Received: 5 November 2020 | Revised: 18 March 2021 | Accepted: 26 May 2021

DOI: 10.1111/biom.13505

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator

Lorenzo Fattorini[1] | **Marzia Marcheselli**[1] | **Caterina Pisani**[1] | **Luca Pratelli**[2]

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

[2] Naval Academy, Livorno, Italy

Correspondence
Marzia Marcheselli, Department of Economics and Statistics, University of Siena, P.zza S. Francesco 8, 53100 Siena, Italy.
Email: marzia.marcheselli@unisi.it

[Correction added on May 16, after first online publication: CRUI-CARE funding statement has been added.]

Abstract

Nearest neighbor spatial interpolation for mapping continuous populations and finite populations of areas or units is approached from a design-based perspective, that is, populations are fixed, and uncertainty stems from the sampling scheme adopted to select locations. We derive conditions for design-based pointwise and uniform consistency of the nearest neighbor interpolators. We prove that consistency holds under certain schemes that are widely applied in environmental and forest surveys. Furthermore, we propose a pseudopopulation bootstrap estimator of the root mean squared errors of the interpolated values. Finally, a simulation study is performed to assess the theoretical results.

KEYWORDS

environmental sampling, pointwise consistency, pseudopopulation bootstrap, spatial populations, uniform consistency

1 | INTRODUCTION

The nearest neighbor (NN) criterion is widely adopted in several fields of statistical analysis. The criterion has the appealing property of being nonparametric: it is simply based on the supposition that data that are near in some sense tend to be similar. Among other areas, the NN criterion is adopted in pattern recognition and clustering problems, where an object of unknown category is classified in the same category of its nearest observed object (e.g., Devroye *et al.*, 1996; Bremner *et al.*, 2005; Everitt *et al.*, 2011), in nonparametric regression, where the predicted value of the variable of interest for a nonobserved unit is that attached to the nearest unit in the covariate space (e.g., Stone, 1977; Altman, 1992; Terrell and Scott, 1992), and in outlier and anomaly detection, where the larger the distance of an observation is to its NN, the more likely the observation is to be an outlier (e.g., Campos *et al.*, 2016).

The NN criterion has been adopted for spatial interpolation for a long time. In the case of continuous surfaces, given a set of n sampled locations for which the surface values have been recorded, the interpolated value at any other location is the value observed at the nearest sampled location. Practically speaking, the interpolated surface is a piecewise constant function assigning the value recorded at a sampled location to each location inside the Voronoi cell around the sampled location. The NN criterion can be extended to the interpolation of values in finite populations of areas or units.

Owing to its simplicity, mapping by NN interpolation constitutes a widely extended practice in many fields of research such as, among others, environmental and ecological surveys (e.g., Li and Heap, 2008), epidemiology and air quality (e.g., Wong *et al.*, 2004, and references therein), atmospheric sciences (e.g., Chen *et al.*, 2010), and high-resolution imaging (e.g., Ashraf *et al.*, 2017).

Despite its large use, the NN spatial interpolator has been invariably adopted as a descriptive technique. From a model-dependent perspective, Cressie (1993, section 5.9) classified descriptive mapping techniques for which no

stochastic model is assumed and, as such, no uncertainty is associated, as "nonstochastic methods of spatial prediction."

However, in recent years, another nonstochastic method of spatial prediction, that is, the widely applied inverse distance weighting (IDW) interpolator, has been approached from a design-based perspective: the population to be interpolated is considered fixed, and uncertainty only stems from the probabilistic sampling scheme adopted to select locations. In IDW interpolation, the interpolated value is achieved as a convex combination of the values observed at sampled locations with weights decreasing with the distances to the location to be interpolated. Conditions ensuring design-based asymptotic unbiasedness and consistency of the IDW interpolator have been proven for continuous populations (Fattorini *et al.*, 2018a), finite populations of areas partitioning a region (Fattorini *et al.*, 2018b), and finite populations of units located in a region (Fattorini *et al.*, 2019). For the three scenarios, design-based asymptotic unbiasedness and consistency are achieved at the cost of supposing (*i*) some forms of smoothness of the survey variable throughout the study region; (*ii*) asymptotically balanced spatial sampling schemes; (*iii*) some mathematical properties of the distance functions adopted for weighting sampled observations; and (*iv*) some sort of regularities, such as in the shape of areas or in the enlargement of the populations of units, in the case of finite populations.

Regarding the distance function to adopt for weighting sample observations, it has been proven that negative powers of type $\phi(d) = d^{-\alpha}$, where d is a positive real number representing a distance and α is a positive real number, satisfy asymptotic unbiasedness and consistency of the IDW interpolator for $\alpha > 2$ (Fattorini *et al.*, 2018a; Fattorini *et al.*, 2018b; Fattorini *et al.*, 2019). Therefore, NN interpolator can be viewed as the limiting case of the IDW interpolator with weights of type $\phi(d) = d^{-\alpha}$ for α approaching infinity.

The purpose of this paper is to derive conditions sufficient to extend the asymptotic properties proven for the IDW interpolator to the NN interpolator. We do so in a unifying approach that includes the three types of spatial populations. Indeed, design-based asymptotic unbiasedness and consistency of NN interpolator cannot be straightforwardly achieved from the inequalities regarding the IDW interpolator under the three scenarios as α approaches infinity. In this way, three different results would be obtained. On the other hand, because the NN interpolator only involves the nearest sample location, asymptotic results are achieved in a more direct and less cumbersome way and are jointly valid for the three types of spatial populations. Moreover, a pseudopopulation bootstrap approach is adopted to obtain a reliable, conservative estimator of the accuracy of the NN interpolator, for which no consistency result is presently available and uncertainty assessment has been traditionally neglected or has not yet gone beyond the simple application of leave-one-out or cross-validation techniques (e.g., Chen *et al.*, 2010) without any theoretical investigation.

The paper is organized as follows. Notation and setting are introduced in Section 2. Section 3 contains some finite sample results useful for determining the asymptotic properties of the NN interpolator, which are detailed in Section 4. In Section 5, the asymptotic properties are proven to hold under familiar spatial schemes, and in Section 6, a pseudopopulation bootstrap estimator of the precision of the NN interpolator is proposed. A simulation study is described in Section 7, whereas the application of the NN interpolator for providing the forest map in a region in Casentino Valley is illustrated in Section 8. Finally, concluding remarks are given in Section 9. Supporting Information contains technical details and proofs, tables, and figures referring to the simulation study.

2 | NOTATION AND SETTING

Denote by λ the Lebesgue measure on $\mathbb{R}^2$ and by $I(E)$ the indicator function of the event E. Let Y be a survey variable, and consider a study region A that is assumed to be a compact set of $\mathbb{R}^2$. Moreover, let f be a measurable function defined on a Borelian subset B of A, with values on $[0, L]$ and such that, for any Borelian subset C of B, $\int_C f(\mathbf{p})\mu(d\mathbf{p})$ yields the amount of Y in the region C, where μ is the Lebesgue measure λ under continuous populations and population of spatial areas, whereas it is the counting measure under a population of units. Then, in accordance with the features of spatial populations, there are the following three settings.

Continuous populations

B coincides with A, and f is the density of Y, that is, $\int_C f(\mathbf{p})\lambda(d\mathbf{p})$ is the amount of Y in C. Therefore, mapping necessitates the knowledge of $f(\mathbf{p})$ for (almost) each $\mathbf{p} \in B$.

Finite populations of spatial areas

B coincides with A, which is partitioned into N areas $a_1, \ldots, a_N$, and y_j is the amount of the survey variable Y within a_j. Therefore, mapping requires knowledge of y_j for each $j = 1, \ldots, N$. As the area size $\lambda(a_j)$ is usually known for each $j = 1, \ldots, N$, mapping actually requires knowledge of the density of Y within area j, $y_j/\lambda(a_j)$, for each $j = 1, \ldots, N$, that is equivalent to the knowledge of the piecewise constant function

$$f(\mathbf{p}) = \sum_{j=1}^{N} \frac{y_j}{\lambda(a_j)} I(\mathbf{p} \in a_j)$$

for each $\mathbf{p} \in B$. In particular, $\int_B f(\mathbf{p})\lambda(d\mathbf{p}) = y_1 + \cdots + y_N$.

Finite populations of units

B is the set $\{\mathbf{p}_1, \dots, \mathbf{p}_N\}$ of N unit locations, and $y_j = f(\mathbf{p}_j)$ is the value of the survey variable for the unit j. Therefore, mapping requires the knowledge of $f(\mathbf{p}_j)$ for each $j = 1, \dots, N$. It is worth noting that $\int_B f(\mathbf{p})\mu(d\mathbf{p}) = y_1 + \cdots + y_N$, where μ is the counting measure that yields mass 1 at every $\mathbf{p}_j$ for $j = 1, \dots, N$.

Let $\mathbf{P}_1, \dots, \mathbf{P}_n$ be n random variables with values in B that represent the n locations selected from B by means of a probabilistic fixed-size sampling scheme. In the case of continuous populations, $\mathbf{P}_1, \dots, \mathbf{P}_n$ denote n locations selected in the continuum B, and $f(\mathbf{P}_1), \dots, f(\mathbf{P}_n)$ are the densities of Y recorded at those locations. In the case of finite populations of areas, $\mathbf{P}_1, \dots, \mathbf{P}_n$ denote the centroids identifying the n sampled areas, and $f(\mathbf{P}_1), \dots, f(\mathbf{P}_n)$ are the densities recorded within the corresponding areas. Finally, in the case of finite populations of units, $\mathbf{P}_1, \dots, \mathbf{P}_n$ denote the locations of n sampled units, and $f(\mathbf{P}_1), \dots, f(\mathbf{P}_n)$ are the values of Y for these units. The NN spatial interpolator $\hat{f}$ of f is

$$\hat{f}(\mathbf{p}) = I(Q_{\mathbf{p}})f(\mathbf{p}) + \frac{I(Q_{\mathbf{p}}^c)}{\mathrm{Card}(H_{\mathbf{p}})} \sum_{i \in H_{\mathbf{p}}} f(\mathbf{P}_i)\,, \ \mathbf{p} \in B, \quad (1)$$

where $Q_{\mathbf{p}} = \bigcup_{i=1}^n \{\mathbf{P}_i = \mathbf{p}\}$ and $H_{\mathbf{p}} = \{i : \ \|\mathbf{P}_i - \mathbf{p}\| = \min_{h=1,\dots n}\|\mathbf{P}_h - \mathbf{p}\|\}$.

In the continuous case, $Q_{\mathbf{p}}$ has probability 0 and $\mathrm{Card}(H_{\mathbf{p}})$ is equal to 1 almost surely, in such a way that (1) reduces almost surely to

$$\hat{f}(\mathbf{p}) = f(\mathbf{P}_i)\,, \ \mathbf{p} \in B, \quad (2)$$

where $\|\mathbf{P}_i - \mathbf{p}\| = \min_{h=1,\dots n}\|\mathbf{P}_h - \mathbf{p}\|$.

In the case of finite populations of areas, $\hat{f}(\mathbf{p}) = \hat{f}(\mathbf{b}_j)$ for each $\mathbf{p} \in a_j$, where $\mathbf{b}_j$ denotes the centroid of the jth area, and

$$\hat{f}(\mathbf{b}_j) = I(Q_{\mathbf{b}_j})f(\mathbf{b}_j) + \frac{I(Q_{\mathbf{b}_j}^c)}{\mathrm{Card}(H_{\mathbf{b}_j})} \sum_{i \in H_{\mathbf{b}_j}} f(\mathbf{P}_i)\,, \quad j = 1, \dots, N, \quad (3)$$

where $\mathrm{Card}(H_{\mathbf{b}_j})$ may be greater than 1, as, for example, in the case of populations of regular polygons (e.g., pixels).

Finally, in the case of finite populations of units, the NN interpolator is

$$\hat{f}(\mathbf{p}_j) = I(Q_{\mathbf{p}_j})f(\mathbf{p}_j) + \frac{I(Q_{\mathbf{p}_j}^c)}{\mathrm{Card}(H_{\mathbf{p}_j})} \sum_{i \in H_{\mathbf{p}_j}} f(\mathbf{P}_i)\,, \quad j = 1, \dots, N, \quad (4)$$

where, if units are settled on regular grids (e.g., net nodes), NNs may be more than 1.

Interpolator (1) is the limit of the IDW interpolator with distance function $d^{-\alpha}$

$$\hat{f}_\alpha(\mathbf{p}) = I(Q_{\mathbf{p}})f(\mathbf{p}) + I(Q_{\mathbf{p}}^c)\frac{\sum_{i=1}^n f(\mathbf{P}_i)\|\mathbf{P}_i - \mathbf{p}\|^{-\alpha}}{\sum_{i=1}^n \|\mathbf{P}_i - \mathbf{p}\|^{-\alpha}},$$

when α approaches infinity.

3 | SOME FINITE SAMPLE RESULTS

We derive some results for n finite, to be subsequently exploited for determining the asymptotic properties of the NN interpolator.

Denote by $\|\hat{f} - f\|_\infty = \sup_{\mathbf{p} \in B}|\hat{f}(\mathbf{p}) - f(\mathbf{p})|$. In the following, without loss of generality, we suppose that

$$\|\hat{f} - f\|_\infty = \sup_{\mathbf{p} \in D}|\hat{f}(\mathbf{p}) - f(\mathbf{p})| \quad (5)$$

for a suitable countable subset $D \subset B$. Indeed, (5) is true when f is continuous or B is a countable set. For any $\delta > 0$ and $\mathbf{p} \in B$, denote by

$$\Delta(\mathbf{p}, \delta) = \sup_{\mathbf{p} \in B:\, \|\mathbf{p} - \mathbf{q}\| \le \delta}|f(\mathbf{q}) - f(\mathbf{p})|$$

the largest jump of f in the δ-ball of $\mathbf{p}$ and $\Delta(\delta) = \sup_{\mathbf{p} \in D}\Delta(\mathbf{p}, \delta)$ the largest jump on D.

Moreover, denote by $A_i(\mathbf{p}, \delta) = \{\|\mathbf{P}_i - \mathbf{p}\| > \delta\}$ the event that the ith sampled location is outside the δ-ball of $\mathbf{p}$, in such a way that

$$A(\mathbf{p}, \delta) = \bigcap_{i=1}^n A_i(\mathbf{p}, \delta) = \bigcap_{i=1}^n \{\|\mathbf{P}_i - \mathbf{p}\| > \delta\}$$

is the event that no sampled location is within the δ-ball of $\mathbf{p}$.

Theorem 1. *For any $\delta > 0$ and $\mathbf{p} \in B$*

$$E\{|\hat{f}(\mathbf{p}) - f(\mathbf{p})|\} \le \Delta(\mathbf{p}, \delta) + L\Pr\{A(\mathbf{p}, \delta)\}. \quad (6)$$

Moreover, under condition (5)

$$E\{\|\hat{f} - f\|_\infty\} \le \Delta(\delta) + L\Pr\left\{\bigcup_{\mathbf{p} \in D} A(\mathbf{p}, \delta)\right\}. \quad (7)$$

Both the inequalities highlight that expectations of absolute errors are bounded by the sum of two terms: the first depending on the roughness of f and the second depending on the sampling design. Therefore, precise interpolation takes hold when both terms are small. If f is continuous at $\mathbf{p}$ or on the whole B, then the first term on the right-hand sides of (6) and (7) approaches zero with δ, and accordingly, the precision of the interpolation depends on

the features of the sampling design throughout the second terms. Practically speaking, the sampling scheme should be able to ensure a spatial balance, that is, to evenly spread sampled locations in such a way that a location, in the case of (6), or any location on the whole B, in the case of (7), is likely to have neighboring locations sampled. In turn, regarding the right-hand side of (6), the second term can be bounded on the basis of the random variable $Z(\mathbf{p},\delta) = \sum_{i=1}^{n} I\{A_i^c(\mathbf{p},\delta)\}$ representing the number of sampled locations falling in the δ-ball of $\mathbf{p}$.

Theorem 2. *For any fixed n and for any $\delta > 0$ and $\mathbf{p} \in B$*

$$\Pr\{A(\mathbf{p},\delta)\} = \Pr\{Z(\mathbf{p},\delta) = 0\} \leq \frac{1}{\sum_{i=1}^{n} \Pr\{A_i^c(\mathbf{p},\delta)\}} + \sup_{h\neq i=1,\ldots,n} \left[\frac{\Pr\{A_i^c(\mathbf{p},\delta)\cap A_h^c(\mathbf{p},\delta)\}}{\Pr\{A_i^c(\mathbf{p},\delta)\}\Pr\{A_h^c(\mathbf{p},\delta)\}} - 1\right]^{+}. \quad (8)$$

The precision of the NN interpolator deteriorates where discontinuities are present. However, the precision of the whole map is preserved if these discontinuities, as usual in practical situations, occur for sets of measure zero. Indeed, in this case, the mean integrated absolute error

$$MIAE(\hat{f}) = \int_B E\{|\hat{f}(\mathbf{p}) - f(\mathbf{p})|\}\lambda(d\mathbf{p}) \quad (9)$$

strictly depends on $\int_B \Pr\{A(\mathbf{p},\delta)\}\lambda(d\mathbf{p})$ that, in turn, will be small if the second term of (6) is small due to the effectiveness of the sampling design.

More compelling results are achieved if we suppose f to be Lipschitz continuous at $\mathbf{p}$, that is, if $|f(\mathbf{q}) - f(\mathbf{p})| \leq \beta\|\mathbf{p}-\mathbf{q}\|$ for each $\mathbf{q} \in B$, where $\beta > 0$ is the Lipschitz constant. In this case, taking $\delta = tn^{-1/2}$ for any $t > 0$, from inequalities (6) and (7), it follows that

$$E\{|\hat{f}(\mathbf{p}) - f(\mathbf{p})|\} \leq \beta t n^{-1/2} + L\Pr\{A(\mathbf{p}, tn^{-1/2})\}, \quad (10)$$

$$E(\|\hat{f} - f\|_\infty) \leq \beta t n^{-1/2} + L\Pr\{\bigcup_{\mathbf{p}\in D} A(\mathbf{p}, tn^{-1/2})\}, \quad (11)$$

respectively. The previous inequalities will be useful in investigating the asymptotic properties of interpolator (1) under suitable spatial schemes.

4 | ASYMPTOTIC RESULTS

To achieve design-based asymptotic unbiasedness and consistency of (1), the following three asymptotic scenarios are considered. All of them refer to the infill asymptotics paradigm (Cressie, 1993) and have already been exploited in Fattorini *et al.* (2018a), Fattorini *et al.* (2018b), and Fattorini *et al.* (2019).

In the case of continuous populations, a sequence of fixed-size designs to select samples of increasing size on the fixed subset B is assumed. In particular, for any natural number k, a fixed-size design selecting a sample of n_k locations $\mathbf{P}_{k,1},\ldots,\mathbf{P}_{k,n_k}$ from B is considered, with $n_k \to \infty$ as k increases, and for each $\mathbf{p} \in B$, $\hat{f}_k(\mathbf{p})$ is the NN interpolator (2) of $f(\mathbf{p})$.

In the case of finite populations of areas, B is fixed, and for any natural number k, B is partitioned into N_k units $a_{k,1},\ldots,a_{k,N_k}$ with centroids $\mathbf{b}_{k,1},\ldots,\mathbf{b}_{k,N_k}$, where, as k increases, $N_k \uparrow \infty$ and all the units decrease in size such that $\sup_{j=1,\ldots,N_k}\mathrm{diam}(a_{k,j}) \to 0$. Then, a sequence of fixed-size designs is considered to select samples of $n_k < N_k$ areas identified by their centroids $\mathbf{P}_{k,1},\ldots,\mathbf{P}_{k,n_k}$, with $n_k \to \infty$. Therefore, referring to the kth partition, for each $\mathbf{p} \in a_{k,j}$ and $j = 1,\ldots,N_k$, $\hat{f}_k(\mathbf{p}) = \hat{f}_k(\mathbf{b}_{k,j})$ is the NN interpolator (3) of the piecewise constant function $f_k(\mathbf{p})$.

Finally, in the case of finite populations of units, as is customary in the finite population asymptotic framework (Särndal *et al.*, 1992), let $\mathcal{V} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots\}$ be an infinite sequence of points onto A. A sequence $\{B_k\}$ of populations is considered where B_1 consists of the first N_1 points from $\mathcal{V}$, B_2 consists of the first N_2 points from $\mathcal{V}$ with $N_2 > N_1$, and so on, in such a way that $\{B_k\}$ turns out to be a sequence of nested populations of increasing sizes. Finally, suppose a sequence of fixed-size designs to select a sample of size n_k of units identified by the locations $\mathbf{P}_{k,1},\ldots,\mathbf{P}_{k,n_k}$ from B_k with $n_k \to \infty$. Therefore, referring to the kth population, for each $\mathbf{p}_j \in B_k$, $\hat{f}_k(\mathbf{p}_j)$ is the NN interpolator (4) of $f(\mathbf{p}_j)$.

A unique definition of design-based consistency can be given for all the asymptotic scenarios. In particular, NN interpolator (1) is pointwise design consistent at $\mathbf{p} \in B_k$ if for any $\varepsilon > 0$

$$\lim_{k\to\infty} \Pr\{|\hat{f}_k(\mathbf{p}) - f_k(\mathbf{p})| > \varepsilon\} = 0,$$

and it is uniformly consistent if

$$\lim_{k\to\infty} \Pr\{\|\hat{f}_k - f_k\|_\infty > \varepsilon\} = 0,$$

where for any k, $B_k = B$ in the cases of continuous and area populations and $f_k = f$ in the cases of continuous and unit populations. Because in all cases, the f_ks are bounded with values in $[0, L]$, pointwise or uniform design consistency also entails pointwise or uniform design asymptotic unbiasedness.

From inequalities (6) and (7), taking $\delta_k = tn_k^{-1/2}$ for any $t > 0$, the first terms of (6) and (7) approach 0 with δ_k. Therefore, pointwise and uniform consistency of (1) is obviously achieved if the sequence of sampling designs ensures that, for any $\epsilon > 0$, there exist a real $t > 0$ and an integer k_0 such that

$$\Pr\{A_k(\mathbf{p}, tn_k^{-1/2})\} < \epsilon, \quad \forall k > k_0 \tag{12}$$

or if

$$\Pr\left\{\bigcup_{\mathbf{p}\in D_k} A_k(\mathbf{p}, tn_k^{-1/2})\right\} < \epsilon, \quad \forall k > k_0, \tag{13}$$

respectively, where $A_k(\mathbf{p}, \delta) = \bigcap_{i=1}^{n_k} \{\|\mathbf{P}_{k,i} - \mathbf{p}\| > \delta\}$ and D_k is a suitable countable subset of B_k.

5 | ASYMPTOTIC BEHAVIOR UNDER FAMILIAR SPATIAL SCHEMES

Sampling locations from a continuous population can be performed by uniform random sampling (URS), that is, the random and independent selection of n locations. We prove that under URS, condition (12) invariably holds, ensuring pointwise consistency of (1). URS is probably the most straightforward scheme but may lead to uneven surveying of B.

Many schemes are available for sampling spatial locations from a continuum that are able to achieve even coverage of the study region, so-called spatial balance. Spatial balance can be obtained by the use of quite complex, explicitly tailored schemes (e.g., Stevens and Olsen, 2004; Lister and Scott, 2009). Alternatively, spatial balance can be readily obtained by simple schemes involving the tessellation of the study region into n regular polygons and the random or systematic selection of one location per polygon. The two schemes are referred to as tessellation stratified sampling (TSS) and systematic grid sampling (SGS), respectively, and are widely applied in environmental surveys, especially forest surveys at large scale (e.g., Tomppo *et al.*, 2010). Indeed, these schemes have the appealing property that, for a suitable $t > 0$, they ensure $\Pr\{\bigcup_{\mathbf{p}\in D_k} A_k(\mathbf{p}, tn_k^{-1/2})\} = 0$ and therefore the uniform consistency of the NN interpolator. Moreover, from (10) or (11), under the Lipschitz condition at $\mathbf{p}$ or for the whole B, $E\{|\hat{f}_k(\mathbf{p}) - f_k(\mathbf{p})|\}$ or $E\{\|\hat{f}_k - f_k\|_\infty\}$ are $O(n_k^{-1/2})$, that is, consistency occurs at a rate of $n_k^{-1/2}$.

Similarly, many schemes are available for sampling finite populations of areas and units that are able to achieve spatial balance. Also, in these cases, spatial balance can be obtained by the use of explicitly tailored schemes (see, e.g., Grafström and Tillé, 2013, and references therein) or by the use of simple schemes that involve the stratification of the population into n regular blocks of contiguous areas or units and the random or systematic selection of one area or one unit per block. The two schemes are referred to as one-per-stratum stratified sampling (OPSS) and systematic sampling (SYS), respectively, and have long history in the statistical literature (e.g., Breidt, 1995). Moreover, in this case, the two schemes ensure that, for a suitable $t > 0$, $\Pr\{\bigcup_{\mathbf{p}\in D_k} A_k(\mathbf{p}, tn_k^{-1/2})\} = 0$. Therefore, they ensure uniform consistency and, under the Lipschitz condition, consistency occurs at a rate of $n_k^{-1/2}$.

Regarding the concept of spatial balance in finite populations of spatial areas and units, Stevens and Olsen (2004) link this concept to the NN structure, proposing to quantify the spatial balance of a sample by the variance of the sums of inclusion probabilities of those units lying in the Voronoi polygons determined by the sample units (on the issue, see also Grafström *et al.*, 2014). However, it should be noted that the asymptotical spatial balance involved by condition (12) or (13) can be achieved even by schemes not explicitly intended to achieve spatial balance. One of these schemes is the so-called 3P sampling (from the acronym of probability proportional to prediction). Indeed, regarding populations of units, their mapping is precluded in the absence of population lists and locations. Therefore, mapping is unfeasible in forest and environmental surveys where populations are communities scattered over large areas without any possibility of having lists. Probably, the unique relevant case in which the mapping of natural populations is possible is under 3P sampling. The scheme is a variation of Poisson sampling: all the units in the population are visited by a crew of experts (and hence listed and mapped), a prediction x_j for the value of the survey variable is given for each unit j, and units are independently included in the sample with probabilities $\pi_j = x_j/L^*$, where $L^* > L$ is chosen to ensure $\pi_j \leq 1$ for each j and adequate values of expected sample size (e.g., Gregoire and Valentine, 2008). Because prediction errors $e_j = y_j - x_j$ are known for each sampled unit, Fattorini *et al.* (2019) suggest interpolating the e_js instead of the y_js and then achieving the interpolated Y-values by means of $\hat{y}_j = x_j + \hat{e}_j$ for each $j \in B$. Even if prediction errors can take negative values, they are bounded by L in such a way that all the consistency results continue to hold. The mapping improvement with respect to the direct interpolation of the y_js has been investigated by Fattorini *et al.* (2020) and has been proven to be relevant.

To prove the pointwise consistency of (1) under 3P sampling, analogously to Fattorini *et al.* (2019), we further assume the following condition: $V = \{\mathbf{p}_1, \mathbf{p}_2, ...\}$ is regular, that is, for any $\mathbf{p}_j \in V$ and for any natural number m, there exist a real number $t > 0$ and an integer k_0 such that

$$\text{Card}\{B_j(tN_k^{-1/2}) \cap B_k\} > m, \quad \forall k > k_0, \tag{14}$$

where $B_j(\delta)$ is the set of units of V in the δ-ball of unit j. Condition (14) requires that the populations in the sequence increase in such a way that, for a sufficiently large k, any unit has many neighboring units around it. Under this condition, we prove that 3P sampling ensures the consistency of (1) when the inclusion probabilities $\pi_{j,k}$ are invariably greater than a threshold $\pi_0 > 0$ for any $j \in B_k$ and any k. A lower bound for Y is common in forest and environmental surveys in which units with Y-values (e.g., tree height or basal area) smaller than a given threshold $l > 0$ are not considered in the population such that $\pi_0 = l/L^*$.

6 | PSEUDOPOPULATION BOOTSTRAP ESTIMATION OF PRECISION

Mashreghi *et al.* (2016) provide extended surveys of the bootstrap methods adopted for design-based inference. Among them, the pseudopopulation bootstrap is based on constructing a pseudopopulation likely to resemble the true population from which bootstrap samples are selected using the same sampling scheme adopted in the survey. In this setting, the key problem is to reconstruct pseudopopulations able to mimic the characteristics of the unknown populations in such a way that the bootstrap distribution of a statistic resembles the true distribution with bootstrap mean squared error approaching the true one (e.g., Quatemberg, 2016). Accordingly, to estimate the precision of (1), we use the estimated maps as pseudopopulations from which bootstrap samples are selected by means of the same spatial scheme adopted to select the original sample. If estimated maps converge to true ones, the bootstrap distributions of the NN interpolator achieved from resampling from these maps should converge to the true distributions, also providing reliable estimators of their mean squared errors.

Let $\hat{f}(B) = \{\hat{f}(\mathbf{p}), \mathbf{p} \in B\}$ be the estimated map based on $f(\mathbf{P}_1), \dots, f(\mathbf{P}_n)$. For each $\mathbf{p} \in B$, the pseudopopulation bootstrap estimator of the root mean squared error of $\hat{f}(\mathbf{p})$ is

$$\hat{rmse}^*_M(\mathbf{p}) = \left[\frac{1}{M}\sum_{m=1}^{M}\{\hat{f}^*_m(\mathbf{p}) - \hat{f}(\mathbf{p})\}^2\right]^{1/2}, \tag{15}$$

where M is the number of bootstrap samples and $\hat{f}^*_m(\mathbf{p})$ is the bootstrapped value of the NN interpolator at $\mathbf{p} \in B$ based on $\hat{f}(\mathbf{P}^*_{1,m}), \dots, \hat{f}(\mathbf{P}^*_{n,m})$ (obtained from the estimated map $\hat{f}(B)$), that is, for any $\mathbf{p} \in B$ and $m = 1, \dots, M$

$$\hat{f}^*_m(\mathbf{p}) = I(Q^*_{\mathbf{p},m})\hat{f}(\mathbf{p}) + \frac{I(Q^{*c}_{\mathbf{p},m})}{\text{Card}(H^*_{\mathbf{p},m})}\sum_{i \in H^*_{\mathbf{p},m}}\hat{f}(\mathbf{P}^*_{i,m}), \tag{16}$$

where $\mathbf{P}^*_{1,m}, \dots, \mathbf{P}^*_{n,m}$ are the locations selected in the mth bootstrap resampling, $Q^*_{\mathbf{p},m} = \cup_{i=1}^{n}\{\mathbf{P}^*_{i,m} = \mathbf{p}\}$ and $H^*_{\mathbf{p},m} = \{i : \|\mathbf{P}^*_{i,m} - \mathbf{p}\| = \min_{h=1,\dots n}\|\mathbf{P}^*_{h,m} - \mathbf{p}\|\}$.

We obtain a finite sample result about (15) supposing two further conditions: (i) for a given sample size n, the sampling design ensures the existence of a $\delta > 0$ such that

$$\Pr\{A(\mathbf{p}, \delta)\} = 0, \tag{17}$$

(ii) there exist a vector $\mathbf{a} \in \mathbb{R}^2$, $\mathbf{a} \neq \mathbf{0}$ and a function $\mathbf{q} \mapsto o(\|\mathbf{q} - \mathbf{p}\|)$ negligible with respect to $\|\mathbf{q} - \mathbf{p}\|$, such that

$$f(\mathbf{P}_i) = f(\mathbf{p}) + \langle \mathbf{a}, \mathbf{P}_i - \mathbf{p}\rangle + o(\|\mathbf{P}_i - \mathbf{p}\|), \quad i = 1, \dots, n. \tag{18}$$

Theorem 3. *For a given n, under conditions (17) and (18) and for M large enough,*

$$\frac{E\{\hat{rmse}^*_M(\mathbf{p})\}}{E[\{\hat{f}(\mathbf{p}) - f(\mathbf{p})\}^2]^{1/2}} \leq 3. \tag{19}$$

The requirement of M being large enough can be readily satisfied by increasing the computational effort. Condition (17) is less restrictive than condition (12), and it holds for all the sampling schemes discussed in Section 5 that ensure the pointwise consistency of the NN interpolator. On the other hand, condition (18) requires that in the case of continuous populations and finite populations of areas, f is differentiable at $\mathbf{p}$ with $\nabla f(\mathbf{p}) \neq 0$, whereas this requirement is not necessary for finite populations of points. Practically speaking, Theorem 3 states that, under suitable conditions, the pseudopopulation bootstrap estimator (15) tends to be conservative, with its expectation being at most three times greater than the true root mean squared error. Even if the result may induce one to suspect substantial overestimation that may mask the effectiveness of interpolation, 3 is just a threshold limiting possible overestimation. Finally, for $n \to \infty$ and for M sufficiently large, the consistency of (15) is obvious from (19). Indeed, from (19), for any n, it holds that (15) is bounded by three times the true root mean squared error. However, owing to the consistency of the NN interpolator under the required conditions, the true root mean squared error tends to 0 as n increases, so that (15) tends to 0 a fortiori.

7 | SIMULATION STUDIES

We consider three artificial surfaces on the unit square A to generate continuous populations, finite populations of areas, and finite populations of units, referred to as surface 1, surface 2, and surface 3, and, respectively, defined at any location $\mathbf{p} = (p_1, p_2)$ as

$$f(\mathbf{p}) = \frac{C_1}{2}(\sin^2 p_1 + \cos^2 p_2 + p_1), \quad f(\mathbf{p}) = C_2(\sin 3p_1 \sin^2 3p_2),$$

$$f(\mathbf{p}) = \begin{cases} C_3 p_1 p_2 & \text{if } \min(p_1, p_2) \le 1/2 \\ C_3(1 + p_1)p_2 & \text{otherwise.} \end{cases}$$

where the constants C_1, C_2, and C_3 ensure a maximum value $L = 10$. The three surfaces are represented in Web Figure 3 in the Supporting Information.

Regarding continuous populations, the three surfaces were taken as population values on $B = A$. Sampling was performed selecting $n = 16, 36, 64, 100$ locations on B by means of URS, TSS, and SGS. The last two schemes were performed by partitioning B into 4×4, 6×6, 8×8, and 10×10 grids of equal-sized quadrats and selecting a location in each quadrat.

Regarding finite populations of areas, the three surfaces were used to generate the Y-values within the areas. For each surface, four populations of $N = 100, 400, 900, 1600$ areas were constructed by partitioning $B = A$ into grids of 10×10, 20×20, 30×30, 40×40 quadrats and taking the integrals of the surface within quadrats as population values from which densities are derived. Sampling was performed by selecting $n = 0.1N$ quadrats by means of simple random sampling without replacement (SRSWOR), OPSS, and SYS. The last two schemes were performed by partitioning grids into blocks of 2×5 contiguous quadrats and selecting one quadrat per block. Regarding unit populations, three nested populations of 500, 1000, and 1500 units were located on A in accordance with four spatial patterns referred to as regular, random, trended, and clustered patterns. For the regular pattern, populations were constructed by independently generating the first 500 locations at random but discarding those having distances smaller than $0.5 \times 500^{-1/2}$ to those previously generated, then adding 500 further locations at random but discarding those with distances smaller than $0.5 \times 1000^{-1/2}$ to those previously generated, and finally randomly adding a further 500 locations but discarding those having distances smaller than $0.5 \times 1500^{-1/2}$ to those previously generated. For the random pattern, populations were constructed by independently generating 1500 locations at random on A and then assigning the first 500 to the smaller population, the first 1000 to the second, and all of them to the largest. For the trended pattern, populations were constructed by independently generating 1500 pairs of random numbers u_1, u_2 uniformly distributed on [0,1], performing the transformation $(1 - u_1^2, 1 - u_2^2)$ to determine locations, and then assigning the first 500 locations to the smallest population, the first 1000 to the second, and all of them to the largest. For the clustered pattern, populations were constructed by independently generating 10 cluster centeres at random on A and assigning 50 locations to each cluster generated from a spherical normal distribution centered at the cluster center with variance 0.025, adding a further 50 locations to each cluster from the same distribution and finally adding a further 50 locations to each cluster from the same distribution. Points falling outside A were discarded and newly generated.

The three surfaces were used for assigning the Y-values in the populations. 3P sampling with $L^* = 50$ was adopted to select units. Units with Y-values smaller than $l = 4$ were discarded from the populations to ensure a lower bound of $\pi_0 = 0.08$ for the inclusion probabilities. Expert predictions for the y_js were generated using the relationship $x_j = a + by_j$ with $b = 1 - \rho(L + l)/(L - l)$ and $a = (1 + \rho)l - bl$ in accordance with Fattorini *et al.* (2020), assuming that predictions increased linearly with Y-values with a maximum error rate $\rho = 0.10$ occurring at the extremes. The predictions, joined with the l and L^* choices, ensured an expected sampling fraction of approximately 12% in all cases.

For each combination of population, sampling scheme, and sample size, sampling was replicated $R = 10,000$ times. At each simulation run r, the estimated map $\hat{f}_r(\mathbf{p})$, $\mathbf{p} \in B$ was obtained from (1), and $M = 1000$ bootstrap samples were independently selected from the estimated map adopting the same scheme adopted to select the original sample to compute the bootstrap root mean squared error of $\hat{f}_r(\mathbf{p})$ for each $\mathbf{p} \in B$ by means of Equation (15). In the case of continuous populations, mapping was performed by computing $\hat{f}_r(\mathbf{p})$ for a regular grid of 100×100 locations on B.

Based on the Monte Carlo distributions, Web Tables 1–9 of the Supporting Information report the minima, averages, and maxima of the absolute bias and root mean squared error of (1) and of the ratio of the expectation of bootstrap root mean squared error (15) to the true value of root mean squared error. Web Figures 4–33 of the Supporting Information show the spatial patterns of these performance indicators.

Simulation results confirm the theoretical findings.

For populations generated from the continuous surfaces 1 and 2, a sharp decrease in the minima, averages, and maxima of absolute bias values and of mean squared errors occurs as the sample size (continuous populations) or population and sample sizes (finite populations of areas and units) increase.

For continuous populations and finite populations of areas generated from surface 3, with discontinuity at the internal edges of the upper-right quadrant of the unit square, decreases occur only for minima and averages, whereas maxima decrease more slowly. However, as mean squared error averages can be viewed as the empirical counterparts of mean integrated absolute error (9), the consistency of maps is preserved overall, notwithstanding discontinuities. That is also apparent from Web Figures 10–12 and 19–21 of the Supporting Information.

For finite populations of units generated from surface 3, the slow decrease in maxima disappears. In these cases, maxima decrease as the minima and averages because discontinuities in Y-values are absorbed by the corresponding predictions, providing negligible jumps in prediction errors. On the whole, decreases that occurred in finite populations of units are less marked than those that occurred in continuous populations and finite populations of areas because the use of prediction errors provides formidable gains in precision even for small population and sample sizes, leaving limited room for improvement as sizes increase.

Regarding bootstrap root mean squared errors, the ratios of their expectations to the true root mean squared error are, with very few exceptions, invariably greater than one on average and tend to asymptotically increase toward 1.2–1.5. Minima of this ratio evidence the presence of underestimation that tends to decline asymptotically, thus confirming the conservative nature of the bootstrap root mean squared error estimator. That is apparent from the web figures, where the lighter zones, corresponding to underestimation, become continually decrease as population and sample sizes increase. Maxima of this ratio evidence the possibilities of large overestimation that occur in presence of discontinuities (surface 3) and especially under systematic schemes. In the other cases, the maxima rarely exceed 2.

8 | CASE STUDY

The NN interpolator was adopted to provide the forest map in a region A of 4900 ha located in Casentino Valley, the Eastern part of the Tuscany Region (Central Italy). The forest land is mainly characterized by mountainous beech forest, coniferous forest, and thermophilous deciduous forest. The climate is temperate-humid: mean annual temperature is approximately 10°C, and total annual rainfall is greater than 1000 mm, with an average of more than 55 mm in the summer months (June–August). The forest grows on sandy-loamy or loamy soils, rich in humus on the surface horizons. Soil depth varies. The slopes are generally steep or very steep. The area is characterized

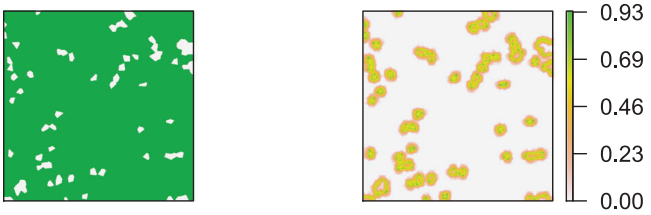

FIGURE 1 (a) Estimated map of forest (not-white) and not-forest (white) presence (left). (b) Map of estimated precision (right). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

by forest exploitation, and thus, the estimation of a map for the dichotomous survey variable forest/not-forest land is essential for analyzing the effects of human activities. Specifically, the function f related to the dichotomous survey variable is such that, for each $\mathbf{p} \in A$, $f(\mathbf{p})$ is equal to 1 if $\mathbf{p}$ is in the forest and equal to 0 otherwise.

The survey was performed in 2013, and sample locations were selected by means of TSS. The study region was partitioned into 1225 quadrats of 200 m side, and a location was randomly selected within each quadrat. Each sample location was assigned a value of 1 if lying in the forest and a value of 0 otherwise. Based on the 1225 sample locations, the NN interpolator (2) was adopted to estimate $f(\mathbf{p})$ for each location in the regular network of 1000×1000 nodes within A. The large number of locations at which estimation was performed ensured a good resolution of the resulting map displayed in Figure 1(a), which evidences the massive presence of forested land in the study region, notwithstanding the intensive management.

Moreover, at each location of the network, we also estimated the root mean squared error based on $B = 1000$ bootstrap samples of size 1225 selected from the pseudopopulation of the interpolated values using the same sampling scheme adopted to select the original sample. The map of bootstrap root mean squared errors is reported in Figure 1(b), which shows that uncertainty increases when there is a change from forest to not-forest, as expected owing to the theoretical findings. Indeed, when the dichotomous variable forest/not-forest jumps from 1 to 0 along forest edges, f exhibits discontinuities, and the precision of the NN interpolator deteriorates.

9 | CONCLUSIONS

Conditions for design-based consistency of maps achieved from NN interpolation for continuous populations and finite populations of areas or units are given. Beyond the condition on the smoothness of surfaces generating populations, consistency conditions only regard the features

of the sampling schemes adopted to select points, areas, or units. The use of TSS or SGS in continuous populations, the use of OPSS and SYS in finite populations of areas and of 3P sampling in finite populations of units ensures consistency. The focus on these schemes is not incidental; they nearly cover the range of possibilities to be adopted in forest and environmental surveys, as naturalists tend to avoid complex schemes, preferring schemes that are simple to be implemented and achieve spatial balance. Therefore, the achieved consistency results add statistical rigor to an interpolation technique widely used in environmental surveys with descriptive aims without attempting inference.

In addition, apart from when NN interpolation is applied in finite populations of regular polygons or finite populations of units located on a regular network, when NNs may be more than one and the interpolator is the convex combination of sample data recorded at these nearest locations, in most cases, there is only a single neighbor. Thus, the interpolated values have the same support as the Y variable, even when the support is discrete. This allows the application of NN interpolation for constructing maps of dichotomous 0–1 variables as in the case study, where a map of forest/not-forest land is obtained. Relevantly, the resulting surfaces are piecewise constant with discontinuities along borders of 0 measure, and as such are Lipschitizian almost everywhere, thus providing consistency at a rate of $n^{-1/2}$ under suitable stratified and SYS schemes. This is of practical importance in certain applications such as land use and land cover mapping, a vital issue in the present period of substantial deforestation and urban sprawl. Indeed, the accuracy of land cover maps and its reliable estimation, which has a long tradition in the literature (e.g., Stehman and Czaplewski, 1998; Stehman, 2009; and references therein) can be straightforwardly and rigorously addressed in a design-based approach by means of NN interpolation and bootstrap estimation of mean squared errors. Even if in this case, the survey variable is of multivariate nature being equal to the kth vector of the standard basis of $\mathbb{R}^K$ when the point is in the kth land category ($k = 1, \dots, K$), the consistency results continue to hold marginally for each map of the K land categories.

Finally, when auxiliary variables are available for the whole study region at little or no cost, as usually occurs in forest inventories (e.g., Opsomer *et al.*, 2007), NN interpolation can be performed in the auxiliary space, that is, the interpolated value at a location is the value observed at the sample location that is nearest in the auxiliary space. This intriguing idea has been empirically investigated by Grafström *et al.* (2014), achieving promising results that, however, necessitate further theoretical investigations to be fully confirmed.

ACKNOWLEDGMENTS

The authors wish to thank Piermaria Corona, Director of the Forestry Research Centre, for stimulating this research and providing many suggestions and data and Nicola Puletti, from Forestry Research Centre, for his assistance with the case study.

OpenAccess Funding provided by Universita degli Studi di Siena within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information of this article.

ORCID

Marzia Marcheselli https://orcid.org/0000-0003-2361-5289

REFERENCES

Altman, N.S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46, 175–185.

Ashraf, I., Hur, S., and Park, Y. (2017) An investigation of interpolation techniques to generate 2D intensity image from LIDAR data. *IEEE Access*, 5, 8250–8260.

Breidt, F.J. (1995) Markov chain designs for one-per-stratum sampling. *Survey Methodology*, 21, 63–70.

Bremner, D., Demaine, E., Erickson, J., Iacono, J., Langerman, S., Morin, P. and Toussaint, G.T. (2005) Output-sensitive algorithms for computing nearest-neighbor decision boundaries. *Discrete & Computational Geometry*, 33, 593–604.

Campos, G.O., Zimek, A., Sander, J., Campello, R.J.G.B., Micenková, B., Schubert, E., Assent, I. and Houle, M.E. (2016) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30, 891–927.

Chen, D., Ou, T., Gong, L., Xu, C.Y., Li, W., Ho, C.H. and Qian, W. (2010) Spatial interpolation of daily precipitation in China: 1951–2005. *Advances in Atmospheric Sciences*, 27, 1221–1232.

Cressie, N. (1993) *Statistics for Spatial Data*. New York: Wiley.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. Berlin: Springer.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th ed. New York: Wiley.

Fattorini, L., Franceschi, S. and Corona, P. (2020) Design-based mapping of tree attributes by 3P sampling. *Biometrical Journal*, 62, 1810–1825.

Fattorini, L., Marcheselli, M., Pisani, C. and Pratelli, L. (2018a) Design-based maps for continuous spatial populations. *Biometrika*, 105, 419–429.

Fattorini, L., Marcheselli, M. and Pratelli, L. (2018b) Design-based maps for finite populations of spatial units. *Journal of the American Statistical Association*, 113, 686–697.

Fattorini, L., Marcheselli, M., Pisani, C. and Pratelli, L. (2019) Design-based mapping for finite populations of marked points. *Electronic Journal of Statistics*, 13, 2121–2149.

Grafström, A., Saarela, S. and Ene, L.T. (2014) Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research*, 44, 1156–1164.

Grafström, A. and Tillé, Y. (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24, 120–131.

Gregoire, T.G. and Valentine, H.T. (2008) *Sampling Strategies for Natural Resources and the Environment*. Boca Raton: Chapman & Hall/CRC.

Li, J. and Heap, A.D. (2008) *A review of spatial interpolation methods for environmental scientists*. Record 2008/23. Camberra: Geoscience Australia.

Lister, A.J. and Scott, C.T. (2009) Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment*, 149, 71–80.

Mashreghi, Z., Haziza, D. and Léger, C. (2016) A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1–52.

Opsomer, J.D., Breidt, F.G., Moisen, G.G., and Kauermann, G. (2007) Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102, 400–416.

Quatemberg, A. (2016) *Pseudo-Populations. A Basic Concept in Statistical Surveys*. Berlin: Springer.

Särndal, C.E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Stehman, S.V. (2009) Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, 30, 5243–5272.

Stehman, S.V. and Czaplewski, R.L. (1998) Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, 64, 331–344.

Stevens, D.L. and Olsen A.R., (2004) Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99, 262–278.

Stone, C.J. (1977) Consistent nonparametric regression. *Annals of Statistics*, 5, 595–620.

Terrell, G.R. and Scott, D.W. (1992) Variable kernel density estimation. *Annals of Statistics*, 20, 1236–1265.

Tomppo, L.M., Gschwantner, R.E. and McRoberts, R.E. (2010) *National Forest Inventories: Pathways for Common Reporting*. Heidelberg: Springer.

Wong, D.W., Yuan, L. and Perlin, S.A. (2004) Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science & Environmental Epidemiology*, 14, 404–415.

SUPPORTING INFORMATION

Web Appendices containing technical details and proofs, tables, and figures referenced in Sections 3, 5, 6, and 7 are available with this paper at the Biometrics website on the Wiley Online Library. In addition to this, the Fortran code implementing the simulation study and the case study are available at the Biometrics website on Wiley Online Library.

How to cite this article: Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L. Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. *Biometrics*. 2022;78:1454–1463. https://doi.org/10.1111/biom.13505