R packages for optimal stratified sampling: a review and compared evaluation

Giulio Barcaroli, Marco Ballin

Methods, Quality and Metadata Division

Italian National Institute of Statistics (Istat)

Outline

- Optimal design of stratified sampling
- Packages for optimal design: univariate and multivariate
- Package `stratification`
- Package `SamplingStrata`
- Comparison in the univariate case
- Comparison in the multivariate case:
  - correlated survey variables
  - uncorrelated survey variables
- Conclusions

Optimal design of stratified sampling

A possible definition of “best stratification”:

“*The partition of the sampling frame that ensures the minimum sample cost under the condition to satisfy precision constraint(s); or, conversely, that maximizes the precision of target estimate(s) under budget constraints*”.

Optimization can operate in an univariate (only one target estimate) or multivariate (more than one target estimate) setting.

Packages for the optimal design of stratified sampling

```
findPackage(c("sampling","stratification"))
402 out of 12903 CRAN packages found in 11 seconds.
```

SCORE	NAME	DESC_SHORT
100.0	SamplingStrata	Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys
46.2	stratifyR	Optimal Stratification of Univariate Populations
38.5	stratbr	Optimal Stratification in Stratified Sampling
38.5	stratvns	Optimal Stratification in Stratified Sampling Optimization Algorithm
19.2	stratification	Univariate Stratification of Survey Populations
19.2	RcmdrPlugin.sampling	Tools for sampling in Official Statistical Surveys

Package	Method	Number of Y's	Precision	Number of strata	Sample size	Strata boundaries	Allocation
RcmdrPlugin. sampling	Fixed and proportional allocation	One	No	Fixed	Fixed	Fixed	**Output**
stratbr	Biased Random Key Genetic Algorithms	One	**Maximized**	Fixed	Fixed	**Output**	**Output**
stratvns	Variable Neighborhood Decomposition Search	One	Fixed	Fixed	**Minimized**	**Output**	**Output**
stratifyR	Dynamic Programming	One	**Maximized**	Fixed	Fixed	**Output**	**Output**
stratification	Frequency rule; Geometric rule; Lavallée-Hidiroglou	One	Fixed	Fixed	**Minimized**	**Output**	**Output**
SamplingStrata	Genetic algorithm	Many	Fixed	**Output**	**Minimized**	**Output**	**Output**

Univariate optimization

`stratification`

This package implements three different methods:

- the Lavallee-Hidiroglou method;
- the cumulative root frequency rule of Dalenius and Hodges;
- the geometric rule of Gunning and Horgan

There is also the possibility to incorporate in the determination of the stratum boundaries

- an *anticipated non-response*,
- a *take-all stratum for large units*,
- a *take-none stratum for small units*,
- a *certainty stratum* to ensure that some specific units are in the sample.

Multivariate optimization

SamplingStrata

Making use of the Genetic Algorithm, this package allows the multivariate stratification of population frames, minimizing the sample size on the basis of constraints on precision levels of the survey variables Y's.

The package covers all the necessary steps, from the building of the sampling fame and sampling strata, to the optimization of the stratification and allocation, to the selection of the sample.

Comparison in the univariate case

A first step for the evaluation of the performance of the two package is to compare them in the univariate case. Considering the different datasets available in the package `stratification`, these are the results:

Dataset	CV	Strata	Geometric	CumFreq	LH (Sethi)	LH (Kozak)	Genetic
`UScities`	0.01	5	180	175	183	172	172
`UScolleges`	0.01	5	197	190	162	158	161
`USbanks`	0.01	5	109	113	107	91	92
`Debtors`	0.0359	5	103	84	81	80	81

Performances between the best results from `stratification` (obtained with the Kozak implementation of the Lavallée-Hidiroglou method) are equivalent to the results obtained with the genetic algorithm in `SamplingStrata`.

Comparison in the multivariate case

Let us consider the dataset «Sweden», containing data on population of 284 Sweden Municipalities from Sarndal et al. (available with the `stratification` package).

Variable	Definition
P85	1985 population (in thousands)
P75	1975 population (in thousands)
RMT85	Revenues from the 1985 municipal taxation (in millions of kronor)
CS82	Number of Conservative seats in municipal council
SS82	Number of Social-Democratic seats in municipal council
ME84	Number of municipal employees in 1984
REV84	Real estate values according to 1984 assessment (in millions of kronor)

Multivariate case

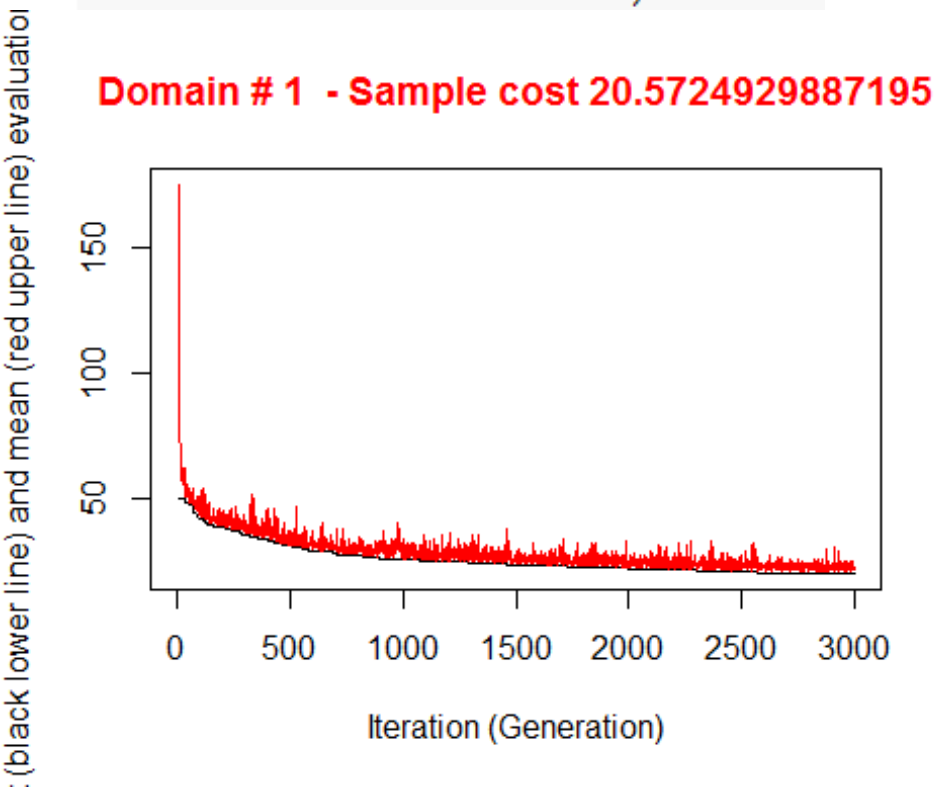
Let us suppose we want to plan a stratified sampling design that ensures expected CV's not exceeding 0.05 (5%) on all the seven variables.

Multivariate case: SamplingStrata

The multivariate optimization is a straightforward operation for the package `SamplingStrata`.

Once rounded, the total sample size required is 25, with 12 strata.

```
solution <- optimizeStrata(strata=strata,
                errors=errors,
                iter=3000,
                pop=20,
                minnumstr = 1,
                suggestions = kmeans,
                showPlot = FALSE)
```



Multivariate case: SamplingStrata

Results are in line with the set of precision constraints (0.05 on all variables):

```
results$coeff_var

# CV1         CV2         CV3         CV4         CV5         CV6         CV7

# 1 0.04544416 0.04597138 0.04814187 0.04663923 0.04302946 0.04706113 0.04992981
```

Multivariate case: stratification

In principle, the package `stratification` cannot handle a multivariate problem, as the optimization is carried out on only one survey variable.

However, it is possible to define a strategy that may overcome this limit:

1. run the stratification step by considering as target one of the variables;
2. calculate the expected CV on the other variables and compare to precision constraints: are they compliant?;
3. if not, make the precision constraint on the target variable more restrictive and run again steps 1 and 2 until all constraints are respected;
4. Repeat steps 1-3 changing the target variable and consider as solution the best results obtained.

Multivariate case: stratification

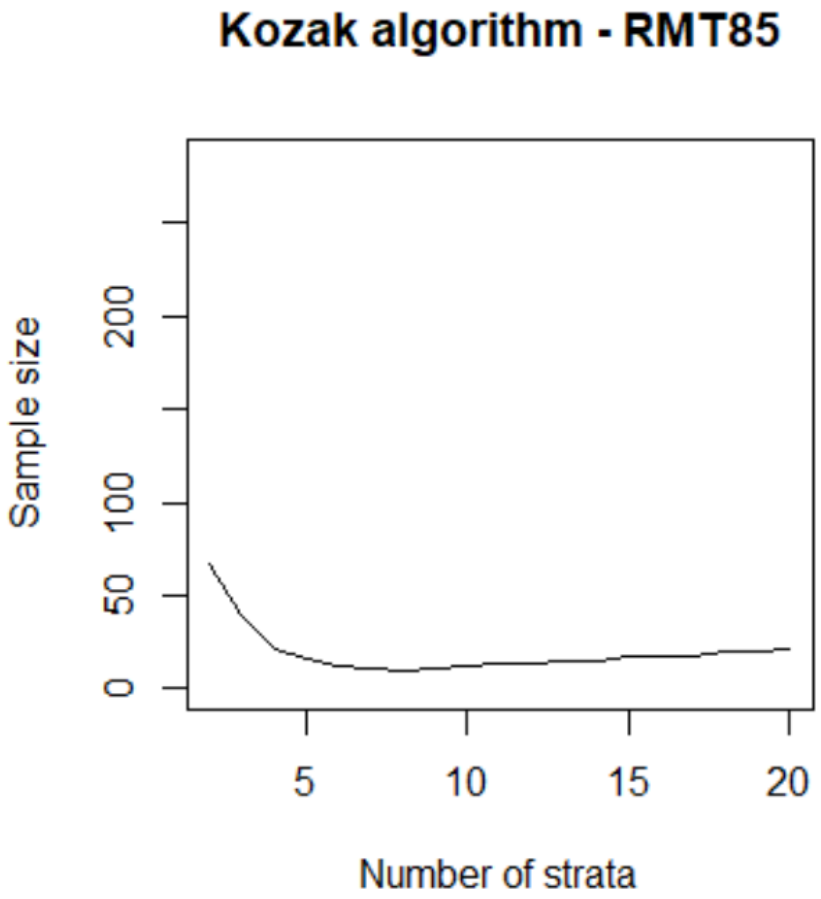
The dataset Sweden is characterized by high correlation levels among variables:

```
##             P85      P75    RMT85     CS82     SS82      S82     ME84
## P85
## P75    1.00***
## RMT85  0.96***  0.97***
## CS82   0.57***  0.55***  0.49***
## SS82   0.48***  0.47***  0.40***  0.19**
## S82    0.69***  0.67***  0.58***  0.63***  0.76***
## ME84   0.96***  0.97***  1.00***  0.49***  0.41***  0.59***
## REV84  0.98***  0.97***  0.94***  0.55***  0.47***  0.68***  0.94***
```

Two variables are the most correlated with the others: RMT85 and REV84.
We start with RMT85.

Multivariate case: stratification

To overcome a limit of the package, i.e. the need to indicate a given number of strata, we iterate the optimization varying this number from 2 to 20, and obtain a minimum sample size (10) in correspondence of 8 strata.



Multivariate case: stratification

With 8 strata and a precision constraint of 0.05 on RMT85, we obtain a sample size of 10 together with the following expected CV's:

P85	P75	RMT85	CS82	SS82	ME84	REV84
0.05622861	0.05871432	0.04273058	0.1711964	0.1013585	0.0490159	0.1486845

The CV's of only two variables are compliant (RMT85 and ME84).
Two others are slightly beyond the limits (P85 and P75).
Three CV's (CS82, SS82 and REV84) are far beyond the precision constraint.

Multivariate case: stratification

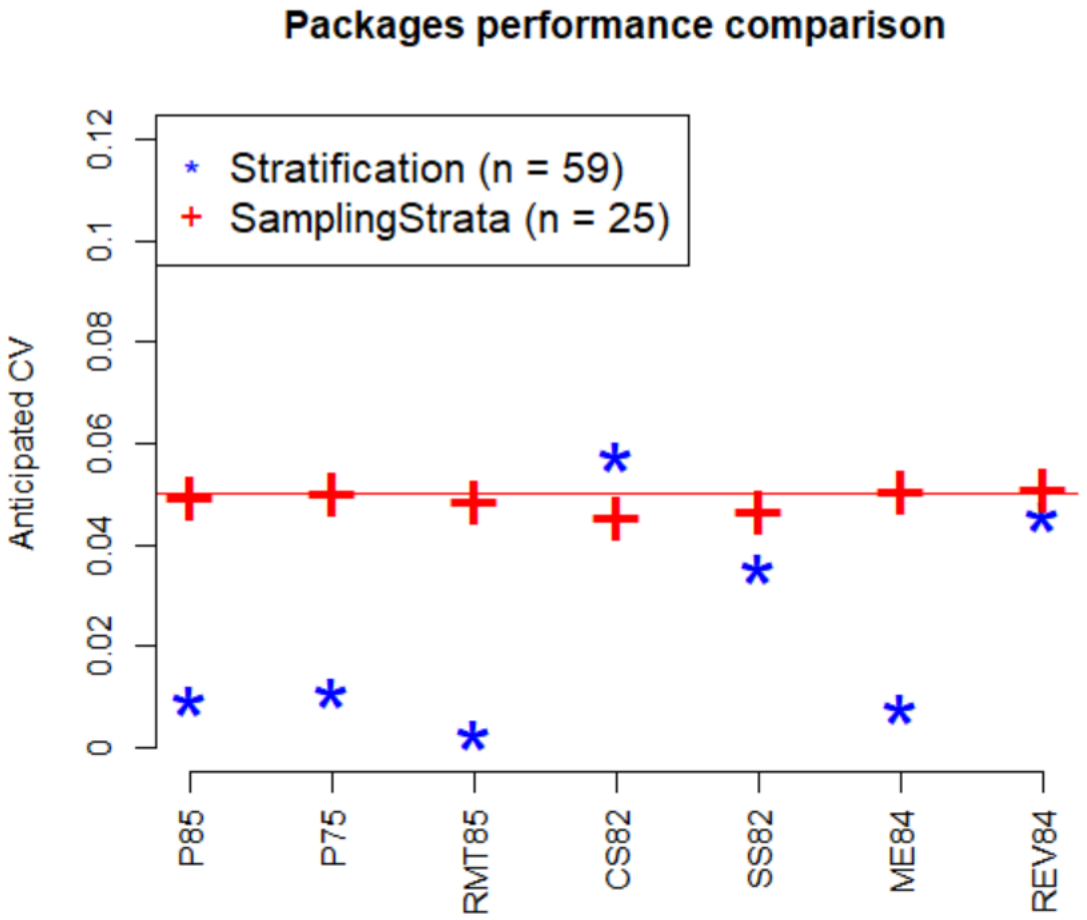
So, accordingly to our strategy we progressively set tighter values to the precision constraint on RMT85, until we obtain, with 0.003, a sample size of 60 (with 22 strata), and expected CV's :

P85	P75	RMT85	CS82	SS82	ME84	REV84
0.009440633	0.0112179	0.002680708	0.05482607	0.03407485	0.007172318	0.04591209

Now, only the CS82 expected CV slightly exceeds the 0.05 limit, while all the other are compliant.

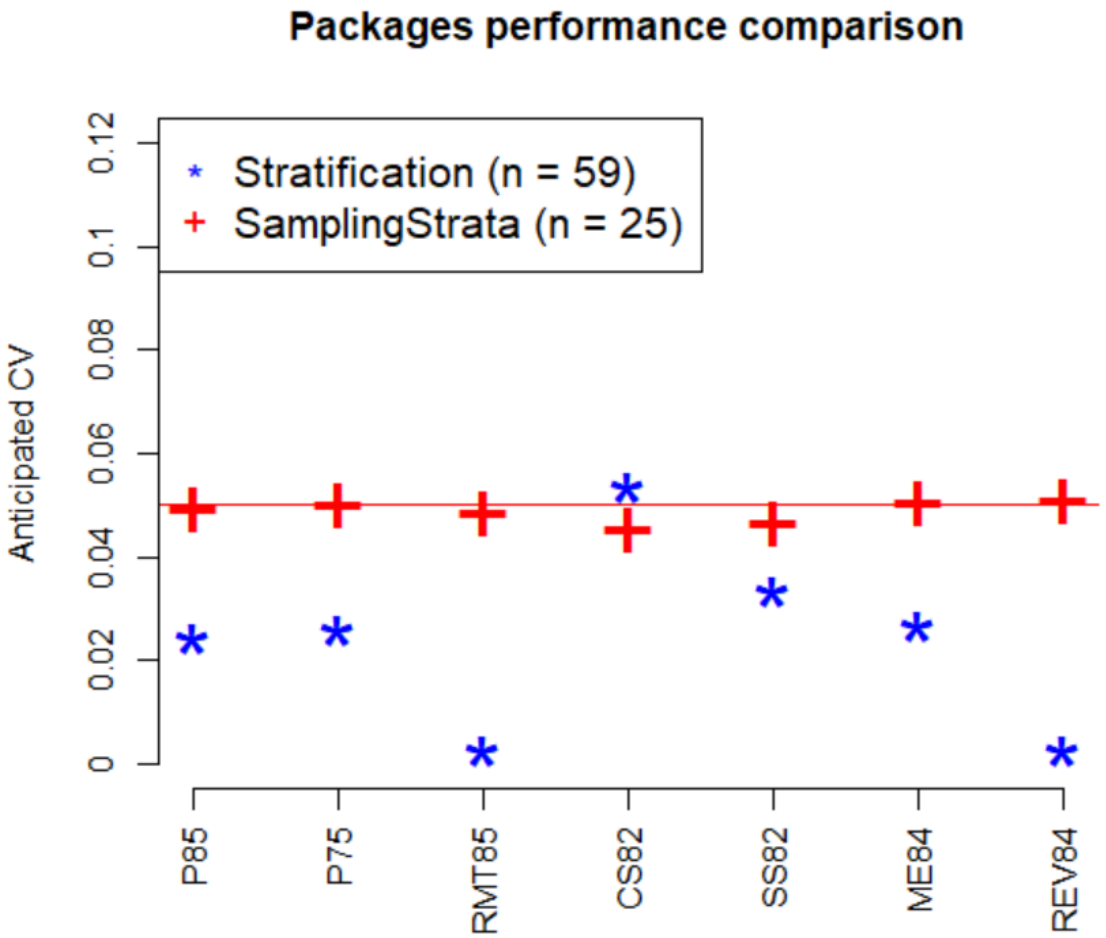
Multivariate case (correlated data)

This is the overall comparison of the results obtained by the two packages. Both solutions are compliant, but sample size required by `stratification` is much higher.



Multivariate case (correlated data)

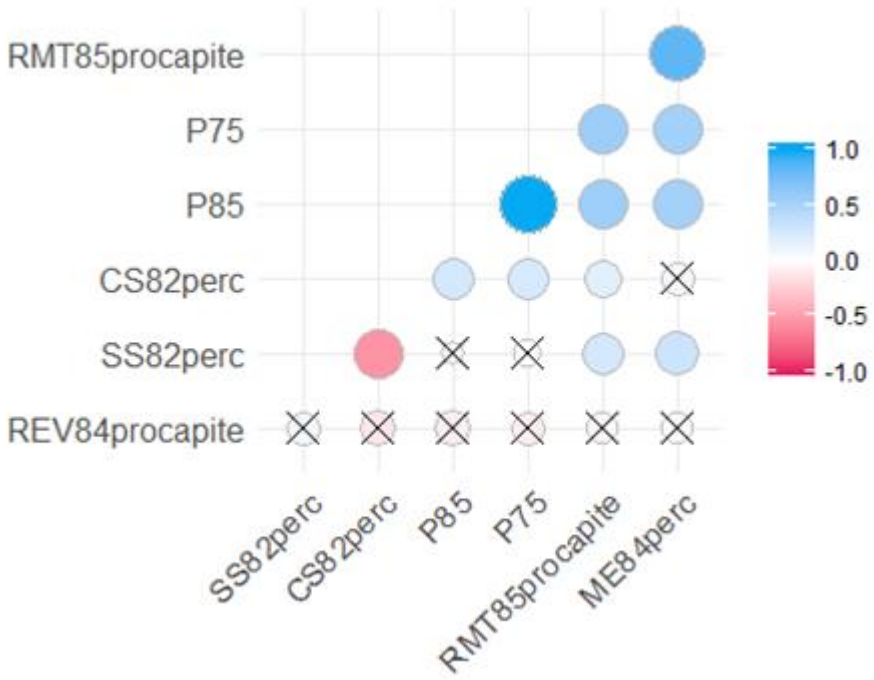
Using REV84 as «pivot» variable for package `stratification` did not change the results.
All other variables tested gave even worse results for `stratification`.



Multivariate case (uncorrelated data)

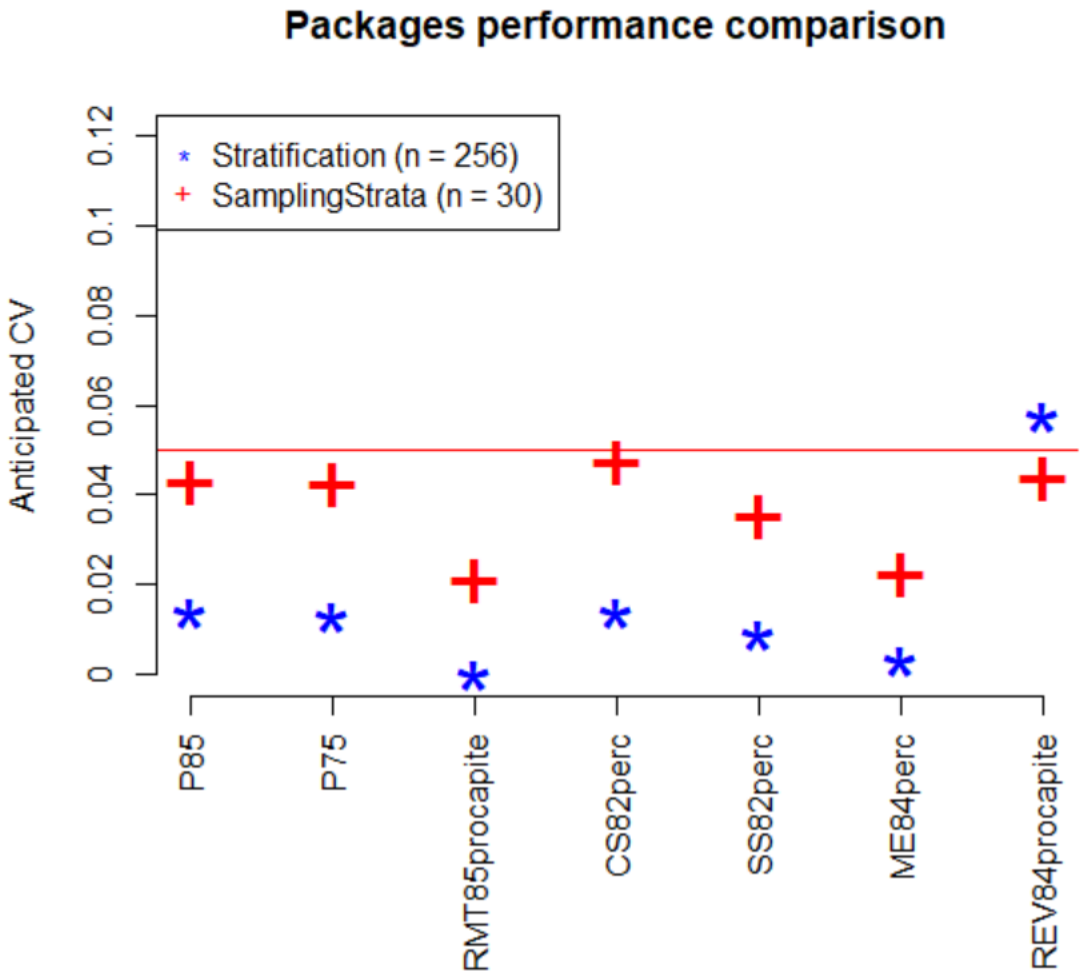
What if we consider data that are not correlated?
To investigate this different situation, we modified the original dataset by deriving new variables:

Variable	Definition
P85	Not transformed
P75	Not transformed
RMT84procapite	RMT84 / P85
CS82perc	CS82 / S82
SS82perc	SS82 / S82
ME84perc	ME84 / P85
REV84procapite	REV84 / P85



Multivariate case (uncorrelated data)

While there is a slight increment in the sample size of the `SamplingStrata` solution, this increment is dramatic for `stratification`.



Conclusions

In the univariate case, the two packages can be said to be equivalent, though `stratification` is more efficient in terms of processing time.

In the multivariate case, on the contrary, when the number of survey variables is not small the convenience of `SamplingStrata` is evident.

This is true even when there is a relatively high correlation among variables. ***When correlation is low***, the use of an univariate approach is not advisable.

References

Baillargeon S. and Rivest L.-P. (2012). The construction of stratified designs in R with the package stratification. Survey Methodology, Vol. 37, No. 1, pp. 53-65

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

Barcaroli G., Pagliuca D., Willighagen E., Zardetto D. (2018). SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys. R package version 1.2. http://cran.r-project.org/web/packages/SamplingStrata/index.html

Ballin M., Barcaroli G. (2016). Optimization of stratified sampling with the R package SamplingStrata: Applications to network data. Computational Network Analysis with R: Applications in Biology, Medicine and Chemistry, Wiley

Ballin M., Barcaroli G., Catanese E., D'Orazio M. (2016). Stratification in Business and Agriculture Surveys with R. Romanian Statistical Review 2/2016, pp. 43-58

Barcaroli, G. (2014). SamplingStrata: An R package for the optimization of stratified sampling. Journal of Statistical Software 61 (4), 1-24.

Bethel J. (1989). Sample Allocation in Multivariate Surveys. Survey Methodology, Vol. 15, pp. 47-57

Kozak M., Wang H.Y. (2010). On stochastic optimization in sample allocation among strata. Metron – International Journal of Statistics 2010, vol. LXVIII, n.1, pp. 95-103

Lavallée P., Hidiroglou M.A. (1988). On the stratification of skewed populations. Survey Methodology, Vol.14, pp.33-43

Thank you for your attention

barcarol@istat.it

More information on SamplingStrata:

https://barcaroli.github.io/SamplingStrata/index.html