4Th International Conference
New Challenges for Statistical Software - The Use of R in Official Statistics
Bucharest, 7-8 April 2016

STRATIFICATION IN BUSINESS AND AGRICULTURE SURVEYS WITH R

Marco Ballin *(ballin@istat.it), *Giulio Barcaroli* *(barcarol@istat.it),
Elena Catanese *(catanese@istat.it), *Marcello D’Orazio* *(madorazi@istat.it)

*Italian National Institute of Statistics

Introduction

Sample design of surveys on enterprises or farms is usually **one stage stratified**

Stratification based on **categorical variables** such as geographical regions, type of activity is straightforward, while not on **continuous auxiliary variables**

In general the higher is the association/correlation between the auxiliary and the target variables the higher is gain in efficiency

In multipurpose surveys where there are many target variables it is hard to find a choice of auxiliary variables that are simultaneously optimal for all variables (may be not correlated or even negatively correlated to some of the them)

The work explores **univariate** and **multivariate** methods (and *R packages* !) to stratify and allocate sample when the categorization of continuous variables is required

A comparative simulation study is provided

Stratified sampling: overview

In stratified sampling the sampling frame is divided into non-overlapping subpopulations or *strata* and sampling is performed independently in each stratum

The designer must choose:

(i) how to stratify the population and how many strata to create;

(ii) the probabilistic criterion to select the sample in each stratum;

(iii) the size of the whole sample and corresponding partitioning among the strata (so called *allocation*).

This work focuses mainly on problem (i)

In the present study it has been assumed:

(ii) SRSWOR , (simple random sampling without replacement) in each stratum

(iii) Generalized *optimal allocation* for Multivariate Surveys (Bethel, 1989; Chromy, 1987 algorithm)

Alternative methods for (iii) : *proportional allocation*, *equal allocation*, *power allocation*

Univariate Stratification

These methods determine simultaneously the best stratification and allocation, **but allow only one target variable and one stratification variable**:

a. method of the square root of the cumulative frequency
cumulative $\sqrt{f}$ *rule* (Dalenius and Hodges ,1959);
b. Generalized Lavallee-Hidoroglou method (Lavallée and Hidiroglou, 1988).
c. geometric method (Gunning and Horgan, 2004);

a. The cumulative rule is not suitable for variables showing a highly skewed distributions (typical situation in most business and agriculture surveys)

c. The method Lavallée-Hidoroglou is of wide use, and also serves to determine the units to be included in the *take-all strata* (units with inclusion probability=1) .Some convergence problems may be encountered

b. The geometric stratification method is suitable to stratify skewed populations and does not suffer of convergence problems, but it does not separate large units in a take-all stratum and requires the specification of an allocation criterion.

The R package "*Stratification*"

This package (Baillargeon and Rivest, 2011, 2014) implements all the univariate methods described before. In particular:

- ***strata.cumrootf*** implements *cumulative* $\sqrt{f}$ *rule*;

User must define H and (CV or sample size), allocation criterion can be chosen

- ***strata.geo*** implements Gunning and Horgan geometric stratification.

- ***strata.LH*** implements the generalized Lavallée-Hidiroglou (1998) method.

By default the optimization problem is solved by means of the *Kozak algorithm*, which solves many convergence problems of the original LH method (in alternative it is possible to apply the Sethi algorithm)

In addition the function ***strata.LH*** permits to identify, if required, the take-none (argument take-none), the take-all (argument take all) strata or both

The user must choose the allocation method (alloc), number of strata (Ls) and CV

Multivariate stratification: the genetic algorithm based method

The application of the genetic algorithm is based on the following choices and steps:

1) The population is first partitioned into the most detailed stratification (**atomic strata)** by cross-classifying units in the sampling frame using all the values of available auxiliary variables (after categorizing continuous ones);

2) Each stratification is considered as an **individua**l in a population subject to evolution;

3) Each individual is characterized by a genome

4) For each individual his **fitness** (the lower the associated **cost function** the higher the fitness) is calculated by solving the corresponding problem of optimal algorithm through Bethel;

5) In the transition from one generation to the next, a percentage of those individuals with higher fitness are directly moved to the next generation (*elitism*), the others are subject to ***selection***, i.e. they are randomly selected with probability proportional to their fitness, in order to let them procreate *children*; each child is procreated by applying ***crossover*** to their parents (a swap of the genes contained in the two genomes), and applying ***mutation*** to the resulting genome.

6)At the end of the process of evolution (iterations), the individual with the best fitness in absolute is the optimal solution

The R package *SamplingStrata*

This package (Barcaroli 2014, Barcaroli at al 2016) offers an approach for the determination of the *best* stratification of a sampling frame (constrained optimization problem: minimum sample cost subject to satisfy precision requirements) in a multivariate and multidomain case.

The solution is based on the use of a genetic algorithm: each solution (i.e. a particular partition of the sampling frame) is considered as an individual in a population; the fitness of all individuals is evaluated applying the Bethel-Chromy algorithm to calculate the sampling size satisfying precision constraints on the target estimates.

Key function:

- *optimizeStrata:* to perform the optimization of the strata

Main parameters of *optimizeStrata*

1. number of iterations (**iter**),
2. the population size (**pop**),
3. the mutation rate (**mut_chance**),
4. the minimum number of units per stratum (**minnumstr**).

Application to Farm Structure Survey

- **FSS** is carried out every 3 years (every 10 years is the Agricultural Census itself)**.** It investigates: crop characteristics, livestock, labour force etc.
- EU regulations:
  - stratified one stage sampling
  - characteristics (crops and livestock) subject to precision requirements (CV<=5%) at NUTS2 level.

In our simulation it is considered one NUTS2 region (Veneto)

Table 1 – Auxiliary and target variables used for stratification and allocation purposes.

Y variables	CVs	Variables X used for stratification	
		Set 1	Set 2
UAA	0.04	UAA	UAA
Cereals	0.05	LSU	Cereals
Oil seed crops	0.05		Industrial crops
Harvested green	0.05		Harvested green
Permanent grassland	0.05		Permanent grassland
Vineyards	0.05		Vineyards
LSU	0.04		LSU
Dairy cows	0.05		Dairy Cows
Other bovines	0.05		Other bovines
Pigs	0.05		Pigs
Poultry	0.05		Poultry

The target population consisted of 119,384 farms.

The frame was constituted by the 2010 Census List

The stratification strategies and the settings

- **Traditional strategy (A)**: stratification has been carried out independently on each continuous auxiliary variable, where strata boundaries obtained by the stratification package (generalized LH procedure) have been cross classified to get the final stratification, then the computation of the optimal allocation has been carried out through the Bethel algorithm, to ensure comparability with the package sampling strata (*R package Stratification+Bethel*)

- **New strategy (B)** application of the genetic algorithm, (*R package SamplingStrata*) introduced by Ballin and Barcaroli (2013), which performs jointly stratification and allocation.

- Two different set of continuous auxiliary stratification variables

1. **Set1: Utilized Agricultural Area** (proxy for crop characteristics; the sum) and **Livestock Units** (proxy for livestock characteristics; a weighted sum) have been used as stratification variables
2. **Set 2:** all the **11** target variables subject to precision requirements were used

Comparison Set1

Dependency of overall sample size has been studied by varying H:

- the univariate *Stratification* package does not support more than 20
- In addition take-all strata efficiency was checked for strategy A, while the package *SamplingStrata* automatically performs this operation

- In the package *SamplingStrata* the atomic stratification was obtained through the *K-means algorithm*(default of the package) even if in some cases it was tested as starting point the stratificaion obtained through strategy A

Comparison Set1

Strategy B

Categories for LSU, UAA	Atomic strata	Optimized strata H	Overall n	of which n take-all
5	19	10	10356	537
10	74	36	3132	186
15	169	91	2680	112
20	298	**155**	**2507**	89
25	443	**213**	**2472**	73
30	642	321	2495	70
35	828	385	2567	76
40	1066	522	2795	122

Strategy A

Categories	No take-all stratum		With a take-all stratum		
for LSU,UAA	H	n	H	Overall n	n take-all
5	25	3358	23	3385	62
6	34	3140	34	3217	42
7	47	2992	47	3065	36
8	61	2880	60	2848	20
9	77	2819	76	2862	17
10	95	2715	93	2752	15
12	**136**	**2638**	**126**	**2647**	9
15	**202**	**2603**	**199**	**2631**	7
20	352	2656	336	2652	6

1. Best absolute performances are obtained with near 200 strata in both settings
2. Relative (sampled units/strata) best results are obtained with 120/150 strata in both settings
3. Strategy B provides better results (2472 units), more take-all units recognized (73 vs 7)
4. Good results are achieved with strategy B using as starting point 15X15 strategy A (202,2603) giving **106** strata and **2541** sampled units

Comparison Set2

- Strategy A using 11 stratification variables provides unfeasible results: 3 categories per variable would lead to 1048 strata and 2090 sampled units

- Strategy B reaches its best absolute results with 1516 sample units but with 486 strata, which in real sample surveys would also be avoided bearing in mind the unit non-response treatment (roughly 3 units per stratum)

- Best relative results are achieved with **1571 units and 312 strata**

- For strategy B the reduction from atomic strata to final strata is always roughly almost ½, but for istance in set 1, using Lavaleeè Hidiroglu as starting point instead of Kmeans 2541 units, 106 strata, instead of 2680 and 91

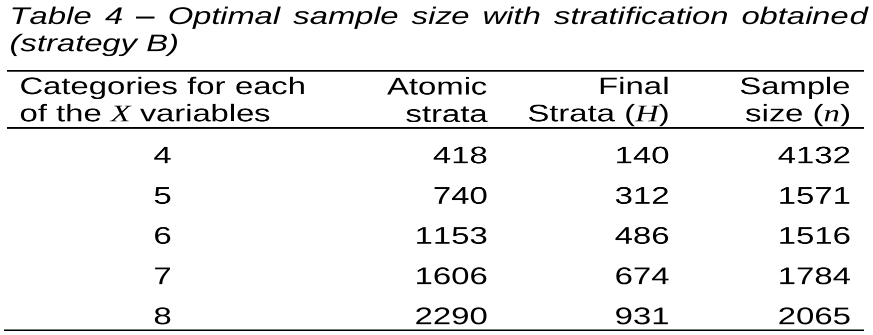

Table 4 – Optimal sample size with stratification obtained by using the Genetic Algorithm (GA) based method (strategy B)

Categories for each of the X variables	Atomic strata	Final Strata (H)	Sample size (n)
4	418	140	4132
5	740	312	1571
6	1153	486	1516
7	1606	674	1784
8	2290	931	2065

Importance of the **atomic stratification** as key parameter for *samplingstrata*

Conclusions

- In this work we analysed two *R packages* which may be useful for stratification purposes in presence of continuous auxiliary variables by applying them for FSS sample design were it was not adequate to categorize only one variable
- We analysed two extremal settings, from 2 to 11 stratification variables

- Concerning take-all units and strata which are often utilized in presence of skewed distribution, the use of the *genetic algorithm* allows the user not tackle the problem because they are automatically identified, while in *strategy A* this may be not adequate in presence of many target variables
- The use of the *strategy B* package gave better results in both settings
- Best relative results are achieved with **1571 units and 312 strata in set2** for strategy B, while *strategy A* gave its best results in **set1** with **2603 units and 202 strata,** thus providing an absolute gain of 40% in terms of sampling units
- The genetic algorithm, given a fine atomic stratification, allows to automatically explore the space of all the possible partitions of the population thus achieving an optimal solution in terms of sample size that takes into account more complex relationships between auxiliary variables, which is not the case of *strategy A* (only cross-classification employed)

Future perspectives

- The *SamplingStrata* algorithm strongly depends on the initial atomic stratification:

1. A too finer atomic stratification implies to the risk of getting on average too few sampled units per stratum which would render the solution unfeasible (see step 2: 931 strata and 2065 units)
2. An inadequate even finer initial stratification may not get the optimal sample size (see step1: i.s. 443 achieves 213 strata and 2472 units, while 828 achieves 385 strata and 2567 units)

- Possible solutions to overcome these problems

1. it may be useful to combine the strategies (i.e. use LH method as starting point instead of k.means) to achieve a good performance in terms of both sample size and number of strata
2. It may be also adequate to choose an intermediate number of auxiliary variables between set1 and set2 and use *strategy B*, which is currently occurring in **real FSS2016 sample design**

References

Baillargeon S and Rivest L.-P. 2009 A general Algorithm for Univariate Stratification. *International Statistical Review*, 77: 331-344.

Baillargeon S and Rivest L.-P. 2011 The Construction of Stratified designs in R with the package stratification. *Survey Methodology*, 37: 53-65.

Baillargeon S and Rivest L.-P. 2014 stratification: Univariate Stratification of Survey Populations. R package version 2.2-5. http://CRAN.R-project.org/package=stratification

Ballin M and Barcaroli G. 2013. Joint Determination of optimal Stratification and Sample Allocation Using Genetic Algorithm, *Survey Methodology*, 39: 369-393

Barcaroli G. 2014. SamplingStrata: An R Package for the Optimization of Stratified Sampling. Journal of Statistical Software, 61(4), 1-24. URL http://www.jstatsoft.org/v61/i04/

Barcaroli G., Pagliuca D., Willighagen E. and Zardetto D.. 2016 SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys. R package version 1.1 https://CRAN.R-project.org/package=SamplingStrata

Cochran, W.G. 1977 *Sampling Techniques, 3rd Edition*, John Wiley & Sons, New York.

Dalenious T. and Hodges J.L. 1959. Minimum variance Stratification. *Journal of the American Statistical Association*, 54: 88-101.

DeJong K.A. 2006. Evolutionary Computation: a Unified Approach. MIT Press, Boston, MA

Hidiroglou M.A. 1986. The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40: 27-31.

Hidiroglou M.A. and Lavallée P. 2009 Sampling and Estimation in Business Surveys, in *Sample Surveys: Design, Methods and Applications, Vol. 29A*, Elsevier

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

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

Rivest L.-P. 2002 A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28: 191-198.