The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3133–3152
https://doi.org/10.1214/23-AOAS1754

DESIGN-BASED MAPPING OF LAND USE/LAND COVER CLASSES WITH BOOTSTRAP ESTIMATION OF PRECISION BY NEAREST-NEIGHBOUR INTERPOLATION

BY AGNESE MARCELLI[1,a], ROSA MARIA DI BIASE[2,b], PIERMARIA CORONA[3,c], STEPHEN V. STEHMAN[4,d] AND LORENZO FATTORINI[5,e]

[1]*Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia,* [a]*agnese.marcelli@unitus.it*
[2]*Department of Sociology and Social Research, University of Milano Bicocca,* [b]*rosamaria.dibiase@unimib.it*
[3]*CREA, Research Centre for Forestry and Wood,* [c]*piermaria.corona@crea.gov.it*
[4]*Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry,* [d]*svstehma@syr.edu*
[5]*Department of Economic and Statistics, University of Siena,* [e]*lorenzo.fattorini@unisi.it*

Land use/land cover mapping is usually performed by classifying satellite imagery (e.g., Landsat, Sentinel) 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 (i.e., categorical variables) 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)—taken as the true map—to check the finite-sample performance of the proposal as well as the matching of the area coverage estimates arising from the map with those achieved by traditional estimators. The Italian land use map arising from the IUTI surveys and the U.S. land cover map arising from the LCMAP program are considered as case studies.

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. The National Research Council (2001) of the United States proclaimed land change dynamics as one of the major challenges facing environmental science, and Turner, Lambim and Reenberg (2007) further highlighted the critical importance of land change science. As LULC varies in space, its mapping is an essential part of LULC analysis.

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 addition to the increasing availability of remote sensing information, the improvement of computational facilities has allowed the processing of large data sets even by small teams of researchers (e.g., Cihlar (2000), Aldoski et al. (2020)). In this framework satellite data, acquired for the whole survey region,

Received October 2022; revised February 2023.
Key words and phrases. Spatial sampling, spatial interpolation, consistency, pseudo-population bootstrap, simulation, case studies.

are transformed into LULC labels by a wide variety of image classification algorithms such as decision trees (e.g., Hansen, Dubayah and De Fries (1996)), neural networks (e.g., Yool (1998)), fuzzy classification (e.g., Mannan, Roy and Ray (1998)), mixture modelling (e.g., van der Meer (1995)), and many others (see, e.g., Nguyen et al. (2020)). At the end of these classification procedures, LULC maps are produced, usually referred to as map data. 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 but instead is selected because of convenience or judgment (e.g., Nguyen et al. (2020), Section 2.3.2). Finally, the last step of any LULC mapping is an assessment of the quality of the resulting map. As stated by Cihlar (2000), "No land cover classification project would be complete without an accuracy assessment." In the traditional assessments, the map labels are compared to the reference labels where the reference labels represent the best assessment of ground condition. Because the reference labels are typically more expensive and time consuming to obtain, it is impractical to completely census the entire survey region, so assessments are usually performed using the reference class labels recorded for a sample of locations. The most widely used summarization of these assessments is the confusion matrix, that is, a cross-tabulation of the predicted classes at the sample points against the reference class labels. Several synthetic measures are then derived from the confusion matrix, in particular, measures of the accuracy regarding each class (Stehman (1997)). If sample locations are selected by a probabilistic sampling scheme, the reference data also allow—as a secondary aim—for the design-based estimation of the areas of LULC classes. There is a vast literature dealing with the sampling strategies suitable for performing accuracy assessments (e.g., Fitzpatrick-Lins (1981), Stehman and Czaplewski (1998), Nusser and Klaas (2003), Fattorini, Marcheselli and Pisani (2004), Stehman (2009), Stehman and Foody (2019)).

A serious drawback of these assessment procedures is that they invariably provide estimates of aggregate areas correctly classified or misclassified giving no insight on the precision of the map at any single point of the survey region. Uncertainty measures at any point of the survey region have been proposed in literature and reviewed by Stehman and Foody (2019, Section 4.7). These approaches are inherently model-based representations of per location uncertainty. For example, the classification procedure of satellite data may produce a measure of uncertainty for each location, such as the class membership probabilities of a maximum likelihood classifier, or spatial modelling techniques could be applied to the validation sample data to generate a spatial map of uncertainty (e.g., Comber et al. (2013); Khatami, Mountrakis and Stehman (2017); Rodríguez-Jeangros et al. (2016), (2017)). In general, the importance of inference for LULC mapping applications has been emphasized (McRoberts (2011)), and this includes inference for estimates of area coverage as well as for precision of the map class labels at points in the study region.

The main proposal of this paper is to directly exploit reference data to construct LULC maps in a design-based framework without the use of satellite data. Toward this end, we have adapted the design-based treatment of the well-known nearest-neighbour (NN) interpolator, recently proposed by Fattorini et al. (2022) for mapping a quantitative variable, to the interpolation of LULC classes. In addition to providing the NN LULC map, we have proposed 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, originally proposed by Fattorini et al. (2022) 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. Whereas the traditional approach to mapping requires a training sample (typically not a probabilistic sample) to produce the classification and an additional, separate probabilistic sample to assess the accuracy of the resulting map. Conversely, our design-based proposal requires only a probabilistic sample to produce the map and to estimate its precision. The design-based assessment of the map precision at any point of the survey region also differs from the previously mentioned model-based procedures usually adopted for remote sensing-based maps.

As stated before, probabilistic samples are obtained in most LULC satellite-based surveys to assess the accuracy of the resulting maps, and such a sample is, therefore, commonly available to produce a design-based map. Moreover, probabilistic samples from which we can achieve design-based NN maps, can be obtained from the first phases of most forest inventories. Usually, forest inventories are performed by selecting a first-phase sample of locations evenly distributed throughout the whole region, often by using spatially stratified or systematic sampling schemes. Subsequently, the classes of these points are recorded from aerial photos, thus providing the information for performing the NN interpolation, while only forest points are considered in subsequent phases (e.g., Fattorini (2015); Rizzo and Gasparini (2022)).

In Section 2 some preliminaries and notations are given about LULC maps. In Section 3 the theoretical results on the design-based NN interpolator achieved by Fattorini et al. (2022) regarding quantitative variables are extended to categorical variables to prove the design-based consistency of the LULC maps and to provide a pseudo-population bootstrap estimator of the precision at any point of the survey region and within subregions. In Section 4 the estimators of area coverage arising from the maps are theoretically compared with the traditional ones. In Section 5 a simulation study is performed on an estimated land use map in Southern Tuscany (Italy)—taken as the true map—to check the finite-sample performance of the proposal as well as the matching of the area estimates arising from the map with those achieved by traditional estimators. In Section 6 the Italian land use map, arising from the IUTI survey (Corona et al. (2012)), and the U.S. land cover map, arising from the LCMAP program, are considered as case studies. Final remarks are provided in Section 7. Technical details are provided by Appendices A and B in the Supplementary Material, Marcelli et al. (2022b).

2. Preliminaries and notation. Denote by A the survey region of size $|A|$, partitioned into K LULC classes $c_1, \ldots, c_K$. For any point $p \in A$, let $y(p)$ be the LULC class at p in such a way that

$$\{y(p), p \in A\} \tag{1}$$

defines the LULC map of A. As exemplified in Figure 1, the survey region A is partitioned into K sets, $D_1, \ldots, D_K$, where $D_k = \{p : p \in A, y(p) = c_k\}$ denotes the portion of the survey region occupied by class c_k. In a design-based approach, the LULC map is a fixed, unknown characteristic of the survey region that must be estimated from the classes recorded at the sampled locations rather than a realization of a spatial model, as assumed in model-dependent approaches.

Regarding (1), it is at once apparent from Figure 1 that the map comprises disjoint regions having constant values within regions, jumping from a class to another along borders, which are sets of measure 0. Therefore, even with this categorical nature of the data, in mathematical terms any LULC map can be viewed as a function that is continuous almost everywhere. This feature will play an important role in determining the design-based inferential properties of the NN LULC interpolator. In this framework denote by $\Delta = \bigcup_k \delta D_k$ the union of borders where class changes occur and define a point $p \in A$ to be a "boundary" point if $p \in \Delta$ and to be an "interior" point if $p \in A \setminus \Delta$.

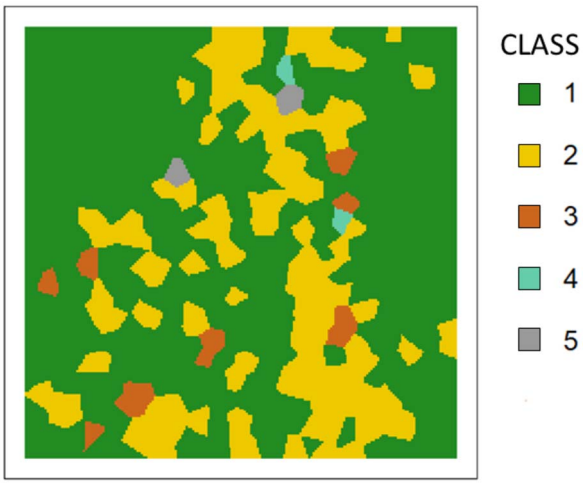

FIG. 1. *Land use map in a square region located in Southern Tuscany (Italy) at the year* 2008 *estimated during the IUTI survey. The five classes are those adopted in the IUTI classification.*

3. Nearest-neighbour interpolation of land use/land cover maps. As stated in the Introduction, it is usually impractical to completely census the entire survey region. In most cases the reference LULC classes are recorded at n sample locations $P_1, \ldots, P_n$. Then a criterion is necessary to estimate the LULC classes 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 $P_1, \ldots, P_n$.

For the interpolation of LULC classes, we modify the NN interpolator that was originally developed by Fattorini et al. (2022) to estimate the values of a quantitative variable at nonsampled locations. Therefore, the LULC class at a nonsampled location $p \in A$ is estimated by

$$\hat{y}(p) = y(P_{NN(p)}), \tag{2}$$

where $P_{NN(p)} = \operatorname{argmin}_{i=1,\ldots,n} \|p - P_i\|$ and $\|\|$ denotes a norm in $\mathbb{R}^2$. Based on (2), we propose indices suitable for quantifying the precision of categorical maps (it should be noticed that we refer to categorical maps because, unlike with choropleth/patch maps, our regions are not predefined). Subsequently, based on these indices, we derive conditions ensuring the design-based consistency of NN maps to support the statistical soundness of (2). 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

$$\operatorname{Err}(p) = E\{z(p)\} = \Pr\{\hat{y}(p) \neq y(p)\} \tag{3}$$

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. Using (3), the precision can be determined for the whole map or for any measurable subregion $A_0 \subset A$ by

$$\operatorname{Err}(A_0) = \frac{1}{|A_0|} \int_{A_0} \operatorname{Err}(p)\, dp. \tag{4}$$

Based on (3) and (4), we define pointwise consistency of the NN interpolator at $p \in A$ if

$$\lim_{n \to \infty} \operatorname{Err}(p) = 0, \tag{5}$$

and we define consistency in mean if

$$\lim_{n\to\infty} \mathrm{Err}(A) = 0. \tag{6}$$

Without delving into the theoretical issues detailed in Appendix A of the Supplementary Material (Marcelli et al. (2022b)), we have proven that consistency holds under suitable sampling schemes at any interior point of the map, that is, almost everywhere. Hence, in these cases consistency in mean holds.

In terms of the sampling schemes needed for consistency, in Appendix A of the Supplementary Material (Marcelli et al. (2022b)) we have proven that consistency holds under those schemes widely applied in environmental surveys, such as: (1) uniform random sampling (URS) in which n locations are randomly and independently selected in the survey region, (2) tessellation stratified sampling (TSS) in which the survey region is partitioned into n tassels of equal size and one location is randomly selected within each patch, and (3) systematic grid sampling (SGS) in which the survey region is partitioned into n regular polygons, one location is randomly selected in one polygon and then repeated in the others. Moreover, in Appendix A of the Supplementary Material (Marcelli et al. (2022b)) we have proven that, under URS, the error probability at any inner point approaches zero at least as c^n with $c \in (0, 1)$, while under TSS and SGS, the error probability is definitively equal to 0 for a sufficiently large sample size.

Regarding these schemes, SGS is widely used in forest surveys (e.g., Opsomer et al. (2007), Tomppo, Gschwantner and McRoberts (2010)), even though its performance may be worse than that of URS when spatial regularities occur. More recently, TSS has become increasingly popular. It has been applied in the 2000–06 Italian National Forest Inventory (Fattorini, Marcheselli and Pisani (2006)) and some years later in the USDA Forest Inventory. TSS does not necessitate partitioning into regular polygons and does not suffer a loss of precision under spatial regularities, as occurs under SGS.

Regarding the estimation of map precision, we adapt the procedure by Fattorini et al. (2022), originally proposed to estimate the root mean squared errors of NN interpolations to the estimation of the error probabilities (3). 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. Therefore, the key problem under PPB is to reconstruct pseudo-populations able to mimic the characteristics of the unknown population in such a way that the bootstrap distribution of any statistic can resemble the true distribution with indices of precision approaching the true ones (e.g., Quatember (2015)).

Accordingly, to estimate (3) we pursue the idea by Fattorini et al. (2022) of using the estimated maps as pseudo-populations from which bootstrap samples are selected using the same spatial scheme adopted to select the original sample. 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 $\mathrm{Err}(p)$.

To this purpose, let $\hat{y}(A) = \{\hat{y}(p),\, p \in A\}$ be the estimated LULC map based on the reference data $y(P_1), \ldots, y(P_n)$. Then, we propose

$$\widehat{\mathrm{Err}}^*_B(p) = \frac{1}{B}\sum_{b=1}^{B} z^*_b(p) \tag{7}$$

as the bootstrap estimator of $\mathrm{Err}(p)$, where B is the number of bootstrap samples, $P^*_{1,b}, \ldots, P^*_{n,b}$ are the locations selected in the bth bootstrap resampling using the same

scheme adopted to select the original sample, $\hat{y}(P^*_{1,b}), \ldots, \hat{y}(P^*_{n,b})$ are the sample observations at these locations derived from the estimated map $\hat{y}(A)$, and $\hat{y}^*_b(p)$ is the bootstrapped value of the nearest neighbour interpolator at $p \in A$, based on $\hat{y}(P^*_{1,b}), \ldots, \hat{y}(P^*_{n,b})$, that is,

$$\hat{y}^*_b(p) = \hat{y}\big(P^*_{NN(p),b}\big), \quad p \in A, b = 1, \ldots, B, \tag{8}$$

where $P^*_{NN(p),b} = \text{argmin}_{i=1,\ldots,n} \|P^*_{i,b} - p\|$ and $z^*_b(p)$ is the dichotomous quantity equal to 0 if $\hat{y}^*_b(p)$ is equal to $\hat{y}(p)$ and equal to 1 otherwise.

Because the NN interpolator is design-consistent at inner points, $\text{Err}(p)$ converges to 0 as n increases in such a way that $E\{\widehat{\text{Err}}^*_B(p)\}$ should also converge to 0. Therefore, in accordance with Fattorini et al. (2022) we evaluate the effectiveness of the bootstrap estimator of precision $\{\widehat{\text{Err}}^*_B(p)\}$ by

$$\text{borat}_B(p) = \frac{E\{\widehat{\text{Err}}^*_B(p)\}}{\text{Err}(p)} = E\left\{\frac{\widehat{\text{Err}}^*_B(p)}{\text{Err}(p)}\right\}. \tag{9}$$

Unfortunately, because of the categorical nature of the LULC variable $y(p)$, the requirements necessary for proving the conservative nature of the bootstrap estimator (7) do not hold (see Fattorini et al. (2022), Proposition 3). Rather, as argued in Appendix B of the Supplementary Material (Marcelli et al. (2022b)), the bootstrap estimator (7) may be quite unstable, especially in the inner parts, those far from the boundaries where class changes occur, where the error probabilities are likely to vanish. Obviously, (9) is undefined when $\text{Err}(p) = 0$. In these cases, which frequently occur for large samples when the NN interpolation becomes consistent, we are forced to consider the bias of (7),

$$\text{bias}_B(p) = E\big\{\widehat{\text{Err}}^*_B(p)\big\} - \text{Err}(p), \tag{10}$$

to obtain insights about the performance of the bootstrap.

4. Area estimation of land use/land cover classes. Given the K LULC classes $c_1, \ldots, c_K$ that are present on A, a common use of LULC analysis has been the estimation of class coverages $\gamma_1, \ldots, \gamma_K$, where

$$|D_k| = \int_A I(p \in D_k)\, dp \tag{11}$$

is the area of the kth class and $\gamma_k = |D_k|/|A|$ is the proportion of the survey region covered by the kth class ($k = 1, \ldots, K$).

4.1. *Traditional estimation*. Estimation of class coverage has been traditionally performed by counting the sample locations within the classes, say $n_1, \ldots, n_K$, and considering the frequencies of these locations in the sample, that is,

$$f_k = \frac{n_k}{n}, \quad k = 1, \ldots, K. \tag{12}$$

Even though straightforward, the estimator (12) has appealing properties. Indeed, Gregoire and Valentine (2008) rephrase (12) as a Monte Carlo integration of (11). Therefore, under URS the estimator (12) coincides with the "basic Monte Carlo integration" and as such it is unbiased, asymptotically ($n \to \infty$) normal, and consistent with a variance of order n^{-1} that can be unbiasedly estimated by

$$\hat{V}^2_k = \frac{f_k(1 - f_k)}{n - 1}. \tag{13}$$

Moreover, under TSS the estimator (12) coincides with the "modified Monte Carlo integration" and as such it is unbiased, asymptotically ($n \to \infty$) normal if the n tassels partitioning

the survey region are not too "stretched," and consistent with a variance of order $n^{-\gamma}$ with $1 < \gamma \leq 2$ that is invariably smaller than the variance under URS and that can be conservatively estimated by (13). Therefore, under TSS the estimator (12) is "super-efficient," having a variance of order greater than n^{-1} (Barabesi, Franceschi and Marcheselli (2012)). Less appealing results are achieved under SGS. That is, the variance of the estimator (12) under SGS is not invariably smaller than the variance under URS, owing to the large loss of precision that occurs under SGS when some spatial regularities are present (Barabesi (2003)). Moreover, under SGS normality does not hold, and there is no way to prove that (13) is a conservative estimator of variance.

4.2. *Estimation from maps*. The NN map of type (2) creates an alternative way to estimate the area of each class. Indeed, denoting by

$$\hat{D}_k = \{p : p \in A, \hat{y}(p) = c_k\} \tag{14}$$

the set where the LULC class is estimated to be c_k with $k = 1, \ldots, K$ (in practice, the $\hat{D}_k$'s are the sample counterparts of the D_k's), we can estimate (11) by

$$|\hat{D}_k| = \int_A I(p \in \hat{D}_k)\, dp \tag{15}$$

in such a way that

$$\hat{\gamma}_k = \frac{|\hat{D}_k|}{|A|}, \quad k = 1, \ldots, K \tag{16}$$

is the map estimator of γ_k. For finite samples the estimator (16) is biased with an intractable variance, but under URS, TSS, and SGS,it asymptotically converges to γ_k owing to the consistency of NN maps.

Regarding the estimation of the design-based variance of (16), it necessitates the bootstrap procedure, described in Section 3, in which we resample from the estimated map using the same scheme implemented to select the original sample. Therefore, the bootstrap estimator of the root mean squared error of (16) is given by

$$\widehat{\mathrm{rmse}}^*_{k,B} = \left\{ \frac{1}{B} \sum_{b=1}^{B} (\hat{\gamma}^*_{k,b} - \hat{\gamma}_k)^2 \right\}^{1/2}, \tag{17}$$

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.

5. Simulation study. To check the performance of the proposed methodology, we conducted a simulation study. The population we considered was a 10 km $\times$ 10 km region located in Southern Tuscany between the provinces of Siena and Grosseto (Central Italy). This region provides a realistic composition (i.e., distribution of classes) and spatial pattern of land cover. 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 IUTI project (see the case study in Section 6.1). The quadrat included five of the six land use classes adopted in the IUTI classification: Forest land (1), Cropland (2), Grassland (3), Wetland (4), and Settlements (5). The quadrat was taken as the true population in the simulation study (see Figure 1).

Sampling was simulated selecting $n = 100; 400; 1600; 10{,}000$ locations within the quadrat by URS, TSS, and SGS, all ensuring consistency of the NN interpolator. URS was performed by randomly and independently selecting n points in the quadrat. TSS was performed by partitioning the quadrat into a grid of $10 \times 10, 20 \times 20, 40 \times 40, 100 \times 100$ tassels of equal size and randomly and independently selecting a point in each subquadrat. SGS was performed by selecting a point in one subquadrat and systematically repeating it in the remaining ones.

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. The grid density was determined in such a way to provide a suitable resolution of the resulting maps. Therefore, at each simulation run the classes at the selected points were recorded from the map of Figure 1, and the NN interpolator (2) was performed at each node of the grid, assigning to each node $p \in G$ the land use class of the nearest location in the sample. Moreover, at each simulation run, $B = 1000$ 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 (2) was performed for each node $p \in G$ to compute the bootstrap estimates (7). Finally, to check the discrepancies from the traditional estimators of area coverage and those achieved from the map, the estimator (12) and the corresponding variance estimator (13), together with the estimator (16) and the corresponding bootstrap estimator of root mean squared error (17), were computed for each land use class.

At the completion of the simulation runs and for each combination of sampling scheme and sample size, we had available $R = 10{,}000$ independent determinations $z_1(p), \ldots, z_R(p)$ for each node $p \in G$, where $z_i(p)$ was equal to 1 if the interpolated class assigned to p at the ith simulation run was wrong and 0 otherwise. From these Monte Carlo distributions, the values of $\mathrm{Err}(p)$ were empirically determined by

$$\mathrm{Err}(p) \sim \frac{1}{R}\sum_{i=1}^{R} z_i(p), \quad p \in G. \tag{18}$$

In addition, for each node $p \in G$ we had available $R = 10{,}000$ independent bootstrap estimates of $\mathrm{Err}(p)$, $\widehat{\mathrm{Err}}^*_{B,1}(p), \ldots, \widehat{\mathrm{Err}}^*_{B,R}(p)$. From these Monte Carlo distributions, the expectation of $\widehat{\mathrm{Err}}^*_B(p)$ was empirically determined by

$$E\{\widehat{\mathrm{Err}}^*_B(p)\} \sim \frac{1}{R}\sum_{i=1}^{R} \widehat{\mathrm{Err}}^*_{B,i}(p), \quad p \in G. \tag{19}$$

From the quantities (18) and (19), the values of $\mathrm{borat}_B(p)$ and $\mathrm{bias}_B(p)$ were achieved by equations (9) and (10), respectively. For each combination of sampling scheme, sample size, and land use class, we had available $R = 10{,}000$ independent traditional estimates of coverage, $f_{k,1}, \ldots, f_{k,R}$, together with the corresponding standard error estimates $\hat{\upsilon}_{k,1}, \ldots, \hat{\upsilon}_{k,R}$, and as many coverage estimates arising from map $\hat{\gamma}_{k,1}, \ldots, \hat{\gamma}_{k,R}$, together with the corresponding bootstrap estimates of root mean squared errors $\widehat{\mathrm{rmse}}^*_{k,B,1}, \ldots, \widehat{\mathrm{rmse}}^*_{k,B,R}$. Then we empirically computed the expectation and the standard error of the traditional estimators

$$\begin{aligned} E(f_k) &\sim \frac{1}{R}\sum_{i=1}^{R} f_{k,i}, \quad k = 1, \ldots, K, \\ \mathrm{SE}(f_k) &\sim \left\{\frac{1}{R}\sum_{i=1}^{R}(f_{k,i} - \gamma_k)^2\right\}^{1/2}, \quad k = 1, \ldots, K \end{aligned} \tag{20}$$

and the expectation of the standard error estimators

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

The unbiasedness of the f_k's is established by sampling theory, but we computed their expectations for the purpose of checking the reliability of the simulation. Finally, we empirically computed the expectation and the root mean squared error of the area estimators, based on maps

$$E(\hat{\gamma}_k) \sim \frac{1}{R}\sum_{i=1}^{R}\hat{\gamma}_{k,i}, \quad k = 1, \ldots, K$$
$$\mathrm{SE}(\hat{\gamma}_k) \sim \left\{\frac{1}{R}\sum_{i=1}^{R}(\hat{\gamma}_{k,i} - \gamma_k)^2\right\}^{1/2}, \quad k = 1, \ldots, K \tag{21}$$

and the expectation of the bootstrap estimator of root mean squared error

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

From the expectations (20) and (21) we computed the ratio $R_k = E(\hat{\gamma}_k)/E(f_k)$ to quantify the discrepancies between the two estimators.

For each combination of sampling scheme and sample size, Table 1 gives the minima, averages, and maxima of the sets $\{\mathrm{Err}(p),\ p \in G\}$ and $\{\mathrm{bias}_B(p),\ p \in G\}$. Note that the averages in the sets of error probabilities constitute the empirical evaluation of the overall precision of the maps on the whole quadrat, theoretically given by equation (4). For each sample size, Figure 2 displays the spatial patterns of the sets $\{\mathrm{Err}(p),\ p \in G\}$, and $\{\mathrm{bias}_B(p),\ p \in G\}$ under TSS. The same patterns under URS and SGS are displayed in Figures C1 and C2 in Appendix C of the Supplementary Material (Marcelli et al. (2022b)). Figure 3 displays the cumulative frequencies of $\{\mathrm{Err}(p),\ p \in G\}$ under URS, TSS, and SGS.

TABLE 1

Values of minima, averages, and maxima of the error probabilities (ERR) and the bias of their bootstrap estimator (BOBIAS) evaluated at each node of the regular grid of 201×201 *locations within the quadrat of Figure 1*

		ERR			BOBIAS		
SCHEME	n	MIN	AVERAGE	MAX	MIN	AVERAGE	MAX
URS	100	0.00	0.28	0.88	−0.74	−0.09	0.10
	400	0.00	0.18	0.84	−0.69	−0.02	0.13
	1600	0.00	0.10	0.84	−0.71	0.01	0.15
	10,000	0.00	0.04	0.73	−0.54	0.01	0.13
TSS	100	0.00	0.22	0.75	−0.68	−0.06	0.16
	400	0.00	0.16	0.85	−0.76	−0.02	0.18
	1600	0.00	0.08	0.87	−0.79	0.00	0.16
	10,000	0.00	0.04	0.74	−0.55	0.01	0.14
SGS	100	0.00	0.26	0.89	−0.78	−0.10	0.20
	400	0.00	0.15	0.91	−0.81	−0.01	0.30
	1600	0.00	0.07	0.90	−0.79	0.01	0.28
	10,000	0.00	0.03	0.81	−0.70	0.00	0.15

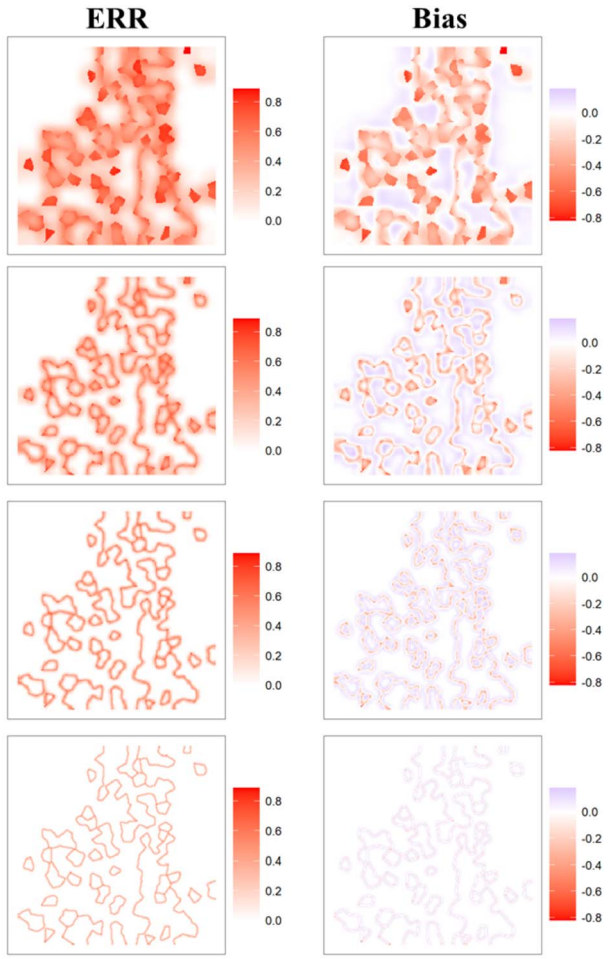

FIG. 2. *Spatial patterns of the error probabilities* (*left column*) *and the bias of their bootstrap estimator* (*right column*) *evaluated at each node of the regular grid of* 201×201 *locations within the quadrat of Figure* 1 *under TSS and sample sizes* $n = 100, 400, 1600, 10{,}000$ (*rows*).

For each combination of sampling scheme, sample size, and land use class, Table 2 gives the expectations and the standard errors of the traditional estimators and the expectations of the standard error estimators, the expectations and the root mean squared errors of the estimators based on maps, and the expectations of the bootstrap estimators of root mean squared errors and the ratio of expectations R_k.

The results of the simulation study confirm the theoretical results. The NN interpolator (2) proves to be consistent under the three sampling schemes with error probabilities that quickly decrease on averages and maxima as the sample size increases. Comparatively, TSS proves to be superior to URS and SGS in terms of maxima. Consistency is clearly apparent from Figure 2 and Figures C1–C2 in the Supplementary Material (Marcelli et al. (2022b)), where, as the sample size increases, the maps on the left side become whiter and whiter with large zones in which interpolation is performed without error, and errors are restricted along the borders where changes of class occur. Especially encouraging are the plots of cumulative frequencies of errors (Figure 3), showing 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 (Table 1) may appear quite discouraging with minimum values always smaller than -0.5, even for the very large sample size of 10,000

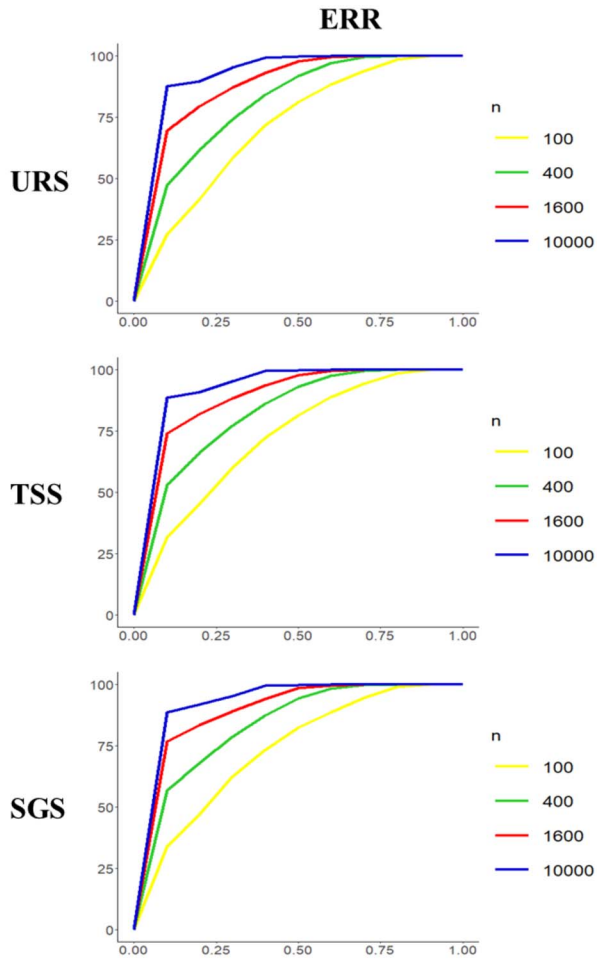

FIG. 3. *Cumulative frequencies of the error probabilities evaluated at each node of the regular grid of* 201×201 *locations within the quadrat of Figure* 1 *under URS, TSS, and SGS* (*rows*) *and sample sizes* $n = 100, 400, 1600, 10{,}000$ (*colors*).

points. However, it is apparent from the maps on the right side of Figure 2 and Figures C1–C2 that, as sample size increases, the maps become whiter and whiter with large zones in which the estimator is unbiased, while 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. The overestimation of the variance and the subsequent masking of the precision, a well-known problem of the traditional estimation under TSS and SGS (Barabesi, Franceschi and Marcheselli (2012)), is less marked for map estimator.

6. Case studies.

6.1. *The IUTI land use survey.* In 2008, the Italian Ministry of Environment and Protection of Land and Sea promoted and carried out a land use pure panel survey, referred to as

TABLE 2

Percent values of expectation (E_{TR}), standard error (SE_{TR}), and expectation of standard error estimators (ESEE) for the traditional estimator and percent values of expectation (E_{MAP}), root mean squared error ($RMSE_{MAP}$), and expectation of bootstrap estimator of root mean squared error (EBRMSEE) for the estimator based on maps together with the ratio of expectations ($RAT = E_{MAP}/E_{TR}$)

CLASS	SCHEME	n	E_{TR}	SE_{TR}	ESEE	E_{MAP}	$RMSE_{MAP}$	EBRMSEE	RAT
Forest	URS	100	70.40	4.51	4.56	70.21	4.34	3.74	1.00
land		400	70.39	2.29	2.28	70.42	1.85	1.70	1.00
% coverage		1600	70.41	1.16	1.14	70.51	0.67	0.68	1.00
70.51		10,000	70.39	0.47	0.46	70.51	0.17	0.18	1.00
	TSS	100	70.34	3.70	4.57	70.18	3.74	2.73	1.00
		400	70.37	1.49	2.28	70.48	1.39	1.24	1.00
		1600	70.38	0.56	1.14	70.50	0.47	0.49	1.00
		10,000	70.39	0.14	0.46	70.51	0.13	0.15	1.00
	SGS	100	70.38	3.18	4.58	70.03	2.86	0.82	1.00
		400	70.40	1.11	2.28	70.46	1.12	0.23	1.00
		1600	70.39	0.41	1.14	70.51	0.38	0.08	1.00
		10,000	70.39	0.17	0.46	70.51	0.11	0.04	1.00
Cropland	URS	100	25.88	4.33	4.37	26.02	4.18	3.62	1.01
		400	25.89	2.20	2.19	25.86	1.76	1.62	1.00
% coverage		1600	25.85	1.11	1.09	25.79	0.64	0.65	1.00
25.79		10,000	25.87	0.45	0.44	25.79	0.16	0.17	1.00
	TSS	100	25.94	3.56	4.39	26.05	3.59	2.64	1.00
		400	25.90	1.39	2.19	25.81	1.29	1.18	1.00
		1600	25.87	0.52	1.10	25.79	0.44	0.46	1.00
		10,000	25.88	0.17	0.44	25.79	0.13	0.14	1.00
	SGS	100	25.90	3.15	4.39	26.21	2.98	0.74	1.01
		400	25.86	1.12	2.19	25.79	1.12	0.20	1.00
		1600	25.88	0.34	1.10	25.79	0.31	0.08	1.00
		10,000	25.88	0.14	0.44	25.79	0.10	0.04	1.00
Grassland	URS	100	2.64	1.61	1.50	2.67	1.71	1.34	1.01
		400	2.65	0.81	0.79	2.65	0.73	0.65	1.00
% coverage		1600	2.65	0.40	0.40	2.63	0.26	0.26	0.99
2.63		10,000	2.65	0.16	0.16	2.63	0.07	0.07	0.99
	TSS	100	2.64	1.36	1.53	2.68	1.41	0.99	1.01
		400	2.66	0.60	0.80	2.65	0.56	0.48	1.00
		1600	2.66	0.21	0.40	2.63	0.18	0.19	0.99
		10,000	2.65	0.06	0.16	2.63	0.05	0.06	0.99
	SGS	100	2.65	1.89	1.42	2.67	1.91	0.18	1.01
		400	2.66	0.37	0.80	2.68	0.39	0.07	1.01
		1600	2.65	0.15	0.40	2.63	0.16	0.01	0.99
		10,000	2.66	0.04	0.16	2.63	0.03	0.00	0.99
Wetland	URS	100	0.42	0.65	0.38	0.44	0.72	0.34	1.04
		400	0.42	0.33	0.28	0.42	0.31	0.25	0.99
% coverage		1600	0.42	0.16	0.16	0.42	0.12	0.12	0.99
0.42		10,000	0.42	0.07	0.06	0.42	0.03	0.03	0.99
	TSS	100	0.42	0.61	0.39	0.43	0.63	0.26	1.02
		400	0.42	0.27	0.29	0.42	0.26	0.19	0.99
		1600	0.42	0.10	0.16	0.42	0.08	0.08	0.99
		10,000	0.42	0.03	0.06	0.42	0.02	0.03	0.99
	SGS	100	0.42	0.52	0.41	0.45	0.57	0.06	1.09
		400	0.42	0.12	0.32	0.42	0.12	0.00	0.99
		1600	0.42	0.05	0.16	0.42	0.05	0.00	0.99
		10,000	0.42	0.01	0.06	0.42	0.01	0.00	0.99

TABLE 2
(*Continued*)

CLASS	SCHEME	n	E_{TR}	SE_{TR}	ESEE	E_{MAP}	$RMSE_{MAP}$	EBRMSEE	RAT
Settlements	URS	100	0.66	0.80	0.55	0.67	0.85	0.50	1.01
		400	0.65	0.40	0.37	0.65	0.35	0.31	1.00
% coverage		1600	0.66	0.20	0.20	0.65	0.12	0.13	0.99
0.65		10,000	0.66	0.08	0.08	0.65	0.03	0.03	0.99
	TSS	100	0.66	0.70	0.59	0.66	0.72	0.38	1.00
		400	0.66	0.30	0.39	0.65	0.27	0.24	0.99
		1600	0.66	0.11	0.20	0.65	0.09	0.09	0.99
		10,000	0.66	0.03	0.08	0.65	0.03	0.03	0.99
	SGS	100	0.65	0.60	0.62	0.65	0.59	0.01	0.99
		400	0.66	0.19	0.40	0.65	0.18	0.00	0.99
		1600	0.66	0.06	0.20	0.65	0.06	0.00	0.99
		10,000	0.66	0.01	0.08	0.65	0.01	0.00	0.99

IUTI from the Italian acronym of "Inventario dell'Uso delle Terre d'Italia," with the main purpose of implementing the national greenhouse gas assessment under the Kyoto Protocol framework. A spatially balanced sample of points was selected throughout the Italian territory by TSS. More precisely, the Italian territory was covered by a network of $n = 1{,}217{,}032$ quadrats of size 25 ha, and a point was randomly selected within each quadrat. The sample points were accurately photo-interpreted on the very high-resolution imagery available for the years 1990 and 2008 and then assigned to land use classes 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 (International Panel on Climate Change (2003)). The coarsest classification adopted six land use classes: Forest land (1), Cropland (2), Grassland (3), Wetland (4), Settlements (5), and Other lands (6). The classification was performed and officially released for the years 1990 and 2008 (ISPRA (2014)), and it is currently available on the Geoportale Nazionale at the web site http://www.pcn.minambiente.it/GN/accesso-ai-servizi/servizi-di-visualizzazione-wms.

The primary aim of the survey was the estimation of land use proportions at both years and their changes over time together with the estimation of the corresponding standard errors (Corona et al. (2012)), but no attempt was made to make inference on the land use maps. Therefore, data from the year 2008 were here adopted for the first time to provide a map with accompanying inference. To this purpose, based on the 1,217,032 sample locations, the NN interpolator (2) was adopted to estimate $y(p)$ for each location p in the network of 2,800,360 nodes within the Italian territory with a map resolution of 10.86 ha (see Figure 4a). From the estimated map, $B = 1000$ 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 the bootstrap estimate $\widehat{Err}^*_B(p)$ was derived by equation (7) for each location in the network of nodes where estimation was performed (see Figures 4b and 4c). For a clearer depiction of the results, we focused on a portion of the land use map and of the map of the estimated precision and their cumulative frequencies for the region of the Gargano Promontory in Southern Italy (see Figure D1 in Appendix D of the Supplementary Material, Marcelli et al. (2022b)).

From Figures 4b and 4c, the precision of the strategy is apparent, due in most part to the intensive sampling effort of one point every 25 ha. 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.

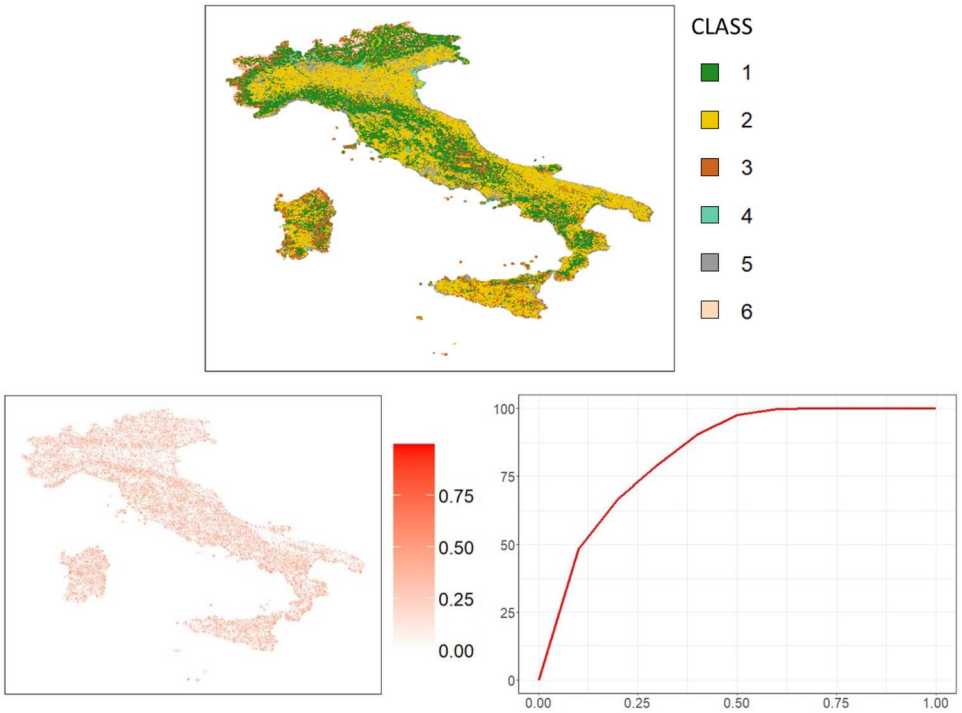

FIG. 4. (*a*) *Italian map of the six land use classes estimated from the IUTI TSS sample of* 1,217,032 *points classified at the year* 2008; (*b*) *Map of the estimates of the error probabilities achieved by* $B = 1000$ *bootstrap samples*; (*c*) *Cumulative frequencies of the estimates of the error probabilities*.

Precision is even better in the Gargano Promontory (Figures D1b and D1c in Appendix D of the Supplementary Material, Marcelli et al. (2022b)) where estimated errors smaller than 10% occur on about 70% of the promontory. Finally, Table 3 shows the high precision of the six area estimates achieved by the traditional estimator (12) with estimated standard errors invariably smaller than 0.05%. Table 3 also shows the similarity between the traditional estimates with those achieved from the map, with discrepancies always smaller than 0.25%. The convergence of the two estimators for intensive sampling effort is consistent with the simulation results of Section 5 (Table 2). Also consistent with the simulation results, the map estimator had better precision than the traditional estimator (BRMSE $<$ SE).

TABLE 3

Percent values of the traditional estimates (*TR*) *of the areas of the six land use classes achieved from the IUTI TSS sample of* 1,217,032 *points classified at the year* 2008 *with their standard error estimates* (*SEE*), *together with the percent values of the estimates achieved from the estimated map* (*MAP*) *and their bootstrap estimates of root mean squared error* (*BRMSE*) *based on* $B = 1000$ *bootstrap samples*

Land-use class	TR	SEE	MAP	BRMSE
Forest Land	32.03	0.042	32.10	0.021
Cropland	44.17	0.045	44.32	0.024
Grassland	12.82	0.030	12.86	0.018
Wetland	1.72	0.012	1.49	0.006
Settlements	7.10	0.023	7.11	0.014
Other lands	2.17	0.013	2.14	0.006

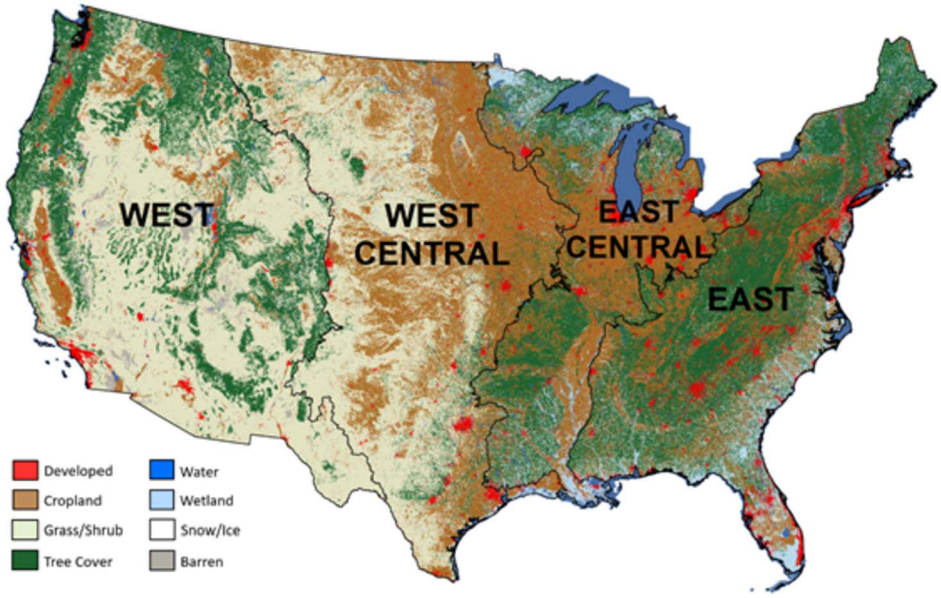

FIG. 5. *U.S. map of the eight land cover classes at the year* 2017, *estimated from the LCMAP classification algorithm of Landsat data.*

6.2. *The LCMAP survey.* The LCMAP program of the U.S. Geological Survey, from the acronym of Land Change Monitoring, Assessment, and Projection, has a main objective of mapping annual land cover in the U.S. with the time period starting in 1985 (Brown et al. (2020)). The mapping is currently done by classifying satellite imagery (Landsat) using an LCMAP classification algorithm (see Figure 5). The population under study is the set of Landsat pixels (30 m $\times$ 30 m) in the conterminous US (excluding Alaska and Hawaii), an area of about eight million km^2. A sample of $n = 24{,}971$ pixels was selected from the population by simple random sampling without replacement (SRSWOR). For each sample pixel, interpreters determined land cover annually from 1985 to 2017 following protocols described in Pengra et al. (2020). These sample data represented the reference (true) conditions and were used to assess map accuracy (Stehman et al. (2021)) as well as to estimate the area of different land cover classes and area of change in land cover classes (Auch et al. (2022)). There are eight land cover classes in LCMAP: Developed (1), Cropland (2), Grass/shrub (3), Tree cover (4), Water (5), Wetland (6), Snow/ice (7), and Barren (8).

The 30 m $\times$ 30 m pixels were so small with respect to the area of the conterminous U.S. that the pixels were considered as points on a continuum, while SRSWOR of pixels was assimilated to URS of points. Therefore, with this approximation in mind, the sample data for the year 2017 were adopted for the first time to provide a U.S. map of land cover together with the corresponding map of the estimated precision. To this purpose and based on the 24,971 sample pixels, the NN interpolator (2) was implemented to estimate $y(p)$ for each location p in a network of 738,512 nodes (one each 3.28 km) within the conterminous U.S. (see Figure 6a) with a map resolution of 10.76 km^2. This resolution was by far less accurate with respect to that of Figure 5 in which a point every 900 m^2 was mapped. However, resolution was only a practical issue, and a much denser map could be produced just by increasing the computational time. From the estimated map, $B = 1000$ bootstrap samples of size 24,971 were selected from the U.S. shapefile by URS, giving rise to as many bootstrapped maps from which the bootstrap estimate $\widehat{\mathrm{Err}}^*_B(p)$ was derived by equation (7) for each location in the network of nodes where estimation was performed (see Figures 6b and 6c).

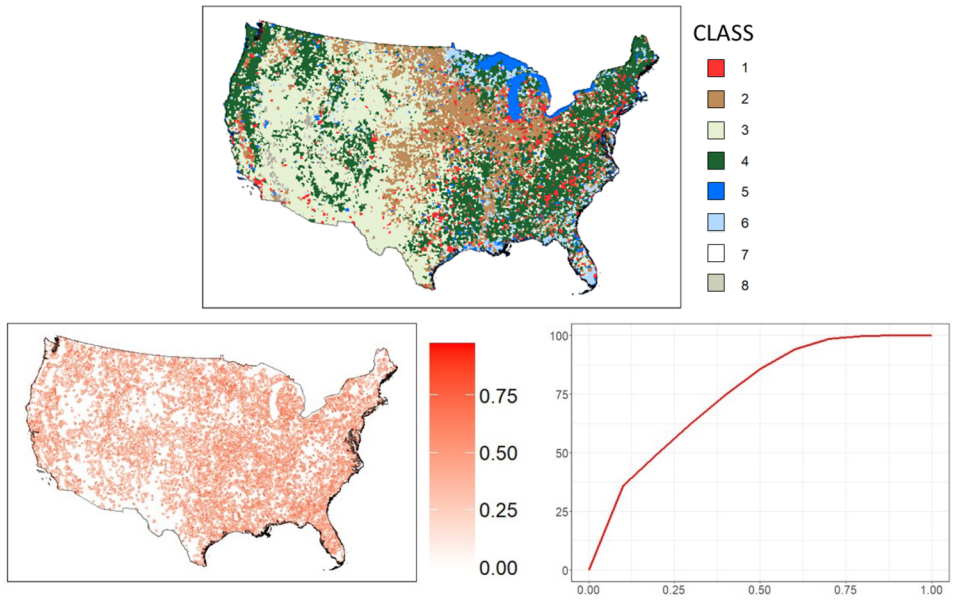

FIG. 6. (*a*) *U.S. map of the eight land cover classes estimated from the LCMAP SRSWOR sample of* 24,971 *points classified at the year* 2017 (*see Figure* 5 *for correspondence between colours and land cover classes*); (*b*) *Map of the estimates of the error probabilities achieved by* $B = 1000$ *bootstrap samples*; (*c*) *Cumulative frequencies of the estimates of the error probabilities*.

As it is apparent from Figure 6a, the U.S. eastern regions generally have greater heterogeneity of land cover, which gives rise to larger error values relative to the western regions. To provide a comparison, we focused on portions of the land cover map in the states of Florida (Figures D2a and D2b in Appendix D of the Supplementary Material, Marcelli et al. (2022b)) and Texas (Figures D3a and D3b in Appendix D of the Supplementary Material, Marcelli et al. (2022b)). The design-based, objective evaluation of precision that accompanies the NN interpolated map is not available from the LCMAP classification of land cover.

It is apparent that the precision of the LCMAP strategy (Figures 6b and 6c) is lower than that achieved in the IUTI survey, and this low precision is due primarily to the less intensive sampling effort and to the use of SRSWOR that does not ensure an even coverage of the U.S. territory. Estimated errors smaller than 10% occur on about 30% of the U.S. territory, compared to 50% in the IUTI case. Estimated errors greater than 25% occur for about half of the U.S. territory vs. 25% in the IUTI case, and estimated errors smaller than 50% occur on about 80% of the U.S. territory relative to about 100% in the IUTI surey region. Precision is worse in Florida, owing to the great variability of land cover classes throughout the state (Figure D2 in Appendix D of the Supplementary Material, Marcelli et al. (2022b)), where estimated errors smaller than 10% occur only on 10% of the state. Conversely, precision is better in Texas (Figure D3 in Appendix D of the Supplementary Material, Marcelli et al. (2022b)) because of the low variability of land cover classes throughout the state, where estimated errors smaller than 10% occur on more than 50% of the state.

Finally, Table 4 shows the good precision of the eight area estimates achieved by the traditional estimator (12) with most of the estimated standard errors smaller than 0.3%, notwithstanding the moderate sampling effort and the straightforward use of SRSWOR. Interestingly, Table 4 also shows the similarity between the traditional estimates with those achieved from the map, with most of the discrepancies smaller than 1%, indicating that the two estimators yield similar results also with moderate sample sizes.

TABLE 4

Percent values of the traditional estimates (TR) of the area of the eight land cover classes achieved from the LCMAP SRSWOR sample of 24,971 *points classified at the year* 2017 *with their standard error estimates (SE), together with the percent values of the estimates achieved from the estimated map (MAP) and their bootstrap estimates of root mean squared error (BRMSE) based on* $B = 1000$ *bootstrap samples*

Land-cover class	TR	SE	MAP	BRMSE
Developed	5.55	0.145	5.52	0.115
Cropland	17.44	0.240	17.50	0.173
Grass/shrub	38.12	0.307	38.65	0.209
Tree cover	27.83	0.284	28.32	0.205
Water	5.24	0.141	4.17	0.114
Wetland	4.95	0.137	4.95	0.112
Snow/ice	0.01	0.007	0.02	0.007
Barren	0.86	0.058	0.87	0.048

7. Final remarks. Design-based estimates of LULC maps are performed exploiting the reference data selected by 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 reference sample rather than from model assumptions. As emphasized by Särndal, Swensson and Wretman (1992), "The probability distribution associated with the design is real, not modelled or assumed."

Regarding computational requirements, the proposed methodology was parallelized. Therefore, the computational effort depends on the system requirements of the machine on which the code runs. Both the simulation study and the case studies were conducted using 10 cores out of the 20 available in our computer. The simulation study was conducted for a relatively small area with an increasing number of selected locations and different sampling schemes. Therefore, the computational times depended on the combination of sampling scheme (with SGS being the slowest) and sample size. In particular, using 10,000 sample locations and SGS, the simulation time was more than 24 hours for a total of 10 days needed to perform the complete simulation scenario. As for the case studies, mapping from the LCMAP survey was faster than mapping from the IUTI survey because the former had a smaller number of nodes and sample locations. The U.S. maps were obtained in 16 hours, while the Italian maps were obtained in five days. Obviously, the number of bootstrap samples heavily impacts the computational effort. If only 100 bootstrap samples were used for the Italian maps, results would be obtained in about 12 hours, while about 50 days would be necessary with 10,000 bootstrap samples.

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. Opsomer et al. (2007) point out that "tremendous" improvement can be achieved by the exploitation of such auxiliary variables. Indeed, incorporating other auxiliary variables is straightforward in design-based mapping of quantitative variables in which the survey variable is predicted by a suitable function of the auxiliary variables and the design-based interpolation is performed on the residuals (Di Biase et al. (2022)). However, this approach is less straightforward in LULC mapping because classes are

categorical variables. An alternative way to exploit auxiliary information is to perform NN interpolation in the space of auxiliary variables; that is, the interpolated class at a location is the class observed at the sample location that is nearest in the auxiliary space rather than in the coordinate space. This intriguing idea has been empirically investigated by Grafström, Saarela and Ene (2014), achieving promising results that, however, necessitate further theoretical investigations to be fully confirmed, especially concerning the property of consistency.

A second concern, pointed out in Section 4 and also considered in the simulation and case studies, 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. Once again, a solution is possible when we map quantitative variables because the predictions at single points can be rescaled in such a way that their totals match the estimates of totals achieved by traditional estimators such as Horvitz-Thompson, ratio, or regression estimators (Marcelli et al. (2022)). Obviously, a similar harmonization of the two estimates is not possible with categorical maps because classes cannot be rescaled. However, 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.

Current "good practice" guidance for LULC change studies (Olofsson et al. (2014)) recommends use of a probability sample of reference class data for estimating area with accompanying standard errors produced in the context of design-based inference. The new capacity introduced in this article for design-based mapping of LULC classes creates the opportunity to produce maps and area estimates within a single inference framework. Current practice mixes the type of inference invoked as uncertainty of the map classification, if addressed at all, is derived within a model-based inference framework, whereas area estimation invokes design-based inference. Although additional 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 are grateful to Luca Pratelli from the Naval Accademy of Livorno (Italy) for his help in the theoretical issues of the paper and to the former Editor-in-Chief Karen Kafadar, the anonymous Associate Editor, and the two anonymous referees for their comments that have been essential for improving the manuscript.

Funding. This paper was carried out in the frame of the subproject "Precision Forestry" (AgriDigit program) (DM 36503.7305.2018 of December 2, 2018) funded by the Italian Ministry of Agricultural, Food, and Forest Policies (MiPAAF, Italy) and managed by the CREA–Research Centre for Forestry and Wood, Italy. The authors acknowledge research support from the Research and Innovation Centre, Fondazione Edmund Mach (San Michele all'Adige, Trento, Italy).

SUPPLEMENTARY MATERIAL

Supplement (DOI: 10.1214/23-AOAS1754SUPP; .pdf). The supplement consists of four sections, containing, respectively, some results about consistency, features of bootstrap estimators of precision, figures from the simulation study and figures from case studies.

REFERENCES

AL-DOSKI, J., MANSOR, S. B., SAN, H. P. and KHUZAIMAH, Z. (2020). Land cover mapping using remote sensing data. *Am. J. Geogr. Inf. Syst.* **9** 33–45.

AUCH, R. F., WELLINGTON, D. F., TAYLOR, J. L., STEHMAN, S. V., TOLLERUD, H. J., BROWN, J. F., LOVELAND, T. R., PENGRA, B. W., HORTON, J. A. et al. (2022). Conterminous United States land-cover change (1985–2016): New insights from annual time series. *Land* **11** 298.

BARABESI, L. (2003). A Monte Carlo integration approach to Horvitz–Thompson estimation in replicated environmental designs. *Metron* **LXI** 355–374.

BARABESI, L., FRANCESCHI, S. and MARCHESELLI, M. (2012). Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation. *Ann. Appl. Stat.* **6** 210–228.

BROWN, J. F., TOLLERUD, H. J., BARBER, C. P., ZHOU, Q., DWYER, J. L., VOGELMAN, J. E., LOVELAND, T. R., WOODCOK, C. E., STEHMAN, S. V. et al. (2020). Lessons learned implementing an operational continuous United States national land change monitoring capability: The land change monitoring, assessment, and projection (LCMAP) approach. *Remote Sens. Environ.* **238** 111356.

CIHLAR, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities. *Int. J. Remote Sens.* **21** 1093–1114.

COMBER, A., SEE, L., FRITZ, S., VAN DER VELDE, M., PERGER, C. and FOODY, G. M. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *Int. J. Appl. Earth Obs. Geoinf.* **23** 37–48.

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.

DI BIASE, R. M., FATTORINI, L., FRANCESCHI, S., GROTTI, M., PULETTI, N. and CORONA, P. (2022). From model selection to maps: A completely design-based data-driven inference for mapping forest resources. *Environmetrics* **33** e2750.

FATTORINI, L. (2015). Design-based methodological advances to support national forest inventories: A review of recent proposals. *iForest* **8** 6–11.

FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2004). Two-phase estimation of coverages with second-phase corrections. *Environmetrics* **15** 357–368.

FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2006). A three-phase sampling strategy for large-scale multiresource forest inventories. *J. Agric. Biol. Environ. Stat.* **11** 296–316.

FATTORINI, L., MARCHESELLI, M., PISANI, C. and PRATELLI, L. (2022). Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. *Biometrics* **78** 1454–1463.

FITZPATRICK-LINS, K. (1981). Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogramm. Eng. Remote Sens.* **47** 343–351.

GRAFSTRÖM, A., SAARELA, S. and ENE, L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Can. J. For. Res.* **44** 1156–1164.

GREGOIRE, T. G. and VALENTINE, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. CRC Press/CRC, Boca Raton, FL.

HANSEN, M., DUBAYAH, R. and DE FRIES, R. (1996). Classification trees: An alternative to traditional land cover classifier. *Int. J. Remote Sens.* **17** 1075–1081.

INTERNATIONAL PANEL ON CLIMATE CHANGE (2003). *Good Practice Guidance for Land Use, Land Use Change and Forestry. IPCC National Greenhouse Gas Inventories Program.*

ISPRA (2014). Italian Greenhouse Gas Inventory 1990–2012. National Inventory Report 2014 ISPRA Rapporti 198/14.

KHATAMI, R., MOUNTRAKIS, G. and STEHMAN, S. V. (2017). Mapping per-pixel predicted accuracy of classified remote sensing images. *Remote Sens. Environ.* **191** 156–167.

MANNAN, B., ROY, J. and RAY, A. K. (1998). Fuzzy ARTMAP supervised classification of multi-spectral remotely-sensed images. *Int. J. Remote Sens.* **19** 767–774.

MARCELLI, A., FATTORINI, L. and FRANCESCHI, S. (2022). Harmonization of design-based mapping for spatial populations. *Stoch. Environ. Res. Risk Assess.*. **36** 3171–3182.

MARCELLI, A., DI BIASE, R. M., CORONA, P., STEHMAN, S. V. and FATTORINI, L. (2023). Supplement to "Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation." https://doi.org/10.1214/23-AOAS1754SUPP

MCROBERTS, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sens. Environ.* **115** 715–724.

NATIONAL RESEARCH COUNCIL (2001). *Grand Challenges in Environmental Sciences*. The National Academy Press, Washington, DC.

NGUYEN, H. T. T., DOAN, T. M., TOMPPO, E. and MCROBERTS, R. E. (2020). Land use/land cover mapping using multitemporal Sentinel-2 imagery and four classification methods—a case study from Dak Nong, Vietnam. *Remote Sens.* **12** 1367.

NUSSER, S. M. and KLAAS, E. E. (2003). Survey methods for assessing land cover map accuracy. *Environ. Ecol. Stat.* **10** 309–331.

OLOFSSON, P., FOODY, G. M., HEROLD, M., STEHMAN, S. V., WOODCOCK, C. E. and WULDER, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **148** 42–57.

OPSOMER, J. D., BREIDT, F. J., MOISEN, G. G. and KAUERMANN, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *J. Amer. Statist. Assoc.* **102** 400–409.

PENGRA, B. W., STEHMAN, S. V., HORTON, J. A., DOCKTER, D. J., SCHROEDER, T. A., YANG, Z., COHEN, W. B., HEALEY, S. P. and LOVELAND, T. R. (2020). Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. *Remote Sens. Environ.* **238** 111261.

QUATEMBER, A. (2015). *Pseudo-Populations. A Basic Concept in Statistical Surveys*. Springer, Cham.

RIZZO, M. and GASPARINI, P. (2022). Land use and land cover photointerpretation. In *Italian National Forest Inventory-Methods and Results of the Third Survey* (P. Gasparini, L. Di Cosmo, A. Floris and D. De Laurentis, eds.) 49–59. Springer, Cham, CH.

RODRÍGUEZ-JEANGROS, N., HERING, A. S., KAISER, T. and MCCRAY, J. (2016). Fusing multiple existing space-time land cover products. *Environmetrics* **28** e2429.

RODRÍGUEZ-JEANGROS, N., HERING, A. S., KAISER, T. and MCCRAY, J. (2017). SCaMF-RM: A fused high-resolution land cover product of the Rocky Mountains. *Remote Sens.* **9** 1015.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling. Springer Series in Statistics*. Springer, New York.

STEHMAN, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **62** 77–89.

STEHMAN, S. V. (2009). Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **30** 5243–5272.

STEHMAN, S. V. and CZAPLEWSKI, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sens. Environ.* **64** 331–344.

STEHMAN, S. V. and FOODY, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sens. Environ.* **231** 111199.

STEHMAN, S. V., PENGRA, B. P., HORTON, J. A. and WELLINGTON, D. F. (2021). Validation of the United States geological survey's land change monitoring, assessment and projection (LCMAP) annual land cover products 1985–2017. *Remote Sens. Environ.* **265** 112646.

TOMPPO, L. M., GSCHWANTNER, T. and MCROBERTS, R. E. (2010). *National Forest Inventories: Pathways for Common Reporting*. Springer, Heidelberg, DE.

TURNER, B. L., LAMBIM, E. F. and REENBERG, A. (2007). The emergence of land change science for global environmental change and sustainability. *Proc. Natl. Acad. Sci. USA* **104** 20666–20671.

VAN DER MEER, F. (1995). Spectral unmixing of landsat thematic mapper data. *Int. J. Remote Sens.* **16** 3189–3194.

YOOL, S. R. (1998). Land cover classification in rugged areas using simulated moderate-resolution remote sensor data and an artificial neural network. *Int. J. Remote Sens.* **19** 85–96.