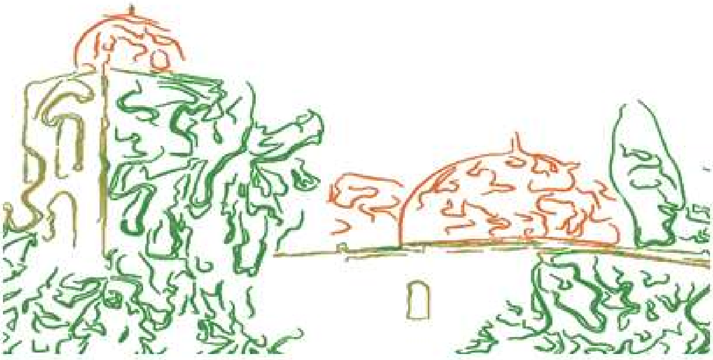
PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023



GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
-
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation

Marcelli A[1,*], Di Biase RM[2,3], Corona P[4], Stehman SV[5] and Fattorini L[2]

[1] *Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Italy; agnese.marcelli@unitus.it*
[2] *Department of Economics and Statistics, University of Siena, Italy; rosa.dibiase@unisi.it, lorenzo.fattorini@unisi.it*
[3] *NBFC, National Biodiversity Future Center, Palermo, Italy*
[4] *CREA, Research Centre for Forestry and Wood, Arezzo, Italy; piermaria.corona@crea.gov.it*
[5] *Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, New York, United States; svstehma@syr.edu*
[*] *Corresponding author*

Abstract. *Land use/land cover mapping is usually performed by classifying satellite imagery for the whole survey region using classification algorithms implemented with training data. Subsequently, probabilistic samples are usually implemented with the main purpose of assessing the accuracy of these maps by comparing the map class and the ground condition determined for the sampled units. The main proposal of this paper is to directly exploit these probabilistic samples to estimate the land use/land cover class at any location of the survey region in a design-based framework by the well-known nearest-neighbour interpolator. For the first time, the design-based consistency of nearest-neighbour maps is theoretically proven and a pseudo-population bootstrap estimator of their precision is proposed and discussed. These nearest-neighbour maps provide the ability to place mapping within a rigorous design-based inference framework, in contrast to most traditional mapping approaches which often are implemented with no inferential basis or by necessity (due to lack of a probabilistic sample) model-based inference. A simulation study is performed on an estimated land use map in Southern Tuscany (Italy). The Italian land use map arising from the IUTI surveys is considered as case study.*

Keywords. *Spatial sampling; Spatial interpolation; Consistency; Pseudo-population bootstrap; Simulation study.*

1 Introduction

Land use/land cover (LULC) refer to the composition of land surface by classes with different characteristics. Their knowledge is of basic importance in a wide range of human activities such as scientific research, landscape management, and political decisions.

In the last few decades, LULC mapping is, in most cases, the product of satellite data, whose costs are continuously decreasing owing to the continuous improvement of remote sensing technologies and the launch of new satellites. In this framework, satellite data acquired for the whole survey region are transformed into LULC labels using image classification algorithms, such as decision trees. At the end of these classification procedures LULC maps are produced. These classification algorithms are typically applied using a "training sample" that consists of a set of locations for which the reference LULC class labels have been obtained where reference labels are considered the best available assessment of ground conditions. These reference labels along with the remote sensing-based covariates are then used to create the predictive models that produce the class labels at all locations in the survey region. The training sample is typically not a probability sample. Finally, the last step of any LULC mapping is an assessment of the quality of the resulting map. In the traditional assessments, the map labels are compared to the reference labels. Because the reference labels are typically more expensive and time consuming to obtain, assessments are usually performed using the reference class labels recorded for a sample of locations.

The main proposal of this paper is to directly exploit reference data, if sample locations are selected by a probabilistic sampling scheme, to construct LULC maps in a design-based framework, without the use of satellite data. Towards this end, we have adapted the design-based treatment of the well-known nearest-neighbour (NN) interpolator to the interpolation of LULC classes, together with a quantification of the map precision at any point of the survey region in terms of the probability of assigning a wrong class. Then, we have adapted the bootstrap procedure for estimating root mean squared errors of NN interpolations, to the estimation of the error probabilities at any point of the survey region. In this way, the procedure provides a map depicting the design-based precision at each point.

The paper is organized as follows: in section 2 some preliminaries and notations are given about LULC maps, together with the theoretical results on the design-based NN interpolator. In section 3 the estimators of area coverage arising from the maps are theoretically compared with the traditional ones. In section 4, a simulation study is performed on an estimated land use map in Southern Tuscany (Italy), while in section 5, the Italian land use map arising from the IUTI survey is considered as case study. Final remarks are provided in section 6.

2 Nearest neighbour interpolation of land use / land cover maps

Let A be the survey region of size $|A|$, partitioned into K LULC classes $c_1,...,c_K$. For any point $p \in A$, let $y(p)$ be the LULC class at p in such a way that $\{y(p), p \in A\}$ defines the LULC map of A. The survey region A is partitioned into K sets, $D_1,...,D_K$, where $D_k = \{p : p \in A, y(p) = c_k\}$ denotes the portion of the survey region occupied by class c_k.

In most cases, the reference LULC classes are recorded at n sample locations $P_1,...,P_n$ and then they are estimated at any unsampled location $p \in A$. We estimate LULC surfaces in a design-based framework in such a way that the properties of the resulting maps are solely determined by the probabilistic sampling scheme implemented to select the sample locations.

For the interpolation of LULC classes, we modified the NN interpolator proposed by [2] for mapping quantitative variables. Therefore, the LULC class at a non sampled location $p \in A$ is estimated by $\hat{y}(p) = y\left(P_{NN(p)}\right)$, where $P_{NN(p)} = argmin_{i=1,...,n}\|p - P_i\|$ and $\|\|$ denotes a norm in $\mathbb{R}^2$. Based on this estimator, we propose indices suitable for quantifying the precision of categorical maps. Subsequently, based on these indices, we derive conditions ensuring the design-based consistency of NN maps. To this purpose, for each point $p \in A$ denote by $z(p)$ the random variable that is equal to 1 if $\hat{y}(p) \neq y(p)$ and equal to 0 otherwise. Therefore, the expectation $Err(p) = E\{z(p)\} = Pr\{\hat{y}(p) \neq y(p)\}$ gives the probability of providing a wrong interpolation at p and as such it can be taken as a suitable index of the NN interpolator precision at p. We have proven that consistency holds at any interior point of the map under suitable sampling schemes, such as uniform random sampling (URS), tessellation stratified sampling (TSS) and systematic grid sampling (SGS).

Regarding the estimation of map precision, we adapted once again the procedure by [2] to the estimation of $Err(p)$. The procedure is based on a pseudo-population bootstrap (PPB) approach, in which a pseudo-population likely to resemble the true population is constructed. Bootstrap samples are then selected from the pseudo-population using the same sampling scheme adopted in the survey. Because under suitable schemes the estimated maps converge to the true map, the bootstrap distributions of the NN interpolator achieved by resampling from these maps should converge to the true distributions, also providing consistent estimators of $Err(p)$.

To obtain insights about the performance of this bootstrap procedure, we considered this bias: $bias_B(p) = E\{\widehat{Err}^*_B(p)\} - Err(p)$, where $\widehat{Err}^*_B(p) = \frac{1}{B}\sum_{b=1}^{B} z^*_b(p)$ is the bootstrap estimator of $Err(p)$.

3 Area estimation of land use / land cover classes

Given the K LULC classes that are present on A, a common use of LULC analysis has been the estimation of class coverages $\gamma_1,...,\gamma_K$, where $|D_k| = \int_A I(p \in D_k)dp$ is the area of the k-th class and $\gamma_k = |D_k|/|A|$ is the proportion of the survey region covered by the k-th class.

Estimation of class coverage has been traditionally performed by counting the sample locations within the classes, say $n_1,...,n_K$, and considering the frequencies of these locations in the sample

$$f_k = \frac{n_k}{n}, k = 1,...,K$$

The corresponding variance estimator is

$$\hat{V}_k^2 = \frac{f_k(1-f_k)}{n-1}$$

Under URS and TSS this estimator is proven to be unbiased, asymptotically normal, and consistent, with an unbiased variance estimator. Moreover, under TSS, $\hat{V}_k^2$ is invariably smaller than the variance under URS. Less appealing results are achieved under SGS, where $\hat{V}_k^2$ is not invariably smaller than the variance under URS.

The NN map procedure previously described creates an alternative way to estimate the area of each class. Indeed, denoting by $\hat{D}_k = \{p : p \in A, \hat{y}(p) = c_k\}$ the set where the LULC class is estimated to be c_k, we can estimate $|D_k|$ by $|\hat{D}_k| = \int_A I(p \in \hat{D}_k)dp$ in such a way that

$$\hat{\gamma}_k = \frac{|\hat{D}_k|}{|A|}, k = 1,...,K$$

is the map estimator of γ_k. The bootstrap procedure described above can be exploited to estimate the design-based variance of $\hat{\gamma}_k$ through the the bootstrap estimator of the root mean squared error

$$\widehat{rmse}^*_{k,B} = \left\{ \frac{1}{B} \sum_{b=1}^{B} (\hat{\gamma}^*_{k,b} - \hat{\gamma}_k)^2 \right\}^{1/2}$$

where $\hat{\gamma}^*_{k,b}$ denotes the map estimate of γ_k achieved from the bootstrap sample b.

Note that we are not suggesting the use of the second method, as the traditional method is more straightforward and it has several appealing properties. Rather, a problem arises when the estimates achieved from the map differ substantially from the traditional estimates, a situation that would create a dilemma in a reporting phase. Even if the problem disappears asymptotically because both methods provide consistent estimators, it is however necessary to verify that the two methods do not differ in a relevant way for finite moderate sample sizes.

4 Simulation study

To check the performance of the proposed methodology we conducted a simulation study on a 10 km×10 km region located in Southern Tuscany. The quadrat was selected from the land use map of the whole of Italy for the year 2008 that was estimated from a TSS sample of 1,217,032 points during the land use pure panel survey IUTI. The IUTI project, the Italian acronym of "Inventario dell' Uso delle Terre d'Italia", was carried out by the Italian Ministry of Environment and Protection of Land and Sea. Its main purpose was the implementation of the national greenhouse gas assessment under the Kyoto Protocol framework. In accordance with a land use classification based on the greenhouse gas reporting system introduced by the Good Practice Guidance for Land Use, Land Use Change and Forestry [3], the IUTI coarsest classification adopted six land use classes: Forest land (1), Cropland (2), Grassland (3), Wetland (4), Settlements (5) and Other lands (6). The quadrat considered for the simulation study only included five of these six land use classes.
Sampling was simulated selecting $n = 100; 400; 1,600; 10,000$ locations within the quadrat by URS, TSS and SGS. For each combination of sampling scheme and sample size, $R = 10,000$ samples were independently selected. Because it was impossible to perform interpolation in the continuum of the survey region, we created a regular grid G of 201×201 nodes and we performed interpolation at each node of the grid. At each simulation run, the classes at the selected points were recorded from the map and the NN interpolator was performed at each node of the grid, assigning to each node the land use class of the nearest location in the sample. Moreover, at each simulation run, $B = 1,000$ bootstrap samples were independently selected using the same scheme adopted to select the original sample. Then, the classes at the selected points in the bootstrap samples were assigned from the map estimated from the original sample and the NN interpolation was performed for each node to compute $\widehat{Err}^*_B(p)$. Finally, to check the discrepancies from the traditional estimators of area coverage and those achieved from the map, f_k and $\hat{V}^2_k$ together with $\hat{\gamma}_k$ and $\widehat{rmse}^*_{k,B}$ were computed for each land use class.
At the completion of the simulation runs, for each combination of sampling scheme and sample size, we empirically determined the values of $Err(p)$ and the expectation of $\widehat{Err}^*_B(p)$. From these quantities we achieved the values of $bias_B(p)$. As to coverage estimation, we empirically computed the expectation and the standard error of the traditional estimators

$$E(f_k) \sim \frac{1}{R}\sum_{i=1}^{R} f_{k,i}, k = 1, ..., K$$

$$SE(f_k) \sim \left\{\frac{1}{R}\sum_{i=1}^{R}(f_{k,i} - \gamma_k)^2\right\}^{1/2}, k = 1, ..., K$$

and the expectation of the standard error estimators

$$E(\hat{\upsilon}_k) \sim \frac{1}{R}\sum_{i=1}^{R} \hat{\upsilon}_{k,i}, k = 1, ..., K$$

The expectation $E(\hat{\gamma}_k)$ and the root mean squared error $SE(\hat{\gamma}_k)$ of the map estimator were empirically computed by substituting in the formulas above f_k with $\hat{\gamma}_k$, and the expectation of the bootstrap estimator of root mean squared error was empirically determined by

$$E(\widehat{rmse}^*_{k,B}) \sim \frac{1}{R}\sum_{i=1}^{R} \widehat{rmse}^*_{k,B,i}, k = 1, ..., K$$

Finally, we computed the ratio $R_k = E(\hat{\gamma}_k)/E(f_k)$ to quantify the discrepancies between the two estimators. The results of the simulation study confirm the theoretical results. The NN interpolator $\hat{y}(p)$ proves to be consistent under the three sampling schemes with error probabilities that quickly decrease as the sample size increases. Comparatively, TSS proves to be superior to URS and SGS. Cumulative frequencies of errors show that, even with the smallest sample size of $n = 100$, in about 50% of the survey region the error probabilities are smaller than 0.25 and that percentage quickly increases with sample sizes. Regarding the bootstrap estimator of the error probabilities, the minimum values of bias may appear quite discouraging with minimum values always smaller than -0.5 even for the very large sample size of 10,000 points. However, as sample size increases, in large zones of the map the estimator is unbiased and large underestimation is restricted to isolated points along the borders where changes of class occur. Regarding the estimation of areas, the similarity in the estimates arising from the traditional and map methods confirms that irrespective of the coverage extent, sampling scheme and sample size, the two estimators show very similar expectations with their ratio invariably equal to one. Even if our purpose is not the comparison of the two methods, as sample size increases, the map estimator tends to outperform the traditional estimator in terms of precision, with improvements that are sometimes substantial.

5 Case study: the IUTI land use survey

As case study we considered the whole sample of quadrats selected by TSS for the IUTI project (see section 4). The primary aim of the survey was the estimation of land use proportions together with the estimation of the corresponding standard errors [1], but no attempt was made to make inference on the land use maps. Therefore, IUTI data were here adopted for the first time to provide a map with accompanying inference. To this purpose, the NN interpolator was adopted to estimate $y(p)$ for each location p in the network of 2,800,360 nodes within the Italian territory (see Figure 1a). From the estimated map, $B = 1,000$ bootstrap samples of size 1,217,032 were selected by the same TSS scheme adopted to select the original sample, giving rise to as many bootstrapped maps from which $\widehat{Err}^*_B(p)$ was derived for each location in the network of nodes where estimation was performed (see Figures 1b and 1c). From Figures 1b and 1c, the precision of the strategy is apparent. Estimated errors smaller than 10% occur on about half of the Italian territory. Greater uncertainties arise along the boundaries corresponding to land cover class changes with estimated errors greater than 25% occurring on about 25% of the territory. Estimated errors smaller than 50% occur almost everywhere. The convergence of the traditional and map coverage estimators for intensive sampling effort is consistent with the simulation results. Also in this case, the map estimator had better precision than the traditional estimator.

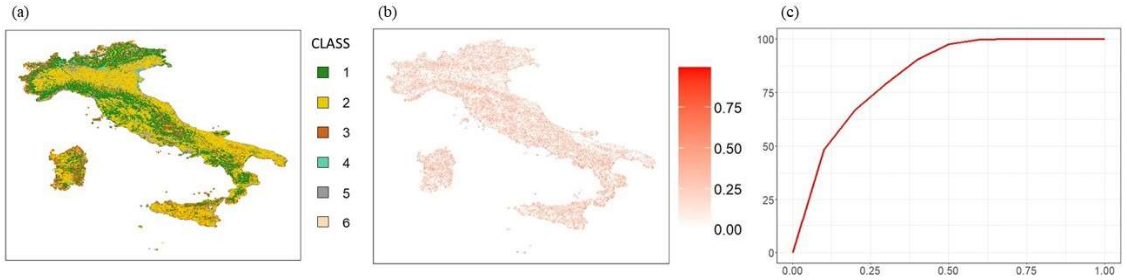


Figure 1: (a) Italian map of the six land use classes estimated from the IUTI TSS sample; (b) Map of the estimates of the error probabilities achieved by $B = 1000$ bootstrap samples; (c) Cumulative frequencies of the estimates of the error probabilities.

6 Final remarks

Design-based estimates of LULC maps are performed using a probabilistic sampling scheme. Mapping is performed from a design-based framework adopting the NN interpolator. Sampling schemes ensuring the design-based consistency of maps are identified and a bootstrap estimator of the design-based precision, measured by the probability of assigning the true LULC class at any point of the survey region is proposed. This allows for the novel ability to construct a map of estimated precision (within a design-based inference setting) to accompany the resulting LULC map. The design-based nature of the approach is appealing because the properties of the LULC maps stem from the sampling scheme implemented to select the sample, rather than from model assumptions.

Several drawbacks of the procedure indicate directions for future developments. First, the mapping performed by NN interpolation exploits only information arising from space without taking advantage of the knowledge of remote-sensing covariates (i.e., auxiliary variables) often available for the whole survey region. A second concern is that the areas of LULC classes can be estimated by traditional methods and also from the resulting design-based maps. Therefore, there may be a dilemma of which estimates to report if the map estimates differ greatly from the traditional estimates. This second concern is less crucial than the first, because both estimators arising from the two procedures are design-based consistent in such a way that the differences in the resulting estimates tend to disappear asymptotically and, as demonstrated in the simulation study, they tend to be negligible for finite, moderate sample sizes. Although both these concerns necessitate additional investigations and further developments are needed to improve the practical utility of the design-based maps, this article establishes the conceptual basis for a unified approach to inference for mapping and area estimation of land cover and land cover change.

Acknowledgments. The authors acknowledge the funding by PRIN 2020 (cod 2020E52THS) - Research Projects of National Relevance funded by the Italian Ministry of University and Research entitled: "Multi-scale observations to predict Forest response to pollution and climate change" (MULTIFOR, project number 2020E52THS). The authors also acknowledge the support of NBFC to University of Siena, funded by the Italian Ministry of University and Research PNRR, Missione 4 Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, Project CN00000033.

References

[1] Corona, P., Barbati, A., Tomao, A., Bertani, R., Valentini, R., Marchetti, M., Fattorini, L. and Perugini, L. (2012). Land use inventory as framework for environmental accounting: an application in Italy. *iForest* **5**, 204–209.

[2] Fattorini, L., Marcheselli, M., Pisani, C. and Pratelli, L. (2021). Land use inventory as framework for environmental accounting: an application in Italy. Design-based properties of the nearest neighbour spatial interpolator and its bootstrap mean squared error estimator. *Biometrics*. Online first.

[3] International Panel on Climate Change (2003). Good Practice Guidance for Land Use, Land Use Change and Forestry. *IPCC National Greenhouse Gas Inventories Program*.