Computing Efficiency of GA[(1)] versus Simplicity of Geometric Method[(2)] in Boundary Determination in Stratified Sampling

[(1)]Timur Keskintürk & Şebnem ER

Research Assistants at the Faculty of Business Administration, Department of Quantitative Methods in Istanbul University, Avcilar Campus, Istanbul, Turkiye

sebnemer@istanbul.edu.tr, tkturk@istanbul.edu.tr

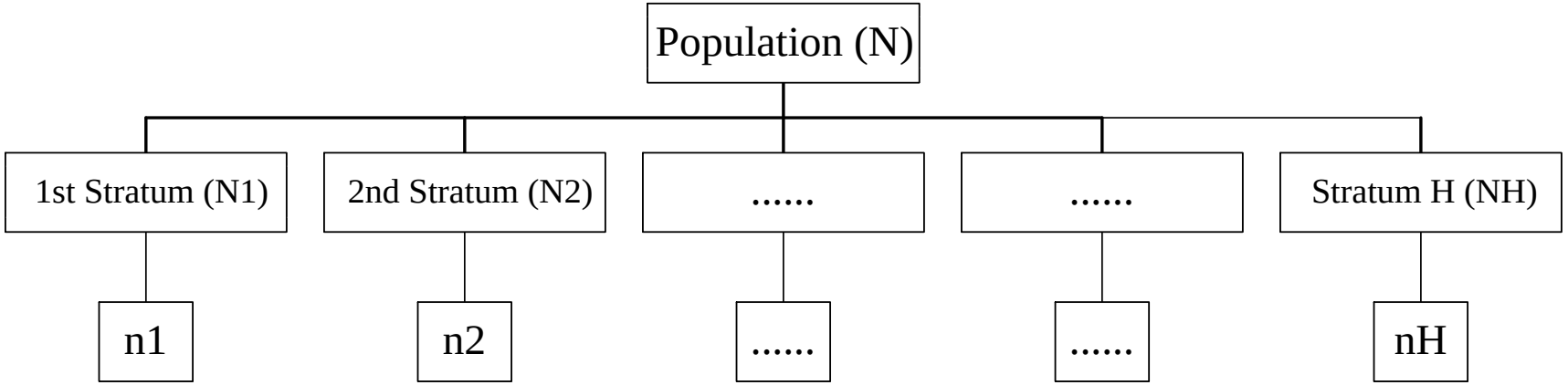
[(2)]Patricia Gunning & Jane M. Horgan

School of Computing, Dublin City University, Dublin 9, Ireland.

Stratified Sampling

“Stratified Sampling” is a methodology in which the elements of a heterogeneous population (N) are classified into mutually exclusive and exhaustive subgroups (strata - H) based on one or more important characteristics.

One of the main objectives of stratified sampling is **to reduce the variance of the estimator** and to get more statistical precision than with the simple random sampling (Cochran, 1977).



The determination of the stratum boundaries and sample allocation

We adopt the general strategy of minimizing the variance of the estimator and introduce a GA approach for the determination of stratum boundaries and sample size allocation.

Several methods are given in the literature for the problem of boundary determination such as

- Dalenius and Hodges' (1959) cumulative square root of the frequency method,
- Ekman's (1959) rule,
- Sethi's (1963) rule,
- Lavallée & Hidiroglou's (1988) algorithm,
- Nicolini's (2001) natural classes method (NCM),
- Gunning and Horgan's (2004) geometric approach,
- Kozak's (2004) random search method.

In our study sampling costs are assumed to be equal for all strata.

Y	Stratification variable
N	Population size
n	Sample size
H	Number of strata
N_h	Number of elements in stratum h (h=1,…,H)
n_h	Sample size in stratum h
σ^2_{yh}	Variance of the elements in stratum h
$\bar{Y}_h$	Mean of elements in stratum h

Estimated variance of the mean in stratified sampling

$$S^2_{\bar{Y}_{strat}} = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{\sigma^2_{yh}}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

Sample Size Allocation Methods

Equal Allocation	(m1) :	$n_h = \frac{n}{H} \qquad n_1 = \cdots = n_h \qquad h = 1, 2, \cdots, H$
Proportional Allocation	(m2) :	$n_h = n \cdot \frac{N_h}{N} \qquad h = 1, 2, \cdots, H$
Neyman's Optimum Allocation	(m3) :	$n_h = n \cdot \frac{N_h \sigma_{yh}}{\sum_{h=1}^{H} N_h \sigma_{yh}} \qquad h = 1, 2, \cdots, H$
Genetic Algorithm	(m4) :	sample sizes n_1, …, n_h are determined with GA.

Genetic Algorithm

The genetic algorithm (GA), developed initially by J. Holland, is a heuristic optimization method (Holland, 1975; Goldberg, 1989; Michalewicz, 1992; Reeves, 1995; Haupt and Haupt, 1998).

This algorithm encodes a potential solution to a specific problem on a simple chromosome and applies genetic operators (selection, crossover and mutation) to these structures so as to preserve critical information.

Individuals yielding better solutions to the problem are likely to survive in a competing environment and tend to result in a good quality offspring (Man and Kwong, 1996).

This process is repeated until a predetermined number of iterations is reached. The best individual in the last generation becomes the solution of the problem.

The principle of any genetic algorithm is given as follows:

Start *Generate random initial generation.*

Fitness Function: *Evaluate the fitness of each chromosome.*

Selection: *Select the better individuals for the next generation.*

Crossover: *With a crossover probability, exchange the parents to form new offspring.*

Mutation: *With a mutation probability mutate new offspring.*

Loop: *If stopping criterion is not reached go to fitness function.*

Stop *and return the best solution in current generation*

GA Approach in Stratified Sampling

The first step of GA is the representation of the combined problem of stratum boundaries and stratum allocation with finite length strings called chromosomes or individuals.

In order to solve the stratification problem with GA, stratification values must be encoded into chromosomes.

The range of ascending values subject to stratification must be divided into H parts by points $Y_1<Y_2<\ldots..<Y_{H-1}$. Each such part corresponds to a stratum boundary.

In GA, several types of encoding can be used to represent this structure. The most commonly used representation is the binary "bit strings", however, real-valued, integer, ternary strings can also be used.

Representation structure (encoding)

In the paper binary encoding is used for boundary determination with sample allocation methods m1, m2 and m3; with method 4 both binary and real-valued encoding is used.

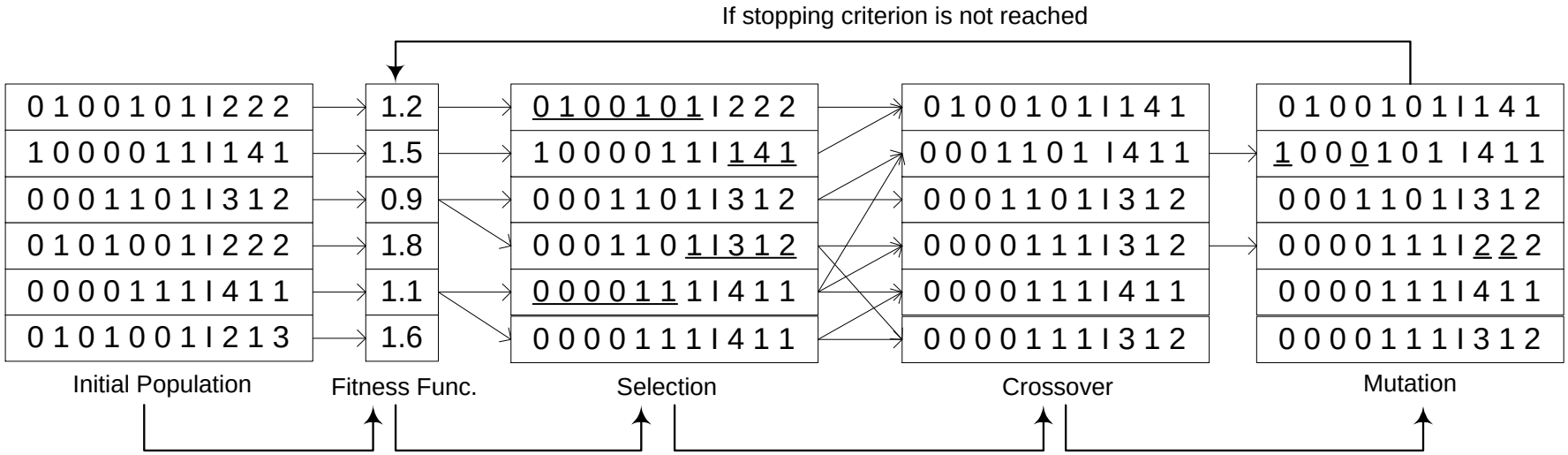
	1.2	2.0	3.2	3.8	4.0	4.9	5.2	5.3	5.8	6.0
Binary Encoding (m 1,2,3)	0	0	0	1	0	0	1	0	0	1

Binary & Real-Valued Encoding (m4)	0	0	0	1	0	0	1	0	0	1	2	2	3

1st Strata → 1.2, 2.0, 3.2, 3.8 (N=4)
2nd Strata → 4.0, 4.9, 5.2 (N=3)
3rd Strata → 5.3, 5.8, 6.0 (N=3)

Boundaries → 3.8, 5.2, 6.0

After constructing the initial generation, each chromosome is evaluated by an objective function, referred as a fitness function, from which a fitness value is derived. In our algorithm the fitness value is the variance of the estimator in stratified sampling. In the first and fourth combined problems where equal and GA sample allocation methods are utilized, a penalty function is defined relating to each stratum's sample sizes in order to avoid unfeasible solutions obtained whenever the sample size of each stratum is greater than the size of that stratum. Selection (roulette wheel selection) determines whether chromosomes will survive in the next generation or not, according to their fitness values.



Selection determines whether chromosomes will survive in the next generation or not, according to their fitness values. Chromosomes with a better fitness value have more chance to survive than the weaker ones. This replicates nature in that fitter individuals will tend to have a better probability of survival and will go forward. Weaker individuals are not without a chance. In nature such individuals may have genetic coding that may prove useful to future generations. There are several widely used methods of selection. In this paper, one of the most popular methods, roulette wheel selection is used.

In this method, the roulette wheel is divided into parts according to the chromosomes' fitness values.

Crossover, one of the main operators of any GA, provides exchange of individual characteristics between chromosomes. In this paper, according to the size of the examples, single-point, 2-point or multi-point crossover methods are used.

Individual 1	0	0	1	0	0	0	1	0	0	1
Individual 2	0	1	0	0	1	0	0	0	0	1

Offspring 1	0	0	1	0	1	0	0	0	0	1
Offspring 2	0	1	0	0	0	0	1	0	0	1

Individual 1	0	0	1	0	0	0	1	0	0	1	**2**	**3**	**2**
Individual 2	0	1	0	0	1	0	0	0	0	1	**1**	**2**	**4**

Offspring 1	0	0	1	0	1	0	0	0	0	1	**1**	**2**	**4**
Offspring 2	0	1	0	0	0	0	1	0	0	1	**2**	**3**	**2**

After crossover, random exchange mutation is applied so that two positions are selected at random along the chromosome and the genes contained in these positions are exchanged. The reason for using this mutation operator is to guarantee the number of strata be held fixed after mutation.

Indiv.	0	0	1	0	0	0	1	0	0	1	2	3	2

Mut. Indv	0	0	1	0	1	0	0	0	0	1	2	2	3

This GA process is repeated until a predetermined number of iterations is achieved. As a heuristic method convergence to exact minimum is not guarenteed, yet emprical evidence suggests that GAs are a robust way to find “near-optimal” solutions.

Geometric Method

Gunning & Horgan's Geometric Method is based on the principle of making the coefficients of the variation (CV) among strata equal in order to find the break points of the stratification variable.

Recall: CV $\frac{\sigma}{\mu}$

In Surveys: $\frac{S}{\bar{X}}$

- Each Stratum has a uniform distribution
- The CVs are equal in each stratum

Standard deviation for a uniform distribution = $\frac{1}{\sqrt{12}}(b-a)$

Mean of uniform distribution $\frac{a+b}{2}$

Boundary Notation

L = number of strata

$k_0 = a$ minimum value

$k_L = ar^L$ maximum value

k_1, k_2, k_3 etc. = intermediate boundaries

k_h = stratum *h* boundary

So Coefficient of Variation (CV) for stratum h= $\frac{\frac{1}{\sqrt{12}}(k_h - k_{h-1})}{\frac{k_h + k_{h-1}}{2}}$

With equal Coefficient of Variations among strata one can get:

$$\frac{k_{h+1}-k_h}{k_{h+1}+k_h}=\frac{k_h-k_{h-1}}{k_h+k_{h-1}} \Longrightarrow k_h^2 = k_{h-1}k_{h+1}$$

The Stratum Breaks

k_0 k_1 k_2 k_3 k_4 k_5

a ar ar^2 ar^3 ar^4 ar^5 *Geometric Progression*

Calculation of ratio r

$$k_L = k_0 r^L$$

$$r = \left(\frac{k_L}{k_0}\right)^{1/L} = \left(\frac{max}{min}\right)^{\frac{1}{strata}}$$

$$k_h = ar^h = k_0 \left(k_L/k_0\right)^{h/L}$$

Geometric method is a very simple method of boundary determination having some drawbacks:

This method highly depends on the assumption that the distributions within strata are uniform.

This method will not work well for normal distributions.

Also since the breaks increase geometrically, it will not work well with variables that have very low starting points: this will lead to too many small strata which will result in finding strata boundaries where the sample sizes are greater than the size of the strata or stratifying a population in a way that some strata contain very few or no elements at all, eventuating in an unavailable value of the variance of the estimator .

However these drawbacks can be avoided by implementing various modifications. After these modifications, we implemented both of the methods to several real life data.

Numerical Applications

Data of numerical application:

- The first two examples (iso2004 and iso2005) consists of the net sales data of 487 and 485 Turkish manufacturing firms from the first 500 largest corporations belonging to Istanbul Chamber of Industry (ICI) in year 2004 and 2005 respectively. (N=487);
- The rest of the examples are obtained from Horgan's paper.

Each of the examples is divided into 2, 3, 4, 5 and 6 strata. The total sample size is 80 for iso2004, whereas 100 for the other examples.

The skewness of the data can be summarized as follows:

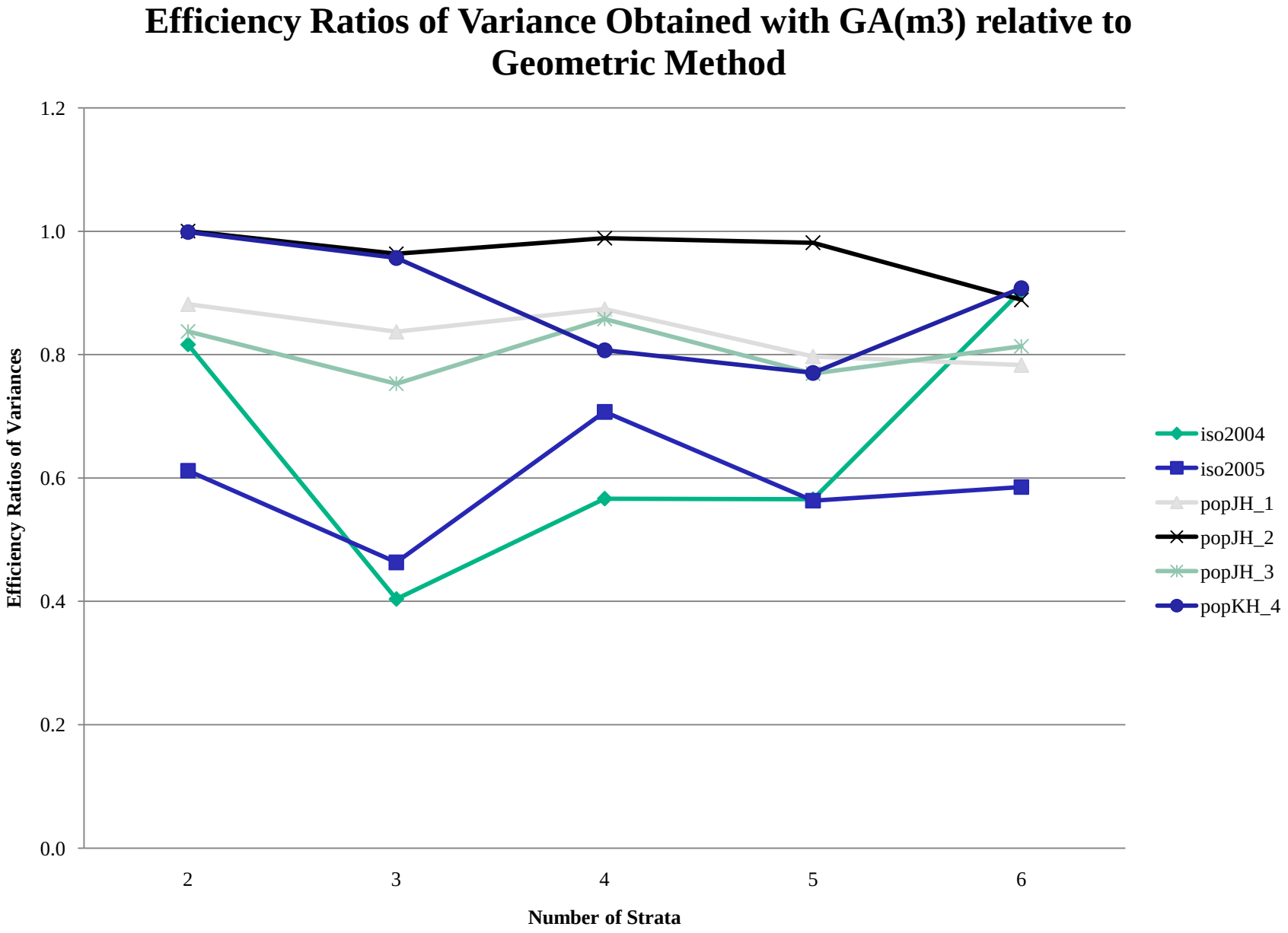
Data	Skewness
iso2004	10.059
iso2005	12.672
popJH_1	6.440
popJH_2	2.872
popJH_3	2.457
popJH_4	2.076

Efficiency Ratios of Variance of the Estimator Obtained with GA & Geometric Method

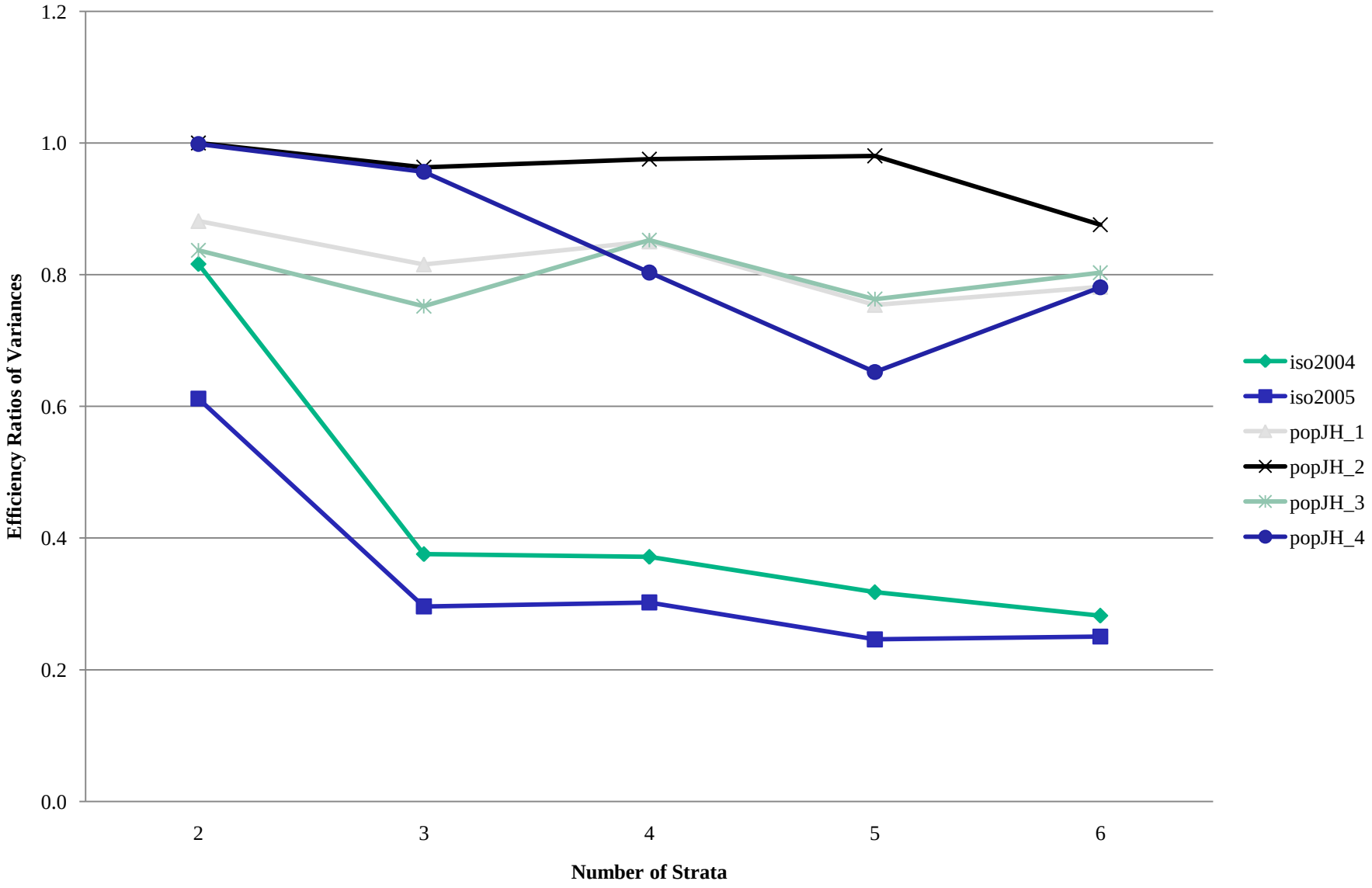
		GA(m3) / Geom	GA(m4) / Geom
iso2004	H=2	0.81625	0.81625
	H=3	0.40375	0.37560
	H=4	0.56646	0.37144
	H=5	0.56546	0.31775
	H=6	0.92885	0.38878
iso2005	H=2	0.61187	0.61187
	H=3	0.46313	0.29616
	H=4	0.70714	0.30223
	H=5	0.56305	0.24605
	H=6	0.58817	0.28262

		GA(m3) / Geom	GA(m4) / Geom
popJH_1	H=2	0.88172	0.88172
	H=3	0.83703	0.81578
	H=4	0.87383	0.85068
	H=5	0.79699	0.75406
	H=6	1.10000	0.78161
popJH_2	H=2	1.00008	1.00000
	H=3	0.96331	0.96331
	H=4	0.98880	0.97565
	H=5	0.98141	0.98069
	H=6	0.88874	0.87592

		GA(m3) / Geom	GA(m4)/ Geom
popJH_3	H=2	0.83754	0.83720
	H=3	0.75285	0.75234
	H=4	0.85767	0.85248
	H=5	0.76931	0.76292
	H=6	0.81342	0.80304
popJH_4	H=2	0.99864	0.99864
	H=3	0.95662	0.95625
	H=4	0.80697	0.80341
	H=5	0.77029	0.65227
	H=6	0.90768	0.78090


Efficiency Ratios of Variance Obtained with GA(m3) relative to Geometric Method
Efficiency Ratios of Variances
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
Number of Strata
iso2004
iso2005
popJH_1
popJH_2
popJH_3
popKH_4

Efficiency Ratios of Variance Obtained with GA(m4) relative to Geometric Method



Interpretation of the Results

The smallest variance results for all of the numerical examples are obtained when both stratum boundaries and strata sample sizes are determined with GA. This confirms that GA can be efficiently utilized in the stratification of heterogeneous populations.

Because of the assumption and constraints, the applicability of geometric method is respectively limited. On the contrary, GA can be easily applied in stratifying populations with a wide variety of characteristics due to the inherent flexibility of the approach. Besides, even for those populations that geometric method can be applied to, GA evidently yields reasonably more efficient solutions.

All the desirable features aside, it should be noted that GA is a computer intensive method, which makes it more complex especially compared to the simplicity of geometric method.

Further Research Topics

Our GA approach is proposed in the context of a fixed cost with a predetermined number of strata and sample size. Future research might develop the GA approach where factors such as sample cost, the number of strata and the sample size vary.

Since Geometric Method is a simple way of boundary determination, the results of the Geometric Method can be utilized as the initial population of GA in order to reduce the iteration time and make it converge better to the optimum.

Thanks for listening....

Timur Keskinturk
Sebnem ER
Prof. Jane Horgan