Geometric Stratification Revisited

Horgan, Jane M.
Dublin City University, School of Computing
Glasnevin
Dublin 20, Ireland
E-mail: jhorgan@computing.dcu.ie

ABSTRACT Geometric stratification is an absurdly simple way of stratifying skewed populations, taking the boundaries in geometric progression. Implementation difficulties have recently been highlighted, giving rise to unfeasible solutions, in particular with strata which are too small or even empty. In this paper we suggest a modification, adding empirical rules for determining end points, outliers, take-none and take-all strata in order to improve the efficiency and ensure a feasible set of boundaries.

1 Introduction

A stratified sample design partitions a population into H mutually exclusive groups called *strata*. The population mean is

$$\overline{X} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} X_{hi}, \tag{1}$$

where X_{hi} is the i^{th} unit in the h^{th} stratum which contains N_h units ($h = 1, 2, \cdots, H$) units, and $N = \sum_{h=1}^{H} N_h$ is the total population size.

From each stratum a simple random sample of size $n_h \leq N_h$ is drawn without replacement. The total sample size is the sum $n = \sum_{h=1}^{H} n_h$ of the units selected from each stratum.

The mean of the sample selected from stratum h is

$$\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}, \tag{2}$$

where x_{hi} is the i^{th} unit selected from the h^{th} stratum. The overall stratified sample mean is

$$\overline{x}_{\text{strat}} = \sum_{h=1}^{H} W_h \overline{x}_h, \tag{3}$$

where $W_h = N_h/N$ is the weight of stratum h. It is easy to show (Cochran, 1977) that (3) is an unbiased estimator of the population mean $\overline{X}$, with variance

$$V(\overline{x}_{\text{strat}}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}. \tag{4}$$

The objectives of stratification include choosing the sample sizes in each stratum (n_h) and the boundaries ($k_1, k_2 \cdots, k_{H-1}$) to minimise (4). We denote the minimum value in the population by k_0 and the maximum as k_H.

Neyman (1934) showed that for a given sample size n, (4) is minimised when the n_h satisfy

$$n_h = \frac{nW_hS_h}{\sum_{i=1}^{H} W_iS_i}. \tag{5}$$

Dalenius (1950) showed that (4) is minimised when the stratum boundaries k_h satisfy

$$\frac{S_h^2 + (k_h - \overline{X}_h)^2}{S_h} = \frac{S_{h+1}^2 + (k_h - \overline{X}_{h+1})^2}{S_{h+1}}, \quad 0 \leq h \leq H-1, \tag{6}$$

However, these equations are ill adapted to practical computations because $\overline{X}_h$ and S_h depend on k_h. To this day, they remain intractable, and the best that can be done is to obtain approximations to (6) or iterative, computational algorithms to approach a solution which minimises the variance given in (4).

Gunning and Horgan (2004) developed geometric stratification, an extremely simple way of finding boundaries for stratifying skewed populations which approximately minimise (4). They tested it on real data and showed that it compared favourably in terms of efficiency and sample size to the cumulative root frequency approximation of Dalenius and Hodges (1959), and to the Lavallée-Hidiroglou iterative method (Lavallée and Hidiroglou, 1988).

Subsequent work has uncovered some problems which did not emerge during the original testing. Kozak and Verma (2006) implemented the method on five positively skewed artificial populations and found that with Neyman allocation in (5), sample sizes in some strata were too small to allow the calculation of the variance ($n_h < 2$) or were greater than the stratum sizes ($n_h > N_h$). Similar problems arose in the work of Keskinturk and Er (2007) and of Brito, Maculan, Lila and Montenegro (2010). Baillergeon and Rivest (2009) found that when populations contained very small X values, the geometric method performed poorly. All researchers reported decreased efficiency in the geometric method.

The aim of this paper is to revisit geometric stratification, and develop a modification to restore the level of efficiency observed in Gunning and Horgan (2004). After a brief overview of geometric stratification in Section 2, we describe the proposed adjustments in Section 3, and compare the efficiencies of the estimators obtained with the modified method with those of the Lavallée and Hidiroglou (1988) optimisation algorithm. All the comparisons are implemented using the R package called Stratification devised by Baillargeon and Rivest (2010). The final Section 4 summarises our developments.

2 Geometric Stratification

Geometric stratification (Gunning and Horgan, 2004) is based on an observation of Lavallée and Hidiroglou (1988):

> *"for skewed populations, stratum coefficients of variation tend to be equalised with optimal design."*

Some years previously Dalenius and Hodges (1959) hinted at the same conjecture:

> *"for many populations, and for reasonable locations of the stratum boundaries, the relative variance does not vary much from stratum to stratum"*

When we investigated the consequence of this assumption, we made a curious discovery: setting equal the coefficients of variation in each stratum, i.e.

$$\frac{S_1}{\overline{X}_1} = \frac{S_2}{\overline{X}_2} = \ldots = \frac{S_H}{\overline{X}_H}, \tag{7}$$

produces boundaries that are in geometric progression (Horgan, 2006). We briefly outline the argument which leads to geometric stratification.

2.1 The Argument

Following Dalenius and Hodges (1959), we assume that X is approximately uniformly distributed in each stratum, which implies that

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2}, \quad 1 \le h \le H, \tag{8}$$

and

$$S_h \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1}), \quad 1 \le h \le H. \tag{9}$$

The coefficient of variation of stratum h is therefore

$$CV_h = \frac{S_h}{\overline{X}_h} \approx \frac{2(k_h - k_{h-1})}{\sqrt{12}(k_h + k_{h-1})}. \tag{10}$$

With approximately equal CV_h it follows that

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} \approx \frac{k_h - k_{h-1}}{k_h + k_{h-1}}, \tag{11}$$

which reduces approximately to

$$k_h^2 = k_{h+1}k_{h-1}, \tag{12}$$

which would mean that the stratum boundaries are the terms of a geometric progression,

$$k_h = ar^h \quad h = 0, 1, \cdots, H. \tag{13}$$

Thus $a = k_0$, the minimum value of the variable, and $k_H = ar^H$, the maximum value of the variable, so that the constant ratio $r = (k_H/k_0)^{1/H}$.

The somewhat artificial example given in Gunning and Horgan (2004) illustrates its simplicity:

A population ranging from 5-50,000 is to be divided into 4 strata.

$$H = 4 \quad k_0 = 5 \quad k_4 = 50,000$$

Thus

$$r = (50,000/5)^{1/4} = 10$$

and so $k_h = 5 \cdot 10^h$ which means the breaks are

$$5, \quad 50, \quad 500, \quad 5,000, \quad 50,000.$$

Geometric stratification does not involve iteration, overcomes the pain of optimisers, and is obtained in one run through of the data file.

2.2 The Cochran-Horgan Data

Initial tests by Gunning and Horgan (2004) indicated that geometric stratification compared favourably in terms of efficiency to the Lavallée-Hidiroglou iterative algorithm (LH) for obtaining optimum boundaries. They tested it on four real skewed populations:

- Debtors: An accounting population of debtors in an Irish firm, detailed in Horgan (2003);
- UScities: The population in thousands of US cities from Cochran (1961);
- USColleges: The number of students in four-year US colleges from Cochran (1961);
- USbanks: The resources in millions of dollars of a large commercial bank in the US from Cochran (1961),

The four populations are summarised in Table 1 and boxplots are provided in Figure 1.

Table 1: *Summary Statistics for the Cochran-Horgan Populations*

Population	N	Range	Skew	Mean	SD
Debtors	3369	40-28,000	6.44	838.64	1873.99
UScities	1038	10-198	2.87	32.57	30.40
UScolleges	677	200-9,623	2.45	1563.00	1799.06
USbanks	357	70-977	2.07	225.62	190.46

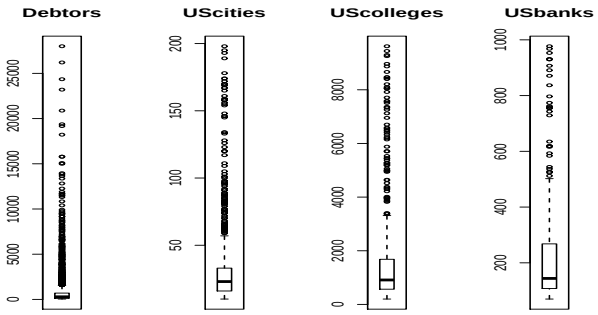

Figure 1: Cochran-Horgan Data

These data are now available in the stratification package of Baillargeon and Rivest (2010).

2.3 Efficiency Comparisons in the Cochran-Horgan Data

We implement the geometric and LH methods of stratification on these populations and compare them in terms of their relative efficiency.

For a specified sample size, the relative efficiency is the ratio of the variance of the mean obtained with LH and with geometric stratification. Conversely the relative efficiency can be defined as the ratio of the sample size required with each design to obtain the same specified variance ($V(\overline{x}_{strat})$ or equivalently the coefficient of variation ($CV = V(\overline{x}_{strat})/\overline{X}$).

Table 2 gives the efficiency of LH relative to the geometric with CV levels of 0.01, 0.015 and 0.02 and 4, 5, and 6 strata.

Table 2: *Efficiency of LH Algorithm relative to Geometric for Cochran-Horgan Data*

Pop	No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02
Debtors	4	0.76	0.81	0.87
	5	0.70	0.77	0.83
	6	0.75	0.79	0.83
USCities	4	0.84	0.94	0.99
	5	0.99	0.82	0.97
	6	0.87	0.89	0.99
USColleges	4	0.90	0.91	0.91
	5	0.82	0.90	0.94
	6	0.92	1.05	0.90
USbanks	4	0.81	0.95	1.05
	5	0.98	0.99	0.98
	6	0.98	1.20	1.31

Here we see that the efficiency is nearly always over 70% and in many cases in the high nineties. With six strata in the US banks data, the efficiency is 1.2 and 1.31 with CV = 0.015 and 0.02 respectively, indicating that LH needs 20% more sample values than the geometric to achieve CV = 0.015, and 31% more to achieve CV = 0.02. It is likely in these cases that the LH reached a local rather than a global minimum, one of the hazards of iterative procedures.

We used Kozak (2004) optimisation in the R stratification package to implement LH stratification; this is the default in the stratification package.

3 Data with Outliers

One of the first critics of geometric stratification, Kozak and Verma (2006), showed that the geometric method may not only lead to poor precision but also that some strata may be empty, and sample sizes may be less than 2 or greater than the stratum sizes. Horgan (2010) pointed out that geometric stratification uses just two values of the population to get the boundaries, the minimum and the maximum, and, as the efficiency depends critically on these, things will go wrong if either the minimum k_0 is too small, giving too many small strata, or the maximum k_H too large, dragging boundaries up.

Even in their original paper, Gunning and Horgan (2004) state:

> "*since the boundaries increase geometrically, it will not work with variables that have very low starting points: this will lead to too many small strata*"

Horgan (2010) cautioned that modifications of the geometric algorithm are necessary to address outliers and small starting points, and suggested a take-all stratum in the case of large outliers, and a take-none stratum when the starting points are very small.

To look at the problems that may arise with geometric stratification, we implement it on the skewed data provided in Baillergeon and Rivest (2010)

Our first population is the size measure used for Canadian retailers from the Monthly Retail Trade Survey (MRTS), consisting of 2,000 observations. In addition we use five of populations from the data set SWEDEN, the 284 data points of Sweden Municipalities from Sarndal et al. (1992). These are:

- P85: 1985 population in thousands;
- P75: 1975 population in thousands;
- RMT85: Revenues from the 1985 municipal taxation (in millions of kronor);
- ME84: Revenue of municipal employees in 1984;
- REV84: Real estate values according to 1984 assessment (in millions of knonor).

The other populations given in Baillargeon and Rivest (2010) turned out to be unsuitable for use with our algorithm; some were not sufficiently skewed, others were essentially discrete and yet others contained negative values.

The population used are summarised in Table 3 with boxplots provided in Figure 2

Table 3: *Summary Statistics for data sets from MRTS and Sweden*

Population	N	Range	Skew	Mean	SD
MRTS	2000	141.2-486,400	8.62	16880	21574.88
P85	284	3-653	8.23	29.36	51.56
P75	284	4-671	8.47	28.81	52.87
RMT85	284	21-6720	8.79	245.10	51.56
ME84	284	173-47,070	8.78	30.88	4253.13
REV84	284	347-59,880	7.88	3088	4746.16

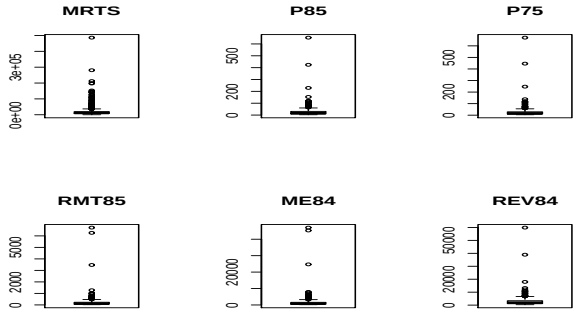

Figure 2: MRTS and Sweden Data

The first thing to notice from Figure 2 is that, unlike the Cochran-Horgan data, all the populations contain extremely large outliers.

3.1 The MRTS data

Initial applications of geometric stratification to the MRTS data yield inefficient results. Table 4 give the efficiencies of the LH method compared to the geometric, for 4, 5 and 6 strata.

Table 4: *Efficiency of LH relative to the Geometric with MRTS*

Pop	No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02
MRTS	4	0.40	0.40	0.40
	5	0.41	0.41	0.42
	6	0.38	0.37	0.39

The efficiency levels observed in Table 4 are too low for serious consideration of the geometric method as an alternative to the LH. The maximum relative efficiency is just 0.42 indicating that the LH method will require a sample size of just 42% of that required by the geometric stratification to attain the same precision.

On closer inspection, we found that there are three issues in this population that may have led to the inefficiency; large outliers, small starting points and over-allocation. We look at each one separately.

3.1.1 Large Outliers

Since the geometric method is critically dependent on the maximum value, we should exclude extreme outliers before implementing geometric stratification. Figure 3 give two boxplots of the $MRTS$ data. The first one on the left is the whole population from which we can see that there are in fact five outliers. The boxplot on the right in Figure 3 is the $MRTS$ data with these outliers removed.

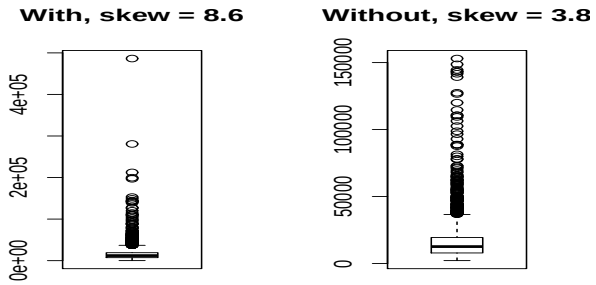

Figure 3: Outliers in MRTS

Notice that when the five top values are removed the skewness of $MRTS$ reduces from 8.6 to 3.8, illustrating the huge influence five data points can have on the distribution. Large outliers are usually put into a take-all stratum.

3.1.2 Small starting points.

It is not as easy to establish the influential small starting points as it is to establish the large outliers. A possible approach is to examine the sample size necessary using Neyman allocation with the highest likely number of strata ($H = 6$) and the highest likely coefficient of variation (CV = .02). In the MRTS population H = 6 and CV = .02 will require $n = 239$ allocated with Neyman allocation yielding:

Table 5: *Sample Allocation with MRTS[1:1995]*

N_h	9	70	533	1201	173	14
n_h	1	1	19	139	65	14

As $n_h < 2$ in each of the first two strata, we will take the starting point for geometric as the 80th value.

3.1.3 Over allocation

We also found that with Neyman allocation $n_h > N_h$ in some strata. Cochran (1977) points out that this is not unusual, and explains that the problem arises when the sampling fraction n/N is substantial, and some strata are much more variable that others. It occurs frequently in practice, and in skewed populations usually in the top stratum, where S_H can often be very large. The way to rectify the problem is to set $n_h = N_h$, and allocate the remaining sampling units optimally among the other strata:

$$n*_h = \frac{(n - N_H)W_h S_h}{\sum_{i=1}^{H} W_i S_i}. \tag{14}$$

provided that $n*_h \leq N_h$. If it should happen that some $n*_h > N_h$, the allocation is changed to include all of this stratum and reallocate the remaining elements optimally. This process is continued until every $n*_h \leq N_h$. The resultant allocation may be shown to be optimum for a given n (Cochran 1977).

3.2 Modified geometric stratification with $MRTS$ data

The excluded extreme top values will be examined 100% (a take-all stratum), can be put into a take-all stratum, and the excluded small values will not sampled at all (a take-none stratum). The population to be stratified by geometric stratification are the values of $MRTS$ in the interval [80 1995].

Table 6 gives the efficiency of LH compared to the geometric for CV levels 0.01, 0.15 and 0.02 with 4, 5 and 6 strata in the reduced population.

Table 6: *Efficiency of LH relative to the Geometric in MRTS[80:1995].*

Pop	No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02
MRTS	4	0.72	0.87	0.76
	5	0.73	0.80	0.81
	6	0.73	0.75	0.75

We see from Table 6 that, although the efficiency is less than one in all cases indicating that LH is more efficient than the geometric, the efficiency never drops below 0.7 and is sometimes 0.8 or more. Comparing the efficiency levels obtained in the reduced population (Table 6) with those obtained in the full population (Table 4), we see that the modification leads to substantial improvements in the efficiency of geometric stratification.

3.3 Efficiency Comparisons with Sweden data

In the Sweden data set, there are three clear extreme outliers which will go into the take-all stratum, and the take-none stratum are determined as shown above for the MRTS. The column on the left of

Table 7 gives the efficiency of the LH relative to the geometric for each of the populations with strata 4, 5 and 6 and CV levels .01, .015 and .02, while the column on the right gives the corresponding efficiencies for the reduced populations.

Table 7: Relative Efficiencies of LH relative to Geometric for Sweden Data

	P85			P85[14:281]		
No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02	CV = 0.01	CV = 0.015	CV = 0.02
4	0.62	0.63	0.66	0.74	0.78	0.84
5	0.57	0.59	0.57	0.76	0.91	1.05
6	0.64	0.75	0.66	0.79	0.94	0.87
	P75			P75[1:281]		
No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02	CV = 0.01	CV = 0.015	CV = 0.02
4	0.64	0.66	0.68	0.76	0.79	0.85
5	0.52	0.62	0.72	0.75	0.83	0.95
6	0.60	0.70	0.68	0.80	0.97	1.13
	RMT85			RMT85[22:281]		
No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02	CV = 0.01	CV = 0.015	CV = 0.02
4	0.64	0.67	0.71	0.74	0.78	0.85
5	0.59	0.61	0.57	0.79	0.92	1.04
6	0.66	0.67	0.67	0.71	0.83	0.94
	ME84			ME84[26:281]		
No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02	CV = 0.01	CV = 0.015	CV = 0.02
4	0.64	0.68	0.72	0.75	0.84	0.88
5	0.60	0.62	0.72	0.82	0.88	0.92
6	0.58	0.63	0.71	0.71	0.83	0.96
	REV84			REV84[13:281]		
No. of Strata	CV = 0.01	CV = 0.015	CV = 0.02	CV = 0.01	CV = 0.015	CV = 0.02
4	0.66	0.70	0.69	0.80	0.92	0.88
5	0.64	0.62	0.64	0.79	0.89	0.93
6	0.63	0.67	0.61	0.84	0.98	0.91

We see from Table 7 that, in all cases, the modification has led to improved efficiency of the geometric relative to the LH. Before modification, the efficiency of LH relative to the Geometric is less than 0.7 in most cases, and in the modified population, it is greater tha 0.7 in all cases. In many cases, the relative efficiency is greater than 0.9, and is greater than one in three cases when CV = 0.02.

4 Discussion

Geometric stratification is an very simple procedure; just take the stratum boundaries in geometric progression. There is no need for iteration and it can be implemented by hand. There is however a catch; if outliers, either too small or too large, are included, the method could be too inefficient to be of any practical value.

In this paper we have given empirical rules for determining outliers, take-all, and take-none strata so that geometric stratification returns feasible solutions, and its efficiency is improved. What we have shown is that geometric stratification is useful if we examine the population for irregularities before

applying it, notably for extremely large and small values, which should be removed before applying the geometric method. This is neither more nor less than what the practicing survey designer would do. It is also what the auditor would do before carrying out an audit; find the extremely large values for complete enumeration, and exclude the extremely small values. We have shown that simple boxplots are an effective way of finding large outliers and an initial inspection of the smallest sampling fraction to be used will cast light on the unacceptably small starting points.

Conceivably, instead of either geometric stratification, Lavallée-Hidiroglou, or other method, stratification boundaries could simply be chosen entirely as a value judgement, by a practitioner with "a good eye". In a sense this primitive idea is what we need to salvage the geometric method: before implementation first discard both large and small outliers. Indeed it is good practice to do this before implementation of any stratification method. As the geometric method is based completely on the minimum and maximum, it is more sensitive to outliers than any of the other procedures.

REFERENCES (RÉFERENCES)

Baillargeon, S. and Rivest, L.-P. (2009). A General Algorithm or Univariate Stratification, *International Statistical Review*, 77, 3, 331-344.

Baillargeon, S. and Rivest L.-P. (2010). *Univariate Stratification of Survey Populations, R Package*, available on the CRAN website at $http://www.r-project.org/$.

Brito, J. Ochi, L. Montenegro, F. and Maculan, N. (2009). An ILS Approach applied to Optimum Stratification Problem.

Cochran, W.G. (1961). Comparison of Methods for Determining Stratum Boundaries. *Bulletin of the International Statistical Institute*, 32, 2, 345-358.

Cochran (1977), *Sampling Techniques*, New York: Wiley

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, 88-101.

Gunning, P. and Horgan J. M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations *Survey Methodology*, 2, 30, 159-166.

Horgan, J. M. (2010) Choosing the Stratification Boundaries: The Elusive Optima, Istanbul Universitesi Isletme Fakultesi Dergisi, 39, 2, 195-204

Horgan, J. M. (2006). Stratification of Skewed Populations: A Review, *The International Statistics Review*, 74, 67-76.

Horgan, J. M. (2003). A List Sequential Sampling Scheme with Applications in Financial Auditing, *IMA Journal of Management Mathematics*, 14, 1-18.

Keskinturk, T and Er, S.(2007). A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling, *Computational Statistics and Data Analysis*, 52, 1, 58-67.

Kozak, M. (2004). Optimal Stratification Using Random Search Method in Agricultural Surveys, *Statistics in Transition*, 6, 5, 797-806.

Kozak, M. and Verma, M. (2006). Geometric versus Optimization to Stratification: A Comparison of Efficiency, *Survey Methodology*, 32, 2, 157-183.

Lavallée, P. and Hidiroglou. M. (1988). On the Stratification of Skewed Populations, *Survey Methodology* , 14, 33-43.

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 the Royal Statistics Society*, 97, 558–606.

Sarndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer.