Supplementary materials for this article are available at https://doi.org/10.1007/s13253-024-00664-4.

Design-based Consistent Strategies Exploiting Auxiliary Information in Environmental Mapping

Lorenzo FATTORINI, Sara FRANCESCHI, and Caterina PISANI

Mapping continuous populations and finite populations of areas is approached in a design-based framework. The Tobler's law is adopted as assisting model, suggesting the use of inverse distance weighting (IDW) and k nearest neighbor (kNN) interpolators. The two interpolators are also able to exploit information provided by the huge list of inexpensive auxiliary data deriving from remote sensing sources. Conditions ensuring design-based consistency of kNN and IDW interpolators exploiting auxiliary information are derived under very simple tessellated schemes widely applied in environmental surveys. Simulation studies performed on a real population fully confirm the theoretical findings. Consistency results about kNN can be extended to random forest imputation techniques, that in the last years have been increasingly applied in mapping forest and environmental resources.

Supplementary materials accompanying this paper appear online.

Key Words: Design-based consistency; k nearest neighbor interpolator; Inverse distance weighting interpolator; Remote sensing information; Tessellated sampling schemes.

1. INTRODUCTION

Mapping is essential to understand the spatial pattern of environmental phenomena and resources, such as species composition and diversity measures in vegetation sciences and zoology or density of contaminating pollutants in ecology (e.g., Su et al. 2021; Choi and Chong 2022). Populations to be mapped can be of three types: continuous populations, i.e., the continuum of points constituting the survey region, finite collections of areas partitioning the survey region (e.g., administrative districts or pixels) and finite collections of units (e.g., plants, shrubs, or trees). Populations of units will not be considered since, in most cases, the list and locations of the units are not available and thus their mapping is precluded. Usually, in continuous populations and finite populations of areas the survey variable is recorded for a subset of locations/areas and an estimation criterion is adopted for obtaining maps

L. Fattorini · S. Franceschi (✉)· C.Pisani Department of Economics and Statistics, University of Siena, Siena, Italy (E-mail: *franceschi2@unisi.it*). C. Pisani NBFC, National Biodiversity Future Center, 90133 Palermo, Italy.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-024-00664-4

Published online: 25 November 2024

depicting the spatial pattern of the variable throughout the whole survey region (Cressie 1993). Often the survey variable is measured not at a single location, but in a plot of pre-fixed shape (e.g., Tomppo et al. 2010). In this case, the survey variable is generally the overall amount, or the density, of an attribute (e.g. dead wood, biomass) within the plot. As a plot can be centered at any point of the region, these centers constitute the continuum of points forming the survey region and thus techniques for mapping continuous populations can be adopted. Mapping is traditionally addressed in a model-based framework: locations/areas where the variable of interest is recorded are considered as fixed, while values of the variable are assumed to be outcomes of a superpopulation probability model (Cressie 1993). As an alternative to model-based approaches, Fattorini et al. (2018a, b) introduce a design-based approach to mapping in which the values of the variable are viewed as fixed constants while uncertainty is entailed by the probabilistic sampling scheme adopted for selecting locations/areas. The main attraction of a design-based approach is that inference is objective, because the properties of the resulting maps stem from the characteristics of the sampling scheme actually adopted in the field, without the need for the huge sequences of assumptions usually introduced in model-based approaches (e.g., geostatistical mapping). However, mapping in a design-based framework is challenging, because when estimating the value at a single location/area, either the location/area is sampled and there is no need for estimation, or it is unsampled so that we have no information for performing estimation. Thus, the use of an assisting model to estimate at unsampled locations/areas seems to be the sole way to fill the information vacancy. As to the assisting model to be used, Fattorini et al. (2018a, b) suggest using the well-known Tobler's first law of geography, i.e., locations/areas close in space tend to be more similar than those that are far apart (Tobler 1970). Therefore, based on this principle, estimation at unsampled locations/areas can be suitably performed exploiting the sample values observed at locations/areas that are neighbor (in some sense) to the location/area under estimation. The class of this kind of estimators is quite large and contains the inverse distance weighting (IDW) interpolators, the k nearest neighbor (kNN) interpolators and, subsequently, the random forest imputation (RFI) techniques, because, as pointed out by Lin and Jeon (2006), they can be viewed as adaptively weighted kNN methods. At the end of the past century, kNN techniques became increasingly popular especially in forest studies (Tomppo 1990; Tomppo and Katila 1991), probably owing to their simplicity and pliancy. Indeed, kNN interpolators are readily achieved by the linear combination of the k sample observations that are "nearest", by some distance criterion and in some space, to the location where interpolation is performed. Moreover, the choice of the number of neighbors k, which in turn determines map smoothness, and the choice of different distance metrics allow the implementation of several interpolators. After the two seminal papers by Tomppo, hundreds of applications of kNN stemmed from forest studies (see Chirici et al. 2016, for a review).

Obviously, kNN, RFI and IDW interpolators and the subsequent maps are destined to be design-biased and therefore the sole way to render statistically sound these design-based model-assisted mapping methods is to determine the conditions ensuring some sort of design-based consistency (Fattorini et al. 2018a, b). Indeed, consistency, though often overlooked in environmental applications, is crucial because, when it holds, the sampling distributions of the interpolators are expected to be tightly concentrated around the true

values, provided that the sample size and the population size (in the case of finite populations of areas) are sufficiently large (Särndal et al. 1992). Moreover, the increasing availability of a huge list of inexpensive auxiliary data deriving from remote sensing sources constitutes a "tremendous" opportunity to reduce costs and improve inference in environmental surveys (Opsomer et al. 2007). Thus, effective mapping strategies should be able to take advantage of this additional information. In this sense, the great appeal of kNN interpolators and related RFI techniques is the easy exploitation of auxiliary information by determining the nearest neighbors in the space of auxiliary variables, rather than in the geographic space. Among the plethora of articles listed in Chirici et al. (2016), Landsat imagery, airborne laser scanning metrics and digital aerial imagery are the most frequent source of auxiliary information and there is no mention of kNN directly performed on the geographic space or in spaces including spatial coordinates.

Recently, Di Biase et al. (2022) propose an alternative exploitation of auxiliary information estimating population values as linear functions of the auxiliary variables and then interpolating the regression errors in the geographical space by the IDW interpolator to be added to the regression estimates. Grafström and Tillé (2013) exploit the auxiliary information similarly, though not for the purpose of mapping but for sampling and estimating of population totals.

There is a vast body of literature on the asymptotic properties of estimators and classifiers based on kNN and random forests. Seminal works include those published by Stone (1977) and Devroye et al. (1994), as well as more recent papers, such as those by Hall et al. (2008), by Samworth (2012), by Gadat et al. (2016) and Györfi and Weiss (2021). However, in these studies the properties are derived in a model-based framework assuming a superpopulation probability model and auxiliary and sample data are considered as realizations of independent and identically distributed random vectors from that model. Thus the use of these results in a design-based framework is obviously precluded. The purpose of this paper is to derive conditions ensuring the design-based consistency of kNN and IDW interpolators, based on alternative exploitations of auxiliary information, and to evaluate their performance.

2. PRELIMINARIES AND NOTATION

Consider a study region $\mathcal{A}$ that is assumed to be a compact set of $\mathbb{R}^2$, denote by λ the Lebesgue measure on $\mathbb{R}^2$ and, for any $\mathbf{p}, \mathbf{q} \in \mathcal{A}$, by $d_g(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|$ the Euclidean distance on $\mathbb{R}^2$, where the suffix g evidences that the distance is referred to the geographic space. Moreover, denote by $I(E)$ the indicator of the event E. Let Y be the variable of interest and f be a measurable function defined on $\mathcal{A}$, with values in $[0, L]$ and related to Y in such a way that, for any borelian set $B \subset \mathcal{A}$, $\int_B f(\mathbf{p})\lambda(d\mathbf{p})$ gives the amount of Y in B. We consider two types of spatial populations giving rise to the following scenarios.

Continuous populations: f is the density of Y on $\mathcal{A}$. Thus, in principle, mapping necessitates the knowledge of $f(\mathbf{p})$ for each location $\mathbf{p} \in \mathcal{A}$.

Finite populations of areas: $\mathcal{A}$ is partitioned into N areas $a_1, \ldots, a_N$, y_j is the amount of Y within a_j and mapping requires knowledge of y_j for each $j = 1, \ldots, N$. Since the area size $\lambda(a_j)$ is usually known for each $j = 1, \ldots, N$, mapping equivalently requires

knowledge of the density of Y within areas, say $f_j = y_j/\lambda(a_j)$ for each $j = 1, \ldots, N$, that is the knowledge of the piecewise constant function $f(\mathbf{p}) = \sum_{j=1}^{N} f_j I(\mathbf{p} \in a_j)$ for each $\mathbf{p} \in \mathcal{A}$.

Let $\mathbf{P}_1, \ldots, \mathbf{P}_n$ be n random vectors with values in $\mathcal{A}$ that represent the n locations selected onto $\mathcal{A}$ by means of a probabilistic fixed-size sampling scheme. In the case of continuous populations, $\mathbf{P}_1, \ldots, \mathbf{P}_n$ denote n locations selected in the continuum $\mathcal{A}$ and $f(\mathbf{P}_1), \ldots, f(\mathbf{P}_n)$ are the densities recorded at those locations. In the case of finite populations of areas, $\mathbf{P}_1, \ldots, \mathbf{P}_n$ denote the centroids identifying the n sampled areas and $f(\mathbf{P}_1), \ldots, f(\mathbf{P}_n)$ are the densities recorded within these areas. While this notation will be exploited in a unifying perspective that includes both types of populations, for finite populations of areas it is standard to denote by S the set of the n selected areas and to refer to densities as f_j, $j = 1, \ldots, N$.

From a general point of view, $\forall \mathbf{p} \in \mathcal{A}$, an interpolator $\hat{f}(\mathbf{p})$ of $f(\mathbf{p})$ can be written as

$$\hat{f}(\mathbf{p}) = I(Q_p) f(\mathbf{p}) + I(Q_p^c) \sum_{i=1}^{n} w_i(\mathbf{p}) f(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A} \tag{1}$$

where $Q_p = \cup_{i=1}^{n} \{\mathbf{P}_i = \mathbf{p}\}$ is the event that $\mathbf{p}$ is a sampled location and $w_i(\mathbf{p})$s are suitable standardized weights summing to one, and thus depending, not only on the distance from $\mathbf{p}$ to $\mathbf{P}_i$, but also on the distances from $\mathbf{p}$ to all the other sampled locations/areas. In the continuous case, Q_p has probability 0, in such a way that (1) reduces almost surely to

$$\hat{f}(\mathbf{p}) = \sum_{i=1}^{n} w_i(\mathbf{p}) f(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A}. \tag{2}$$

In the case of finite populations of areas, $\hat{f}(\mathbf{p}) = \hat{f}_j$ for each $\mathbf{p} \in a_j$, and, denoting by $\mathbf{c}_1, \ldots, \mathbf{c}_N$ the centroids of $a_1, \ldots, a_N$, from (1)

$$\hat{f}_j = I(j \in S) f_j + (1 - I(j \in S)) \sum_{i \in S} w_i(\mathbf{c}_j) f_i, \ j = 1, \ldots, N \tag{3}$$

where $w_i(\mathbf{c}_j)$s now depend on the distances from the centroids of the sampled areas to $\mathbf{c}_j$.

Since f is measurable with values on $[0, L]$, interpolators of type (1) are random variables with values in $[0, L]$. Determining their finite-sample design-based properties is challenging because their design-based expectation and variance cannot be analytically obtained, which limits direct insights into their bias and precision. Consequently, it is necessary to derive conditions under which design-based asymptotic unbiasedness and consistency hold. As interpolators (1) are bounded, consistency also entails asymptotic unbiasedness. Thus, for brevity we only refer to design consistency.

The asymptotic scenarios introduced by Fattorini et al. (2018a, b) are considered to define design consistency of (1) (see section S5 of the Supplementary Materials for additional details). In the case of continuous populations, for any natural number m, a fixed-size design selecting a sample of n_m locations $\mathbf{P}_{m,1}, \ldots, \mathbf{P}_{m,n_m}$ on $\mathcal{A}$, with $n_m \to \infty$ as m increases is supposed. Therefore, from (2), for each $\mathbf{p} \in \mathcal{A}$, $\hat{f}_m(\mathbf{p})$ is the interpolator of

$f(\mathbf{p})$ under the mth design. In the case of finite populations of areas, for any natural number m, $\mathcal{A}$ is partitioned into an increasing number of N_m areas $a_{m,1}, \ldots, a_{m,N_m}$ with centroids $\mathbf{c}_{m,1}, \ldots, \mathbf{c}_{m,N_m}$, where $N_m \to \infty$ and all the areas decrease in size as m increases in such a way that $\sup_{j=1,\ldots,N_m} \text{diam}(a_{m,j}) \to 0$. Then a sequence of fixed-size designs is considered to select samples S_m of $n_m < N_m$ areas with $n_m \to \infty$. Therefore, for each $\mathbf{p} \in a_{m,j}$ and $j = 1, \ldots, N_m$, $\hat{f}_m(\mathbf{p})$ is the interpolator (3) of the piecewise constant function $f_m(\mathbf{p}) = \sum_{j=1}^{N_m} f_{m,j} I(\mathbf{p} \in a_{m,j})$ where $f_{m,j}$ is the density within $a_{m,j}$.

From these asymptotic scenarios, we can give a unique definition of design consistency, i.e., interpolators of type (1) are pointwise design consistent at $\mathbf{p} \in \mathcal{A}$ if for any arbitrary real value $\epsilon > 0$

$$\lim_{m \to \infty} \Pr\{|\hat{f}_m(\mathbf{p}) - f_m(\mathbf{p})| > \epsilon\} = 0 \tag{4}$$

where $f_m = f$ in the case of continuous populations.

3. DESIGN CONSISTENCY OF *K*NN INTERPOLATORS

For the first time, we here derive conditions for the design consistency of kNN, distinguishing the cases in which the k neighbors are determined in the geographic space, in the space of the auxiliary variables, and in the composite space formed by the auxiliary variables and the geographic coordinates. Owing to the analogy between kNN and RFI techniques, the achieved results should also hold for these techniques.

3.1. *k*NN DESIGN CONSISTENCY IN THE GEOGRAPHIC SPACE

In a unifying perspective, for a fixed k (usually much smaller than the sample size n), the kNN interpolator of $f(\mathbf{p})$ can be written from (1) as

$$\hat{f}_g(\mathbf{p}) = I(Q_p) f(\mathbf{p}) + I(Q_p^c) \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) f(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A} \tag{5}$$

where $d_{g(1)} < \ldots < d_{g(k)}$ denote the k smallest distances to $\mathbf{p}$ (in geographic space) occurred in the sample, $H_{g(l)}(\mathbf{p}) = \{\mathbf{P}_i : d_g(\mathbf{P}_i, \mathbf{p}) = d_{g(l)}\}$ is the set of sample locations with distances to $\mathbf{p}$ equal to the l-th ranked distance, henceforth referred to as the sample neighbors to $\mathbf{p}$ of order l, so that $H_{g,k}(\mathbf{p}) = \cup_{l=1}^{k} H_{g(l)}(\mathbf{p})$ is the sample neighbors to $\mathbf{p}$ of order less than or equal to k, i.e. the set of sample locations involved in estimation, and $w_{g,i}(\mathbf{p})$s are the weights attached to the f values corresponding to these locations, with $\sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) = 1$. The class of kNN interpolators (5) is large and contains, among others, the kNN interpolators with equal weights or weights proportional to the inverse of the squared distances to $\mathbf{p}$ and, for $k = 1$, the well-known nearest neighbor interpolator (Fattorini et al. 2022).

In the continuous case, $\text{card}\{H_{g(l)}(\mathbf{p})\} = 1$ for each $l = 1, \ldots, k$ almost surely, in such a way that from (2)

$$\hat{f}_g(\mathbf{p}) = \sum_{l=1}^{k} w_{g,l}(\mathbf{p}) f(\mathbf{P}_{g(l)}) \tag{6}$$

where $\mathbf{P}_{g(l)}$ denotes the sample location having distance $d_{g(l)}$ to $\mathbf{p}$. In the case of finite populations of areas, from (3) it follows that

$$\hat{f}_{g,j} = I(j \in S) f_j + (1 - I(j \in S)) \sum_{h \in H_{g,k.j}} w_{g,j,h} f_h, \; j = 1, \ldots, N \tag{7}$$

where $H_{g,k,j} = \{h : d_g(\mathbf{c}_h, \mathbf{c}_j) \leq d_{g(k)}\}$ is the set of sampled areas whose centroids have geographic distances to $\mathbf{c}_j$ less than or equal $d_{g(k)}$, and $w_{g,j,h}$s are the weights attached to these centroids, with $\sum_{h \in H_{g,k,j}} w_{g,j,h} = 1$. In the case of populations constituted by grids of regular polygons (e.g., pixels), sampled polygons having the same distance to the polygon under estimation may be more than one.

For finite-sample size n, we derive an upper bound for the expected absolute error of (5) that will be crucial in determining the asymptotic properties. For any arbitrary real value $\delta > 0$ and $\mathbf{p} \in \mathcal{A}$ denote by $B_g(\delta, \mathbf{p}) = \{\mathbf{q} : \mathbf{q} \in \mathcal{A}, d_g(\mathbf{p}, \mathbf{q}) \leq \delta\}$ the geographic δ-neighborhood of $\mathbf{p}$ in $\mathcal{A}$, and by $\Delta_g(\mathbf{p}, \delta) = \sup_{\mathbf{q} \in B_g(\delta, \mathbf{p})} |f(\mathbf{q}) - f(\mathbf{p})|$ the largest jump of f in $B_g(\delta, \mathbf{p})$. Moreover, denote by $Q_{g,i}(\mathbf{p}, \delta) = \{d_g(\mathbf{P}_i, \mathbf{p}) > \delta\}$ the event that the i-th sample location is outside the geographic δ-neighborhood of $\mathbf{p}$, in such a way that $Q_{g(l)}(\mathbf{p}, \delta) = \cap_{i \in H_{g(l)}(\mathbf{p})} Q_{g,i}(\mathbf{p}, \delta)$ is the event that the sample neighbors to $\mathbf{p}$ of order l are outside the geographic δ-neighborhood of $\mathbf{p}$ ($l = 1, \ldots, k$). For any finite $n > k$, the following inequality holds

$$\mathrm{E}[|\hat{f}_g(\mathbf{p}) - f(\mathbf{p})|] \leq \Delta_g(\mathbf{p}, \delta) + L \Pr\{Q_{g(k)}(\mathbf{p}, \delta)\} \tag{8}$$

where E denotes the design-based expectation, i.e., expectation with respect to the possible samples of size n that can be selected by the sampling design (see Section S1 of the Supplementary Materials for the proof).

By (8), the expected errors are bounded by the sum of two terms, the first depending on the roughness of f on the geographic space, the second on the sampling design by means of the probability that the sample neighbors to $\mathbf{p}$ of order less than or equal to k, i.e., those involved in the estimation, are outside the geographic δ-neighborhood of $\mathbf{p}$ ($l = 1, \ldots, k$). Therefore, a precise interpolation takes rise when both terms are small.

From the asymptotic scenarios described in section 2, the kNN interpolators (5) are pointwise design consistent at $\mathbf{p} \in \mathcal{A}$ if for any arbitrary real value $\epsilon > 0$

$$\lim_{m \to \infty} \Pr\{|\hat{f}_{g,m}(\mathbf{p}) - f_m(\mathbf{p})| > \epsilon\} = 0 \tag{9}$$

where $\hat{f}_{g,m}(\mathbf{p})$ is the kNN interpolator of type (6) or (7) achieved from the mth design. Then, design consistency of (5) at any continuity point of f_m straightforwardly follows from inequality (8). Taking $\delta_m = t n_m^{-1/2}$, the first term in the right side of (8) approaches 0 with δ_m owing to the continuity of f_m at $\mathbf{p}$, i.e., for any arbitrary real value $\epsilon > 0$ there exist a real value $t > 0$ and an integer m_0 such that

$$\Delta_g(\mathbf{p}, \delta_m) < \epsilon, \; m > m_0. \tag{10}$$

Therefore, consistency on the geographic space is achieved at any continuity point of f_m if the sequence of sampling designs also ensures that

$$\Pr\{Q_{g(k)}(\mathbf{p}, \delta_m)\} < \epsilon,\ m > m_0 \tag{11}$$

i.e. the probability that the sample locations/areas involved in the estimation are outside the geographic δ_m-neighborhood of $\mathbf{p}$ approaches 0 with δ_m.

In practice, consistency of kNN in the geographic space occurs at any continuity point of f_m when the sampling scheme ensures that the sample locations/areas involved in the estimation approach $\mathbf{p}$ as m increases. This feature, joined with the continuity of f_m at $\mathbf{p}$, ensures that the sample values involved in the estimation approach $f_m(\mathbf{p})$, so that the kNN estimator, being a convex combination of these values, also approaches $f_m(\mathbf{p})$.

Obviously, if f_m is discontinuous at $\mathbf{p}$, condition (10) does not hold and consistency in the geographic space is not ensured. Thus, when f_m exhibits many discontinuities, the precision of kNN estimators on $\mathcal{A}$ deteriorates. However, under suitable sampling schemes, the precision of the whole map is preserved if these discontinuities occur on sets of measure zero. Indeed, from (8) the mean integrated absolute error is bounded by

$$MIAE(\hat{f}_{g,m}) = \int_{\mathcal{A}} \mathrm{E}[|\hat{f}_{g,m}(\mathbf{p}) - f_m(\mathbf{p})|]\lambda(d\mathbf{p}) \leq \int_{\mathcal{A}} \Delta_g(\mathbf{p}, \delta_m)\lambda(d\mathbf{p}) + \int_{\mathcal{A}} \Pr\{Q_{g(k)}(\mathbf{p}, \delta_m)\}\lambda(d\mathbf{p}).$$

Therefore, if discontinuities occur on sets of measure 0, the first integral approaches 0 and the precision of the whole map strictly depends on the second integral, that in turn will be small if the sampling scheme is able to ensure an asymptotical spatial balance, i.e., to evenly spread sample locations in such a way that for any location $\mathbf{p} \in \mathcal{A}$, the sample locations involved in the estimation are near to $\mathbf{p}$. The assumption of continuity except sets of measure zero is reasonably valid in many natural scenarios where the density of an attribute changes smoothly throughout space (continuity) and when it changes abruptly, that occurs along borders delineating variations in the characteristics of the study region (e.g., forest-meadows). Therefore, borders may be realistically approximated by curves well approaching the theoretical condition of discontinuity over a region of measure zero.

Many schemes are available for selecting sample locations on the continuum of the geographic space ensuring spatial balance. If a regular tessellation of $\mathcal{A}$ into n regular polygons (e.g., quadrats, hexagons) can be performed, spatial balance can be simply achieved by systematic grid sampling (SGS), which consists of randomly selecting a location in the first polygon and systematically repeating it in the remaining ones. SGS is widely used in forest surveys (e.g., Opsomer et al. 2007; Tomppo et al. 2010), even though, when spatial regularities occur, its performance may be even worse than that of uniform random sampling (URS), i.e., the most straightforward scheme achieved by selecting sample locations independently and at random on the study region. More recently, tessellation stratified sampling (TSS) has become increasingly popular as a spatially balanced scheme. TSS does not necessitate partitions into regular polygons and does not suffer a loss of precision under spatial regularities. TSS consists of partitioning the study region into regular or irregular polygons of

equal area and then randomly and independently selecting a location in each polygon. URS, SGS and TSS satisfy (11) (see Section S2 of the Supplementary Materials) and thus, design consistency of kNN on the geographic space occurs at any continuity point of f.

Similarly, when sampling finite populations of areas, spatial balance can be ensured using very simple schemes. If a stratification into n regular blocks of contiguous areas is possible, systematic sampling (SYS) can be adopted, which consists of randomly selecting an area in the first block and systematically repeating it in the remaining blocks. SYS suffers from drawbacks similar to those of SGS, i.e., when spatial regularities occur, its performance may be even worse than that of simple random sampling without replacements (SRSWOR) (see e.g., Särndal et al. 1992). Alternatively, one-per-stratum stratified sampling (OPSS) does not necessitate partitioning into regular blocks and does not suffer a loss of precision under spatial regularities. OPSS consists of partitioning the population into blocks of contiguous areas, neither necessarily regular nor of the same size, and then randomly and independently selecting an area in each block. Both SYS and OPSS have long history in the statistical literature (e.g., Breidt 1995) and are proven to satisfy condition (11) (see Section S2 of the Supplementary Materials). Therefore, under these schemes, consistency of kNN interpolators on the geographic space occurs at any continuity point of f_m.

3.2. kNN DESIGN CONSISTENCY IN THE AUXILIARY VARIABLES SPACE

Even if we proved that, under suitable sampling schemes, the kNN interpolator performed on the geographic space is design consistent at any continuity point, it does not take advantage of the information provided by inexpensive auxiliary data. To exploit this information, myriads of applications perform kNN interpolation in the space of auxiliary variables (henceforth auxiliary space), i.e., for each unsampled location/area the values of the sampled locations/areas having the k smallest distances in the auxiliary space are used (see e.g., Chirici et al. 2016). In practice, it is presumed that locations/areas that are near in the auxiliary space tend to have similar values of the interest variable.

Denote by $\mathbf{x}(\mathbf{p})$ a vector of G auxiliary variables freely or cheaply available for each $\mathbf{p} \in \mathcal{A}$. In the case of finite populations of areas, $\mathbf{x}$ is a piecewise constant function taking values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ within $a_1, \ldots, a_N$, respectively. The geographic distances on $\mathcal{A}$ are replaced by the Euclidean distances in the auxiliary space $d_x(\mathbf{p}, \mathbf{q}) = ||\mathbf{x}(\mathbf{p}) - \mathbf{x}(\mathbf{q})||$, $\mathbf{p}, \mathbf{q} \in \mathcal{A}$. Accordingly, the kNN interpolator of $f(\mathbf{p})$ based on auxiliary data is

$$\hat{f}_x(\mathbf{p}) = I(Q_p) f(\mathbf{p}) + I(Q_p^c) \sum_{i \in H_{x,k}(\mathbf{p})} w_{x,i}(\mathbf{p}) f(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A} \tag{12}$$

where all the quantities involved in (12) are analogous to those involved in (5) with the distances computed in the auxiliary space.

Despite its large use, the d_x distance is not theoretically suitable for identifying the k nearest neighbors to be used in kNN interpolation. First of all, d_x is not a distance on $\mathcal{A}$ because it may occur that $d_x(\mathbf{p}, \mathbf{q}) = 0$ even when $\mathbf{p} \neq \mathbf{q}$. In addition, if we denote by $B_x(\delta, \mathbf{p}) = \{\mathbf{q} : \mathbf{q} \in \mathcal{A}, d_x(\mathbf{p}, \mathbf{q}) \leq \delta\}$ the δ-neighborhood of $\mathbf{x}(\mathbf{p})$ in the auxiliary space, and by $\Delta_x(\mathbf{p}, \delta) = \sup_{\mathbf{q} \in B_x(\delta, \mathbf{p})} |f(\mathbf{q}) - f(\mathbf{p})|$ the largest jump of f in $B_x(\delta, \mathbf{p})$, we cannot

claim that for any arbitrary real value $\epsilon > 0$ there exists a real value $t > 0$ and an integer m_0 such that, taking $\delta_m = t n_m^{-1/2}$

$$\Delta_x(\mathbf{p}, \delta_m) < \epsilon, \ m > m_0. \tag{13}$$

Indeed, even if δ_m approaches 0, i.e., the distances to $\mathbf{x}(\mathbf{p})$ become smaller and smaller, nothing ensures that the same occurs for $\Delta_x(\mathbf{p}, \delta_m)$. In practice, although a sampling scheme ensuring (11) is adopted, i.e., a scheme able to select neighboring locations to $\mathbf{p}$ in the geographic space, kNN interpolators of type (12) do not exploit these locations but those locations that are nearest to $\mathbf{x}(\mathbf{p})$ in the auxiliary space. This inconsistency between the scheme and the interpolators of type (12) precludes design consistency. In other words, because (13) does not hold, (9) does not hold when $\hat{f}_{g,m}(\mathbf{p})$ is replaced by $\hat{f}_{x,m}(\mathbf{p})$.

However, this problem disappears as the number G of auxiliary variables increases since $\hat{f}_x(\mathbf{p})$ tends to coincide with $\hat{f}_g(\mathbf{p})$. This feature can be heuristically explained by the fact that $f(\mathbf{p})$ can be naturally linked with the auxiliary variables $\mathbf{x}(\mathbf{p})$ by a linear function plus a term $e(\mathbf{p})$ that quantifies the error achieved by deterministically predicting $f(\mathbf{p})$ as a linear function of $\mathbf{x}(\mathbf{p})$. In practice, a very general scenario for $f(\mathbf{p})$ can be

$$f(\mathbf{p}) = \mathbf{b}^t \mathbf{x}(\mathbf{p}) + e(\mathbf{p}), \ \mathbf{p} \in \mathcal{A}. \tag{14}$$

It is worth noting that (14) is not an assumption but just an identity. Therefore, under the very general scenario depicted by (14), the computation of distances in the auxiliary space neglects a geographic component of f that may be relevant and may deteriorate the choice of nearest neighbors. Obviously, if G increases, the fitting of $\mathbf{b}^t \mathbf{x}(\mathbf{p})$ improves, so that the error terms $e(\mathbf{p})$ tend to vanish. Therefore, in this case consistency is achieved as $f(\mathbf{p})$ tends to coincide with a linear function of $\mathbf{x}(\mathbf{p})$. Indeed, since $\mathbf{x}(\mathbf{p})$ is continuous on $\mathcal{A}$ with respect to the pseudo-distance d_x (it is a distance only if $\mathbf{x}$ is one-to-one), when δ approaches 0 the f values tend to $f(\mathbf{p})$ thanks to (14) and the negligibility of error terms.

A more rigorous explanation of the coincidence of kNN interpolators in the geographic and auxiliary space when the number of auxiliary variable increases can be given. Indeed, as G increases, it is more likely that $\mathbf{x}$ becomes one-to-one onto $\mathcal{A}$ (see Section S3 of the Supplementary Materials), in such a way that d_x becomes a distance. Then, if $\mathbf{x}$ is differentiable (and then continuous) at $\mathbf{p}$, the two distances d_g and d_x are equivalent under appropriate conditions (see Section S4 of the Supplementary Materials).

3.3. kNN DESIGN CONSISTENCY IN THE COMPOSITE SPACE

A compromise solution to achieve consistency of kNN interpolators and, at the same time, to exploit the information provided by auxiliary variables, is to consider, for each $\mathbf{p} \in \mathcal{A}$, the $(G + 2)$-vector $\mathbf{z}(\mathbf{p}) = [\mathbf{p}, \mathbf{x}(\mathbf{p})]^T$ that joins the spatial coordinates with the G auxiliary variables. Then, distances on $\mathcal{A}$ are measured by the Euclidean distances in the $\mathbf{z}$-space, henceforth referred to as the composite space, i.e. $d_z(\mathbf{p}, \mathbf{q}) = ||\mathbf{z}(\mathbf{p}) - \mathbf{z}(\mathbf{q})||$, $\mathbf{p}, \mathbf{q} \in \mathcal{A}$. Accordingly, the kNN interpolator of $f(\mathbf{p})$ based on the composite space is given

by

$$\hat{f}_z(\mathbf{p}) = I(Q_p) f(\mathbf{p}) + I(Q_p^c) \sum_{i \in H_{z,k}(\mathbf{p})} w_{z,i}(\mathbf{p}) f(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A} \tag{15}$$

where all the quantities involved in (15) are analogous to those involved in (5) with the distances computed in the composite space. In practice, using d_z the auxiliary information is exploited in the sense that, if some locations are equidistant to $\mathbf{p}$ in the geographic space, then the nearest locations are those that are nearest in the auxiliary space. In finite populations of areas, this feature reduces the possibility that the cardinality of $H_{z,k}(\mathbf{p})$ is greater than k.

The use of d_z solves the issue of choosing an appropriate distance for kNN interpolators. Indeed, d_z is a distance on $\mathcal{A}$ because $d_z(\mathbf{p}, \mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$. In addition, if both f and $\mathbf{x}$ are continuous at $\mathbf{p}$ with respect to d_z then, if we denote by $B_z(\delta, \mathbf{p}) = \{\mathbf{q} : \mathbf{q} \in \mathcal{A}, d_z(\mathbf{p}, \mathbf{q}) \leq \delta\}$ the δ-neighborhood of $\mathbf{p}$ in the composite space, and by $\Delta_z(\mathbf{p}, \delta) = \sup_{\mathbf{q} \in B_z(\delta, \mathbf{p})} |f(\mathbf{q}) - f(\mathbf{p})|$ the largest jump of f in $B_z(\delta, \mathbf{p})$, $\Delta_z(\mathbf{p}, \delta)$ tends to 0 when δ approaches 0. Then, for $\delta_m = t n_m^{-1/2}$ and for any arbitrary real value $\epsilon > 0$ there exist a real value $t > 0$ and an integer m_0 such that

$$\Delta_z(\mathbf{p}, \delta_m) < \epsilon, \ m > m_0. \tag{16}$$

In practice, when a sampling scheme ensuring condition (11) is adopted, i.e., the scheme is able to select neighboring locations to $\mathbf{p}$ in the geographic space, the kNN interpolators of type (15) not only exploit these locations but, among them, choose those that are nearest to $\mathbf{p}$ in the auxiliary space. Therefore, condition (11) joined with (16) ensures the consistency of the interpolator $f_z(\mathbf{p})$, i.e., (9) holds when $\hat{f}_{g,m}(\mathbf{p})$ is replaced by $\hat{f}_{z,m}(\mathbf{p})$.

4. DESIGN CONSISTENCY OF IDW INTERPOLATORS OF REGRESSION ERRORS

Quoting from Di Biase et al. (2022), the set of auxiliary variables $\mathbf{x}(\mathbf{p})$ freely or cheaply available for each $\mathbf{p} \in \mathcal{A}$ can be used to construct a proxy for $f(\mathbf{p})$ and Särndal et al. (1992) suggest the use of linear functions of the auxiliary variables, even thought other options can be considered (e.g., Breidt and Opsomer 2017 and references therein). Adopting the assisting model in Särndal et al. (1992, section 6.4) to build a proxy for $f(\mathbf{p})$, an effective choice for the vector of the coefficients of the auxiliary variables is the least-square vector

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{a} = \left(\int_{\mathcal{A}} \mathbf{x}(\mathbf{p}) \mathbf{x}^t(\mathbf{p}) \lambda(d\mathbf{p}) \right)^{-1} \int_{\mathcal{A}} f(\mathbf{p}) \mathbf{x}(\mathbf{p}) \lambda(d\mathbf{p})$$

in such a way that $f(\mathbf{p})$ can be expressed as in (14).

If the vector $\mathbf{b}$ would be known, the residuals for each sampled location $\mathbf{P}_1, \ldots, \mathbf{P}_n$ were known and residuals at non sampled locations could be interpolated by means of the IDW technique, i.e., adopting a unifying notation

$$\hat{e}_\alpha(\mathbf{p}) = I(Q_p) e(\mathbf{p}) + I(Q_p^c) \sum_{i=1}^{n} w_i(\alpha, \mathbf{p}) e(\mathbf{P}_i), \ \mathbf{p} \in \mathcal{A} \tag{17}$$

where, for a selected real value $\alpha > 2$, $w_i(\alpha, \mathbf{p}) = ||c(\mathbf{p}) - \mathbf{P}_i||^{-\alpha} / \sum_{h=1}^{n} ||c(\mathbf{p}) - \mathbf{P}_h||^{-\alpha}$ and c is the identity function for continuous populations while $c(\mathbf{p})$ is the nearest centroid to $\mathbf{p}$ for finite populations of areas (Fattorini et al. 2018a, b). Accordingly, the IDW interpolator of $f(\mathbf{p})$ achieved exploiting auxiliary information would be given by

$$\hat{f}_\alpha(\mathbf{p}, \mathbf{b}) = \mathbf{b}^t \mathbf{x}(\mathbf{p}) + \hat{e}_\alpha(\mathbf{p}) \tag{18}$$

which is a genuine interpolator, as $\hat{f}_\alpha(\mathbf{P}_i, \mathbf{b}) = f(\mathbf{P}_i)$ for each $i = 1, \ldots, n$.

Because $\mathbf{b}^t \mathbf{x}(\mathbf{p})$ is constant with respect to sampling, the design-based uncertainty of (18) would only depend on the uncertainty arising from the IDW interpolation of errors by means of (17). Therefore, in this framework, the design consistency of (18) immediately follows from the design consistency of the IDW interpolator under TSS, SGS and URS for continuous populations (Fattorini et al. 2018a) and under OPSS and SYS for finite populations of areas (Fattorini et al. 2018b) under the same asymptotic scenarios introduced in section 2. Moreover, if $\mathbf{b}^t \mathbf{x}(\mathbf{p})$ offers reliable predictions, it should mirror any irregularities in f resulting in smoother residuals. This makes IDW interpolation of the residuals better meet the requirement of continuity, thereby providing considerable gains in precision.

Because the least-square vector $\mathbf{b}$ is unknown, involving the knowledge of $f(\mathbf{p})$ for each $\mathbf{p} \in \mathcal{A}$, Di Biase et al. (2022) propose to estimate $\mathbf{b}$ as function of the sample estimators of $\mathbf{a}$ and $\mathbf{A}$, i.e. $\hat{\mathbf{b}} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{a}}$, where $\hat{\mathbf{A}}$ and $\hat{\mathbf{a}}$ are achieved by the continuous extensions of the Horvitz-Thompson (HT) estimator (Cordy 1993) or by the Monte Carlo estimator (Barabesi and Marcheselli 2005) in the continuous case and by the HT estimator in the case of finite populations of areas (Brus 2000). Therefore, the definitive interpolator turns out to be

$$\hat{f}_\alpha(\mathbf{p}, \hat{\mathbf{b}}) = \hat{\mathbf{b}}^t \mathbf{x}(\mathbf{p}) + \hat{e}_\alpha(\mathbf{p}, \hat{\mathbf{b}}), \ \mathbf{p} \in \mathcal{A}. \tag{19}$$

Fattorini et al. (2020) prove the design consistency of the Monte Carlo estimator under URS, TSS and SGS and of the HT estimator under OPSS and SYS under the same asymptotic scenarios adopted in this paper. Then, based on these results, $\hat{\mathbf{b}}$ converges to $\mathbf{b}$, in such a way that $\hat{f}_\alpha(\mathbf{p}, \hat{\mathbf{b}})$ converges to $\hat{f}_\alpha(\mathbf{p}, \mathbf{b})$. That proves the design consistency of (19).

Regarding the choice of α, that impacts interpolation as a smoothing parameter, Di Biase et al. (2022) suggest a leave-one out procedure. Because consistency of (19) holds for each real value $\alpha > 2$, it should also hold for the leave-one-out choice (Fattorini et al. 2023). At to the choice of the auxiliary variables to be used, usually there is a large set available and Di Biase et al. (2022) adopt the rule by Burman and Nolan (1995) that select the linear model with the best predictive capability for new independent observations, ensuring the model always includes an intercept by incorporating a constant variable equal to 1.

5. SIMULATION STUDY

For both types of spatial populations, a simulation study is performed to empirically investigate the design consistency of kNN, RFI and IDW interpolators.

5.1. POPULATIONS AND SAMPLING

The Harvard Forest is a 35ha rectangular region located at Petersham (New England). This region has been completely censused for the first time from 2010 to 2014. More precisely, all the living trees were enumerated, and, among several features, their spatial coordinates and above ground biomass were recorded. Data are available for download from the Harvard Forest Data Archive (2014).

In order to construct the continuous population and the finite populations of areas for implementing the simulation study, we consider a portion of 27ha of the Harvard forest where both Landasat spectral bands values and above ground biomass are available.

In particular, spectral information of Landsat8 l2sp (Level 2 Science Products) bands (called here Band 1-Band 7) are freely downloadable from USGS LandsatLook for year 2014 (data acquired in august 2023) for pixels of side 30m. For each pixel, the density of above ground biomass per ha (AGB) is derived and correlation coefficients between AGB and Landsat spectral bands are reported in Table 1 of the Supplementary Materials.

As to the continuous population, the AGB values to be mapped and the 7 spectral bands values, subsequently adopted as auxiliary information, are artificially achieved by means of ordinary kriging prediction performed on AGB, Band 1, Band 2, Band 3, Band 4, Band 5, Band 6, Band 7, respectively, in correspondence of the pixels centroids. Spectral bands values are rescaled between 0 and 1, while geographical coordinates are rescaled in such a way that horizontal coordinates range between 0 and 1 (see Figure 6 of the Supplementary Materials). As to the sampling scheme, TSS is considered, owing to its frequent implementation in natural resources surveys and its comparable performances with respect to more sophisticated sampling schemes (Di Biase et al. 2024). Sampling is performed selecting $n = 100, 200, 300, 400, 500$ locations by means of TSS partitioning the study area in into 10×10, 10×20, 12×25, 16×25, 20×25 grids of equal-sized rectangles and by independently selecting a location in each rectangle.

Regarding finite populations of areas, six populations of $N = 432, 972, 1728, 2700, 3888$ areas are constructed by partitioning the study region into grids of quadrats. AGB values for each quadrat are obtained by suitably integrating the AGB values of the continuous population. The same procedure is adopted to obtain the auxiliary variables values from the seven Landsat spectral bands. Figure 7 of the Supplementary Materials depicts AGB and spectral bands values for the finite population of 432 areas. Sampling is performed by selecting $n \approx 0.08N$ quadrats by means of OPSS by partitioning the grids into blocks of 3×4 quadrats and selecting one quadrat per block.

5.2. SIMULATION

For each combination of population and sample size, sampling is replicated 10, 000 times. At each simulation run, kNN, RFI and IDW interpolators are performed. For the continuous population, interpolations are achieved on a regular grid of 7500 locations in the study area. The Euclidean norm is used to define the set of neighbors, and, to reduce computation burden, the tuning parameters required for interpolations have been set, even though in real applications they are commonly selected by means of data-driven procedures. In particular

for kNN interpolation, $k = 4$ neighbors are considered, and constant weights are adopted. As to the RFI technique, the default parameters of the function randomForest of the R-package `randomForest` (R Core Team 2021) are considered, while when performing IDW interpolation, the smoothing parameter is fixed to $\alpha = 3$.

Distances necessary for performing kNN interpolation and to implement RFI are calculated in the auxiliary space first considering the auxiliary variable with the largest correlation coefficient with the interest variable, then adding the auxiliary variable with the largest correlation coefficient among the remaining ones and so on. Similarly, as to the composite space, distances for performing kNN and implementing RFI technique are calculated considering the spatial coordinates plus the auxiliary variable with the largest correlation coefficient with the interest variable, then adding the auxiliary variable with the largest correlation coefficient among the remaining ones and so on. Also in the case of the IDW technique, the auxiliary variables are sequentially included in the interpolator following the same criterion. For each location/area where interpolation is performed, the root mean squared error of kNN, RFI and IDW interpolators is computed from the Monte Carlo distribution of the corresponding estimates. Finally, as a measure of global precision of the estimated maps, the averages of the root mean squared errors (ARMSE) are derived.

5.3. RESULTS

Simulation results fully confirm the theoretical findings. ARMSEs values for the continuous population and for the populations of areas are reported in Tables 2 and 3 of the Supplementary Materials and depicted in Fig. 1 and 2.

Specifically, as to the performances of the kNN and RFI interpolators, for both continuous and finite populations, ARMSEs remarkably decrease when the sample size increases and the geographic space is considered, while, in the case of the auxiliary space, slight decreases occur only when at least the three more correlated auxiliary variables are jointly considered. When distances are computed in the composite space, the performance of the kNN and RFI interpolators improves abruptly in comparison to the performance achieved in the corresponding auxiliary space, for any sample size. Additionally, as to the IDW interpolator performance, simulation results fully confirm the theoretical findings on design consistency as, for both continuous population and finite populations of areas, sharp decreases in the ARMSE values occur as the sample size increases.

Finally, simulation results show noticeable increases in the precision of the IDW interpolator compared to the kNN and RFI interpolators, even if all those tuning procedures commonly adopted in real case studies (e.g. for setting the number of neighbors in kNN interpolation, the parameters involved in RFI and the smoothing parameter for IDW interpolation) have not been considered since the primary focus of this work is not on comparing the performance of these interpolators but rather on determining their consistency conditions.

6. CONCLUDING REMARKS

When mapping environmental resources from a design-based perspective, the design-based finite-sample properties of the most widely applied mapping techniques (IDW, kNN,

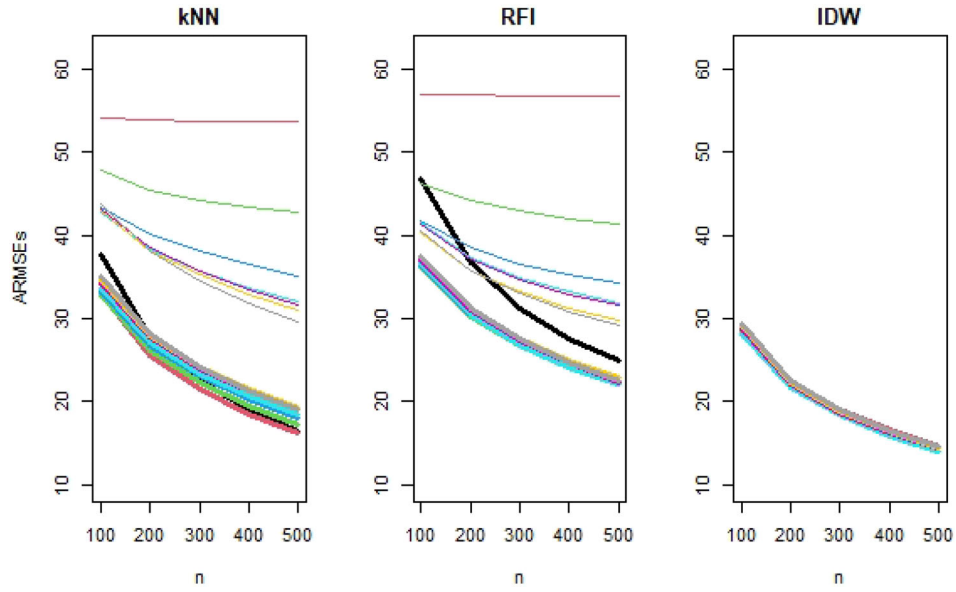

Figure 1. Values of ARMSE of the *k*NN, RFI and IDW interpolators for the continuous population for n = 100, 200, 300, 400, 500 (horizontal axis) considering the geographic space (thick black line), the auxiliary space composed by the first (thin red lines), the two (thin green lines), the three (thin blue lines), the four (thin sky blue lines), the five (thin purple lines), the six (thin yellow lines) auxiliary variables more correlated with AGB and all the seven auxiliary variables (thin gray lines). Averages obtained under the corresponding composite spaces are, respectively, depicted by red, green, blue, sky blue, purple, yellow and gray thick lines (Color figure online).

RFI) are not clearly delineated. Little is known about the design-based bias and variances of the mapped values and how bias and variances are related to the sampling effort. However, if design consistency holds and the sample size and the population size (in the case of finite populations of areas) are large, then the estimated maps can be considered good pictures of the true ones, i.e. the sampling distributions of the estimators of the population values at any point of the study region are tightly concentrated around the true values. On the other hand, little can be said about the resulting maps in absence of design consistency. That is the main reason for which we have considered the design consistency of the most common mapping strategies, showing the crucial role of the sampling schemes adopted to select locations/areas. In particular, we have proven that the tessellated schemes which are widely applied in environmental and forest surveys and straighforwardly achieve spatial balance, i.e., TSS and SGS for continuous populations and OPSS and SYS for finite populations of areas, ensure design consistency and asymptotic unbiasedness of *k*NN interpolators based on composite spaces as well as of IDW interpolators of regression errors. Neverthless, while computing distances using auxiliary space is a common practice (see e.g. Tomppo 1990; Tomppo and Katila 1991), we have proven that the resulting *k*NN interpolators may still lack asymptotic unbiasedness and consistency, thus highlighting the necessity of incorporating geographic coordinates into the distance calculations.

Interestingly, as confirmed by the simulation study, results about *k*NN can be extended to RFI techniques, that in the last years have been widely applied as alternatives to *k*NN (e.g., Chirici et al. 2020; Sun et al. 2020).

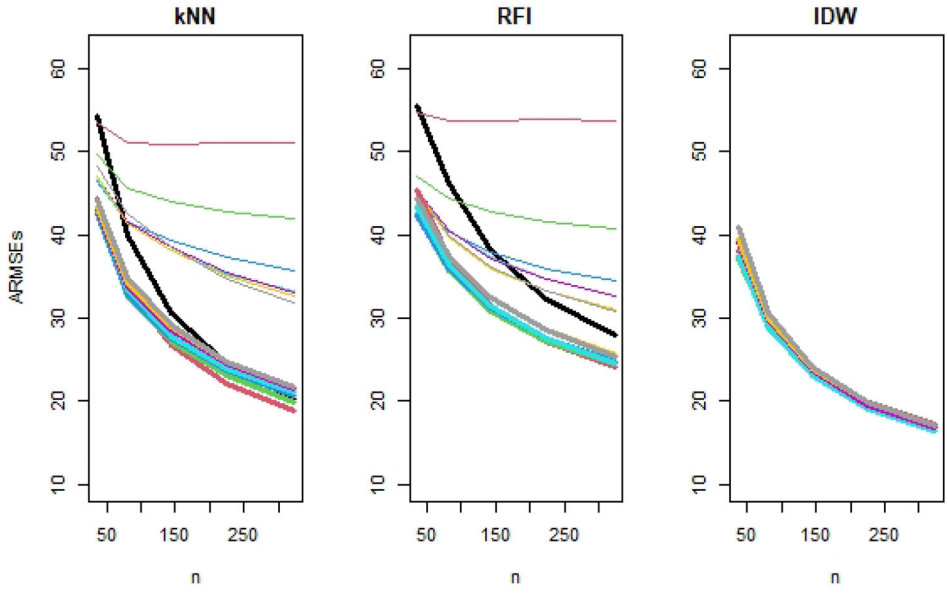

Figure 2. Values of ARMSE of the *k*NN, RFI and IDW interpolators for populations of areas for n = 36, 81, 144, 225, 324 (horizontal axis) considering the geographic space (thick black line), the auxiliary space composed by the first (thin red lines), the two (thin green lines), the three (thin blue lines), the four (thin sky blue lines), the five (thin purple lines), the six (thin yellow lines) auxiliary variables more correlated with AGB and all the seven auxiliary variables (thin gray lines). Averages obtained under the corresponding composite spaces are, respectively, depicted by red, green, blue, sky blue, purple, yellow and gray thick lines (Color figure online).

Finally, we have to point out that spatial balance can be obtained not only by the very simple tessellated schemes, but also by a plethora of more complex, explicitly tailored schemes (e.g. Stevens and Olsen 2004; Grafström et al. 2012; Grafström 2012; Grafström and Tillé 2013). Owing to their capacity in providing spatial balance, design consistency can be probably achieved under these schemes. Notwithstanding this, in this paper we have deliberately neglected them because, owing to their greater complexity with respect to the very simple tessellated schemes, they are not well understood by naturalists and rarely implemented in real surveys.

Funding 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"; National Recovery and Resilience Plan(NRRP); European Union-NextGenerationEU-Statistics for vegetation biodiversity: estimation and mapping (SVeBio), Grant/Award Number:P2022AW4LX-CUP B53D23029510001.

Declarations

Conflict of interest The authors declare no conflict of interest.

[Received May 2024. Revised September 2024. Accepted October 2024.]

REFERENCES

Barabesi L, Marcheselli M (2005) Monte Carlo integration strategies for design-based regression estimators of the spatial mean. Environmetrics 16(8):803–817. https://doi.org/10.1002/env.735

Breidt FJ (1995) Markov chain designs for one-per-stratum spatial sampling. Surv Methodol 21(1):63–70

Breidt FJ, Opsomer JD (2017) Model-assisted survey estimation with modern prediction techniques. Stat Sci 32(2):190–205. https://doi.org/10.1214/16-STS589

Brus D (2000) Using regression models in design-based estimation of spatial means of soil properties. Eur J Soil Sci 51(1):159–172. https://doi.org/10.1046/j.1365-2389.2000.00277.x

Burman P, Nolan D (1995) A general Akaike-type criterion for model selection in robust regression. Biometrika 82(4):877–886. https://doi.org/10.1093/biomet/82.4.877

Chirici G, Mura M, McInerney D et al (2016) A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data. Remote Sens Environ 176:282–294. https://doi.org/10.1016/j.jag.2019.101959

Chirici G, Giannetti F, McRoberts RE et al (2020) Wall-to-wall spatial prediction of growing stock volume based on Italian national forest inventory plots and remotely sensed data. Int J Appl Earth Obs Geoinf 84:101959. https://doi.org/10.1016/j.jag.2019.101959

Choi K, Chong K (2022) Modified inverse distance weighting interpolation for particulate matter estimation and mapping. Atmosphere 13(5):846. https://doi.org/10.3390/atmos13050846

Cordy CB (1993) An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Stat Probab Lett 18(5):353–362. https://doi.org/10.1016/0167-7152(93)90028-H

Cressie N (1993) Statistics for spatial data. Wiley, New York

Devroye L, Gyorfi L, Krzyzak A et al (1994) On the strong universal consistency of nearest neighbor regression function estimates. Ann Stat 22(3):1371–1385. https://doi.org/10.1214/aos/1176325633

Di Biase RM, Fattorini L, Franceschi S et al (2022) From model selection to maps: a completely design-based data-driven inference for mapping forest resources. Environmetrics 33(7):e2750. https://doi.org/10.1002/env.2750

Di Biase RM, Marcheselli M, Pisani C (2024) Achieving spatial balance without tears in environmental and ecological surveys: the tessellation sampling schemes. Environmetrics. https://doi.org/10.1002/env.2869

Fattorini L, Marcheselli M, Pisani C et al (2018a) Design-based maps for continuous spatial populations. Biometrika 105(2):419–429. https://doi.org/10.1093/biomet/asy012

Fattorini L, Marcheselli M, Pratelli L (2018b) Design-based maps for finite populations of spatial units. J Am Stat Assoc 113(522):686–697. https://doi.org/10.1080/01621459.2016.1278174

Fattorini L, Marcheselli M, Pisani C et al (2020) Design-based consistency of the Horvitz-Thompson estimator under spatial sampling with applications to environmental surveys. Spat Stat 35:100404. https://doi.org/10.1016/j.spasta.2019.100404

Fattorini L, Marcheselli M, Pisani C et al (2022) Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. Biometrics 78(4):1454–1463. https://doi.org/10.1111/biom.13505

Fattorini L, Franceschi S, Marcheselli M et al (2023) Design-based spatial interpolation with data driven selection of the smoothing parameter. Environ Ecol Stat 30(1):103–129. https://doi.org/10.1007/s10651-023-00555-w

Gadat S, Klein T, Marteau C (2016) Classification in general finite dimensional spaces with the k-nearest neighbor rule. Ann Stat 44(3):982–1009. https://doi.org/10.1214/15-AOS1395

Grafström A (2012) Spatially correlated Poisson sampling. J Stat Plan Inference 142(1):139–147. https://doi.org/10.1016/j.jspi.2011.07.003

Grafström A, Tillé Y (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24(2):120–131. https://doi.org/10.1002/env.2194

Grafström A, Lundström NL, Schelin L (2012) Spatially balanced sampling through the pivotal method. Biometrics 68(2):514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Györfi L, Weiss R (2021) Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. J Mach Learn Res 22(151):1–25

Hall P, Park BU, Samworth RJ (2008) Choice of neighbor order in nearest-neighbor classification. Ann Stat 36(5):2135–2152. https://doi.org/10.1214/07-AOS537

Harvard Forest Data Archive (2014). https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF253

Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. J Am Stat Assoc 101(474):578–590. https://doi.org/10.1198/016214505000001230

Opsomer JD, Breidt FJ, Moisen GG et al (2007) Model-assisted estimation of forest resources with generalized additive models. J Am Stat Assoc 102(478):400–409. https://doi.org/10.1198/016214506000001491

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna

Samworth RJ (2012) Optimal weighted nearest neighbour classifiers. Ann Stat 40(5):2733–2763. https://doi.org/10.1214/12-AOS1049

Särndal C-E, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer Science & Business Media

Stevens DL, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99(465):262–278. https://doi.org/10.1198/016214504000000250

Stone CJ (1977) Consistent nonparametric regression. Ann Stat 5(4):595–620. https://doi.org/10.1214/aos/1176343886

Su H, Bista M, Li M (2021) Mapping habitat suitability for Asiatic black bear and red panda in Makalu Barun national park of Nepal from Maxent and GARP models. Sci Rep 11(1):14135. https://doi.org/10.1038/s41598-021-93540-x

Sun D, Wen H, Wang D et al (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362:107201. https://doi.org/10.1016/j.geomorph.2020.107201

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(sup1):234–240. https://doi.org/10.2307/143141

Tomppo E (1990) Designing a satellite image-aided national forest survey in Finland [NFI]. Rapport-Sveriges Lantbruksuniversitet, Institutionen foer Biometri och Skogsindelning, Avdelningen foer Skoglig Fjaerranalys (Sweden)

Tomppo E, Katila M (1991) Satellite image-based national forest inventory of Finland. ISPRS J Photogramm 28(7–1):419–424

Tomppo E, Gschwantner T, Lawrence M et al (2010) National forest inventories. Pathways Common Report Eur Sci Found 1:541–553

USGS LandsatLook (2014) https://landsatlook.usgs.gov/