

Spatially balanced samples for land use/land cover surveys

Disegni campionari spazialmente bilanciati per indagini su uso e copertura del suolo

Roberto Benedetti, Federica Piersimoni and Paolo Postiglione¹

Abstract The particular characteristics of geographically distributed data should be taken into account in designing a land use/land cover survey. This paper discusses several methods for sampling spatial units that have been recently introduced in literature. The methodological framework followed is of design-based typology. The techniques outlined are: the GRTS, the Cube, the SPCS, and the LPMs. These methods will be verified on data deriving from LUCAS 2012.

Abstract *Le particolarità dei dati territoriali dovrebbero essere presi in considerazione nella progettazione di una indagine di copertura / uso del suolo. Questo articolo discute diversi metodi per il campionamento di unità spaziali recentemente introdotte in letteratura. Il quadro metodologico seguito è quello basato sul disegno. Le tecniche descritte sono: GRTS, Cube, SCPS e le due varianti del LPM. Questi metodi sono stati verificati su dati derivanti da LUCAS 2012.*

Key words: spatial sampling, spatially balanced samples, MCMC, variogram, spatial stratification, Generalized Random Tessellation Stratified design.

1 Introduction

In the last decades, the definition of methods for sampling spatial units has become so popular that several authors introduced ideas and techniques about this argument. In this paper, our interest will be focused on probability samples that are well spread over the population of interest. The samples that match this property are defined as

¹ Roberto Benedetti, University of Chieti-Pescara, benedett@unich.it
Federica Piersimoni, ISTAT, piersimo@istat.it
Paolo Postiglione University of Chieti-Pescara, postigli@unich.it

spatially balanced. Surprisingly, this concept is mainly based on intuitive considerations, and it is not so clear if it could have an impact on the efficiency of the estimates. Furthermore, the well-spread property is not uniquely defined, and so the methods that aim at satisfying this requirement are based on several interpretations of this concept. In design-based sampling theory, if we assume that there is not a measurement error, the hypothetical observations that can be surveyed cannot be considered dependent. On the other hand, a typical characteristic of spatial data is the dependence. It is evident that, within a model-based or a model-assisted framework, spatial sampling can be explored postulating an appropriate model of spatial dependence. In the past, some survey scientists try to develop methods following the intuition to spread the selected units over the space because closer observations will provide overlapping information as an immediate consequence of the dependence (Benedetti and Palma 1995, Rogerson and Delmelle 2004). This approach leads to the research of a *sample* that is the best representative of the whole population. This *optimal* selection is not surely acceptable in the framework of design-based inference. A possible intuitive solution to the problem of the production of a well-spread sample is to stratify the units of the population on the basis of their location. This strategy has an impact only on the first-order inclusion probabilities, but not on the second-order inclusion probabilities. Furthermore, it is not very evident how to obtain a good partition of the area under investigation. To overcome these drawbacks, the survey practitioners partition the area in as many strata as possible and select one or two units per stratum. Unfortunately, this simple plan is arbitrary, and so it is needed to move some steps further to define other appropriate sampling designs. The main aim of this paper is the application of spatially balanced samples to Land Use/Land Cover (LULC) Surveys. Land is very important for most biological and human activities on the earth. Land is the one of the main economic resource for agriculture, forestry, industries, and transport. Land can be divided into two interconnected concepts. The first is land cover that refers to the biophysical coverage of land (e.g. crops, grass, broad-leaved forest, or build-up area). The second is land use that specifies the socio-economic use of land (e.g. agriculture, forestry, recreation or residential use). The layout of this paper is the following. Section 2 will be devoted to an outline of the main spatial sampling designs that have been recently presented in literature with particular reference to Land Use/Land Cover (LULC) Survey. Finally, Section 3 reports some considerations about some preliminaries results and concludes the paper.

2 Methodologies

The main problems connected to the definition of well-spread samples have stimulated many contributions in literature. We start our description introducing the Generalized Random Tessellation Stratified (GRTS) design. For the definition of GRTS, the underlying idea is the extension of the use of systematic sampling to two or more dimensions (Stevens and Olsen 2004). This design systematically selects the

units, transforming the two-dimensional population into one dimension while trying to preserve some multi-dimensional order. The GRTS is a form of spatially-balanced sampling, and ensures design-based inferences to the entire area. The aim is that no points in the target population are too far from a sampled point, and few sampled points are close together. The concept underlying the GRTS is to apply recursive partitioning to create a spatial address. At each step in the procedure, the total inclusion probability for each cell is computed as the sum or integral of the inclusion probability of all population elements within the cell. The inclusion probability need not be constant, and very general variable probability designs can be adapted. The recursive procedure is performed until every cell has total inclusion probability less than one, and then hierarchical randomization is applied. To each cell is assigned a length equal to its inclusion probability, and then the lengths are linked together, forming a line with length equal to the total sample size. A systematic sample is selected along the line. This one to one recursive map ensures that every point on the line agrees to some population elements. The GRTS is essentially based on the use of Voronoi polygons that are used to define an index of *spatial balance*. Denote with $s = \{s_1, s_2, \dots, s_N\}$ the generic element of the set of all the possible random samples S of fixed size n , which can be selected from the population U , where s_i is equal to 1 if the unit with label i is selected, and 0 otherwise. Furthermore, define with π_i the first-order inclusion probability, and with π_{ij} the second-order inclusion probability, respectively. For a generic sample s the Voronoi polygon for the sample unit $s_i=1$ includes all population units closer to s_i than to any other sample unit $s_j=1$. If we let v_i be the sum of the inclusion probabilities of all units in the i -th Voronoi polygon, for any sample unit u_i , we have $E(v_i)=1$ and for a spatially balanced sample all the v_i should be close to 1. Thus the index $Var(v_i)$ can be used as a measure of spatial balance for a sample. For greater details about the GRTS see Stevens and Olsen (2004). The GRTS design is a commonly used by the Environmental Protection Agency of the United States for aquatic resource monitoring. A possible alternative scheme is represented by the balanced sampling and the cube method (Deville and Tillé 2004). The rationale behind this method is the following. A practitioner might ask to check the quality of the selected sample by verifying how it works on some covariates X known for every unit of the population U . The interest is to know whether the HT estimator $\hat{t}_{HT, x_j}$ is close to the known population total t_{x_j} for each of the available q covariates. The idea underlying this approach is based on the expectation that an error committed in estimating an auxiliary could indicate that there is a risk to reproduce the same error on the survey variables. This proposal lead to the need of procedures aimed to select samples with the important property:

$$\sum_{k \in s} w_k x_{kj} = \hat{t}_{HT, x_j} = t_{x_j} = \sum_{k \in U} x_{kj} \quad \forall j = 1, \dots, q, \quad (1)$$

for all the $s \in S$ such that $p(s) > 0$, and where x_{kj} is the value of j -th variable for the k -th unit. Note that the balanced sampling has been defined in a non-spatial context,

but for example, it can be applied in a spatial context by using the coordinates of the units as covariates and by imposing that any selected sample should respect for each coordinate the first p moments, assuming implicitly that the survey variable y follows a polynomial spatial trend of order p (Breidt and Chauvet 2012). However, these methods do not explicitly use the concept of distance, which is a basic statistics to describe the spatial distribution of the sample units. A design based on the distance between units supports the intuitive criterion that units that are close, seldom appear simultaneously in the sample. This last condition can be considered as reasonable under the assumption that, increasing the distance between two units i and j , always increases the difference $|y_i - y_j|$ between the values of the survey variable. In this situation it is clear that the HT variance of the estimates will necessarily decrease if we set high second-order inclusion probabilities to couples with very different y values as they are far each other. Many scientists try to address the problem of selecting spatially balanced samples. Grafström (2012) proposed a method called Spatially Correlated Poisson Sampling (SCPS) by modifying the Correlated Poisson Sampling (CPS), introduced by Bondesson and Thorburn (2008) to select units with unequal inclusion probabilities. The CPS is based on a list sequential of the visit of the units of the population. Each unit k can be selected in the sample according to inclusion probabilities that are adapted step by step in order to produce correlations between the indicator variable of the unit visited, I_k , and the indicator variables relative to all the other units of the population, I_l with $l \neq k$. In the case of negative correlation between the indicator variables of the units that are closer to those already selected; it is very difficult that these units are included in the sample. Obviously, this situation is desirable in order to define spatially balanced samples. If unit 1 is included with probability $\pi_1^{(0)} = \pi_1$, we will set $I_1 = 1$ and $I_l = 0$ otherwise. After each step, the inclusion probabilities for the remaining units in the list are updated according to a specific rule. Starting with $\pi_k^{(0)} = \pi_k$, $k \geq 1$. At step t , the values of $I_1, I_2, \dots, I_{t-1}$ are known, we will select the unit t with probability $\pi_t^{(t-1)}$. We update the generic unit $k \geq t+1$, according to $\pi_k^{(t)} = \pi_k^{(t-1)} - (I_t - \pi_t^{(t-1)})w_{k-t}^{(t)}$, where $w_{k-t}^{(t)}$ are weights that depend on $I_1, I_2, \dots, I_{t-1}$ but not on I_t (Bondesson and Thorburn 2008). Grafström (2012) follows the logic of CPS to define SCPS method that is based on two different strategies for choosing the weights: maximal weights and Gaussian preliminary weights. Maximal weights strategy produces samples of fixed size if the inclusion probabilities sum to an integer. Two alternative procedures to select samples with fixed π_k and correlated inclusion probabilities were derived (Grafström et al. 2012). These methods constitute an extension of the Pivotal method introduced to select πps samples (Dewille and Tillé 1998). They are essentially based on an updating rule of the probabilities π_k and π_l that at each step, should locally keep the sum of the updated probabilities as near constant and differ from each other in a way to choose the two nearby units k and l . These two methods are referred to as the Local Pivotal Method 1 (LPM1), which, according to the suggestion of the authors, is better *spatially balanced*, and the Local Pivotal Method 2 (LPM2), which is simpler and faster. The LPM1 randomly chooses the first unit k

and then the closer unit l (if two or more units have the same distance to k , the method randomly chooses between them) under the assumption that k is the nearest neighbor of l . The LPM 2 is very similar to the LPM 1, the only difference is represented from the absence of the restriction that the two units should be the nearest neighbors each other. If we assume that the phenomenon to be surveyed is affected by some spatial effects, the problem of selecting spatially balanced samples consists in using a design that will give higher probabilities to samples with higher within distance. This design $p(s)$ can be obtained by setting each $p(s) = M(D_s) / \sum_s M(D_s)$ proportional to some synthetic index $M(D_s)$ of the matrix D_s of the distances observed within each possible sample s . The algorithm starts at iteration $t=0$, with an initial point $s(0)$, randomly selected from $\{0,1\}^N$ according to a *SRS* with constant inclusion probabilities. In a generic iteration t the elements of $s(t)$ are updated in the subsequent steps:

1. select at random two units included and not included in the sample in the previous iteration, say i and j . Formally one respectively among the units within the sample, for which $s_i^{(t)} = 1$, and another among the units outside the sample for which $s_i^{(t)} = 0$;
2. denote with $^*s_i^{(t)}$ the sample where the units in the position i and j exchange their status. Randomly decide whether to adopt $^*s_i^{(t)}$, that is:

$$s^{(t+1)} = \begin{cases} ^*s^{(t)} & \text{with probability } p = \min\left\{1, \left(M(D_{s^{(t+1)}}) / M(D_{s^{(t)}})\right)^\beta\right\}, \\ s^{(t)} & \text{otherwise} \end{cases}, \quad (2)$$

3. repeat steps (1) and (2) mq times.

For a detailed review of spatially balanced samples, the reader interested can refer to Benedetti et al. (2014).

3 Preliminary results and conclusions

The methods described in the previous sections have been verified on LULC data. The aim is to compare the schemes highlighting advantages and drawbacks of each method. In order to reach this objective, we have constructed an artificial data set based on the well-known LUCAS survey. The Land Use/Cover Area frame Survey (LUCAS, Gallego and Delincè 2010) is a task by EUROSTAT that was initially defined to deliver, on a yearly basis, European crop estimates.

Our artificial data set has been built considering 2012 as reference year and Italy as country under investigation. A sample of 21,013 points was drawn from a population (or a first phase sample) obtained overlaying a regular grid of points selected every 2 km to the Italian national boundaries map. This sample constitutes

our reference population on which it is possible to test the different spatially balanced samples methods. We have performed a simulated exercise, comparing the different results obtained. The first preliminary results are very encouraging. Note that the gain in terms of the relative efficiency of the sample mean compared with simple random sampling for each design estimated in many replicated samples is remarkable (at least 20%).

Table 1: Relative efficiency of the area estimates for each Land Use (MSE/MSE_{SRS}) and for each design estimated in 10,000 replicated samples in the Emilia Romagna for $n=600$.

Method	Agriculture	Forestry	Urban	Unused	Other
CUBE1	0,893	0,858	1,000	0,965	1,005
CUBE2	0,893	0,854	1,000	0,952	0,997
GRTS	0,780	0,699	0,881	0,834	0,818
LPM1	0,893	0,854	1,000	0,952	0,997
LPM2	0,780	0,699	0,881	0,834	0,818
SCPS	0,780	0,699	0,881	0,834	0,818
PPD1	0,893	0,854	1,000	0,952	0,997
PPD2	0,780	0,699	0,881	0,834	0,818
PPD5	0,780	0,699	0,881	0,834	0,818
PPD10	0,730	0,651	0,841	0,767	0,756

References

1. Benedetti, R, Piersimoni, F, Postiglione, P.: Sampling spatial units for agricultural surveys. *Advances in Spatial Science Series*. Springer (forthcoming) (2014)
2. Benedetti, R, Palma, D: Optimal sampling designs for dependent spatial units. *Environmetrics*, 6: 101-114. (1995)
3. Benedetti, R, Marchetti, S, Piersimoni, F, Pratesi, M.: Spatial network sampling in small area estimation. *Proceedings 59th ISI World Statistics Congress*, 25-30 August 2013, Hong Kong (Session IPS024), pp. 241-246. (2013)
4. Bondesson, L, Thorburn, D.: A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35: 466-483. (2008)
5. Breidt, FJ, Chauvet, G: Penalized balanced sampling. *Biometrika*, 99: 945-958 (2012)
6. Deville, JC, Tillé, Y: Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85: 89-101 (1998)
7. Deville, JC, Tillé, Y: Efficient balanced sampling: the cube method. *Biometrika*, 91: 893-912. (2004)
8. Gallego, J, Delincè, J: The European land use and cover area-frame statistical survey, In: *Agricultural survey methods*, Benedetti R, Bee M, Espa G, Piersimoni F (eds), John Wiley & Sons Ltd, Chichester, UK, pp. 151-168. (2010)
9. Grafström, A: Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142: 139-147. (2012)
10. Grafström, A, Lundström, NLP, Schelin, L: Spatially balanced sampling through the pivotal method. *Biometrics*, 68: 514-520. (2012)
11. Rogerson, P, Delmelle, E: Optimal sampling design for variables with varying spatial importance. *Geographical Analysis*, 36: 177-194. (2004)
12. Stevens, DLJr, Olsen, AR: Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99: 262-278. (2004)