ISSN 2347-1921

Using Particle Swarm Optimization to Determine the Optimal Strata Boundaries

Mowafaq Muhammed AL-kassab, **Ammar Ahmed Ali**

Department of Statistics & Information, College of Mathematics and Computer Sciences, University of Mosul, Mosul, Iraq

ABSTRACT

Stratified random sampling is a commonly used sampling methodology especially for heterogeneous populations with outliers. Stratified sampling is preferably employed due to its capability of improving statistical precision by yielding a smaller variance of the estimator, compared with simple random sampling. In order to reduce the variance of the estimator in stratified sampling, the problems of stratum boundary determination and sample allocation must be resolved initially. This paper proposes a PSO algorithm to solving the problem of stratum boundary determination in heterogeneous populations while distributing the sample size according to Neyman allocation method. The PSO algorithm is tested on two groups of populations and a comparative study with Kozak, GA and Delanius and Hodges methods have been implemented. The numerical results show the ability of the proposed algorithm to find the optimal stratified boundaries for a set of standard populations and various standard test functions compared with other algorithms.

Keywords

Stratified random sampling; Particle Swarm Optimization; Optimal Strata Boundaries; Neyman Allocation.

1.INTRODUCTION

A common procedure in sampling surveys is partitioning the elements of a population, before distributing the sample on it, in such a way to obtain most useful information from the data to be collected. This procedure is called stratification. It may have different aims, such as to guarantee obtaining information for some or all the geopolitical regions of a country, or to provide more precision in estimating population quantities by identifying strata with more homogeneous elements into them, according to one or more variables [3].

A principal use of stratification, in order to obtain a better precision, is in defining what percentage of the sample must be taken from each stratum once we have chosen a non-uniform allocation scheme, that is, a non-trivial functional relation between the size of each stratum and the number of sample units to be collected in it. Thus, it is important to consider the allocation scheme itself in order to do a suitable stratification [6].

Several numerical and computational methods have been developed for obtaining the optimum boundaries in stratified sampling. 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 [8]. More recently Lavallée & Hidiroglou algorithm [13] and Gunning & Horgan's (2004) geometric method[9] have been proposed for highly skewed populations whereas Kozak's (2004) random search method [12] and Keskinturk & Er's (2007) genetic algorithm (GA) method [11] have been proposed for even non-skewed populations.

This study presents the PSO algorithm for the determination of stratum boundaries. In order to explore the efficiency of PSO algorithm, we compare its efficiency with Kozak, GA and Delanius and Hodges methods

The rest of the paper is organized as follows. Section 2 describes stratified random sampling. In section 3, Background of PSO and previous work are summarized. PSO model for optimal stratum boundaries is also discussed. In order to test the efficiency of the proposed PSO, a comparative study with Kozak, GA and Delanius and Hodges methods is performed in Section 4. Conclusions and future research are drawnin Section 5.

2. Stratified random sampling

There are several alternative methods such as equal, proportional [7], and Neyman allocation [14]. The equal allocation method is the simplest method where each stratum sample size is the same. With the proportional allocation method, the sample sizes in each stratum are proportional to the size of that stratum. These two methods are efficient and suitable if the variances withinthe stratum are similar. On the other hand, if the stratum variances differ substantially, asin for example highly skewed populations, the Neyman allocation method should be used. This method is based onthe principle of sampling fewer elements from homogeneous strata and more elements from strata with high internal variability. In this study, distributing the sample size according to Neyman allocation method, and sampling costs are assumed to be equal for all strata.

In this paper, each character expresses the value as follows. *Y:*stratification variable; *N:*population size; *n:* sample size; *L:* number of strata; N_h: number of elements in stratum $h(h = 1, . . . , L)$;n_h: sample size in stratum h;$V(\bar{y}_{st})$:variance of the elements in stratum h; Y_h: mean of elements in stratum h; $\bar{y}_{st}$ estimated mean in stratified sampling.

In stratified sampling [6], a population with N units is divided into L groups

with N1,N2, ...,Ni, ...,NH units respectively. These groups are called strata. There is no overlap among them and together they exhaust the population. Thus, we have

$$N_1 + N_2 + ... + N_h + ... + N_L = N \qquad ...(1)$$

After the strata definition, which is based on characteristics of the population, sampling units are selected in each stratum, independently, according to specific criteria of selection. The sample sizes of the strata are denoted by $n_1, n_2, ..., n_h, ..., n_L$, respectively. The size of the sample taken from the population and symbolized by the n. Thus

$$\sum_{h=1}^{L} n_h = n \qquad ...... (2)$$

The mean of the stratum h, denoted by μ_h.

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

The mean of the sample taken from the stratum h, denoted by $\bar{y}_h$.

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

The variance of stratum h, denoted by ${\sigma^2}_h$.

$$\sigma^2{}_h = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{hi} - \mu_h)^2 \text{ .... (5)}$$

The variance of the sample taken from stratum h, denoted by $S^2{}_h$.

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

The weight of stratum h denoted by W_h is:

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

It also can be obtained the population mean denoted by μ, by multiplying mean of stratum h by weight of stratum h:

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

If we multiplying mean of the sample from stratum h by weight of stratum h, we get the stratified mean denoted by$\bar{y}_{st}$:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad \text{......(9)}$$

Moreover, the variance of stratified sampling mean is:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W^2{}_h \frac{\sigma^2{}_h}{n_h} \qquad \text{......(10)}$$

When total sample size n is allocated using Neyman's optimum allocation method is:

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

The equation (11) is associated Neyman's allocation [14]. Replacing nh in (11) by (10), we have:

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

3. PSO Algorithm for Stratified Sampling

3.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population based stochastic optimization technique developed by Kennedy and Eberhart in 1995 [10], inspired by social behavior of bird flocking. It belongs to Swarm Intelligence (SI), which originates from the study of natural creatures living in a group. Each individual possess little or no wisdom, but by interacting with each other or the surrounding environment, they can perform very complex tasks as a group.

PSO could be explained well in an imagined scenario: a group of birds is flying in an area to look for food, and there's only one piece of food in this area. the easiest way to find the food is to follow the one who is closest to the food.

The basic concept of PSO lies in accelerating each particle toward its *pbest,* which was achieved so far by that particle, and the *gbest,* which is the best value obtained so far by any particle in the neighborhood of the particle, with a random weighted acceleration at each time step.

Each particle tries to modify its position using the following information [10]:

- The current positions (*X* (*t*)),
- The current velocities (*V* (*t*)),
- The distance between the *pbest* and the current position (P_L - *X* (*t*)),
- The distance between the *gbest* and the current position (P_G - *X* (*t*)).

In this paper, we will apply PSO algorithm to determine stratum boundaries of each stratum in stratified sampling.

3.2 Input Information

For stratum boundary determination using Neyman allocation, the software implemented takes into consideration the following parameters:

- Number of strata L.
- Population data D that represents the study population, or population function f(x) in the period [0, 1].

3.3 Fitness Function

Fitness function is a critical factor in the PSO method. Every particle in the PSO's population has a fitness value, and it moves in solution space with respect to its previous position where it has met the best fitness value. In this paper, the fitness value is the variance of Neyman allocation in stratified sampling denoted as Eq. (12) that must be minimized through the iteration process.

3.4 Particle Structure

The composition and shape of the particle in stratified sampling differs in the way of representation from the most representations found in the literature, which represented by a single vector structure. The range of ascending values subject to stratification must be divided into L parts by points $Y_1<Y_2<\cdot\ \cdot\ \cdot<Y_{L-1}$. Each such part corresponds to a stratum boundary. The length of particle equal to the number of the strata L. The first gene in particle refers to the sequence of last observation in the first stratum, so it refers to the size of first stratum N_1. The second gene refers to the sequence of the last observation in second stratum. The difference between the value of the first gene and the second gene refers to the size of second stratum N_2 and so on. Therefore, the gene value refers to the stratified boundaries for each stratum, and the difference between gene and previous gene refers to the size of stratum. Fig. 1 illustrate the particle representation of six strata boundaries for a population contains 30 observations.

4	8	14	18	27	30

Fig 1: Particle representation

3.5 Initial Population Creation

The size of the population (number of particles) and the way the initial population is created have a significant influence in the performance of the algorithm and to the quality of the result. Since each particle must contain a number of genes equal to the number of strata L, The last gene must have a value of "N" because it represents the upper limit of the last stratum. In general, the ideal situation would be to have the greatest possible diversity of particles to better through the search space.

3.6 Particles Movement

In the algorithm of PSO, each solution is called a "particle", and every particle has its position, velocity, and fitness value. At each iteration, every particle moves towards its personal best position and towards the best particle of the swarm found so far. The velocity changes according to formulation (13):

$$V_{i+1}^{n} = round(w * V_i^n + c_1 * r_1(P_L^n - x_i^n) + c_2 * r_2(P_g - x_i^n)) \qquad (13)$$

where *i* is the iteration sequence of the particle *n*, c_1 and c_2 are positive constant parameters called acceleration coefficients which are responsible for controlling the maximum step size, r_1 and r_2 are random numbers between (0, 1), *w* is a constant. and V_{i+1}^{n} is particle *n*'s velocity at iteration *i*+ 1. V_i^n is particle *n*'s velocity at iteration *i*. x_i^n is particle *n*'s position at iteration i. P_L^n is the historical individualbest position of the swarm. Finally, the new position of particle *n*, x_{i+1}^{n} , iscalculated as shown in (14). *The flowchart of which is shown in Fig 2.*

$$x_{i+1}^{n} = x_i^n + V_{i+1}^{n} \qquad (14)$$

ISSN 2347-1921

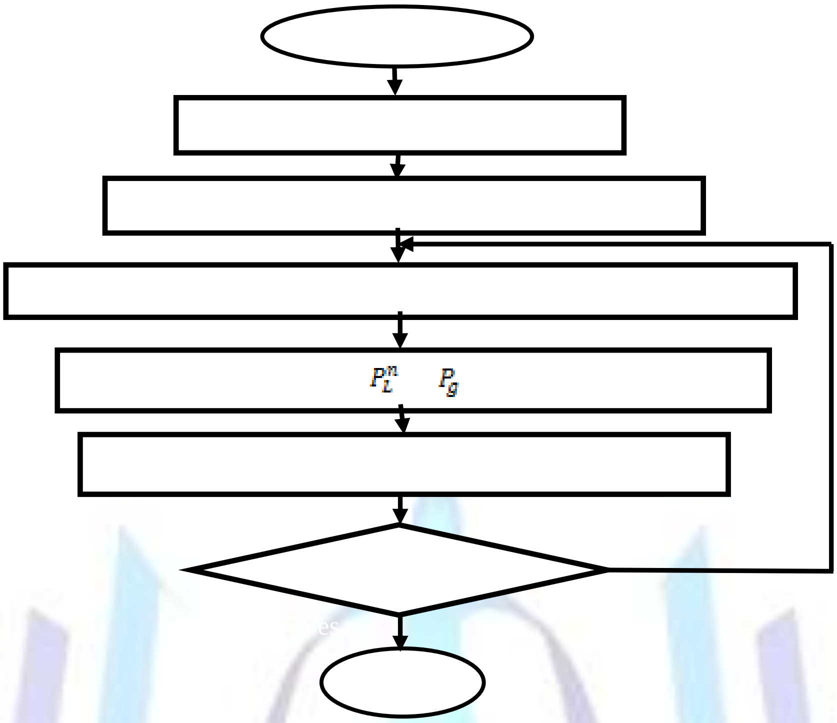

Fig 2: The flowchart of PSO

4. Numerical Results

The PSO experiments for the stratification sampling has been performed on two groups of populations: data and functions, to find optimal strata boundaries based on variance of Neyman allocation. All experiments are implemented using Matlab 8.1.0 (R2013a).

4.1 Testing PSO algorithm to find the stratified boundaries for populations of data

In this section, 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[15] and GA4Stratification[16] packages are used for stratification. Each of the populations are divided into 3, 4, 5 and 6 strata. The total sample size is 100 and the boundaries are obtained with Kozak and GA methods with random initial boundaries.

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

Pop2: Number of municipal employees of 284 municipalities in Sweden in 1984 (ME84).

Pop3: Simulated Data from the Monthly Retail Trade Survey of Statistics Canada (MRTS).

Pop4: Population in thousands of 284 municipalities in Sweden in 1975 (P75).

Pop5: Real estate values in millions of kronor according to 1984 assessment of 284 municipalities in Sweden in 1984 (REV84).

Pop6: The resources in millions of dollars of large com-mercial US banks (USbanks).

Pop7: The population in thousands of US cities in 1940 (UScities).

Pop8: The number of students in four-year US colleges in 1952-1953 (UScolleges).

Table 1. Summary Statistics of the Populations

H	PSO	GA	Kozak

Pop	Name	N	Range	Skewness	Kurtosis	Mean	StdDev.
Pop1	Debtors	3369	27960	6.44	59.00	838.64	1873.99
Pop2	ME84	284	46901	8.64	84.04	1779.07	4253.13
Pop3	MRTS	2000	486225	8.61	136.20	16882.8	21574.88
Pop4	P75	284	667	8.43	88.56	28.81	52.87
Pop5	REV84	284	59530	7.83	81.33	3088.09	4746.16
Pop6	Usbanks	357	907	2.07	4.06	225.62	190.46
Pop7	Uscities	1038	188	2.87	9.12	32.57	30.4
Pop8	Uscolleges	677	9423	2.45	5.80	1563	1799.06

In order to compare the efficiency of three methods, the variance of the estimator given in Eq. (12) is calculated. We implement our proposed algorithm using MATLAB programming language on a PC (CPU 3.00 GHz,3GB RAM). PSO parameter settings used for stratifying these examples are shown in Table 2. Table 3 summarizes the variances of the estimators obtained with PSO, GA, and Kozak's methods

Table 2. PSO parameter

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

Table 3. Variances of the estimators for stratification examples obtained with PSO, GA and Kozak's methods

Pop1 : Debtors			
3	**2467.5912**	2469.5	4090.9
4	**1359.2777**	1369.2	2291.7
5	**822.5217**	831.2	1269.5
6	**572.5113**	588.98	605.58
Pop2 : ME84			
3	**6506.1354**	36797	36797
4	**3115.9730**	34787	34787
5	**2117.1294**	35614	35614
6	**1555.0057**	24207	35577
Pop3 : MRTS			
3	**591721.6322**	593160	1039100
4	**310783.6279**	311190	311190
5	**204832.8365**	207070	207000
6	**148921.0486**	150780	150750
Pop4 : P75			
3	**1.8168**	5.3956	5.3956
4	**0.9076**	4.9031	4.9031
5	**0.5936**	4.0269	5.4344
6	**0.4321**	3.9140	5.3821
Pop5 : REV84			
3	**18231.5254**	47733	47733
4	**9296.0850**	46545	46545
5	**5590.6497**	34483	137400
6	**3815.2788**	31654	42403
Pop6 : Usbanks			
3	**33.5197**	36.850	36.850
4	**17.3272**	27.331	27.331
5	**11.0075**	20.370	20.370
6	**6.7485**	18.435	18.448
Pop7 : Uscities			
3	**0.891952**	0.917173	0.917173
4	**0.472761**	0.473657	0.873657
5	**0.264204**	0.266574	0.569189
6	**0.196972**	0.199325	0.274273
Pop8 : Uscolleges			
3	**2451.4876**	2469.7	2469.7
4	**1500.4899**	1539	1539
5	**928.9271**	1020.9	2763.70
6	**603.2371**	892.40	892.33

Whereas the sample sizes given in Table 4 in Appendix. The results obtained by the PSO algorithm are better than the ones observed by using GA and Kozak's methods.

4.2 Testing PSO algorithm to find the stratified boundaries for populations of functions

The proposed PSO is tested using three benchmark functions. For comparison, Delanius and Hodges [10] is also executed on these functions. Table 5 shows the details of test functions.

Table 5. Benchmark functions (f1-f3)

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

Table 6 to Table 8 list the comparison results of these 2 methods for 3 benchmark functions of 4 different strata.

Table 6. Comparison results of $f_1(x) = xe^{-x}$

H	Delanius and Hodgesmethod		PSO	
	strata boundaries	$V_{Ney}(\bar{y}_{st})$	strata boundaries	$V_{Ney}(\bar{y}_{st})$
2	2.36	**0.6389**	2.280	**0.6177**
3	1.54 3.26	**0.3069**	1.564 3.226	**0.2964**
4	1.20 2.27 3.94	**0.1817**	1.227 2.307 3.873	**0.1732**
5	1.01 1.82 2.86 4.49	**0.1192**	1.026 1.846 2.850 4.364	**0.1133**

Table 7. Comparison results of $f_2(x) = e^{-x}$

H	Delanius and Hodgesmethod		PSO	
	strata boundaries	$V_{Ney}(\bar{y}_{st})$	strata boundaries	$V_{Ney}(\bar{y}_{st})$
2	1.27	**0.2855**	1.260	**0.2835**
3	0.73 2.04	**0.1339**	0.763 2.022	**0.1321**
4	0.52 1.27 2.61	**0.0774**	0.558 1.322 2.581	**0.0761**
5	0.39 0.92 1.68 3.02	**0.0503**	0.453 1.025 1.798 3.073	**0.0495**

Table 8. Comparison results of $f_3(x) = 2\,(1 - x)$

H	Delanius and Hodgesmethod		PSO	
	strata boundaries	$V_{Ney}(\bar{y}_{st})$	strata boundaries	$V_{Ney}(\bar{y}_{st})$
2	0.35	**0.0152**	0.354	**0.2835**
3	0.23 0.50	**0.0069**	0.229 0.502	**0.1321**
4	0.18 0.37 0.62	**0.0039**	0.170 0.361 0.587	**0.0761**
5	0.12 0.25 0.40 0.62	**0.0026**	0.135 0.282 0.447 0.642	**0.0495**

5.Conclusions

ISSN 2347-1921

Stratified sampling is a sampling methodology used for heterogeneous populations in order to gain more precision than other methods of sampling. This paper proposes a PSO algorithm for finding the optimal stratified boundaries with Neyman allocation and its performance, is evaluated using different test problems. The numerical results show the efficiency and capabilities of PSO algorithm in finding the Optimal Strata Boundaries. Amazingly, its performance better than other methods such as Kozak, GA and Delanius and Hodges methods. This confirms that PSO can be efficiently utilized in the stratification of heterogeneous populations. Future research might use PSO algorithm where factors such as sample cost, the number of strata, and the sample size vary.

REFERENCES

[1] Al-Saffawi. S.Y.(1976), Applying Statistical Techniques Analysis and Estimating a certain Agricultural Production ", M.SC. Thesis, University of Baghdad, Iraq.

[2] Azhar Al-Hasow & Al-Kassab ,1996 " *Method to find Stratum Boundaries Using Neyman Allocation*" Master Thesis , University of Mosul , Iraq.

[3] Al-Kassab, MMT & Al-Taay, H, (1994)"*Approximately Optimal Stratification Using Neyman Allocation* ", J , of Tanmiat Al-Rafidain.

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

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

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

[7] Daghistani , THN , (1995) " *An Approximately Optimal Stratification Using Proportional Allocation* " Master Thesis , University of Mosul , Iraq.

[8] Dalenius,T.,Hodges, J.L.Jr. (1959)." *Minimum Variance Stratification*", Journal of the American Statistical Association, 54, 285, pp.88-101.

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

[10] Kennedy J., Eberhart R., and Shi Y.,(2001) "*Swarm Intelligence".* New York: Morgan Kaufmann.

[11] Keskintürk, T., Er, Ş., 2007 "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.

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

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

[14] Neyman, Jerzy. 1934, “On the Two Different Aspects of the Representative Methods :The Method of Stratified Sampling and the Methodof Purposive Selection”, Journal of the Royal Statistical Society, 97 (4), pp.558-625.

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

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

APPENDIX

ISSN 2347-1921

Table 4. Size of the strata (N_h) obtained from PSO, GA and Kozak's methods

H		PSO	GA	Kozak
Pop1 : Debtors				
3	Nh	2740 506 123	2690 545 134	2673 561 135
4	Nh	2079 910 311 69	2085 901 302 81	2071 914 303 81
5	Nh	1917 945 334 138 35	1892 955 339 136 47	1892 954 335 139 49
6	Nh	1620 972 405 237 106 26	1604 956 426 221 118 44	1533 905 493 265 126 47
Pop2 : ME84				
3	Nh	227 54 3	145 78 61	145 78 61
4	Nh	163 85 33 3	115 64 44 61	115 64 44 61
5	Nh	146 77 33 25 3	54 69 56 41 64	54 69 56 41 64
6	Nh	116 66 59 24 16 3	54 69 56 41 19 45	54 61 33 34 37 65
Pop3 : MRTS				
3	Nh	1348 576 76	1227 671 102	1204 688 108
4	Nh	1023 744 204 29	1023 742 203 32	1017 748 303 32
5	Nh	786 701 345 140 28	749 698 371 150 32	774 675 369 105 32
6	Nh	521 593 523 235 104 24	521 573 455 283 136 32	513 580 458 281 136 32
Pop4 : P75				
3	Nh	230 51 3	150 77 57	150 77 57
4	Nh	180 77 24 3	111 73 43 57	111 73 43 57
5	Nh	155 81 34 11 3	123 61 33 19 48	64 68 52 34 66
6	Nh	111 73 52 28 17 3	45 87 52 33 18 49	45 66 39 34 33 67

Table 4 (Continues). Size of the strata (N_h) obtained from PSO, GA and Kozak's methods

H		PSO	GA	Kozak
Pop5 : REV84				
3	Nh	215 67 2	138 81 65	138 81 65
4	Nh	158 88 36 2	64 81 69 70	64 81 69 70
5	Nh	145 87 37 13 3	64 74 53 39 54	61 69 51 34 69
6	Nh	130 76 40 23 13 2	61 60 42 43 26 52	57 51 37 42 28 69
Pop6 : Usbanks				
3	Nh	258 75 24	212 84 61	212 84 61
4	Nh	212 84 73 18	111 112 73 61	111 112 73 61
5	Nh	111 112 74 42 18	110 101 54 32 60	110 101 54 32 60
6	Nh	110 101 54 36 38 18	54 68 90 53 32 60	51 63 97 54 32 60
Pop7 : Uscities				
3	Nh	795 192 51	749 193 96	749 193 96
4	Nh	434 412 153 39	434 409 155 40	434 356 154 94
5	Nh	393 382 135 91 37	393 367 150 89 39	226 271 298 149 94
6	Nh	226 271 285 128 91 37	274 263 245 128 89 39	226 271 285 128 89 39
Pop8 : Uscolleges				
3	Nh	481 135 61	478 130 69	478 130 69
4	Nh	272 231 113 61	256 234 118 69	256 234 118 69
5	Nh	272 225 111 34 35	253 221 82 60 61	192 166 145 105 69
6	Nh	255 219 81 53 34 35	132 180 166 78 52 69	133 179 166 77 53 69