

Efficient statistical sample designs in a GIS for monitoring the landscape changes

Elisabetta Carfagna¹ Patrizia Tassinari², Maroussa Zagoraiou¹,
Stefano Benni², Daniele Torreggiani²

¹Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy, elisabetta.carfagna@unibo.it, maroussa.zagoraiou@unibo.it

²Department of Agricultural Economics and Engineering, Spatial Engineering Division, University of Bologna, v.le Fanin 48, 40127, Bologna, Italy

Abstract: The process of land planning, addressed to operate the synthesis between development aims and appropriate policies of preservation and management of territorial resources, requires detailed analysis of the territory. Generally this analysis is carried out on the whole area of interest by making use of data stored in Geographic Information Systems (GIS), but a detailed analysis of changes in the landscape can be carried out only on a sample of the whole territory. A very efficient sample design is the adaptive sequential procedure with permanent random numbers (ASPRN) proposed by Carfagna and Marzioletti (2009). However, ASPRN obliges the environmental engineer to jump from one side to another of the study area. Carfagna (2007) proposed a two-steps selection procedure with permanent random numbers (TSPRN) which allows to overcome the operational drawbacks of the ASPRN. In this paper we apply both procedures to a study area, in order to compare them from an operational viewpoint and to set up a methodology which enables an efficient estimate of the change of main landscape features on large areas.

Keywords: Geographic Information Systems, permanent random numbers, adaptive sampling, sequential sampling.

1. Introduction

The preservation and management of territorial resources requires detailed knowledge of the landscape and of the evolutionary dynamic that produced its arrangement, as well as the monitoring of the changes occurred in the various landscape features. For these purposes, the analysis of different kinds of geographical data referred to manifold time steps is needed and is currently carried out by making use of databases and maps available for the whole territory. These data are stored and elaborated in Geographic Information Systems (GIS) which have made possible to compile geographical databases with highly detailed content covering large spatial areas. GIS also enable the construction of multi-layer and multi-temporal archives in order to analyse changes in land-use/land-cover patterns. However, when a more detailed measure of the actual variations of main landscape features is required, this detailed analysis can be carried out only on a sample of the whole territory.

Our aim is to provide a statistically sound method for monitoring the landscape transformations in rural areas, for verifying the actual progress in the planned developments and for assessing the exploitation of landscape resources in order to

define guidelines for their management. In order to set up a methodology which enables an efficient estimate of the change on large areas, we have chosen a target study area (787 km²) situated in the eastern part of the Bologna province, Italy, which comprises the group of ten municipalities known as the “Nuovo Circondario Imolese”, a first study can be found in Tassinari et al. 2008. The analysis is focused on the changes in the landscape features in the period from 1975 to 2005. The parameter we estimate is the change of building cover density that is the difference between the area covered by buildings in 2005 and in 1975 divided by the land area. This parameter is proven to be useful in analyses supporting the formulation of the strategic and operational planning tools defined by the most recent legislative purviews concerning land government (Benni et al., 2009).

The analysis we have performed uses a discretization of the territory into areal units defined according to the division borders of the most recent population and housing census (Istat, 2001). Thus, the area frame we have used is based on irregular physical boundaries. The need to maximize the precision of the estimate of landscape change suggests the adoption of a stratification which should be able to distinguish areas characterised by different rates of change of main landscape features. The stratification is based on pedology, elevation and land-use/land-cover. Particularly, a series of GIS operations have been performed and the final result comes out from combinations of land-use/land-cover classes and land suitability classes (see fig.1).

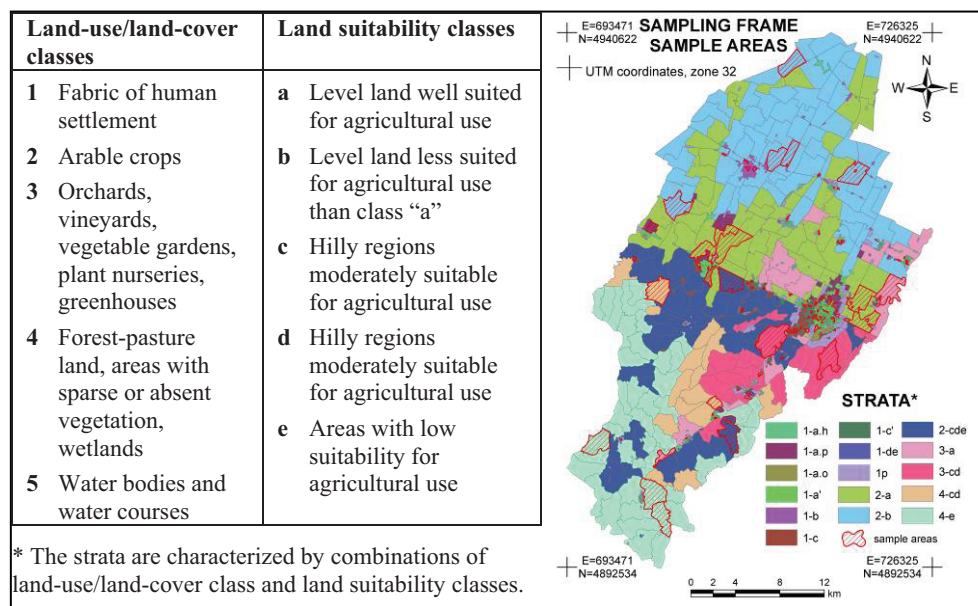


Figure 1: Table of land-use/land-cover classes and land suitability classes considered in the stratification; map of the sampling frame and of sample areas.

Each unit of the stratum characterized by class 1 of land-use and class *a* of land suitability was reclassified according to the attribution of the majority of its surface to areas already urbanized at the beginning of the study period (stratum 1-a') or to areas urbanized during that period (1-a). The same procedure was adopted for the units in stratum 1-c. A pilot sample showed a very high variability in stratum 1-a; thus we decided to subdivide stratum 1-a depending on whether the final destination was predominantly residential (1-a.h), productive (1-a.p), or any other type of urban use (1-a.o).

Since data acquisition and in depth data analysis are time consuming, it is very important to use a very efficient sample design. The efficiency of a sample design can be improved adopting optimal allocation and an adaptive sequential procedure. However, we remind that adaptive sequential procedures do not allow unbiased estimates of parameters because their efficiency is due to selection procedures depending on previously selected units and stopping rules based on the parameters to be estimated.

2. Description of Methods and Comparisons

Carfagna and Marzialetti (2009) proposed a particular kind of adaptive sequential procedure (ASPRN) which involves the permanent random numbers method (Ohlsson 1995) and allows unbiased and efficient estimates. In this sequential procedure the sample size per stratum is dependent on the previously selected units but the sample selection is not. Moreover, the stopping rule is not based on the estimate of the parameter, it is based only on its standard deviation. Therefore the authors proved that the sequential procedure they proposed allows unbiased and efficient estimates of the parameters of interest.

Carfagna and Marzialetti did not use real data, they performed several simulations and noticed that: i) with increasing sample size, the estimator tends to converge to the value of the population parameter, ii) the standard deviation of the estimate decreases as the sample size increases, although this decrease is not always strictly monotone, iii) the ASPRN is always more efficient than:

- 1) stratified sampling with proportional allocation
- 2) stratified random sampling in two phases proposed by Thompson and Seber (1996).

On the other side, from our operational viewpoint, ASPRN requires organization and continuous interaction between the statisticians and the environmental engineers for choosing, at each step, the next area sample unit to be selected. It implies therefore segmentation of data acquisition and pre-processing. Instead, a two-steps selection procedure with permanent random numbers (TSPRN) proposed by Carfagna (2007) minimizes this interaction and reduces this cost and burden but is less efficient than ASPRN although more efficient than Thompson and Seber's.

In this paper we make a comparison of ASPRN with the TSPRN in order to test how much ASPRN is more efficient than TSPRN when real data are used.

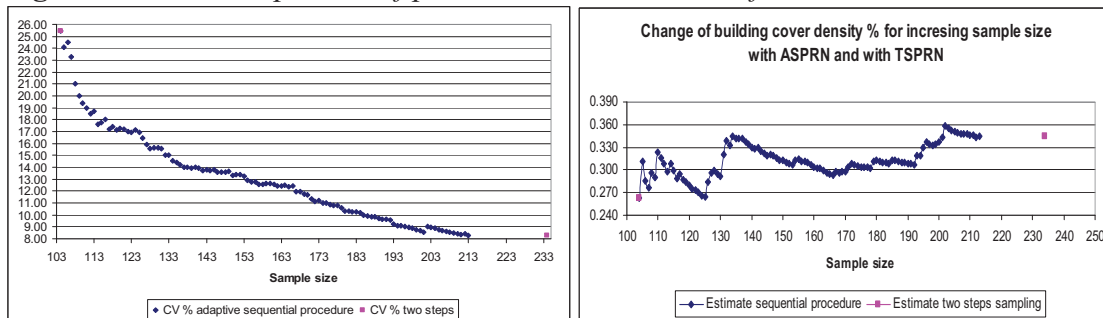
First of all, we have adopted the TSPRN for estimating the building cover density and we have selected the pilot or first step sample which is constituted of 104 census sections. Then, we have estimated the change in built areas and the standard deviation within each stratum in order to compute the number of sample units needed for reaching a coefficient of variation of the estimate equal to 10% (234). Thus, we have selected 130 more units with an allocation as near as possible to Neyman's one of 234 units. The coefficient of variation (CV) we have obtained is 8.3%, lower than the target CV, this means that the first phase sample overestimated the standard deviation in some strata.

Then, we have adopted the ASPRN and we have noticed that a CV equal to 10% is already reached with a sample size equal to 180 and the same CV obtained with the two steps procedure (8.3%) is reached with 213 sample units; thus the adaptive sequential procedure is much more efficient than the two steps sampling with permanent random numbers, as showed in figure 2a, although the trend of the CV% is not strictly

monotone. The efficiency of ASPRN is mainly due to the fact that it allows continuous updating of the estimate of the standard deviation within the strata.

If we compare the estimates of the change in built areas cover given by the two procedures when the CV for both estimates is about 8%, we notice that their difference is very small, it's about 81 square meters (2.9%), that is 0.01% of the study area. The estimate of the change in built areas cover given by ASPRN with CV equal to 10% is similar to the estimate with CV equal to 8%, the difference is 246 square meters (9%), that is 0.03% of the study area. Instead the difference between the estimates given by the pilot sample and by the ASPRN with CV equal to 8.3% is 642 square meters (23.7%), that is 0.08% of the study area, as showed in fig 2b.

Fig. 2a and 2b: Comparison of precisions and estimates of TSPRN and ASPRN



2a: CV% for TSPRN and for ASPRN with increasing sample size

2b: Change of cover density per cent with TSPRN and with ASPRN for increasing sample size

References

- Benni, S., D. Torreggiani, P. Tassinari (2009) Sistemi integrati per l'analisi delle risorse del territorio rurale: alcune note di metodo. *Agribusiness Paesaggio & Ambiente* 12, 16-22.
- Carfagna E. (2007) Crop area estimates with area frames in the presence of measurement errors, *Proceeding of ICAS-IV, Fourth International Conference on Agricultural Statistic*. Invited paper, Beijing, 22-24 October 2007.
- Carfagna E. and Marzioletti J (2009) Sequential Design in Quality Control and Validation of Land Cover Data Bases *Journal of Applied Stochastic Models in Business and Industry*, Volume 25, Issue 2, pp. 195-205, DOI: 10.1002/asmb.742.
- Istat, Italian statistic board (2001) 14th Census of Population and Housing. Istat, Italy.
- Ohlsson E. (1995) Coordination of Samples Using Permanent Random Numbers, in *Business survey methods*, Cox B. G., Binder D.A., Chinnapa B.N., Christianson A., Colledge M.J., Kott P.S. (Eds.), Wiley, New York, 153-169.
- Tassinari P., Carfagna E., Benni S., Torreggiani D. (2008) Wide-area spatial analysis: A first methodological contribution for the study of changes in the rural built environment, *Biosystems Engineering*, 100 (3), 435-447.
- Thompson S.K., Seber G.A.F. (1996) *Adaptive Sampling*, Wiley, New York.