The 8th International Conference
The Use of R in Official Statistics
2-4 December 2020

R package SamplingStrata: new developments and extension to Spatial Sampling

Marco Ballin, Giulio Barcaroli
ballin@istat.it, gbarcaroli@gmail.com

Outline

- Overview of SamplingStrata R package
- Recent developments: models and anticipated variance
- Extension to spatial sampling
- Methods in SamplingStrata for handling spatial sampling
- Case study: the Meuse dataset

Overview of SamplingStrata: problem setting

Problem: given a sampling frame, we need to plan a sampling survey in the most efficient way, i.e. the one that

1. miminizes the **sample size** required to satisfy given **precision constraints** on the target estimates, or
2. maximizes the precision of target estimates given an overall sample size.

SamplingStrata can be applied directly in the first case, and, indirectly, also in the second.
It operates by considering all the available potential stratification variables in the sampling frame, and looking for the best stratification of the frame (the one ensuring condition (1)).
This is done by exploring, with a **genetic algorithm**, the universe of possible solutions, and choosing the one for which the required sample size is the minimum.

Overview of SamplingStrata: optimization methods

The R package *SamplingStrata* offers two different methods for the optimization of the sampling frame, both based on the use of the *genetic algorithm*:

1. the first one is applicable when the stratification variables are all **categorical** (or, if continuous, categorized);
2. the second one is applicable when all the stratification variables are of the **continuous** type.

In the first case, the optimization proceeds by aggregating *atomic strata*, i.e. the ones resulting from cross-classifying units in the frame by their stratification variables values.

In the second case, strata are determined by randomly cutting the interval of acceptable values for each stratification variable, and then cross-classifying units.

Overview of SamplingStrata: anticipated variance

- If the Y's variables available in the frame are proxy of the real target variables (named Z's), a sampling design planned on Y's can't guarantee that the final stratification and allocation are compliant with the set of precision constraints for the real target variables.

- If we are able to define and fit models linking the proxys and the targets, using such models we are able to correctly evaluate the so called *anticipated variance* and in this way we guarantee that the expected sample errors are compliant with the requested constraints.

Overview of SamplingStrata: anticipated variance

- In the current implementation, only models linking continuous variables can be considered.
- The models, their definition and use are the same that have been implemented for the univariate case in the package *stratification*. Among them, the linear model with **heteroscedasticity**:

$$Z = \beta Y + \epsilon$$

- with $\epsilon_i \sim N(0, s_i^2)$
- where $s_i^2 = \sigma_i^2 y_i^{2\gamma}$ and $s_{ij} = 0$

(in case $\gamma = 0$, the model is homoscedastic)

Overview of SamplingStrata: heteroscedasticity

In case of heteroscedasticity ($\gamma > 0$) a crucial point is in correctly quantifying it.

A simple and quick solution has been implemented in a new function *computeGamma* embedded in *SamplingStrata*.

This function receives in input the vectors respectively of residuals and of the explanatory variable, and yields a heteroscedasticity index value together with the value of model variance to be used as values of corresponding parameters.

Extension to spatial sampling

In case Z is the target variable, omitting as negligible the *fpc* factor, the sampling variance of its estimated mean is:

$$V(\hat{\bar{Z}}) = \sum_{h=1}^{H} (N_h/N)^2 S_h^2/n_h \tag{1}$$

We can write the variance in each stratum h as:

$$S_h^2 = \frac{1}{N_h^2} \sum_{i=1}^{N_{h-1}} \sum_{j=i+1}^{N_h} (z_i - z_j)^2 \tag{2}$$

Extension to spatial sampling

Obviously, values z are not known, but only their predictions, obtained by means of the known or estimated linking models.

So, in Equation 2 we can substitute the term $(z_i - z_j)^2$ with:

$$D_{ij}^2 = \frac{(\tilde{z}_i - \tilde{z}_j)^2}{R^2} + V(e_i) + V(e_j) - 2Cov(e_i, e_j) \qquad (3)$$

where R^2 is the squared correlation coefficient indicating the fitting of the regression model and $V(e_i)$, $V(e_j)$ are the model variances of the residuals.

The auto-correlation component is contained in the term $Cov(e_i, e_j)$.

Extension to spatial sampling

In particular, assuming that the autocorrelation function is exponential, the quantity D_{ij} may be calculated in this way:

$$D_{ij}^2 = \frac{(\tilde{z}_i - \tilde{z}_j)^2}{R^2} + (s_i^2 + s_j^2) - 2 s_i s_j e^{-k(d_{ij}/range)} \quad (4)$$

where d_{ij} is the distance between units i and j, s_i^2 and s_j^2 are estimates of the corresponding prediction variance and *range* is the maximum distance below which spatial auto-correlation can be observed among units.

The value of *range* can be determined by an analysis of the spatial *variogram*.

Method to handle spatial sampling

Given the previous formula for the calculation of variance in strata including also the spatial component, an option has been made available in SamplingStrata consisting in supplying for each variable of interest its predicted value for each unit in the frame together with its associate predicted error, both previously determined by a suitable model handling spatial component (for instance, *kriging*).

'Spatial' optimization method

In the *spatial* method:

- any kind of model can be used, parametric or non parametric;
- there is no need to indicate the values of models parameters;
- the only parameter explicitly required is the *range* of spatial correlation, whose value can be determined by analysing the *variogram*.

For each unit in the frame the following information must be given:

1. **geographic coordinates**;
2. values of the **predicted values** of the Z's;
3. values of **prediction errors**.

'Spatial' optimization method

After running the spatial optimization method, yielding the best stratification of the sampling frame, then we can proceed with the **selection of the sample**.

It is important to select the units also taking into account their spatial positioning, in order to obtain a **spatial balanced sample**.

In order to do that, a new function, *selectSampleSpatial*, has been developed, which is a wrapper of the function *lpm2_kdtree* in package *SamplingBigData* (Lisic and Grafström, 2018).

Case study: the Meuse dataset

This case study is reported in the SamplingStrata vignette '*Spatial sampling with SamplingStrata*', at the link:

- https://barcaroli.github.io/SamplingStrata/articles/spatial.html

To go into deeper detail...

- The **paper** from which this presentation has been derived: M. Ballin M. and G. Barcaroli (2020). *R package SamplingStrata: new developments and extension to Spatial Sampling*. arXiv:2004.09366 [stat.ME]
- To get **general information** on SamplingStrata package: https://barcaroli.github.io/SamplingStrata/articles/SamplingStrata
- In particular on **spatial sampling with SamplingStrata**: https://barcaroli.github.io/SamplingStrata/articles/spatial.html

References

- S. Baillargeon and L.-P. Rivest (2014). Stratification: Univariate Stratification of Survey Populations. https://CRAN.R-project.org/package=stratification.
- M. Ballin M. and G. Barcaroli (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey methodology* **39**:369-393.
- G. Barcaroli (2014). SamplingStrata: An R package for the optimization of stratified sampling.*Journal of Statistical Software*, **61**:1—24.
- J. Bethel (1989). Sample allocation in multivariate surveys. *Survey Methodology*, **15**:47—57
- J.J. de Gruijter and B. Minasny and A.B. McBratney (2015). Optimizing Stratification and Allocation for Design-Based Estimation of Spatial Means Using Predictions with Error. *Journal of Survey Statistics and Methodology*, **3**:19–42.
- J.J. de Gruijter and I. Wheeler and B. Malone (2019). Using model predictions of soil carbon in farm-scale auditing - A software tool. *Agricultural Systems*, Elsevier, vol. 169(C), pages 24-30
- K. Henry and R. Valliant (2006). Comparing Strategies to Estimate a Measure of Heteroscedasticity, https://www.irs.gov/pub/irs-soi/06rpphenry.pdf
- Lisic, J., and A. Grafström. 2018. SamplingBigData: Sampling Methods for Big Data. https://cran.r-project.org/package=SamplingBigData