BALANCED, SPATIALLY BALANCED AND DOUBLY BALANCED SAMPLE SELECTION ALGORITHMS BASED ON MCMC

Roberto Benedetti[1] & Maria Michela Dickson[2] & Giuseppe Espa[3] & Federica Piersimoni[4]

[1] *"G. d'Annunzio" University, Department of Economics, Viale Pindaro 42, Pescara, IT-65127, Italy, benedett@unich.it*
[2] *University of Trento, Department of Economics and Management, Via Inama 5, 38100,Trento, Italy, MariaMichela.Dickson@unitn.it*
[3] *University of Trento, Department of Economics and Management, Via Inama 5, 38100,Trento, Italy, giuseppe.espa@unitn.it*
[4] *Istat, Italian National Institute of Statistics, Agricultural Statistical Service, Viale Oceano Pacifico 171, Rome, IT-00144, Italy, piersimo@istat.it*

Summary. The spatial distribution of a population represents an important tool to design samples and its use increased in the last decades as the GIS and GPS technologies made more and more cheap to add information regarding the exact or estimated position for each record in the frame.

In most environmental surveys the spatial attributes are even used as a criterion to define the statistical units. Points, lines and portions of land are widely used in this filed to identify an object to be observed or measured.

However it is a quite common practice within National Statistical Offices to add the available data on the spatial characteristics of each entity to the sampling frames used for social and economic surveys. These data may represent a source of auxiliaries that can be helpful to design effective sampling strategies which, assuming that the observed phenomenon is related with the spatial features of the population, could gather a considerable gain in their efficiency by a proper use of these particular information.

A set of MCMC-based methods for selecting samples from a spatial finite population that are balanced on some covariates and/or well spread over the population in every dimension, without the use of any spatial stratification is presented. The within sample distance matrix is summarized in a descriptive index which is used to define the probability *p(s)* of each sample to be selected. A set of units with higher within sample distance will be selected with higher probability than a set of nearby units. Through the standardization of the distance matrix these algorithms can be used to produce equal and unequal probability samples either exact when a linear index is used to summarize the matrix or approximate when products and powers of the mean distance are used. The high flexibility of the selection algorithm can make possible numerous extensions to deal with some practical topics that are usually met in spatial surveys such as the sample coordination and the spread of units belonging to different domains. Some examples on real and simulated data show that the designs gives estimates that are better than those obtained by using a classical solution as the Generalized Random Tessellation Stratified (GRTS) design and that often are even slightly better than those obtained by using recently proposed selection procedures as the Spatially Correlated Poisson Sampling (SCPS) and the Pivotal method.

With respect to the Cube method and to these distance based methods the proposed algorithms, even if in their nature they are computationally intensive, seems to represent interesting alternative solutions even faster when dealing with large population frames.

Keywords. Cube method, spatially balanced samples, empirical inclusion probabilities, MCMC, Correlated Poisson Sampling, Pivotal Method, Generalized Random Tessellation Stratified design.

Bibliography

Benedetti R, Palma D (1995), Optimal sampling designs for dependent spatial units, *Environmetrics*, 6, 101-114.

Benedetti, R., Piersimoni, F. & Postiglione, P. (2015). *Sampling spatial units for agricultural surveys*. Advances in Spatial Science Series. Berlin Heidelberg: Springer.

Breidt FJ, Chauvet G (2012). Penalized balanced sampling. *Biometrika*, 99: 945–958.

Chauvet G, Tillé Y (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21: 53-62.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142, 139–147.

Grafström A, Lundström NLP, Schelin L (2012). *Spatially balanced sampling through the pivotal method*. Biometrics, 68: 514-520.

Grafström A, Tillé Y (2013). *Doubly balanced spatial sampling with spreading and restitution of auxiliary totals*. Environmetrics, 24: 120-131.

Stevens Jr., D.L. and Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99, 262–278.

Tillé, Y. (2006). *Sampling algorithms*. Springer series in statistics. Springer, New York.

Tillé Y (2011). Ten years of balanced sampling with the cube method: An appraisal. *Survey Methodology*, 37: 215–226.

Traat, I., Bondesson, L., and Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference,* 123, 395 -413.