Numerical Algorithms for Stratification of Skewed Populations

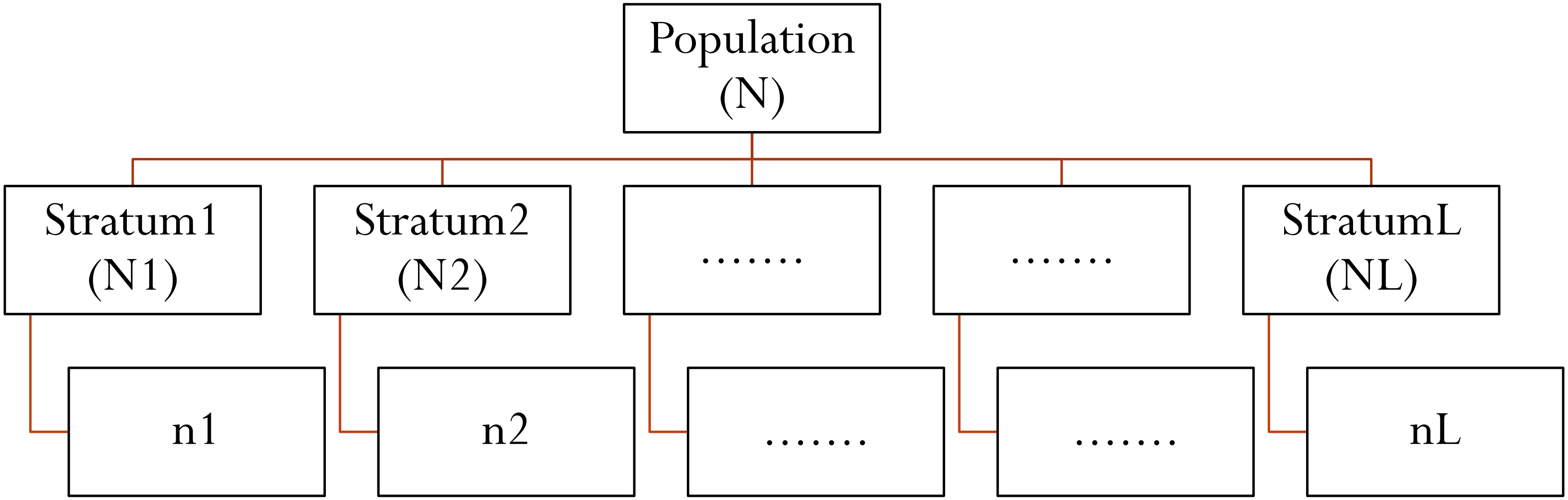
Şebnem Er & Jane M. Horgan

Dublin City University, Ireland

Applied Statistics 2010
September 19 - 22, 2010
Bled, Slovenia

Introduction

- **"Stratified Sampling"** is a sampling method where a heterogeneous population is divided into subpopulations, each of which is internally homogeneous in order to gain more precision than other methods of sampling.



Population (N)

Stratum1 (N1)

Stratum2 (N2)

…….

…….

StratumL (NL)

n1

n2

…….

…….

nL

- The main problem arising in stratified sampling is to obtain optimum boundaries that either
  - **minimise the estimated variance** (maximise the level of precision)
  - or
  - **minimise the total sample size given a level of precision**.

Determination of Stratum Boundaries

- An Exact Solution by Dalenius (1950)

Dalenius showed that when the boundaries satisfy the equations

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

then the variance of the estimate with Neyman allocation method is minimum among all the possible boundary combinations.

Since the Dalenius equations are computationally difficult to solve, many approximations have arisen.

1959: Dalenius & Hodges’ Cumulative Square Root Method

- Dalenius and Hodges’ cum$\sqrt{f}$ rule is based on constructing equal intervals on the cumulative of the square roots of the frequencies of the stratification variable so that nearly optimum points are obtained.

1988: Lavallée & Hidiroglou's Algorithm

- Lavallée & Hidiroglou (1988) derived an iterative algorithm for skewed populations, such that the sample size is minimised for a given level of precision such as the CV equal to a specified level in between 1% and 10% considering a take-all top stratum.

2004: Gunning & Horgan's Geometric Method

- More recently Gunning & Horgan (2004) developed the geometric method which is a combination of the assumptions of Dalenius and Hodges' (1959) cum√f rule of uniform distribution within each stratum and Lavallée-Hidiroglou's method of equal coefficients of variation in each stratum.

- To find the bh, the geometric method is to set the coefficients of variation (CV_h) equal such that the stratum boundaries are the terms of a geometric progression.

2004: Kozak's Random Search Method

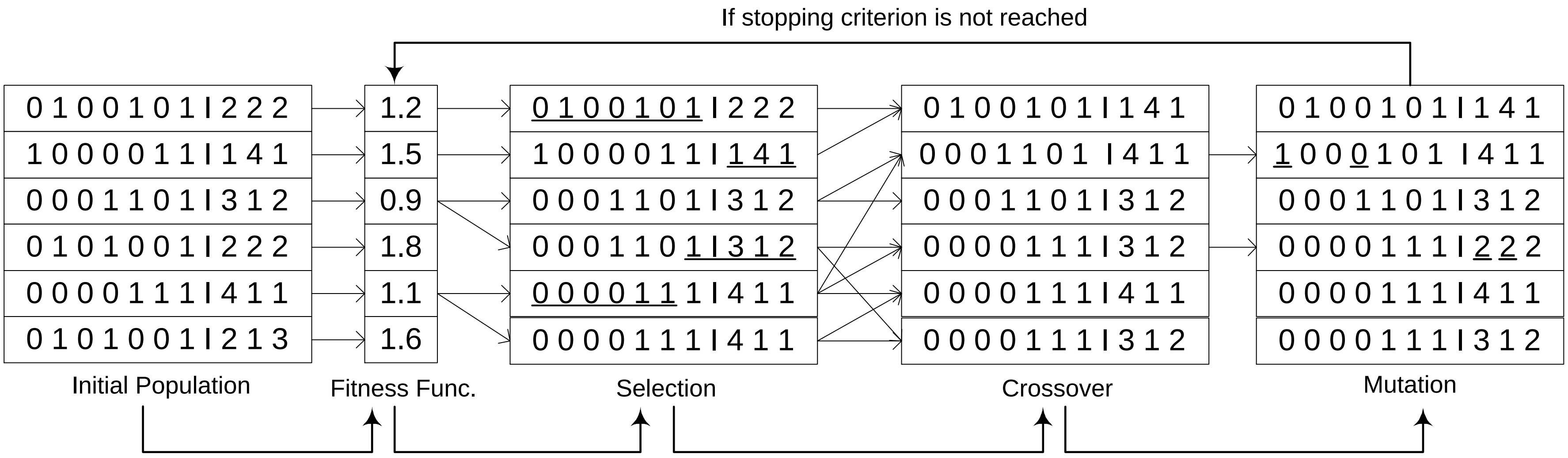
- Kozak’s (2004) Random Search Method chooses a stratum boundary at random and at each iteration makes a random modification on this boundary by an integer according to the size of the population.
- Then the sample size for the new set of boundaries is calculated and if it is smaller, the new boundaries are accepted.
- This is repeated until there is no change in sample size or the number of iterations is reached.

2007: Keskinturk & Er's Genetic Algorithm Method

In order to solve the stratification problem with GA, values of the stratification variable are encoded into chromosomes. In this method binary encoding is used for boundary determination; with GAmethod of allocation 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 & Real-Valued Encoding (mGA)	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**2**	**2**	**3**



If stopping criterion is not reached

Initial Population	Fitness Func.	Selection	Crossover	Mutation
0 1 0 0 1 0 1 I 2 2 2	1.2	0 1 0 0 1 0 1 I 2 2 2	0 1 0 0 1 0 1 I 1 4 1	0 1 0 0 1 0 1 I 1 4 1
1 0 0 0 0 1 1 I 1 4 1	1.5	1 0 0 0 0 1 1 I 1 4 1	0 0 0 1 1 0 1 I 4 1 1	1 0 0 0 1 0 1 I 4 1 1
0 0 0 1 1 0 1 I 3 1 2	0.9	0 0 0 1 1 0 1 I 3 1 2	0 0 0 1 1 0 1 I 3 1 2	0 0 0 1 1 0 1 I 3 1 2
0 1 0 1 0 0 1 I 2 2 2	1.8	0 0 0 1 1 0 1 I 3 1 2	0 0 0 0 1 1 1 I 3 1 2	0 0 0 0 1 1 1 I 2 2 2
0 0 0 0 1 1 1 I 4 1 1	1.1	0 0 0 0 1 1 1 I 4 1 1	0 0 0 0 1 1 1 I 4 1 1	0 0 0 0 1 1 1 I 4 1 1
0 1 0 1 0 0 1 I 2 1 3	1.6	0 0 0 0 1 1 1 I 4 1 1	0 0 0 0 1 1 1 I 3 1 2	0 0 0 0 1 1 1 I 3 1 2

Others

- Brito and Maculan’s Local Search.
- Khan and Ahmad’s Dynamic Programming.
- …

The Aim of the Research

- to make the efficiency comparisons between the numerical iterative algorithms relative to each other

Numerical Applications

Data of numerical application:

- All the data to be stratified using these methods are obtained from the package **R**.
- There are 9 populations
  - Cochran - Horgan Data
    - An accounting population of debtors in an Irish firm (Debtors).
    - The population (in thousands) of UScities
    - The number of students in four-year UScolleges
    - The resources in millions of dollars of a large commercial bank in the US

Numerical Applications

Data of numerical application:

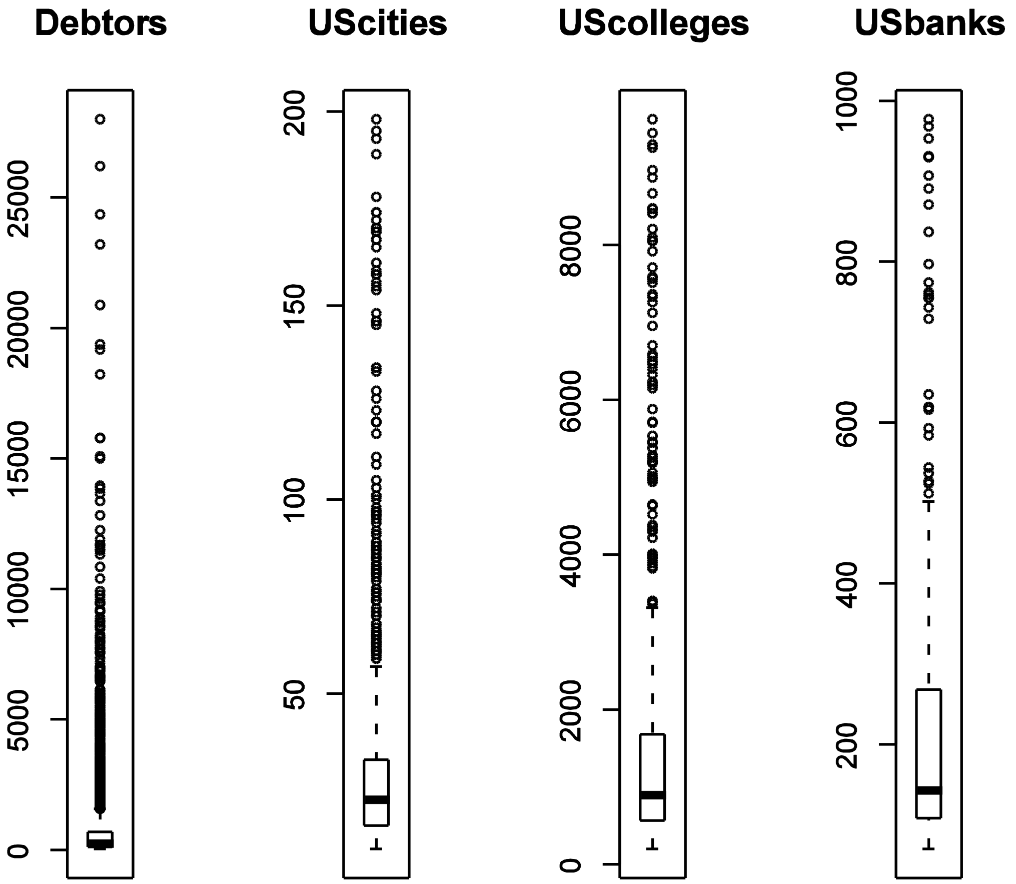
- Swedish Data
  - Number of municipal employees in 1984 in the 284 municipalities in Sweden (me84)
  - 1975 population (in thousands) in the 284 municipalities in Sweden (P75)
  - Real estate values according to 1984 assessment (in millions of kronor) in the 284 municipalities in Sweden (REV84)

Numerical Applications

Data of numerical application:

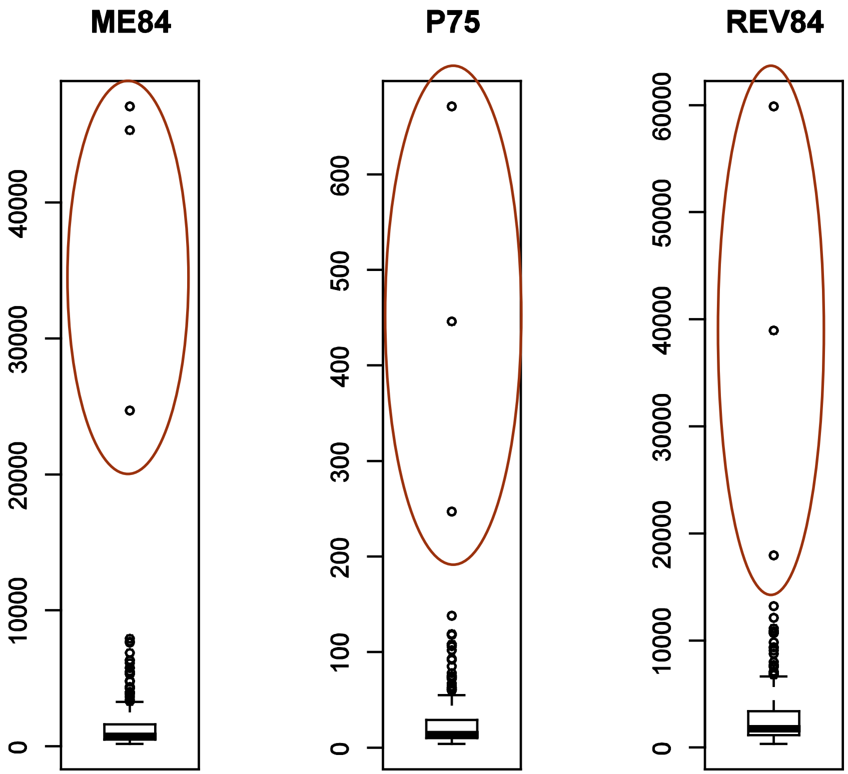
- Canadian Data
    - Simulated Data from the Monthly Retail Trade Survey of Statistics Canada (MRTS)
    - Household income before taxes from the 2001 Survey of Household Spending carried out by Statistics Canada (HHINCTOT)
- and each of them is divided into 3, 4, 5 and 6 strata. The total sample size is 100.

Cochran - Horgan Data



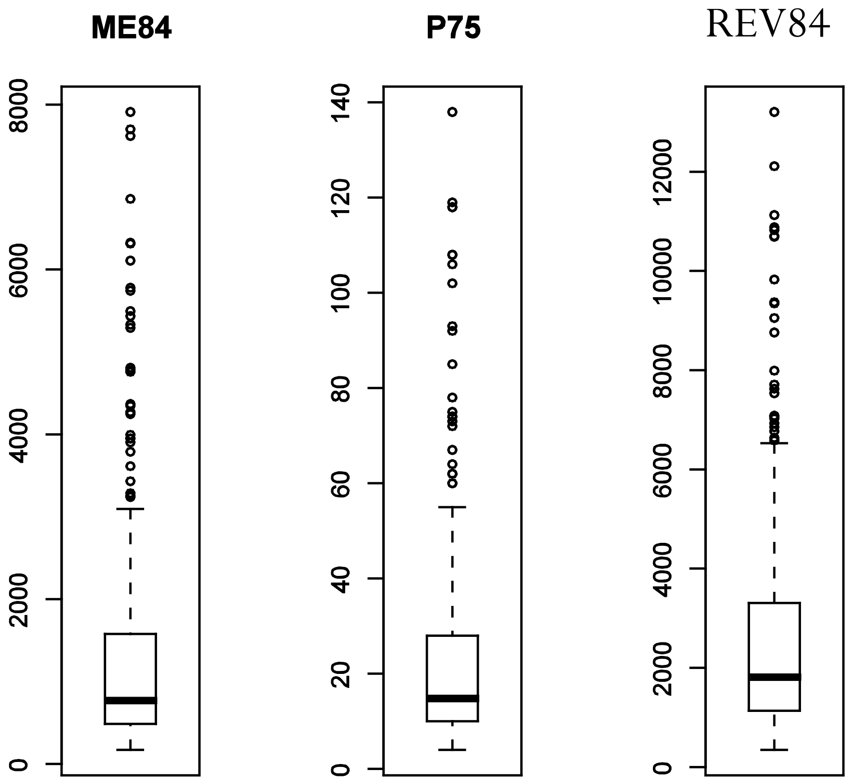
	N	Range	Skew	Kurt
Debtors	3369	40-28000	6.44	59.00
Uscities	1038	10-198	2.87	9.12
Uscolleges	677	200-9623	2.45	5.80
USbanks	357	70-977	2.07	4.06

Swedish Data



	N	Range	Skew	Kurt
ME84	284	173-47074	8.64	84.04
P75	284	4-671	8.43	88.56
REV84	284	347-59877	7.83	81.33

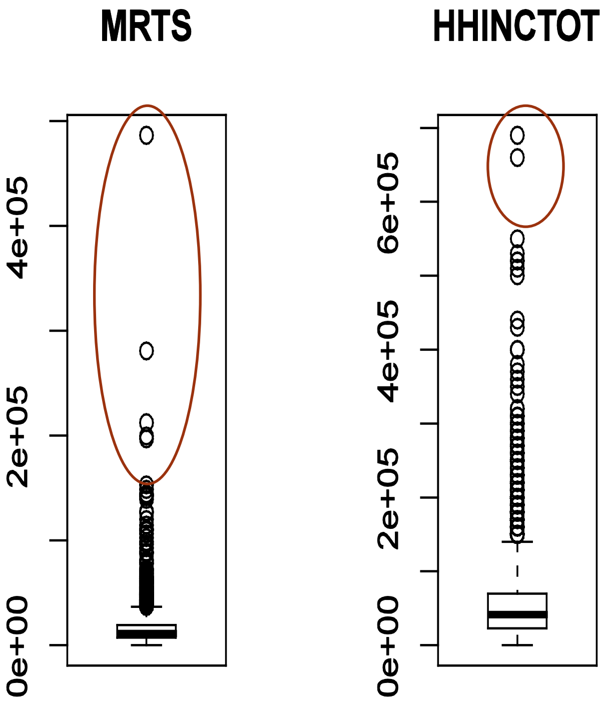
Without Extreme Outliers

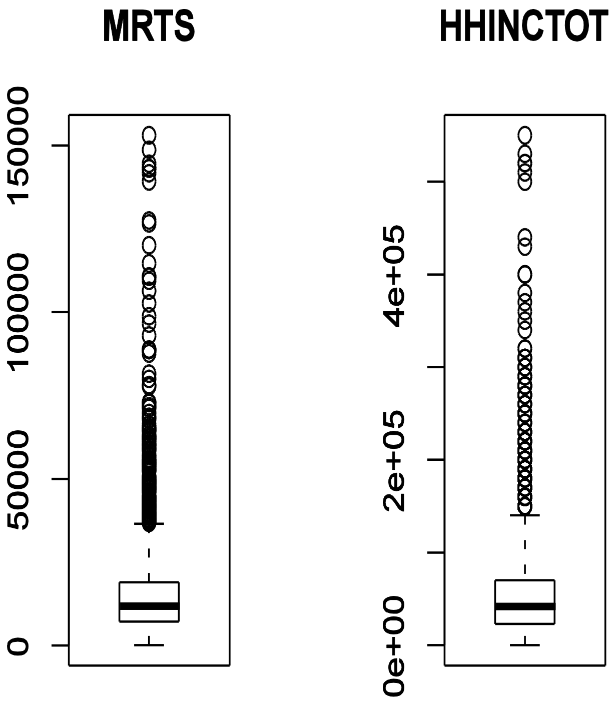


N	Range	Skew	Kurt
281	173-7910	2.23	5.03
281	4-138	2.27	5.63
281	347-13205	1.87	3.48

Canadian Data

Without Extreme Outliers



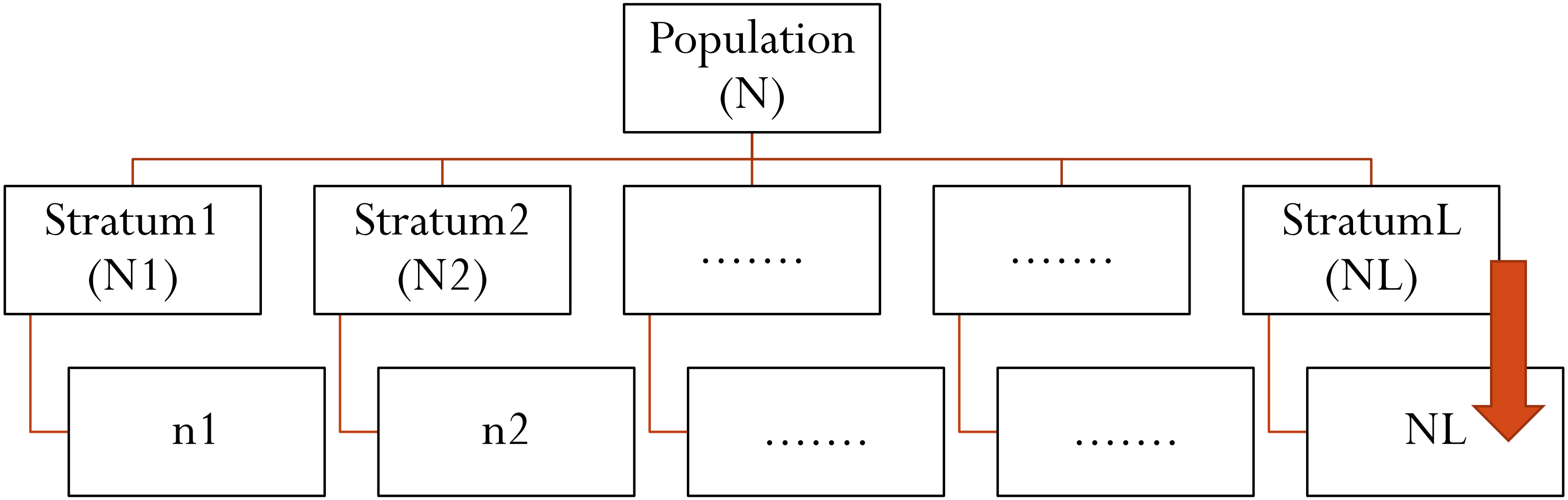


	N	Range	Skew	Kurt	N	Range	Skew	Kurt
MRTS	2000	141-486366	8.61	136.2	1995	141-153008	3.72	20.21
HHINCTOT	16025	100-690000	2.71	18.79	16023	11-550000	2.38	13.12

Efficiency Comparisons Between the Numerical Iterative Algorithms

1. GA / RS
2. RS / LH
3. GA / LH

- Results that give a non-takeall top stratum therefore the results shouldn’t be compared with LH takeall method. Only RS and GA could be compared.



Efficiency Ratios Between The Computational Methods for Horgan-Cochran Data

Data	L	RS/LH	GA/RS	GA/LH
Debtors	3	-	1	-
	4	-	1	-
	5	-	1	-
	6	-	0.986	-
Uscities	3	-	1	-
	4	-	1	-
	5	-	0.736	-
	6	-	1.017	-
Uscolleges	3	-	1	-
	4	-	1	-
	5	-	0.875	-
	6	-	0.750	-
Usbanks	3	-	1	-
	4	1	1	1
	5	0.619	1	0.619
	6	1	1.016	1.018

Efficiency Ratios Between The Computational Methods for Swedish Data

		With Extreme Outliers			Without Extreme Outliers		
Data	L	RS/LH	GA/RS	GA/LH	RS/LH	GA/RS	GA/LH
P75	3	0.928	1	0.928	0.938	1	0.938
	4	0.818	1	0.818	0.820	1	0.820
	5	0.885	1.371	1.213	0.867	1.077	0.934
	6	0.844	1.273	1.075	0.790	1.314	1.038
ME84	3	1	1	1	1	1	1
	4	1	1	1	1	1	1
	5	1	1	1	1	1	1
	6	0.987	1.251	1.235	1	1.052	1.049
REV84	3	1	1	1	1.010	1	1
	4	1	1	1	1.014	1	1.014
	5	1	1.017	1	1	1	1
	6	0.951	0.976	0.928	1	1	1.013

Efficiency Ratios Between The Computational Methods for Canadian Data

		With Outliers			Without Outliers		
Data	L	Kozak/LH	GA/Kozak	GA/LH	Kozak/LH	GA/Kozak	GA/LH
MRTS	3	-	1	-	-	1	-
	4	-	1	-	-	1.055	-
	5	-	1	-	-	1.139	-
	6	-	1	-	-	1	-
HHINCTOT	3	-	1	-	-	1	-
	4	-	1	-	-	1	-
	5	-	1	-	-	1	-
	6	-	1	-	-	1	-

Conclusion

1. The Iterative Algorithms more or less come to the same result.
2. They are computer intensive.
3. LH and RS are publicly available in the R stratification package.
4. GA gets into local minimum when the strata size increases which means the problem gets more complicated to solve.
5. GA is not publicly available yet. GA codes are being written in R.
6. Take-all top stratum is not a parameter in GA before starting the algorithm.

Thank you very much for your interest….

Sebnem Er

er.sebnem@gmail.com