International Journal of Statistics and Applications 2012, 2(1): 1-10
DOI: 10.5923/j.statistics.20120201.01

Comparison of the Efficiency of the Various Algorithms in Stratified Sampling when the Initial Solutions are Determined with Geometric Method

Şebnem Er

Quantitative Methods Department, Istanbul University, Istanbul, 34320, Turkey

Abstract The main aim of this paper is to examine the efficiency of Genetic Algorithm (GA) of Keskintürk and Er (2007)[1], Kozak's (2004) Random Search[2] and Lavallée and Hidiroglou's (1988) Iterative Algorithm method[3] on determination of the stratum boundaries that minimize the variance of the estimate. Initial starting boundaries of the mentioned algorithms are obtained randomly. Here, it is aimed to reach better results in a shorter period of time by utilizing the initial boundaries obtained from Gunning and Horgan's (2004) geometric method[4] compared to the random initial boundaries. Three algorithms are applied on various populations with both random and geometric initial boundaries and their performances are compared. With the stratification of 11 heterogenous populations that have different properties, higher variance of the estimates or infeasible solutions can be observed once the initial boundaries are obtained with geometric method.

Keywords Stratified sampling, Stratum boundaries, Genetic algorithm, Random search, Iterative method

1. Introduction

In stratified sampling,in order to gain more precision than other methods of sampling, a heterogeneous population is divided into subpopulations, each of which is internally homogeneous. As a result the main problem arising in stratified sampling is to obtain the optimum boundaries. Several numerical and computational methods have been developed for this purpose. Some apply to highly skewed populations and some apply to any kind of populations. An early and very simple method is the cumulative square root of the frequency method (cum√f) of Dalenius & Hodges in 1959[5]. More recently Lavallée & Hidiroglou algorithm[3] and Gunning & Horgan's (2004) geometric method[4] have been proposed for highly skewed populations whereas Kozak's (2004) random search method[2] and Keskinturk & Er's (2007) genetic algorithm (GA) method[1] have been proposed for even non-skewed populations. Very recently, Brito et.all[6] proposed an exact algorithm for the stratification problem with only proportional allocation based on the concept of minimum path in graphs and they called their method StratPath. Moreover, developed an iterated local search method to solve the stratification problem of variables with any distribution with Neyman allocation[7].All these methods aim to achieve the optimum boundaries that maximise the level of precision or equivalently minimise the variance of the estimate or the sample size required to reach a level of precision and some of them are available in the stratification package stratification for use with the statistical programming environment R[8]; freely available on the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=stratification.

The main aim of this research is to compare the efficiency ratios of the Lavallée ve Hidiroglou iterative method, Kozak's random search method and Keskinturk and Er's genetic algorithm approach when the initial boundaries are obtained either randomly or from the geometric method of Gunning and Horgan, and to examine the performances of the three methods. The predetermined total sample size (n) is allocated using Neyman[9] optimum allocation method. The paper is structured as follows: In the second section the exact solution of Dalenius[10] and the methods that are developed in order to approximately solve the Dalenius equations are briefly explained. In the third section, the results obtained with Lavallée and Hidiroglou's iterative method, Kozak's random search method and Keskintürk and Er's genetic algorithm are given when the initial boundaries are obtained randomly or from the geometric method of Gunning and Horgan and the performance of the algorithms are compared.

* Corresponding author:
er.sebnem@gmail.com (Şebnem Er)
Published online at http://journal.sapub.org/statistics

2. Dalenius' (1950) Exact Solution

Dalenius (1950)[10] considers a density $f(x)$ with mean

$$\mu=\int_{-\infty}^{+\infty} tf(t)dt \tag{1}$$

The range (X_{max}-X_{min}) of the stratification variable x is divided into L parts at points $b_1<b_2<...<b_{L-1}$, each part corresponding to a stratum. When a sample of $n=\sum n_h$ observations is selected from $f(x)$, the true mean

$$\mu=\sum_{h=1}^{L} W_h\mu_h \tag{2}$$

is estimated by Cochran as [11]

$$\bar{x}_{st}=\sum_{h=1}^{L} W_h\mu_h \tag{3}$$

where for the h_{th} stratum W_h , μ_h, $\bar{x}_{st}$ and are calculated as follows[11]:

$$W_h=\int_{b_{h-1}}^{b_h} f(t)dt=\frac{N_h}{N} \tag{4}$$

$$\mu_h=\frac{\sum_{i=1}^{N_h} x_{hi}}{N_h} \tag{5}$$

$$\bar{x}_h=\frac{\sum_{i=1}^{n_h} x_{hi}}{n_h}\,. \tag{6}$$

The estimate of the mean $\bar{x}_{st}$ has a variance of

$$\sigma^2(\bar{x}_{st})=\sum_{h=1}^{L} W_h^2\frac{\sigma_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \tag{7}$$

where the true variance is

$$\sigma_h^2=\frac{\sum_{i=1}^{N_h}\left(x_{hi}-\mu\right)^2}{N_h-1}\,. \tag{8}$$

If the sampling fractions n_h/N_h are negligible then the variance could be written in short,

$$\sigma^2(\bar{x}_{st})=\sum_{h=1}^{L} W_h^2\frac{\sigma_h^2}{n_h} \tag{9}$$

It is well-known that this variance of the estimate is minimum

$$\sigma_{\min}^2(\bar{x}_{st})=\frac{1}{n}\left(\sum_{h=1}^{L} W_h\sigma_h\right)^2 \tag{10}$$

when total sample size n is allocated using Neyman's optimum allocation method [9]:

$$n_h=n\frac{N_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h}\,. \tag{11}$$

Therefore the variance of the estimate is a function of the boundaries b_h. As a result, it is very difficult to find the boundaries that minimise the variance of the estimate. Dalenius (1950)[10] has shown that the variance of the estimate obtained with Neyman's optimum allocation method is optimum or in other words minimum, when the stratum boundaries satisfy the following equations:

$$\frac{\sigma_h^2+\left(b_h-\mu_h\right)^2}{\sigma_h}=\frac{\sigma_{h+1}^2+\left(b_h-\mu_{h+1}\right)^2}{\sigma_{h+1}} \tag{12}$$

It is very difficult to find the stratum boundaries b_h that satisfy these equations remembered as Dalenius equations since these equations include σ_h^2 and μ_h that both vary with b_h stratum boundaries. As a result, there have been many approximations and algorithms proposed for solving Dalenius equations. The widely known simple method among the proposals is the cumulative square root frequency method of Dalenius and Hodges (1959) ($cum\sqrt{f}$) [5]. Then, in 1988 Lavallée and Hidiroglou's iterative approach[3], in 2004 Gunning and Horgan's geometric method [4] and Kozak's random search method[2], in 2007 Keskintürk and Er's genetic algorithm method[1] are developed in order to find the stratum boundaries. Among these methods, geometric method is the simplest method that does not include any complex algorithms. Therefore, the main aim of this research paper is to set the initial boundaries of the proposed algorithms with geometric method and compare the efficiencies of the algorithms when the boundaries are obtained with or without geometric method since it is believed that these algorithms would reach the solution in a shorter period once they start searching the entire space at a reasonable point. The details of the approaches and algorithms of these methods could be obtained from the original papers of Dalenius and Hodges' (1959)[5], Gunning and Horgan (2004)[4], Kozak (2004)[2] and Keskintürk and Er's (2007)[1]. All of these methods could be applied in R statistical environment using stratification[12] and GA4stratification[13] packages but the GA results given in this studyare obtained in Matlab 7.0 since in the package there is no option for setting the initial boundaries with non-random results.

3. Application

3.1. Populations for Stratification

In this paper, many populations are used for stratification with different skewness, kurtosis, mean, standard deviation and size properties.Those populations that are available in the R stratification[12] and GA4Stratification[13] packages are used for stratification. Each of the populations are divided into 3, 4, 5 and 6 strata and the boundaries are obtained using Lavallée and Hidiroglou, Kozak and GA methods with random and geometric initial boundaries.

Pop1: An accounting population of debtors in an Irish firm (Debtors).

Pop2: The population in thousands of US cities in 1940 (UScities).

Pop3: The number of students in four-year US colleges in 1952-1953 (UScolleges).

Pop4: The resources in millions of dollars of large commercial US banks (USbanks).

Pop5: Number of municipal employees of 284 municipalities in Sweden in 1984 (ME84).

Pop6: Population in thousands of 284 municipalities in Sweden in 1975 (P75).

Pop7: Real estate values in millions of kronor according to 1984 assessment of 284 municipalities in Sweden in 1984 (REV84)

Pop8: Simulated Data from the Monthly Retail Trade

Survey of Statistics Canada (MRTS)

Pop9: Household income before taxes from the 2001 Survey of Household Spending carried out by Statistics Canada (HHINCTOT)

Pop10: Net sales data of 487 Turkish manufacturing firms among the largest 500 firms in 2004 by Istanbul Chamber of Industry (ICI) (iso2004)

Pop11: Net sales data of 485 Turkish manufacturing firms among the largest 500 firms in 2005 by Istanbul Chamber of Industry (ICI) (iso2005)

The boxplots of the populations are displayed between Figures 1 and 3, and the summary statistics of the populations are given in Table 2.

Referring the descriptive statistics in Table 2 and boxplots in Figures 1-3, we see that the populations to be stratified are highly heterogenous which makes stratified sampling efficient to use. For comparison, the initial boundaries are obtained with both random initial boundaries and with geometric method. The populations are divided into 3, 4, 5 and 6 strata and the total sample size is determined as 100 for Pop1-Pop11. For genetic algorithm, the number of iterations is set to 10000, the GA population size to 35, the crossover rate to 0.99 and the mutation rate to 0.15. For efficiency (efficiency – eff) comparisons of the ratio of variance of the estimates or the ratios of squares of coefficient of variations (CV) are calculated and given in Appendix 1. Since Lavallée and Hidiroglou's (LH) method is based on sampling all of the elements in the last stratum (take-all top stratum), the following efficiency ratios are calculated if GA and Kozak's methods provide a take-all top stratum solution:

$$eff_{GA/Kozak} = \frac{\sigma^2_{GA}\left(\bar{x}_{st}\right)}{\sigma^2_{Kozak}\left(\bar{x}_{st}\right)} = \frac{\left(CV_{GA} \times \mu_{\bar{x}_{st}}\right)^2}{\left(CV_{Kozak} \times \mu_{\bar{x}_{st}}\right)^2} = \left(\frac{CV_{GA}}{CV_{Kozak}}\right)^2 \quad (13)$$

$$eff_{GA/LH} = \left(\frac{CV_{GA}}{CV_{LH}}\right)^2 \quad (14)$$

$$eff_{Kozak/LH} = \left(\frac{CV_{Kozak}}{CV_{LH}}\right)^2 \quad (15)$$

For those situtations where some of the last stratum is sampled, only the efficiency ratio between GA and Kozak's method ($eff_{GA/Kozak}$) is calculated.

From the efficiency and the coefficient of variation ratios given in Table 3in Appendix 1 and from the strata and sample sizes given in Table 5 in Appendix 2, it can be seen that the algorithms compared in this paper provide very close results and that the stratum boundaries are very close to each other when the initial boundaries are set randomly.When we look at the summary of the results given in Table 1, we see that the number of cases where GA or Kozak is better than the other one does not differ much and the gains in efficiencies are close to each other.

Table 1. Number of Cases where the Chosen Algorithm Gives Better Results and the Range of the Efficiency Gain (Random Initials)

H	Better results with GA	Better Results with Kozak	Both Same	Total
3	**4** (‰0.1-0.6)	**none**	**7**	**11**
4	**2** (‰0.1-7.2)	**1** (‰1.2)	**8**	**11**
5	**6** (‰0.2-%26)	**3** (‰1.2-%37)	**2**	**11**
6	**8** (‰0.5-%25)	**3** (‰7.6-%27)	**none**	**11**

On the other hand, the results are different with higher coefficient of variations when the initial boundaries are obtained with geometric method (Table 4).Moreover, when the initial boundaries are set to be found with geometric method, many infeasible or nonconverged results are obtained. For example, when we look at Table 4 where the initial boundaries are obtained with geometric method, we see that the coefficient of variations for GA increases in 32 cases among 44 cases. Yet some of these increases in the CVs result from a nonconverged or an infeasible solution. Only in 4 cases there is a gain in efficiency ranging in between ‰0.01 (CV falling from 0.01437 to 0.01436 for H=5 for Pop3-UScolleges) and %0.186 (falling from 0.02485 to 0.02299 for H=5 for Pop8-MRTS), which could be counted as a very minor gain. The results for L&H and Kozak's are more or less the same with the results obtained for GA. When the initial boundaries are obtained with geometric method, with each of Kozak's and L&H's methods there is an efficiency gain in only 5 cases, which are again minor. For these reasons, Lavallée and Hidiroglou's iterative method, Kozak's random search method and Keskintürk and Er's genetic algorithms give more efficient results when the initial boundaries are set randomly due to their nature. As a result, it can be concluded that starting with geometric initial boundaries does not have much contribution on the efficiency ratios or on the stratum boundaries for the computational methods. As proposed by Horgan (2011) [14], in order to obtain feasible solutions in some data sets,some modifications should be applied before utilising the geometric method. Horgan (2011) [14] suggests that the data should be analysed before applying the stratified sampling scheme if there are extreme outliers. In this paper the revitised version of the geometric method is not applied since the algorithms examined here already give good results with random initials. Furthermore, if any researcher wants to use the geometric initial boundaries for data sets with extreme outliers, modified version of the geometric method should be used.

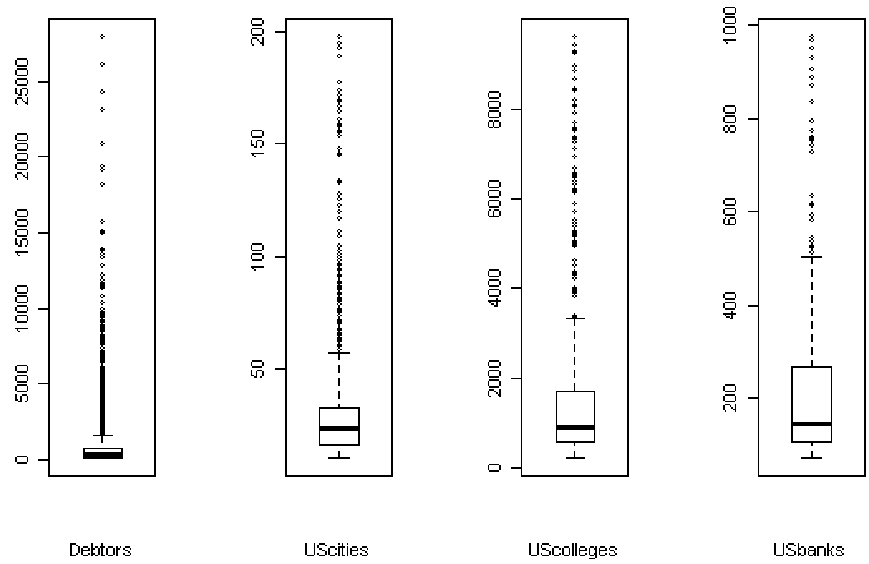

Figure 1. Boxplots of Pop1-Pop4

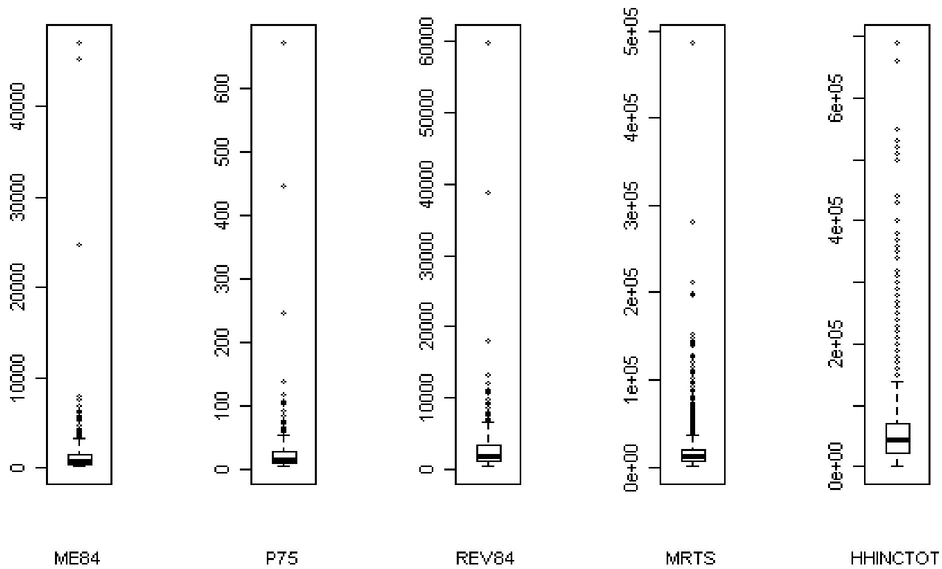

Figure 2. Boxplots of Pop5-Pop9

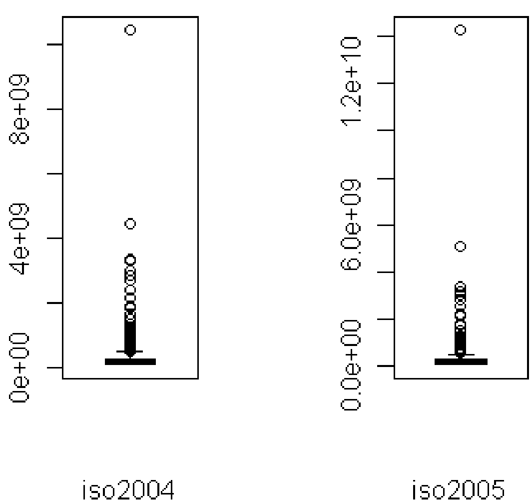

Figure 3. Boxplots of Pop10-Pop11

Table 2. Summary Statistics of the Populations

Pop	Name	N	Range	Skewness	Kurtosis	Mean	StdDev.
Pop1	Debtors	3369	40-28000	6.44	59.00	838.64	1873.99
Pop 2	Uscities	1038	10-198	2.87	9.12	32.57	30.4
Pop 3	UScolleges	677	200-9623	2.45	5.80	1563	1799.06
Pop 4	USbanks	357	70-977	2.07	4.06	225.62	190.46
Pop 5	ME84	284	173-47074	8.64	84.04	1779.07	4253.13
Pop 6	P75	284	4-671	8.43	88.56	28.81	52.87
Pop 7	REV84	284	347-59877	7.83	81.33	3088.09	4746.16
Pop 8	MRTS	2000	141-486366	8.61	136.20	16882.8	21574.88
Pop 9	HHINCTOT[i]	16025	100-690000	2.71	18.79	52123.73	41120.41
Pop 10	iso2004	487	63582908-10446591755	10.03	137.91	278237616.44	637769009.37
Pop 11	iso2005	485	69121110-14239223472	12.63	206.49	305852522.35	785107451.87

4. Conclusions

Stratified sampling is a sampling methodology used for heterogeneous populations in order to gain more precision than other methods of sampling. This paper examines the improvement in the efficiency ratios and stratum boundaries obtained with Lavallée and Hidiroglou [3], Kozak [2] and Keskintürk and Er's (2007) [1] methods once the initial boundaries are obtained with geometric method. With the stratification of 16 heterogenous populations that have different properties, higher variance of the estimates or infeasible solutions can be observed. As a result, researchers should be much more rigorous when using geometric method for the initial boundaries in algorithmic methods or else use the modified version of geometric method once the data has very extreme values.

ACKNOWLEDGEMENTS

I would like to thank the reviewer of this article whose comments and suggestions have helped improve the paper.

REFERENCES

[1] Keskintürk, T., Er, Ş., A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. Computational Statistics & Data Analysis, 52, 1, pp.53-67, 2007.

[2] Kozak, M., Optimal Stratification Using Random Search Method in Agricultural Surveys. Statistics in Transition, 6, 5, pp.797-806, 2004.

[3] Lavallée, P., Hidiroglou, M., On the Stratication of Skewed Populations, Survey Methodology, 14, 1, pp.33-43, 1988.

[4] Gunning, P., Horgan, J.M., A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. Survey Methodology, 30, 2, 2004.

[5] Dalenius, Tore, Hodges, Joseph L.Jr., "Minimum Variance Stratification", Journal of the American Statistical Association, 54 (285), pp.88-101, 1959.

[6] Brito, J., Maculan, N. Lila, M., Montenegro, F. An Exact Algorithm for the Stratication Problem with Proportional Allocation. Optimization Letters, 4, 2, pp.185-195, 2010.

[7] Brito, J., Ochi, L., Montenegro, F., Maculan, N. An Iterative Local Search Approach Applied to the Optimal Stratifica-tionProblem. International Transactions in Operational Research. 17, 6, pp.753-764, 2010.

[8] R Development Core Team. R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, (URL) http://www.r-project.org. 2005.

[9] Neyman, Jerzy. "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 (4), pp.558-625, 1934.

[10] Dalenius, T. The problem of optimum stratification, Skandinavisk Aktuarietidskrift, pp. 203-213, 1950.

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

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

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

[14] Horgan, J.M., "Geometric Stratification Revitised". ISI World Congress 2011 Proceedings, 2011.

APPENDIX 1

Table 3. The efficiency and coefficient of variation ratios of LH, GA and Kozak's methods when the initial boundaries are obtained randomly

H	CVLH	CVGA	CVKozak	effGA/Kozak	effGA/LH	effKozak/LH
Pop1: Debtors						
3	0.06930*	0.05554	0.05554	0.9999	-	-
4	0.04721*	0.04049	0.04049	1.0000	-	-
5	0.03331*	0.03131	0.03131	0.9998	-	-
6	0.02678*	0.02562	0.02587	0.9801	-	-
Pop2: Uscities						
3	0.03217*	0.02649	0.02649	1.0000	-	-
4	0.02249*	0.01927	0.01934	0.9928	-	-
5	0.01943*	0.01437	0.01680	0.7312	-	-
6	0.01552* n=110	0.01214	0.01209	1.0076	-	-
Pop3: Uscolleges						
3	0.03460*	0.02749	0.02749	0.9998	-	-
4	0.02399*	0.02018	0.02018	1.0000	-	-
5	0.01995*	0.01607	0.01726	0.8672	-	-
6	0.01715*	0.01323	0.01324	0.9995	-	-
Pop4: USbanks						
3	0.01839*	0.01802	0.01802	1.0000	-	-
4	0.01270*	0.01270*	0.01270*	1.0000	0.9991	0.9991
5	0.01094*	0.00861*	0.00861*	1.0000	0.6198	0.6198
6	0.00710*	0.00710*	0.00711*	0.9981	0.9997	1.0016
Pop5: ME84						
3	0.01296*	0.01296*	0.01296*	1.0000	0.9998	0.9998
4	0.00870*	0.00870*	0.00870*	1.0000	0.9991	0.9991
5	0.00663*	0.00661*	0.00661*	1.0000	0.9944	0.9944
6	0.00525*	0.00577*	0.00522*	1.2217	1.2064	0.9875
Pop6: P75						
3	0.01514*	0.01459*	0.01459*	1.0000	0.9278	0.9278
4	0.01068*	0.00966*	0.00966*	1.0000	0.8179	0.8179
5	0.00765*	0.00835*	0.00713*	1.3705	1.1904	0.8686
6	0.00608*	0.00623*	0.00552*	1.2735	1.0521	0.8261
Pop7: REV84						
3	0.01618*	0.01607*	0.01607*	1.0000	0.9954	0.9954
4	0.01120*	0.01120*	0.01120*	1.0000	0.9996	0.9996
5	0.00840*	0.00836*	0.00837*	0.9971	0.9896	0.9924
6	0.00700*	0.00666*	0.00675*	0.9759	0.9074	0.9298
Pop8: MRTS						
3	0.04559*	0.04167	0.04168	0.9994	-	-
4	0.03025*	0.02960	0.02960	0.9999	-	-
5	0.02307*	0.02485	0.02297	1.1704	-	-
6	0.01837*	0.01836*	0.01836*	0.9995	0.9984	0.9988
Pop9: HHINCTOT						
3	0.04503*	0.03184	0.03184	1.0000	-	-
4	0.03114*	0.02430	0.02429	1.0012	-	-
5	0.02379*	0.01979	0.01977	1.0012	-	-
6	0.01974*	0.01629	0.01630	0.9995	-	-
Pop10: iso2004						
3	0.01895*	0.01894*	0.01894*	1.0000	0.9982	0.9982
4	0.01208*	0.01206*	0.01206*	1.0000	0.9973	0.9973
5	0.00927*	0.00908*	0.00925*	0.9626	0.9584	0.9956
6	0.00820*	0.00703*	0.00811*	0.7516	0.7346	0.9773
Pop11: iso2005						
3	0.01833*	0.01833*	0.01833*	0.9999	0.9997	0.9998
4	0.01245*	0.01244*	0.01244*	1.0000	0.9973	0.9973
5	0.00912*	0.00903*	0.00910*	0.9852	0.9810	0.9958
6	0.00808*	0.00706*	0.00805*	0.7689	0.7630	0.9924

* Where there is a take-all top stratum

Table 4. The coefficient of variation ratios of LH, GA and Kozak's methods when the initial boundaries are obtained with geometric method

H	CVLH	CVGA	CVKozak
		Pop1: Debtors	
3	Same	0.05554^{+}	Same
4	Same	0.04073^{+}	Same
5	Same	0.03122^{+}	Same
6	Same	0.02587^{+}	Same
		Pop2: UScities	
3	Same	Same	Same
4	0.02228^{-}	0.01940^{+}	0.01927^{-}
5	0.01590^{-}	0.01436^{-}	0.01436^{-}
6	0.01377 (n=100)	0.01258^{+}	Same
		Pop3: UScolleges	
3	Same	0.02730^{-}	Same
4	Same	Same	Same
5	0.01750^{-}	0.01595^{-}	0.01724^{-}
6	0.01401^{-}	0.01327^{+}	Same
		Pop4: USbanks	
3	Same	Same	Same
4	0.01322^{+}	0.01343^{+}	0.01325^{+}
5	0.01039^{-}	0.01043^{+}	Same
6	0.00753^{+}	0.00751^{+}	Same
		Pop5: ME84	
3	$0.01378^{+\text{N.C.}}$	Same	Same
4	$0.01596^{+\text{N.C.}}$	$0.01296^{+\text{N.C.}}$	$0.01296^{+\text{I.F.}}$
5	0.01199^{+}	$0.00870^{+\text{N.C.}}$	0.00746^{+}
6	$0.01180^{+\text{N.C.I.F.}}$	$0.00858^{+\text{N.C.}}$	$0.00870^{+\text{I.F.}}$
		Pop6: P75	
3	$0.01558^{+\text{N.C.}}$	0.01459^{+}	Same
4	$0.01710^{+\text{N.C.}}$	0.01191^{+}	$0.01459^{+\text{I.F.}}$
5	0.01385^{+}	0.00847^{+}	0.00829^{+}
6	$0.01243^{+\text{N.C., I.F.}}$	$0.00835^{+\text{I.F.}}$	$0.00966^{+\text{I.F.}}$
		Pop7: REV84	
3	0.01607^{-}	0.01614^{+}	Same
4	$0.01318^{+\text{N.C.}}$	0.01166^{+}	0.01166^{+}
5	$0.01601^{+\text{N.C., I.F.}}$	$0.01120^{+\text{I.F.}}$	0.01041^{+}
6	$0.01306^{+\text{N.C., I.F.}}$	0.01047^{+}	0.00835^{+}
		Pop8: MRTS	
3	Same	0.04169^{+}	Same
4	Same	Same	Same
5	Same	0.02299^{-}	Same
6	Same	0.01837^{+}	Same
		Pop9: HHINCTOT	
3	Same	0.03939^{+}	Same
4	Same	0.03384^{+}	Same
5	Same	0.02531^{+}	Same
6	Same	0.02275^{+}	Same
		Pop10: iso2004	
3	$0.02111^{+\text{N.C.}}$	Same	Same
4	$0.02148^{+\text{N.C, .I.F.}}$	0.01222^{+}	$0.01894^{+\text{I.F.}}$
5	$0.01832^{+\text{N.C., I.F.}}$	$0.01222^{+\text{I.F.}}$	$0.01220^{+\text{I.F.}}$
6	$0.01469^{+\text{N.C., I.F.}}$	$0.01222^{+\text{I.F.}}$	0.00702^{+}
		Pop11: iso2005	
3	0.01835^{+}	Same	Same
4	$0.02282^{+\text{N.C., I.F.}}$	$0.01840^{+\text{I.F.}}$	$0.01840^{+\text{I.F.}}$
5	$0.01858^{+\text{N.C., I.F.}}$	$0.01255^{+\text{I.F.}}$	$0.01255^{+\text{I.F.}}$
6	$0.01483^{+\text{N.C., I.F.}}$	NONE	0.00706^{-}

$^{\text{I.F.}}$: Infeasible; $^{\text{N.C.}}$: Algorithm did not converge; $^{-}$: a decrease in CV; $^{+}$: an increase in CV.

APPENDIX 2

Table 5. Size of the strata (Nh) and the sample sizes (nh) obtained fromLH, GA and Kozak's methods when the initial boundaries are obtained randomly

H		LH						GA						Kozak					
Pop1: Debtors																			
3	Nh	2894	449	26				2690	545	134				2673	561	135			
	nh	36	38	26				35	28	37				34	29	37			
4	Nh	2179	891	271	28			2085	901	302	81			2071	914	303	81		
	nh	17	24	31	28			19	23	26	32			18	24	26	32		
5	Nh	1856	991	350	146	26		1892	955	339	136	47		1892	954	335	139	49	
	nh	14	19	19	22	26		17	21	20	17	25		17	21	19	17	26	
6	Nh	1608	956	423	223	127	32	1604	956	426	221	118	44	1533	905	493	265	126	47
	nh	11	13	12	12	20	32	12	15	14	14	17	28	10	12	14	17	18	29
Pop2: Uscities																			
3	Nh	795	206	37				749	193	96				749	193	96			-
	nh	35	28	37				43	21	36				43	21	36			
4	Nh	393	433	173	39			434	409	155	40			434	356	154	94		-
	nh	11	20	30	39			19	30	37	14			18	15	21	46		
5	Nh	189	270	367	171	41		393	367	150	89	39		226	271	298	149	94	-
	nh	3	6	18	32	41		21	20	20	21	18		6	8	13	22	51	
6	Nh	154	154	271	267	145	47	274	263	245	128	89	39	226	271	285	128	89	39
	nh	3	3	8	18	31	47	12	12	13	18	24	21	9	12	17	18	24	20
Pop3: UScolleges																			
3	Nh	485	137	55				478	130	69				478	130	69			
	nh	26	19	55				42	23	35				43	22	35			
4	Nh	256	242	118	61			256	234	118	69			256	234	118	69		
	nh	9	12	18	61			15	16	25	44			15	16	25	44		
5	Nh	135	201	167	108	66		253	221	82	60	61		192	166	145	105	69	
	nh	4	6	8	16	66		18	16	9	13	44		10	7	11	23	49	
6	Nh	93	151	134	126	104	69	132	180	166	78	52	69	133	179	166	77	53	69
	nh	2	4	4	6	15	69	6	9	10	9	8	58	6	9	10	9	8	58
Pop4: USbanks																			
3	Nh	212	85	60				212	84	61				212	84	61			
	nh	22	18	60				26	20	54				26	20	54			
4	Nh	110	108	76	63			111	112	73	61			111	112	73	61		
	nh	8	9	20	63			8	11	20	61			8	11	20	61		
5	Nh	70	68	85	71	63		110	101	54	32	60		110	101	54	32	60	
	nh	4	4	9	20	63		12	11	10	7	60		12	11	10	7	60	
6	Nh	54	60	97	54	32	60	54	68	90	53	32	60	51	63	97	54	32	60
	nh	4	4	13	11	8	60	4	6	11	11	8	60	3	5	13	11	8	60
Pop5: ME84																			
3	Nh	144	79	61				145	78	61				145	78	61			
	nh	20	19	61				20	19	61				20	19	61			
4	Nh	115	62	45	62			115	64	44	61			115	64	44	61		
	nh	17	12	9	62			17	13	9	61			17	13	9	61		
5	Nh	54	69	54	43	64		54	69	56	41	64		54	69	56	41	64	
	nh	7	7	12	10	64		7	7	13	9	64		7	7	13	9	64	
6	Nh	42	72	32	36	38	64	54	69	56	41	19	45	54	61	33	34	37	65
	nh	6	8	4	8	10	64	9	10	17	13	6	45	8	6	4	8	9	65
Pop6: P75																			
3	Nh	132	89	63				150	77	57				150	77	57			
	nh	16	21	63				24	19	57				24	19	57			
4	Nh	64	91	66	63			111	73	43	57			111	73	43	57		
	nh	7	12	18	63			19	15	9	57			19	15	9	57		
5	Nh	45	66	65	45	63		123	61	33	19	48		64	68	52	34	66	
	nh	6	6	13	12	63		29	14	5	4	48		10	8	11	5	66	
6	Nh	45	34	53	52	42	58	45	87	52	33	18	49	45	66	39	34	33	67
	nh	7	2	7	14	12	58	8	17	15	6	5	49	7	9	5	6	6	67

Table 5. Continues: Size of the strata (Nh) and the sample sizes (nh) obtained fromLH, GA and Kozak's methods when the initial boundaries are obtained randomly

H		LH						GA						Kozak					
Pop7: REV84																			
3	Nh	131	84	69				138	81	65				138	81	65			
	nh	16	15	69				19	16	65				19	16	65			
4	Nh	64	81	70	69			64	81	69	70			64	81	69	70		
	nh	6	10	15	69			6	9	15	70			6	9	15	70		
5	Nh	61	60	47	47	69		64	74	53	39	54		61	69	51	34	69	
	nh	7	6	7	11	69		9	12	11	14	54		7	8	9	7	69	
6	Nh	50	55	40	46	39	54	61	60	42	43	26	52	57	51	37	42	28	69
	nh	7	7	6	10	16	54	11	8	9	12	8	52	8	5	5	7	6	69
Pop8: MRTS																			
3	Nh	1546	426	28				1227	671	102				1204	688	108			
	nh	42	30	28				30	32	38				29	31	40			
4	Nh	1017	749	206	28			1023	742	203	32			1017	748	203	32		
	nh	26	25	21	28			29	28	21	22			29	28	21	22		
5	Nh	749	690	379	153	29		749	698	371	150	32		774	675	369	150	32	
	nh	19	18	16	18	29		20	19	17	28	16		22	18	16	17	27	
6	Nh	513	580	455	280	140	32	521	573	455	283	136	32	513	580	458	281	136	32
	nh	13	13	11	13	18	32	13	12	11	14	18	32	13	12	11	14	18	32
Pop9: HHINCTOT																			
3	Nh	10566	5451	8				8009	5976	2040				8009	5976	2040			
	nh	40	52	8				31	31	38				31	31	38			
4	Nh	7438	6173	2406	8			6452	5176	3303	1094			6281	5230	3420	1094		
	nh	26	30	36	8			26	23	21	30			24	24	22	30		
5	Nh	5473	5093	4009	1442	8		5900	4375	3336	1960	454		5023	4495	3608	2296	603	
	nh	18	21	24	29	8		26	19	18	18	19		19	19	18	20	24	
6	Nh	4144	3865	3725	2841	1440	10	4811	3860	3422	2482	1182	268	4378	3946	3618	2633	1182	268
	nh	13	12	15	17	33	10	21	16	17	17	13	16	18	16	19	19	13	15
Pop10: iso2004																			
3	Nh	306	125	56				312	120	55				312	120	55			
	nh	21	23	56				23	22	55				23	22	55			
4	Nh	221	133	77	56			229	128	74	56			229	128	74	56		
	nh	13	13	18	56			14	13	17	56			14	13	17	56		
5	Nh	158	108	91	72	58		163	129	85	54	56		158	115	87	69	58	
	nh	7	7	9	19	58		8	11	12	13	56		7	8	9	18	58	
6	Nh	86	83	104	84	72	58	158	108	85	42	39	55	95	105	81	76	65	65
	nh	2	3	7	9	21	58	10	9	10	7	9	55	2	5	5	7	16	65
Pop11: iso2005																			
3	Nh	290	136	59				294	132	59				293	133	59			
	nh	18	23	59				19	22	59				19	22	59			
4	Nh	176	148	96	65			223	122	80	60			223	122	80	60		
	nh	6	12	17	65			13	12	15	60			13	12	15	60		

5	Nh	154	119	76	71	65		166	123	85	51	60		157	117	79	67	65	
	nh	6	7	8	14	65		8	9	12	11	60		6	7	8	14	65	
6	Nh	98	78	99	78	67	65	154	116	75	61	33	46	102	76	100	81	61	65
	nh	3	3	6	9	14	65	9	10	11	14	10	46	3	3	6	10	13	65

[i]Observations with values of zero are excluded from the data since geometric method could not be applied with dataset including zeros as a minimum value.