Stochastic Environmental Research and Risk Assessment (2025) 39:1077–1090
https://doi.org/10.1007/s00477-025-02908-2

ORIGINAL PAPER

Design-based mapping of errors in remote sensing-based land use/land cover maps

R. M. Di Biase[1,5] · A. Marcelli[2] · P. Corona[3] · S. V. Stehman[4] · L. Fattorini[1]

Accepted: 4 January 2025 / Published online: 19 January 2025

Abstract
For the first time, spatially explicit representation of classification errors of land use/land cover (LULC) maps is approached from a design-based perspective. Since LULC maps are typically derived from non-probabilistic training samples, these maps, like the true LULC map, are fixed in a design-based scenario so that the error maps achieved by comparing the satellite-based and true maps are fixed. Based on a probabilistic sample of locations where the true or "reference" class is obtained (i.e., the "reference" class is considered the best representation of the true class), errors can be assessed at these sample locations by comparing the map classes to the reference classes. Then, the presence or absence of errors is interpolated across the entire survey area using the nearest neighbour technique. Under very common sampling schemes used to collect reference sample data, the interpolated error maps are design consistent. A simulation study confirms the design consistency of the interpolated error maps, which converge to the true error map as the reference sample size increases. The U.S. land cover map from the LCMAP program and the Italian forest/non forest map serve as case studies.

Keywords Design consistency · Error maps · Nearest-neighbour interpolation · Quality assessment · Reference sample · Satellite maps

1 Introduction

There is substantial interest in producing land use/land cover (LULC) maps at regional and national scales. Progress in remote sensing technologies and improvement of computational capabilities have created the ability to process large data sets allowing production of LULC maps via the classification of satellite images. In this process, satellite data, acquired for the entire survey area, are transformed into LULC classes using a wide variety of image classification algorithms. Comprehensive reviews of these algorithms are provided by Gomez et al. (2016) and Khatami et al. (2016). In practice, classification algorithms are typically applied to a "training sample", i.e., a set of locations for which the LULC classes have been accurately identified and can be considered ground truth. The training sample classes, together with remote sensing-based covariates, are then used in predictive models to predict LULC classes for all locations in the survey area, resulting in a map, hereafter referred to as the "satellite map". Relevant to subsequent developments, the training sample is typically not a probability sample but is instead selected for convenience or judgment (e.g., Nguyen et al. 2020).

As pointed out by Khatami et al. (2017), satellite maps contain misclassifications that require additional quality information to describe the magnitude and spatial distribution of classification errors. Conventional accuracy assessments address only the magnitude of errors by comparing the map labels to the reference class labels, where a "reference class" represents the best assessment of ground truth. Since reference classes are expensive to obtain, they are usually recorded for a probabilistic sample of locations,

✉ R. M. Di Biase
rosa.dibiase@unisi.it

1 Department of Economics and Statistics, University of Siena, Siena, Italy

2 Department for Innovation in Biological, Agro-Food and Forest Systems, University of Tuscia, Viterbo, Italy

3 CREA, Research Centre for Forestry and Wood, Arezzo, Italy

4 Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, USA

5 NBFC, National Biodiversity Future Center, Palermo, Italy

hereafter referred to as "reference locations". The most used summary of these assessments is the confusion matrix, a cross-tabulation of the predicted classes (i.e., from the satellite-based map) against the reference classes at the reference locations. Several indices of map quality are then derived from the confusion matrix (e.g., Stehman 1997, 2009).

Accuracy assessments based on confusion matrices do not assess the spatial distribution of errors. Indeed, like most spatial phenomena, misclassifications are likely to be spatially clustered (spatial autocorrelation) and their presence may vary across different parts of the survey area (spatial heterogeneity). Therefore, ideally, any satellite map should be accompanied by a map showing the spatial distribution of the quality of satellite-based predictions (e.g., Comber et al. 2012).

Two main approaches have been proposed to spatialize the quality of satellite maps: certainty maps (C-maps) and accuracy maps (A-maps). C-maps are based on the concept that, for a given location, the higher the probability of class membership for a predicted class, the greater the certainty associated with that prediction (Khatami et al. 2017). The probability of belonging to the most probable class is typically used as an indicator of certainty at any predicted location. Recently, Valle et al. (2023) proposed quantifying certainty using conformal statistics. Since these probabilities arise directly from the classification algorithms applied to the training sample (e.g., posterior probabilities from maximum likelihood classifiers or neural network classifiers), an attractive aspect of C-maps is that they can be constructed without the need for expensive reference class data. Khatami et al. (2017) provide a comprehensive list of methodologies for building C-maps.

Since C-maps provide information about the spatial distribution of the degree of confidence in the predicted classes, they are primarily of interest to analysts for improving their classification algorithms. Conversely, according to the universally accepted definition of accuracy in satellite mapping, i.e., the probability of correct classification for a location (e.g., Khatami et al. 2017), an A-map should provide the probabilities of correct classification for all locations within the entire survey area. In this sense, map users may be more interested in A-maps than in C-maps. Because accuracy is unknown at any location not included in the reference sample, accuracy must be estimated at the unsampled locations. These estimates are based on the dichotomous quantities indexing the agreement (i.e., correctly classified or not) between the predicted class and the true reference class at any reference location, henceforth referred to as "e-quantities".

From a methodological perspective, accuracy estimation has been approached in different ways. Foody (2005) proposed creating a grid of spatially constrained confusion matrices, deriving an overall accuracy index for each grid point, and then extrapolating these accuracies over the entire survey area using the inverse squared distance function. In the same scenario, other authors have adopted different extrapolation techniques, such as geographically weighted logistic regression (Comber 2013) and kriging (Steele et al. 1998). Regarding models that link accuracy to predictors available for the entire survey area, logistic regression models are the most commonly used. Accuracy is treated as the dependent variable linked to a set of covariates, such as landscape and/or topographic measures, using logistic regression and then estimated at any unsampled location from the e-quantities observed in the reference locations closest to the unsampled one (e.g., Burnicki 2011 and references therein). Alternatively, without resorting to specific models, interpolation of the e-quantities recorded at the reference locations can be performed in a nonparametric setting using kernel-based interpolators (e.g., constant, linear, and gaussian kernels), so that accuracy is estimated at any unsampled location by a weighted average of the e-quantities observed at neighbouring reference locations (Scheuerer et al. 2013). Recently, Ebrahimy et al. (2021) criticized the use of spatial interpolation and suggested the alternative use of random forest algorithms, while Comber and Tsutsumida (2023) extended A-map methods based on hard classifications, where each location is assigned to a single class, to fuzzy classifications.

Clearly, all these A-maps are model-dependent in nature, meaning explicitly that accuracy is assumed to be generated by a probabilistic model such as logistic models (geographically weighted or not) or kriging. However, in a design-based framework, where variability arises only from the probabilistic scheme adopted to perform a survey (e.g., Särndal et al. 1992, Sect. 1.10), there is no accuracy in satellite maps. That is, there is no probability of classifying a location correctly or not, as if we were playing a game in which we have a chance to guess the true LULC class. In fact, in a design-based approach, the true LULC map, based on the reference class labels, is constant, i.e., there is no stochastic model that generates it. Furthermore, as previously outlined, the training sample from which satellite maps are derived is typically not a probability sample but instead is purposively selected. Therefore, the satellite maps are also constant in the sense that once the training sample has been fixed and all decisions are made regarding the implementation of the classification algorithm, there is no variability (i.e., the classification algorithm will invariably produce the same map), and the matching of the true map and the satellite map results in a fixed dichotomous surface of error/non error, hereafter called an "e-map". Consequently, from a design-based perspective, we simply need to estimate the e-map based on the e-quantities recorded at the reference

locations. The proposal of this paper is to estimate e-maps using the nearest neighbour interpolator (NN), i.e., assigning the e-quantity observed at the nearest reference location to any unsampled location in the survey area. We refer to the maps arising from the NN interpolation of errors recorded at reference locations as the "NN e-maps".

The choice of the NN interpolator rather than other techniques, such as the inverse distance weighting interpolator, is due to the NN property that the interpolated values have the same support of the survey variable, as when the support is dichotomous, like in our case. Because of its simplicity, mapping by NN interpolation is a widely applied practice in environmental surveys (e.g., Li and Heap 2008). However, despite its common use, the NN interpolator has typically been adopted merely as a descriptive technique. From a model-dependent perspective, Cressie (1991, Sect. 5.9) relegates the NN interpolator to a class of techniques called "non-stochastic methods of spatial prediction", for which no stochastic model is assumed. This is likely due to the widespread opinion that inference in spatial mapping is difficult to perform without the use of models. Alternatively, Fattorini et al. (2022) approach the NN interpolator from a design-based perspective in which the surface to be mapped is viewed as constant and the uncertainty arises from the probabilistic sampling scheme adopted to select the reference locations. For finite sample sizes the design-based properties of the NN interpolator are unknown. However, the asymptotic design-based properties of the interpolator have been derived by Fattorini et al. (2022). This derivation requires mild conditions for the surfaces to be interpolated and the ability of the adopted sampling scheme to uniformly distribute sample locations such that as the number of sample locations increases, any unsampled location in the continuum of the survey area is likely to have neighbouring sampled locations. This feature is usually referred to as spatial balance in the sampling literature (e.g., Brown et al. 2015). In accordance with this novel approach, in this paper e-maps are estimated from a design-based perspective, where the statistical properties of the NN e-maps are uniquely determined by the sampling scheme adopted to select the reference locations.

Additional insight regarding assumptions is gained by recognizing that even though the NN interpolation of errors is tacitly based on the assumption that the presence/absence of errors in the nearest locations tends to be similar, this assumption simply constitutes an assisting model adopted only for the purpose of constructing the interpolator. This assumption is destined to be wrong to some degree because in any real situation, there will be locations where the error presence/absence will differ from the error presence/absence in the neighbouring locations. However, under suitable sampling schemes, the NN e-maps converge to the e-maps as the sample size increases. This convergence is attributable to the sampling schemes, irrespective of the validity of the assumption underlying the NN interpolator. In practice, we are in the framework of model-assisted inference in which the model is only adopted to find an appropriate estimator. However, the design-based properties of the estimator do not depend on model validity but only on the sampling scheme actually adopted in the field. In this sense, our proposal is model-assisted but not model-dependent (Särndal et al. 1992, Remark 6.4.1).

Although design-based consistency is ensured in any situation by choice of suitable sampling schemes, the precision of the NN e-maps for finite sample sizes obviously depends on the features of the e-maps to be interpolated. With respect to the two main characteristics of spatial phenomena, spatial autocorrelation and spatial heterogeneity, the precision of the NN e-maps increases with the spatial autocorrelation of the error presence. That is, the similarity of error presence in nearby locations aligns well with the NN assumption. On the other hand, the heterogeneity of errors (i.e., the different amounts of error presence in different parts of the study region), can be addressed by using sampling schemes that evenly spread reference locations throughout the study area. In practice, for finite samples, the spatial autocorrelation of errors is the crucial feature that determines the precision of the NN e-maps.

The paper is organized as follows. In Sect. 2, e-maps are defined, pointing out that these dichotomous surfaces are likely to be continuous almost everywhere. This feature is sufficient, under suitable sampling schemes, to ensure the design-based consistency of the NN e-maps at any continuous point as well as to determine their convergence rate. In Sect. 3, a simulation study is carried out by exploiting satellite maps of a square region located in Northern Tuscany (Central Italy) at years 2012 and 2021. The 2012 map is taken as the true map and the 2021 map is taken as the satellite map so that the e-map is constructed by coupling the maps of the two years. In Sect. 4, a simple descriptive measure is proposed for assessing the fraction of area estimated as error in the NN e-maps, thus providing a synthetic and intuitive evaluation of the accuracy of satellite maps from which NN e-maps are derived. In Sect. 5, two case studies are used to illustrate estimation of an e-map. One case is based on the 2017 US land cover map arising from the LCMAP program (Brown et al. 2020), and the other estimated e-map is based on the forest/non-forest map of Italy (D'Amico et al. 2021). Concluding remarks are provided in Sect. 6.

2 Error maps and their NN interpolations

Denote by A the survey area. For any point $p \in A$, let $y(p)$ be the LULC true class at p and let $y_{sat}(p)$ the LULC class predicted for p by a classification algorithm from the satellite information and training sample. By matching true and satellite maps, the e-map is given by

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

where the e-quantity $e(p) = 1$ if $y_{sat}(p) \neq y(p)$ and 0 otherwise. In the design-based approach, any e-map is a fixed and unknown characteristic of the survey area rather than the realization of a spatial model as would be assumed in the model-dependent approaches mentioned in the Introduction.

Regarding the nature of (1), it is apparent from Fig. 1 that any LULC map comprises disjoint regions with each region comprised of a single class. Changes from one class to another occur along edges, which are sets of measure 0. Likewise, any satellite map shares the same features. As a result, pairing the two maps gives rise to disjoint regions of 0's or 1's which are determined by the intersections of the regions on the two maps and changes from 0 to 1 that occur along the edges of these intersections (see Fig. 1). In mathematical terms, any e-map is continuous almost everywhere. This feature plays an important role in determining the design-based properties of the e-maps estimated from the NN interpolator.

To this purpose, denote by $P_1, \ldots, P_n$ the n reference sample locations at which the best assessments of ground truth are recorded, so that the $e(P_i)$ s can be determined by comparing the reference class with the class predicted by the satellite map. To achieve NN e-maps, we then adopt the NN interpolator, where the e-quantity at an unsampled location $p \in A$ is estimated by

$$\hat{e}(p) = e\left(P_{NN(p)}\right) \tag{2}$$

where $P_{NN(p)} = argmin_{i=1,\ldots,n}\|p - P_i\|$.

As with any other interpolator, the design-based precision of (2) should be determined for any point $p \in A$ by the quantity $E\{[\hat{e}(p) - e(p)]^2\}$. However, because of the dichotomous nature of e-maps, this expectation coincides with the failure probability

$$FP(p) = \Pr\{\hat{e}(p) \neq e(p)\} \tag{3}$$

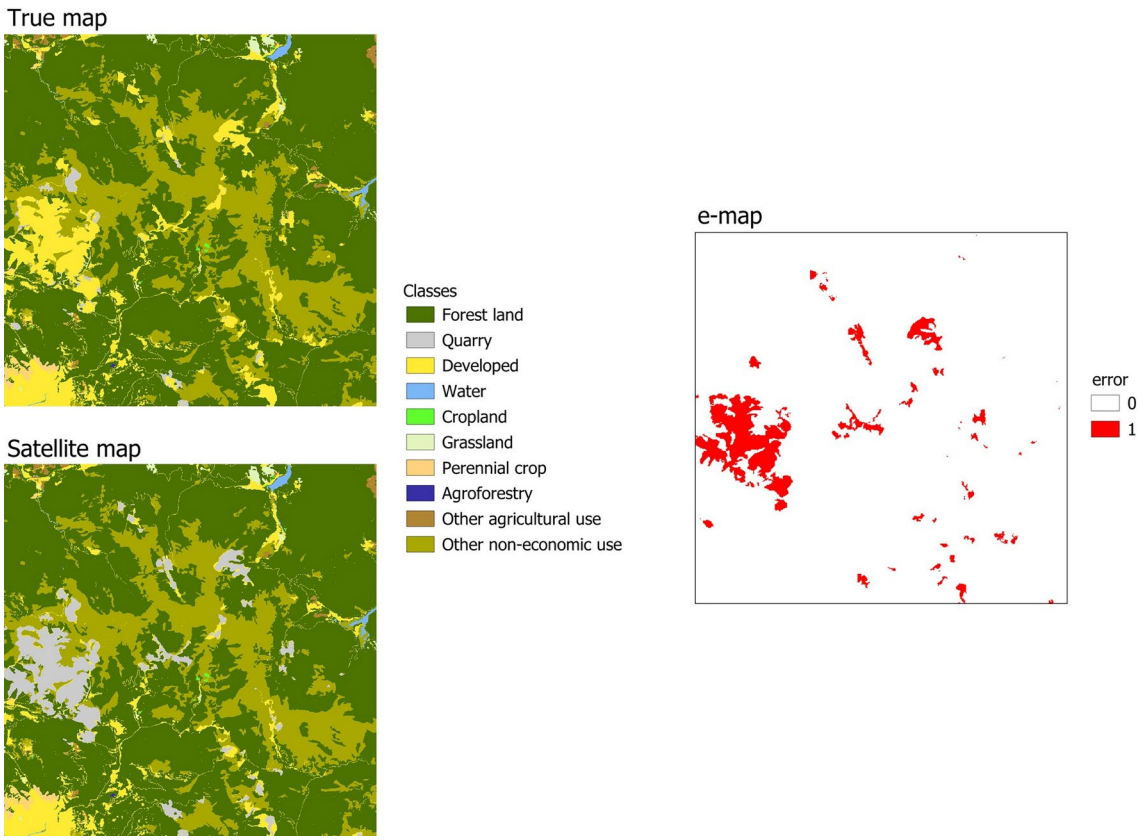

Fig. 1 True and satellite maps of the land use classes presumed for a square region located in Northern Tuscany (Central Italy) and the corresponding e-map

such that pointwise design-based consistency occurs if (3) approaches 0 as n increases. Furthermore, since all random variables involved in the NN interpolation (2) are bounded, with the e-quantities equal to 0 or 1, the design-based consistency implies design-based asymptotical unbiasedness. That is, i.e., for large n, not only are the probability distributions of the interpolated values tightly concentrated around the true e-quantities, but they are also centred on the true values (e.g., Särndal et al. 1992, p. 166).

Regarding the design-based consistency of (2), in Appendix B of the Supplementary Material file it is shown that consistency holds at any continuity point of the e-map (i.e., almost everywhere) under several sampling schemes widely applied in environmental surveys. These schemes include uniform random sampling (URS), in which n reference locations are randomly and independently selected within the survey area; tessellation stratified sampling (TSS), in which the survey area is partitioned into n patches of equal size and one reference location is randomly selected within each patch; and systematic grid sampling (SGS), in which the survey area is partitioned into n regular polygons and one reference location is randomly selected in one polygon and systematically repeated in the same position within the other polygons (e.g., Barabesi et al. 2012). A more detailed description of URS, TSS, and SGS is given in Appendix A of the Supplementary Material file.

In Appendix B, it is shown that under URS $FP(p)$ approaches zero at least as c^n with $c \in (0,1)$ at any continuity point, while under TSS and SGS $FP(p)$ is definitively equal to 0 for a sufficiently large sample size. According to these results, under URS, TSS, and SGS, $FP(p)$ converges to 0 almost everywhere in the survey area so that the integrated failure probability

$$IFP\left(\hat{e}\right) = \int_A FP\left(p\right) dp \tag{4}$$

also converges to 0 as n increases, thus determining the design-based consistency of the entire NN e-map. Often, the reference locations are selected from stratified sampling with different sampling intensities within strata in accordance with their environmental characteristics, aiming to increase the sample size for rare LULC classes (e.g., Olofsson et al. 2014). Obviously, design-based consistency holds within each stratum if the sample sizes within strata are sufficiently large and reference locations within strata are selected by schemes ensuring design-based consistency (e.g., URS, TSS, and SGS).

All these properties hold, *mutatis mutandis*, for the NN e-maps related to a single LULC class. In this case, given a class c_0, for any point $p \in A$, $y(p)$ is the dichotomous surface equal to 1 if p is located in c_0 and equal to 0 otherwise, while $y_{sat}(p)$ is the dichotomous surface equal to 1 if c_0 is the class predicted for p by a classification algorithm and equal to 0 otherwise. Also in this case, the matching of the two maps gives rise to an e-map of type (1), where the e-quantity $e\left(p\right) = 1$ if $y_{sat}(p) \neq y(p)$ and 0 otherwise (see Fig. 2).

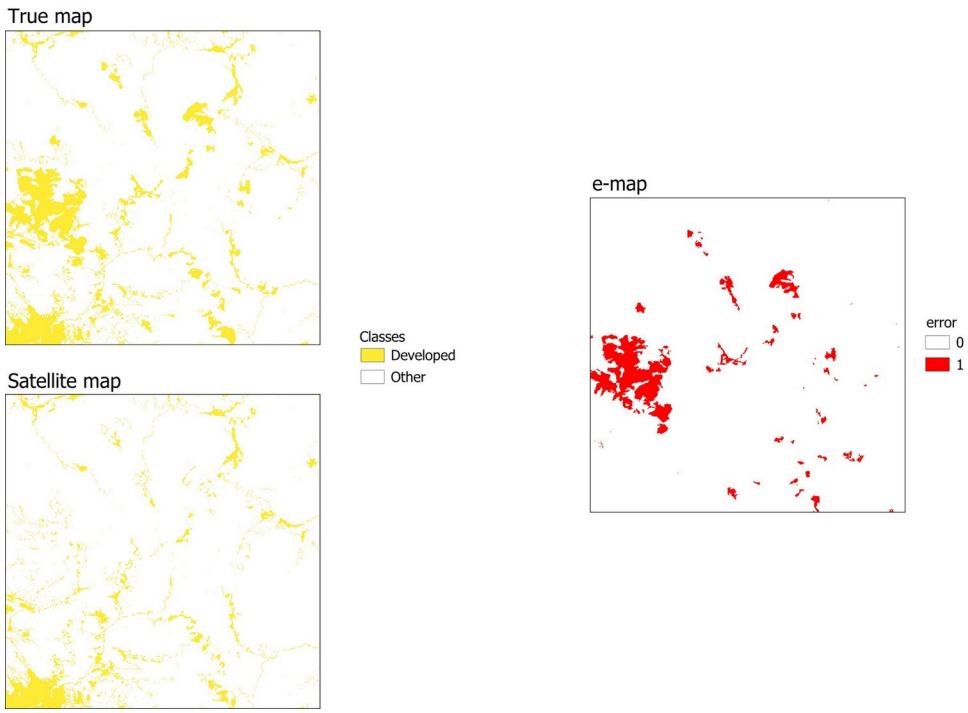

Fig. 2 True and satellite maps of a single land use class (developed) presumed for a square region located in Northern Tuscany (Central Italy) and the corresponding e-map

3 Simulation study

To investigate the performance of NN e-maps arising from interpolator (2), a simulation study was conducted on an artificial e-map. The creation of realistic e-maps would be a quite complex task because the possible sources of error in satellite maps are many and varied. For example, classification errors may be due to the large heterogeneity of LULC classes in the study region, the considered nomenclature, the definition of classes, the spectral similarity between classes, the presence of mixed pixels, and the characteristics of training data. We have simply bypassed the problem by considering a study region of $15\,\mathrm{km} \times 15\,\mathrm{km}$ located in Northern Tuscany (Central Italy) for which we had two land use maps for the years 2012 and 2021. These maps were derived from the land use maps for the entirety of Italy provided by the Italian Institute for Environmental Protection and Research (ISPRA) by exploiting Copernicus data. In particular, the 2012 land use map was obtained by merging the information from the 2012 Copernicus data and the 2012 National soil consumption map, while the 2021 land use map was obtained from the 2018 Copernicus data and the 2021 National consumption map. These maps contain ten land use classes: Forest land, Quarry, Developed, Water, Cropland, Grassland, Perennial crop, Agroforestry, Other agricultural use, Other non-economic use. Maps are available at https://groupware.sinanet.isprambiente.it/uso-copertura-e-consumo-di-suolo/library/copertura-del-suolo/carta-di-copertura-del-suolo.

For the objectives of this simulation study, the 2012 land use map was taken as the true map (i.e., the reference classification), and the 2021 map was taken as the satellite map, in such a way that the e-map was constructed by matching the maps of the two years (Fig. 1). Sampling was performed selecting $n = 100; 400; 1600; 10,000$ locations by implementing URS, TSS, and SGS. For TSS and SGS, the region was partitioned into a grid of 10×10, 20×20, 40×40 and 100×100 quadrats of equal size. For each combination of sampling scheme and sample size, $R = 10,000$ samples were independently selected. At each simulation run, interpolator (2) was computed onto a regular grid G of 1501×1501 points, one every $10m$. This grid was quite dense, providing a satisfactory resolution of the resulting NN e-maps. At the r th simulation run, we then produced the NN e-map $\widehat{e}_r(p)$ for each $p \in G$. At the end of the simulation runs, for each combination of sampling scheme and sample size, $R = 10,000$ NN e-maps were available and used for empirically determining the failure probabilities

$$FP(p) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{e}_r(p) - e(p)\right]^2, \quad p \in G$$

For each combination of sampling scheme and sample size, Table 1 reports the minima, averages, and maxima of the FPs. For TSS, Fig. 3 shows the spatial patterns of the FPs, while Fig. 4 shows the cumulative frequencies of the MSE s. Figures achieved under URS and SGS are very similar and are not reported for brevity.

The results of the simulation study are as would be anticipated from the theoretical findings. The design-consistency of the NN e-maps is confirmed by the mean values of the FP s in Table 1, which decrease with the sample size under the three considered sampling schemes. Furthermore, consistency is also apparent from Fig. 3. Indeed, under TSS, the FP maps become whiter as the sample size increases, with large areas where the FP s vanish, i.e., large areas where the NN e-maps coincide with the true e-maps. In practice, as the sample size increases, failures occur only along the borders delimiting error and non-error areas in the true e-map (Fig. 1). Regarding the cumulative frequencies of FP s (Fig. 4), the results are encouraging, considering that, regardless of the sample size, in about 90% of the area, the failure probabilities are smaller than 0.10.

Table 1 Empirical values of minima, means, and maxima of failure probabilities achieved in the simulation study for sample sizes $n = 100; 400; 1600; 10,000$ under uniform random sampling (URS), tessellation stratified sampling (TSS) and systematic grid sampling (SGS)

Scheme	n	Min	Mean	Max
URS	100	0.0000	0.0614	0.9999
	400	0.0000	0.0494	0.9997
	1600	0.0000	0.0366	0.9995
	10,000	0.0000	0.0209	0.9945
TSS	100	0.0000	0.0602	1.0000
	400	0.0000	0.0465	0.9997
	1600	0.0000	0.0344	0.9994
	10,000	0.0000	0.0189	0.9953
SGS	100	0.0000	0.0581	1.0000
	400	0.0000	0.0459	0.9998
	1600	0.0000	0.0328	0.9996
	10,000	0.0000	0.0174	0.9959

4 Descriptive indices from estimated error maps

Consider a case study in which the quality of a satellite map in a study region A has to be evaluated from a design-based perspective by means of a probabilistic sample of n reference locations $P_1, \ldots, P_n$ at which the error presences $e(P_1), \ldots, e(P_n)$ are recorded. Based on these values, the NN e-map is achieved by means of the NN interpolator (2). In principle, it would be possible to provide the estimate $\widehat{e}(p)$ for any location p in the continuum of locations

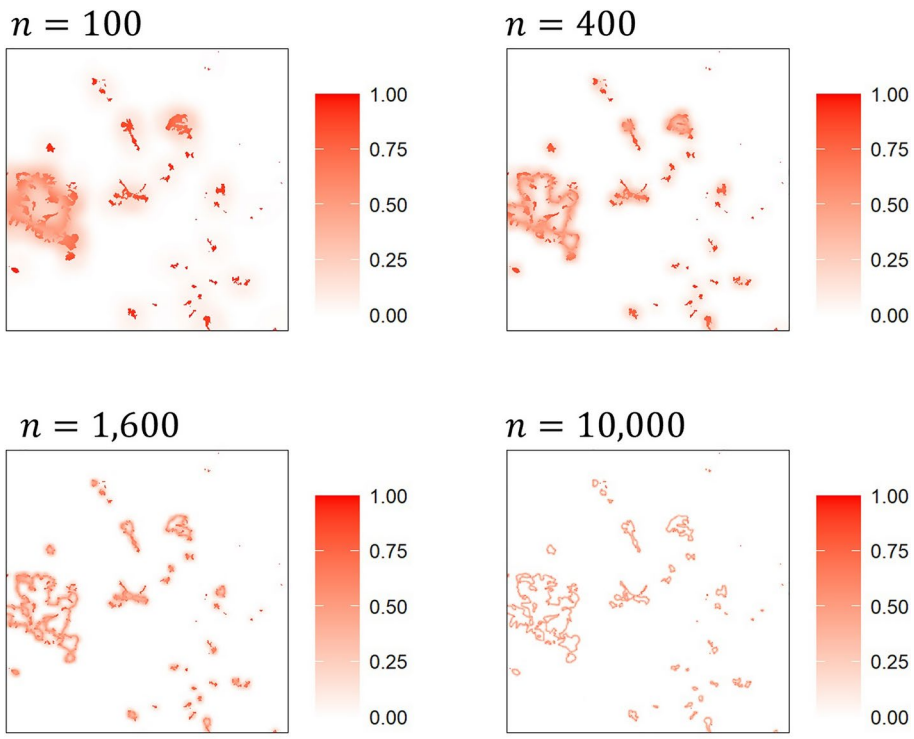

Fig. 3 Maps of empirical values of failure probabilities achieved in the simulation study under tessellation stratified sampling (TSS) for sample sizes $n =$ 100; 400; 1600; 10,000

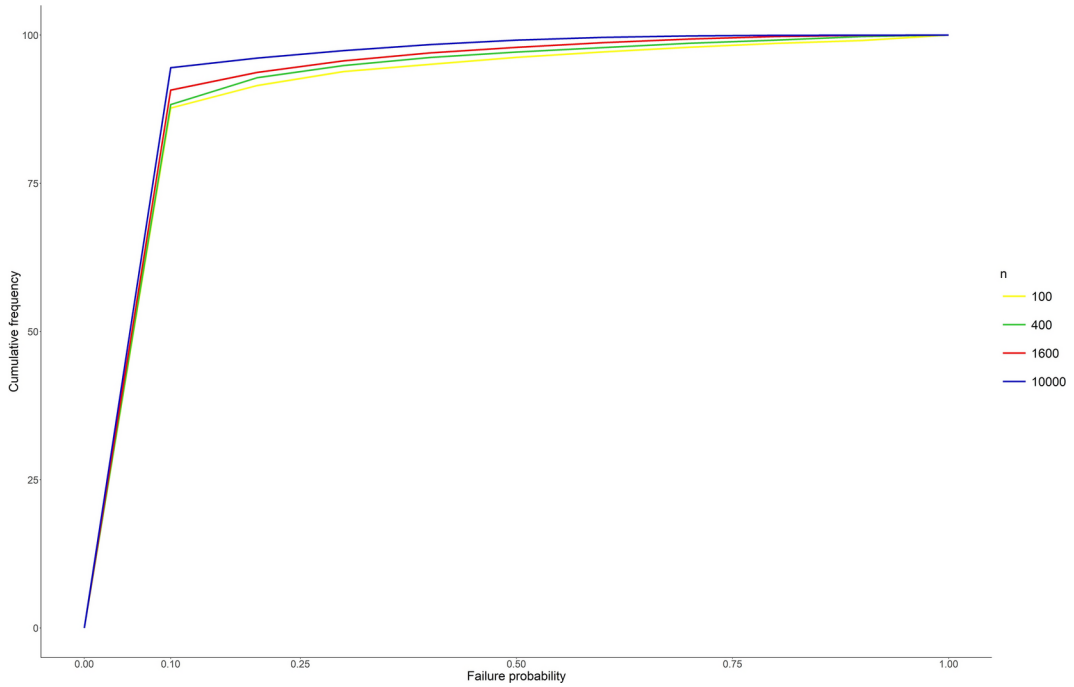

Fig. 4 Cumulative frequencies of the empirical values of failure probabilities achieved in the simulation study under tessellation stratified sampling (TSS) for sample sizes $n = 100$; 400; 1600; 10,000

constituting the study region A. However, in practice, to provide a picture of the NN e-map onto A, the estimation is performed at each location of a regular grid $G \subset A$ of M locations $p_1, \ldots, p_M$. Usually, G is quite dense on A, so that the set of estimates $\widehat{e}(p_1), \ldots, \widehat{e}(p_M)$, henceforth denoted for brevity by $\widehat{e}_1, \ldots, \widehat{e}_M$, provides a satisfactory graphical resolution. The visual display of the $\widehat{e}_g$ s, in which the locations estimated as errors are suitably highlighted by a vivid colour, provides an intuitive evaluation of the quality of the satellite map (e.g., Figs. 6 and 9).

In addition, quantitative descriptive measures combined with the graphical display are informative to support the evaluation of the satellite map quality. To this purpose, a very intuitive measure of the satellite map quality is the error area fraction (EAF). EAF is defined as the fraction of the whole area estimated as error in the NN e-map to the area $|A|$ of the whole study region, expressed as

$$EAF = \frac{1}{|A|}\int_A \hat{e}(p)dp \tag{5}$$

However, because we only know the $\hat{e}_g$ s for $g = 1, \ldots, M$, in practice it is only possible to use their arithmetic mean

$$\widehat{\bar{e}}_M = \frac{1}{M}\sum_{g=1}^{M}\hat{e}_g \tag{6}$$

which constitutes the Monte Carlo integration of (5) and can be exploited to evaluate the proportion of the error area in the NN e-map. Obviously, (6) constitutes a reliable evaluation of (5) only for adequately dense grids. Arithmetic means of type (6) can be computed, *mutatis mutandis*, for sub-portions $G_1, \ldots, G_L$ of $M_1, \ldots, M_L$ locations partitioning the whole grid. In most cases, these sub-portions correspond to sub-regions of interest such as administrative districts. In this case, the resulting sequences of means and their graphical representation will provide useful insights on where satellite maps are estimated to be more or less accurate (e.g., Figs. 7 and 10).

5 Case studies

5.1 The LCMAP survey

The US Geological Survey launched the LCMAP (Land Change Monitoring, Assessment, and Projection) program with the objective of annually mapping the land cover in the US (Brown et al. 2020). This map, henceforth referred to as the LCMAP, was produced by classifying the Landsat satellite imagery. In Fig. 5, the satellite map for the conterminous US (excluding Alaska and Hawaii) at year 2017, available at the website https://www.usgs.gov/special-topics/lcmap/collection-13-conus-science-products, is represented based on a set of $30\,\mathrm{m} \times 30\,\mathrm{m}$ Landsat pixels. The LCMAP contains eight land cover classes: Developed, Cropland, Grass/shrub, Tree cover, Water, Wetland, Snow/ice, and Barren.

To assess map accuracy and simultaneously obtain information about ground conditions, a reference sample of $n = 24,971$ pixels was selected using simple random sampling without replacement (SRSWoR) (Stehman et al. 2021). Given that the size of the pixels was considerably small compared to the entire US surface, the pixels can be considered as points in the continuum, in such a way that SRSWoR of pixels can be viewed as URS of points. This approximation allowed us to adopt the reference sample of pixels as the reference locations at which the comparison between the true on-ground class and the predicted LCMAP class was performed. Practically speaking, for each of $P_1, \ldots, P_n$ locations in the sample, $e(P_i)$ was determined by comparing the reference class with the class predicted by the LCMAP map. Using the NN interpolator (2), the e-quantities were estimated for each location p in a network of $31,751,701$ nodes (one every $500\,\mathrm{m}$) within the conterminous US, with a map resolution of $0.25\,\mathrm{km}^2$ (Fig. 6).

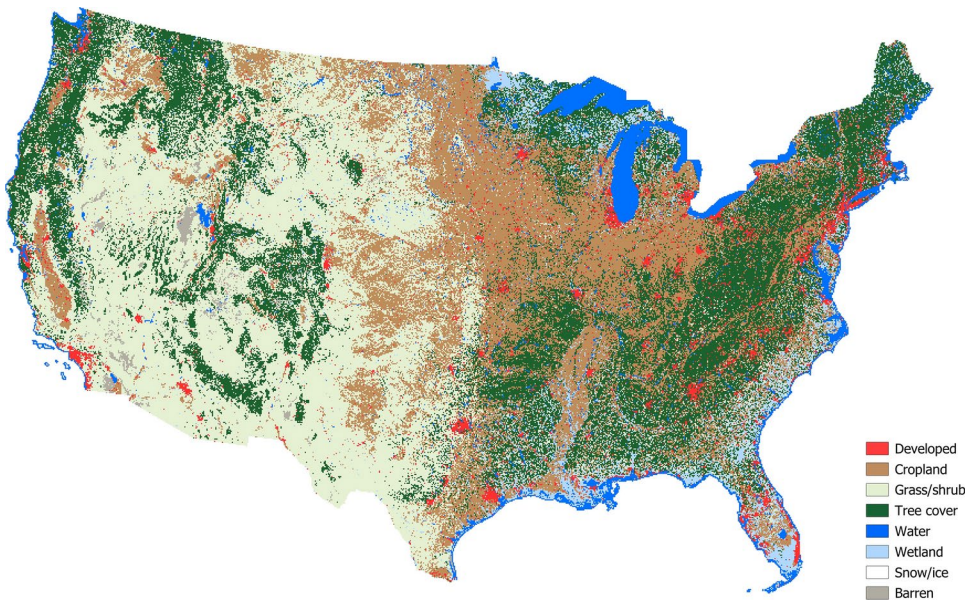

Fig. 5 Satellite map of the eight land cover classes at the year 2017 achieved in LCMAP program

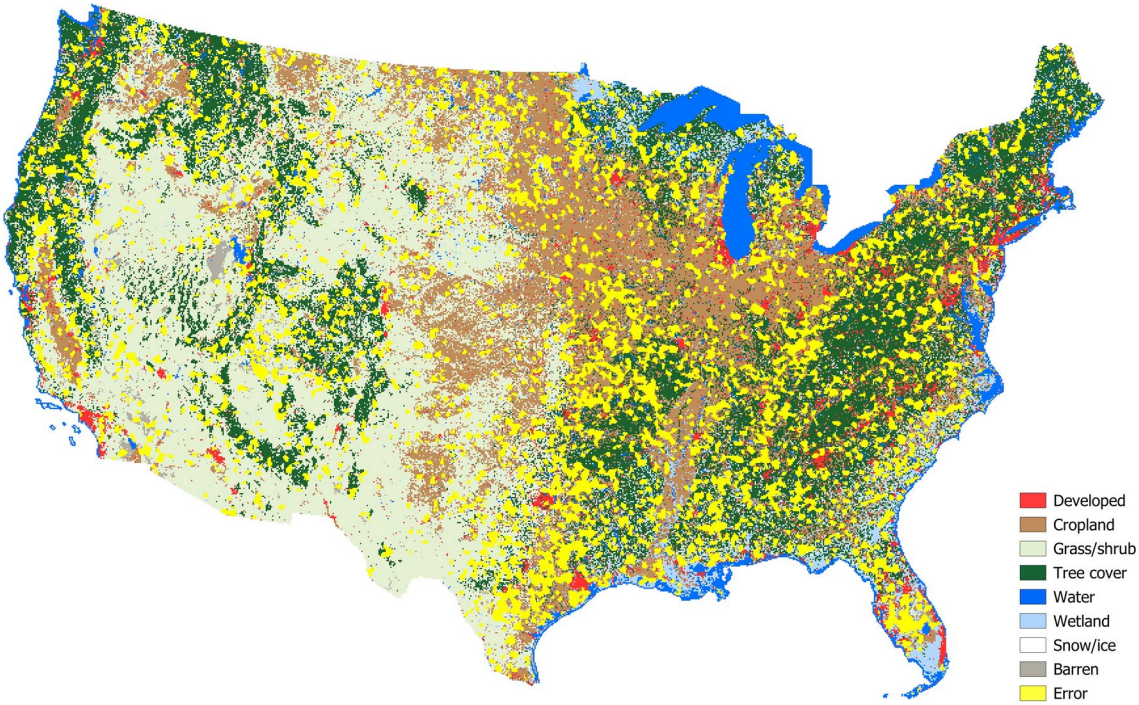

Fig. 6 NN e-map superimposed on the LCMAP for the conterminous US. Locations where an error was estimated are represented in yellow

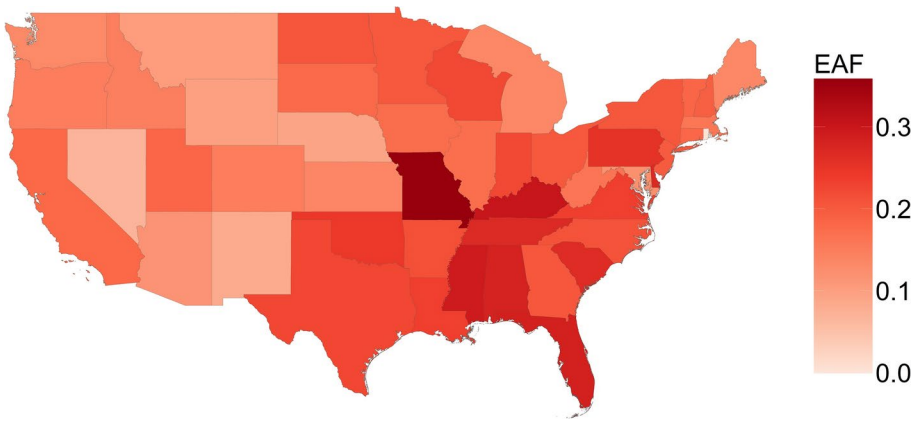

Fig. 7 Map of the fractions of the area wrongly classified by the LCMAP (EAF) in the NN e-map of Fig. 6 for the 48 states in the conterminous U.S

This grid was elaborated using the qGIS software (QGIS Development Team 2023) running on a standard workstation (AMD Ryzen 5 PRO 5650G with 6 cores, 12 logical processes, and 32 GB of RAM). The computational time for generating the NN e-map was less than 45 s using R (R Core Team 2023) on the same workstation. Much denser e-maps could be produced by increasing the computational time or by using a computer with higher performance.

On the whole, by means of (6), the area estimated as error in the NN e-map of Fig. 6 was the 18% of the conterminous US territory. As shown in Fig. 6, most of the errors were estimated where land cover heterogeneity was high, particularly in the eastern region, while few errors were estimated in the western region due to a more homogeneous land cover class distribution. For a clearer representation of the estimated quality of the LCMAP, Fig. 7 reports the map of the fractions of error presence computed by (6) for the 48 states of the conterminous US territory. Error presences of 30% or more were estimated for the southeastern states of Missouri (about 35%), Kentucky, Mississippi, Florida, and Alabama, while error presences smaller than 10% were estimated for the western states characterized by more homogeneous territories such as Nevada, New Mexico, Nebraska, Wyoming, and Montana.

We also considered two enlargements of the NN e-map in Fig. 6 for the states of Florida and Texas (Fig. 8). The large fragmentation of land cover classes in Florida (Fig. 8, right panel) resulted in a large presence of errors in the NN e-map with 28.5% of the state area estimated as error. Conversely, the less fragmented pattern of the land cover classes

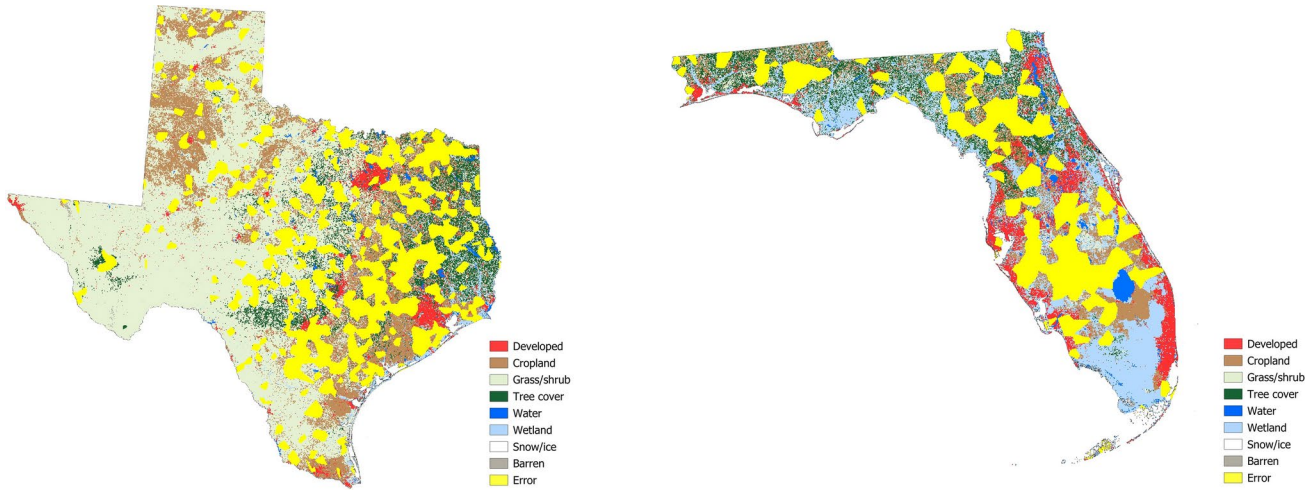

Fig. 8 NN e-maps superimposed on the LCMAP for the state of Texas (left) and Florida (right). Locations where an error was estimated are represented in yellow

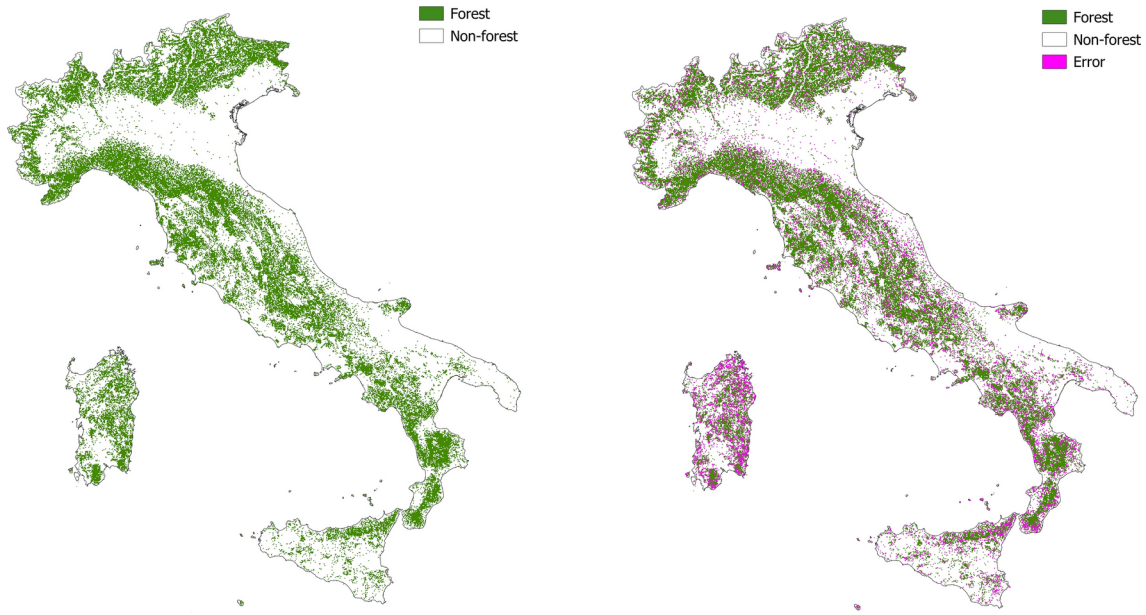

Fig. 9 Left: NFM map of Italy (D'Amico et al. 2021). Right: NN e-map superimposed on the NFM map. Locations where an error was estimated are represented in magenta

in Texas, especially in the western part of the state, led to a smaller 22.6% estimated area of errors in the NN e-map.

The results achieved from LCMAP illustrate an important feature: when the reference sampling density is sparse, the NN e-map will provide a more generalized view of the spatial distribution of classification errors. Precise, very localized error estimation becomes analogous to a small area estimation problem and may not be achievable over a broad area without a very large sample size. Still, the generalized perspective of these spatial error patterns may be informative as a cautionary indicator of where the satellite map may be less accurate.

5.2 The NFM application

For the second case study, we used the National Forest Mask (NFM) of Italy (Fig. 9, left panel). This is a high-resolution forest/non-forest map obtained from manual photo-interpretation of aerial orthophotos available at the regional scale between 2000 and 2016. When multiple years were available, the version closest to 2005 was used. For further details, see D'Amico et al. (2021).

As reference locations, we used the sample of points selected by means of TSS during the IUTI (from the Italian acronym of "Inventario dell'Uso delle Terre d'Italia")

survey of 2008, promoted by the Italian Ministry of Environment and Protection of Land and Sea. More precisely, the entirety of Italy was covered by $n = 1,217,032$ quadrats of $25\,\mathrm{ha}$. A point was randomly selected within each quadrat and then photo-interpreted to record one of six land use classes: Forest land (1), Cropland (2), Grassland (3), Wetland (4), Settlements (5) and Other lands (6). This classification was performed by ISPRA (2014) and is available on the Geoportale Nazionale at the website http://www.pcn.minambiente.it/GN/accesso-ai-servizi/servizi-di-visualizzazione-wms. In particular, the Global Forest Resource Assessment definition of forest was adopted to identify Forest land. To determine the e-quantities for each reference location of the IUTI sample, it was necessary to recode the original IUTI classes distinguishing between forest (IUTI class 1) and non-forest (IUTI classes from 2 to 6) and then compare them with the NFM classes.

Using the NN interpolator (2), the e-quantities were estimated for each location p in a network of $13,396,583$ nodes (one every $150\,\mathrm{m}$) within the Italian territory with a map resolution of $0.02\,\mathrm{km}^2$ (Fig. 9, right panel). Once again, the grid was elaborated using the qGIS software (QGIS Development Team 2023) and the computational time for generating the NN e-map was less than 30 s using R (R Core Team 2023).

On the whole, by means of (6), the area estimated as error in the NN e-map of Fig. 9 (right panel) was 9.8% of the national Italian territory. As shown in the figure, most of the error presences were estimated in the portions of the Italian territory where forest was more fragmented such as Sardinia, Calabria, and some zones of Central and South Italy around the Apennines Mountains. In contrast, a scarce presence of errors was estimated in homogeneous zones such as Apulia where forests were generally absent, and in some northern regions that were homogeneously covered by forests. For a clearer representation of the spatial distribution of the estimated quality of the NFM map, Fig. 10 reports the maps of EAF computed by (6) for the 20 administrative regions partitioning the Italian territory. The largest error presences were estimated for Sardinia (about 25%) and Calabria (about 15%), while the other regions showed estimated error presences of about 10% or less, thus showing a satisfying quality of the NFM map for most parts of the Italian territory.

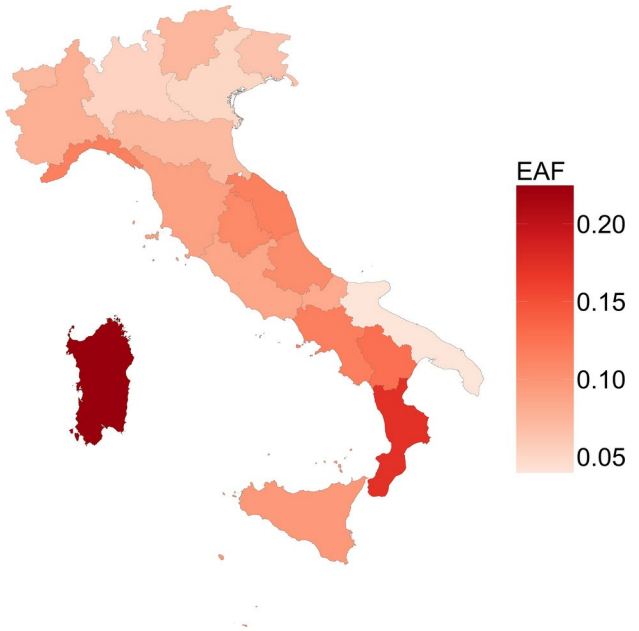

Fig. 10 Map of the fractions of the area wrongly classified by the NFM map (EAF) in the NN e-map of Fig. 9 (right) for the 20 administrative regions of Italy

We also focused on two Italian regions, Tuscany and Apulia. It is apparent from Fig. 11 (left panel) that the fragmented presence of forest in Tuscany resulted in an error presence in the NN e-map of 9% of the whole region. In contrast, the lack of forest in most of the Apulian territory led to few errors in the NN e-map (Fig. 11, right panel), with the area estimated as error being only 4% of the whole region.

Because of the larger sample size and greater density of sampling in this example compared to the LCMAP example, the NN e-map for Italy reported in Fig. 9 (right panel) provides a more reliable and localized depiction of map classification error than the one visible in Fig. 6.

6 Concluding remarks

The work applies to the problem of mapping errors in LULC maps arising from the automated classification of satellite data. The novel feature of this work is providing a way to produce error maps in a design-based setting by means of NN interpolation of the errors observed in reference samples. Under sampling schemes of wide use in environmental surveys, the NN e-maps are asymptotically design-unbiased and consistent meaning that, as the effort in collecting reference samples increases, the NN e-maps approach the true e-maps.

These results represent a valuable theoretical starting point that contradicts the scepticism of Ebrahimy et al. (2021) about spatial interpolation methods in LULC mapping, based on the fact that these methods invariably require assuming that nearby locations are more similar than those farther away. It is indeed possible that in the case of a high degree of heterogeneity in LULC composition, this assumption does not hold. However, the great appeal of our design-based proposal is that even when the similarity between neighbouring locations fails, consistency of NN e-maps is ensured by using sampling schemes that provide

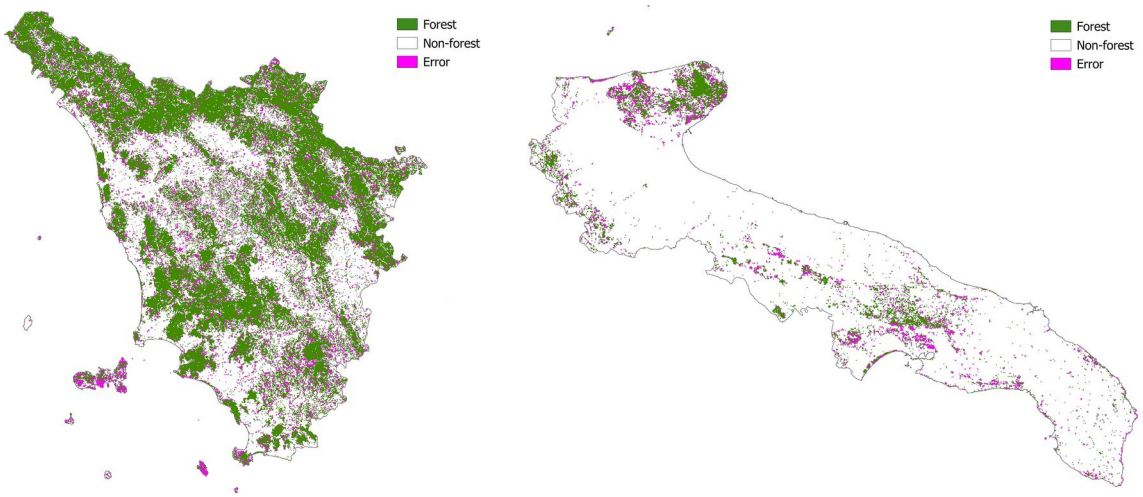

Fig. 11 NN e-maps superimposed on the NFM map for Tuscany (left) and Apulia (right). Locations where an error was estimated are represented in magenta

an asymptotic spatial balance, as in the case of URS, TSS, and SGS. Therefore, under these sampling schemes, NN e-maps become perfect in any situation as the size of reference samples increases.

Although the NN e-maps have the theoretically appealing property of design-based consistency, in real situations, the size of the reference sample is relevant to the practical utility of the resulting NN e-maps. In the presence of favourable situations, i.e., as argued in the Introduction, when there is a strong spatial autocorrelation of errors in which the assumption behind the NN interpolator fits the true e-map reasonably well, the precision of NN e-maps is good even for moderate sample sizes, without recourse to demanding sampling efforts. On the other hand, if the assumption fails, large samples are necessary to ensure a reliable result.

Regarding the case studies, inaccuracies are likely to be present in the NN e-map of the LCMAP land cover project (Fig. 6) because of the small size of the reference sample relative to the extent of the study area (about one reference pixel every 320 km^2). In such situations, the utility of an NN e-map may be limited to a broad indicator of spatial accuracy in the satellite map. Conversely, the NN e-map (Fig. 9, right panel) associated with the forest map of Italy is likely to be akin to the true e-map because the large sampling effort implemented in the IUTI survey of one point every 25 ha ensures consistency. This is also apparent from comparing the fractions of the area estimated as error in the NN e-maps, 18% in the US case compared to 9.8% in the Italian case.

From a practical point of view, our proposal allows LULC map users to be congruent and to fully operate in a design-based mode when assessing the quality of satellite maps. As accuracy estimates and standard errors of the familiar "summary" measures achieved from confusion matrices are traditionally used in a design-based inference setting, it is quite natural that the subsequent analysis of the spatial pattern of accuracy also stems from design-based inference without resorting to models. In addition, as it is apparent from Figs. 6 and 9, NN e-maps demonstrating the areas where estimation is unreliable while ignoring those areas where estimation is reliable and trustworthy are much more intuitive and easier to interpret than model-based C- and A-maps.

In principle, we could employ design-based inference to characterize the precision of NN e-maps by estimating the failure probabilities at any point of the survey area by a bootstrap procedure applied to the NN interpolator originally proposed by Fattorini et al. (2022). However, we have deliberately avoided this procedure, as NN e-maps are estimators of precision themselves, and as it is customary in inferential statistics, estimates of the precision of an estimator of precision are almost never presented in practice. That holds also in the model-based evaluation of satellite maps, where estimates of precision of C- and A-maps are never provided.

Finally, a drawback of our proposal is that NN e-maps exploit only information arising from space without taking advantage of the knowledge of covariates (i.e., auxiliary variables) often available for the whole survey region. Without exploring this possibility, we suggest that a simple way to exploit auxiliary information is to perform NN interpolation in the composite space of spatial coordinates plus auxiliary variables. In practice, among reference locations equally distant from the location where NN interpolation has to be performed, we assign to this location the value of the reference location that is nearest in terms of auxiliary variables. However, we warn against the idea of performing

NN interpolation in the sole space of auxiliary variables, neglecting spatial coordinates, because, as recently shown by Fattorini et al. (2024), design-based consistency is not ensured in these cases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00477-025-02908-2.

Acknowledgements 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). Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Award Number: Project code CN\00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B63C22000650007, Project title "National Biodiversity Future Center—NBFC".

Author contributions L. F., S.S. and P.C. contributed to the conceptualization and development of the methodology. Formal analysis was performed by L. F.. Investigations were carried out by R.D.B., A.M. and P.C.. Data analysis and coding were performed by R.D.B. and A.M.. The first draft of the manuscript was written by L. F., R.D.B., A.M. and S.S. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Siena within the CRUI-CARE Agreement.
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The authors declare that the data supporting the findings of this study are available within the paper.

Declarations

Competing interests The authors declare no competing interests.

References

Barabesi L, Franceschi S, Marcheselli M (2012) Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation. Ann Appl Stat 6:210–228. https://doi.org/10.1214/11-AOAS509

Brown JA, Robertson BL, McDonald T (2015) Spatially balanced sampling: applications to environmental surveys. Procedia Environ Sci 27:6–9. https://doi.org/10.1016/j.proenv.2015.07.108

Brown JF, Tollerud HJ, Barber CP, Zhou Q, Dwyer JL, Vogelman JE, Loveland TR, Woodcock CE, Stehman SV, Zhu Z, Pengra BW, Smith K, Horton JA, Xian G, Auch RF, Sohl TL, Sayler KL, Gallant AL, Zelenak D, Reker RR, Rover J (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. https://doi.org/10.1016/j.rse.2019.111356

Burnicki AC (2011) Modeling the probability of misclassification in a map of land cover change. Photogramm Eng Remote Sens 77:39–50. https://doi.org/10.14358/PERS.77.1.39

Comber AJ (2013) Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. Remote Sens Lett 4:373–380. https://doi.org/10.1080/2150704X.2012.736694

Comber AJ, Tsutsumida N (2023) Geographically weighted accuracy for hard and soft land cover classifications: 5 approaches with coded illustrations. Int J Remote Sens 44:6233–6257. https://doi.org/10.1080/01431161.2023.2264503

Comber AJ, Fisher P, Brunsdon C, Khmag A (2012) Spatial analysis of remote sensing image classification accuracy. Remote Sens Environ 127:237–246. https://doi.org/10.1016/j.rse.2012.09.005

Cressie N (1991) Statistics for spatial data. Wiley, New York

D'Amico G, Vangi E, Francini S, Giannetti F, Nicolaci A, Travaglini D, Massai L, Giambastiani Y, Terranova C, Chirici G (2021) Are we ready for a National Forest Information System? State of the art of forest maps and airborne laser scanning data availability in Italy. iForest 14:144–154. https://doi.org/10.3832/ifor3648-014

Ebrahimy H, Mirbagheri B, Matkan AA, Azadbakht M (2021) Per-pixel land cover accuracy: a random forest-based method with limited reference sample data. ISPRS J Photogramm Remote Sens 172:17–27. https://doi.org/10.1016/j.isprsjprs.2020.11.024

Fattorini L, Marcheselli M, Pisani C, Pratelli L (2022) Design-based properties of the nearest neighbour spatial interpolator and its bootstrap mean squared error estimator. Biometrics 78:1454–1463. https://doi.org/10.1111/biom.13505

Fattorini L, Franceschi S, Pisani C (2024) Design-based consistent strategies exploiting auxiliary information in environmental mapping. Submitted

Foody GM (2005) Local characterization of thematic classification accuracy through spatially constrained confusion matrices. Int J Remote Sens 26:1217–1228. https://doi.org/10.1080/01431160512331326521

Gomez C, White JC, Wulder MA (2016) Optical remotely sensed time series data for land cover classification: a review. ISPRS J Photogramm Remote Sens 116:55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

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

Khatami R, Mountrakis G, Stehman SV (2016) A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: general guidelines for practitioners and future research. Remote Sens Environ 177:89–100. https://doi.org/10.1016/j.rse.2016.02.028

Khatami R, Mountrakis G, Stehman SV (2017) Mapping per-pixel predicted accuracy of remote sensing images. Remote Sens Environ 191:156–167. https://doi.org/10.1016/j.rse.2017.01.025

Li J, Heap AD (2008) A review of spatial interpolation methods for environmental scientists. In: Record 2008/23, Geoscience Australia, Canberra

Nguyen HTT, Doan TM, Tomppo E, McRoberts RE (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. https://doi.org/10.3390/rs12091367

Olofsson P, Foody GM, Herold M, Stehman SV, Woodcock CE, Wulder MA (2014) Good practices for estimating area and assessing accuracy of land change. Remote Sens Environ 148:42–57. https://doi.org/10.1016/j.rse.2014.02.015

QGIS Development Team (2023) QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

R Core Team (2023) R: A language and environment for statistical computing. R Foundation. https://www.R-project.org

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York

Scheuerer M, Schaback R, Schlather M (2013) Interpolation of spatial data—a stochastic or a deterministic problem? Eur J Appl Math 24:601–629. https://doi.org/10.1017/S0956792513000016

Steele BM, Winne JC, Redmond RL (1998) Estimation and mapping of misclassification probabilities for thematic land cover maps. Remote Sens Environ 66:192–202. https://doi.org/10.1016/S0034-4257(98)00061-3

Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ 62:77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Stehman SV (2009) Sampling designs for accuracy assessment of land cover. Int J Remote Sens 30:5243–5272. https://doi.org/10.1080/01431160903131000

Stehman SV, Pengra BP, Horton JA, Wellington DF (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. https://doi.org/10.1016/j.rse.2021.112646

Valle D, Izbicki R, Vieira Leite R (2023) Quantify uncertainty in land-use land-cover classification using conformal statistics. Remote Sens Environ 295:113682. https://doi.org/10.1016/j.rse.2023.113682