimates with known precision for subpopulations;
ii. administrative conveniences may dictate the use of stratified sampling;
most especially when there are sub-stations in a statistical organization and such offices could be directed to collect data on a subject of interest in their domain;
iii. differences between strata may necessitate different sampling methods; and
iv. when increased precision for the whole population is required.

Dalenius and Hodges (1959), Hess et al. (1966), Wang and Aggrawal (1984), Okafor (2002) andHorgan (2006) itemized the following as specific design problems involved in stratification processes:

(a) the choice of a stratification variable;
(b) the choice of number of strata L to be formed;
(c) mode of stratification i.e. the way/manner in which strata boundaries are determined;
(d) the choice of sample size n_h to be taken from the h^{th} stratum. i.e. the problem of allocation of sample size to strata; and
(e) choice of sampling design within strata.

Solutions have also been suggested to most of these problems by different authors. On the choice of stratification variable, Cochran (1977) enjoined the use of the frequency distribution of the study variate itself as stratification variable if available or that of variable X (an auxiliary variable) which is highly correlated with the study variate and perhaps the value of variable Y at a recent census. Same view was expressed by Sukhatme & Sukhatme (1984). Hess et al. (1966) examined the significance of a highly correlated auxiliary variable as choice of stratification variable while Kish and Anderson (1977) took it to a logical conclusion in a multivariate stratification study. However in practice, the estimation variable is commonly used for the purpose of stratification, e.g., Dalenius (1959), Ghosh (1963), Hess et al. (1966) and Hedlin (2000). This study also made use of the estimation variable for stratification.

Next, is the choice of number of strata L to be formed. In most cases, it is predetermined in order to attain a specified level of precision. However, Cochran (1977) developed a model representing the approximate reduction in the precision of stratified sample mean compared to that obtained with simple random sampling and concluded that beyond six strata $L \geq 6$ there is little or no further gain, in terms of precision. This was premised on the following two basic questions Cochran (1977) said should be considered to efficiently determine the number of strata:

(a) at what rate does the variance of $V(\bar{y}_{st})$ decrease as L is increases; and
(b) how is the cost of the survey affected by an increase in L?

Thus, when there is no appreciable gain in precision with additional strata, then optimum number of strata has been reached and when the cost of sampling additional strata is already overshooting the survey budget, it is obvious that the number of strata to be surveyed should be limited to the one covered by the survey budget.

Hess et al. (1966) confirmed that efficient number (L) of strata can be arrived at by observing the $Var(L)/Var(L-1)$ reduction in variance attained when additional stratum is considered, with the remark that in many multipurpose investigations, only marginal stratification gains could be expected from the use of more than six strata.

Ghosh (1963), Hess et al. (1966), Hidiroglou (1986), Hedlin (2000), Gunning and Horgan (2004), Kozark (2004), Keskintur and Er (2007) etc, have all studied methods of strata construction using predetermined number of strata.

Okafor (2002) however stated that, in determining the optimum stratum boundaries, we are absolutely free to choose the number of strata, which is opposed to a situation in which the strata have been predetermined (previously estimated), e.g. geographically or/an administrative stratification. This study allows for optimum number of strata as suggested by Okafor (2002), while deep stratification also dictates number of strata to be considered.

The third design operations in stratified sampling, as earlier stated, is the mode of stratification, i.e. strata boundary determination, reported literature is as reflected in section 2 below.

On the problem of allocation of sample to stratum, Literature had extensively dwelt on the subject matter with the following result:Optimum allocation wasdue to Neyman (1934), proportional and equal allocations have been traditionally long in use. Compromise allocation was by Chatterjee (1968); it was used and improved upon by Khan and Ahsan (2003). Power allocation was used by Lavallee and Hidiroglou (1988) while Genetic Allocation (GA) was developed and used by Keskinturk and Er (2007).This study makes use of the optimum allocations for its empirical investigation for its highest precision and yielding minimum MSE estimate for GS when compared with other competing methods of strata construction.

On the choice of sampling design within the stratum, literature reported the use of random sampling with or without replacement. This study makes use of simple random sampling without replacement in each stratum.

II. METHODS OF STRATA BOUNDARY DETERMINATION

Available methods of strata boundaries determination in the literature include: Dalenius (1950); Equalization of strata Totals (EST) was due to Mahalanobis (1952); Dalenius and Hodge (1959) suggested the cum$\sqrt{f(y)}$ method, here in referred to as Dalenius and Hodge's Rule (DHR). Others are Ekman's Rule (1959) (EKR), Durbin's Rule (1959) (DUR); Sethis Rule (1963) (STR); Thompson Rule (1976) (TNR); Lavalle and Hidiroglou Method (1988) (LHM); Extended Ekman's Rule (EEKR) by Hedlin (2000), Random Search method (RSM) was due to Kozark (2004); Geometric Stratification (GMS) by Gunning and Horgan (2004) and Genetic Algorithm (GA) by Keskintur and Er (2007). Of all the aforementioned, DHR and GMS are popularly in use for easy application and precision.Therefore form the basis of comparison with GS in this study.

2.1 Dalenius and Hodge Rule (DHR)

DHR requires us to choose equal class interval, obtain the cumulative square root of the frequency (cum$\sqrt{f(y)}$) of the study variate and determine the strata boundaries by dividing the total cumulative square root of the frequency by the required number of strata L and the boundary is placed at this division point. It was an approximate solution by Dalenius and Hodge (1959) to Dalenius (1950) equations. For details on derivation of these sets of general equations, see Dalenius (1957, 1959), Murthy (1967, section 10.7a, pp.262), Sukhatme and Sukhatme (1970, section 3.11, pp. 108), Cochran (1977, section 5A.7, pp.127), Raj and Chandhok (1998, section 4.8, pp. 107) and Okafor (2002, section 4.6, pp. 120).

2.2 Geometric Stratification (GMS)

Gunning and Horgan (2004) introduced the new and now the most commonly used method of strata boundary determination called Geometric Stratification (GMS). It was applied to positively skewed populations and results compared with DHR. Stratum boundaries are automatically formed with GMS once the geometric ratio r is determined.

$$r = [\max \quad Y_i / \min \quad Y_i]^{1/L}$$

$$r = [Y_L / Y_0]^{1/L} \qquad (1)$$

where Y_L is the highest value and Y_o is the smallest value of the study variate Y. The boundaries are at the points:

Minimum K_0 = a, ar, ar^2, . . . ,ar^L = Maximum K_L.

The general term is:

$$K_h = ar^h, h = 0,1,2,..., L-1 \qquad (2)$$

Details of the GMS algorithm are in section 2 of Gunning and Horgan (2004). The simplicity of GMS had been extended to Pareto distribution by Gunning and Keogh (2006) and was found to be more efficient than DHR.

2.3 Geographic Stratification (GS)

Cochran (1977, pp.102) identified the problem of "less amenable to mathematical approach" to be responsible for the poor use of GS in practice. He defined GS as strata formations which are compact areas such as counties or neighbourhoods in a city. Therefore, the words of Raj and Chandhok (1998, pp.107) that personal intuitions and experiences have to come to bear if stratification is to be used, could be the only basis for employing GS since there would be so many ways in which strata boundaries may be formed. The most important thing would be to ensure units are internally homogenous within strata. It is often employed for administrative convenience or when separate estimates are required for each stratum.

Thus as the name implies and rightly pointed out by Cochran, GS lacks mathematical approach thus have no algorithm. GS is based on already delineated clear boundaries where there would be no overlapping of units, it cloud be based on the map of the area, type of crops grown, and other natural boundaries that can easilyform strata. It could be a clear administrative unit, a local government area, a county, regional arrangements, geo-political zones or states forminga nation. Jessen (1942) was credited to have first examined performance of GS while Jessen and Houseman (1944) in their empirical investigation reported a moderate performance of the GS in terms of it precision. No appreciable gain in precision was also reported by Judez and Chaya (1999, 2000) for GS. However, this study shows that GS should be employed when accurate estimates are required for data that suits its application for its minimum MSE valuewhen compared with popular stratification methods DHR and GMS respectively.Empirical comparison of strata construction methods had been on the basis of the precision of the population mean or total, however, statistical inference has suggested that most accurate estimators are come by, using the MSE criterion (see Cochran (1977) section 1.9). Hence,the use of relation (5) below as measure of appraisal of the best stratification method in this study. Cochran (1977) section 5A.2 pointed out that using weight that are in error in stratified sampling leads to sample estimate that is biased. Therefore, it is ideal to assess the performance of a procedure by the MSE rather than variance (precision).

III. ESTIMATION PROCEDURE.

This section discusses estimation procedure in stratified random sampling. Symbols and notations of Cochran (1977, pp.90) were adopted in this study.

Notations

The subscript h denotes the stratum and i the unit within the stratum.

L	=	Number of strata.
N_h	=	Total number of population units in stratum h.
n_h	=	Number of sampled units in stratum h.
N	=	Total number of population units in all the L strata
n	=	Sample size of the study
Y_{hi}	=	is the observation of the i^{th} unit in the h^{th} stratum
W_h	=	N_h/N = stratum weight (population units)
w_h	=	n_h/n = stratum weight (sample units)

Optimum allocation due to Neyman (1934) is employed in this study to distribute fixed sample sizes in to the strata.

The expression for the optimum allocation is given as;

$$n_h = \frac{nN_h S_h}{\sum N_h S_h} \tag{3}$$

With the variance as

$$V_{opt} = V_{\min}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{4}$$

$$MSE(\bar{y}_{st}) = \sum W_h^2 (1 - f_h)\frac{s_h^2}{n_h} + [\sum (w_h - W_h)\bar{Y}_h]^2$$

$$= \quad V(\bar{y}_{st}) + [Bias]^2 \tag{5}$$

See Cochran (1977) relation 5A.9 pp. 118.

It should be noted that the true stratum weight is known and applied in this study.

When optimum allocation is used,

$$MSE\ (\bar{y}_{st}) = V(\bar{y}_{st})_{opt} + [\sum (w_h - W_h)\bar{Y}_h]^2 \tag{6}$$

In stratum where optimum allocation produces n_h(stratum sample sizes) which are larger than the stratum size N_h. (i.e. when $n_h > N_h$) the revised optimum allocation is used.Cochran (1977, p.104).

Ropt = $$\tilde{n}_h = (n - N_i)\frac{N_h S_h}{\sum N_h S_h} \tag{7}$$

Where i is the stratum in which $n_h > N_h$.

e.g. if $n_1 > N_1$ then, for $h \geq 2$ $\tilde{n}_h = (n - N_1)\frac{N_h S_h}{\sum N_h S_h}$.

If more than one stratum is involved, the entire affected strata where $n_h > N_h$ are deducted from sample size n to obtain Ropt allocation using relation (7) above. Expression for the variance of Ropt allocation is given as;

$$V_{Ropt}(\bar{y}_{st}) = \frac{(\sum' W_h S_h)^2}{n'} - \frac{\sum' W_h S_h^2}{N} \tag{8}$$

where n' is the revised total sample size and $\sum'$ is the summation over the strata in which $\tilde{n}_h < N_h$.

Thus, relation (8) fits back into relation (6) to obtain $MSE\ (\bar{y}_{st})$ for strata formations where Ropt allocation is used.

IV. EMPIRICAL INVESTIGATION.

Secondary data were collected from the under listed agencies. The data structure reflects varying degree of skewness. Murthy (1967) Section 2.2c, pp. 29; Cochran (1977) Section 5.7 pp.101; had itemized types of data suitable for stratification techniques. The four (4) sets of life data whose features are reflected in Table 1 below are used for this study.

i. Overall cumulative average scores (OCAS) of 145 students that graduated from the Faculty of Engineering University of Ilorin 1989/90 set.
ii. Data of Kano State Ministry of Commerce and Industry Survey (2008) on manpower strength of companies and industries in the five (5) industrial Estates of Kano and those located outside the industrial Estates.
iii. Grants allocation to 774 Local Government's Council in Nigeria for the month of December, 2008 shared in January 2009.(See www.fmf.gov.ng)
iv. Population Census figures for the 774 Local Government Areas of Nigeria during the year 2006 census.(see **www.nigeriastat.gov**.ng)

Table 1: Summary Statistics of the data used in this study.

DATA	N	n	Range	Skewness	Mean	Variance	Standard Deviation
1	145	48	44.7 - 68.8	0.712	55.48	20.05	4.48
2	171	57	3 – 3756	6.581	166	163923	405
3	774	258	72.2 - 365.0	3.239	108.96	700.61	26.47
4	774	258	11.7 - 1277.7	3.218	180	10281	101

5.1 Strata Formation

This study allows for optimum number of strata and the stratification process were continued for DHR and GMS until when deep stratification occurred, i.e., $N_h = 1, \forall h = 1,2,..., L$ (at least one population units in one or more stratum). Fixed sample size **n** is predetermined for all the sets of data used in this study and sample estimation was restricted to strata formation in which $\mathbf{n_h} \geq 2$. Hence, the method of collapsed strata (Cochran (1977), pp.138 5A.12) is not considered.

DHR and GMS procedures described in sections 2.1 and 2.2 above were applied and number of strata formation is as reflected in Table 2. Five (5) strata formations were obtainable for data 1 and data 2 whileSix (6) strata formations for data 3 and data 4 when DHR is used. GMS on its part was estimable in six (6) strata formation for data 1, five (5) for data 3 and in Eight (8) strata formations for data 2 and 4 respectively.

GS as observed by Cochran (1977) is "less amenable to mathematical approach" hence the poor use of GS in practice. Therefore, the words of Raj and Chandhok (1998, pp.107) that personal intuitions and experiences have to come to bear if stratification is to be used, could be the only basis to apply GS besides the need for estimate for a particular sub-group or strata.

GS on its part have the following number of strata per data set. National University Commission (NUC) in Nigeria has delineated boundaries for classes of degree obtainable by university graduate as follows 70% - 100% as First Class; 60.0% - 69.9% as Second Class Upper, 50.0% - 59.9% as Second Class Lower, 45.0% - 49.9% as Third Class while 40.0 44.9% obtained Pass degree. These established boundaries by the NUC formed the geographic strata. It should be noted that this set of students recorded no first class while only a student scored pass degree. Therefore for data 1, three strata formation is obtainable using GS with stratum I for those that scored third class and Pass degree, stratum II has those with second class lower while the third stratum has those with second class upper division.

Data 2is from industrial survey conducted in Kano State Nigeria. Kano State has five industrial estates which naturally formed the geographic strata. Stratum I have industries located outside the five Industrial estates and consistently remain Stratum I for the two through six strata formations. Stratum II of two strata formation contains industries in the five industrial estates.Nearest geographic neighborhood was used to establish two through six strataformation

Data 3 and 4 were stratified geographically into two, three and six strata based on regional and geo-political arrangement of the country. Two strata comprise of Northern and Southern Nigeria, three strata formation comprises of Northern, Western and Eastern regions of post-independence Nigeria, while six strata formation are the present geo–political zones namely: Southwest, Southeast, Southsouth, Northwest, Northeast and Northcentral. It is pertinent to mention that Local Government Areas (LGA) in Federal Capital Territory (FCT) Abuja were merged into Northcentral zone which further implies that, in two and three strata formations FCT–Abuja is stratified North. Furthermore, those states created out of the old regional arrangement were stratified into their old regions in three strata formation, e.g. Edo and Delta States were stratified into Western region, while Cross-River and Akwa-Ibom States were stratified into Eastern region. Stratification boundary is clear in two strata situations in terms of North and South with river Niger and Benue separating the North from the South.Distribution of population units into stratum by DHR, GMS and GS is as shown in Table 2 below.

Table 2: Distribution of Population Units into Stratum by Number of Strata for Data1 to 4.

Number of Strata	Stratum	DATA 1			DATA2			DATA3			DATA 4		
		DHR	**GMS**	**GS**	**DHR**	**GMS**	**GS**	**DHR**	**GMS**	**GS**	**DHR**	**GMS**	**GS**
2	1	79	82		143	117	16	516	744	419	576	205	419
	2	66	63		28	54	115	258	30	355	198	569	355
3	1	79	18	11	116	48	16	516	642	419	398	12	419
	2	43	102	109	38	108	43	157	122	180	266	662	180
	3	23	25	25	17	15	112	101	10	175	110	100	175
4	1	11	11		116	16	16	118	491		148	5	
	2	68	71		27	101	43	398	253		428	200	
	3	43	50		20	46	83	211	24		142	543	
	4	23	13		8	8	29	47	6		56	26	
5	1	11	8		116	9	16	118	324		148	3	
	2	68	37		27	68	29	398	382		250	44	
	3	43	66		11	67	43	157	54		178	478	
	4	16	25		11	21	49	80	13		162	235	
	5	7	9		6	6	34	21	1		36	14	
6	1		5			9	16	118		137	148	3	137
	2		13			39	21	398		95	250	9	95
	3		64			69	34	157		123	178	193	123
	4		38			39	43	54		186	88	469	186
	5		18			11	49	26		112	87	92	112
	6		7			4	8	21		121	23	8	121
7	1		2			4						3	
	2		11			26						5	
	3		46			61						65	
	4		44			48						365	
	5		26			22						293	
	6		9			7						39	
	7		7			3						4	
8	1		2			4						2	
	2		9			12						3	
	3		21			46						24	
	4		50			55						176	
	5		33			32						392	
	6		17			14						151	
	7		7			6						24	
	8		6			2						2	
9	1		2			4						2	
	2		8			11						2	
	3		8			33						8	
	4		46			44						78	
	5		35			46						305	
	6		21			18						279	
	7		16			8						83	
	8		3			5						15	
	9		6			2						2	
10	1		2			4						2	
	2		6			5						1	
	3		7			25						5	
	4		30			43						39	
	5		37			40						158	
	6		29			27						320	
	7		18			17						193	
	8		7			4						42	
	9		3			4						13	
	10		6			2						1	

Table 2 shows the distribution of population units by DHR,GMS and GS.Five (5) strata formations were obtainable for data 1 and data 2 while Six (6) strata formations for data 3 and data 4 when DHR is used. GMS produced ten strata formation for data 1, 2, and 4 while deep stratification occur in five strata formation of data 3. GS is possible in three strata for data 1, two through six strata for data 2 and in two, three and six strata formations for data 3 and 4.

5.2 Sample Estimation

Using relation (3) and (7) above optimum samples are located to each stratum from fixed sample of sizes 48, 57, 258 and 258 for data 1 to 4as shown in Table 3 below. Data 1 and GS straightly made use of optimum allocation. Beyond 2 strata, DHR and GMS demanded the use of revised optimum allocation for data 2, 3 and 4, with oversampling in the last stratum and in some cases the last two strata of their respective strata formations, i.e. besides data 1, optimum allocated sample sizes are bigger than the stratum size itself for data 2 to 4, hence the use of Revised optimum allocation. Cochran (1977, pp.104)

Simple random sampling without replacement was used to select samples from each stratum using **R**packages (generating seed of 123).It should be noted that with deep stratification in five strata formation by GMS for data 3, sample allocation had been exhausted optimally by the first four stratum with zero unit to the last stratum. In the same vein, optimum allocation gives one unit to stratum I of strata formation by GS for data 2 unlike proportional allocation, thus discretion and personal intuition comes to play as stated by Raj and Chandhok (1998, pp.107) by allocating additional unit to Stratum I from the stratum with largest sample. Keskinturk and Er (2007, pp.62) adopted similar approach to enable estimation with GMS. (See * in Table 3). Relevant statistics for the purpose of estimating the population parameters were obtained and the variance $V(\bar{y}_{st})$ and $MSE\ (\bar{y}_{st})$ of the population mean $(\bar{y}_{st})$ were computed. Estimates of the variance and MSE are shown in Table 4 and 5 respectively.

Table 3: Distribution of Sample Units into Stratum by Number of Strata for data 1 to 4 Using Optimum Allocation.

Number of Strata	Stratum	DATA 1			DATA2			DATA3			DATA 4		
		DHR	**GMS**	**GS**	**DHR**	**GMS**	**GS**	**DHR**	**GMS**	**GS**	**DHR**	**GMS**	**GS**
2	1	20	22		29	5	2*	90	233	111	128	19	119
	2	28	26		28	52	55	168	25	147	130	239	139
3	1	28	5	2	23	2	2*	130	187	117	94	2*	123
	2	9	31	35	17	40	17	27	61	104	54	179	90
	3	11	12	11	17	15	38	101	10	37	110	77	45
4	1	4	4		28	2	2*	23	128		32	0	
	2	17	19		3	17	19	97	106		127	26	
	3	12	18		18	30	16	91	18		43	206	
	4	15	7		8	8	20	47	6		56	26	
5	1	5	3		30	0	2*	27	76		43	0	
	2	21	10		7	7	20	114	143		52	4	
	3	14	23		3	23	20	46	28		37	121	
	4	6	8		11	21	9	50	11		90	119	
	5	2	4		6	6	6	21	0		36	14	
6	1		2			0	2*	30		91	51	0	77
	2		5			2	17	128		17	61	2*	22
	3		19			15	8	52		30	44	32	35
	4		12			25	19	16		46	21	157	53
	5		7			11	9	11		34	58	59	33
	6		3			4	2*	21		40	23	8	38
7	1					0						0	
	2					2						0	
	3					10						7	
	4					18						85	
	5					18						127	
	6					7						35	
	7					3						4	
8	1					0						0	
	2					0						0	
	3					5						2	
	4					14						31	
	5					18						123	
	6					12						76	
	7					6						24	
	8					2						2	

Strata formations with $\mathbf{n_h < 2}$ are beyond the scope of this study and were exempted from estimation purpose. However, optimum allocation continuously gives one unit to stratum I of data 2 with GS hence data adjustment from the stratum with largest unit to enable estimation using Keskinturk and Er (2007, pp.62) experience. Beyondthree strata formation with GMS optimum allocation has zero units in early stratum for data 2 and 4.

5.3 Results

Table 4 and 5 below present the Variance and MSE estimate of the population mean for the four data sets by the three methods.

Table 4: Variance of population mean for the three approaches for the four data sets

	Data 1			**Data2**			**Data 3**			**Data 4**		
Strata	DHR	GMS	GS	DHR	GMS	GS	DHR	GMS	GS	DHR	GMS	GS
2	0.098690	0.089851		34.78	128.76	2682.25	0.57960	0.62554	1.36832	8.4273	18.2072	21.1743
3	0.048783	0.067589	0.074031	20.79	36.51	578.96	0.22953	0.36151	1.51920	3.0593	6.9140	25.9403
4	0.024145	0.034797		12.46	23.03	813.70	0.11481	0.24619		2.1646	6.0417	
5	0.015230	0.023983		13.11	3.43	585.92	0.09489	0.20642		1.1201	3.4276	
6		0.016269			5.33	398.56	0.08035		1.06873	0.8765	3.3028	18.8300
7					3.12						2.1526	
8					4.34						1.7707	
9												
10												

As reported by previous studies Jessen (1942), Jessen and Houseman (1944), Judez and Chaya (1999, 2000), the popular methods of strata constructions; DHR and GMS are more precise than GS. But in terms of MSE criterion, GS is more accurate than GMS for data 1 and has the minimum MSE estimates for data 2 to 4. i.e. MSE (DHR)< MSE(GS)<MSE(GMS) for data 1 while MSE(GS) <MSE (GMS)< MSE (DHR) for data 2, 3 & 4 (See Table 5 below). It should be noted in Table 1 that skewness of data 1 is less than 1 and its frequency distribution depicts a normal distribution unlike data 2, 3 and 4 that are highly skewed.

Table 5: Mean Square Errors of population mean for the three approaches for the four data sets

	Data 1			**Data2**			**Data 3**			**Data 4**		
Strata	DHR	GMS	GS	DHR	GMS	GS	DHR	GMS	GS	DHR	GMS	GS
2	0.87617	0.64637		41114.2	50980.6	2732.86	148.217	33.335	1.527	1702.04	533.63	21.321
3	0.10790	0.62326	0.569831	36373.7	34263.7	631.15	206.897	54.794	2.154	3422.86	1432.74	30.481
4	1.78109	0.63904		35700.5	36302.1	2444.54	176.650	74.482		1943.71	1651.97	
5	0.03496	0.12031		30470.8	45593.6	2179.63	115.776	48.911		2310.21	1705.40	
6		0.06573			43252.6	3415.31	52.655		3.919	1485.07	1655.82	27.774
7					42345.5						2339.13	
8					41668.5						2199.67	
9												
10												

V. CONCLUSION

This study has empirically proved that GS is more efficient and should be used when accurate estimates are required for particularly depicted geographic area. This shows that the bias $[\sum (w_h - W_h)\overline{Y}_h]^2$ associated with GS tends to Zero or is at very minimum value when compared with those of DHR and GMS who yielded more precise estimates than GS but have high MSE values. It is also pertinent to mention that $MSE\ (\bar{y}_{st})_{PROP} < MSE\ (\bar{y}_{st})_{OPT}$ meaning that proportional allocation yields better estimates than optimum allocation. Therefore, the use of precision favours optimum allocation while proportional allocation suits the MSE as a measure of accuracy. However GS yield minimum MSE estimates with optimum allocation over other competing methods of strata construction.

Finally, GS performs poorly in term of its precision in comparison with other methods as observed by previous studies and confirmed in this study and therefore not recommended when precision is the basis of assessment. But when high accuracy is desired for positively skewed data that can be stratified geographically, GS is recommended and should be employed for its minimum MSE value using optimum allocation.

REFRENCES

[1]. Cochran, W.G. (1977). Sampling Techniques, Third edition: John Wiley and Sons, New York.
[2]. Dalenius, T. (1950).“The Problem of Optimum Stratification”, SkandinaviskAkturietidskrift, 33, pp.203-213.
[3]. Dalenius, T. and Hodges, J.L., Jr. (1959) “Minimum Variance Stratification” JASA, 54, pp. 88 – 101.
[4]. Durbin, J. (1959), “Review of sampling in Sweden”. Journal of Royal Statistical Societies. A. 122, pp. 246-248.
[5]. Ekman, G. (1959). “An Approximation Useful in Univariate Stratification”. Annals of Mathematical Statistics, 30, pp. 210-229.
[6]. Gunning P, and Horgan J. M (2004) “A New algorithm for the construction of stratum boundaries in skewed population” Survey Methodology, 30, No. 2, pp.159-166.
[7]. Gunning P, Horgan J. M. and Keogh G. (2006) “Efficient Pareto Stratification”. Mathematical Proceedings of Royal Irish Academy 106A (2), PP. 131-138.
[8]. Gunning P. Horgan, J. M. and Yancey W. (2004) “Geometric Stratification of Accounting data”. J.de contaduria Y. Administration, 214, Septembrie – Diciembre.
[9]. Hedlin D. (2000) “A procedure for Stratification by an Extended Ekman rule”. Journal of Official Statistics 6, No. 1, pp.15-29.
[10]. Horgan, J. M. (2006). “Stratification of Skewed Populations: A review”. International Statistical Review, 74, 1, pp.67-76.
[11]. Jessen, R. J. (1942) Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural experiment statistical research bulletin 304.
[12]. Jessen, R.J. and Houseman E.E. (1944) Statistical Investigations of Farm Sample Survey taken in Iowa, Florida and California.Iowa Agricultural experiment statistical bulletin, 329.
[13]. Judez, L. and Chaya, C. (1999) “Effects of Geographical Stratification in a Farm Accountancy Data Network on the Accuracy of the Estimates. Journal of Agricultural Economics, 50, Number 3, pp. 388-399.
[14]. Judez, L. and Chaya, C. (2000) “Comparative Analysis of Alternative Sampling plans to create a farm Accountancy Data Network for the Agricultural Section of Navarra”. Questiio, 21, pp. 137-150.
[15]. Judez L, Chaya C, Miguel T. M and Bru R (2005). Stratification & sample size of data sources for the agricultural mathematical Programming models. Report of the commissioned study by the EU and Spanish Government.
[16]. Keskinturk T. and Er S. (2007) “A Genetic algorithm approach to determine boundaries and Sample size of each stratum in stratified Sampling”. South Pacific Journal of Natural Sciences, B., 21, pp 91-95.
[17]. Kozak, M. (2004) “Optimal Stratification Using Random Search Method in Agricultural Surveys”.Statistics in Transition, 6, No. 5, pp.797-806.
[18]. Lavalle, P. and Hidiroglou, M.A. (1988) “On the Stratification of Skewed Populations”.Survey Methodology 14, 1, pp.33-43.
[19]. Mahalanobis, P.C. (1952). “Some Aspects of the Design of Sample Surveys”.Sankhya, 12, pp 1-7.
[20]. Mendehall, W, Ott, L. and Scheafffer, R. L. (1971); Elementary Survey Sampling; Duxbary Press, Belmont, California.
[21]. Murthy, M. N. (1967) Sampling Theory and methods 2nd Edition Statistical Publishing Society, Calcutta - 35, India.
[22]. 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 Royal Statistical Society, 97, pp.558-606.
[23]. Okafor, F. C. (2002) Sample Survey Theory with Applications Afro-Orbis Publications Ltd. Nsukka, Nigeria.
[24]. Raj, D. and Chandhok, P. (1998) Sample Survey Theory; Narosa publishing House, 6, Community Centre, Panchsheel Park, New Delhi 100 017.
[25]. Sukhatme P.V., Sukhatme B.V. and Asok, C. (1984).Sampling Theory with Applications.3rd Edition, Iowa University Press, USA.
[26]. Thomson, J. (1976) “A comparison of an approximately optimal stratification given proportional allocation with other methods of stratification and allocation”.Metrika, 23, 1, pp.15-25.