Environmental and Ecological Statistics (2023) 30:103–129
https://doi.org/10.1007/s10651-023-00555-w

Design-based spatial interpolation with data driven selection of the smoothing parameter

Lorenzo Fattorini[1] · Sara Franceschi[1] · Marzia Marcheselli[1,3] · Caterina Pisani[1,3] · Luca Pratelli[2]

Received: 6 July 2022 / Accepted: 7 January 2023 / Published online: 16 February 2023
© The Author(s) 2023

Abstract

In the inverse distance weighting interpolation the interpolated, value is a weighted mean of the sampled values, with weights decreasing with the distances. The most widely adopted class of distance functions is the class of negative powers of order α and the appropriate choice of the smoothing parameter α is a crucial issue. In this paper, we give sufficient conditions for the design-based consistency of the inverse distance weighting interpolator when α is selected by cross-validation techniques, and a pseudo-population bootstrap approach is introduced to estimate the accuracy of the resulting interpolator. A simulation study is performed to empirically confirm the theoritical findings and to investigate the finite-sample properties of the interpolator obtained using leave-one-out cross-validation. Moreover, a comparison with the nearest neighbor interpolator, which is the limiting case for $\alpha = \infty$, is performed. Finally, the estimation of the surface of the Shannon diversity index of tree diameter at breast height in the experimental watershed of Bonis forest (Southern Italy) is described.

✉ Sara Franceschi
sara.franceschi@unisi.it

Lorenzo Fattorini
lorenzo.fattorini@unisi.it

Marzia Marcheselli
marzia.marcheselli@unisi.it

Caterina Pisani
caterina.pisani@unisi.it

Luca Pratelli
luca_pratelli@marina.difesa.it

1 Department of Economics and Statistics, University of Siena, P.zza San Francesco 7, 53100 Siena, Italy

2 Naval Academy, Viale Italia 72, 57127 Livorno, Italy

3 NBFC, National Biodiversity Future Center, 90133 Palermo, Italy

Keywords Inverse distance weighting interpolator · Pointwise and uniform consistency · Pseudo-population bootstrap · Spatial populations

1 Introduction

Successful management of natural, social and economic resources requires detailed information about their spatial pattern. For example, mapping soil composition and mineral concentration is essential in geology, and pollutants concentration in ecology, while climatologists are interested in mapping atmospheric variables, such as temperature, humidity, and precipitation. In these cases, the study region is constituted by a continuous set of locations, conceptualized as a continuous spatial population, with the density of the survey variable at each location giving rise to a surface. Occasionally, the study area is partitioned into a finite population of areas, such as a network of regular polygons, as frequently happens in forest inventories, or into a collection of irregular patches, such as administrative districts. The survey variable is the total amount of an attribute within each area, such as the tree biomass in forestry or the volume of a specific agricultural production in economics. Finally, finite populations of units scattered over the study region, such as the factories in a district, the shrubs in a natural reserve, may be of interest. In this case, the survey variable is the value of an attribute attached to each unit.

Spatial prediction enables estimating the value of the survey variable or of its density at unsampled locations on the basis of a sample of locations, thus allowing to construct wall-to-wall maps depicting the spatial pattern of the survey variable throughout the whole study area. The inverse distance weighting (IDW) interpolation is a technique extensively applied by practitioners, also owing to the availability of GIS tools automatically implementing the interpolation. Commonly, IDW interpolation is considered as a non-stochastic method of spatial prediction, and, as such, no uncertainty is associated (Cressie 1993, Sect. 5.9). In the IDW interpolation, according to the first law of geography by Tobler (1970), values recorded at sampled locations do not contribute equally, since the interpolated value is achieved as a weighted mean of the observed values, with weights decreasing with the distances to the location where interpolation has to be performed. Any positive decreasing function of the distances obviously allows giving less weight to observed values further away from the location. In particular, the most widely adopted class of functions is the class of negative powers distance functions of order α, $\phi(d) = d^{-\alpha}$ (see e.g., Gong et al. 2014; Noori et al. 2014; Bărbulescu et al. 2021), where d is a positive real number representing the distance and α is a positive real number playing the role of the smoothing parameter. Therefore the appropriate choice of the smoothing parameter becomes a crucial issue. A default value of 2 is commonly adopted when GIS software are used, nevertheless, it can also be selected by means of cross-validation techniques, such as the leave-one-out cross-validation (LOOCV) (see e.g., Hall and Robinson 2009; Wu et al. 2019).

Recently, the IDW interpolator has been approached under continuous populations (Fattorini et al. 2018a), finite population of areas (Fattorini et al. 2018b) and finite populations of units (Fattorini et al. 2019) in a design-based approach. In this frame-

work, the uncertainty, which is associated to the interpolated values, only stems from the probabilistic sampling scheme adopted to select locations.

Conditions ensuring design-based consistency of the IDW interpolator with negative power distance functions have been proven to hold for any fixed finite $\alpha > 2$ and also for $\alpha = \infty$, which leads to the well-known nearest neighbour (NN) interpolator (Fattorini et al. 2018a, b, 2019, 2021).

The purpose of this paper is to derive sufficient conditions to prove the design-based asymptotic properties of the IDW interpolator when α is selected according to LOOCV. In particular, asymptotic results are achieved in a unifying approach that includes the three types of spatial populations, thus rendering the theoretical developments less burdensome. The finite-sample performance of the corresponding IDW interpolator is empirically compared to that of the NN interpolator. Indeed the latter avoids the computational effort needed for implementing LOOCV, which can be very time-consuming with large study areas where interpolation must be performed for thousands of points, areas or units. Furthermore, a pseudo-population bootstrap approach is introduced to obtain an estimator of the accuracy of the corresponding data driven IDW interpolator. The paper is organized as follows. Notation and setting are given in Sect. 2. In Sect. 3 the IDW interpolator is introduced in a unifying approach for the three types of spatial populations. Section 4 is devoted to the choice of the smoothing parameter by means of LOOCV and to the design-based consistency of the corresponding data driven IDW interpolator. In Sect. 5 a pseudo-population bootstrap estimator of the precision of the data driven IDW interpolator is proposed. A simulation and a case study are respectively described in Sects. 6 and 7, while concluding remarks are reported in Sect. 8. The Appendix contains technical details and proofs. Supplementary Information contains figures referring to the simulation study.

2 Notation and setting

Consider a study region A that is assumed to be a compact set of $\mathbb{R}^2$ and denote by λ the Lebesgue measure on $\mathbb{R}^2$. Interest is in the estimation of the value or density of a survey variable Y on a subset B of A, where B can be a continuum of points, a finite population of areas or a finite population of units. In order to deal with the three types of spatial populations in a unifying approach, consider a function f related to the Y-values, which, without loss of generality, is supposed to be a bounded and measurable function with values on $[0, L]$, with $L \in \mathbb{R}$ and $L < \infty$.

In the case of continuous populations, B coincides with A and $f(p)$ is the density of Y at any point $p \in A$. When finite populations of areas are considered, B coincides with A and is partitioned into N areas $a_1, \ldots, a_N$. In this framework, y_j and $f_j = y_j/\lambda(a_j)$ are the amount and the density of Y within a_j, respectively and f turns out to be

$$f(p) = \sum_{j=1}^{N} f_j I(p \in a_j)$$

for each $p \in B$, where $I(E)$ is the indicator function of the event E. Since the area size $\lambda(a_j)$ is usually known for each $j = 1, \ldots, N$, the knowledge of the piecewise constant surface $f(p)$ is equivalent to the knowledge of $f_1, \ldots, f_N$. Finally, when finite populations of units are considered, B is the set $\{p_1, \ldots, p_N\}$ of N unit locations and $y_j = f(p_j)$ is the value of the survey variable for the unit j.

Let $P_1, \ldots, P_n$ be n random variables with values in B that represent the n locations selected from B by means of a probabilistic sampling scheme. In the case of continuous populations, $P_1, \ldots, P_n$ denote n locations selected in the continuum B and $f(P_1), \ldots, f(P_n)$ are the densities of Y recorded at those locations. In the case of finite populations of areas, $P_1, \ldots, P_n$ denote the centroids identifying the n sampled areas and $f(P_1), \ldots, f(P_n)$ are the densities recorded within the corresponding areas. Finally, in the case of finite populations of units, $P_1, \ldots, P_n$ denote the locations of n sampled units and $f(P_1), \ldots, f(P_n)$ are the values of Y for these units. In all these three settings, the goal is the estimation of $f(p)$ for any $p \in B$ on the basis of the values of f recorded at the sampled locations $P_1, \ldots, P_n$.

3 The IDW interpolator under negative power distance functions

When mapping is considered in a design-based approach and, consequently, uncertainty only stems from the adopted sampling scheme, the use of an assisting model is necessary. Indeed, when estimating $f(p)$ at a single location $p \in B$, either p is sampled and there is no need for estimation, or p is unsampled, so that no information about it is available for performing estimation. The very simple Tobler's first law of geography, i.e. units that are close in space tend to have more similar properties than units that are far apart (Tobler 1970), can be exploited as assisting model. To this end, let $Q_p = \bigcup_{i=1}^{n} \{P_i = c(p)\}$, where $c(p) = p$ for continuous populations and populations of units, while $c(p)$ is the centroid of the area containing p for populations of areas, and denote by $\phi : [0, \infty) \to \mathbb{R}^+$ a non-increasing continuous function on $(0, \infty)$, with $\phi(0) = 0$, $\lim_{d \to 0^+} \phi(d) = \infty$. The IDW interpolator can be expressed as

$$\widehat{f}(p) = I(Q_p) f(p) + I(Q_p^c) \sum_{i=1}^{n} f(P_i) w_i(p) \tag{1}$$

with weights $w_i(p)$s given by

$$w_i(p) = \frac{\phi(\|P_i - c(p)\|)}{\sum_{l=1}^{n} \phi(\|P_l - c(p)\|)}.$$

The properties of the IDW interpolator have been investigated by Fattorini et al. (2018a, b), Fattorini et al. (2019) for continuous populations, populations of areas and populations of units, respectively. In particular, three asymptotic scenarios, all referring to the infill paradigm (Cressie 1993), have been considered and the design-based asymptotic consistency of the IDW interpolator has been proved. Without entering into technical details, design-based asymptotic consistency has been achieved at the

cost of supposing: (i) some forms of smoothness of the survey variable throughout the study region; (ii) the use of a sampling design able to asymptotically achieve spatial balance of the selected locations, that is to ensure that the selected locations are well spread throughout the study region; (iii) in the case of finite populations, some sort of regularities, such as in the shape of areas; (iv) some mathematical properties of the distance functions adopted for weighting sampled observations.

It must be pointed out that the previous conditions are likely to be satisfied in most real environmental surveys. Indeed, the smoothness assumption (i) is very common and it is at the basis of most interpolation techniques (e.g., Cressie 1993, Sect. 3.1). Moreover, this assumption reasonably holds when dealing with natural phenomena where the density of an attribute changes smoothly throughout space and when it changes abruptly, that usually occurs along borders delineating variations in the characteristics of the study region (e.g., forest-meadows). Obviously, design-based consistency does not hold where discontinuities are present. However, since borders may be realistically approximated by curves well approaching the theoretical condition of discontinuity over a region of zero measure, design-based consistency is preserved on the whole.

Regarding condition (ii), the asymptotic spatial balance is ensured under the schemes usually applied in environmental surveys. The achievement of spatial balance has been the main target for long time and it is ensured under very complex schemes (see e.g., Grafström and Tillé 2013; Stevens and Olsen 2004; Jauslin and Tillé 2020) but also under some familiar, widely applied schemes such as systematic grid sampling (SGS) and tesselation stratified sampling (TSS) when dealing with continuous populations, one-per-stratum stratified sampling (OPSS) and systematic sampling (SYS) when dealing with population of areas, stratified spatial sampling with proportional allocation when dealing with population of units.

Furthermore, as to condition (iii), and in particular referring to the regularity of the shape of areas, it is ensured by the fact that in most cases, especially in environmental surveys, areas are regular polygons.

Finally, as to the choice of the distance function, the mathematical condition on ϕ ensuring the design-based consistency of (1) is

$$\lim_{d \to 0^+} d^2 \phi(d) = \infty. \tag{2}$$

Condition (2) does not actually constitute an assumption, since the distance function is chosen by the user.

A widely applied class of distance functions is the class of negative powers distance functions of order α, given by

$$\phi_\alpha(d) = d^{-\alpha}$$

which, for any fixed $\alpha > 2$, satisfies (2) and gives rise to weights of type

$$w_i(p, \alpha) = \frac{\|P_i - c(p)\|^{-\alpha}}{\sum_{l=1}^{n} \|P_l - c(p)\|^{-\alpha}}.$$

Consequently, the corresponding IDW interpolator turns out to be

$$\widehat{f}_{\alpha}(p) = I(Q_p) f(p) + I(Q_p^c) \sum_{i=1}^{n} f(P_i) w_i(p, \alpha) \tag{3}$$

which obvioulsy depends on the chosen α-value. The choice of $\phi_\alpha(d) = d^{-\alpha}$ is particularly appealing owing to its simplicity and because the interpretation of the α parameter, as a smoothing parameter, is rather straightforward. Indeed, as the weights are decreasing functions of α, the larger the values of α, the smaller the contributions of the sampled points with larger distances from p. As a matter of fact, as argued by Fattorini et al. (2021), for $\alpha \to \infty$, $\widehat{f}_{\infty}(p)$ reduces to the well-known NN interpolator in which the interpolated value of $f(p)$ is the value observed at the sampled location nearest to p. More precisely, the NN spatial interpolator of f is the piecewise constant function given by

$$\widehat{f}_{\infty}(p) = I(Q_p) f(p) + \frac{I(Q_p^c)}{\mathrm{Card}(H_p)} \sum_{i \in H_p} f(P_i)\ ,\ \ p \in B \tag{4}$$

where $H_p = \{i : \|P_i - c(p)\| = \min_{h=1,\ldots n} \|P_h - c(p)\|\}$ is the set of sampled locations that are nearest to $c(p)$. Fattorini et al. (2021) prove that the design-based consistency of (4) continues to hold and therefore, any choice of $\alpha > 2$, including $\alpha = \infty$, ensures the design-based consistency of the IDW interpolator to hold. To avoid arbitrariness, the choice of α should be performed using data driven procedures.

4 The choice of the smoothing parameter

An intuitive and widely applied data driven procedure for choosing smoothing parameters can be based on LOOCV methods (e.g., Giraldo et al. 2011; Ignaccolo et al. 2014; Montanari and Cicchitelli 2014). LOOCV is a multipurpose general criterion consisting in removing one sampled location from the dataset, interpolating the value or the density of the Y-variable at the removed sampled location using all other locations and then repeating this process for each sampled location. The interpolated values are then compared with the actual values at the omitted locations. Compared to other cross-validation techniques, LOOCV does not suffer of the random selection of the so-called training set and interpolations at the removed sample locations are obtained on the basis of a sample reduced by only one point. More precisely, according to the LOOCV, α should be selected to minimize

$$\sum_{i=1}^{n} \left[f(P_i) - \widehat{f}_{\alpha,-i}(P_i) \right]^2 \tag{5}$$

where $\widehat{f}_{\alpha,-i}(P_i)$ is the IDW interpolated value by means of the sample of $n-1$ locations obtained by deleting the i-sampled location. It is worth noting that (5), up

to a multiplicative constant depending on n or N, can be considered a naïve estimator of the overall measure of precision given by

$$\int_A \mathrm{E}\big[\widehat{f_\alpha}(p) - f(p)\big]^2 \mu(dp) \tag{6}$$

where μ is the Lebesgue measure when continuous populations and population of areas are considered while it is the counting measure with population of units. In particular, when dealing with continuous populations, (6) reduces to the design-based counterpart of the mean integrated squared error. Furthermore, when population of equal-sized areas are considered, (6) reduces to

$$\frac{\lambda(A)}{N} \sum_{j=1}^{N} \mathrm{E}\big[\widehat{f_\alpha}(c_j) - f(c_j)\big]^2 \tag{7}$$

where c_j is the centroid of a_j. Finally, for population of units, (6) turns out to be

$$\frac{1}{N} \sum_{j=1}^{N} \mathrm{E}\big[\widehat{f_\alpha}(p_j) - f(p_j)\big]^2. \tag{8}$$

Therefore, the choice of α by means of LOOCV can be interpreted as a criterion allowing to minimize the estimate of the overall precision measures. Finally, as (7) and (8) up to the multiplicative constant represent totals of mean squared errors, an alternative approach for choosing α can be based on the minimization of the Horvitz-Thompson (HT) type estimate. In particular, the HT estimate is

$$\frac{1}{N} \sum_{i=1}^{n} \frac{\big[f(P_i) - \widehat{f}_{\alpha,-i}(P_i)\big]^2}{\pi_i} \tag{9}$$

where π_i is the first-order inclusion probability of the i-th area or of the i-th unit. Note that if the sampling design adopted to select the areas or the units ensures equal first-order inclusion probabilities the two approaches are identical.

In the following, the IDW interpolator where α is chosen by means of LOOCV is denoted by $\widehat{f_{\widehat{\alpha}}}$ and henceforth, for sake of brevity, referred to as data driven (DD) interpolator.

Thanks to Proposition 1, reported in the Appendix A, design-based consistency results of the IDW interpolator obtained by Fattorini et al. (2018a, b, 2019) can be extended to the DD interpolator. Moreover, Proposition 1 allows to broaden the asymptotic properties of the IDW interpolator with fixed α not only to the DD interpolator, but also to the IDW interpolator when the smoothing parameter is chosen by any data driven procedure.

5 Pseudo-population bootstrap estimation of precision

Pseudo-population bootstrap is one of the bootstrap methods adopted in design-based inference (Mashreghi et al. 2016). It is based on constructing a pseudo-population, able to mimic the characteristics of the unknown population, from which bootstrap samples are selected using the same sampling scheme adopted in the survey.

Recently, Conti et al. (2020) provided, in a unified framework, the theoretical justification of the use of pseudo-population bootstrap in a wide range of situations. More precisely, they derived conditions on pseudo-populations, ensuring that the bootstrap distributions of plug-in estimators, resampled from these pseudo-populations by means of suitable resampling schemes, asymptotically coincide with the actual distributions of the estimators. However, the crucial condition is that, as the population size increases, the sequence of designs converges to the rejective sampling design of maximum entropy. Unfortunately, as pointed out by Franceschi et al. (2022), these results cannot be exploited in spatial surveys. Indeed, the most effective sampling designs are aimed at achieving spatial balance and do not generally converge to the rejective design of maximum entropy. Thus, we propose to use the DD interpolator for constructing the pseudo-population from which bootstrap samples are selected by means of the sampling scheme adopted to select $P_1, \ldots, P_n$ and to estimate the mean squared error of $\widehat{f}_{\widehat{\alpha}}(p)$ by means of the mean squared error of the bootstrap distribution. The intuition behind this proposal is that, under conditions ensuring consistency of the DD interpolator, the pseudo-population converges to the true one in such a way that the bootstrap distribution should converge to the true distribution, thus providing reliable estimators of the mean squared error.

Accordingly, for each $p \in B$, the pseudo-population bootstrap estimator of the root mean squared error of $\widehat{f}_{\widehat{\alpha}}(p)$ is given by

$$\widehat{V}^*_{\widehat{\alpha},M}(p) = \Big[\frac{1}{M}\sum_{m=1}^{M}\{\widehat{f}^*_{\widehat{\alpha},m}(p) - \widehat{f}_{\widehat{\alpha}}(p)\}^2\Big]^{1/2} \tag{10}$$

where M is the number of bootstrap samples and $\widehat{f}^*_{\widehat{\alpha},m}(p)$ is the bootstrapped value of the IDW interpolator at $p \in B$ based on $\widehat{f}_{\widehat{\alpha}}(P^*_{1,m}), \ldots, \widehat{f}_{\widehat{\alpha}}(P^*_{n,m})$, i.e. for any $p \in B$ and $m = 1, \ldots, M$

$$\widehat{f}^*_{\widehat{\alpha},m}(p) = I(Q^*_{p,m})\widehat{f}_{\widehat{\alpha}}(p) + I(Q^{*c}_{p,m})\sum_{i=1}^{n}\widehat{f}_{\widehat{\alpha}}(P^*_{i,m})w^*_{i,m}(p,\widehat{\alpha})$$

where $P^*_{1,m}, \ldots, P^*_{n,m}$ are the locations selected in the m-th bootstrap resampling using the scheme adopted to select the original sample, $Q^*_{p,m} = \cup_{i=1}^{n}\{P^*_{i,m} = c(p)\}$ and

$$w^*_{i,m}(p,\widehat{\alpha}) = \frac{\|P^*_{i,m} - c(p)\|^{-\widehat{\alpha}}}{\sum_{l=1}^{n}\|P^*_{l,m} - c(p)\|^{-\widehat{\alpha}}}.$$

where, with a slight abuse of notation, $\widehat{\alpha}$ is the smoothing parameter selected by means of LOOCV on the m-th bootstrap sample.

In Proposition 2 of the Appendix it is proven that, for n and M large enough, under suitable conditions,

$$\frac{E\{\widehat{V}^*_{\widehat{\alpha},M}(p)\}}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]^{1/2}} \leq \sqrt{10} \tag{11}$$

that is the pseudo-population bootstrap estimator (10) is not too conservative. Relation (11) continues to hold for random size sampling designs (such as 3P sampling, implemented in the simulation study for the population of units), when the expected value of the reciprocal of the sample size is sufficiently small (see Remark in the Appendix). Moreover, it is worth noting that (11) holds not only for DD but also for any IDW interpolator whatever data driven procedure is adopted to choose the smoothing parameter and for any fixed $\alpha > 2$.

Even if (11) may induce to suspect a large overestimation of the true mean squared error, that may even mask the effectiveness of the DD interpolator, $\sqrt{10}$ is just an upper bound and, as such, it should be viewed as a threshold limiting possible overestimation. As to the conditions for (11) to hold, the first concerns the sampling scheme adopted to select the sample points and basically is satisfied by balanced spatial sampling schemes under which the consistency of the DD interpolator is also ensured, while the second is a mathematical condition on f needed in the case of continuous populations and finite populations of areas. In particular, f is demanded to be differentiable at p with $\nabla f(p) \neq 0$. Finally, the requirement of M large enough can be readily satisfied by simply increasing the computational effort.

6 Simulation studies

An extensive simulation study has been performed in order to empirically check the theoretical findings on the design-based consistency of the DD interpolator and to assess its finite-sample properties. Moreover, the performance of the DD interpolator has been compared to that of the NN interpolator, which, in turn, avoids the intensive computational efforts needed for the selection of the smoothing parameter. More precisely, the simulation study aims to evaluate the absolute bias (AB) and root mean squared error (RMSE) of DD and NN interpolators for any location where interpolation is performed.

As to the estimation of the root mean squared error, the pseudo-population bootstrap estimator (10) has not been implemented in the simulation study owing to the unworkable increase of the computational effort, in addition to that involved in the LOOCV. However, the performance of the pseudopopulation bootstrap estimator has been already investigated by for the NN interpolator. In particular, referring to the same three population scenarios and to the same artificial surfaces adopted in this simulation study, performed an intensive simulation study that suggested the conservative nature of the pseudo-population bootstrap estimator.

The three artificial surfaces, Surf1, Surf2 and Surf3, used for generating continuous populations, finite populations of areas and finite populations of units which, at any

location $p = (p_1, p_2)$, are respectively given by

$$f(p) = C_1(sin^2 p_1 + cos^2 p_2 + p_1)$$
$$f(p) = C_2(sin 3p_1 + sin^2 3p_2)$$
$$f(p) = \begin{cases} C_3 p_1 p_2 & \text{if } \min(p_1, p_2) \leq 1/2 \\ C_3(1 + p_1 p_2) & \text{otherwise} \end{cases}$$

where the constants C_1, C_2 and C_3 ensure a maximum value of 10. The three surfaces are displayed in Fig. 1 of the Supplementary Information.

6.1 Continuous populations

Referring to the asymptotic scenario in Fattorini et al. (2018a), the three surfaces represent artificial population densities on the squared study region $(0, 1) \times (0, 1)$ and samples of increasing size are considered. In particular, sampling is performed by selecting $n = 16, 36, 64, 100$ locations by means of URS, TSS and SGS, all ensuring consistency of DD and NN interpolators. URS is the most straightforward scheme which consists in randomly and independently selecting the n locations. TSS consists of partitioning the study region into n spatial subsets of equal size and randomly and independently selecting a location in each subset. Moreover, if a regular tessellation of the study region into n regular polygons is considered, also SGS, which consists of randomly selecting a location in one polygon and systematically repeating it in the remaining polygons, can be performed.

For implementing the last two schemes, the unit square is partitioned into 4×4, 6×6, 8×8, 10×10 grids of equal-sized quadrats and a location is selected in each quadrat.

6.2 Populations of areas

Following the asymptotic scenario in Fattorini et al. (2018b), the squared study region $(0, 1) \times (0, 1)$ is partitioned into an increasing number N of areas of decreasing size and samples of increasing size are selected. More precisely, for each artificial surface, four populations of $N = 100, 400, 900, 1600$ areas are constructed by partitioning the unit square into grids of 10×10, 20×20, 30×30, 40×40 quadrats and taking, as Y-values, the integrals of the surface within quadrats. The Y-values are rescaled in such a way that the maximum value is 10. Sampling is performed by selecting $n = 0.1N$ quadrats by means of simple random sampling without replacement (SRSWOR), OPSS and SYS, all guaranteeing consistency. OPSS is implemented by partitioning the population into n blocks/strata of contiguous areas and randomly selecting an area from each block/stratum. When populations are constituted by grids of regular polygons, SYS can be alternatively performed by randomly selecting a polygon in one block and then repeating it in the remaining $n - 1$ blocks. The last two schemes are performed by partitioning grids into blocks of 2×5 contiguous quadrats and selecting one quadrat per block.

6.3 Populations of units

According to the asymptotic scenario in Fattorini et al. (2019), nested populations and samples of increasing size are considered on the study region $(0, 1) \times (0, 1)$. More precisely, three nested populations of $N = 500, 1000, 1500$ units are located on the unit square in accordance with four spatial patterns referred to as regular, random, trended and clustered. As to the regular pattern, populations are constructed 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 further 500 locations at random but discarding those with distances smaller than $0.5 \times 1000^{-1/2}$ to those previously generated, and, finally, randomly adding further 500 locations but discarding those having distances smaller than $0.5 \times 1500^{-1/2}$ to those previously generated. As to the random pattern, populations are constructed by independently generating 1500 locations at random and then assigning the first 500 to the smaller population, the first 1000 to the second one and all of them to the largest. As to the trended pattern, populations are constructed 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 one and all of them to the largest. Finally, as to the clus pattern, populations are constructed independently generating 10 cluster centers at random and assign each cluster 50 locations generated from a spherical normal distribution centerd at the cluster center with variance 0.025, adding further 50 locations to each cluster from the same distribution and, at last, adding further 50 locations to each cluster from the same distribution. Points falling outside the unit square are discarded and newly generated. The three surfaces are used for assigning the Y-values in the populations.

As to the choice of the sampling scheme, only 3P sampling, from the acronym of "probability proportional to prediction", is considered. Indeed, some of the most relevant populations of units, whose interpolation is of interest, are probably natural populations such as trees or shrubs. In such situations the list and the units locations are usually not available and mapping is commonly precluded. The sole cases in which mapping is possible, occur for populations located in study regions of limited size and all the units can be visited, and therefore located and listed. In this case, 3P sampling is commonly performed.

Under 3P sampling, all the units of the population are visited by a crew of experts, a prediction x_j for the value of the survey variable is given by the experts for each unit j of the population and units are independently included in the sample with probabilities $\pi_j = x_j / L^*$ where L^* must be large enough to ensure that $\pi_j \leq 1$ for each j (Gregoire and Valentine 2008).

Following Kinnunen et al. (2007), experts' predictions for the y_js are obtained by assuming the existence of a maximum error rate of prediction $\rho \in (0, 1)$, that occurs at the extremes of the values of the survey variable, in such a way that small values near the lower bound l are over evaluated and large values near the upper bound L are under evaluated. In this case, prediction ranges from $(1 + \rho)l$ to $(1 - \rho)L$ when the value of the survey variable is equal to l and L, respectively. Moreover, for simplicity,

experts' predictions for the y_js are generated using the relationship $x_j = a + by_j$ with $b = 1 - \rho(L + l)/(L - l)$ and $a = (1 + \rho)l - bl$, where $L = 10$, $\rho = 0.10$ and $l = 4$. In this case, $L^* = 50$ is adopted. Units with Y-value smaller than $l = 4$ are discarded from populations in order to ensure a lower bound of $\pi_0 = 0.08$ for the inclusion probabilities. Predictions, joined with choice of l, L and L^*, ensure an expected sampling fraction ranging from about 12% to 15% in all cases, according to the spatial units pattern.

6.4 Simulation implementation and results

For each combination of population, sampling scheme, and sample size, sampling is replicated 10,000 times. At each simulation run, for each $p \in B$, the DD interpolator is computed by considering the α value which minimizes (5) for α in $\{2, \ldots, 21\}$. In particular, for continuous populations, interpolation is performed on a regular grid of 100×100 locations on $(0, 1) \times (0, 1)$. As to the population of units, inaccurate and imprecise interpolations of the smallest values of the survey variable may occur (Fattorini et al. 2020). To overcome this drawback, prediction errors, given by the difference between the Y-values and the corresponding experts' predictions, are interpolated. Thus, the interpolated Y-values are given by the sum of the expert predictions and the interpolated errors.

Since for larger values of α the DD interpolator is practically indistinguishable from NN interpolator, if the minimum is reached for $\widehat{\alpha} = 21$, NN interpolation is performed. Notwithstanding for $\alpha = 2$ the asymptotic properties of the IDW are not proven, the value has been considered as it is a rather common choice for practitioners, being the default value of some widely applied GIS software, such as ArcGIS and Surfer. Furthermore, at each simulation run, the NN interpolator is also computed.

For each location where interpolation is performed, the AB and the RMSE of DD and NN interpolators are computed from the Monte Carlo distributions of the corresponding estimates. Furthermore, the mode of the Monte Carlo distribution of the finite values of $\widehat{\alpha}$ selected by means of (5) and the percentage of simulation runs giving rise to $\widehat{\alpha} = \infty$ (F_∞) are calculated. For any combination of population, sampling scheme and sample size, Tables 1, 2 and 3 report the minima, maxima and means of AB and RMSE, together with the mode and F_∞ for the DD interpolator.

Additionally, the performance of DD and NN interpolators is empirically compared by computing, for each location, a measure of relative efficiency (RE) as the ratio between the RMSE of the NN interpolator and of the DD interpolator. The corresponding cumulative distribution function are displayed in Figs. 1, 2 and 3. Figures 2–10 of the Supplementary Information show the spatial pattern of RE for all the populations, sampling schemes and sample sizes.

The simulation results confirm the theoretical findings. Indeed, from Tables 1, 2 and 3, it is at once apparent that for each combination of population, surface and sampling scheme, both AB and RMSE generally decrease as the sample size increases, with very few exceptions for population of areas when surface 3 and SYS are considered. As to the choice of the smoothing parameter, for continuous populations and for populations of areas, DD interpolator generally reduces to the NN interpolator under SGS and SYS,

Table 1 Values of mode of $\widehat{\alpha}$ distribution and F_∞ together with percentage values of mimima, means and maxima for AB and RMSE achieved for continuous populations

Pop	Scheme	n	DD								NN					
			$\widehat{\alpha}$		AB			RMSE			AB			RMSE		
			mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Surf1	URS	16	4	9	0	23	144	43	66	163	0	17	124	53	77	151
		36	5	1	0	11	97	25	39	119	0	8	79	32	50	98
		64	5	0	0	7	74	17	27	91	0	4	57	22	37	72
		100	4	0	0	5	59	13	21	73	0	3	44	17	29	56
	TSS	16	5	2	0	15	106	29	46	121	0	15	94	34	60	108
		36	5	0	0	8	69	16	28	79	0	8	59	21	40	69
		64	5	0	0	5	51	11	20	60	0	6	43	15	29	50
		100	5	0	0	4	41	8	15	48	0	4	33	12	23	39
	SGS	16	–	100	0	13	96	34	55	109	0	13	96	34	55	109
		36	–	100	0	6	60	21	35	69	0	6	60	21	35	69
		64	–	100	0	3	44	15	26	50	0	3	44	15	26	50
		100	–	100	0	2	34	12	21	40	0	2	34	12	21	40
Surf2	URS	16	3	10	0	101	272	141	195	315	0	79	261	158	213	329
		36	5	3	0	49	205	77	120	242	0	38	190	73	138	242
		64	5	4	0	30	163	41	83	195	0	22	146	36	101	188
		100	5	0	0	21	137	25	63	165	0	14	120	19	79	156
	TSS	16	5	7	0	69	257	69	149	297	0	59	257	69	169	311
		36	5	0	0	32	186	22	86	217	0	29	185	23	109	224
		64	5	0	0	20	145	10	60	170	0	18	141	10	81	172
		100	5	0	0	14	118	5	45	139	0	13	113	5	64	139
	SGS	16	–	100	0	61	325	68	165	378	0	61	325	68	165	378
		36	–	100	0	27	227	24	103	265	0	27	227	24	103	265

Table 1 continued

Pop	Scheme	n	DD							NN					
			$\widehat{\alpha}$		AB			RMSE			AB			RMSE	
			mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
		64	–	100	0	15	171	11	74	201	0	15	171	11	74	201
		100	–	100	0	9	133	6	58	158	0	9	133	6	58	158
Surf3	URS	16	3	17	8	66	364	26	124	414	0	50	363	16	126	440
		36	4	11	4	43	365	12	79	409	0	31	364	7	83	436
		64	4	9	2	33	356	7	59	400	0	23	355	4	62	428
		100	4	7	2	26	350	5	47	394	0	18	349	3	50	423
	TSS	16	4	9	3	48	364	13	80	402	0	31	358	7	83	436
		36	3	5	1	33	362	6	53	394	0	21	354	3	56	429
		64	3	3	0	26	348	3	40	381	0	15	341	2	43	420
		100	3	2	0	22	346	3	33	378	0	12	335	1	34	415
	SGS	16	–	100	0	37	361	7	84	436	0	37	361	7	84	436
		36	–	100	0	24	359	3	56	431	0	24	359	3	56	431
		64	–	100	0	17	357	2	42	428	0	17	357	2	42	428
		100	–	100	0	14	350	1	34	422	0	14	350	1	34	422

Table 2 Values of mode of $\widehat{\alpha}$ distribution and F_∞ together with percentage values of mimima, means and maxima for AB and RMSE achieved for populations of areas with 10% sampling fraction

Pop	Scheme	Grid	DD								NN					
			$\widehat{\alpha}$		AB			RMSE			AB			RMSE		
			mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Surf 1	SRSWOR	10 × 10	4	18	0	32	139	57	87	177	0	27	124	68	95	160
		20 × 20	5	1	0	10	73	22	35	95	0	8	58	29	45	78
		30 × 30	4	0	0	5	50	13	21	66	0	4	39	18	29	53
		40 × 40	4	0	0	3	39	9	15	52	0	3	28	13	22	38
	OPSS	10 × 10	5	9	0	26	120	50	72	147	0	27	113	59	86	141
		20 × 20	5	0	0	8	61	18	29	76	0	8	54	25	40	69
		30 × 30	4	0	0	4	41	10	18	51	0	5	36	15	36	46
		40 × 40	4	0	0	3	32	7	12	40	0	2	25	11	20	32
	SYS	10 × 10	–	100	0	45	159	76	110	186	0	45	159	76	110	186
		20 × 20	–	100	0	13	79	33	50	93	0	13	79	33	50	93
		30 × 30	–	100	0	6	55	20	32	65	0	6	55	20	32	65
		40 × 40	–	100	0	4	40	15	24	48	0	4	40	15	24	48
Surf 2	SRSWOR	10 × 10	2	14	5	137	264	165	234	336	0	113	226	213	257	319
		20 × 20	5	3	0	42	146	67	108	192	0	33	131	64	125	194
		30 × 30	5	0	0	22	111	28	65	145	0	16	97	22	80	142
		40 × 40	5	0	0	14	90	14	46	118	0	9	75	10	59	108
	OPSS	10 × 10	2	7	1	124	276	122	196	344	1	89	207	96	217	300
		20 × 20	5	0	0	32	146	21	82	188	0	27	133	15	103	189
		30 × 30	5	0	0	16	105	7	50	136	0	14	93	5	67	132
		40 × 40	5	0	0	10	84	4	35	108	0	8	73	2	50	102

Table 2 continued

Pop	Scheme	Grid	DD								NN					
			$\widehat{\alpha}$		AB			RMSE			AB			RMSE		
			mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
	SYS	10 × 10	–	100	11	119	450	64	200	531	11	119	450	64	200	531
		20 × 20	–	100	3	51	281	14	102	342	3	51	281	14	102	342
		30 × 30	–	100	0	22	194	4	59	237	0	22	194	4	59	237
		40 × 40	–	100	0	17	150	2	46	182	0	17	150	2	46	182
Surf 3	SRSWOR	10 × 10	3	25	14	81	292	40	155	375	0	61	305	26	154	414
		20 × 20	4	11	3	38	289	10	71	356	0	28	303	6	74	401
		30 × 30	4	9	2	26	290	5	47	352	0	16	256	3	48	365
		40 × 40	4	7	0	20	288	3	36	348	0	15	303	2	40	394
	OPSS	10 × 10	4	19	5	59	278	18	111	361	0	44	294	12	112	406
		20 × 20	4	7	0	30	272	5	51	331	0	20	288	3	56	390
		30 × 30	4	6	0	20	282	2	35	343	0	12	240	1	37	354
		40 × 40	4	3	0	17	273	2	26	326	0	12	288	1	31	385
	SYS	10 × 10	–	100	0	58	366	11	106	446	0	58	366	11	106	446
		20 × 20	6	0	0	28	282	2	50	345	0	30	357	3	58	431
		30 × 30	5	0	0	16	277	1	34	352	0	14	207	1	31	336
		40 × 40	5	0	0	14	285	1	25	340	0	16	352	1	32	425

Table 3 Values of mode of $\widehat{\alpha}$ distribution and F_∞ together with percentage values of mimima, means and maxima for AB and RMSE achieved for populations of units under different spatial patterns

Surface	Pattern	N	E(n)	DD								NN					
				$\widehat{\alpha}$		AB			RMSE			AB			RMSE		
				mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Surf 1	Regular	358	45.3	17	87	0	4	16	7	13	28	0	2	12	5	9	15
		729	91.6	16	98	0	1	9	5	7	14	0	1	9	5	6	11
		1102	138.8	–	100	0	1	9	4	5	11	0	1	9	4	5	11
	Random	349	43.7	18	86	0	4	17	7	13	29	0	2	11	6	9	15
		711	88.7	17	97	0	2	9	4	8	16	0	1	8	4	6	11
		1059	131.5	18	99	0	1	7	4	6	11	0	1	7	3	5	9
	Trended	404	54.7	18	90	0	3	18	6	13	24	0	2	16	1	7	22
		807	108.7	18	98	0	1	12	3	7	15	0	1	11	1	5	15
		1231	165.8	19	100	0	1	9	1	4	13	0	1	9	0	4	13
	Clustered	413	53.2	18	93	0	2	12	2	8	18	0	1	10	1	3	17
		827	106.4	19	99	0	1	7	1	3	10	0	1	7	1	2	9
		1237	159.4	–	100	0	1	6	1	2	7	0	1	6	1	2	7
Surf 2	Regular	179	23.1	8	56	0	17	38	19	31	50	0	11	40	8	27	50
		387	50.8	15	80	0	10	29	16	24	37	0	6	29	3	18	38
		579	76.0	13	91	0	6	26	13	19	32	0	4	25	2	15	32
	Random	179	22.9	6	54	0	18	37	18	31	52	0	11	42	11	27	51
		365	47.0	12	75	0	10	32	14	24	39	0	6	31	5	19	40
		571	74.2	20	89	0	6	28	11	18	35	0	4	28	3	15	35
	Trended	105	13.4	5	42	0	22	43	20	36	59	0	16	43	14	35	58
		210	26.7	8	57	0	16	36	18	30	49	0	9	36	6	25	47

Table 3 continued

Surface	Pattern	N	E(n)	DD								NN					
				$\widehat{\alpha}$		AB			RMSE			AB			RMSE		
				mode	F_∞	min	mean	max	min	mean	max	min	mean	max	min	mean	max
		327	41.7	14	70	0	12	35	14	25	42	0	7	35	4	20	44
	Clustered	154	18.8	3	70	0	11	42	8	22	61	0	6	46	3	14	69
		316	38.5	14	86	0	5	19	4	14	34	0	3	17	2	8	34
		473	57.7	17	94	0	3	18	3	10	25	0	2	19	2	6	26
Surf 3	Regular	131	19.1	11	78	0	7	21	8	17	34	0	2	10	4	6	12
		260	37.9	18	92	0	3	9	5	11	20	0	1	5	3	4	7
		373	54.7	18	97	0	1	9	4	7	15	0	1	7	2	3	8
	Random	124	18.1	10	76	0	7	20	9	18	33	0	2	8	4	6	11
		251	36.7	14	91	0	3	11	6	11	21	0	1	7	3	4	8
		366	53.4	18	96	0	1	7	4	8	14	0	1	6	2	3	7
	Trended	251	39.7	15	90	0	3	9	8	15	21	0	1	12	1	4	13
		513	81.1	18	99	0	1	7	4	6	9	0	1	7	0	3	9
		783	123.5	–	100	0	1	5	2	3	7	0	1	5	0	2	7
	Clustered	166	24.9	16	85	0	5	10	11	16	22	0	1	6	1	2	8
		339	50.7	17	97	0	1	5	4	7	11	0	0	4	1	1	5
		511	76.3	18	99	0	0	4	2	4	6	0	0	4	0	1	4

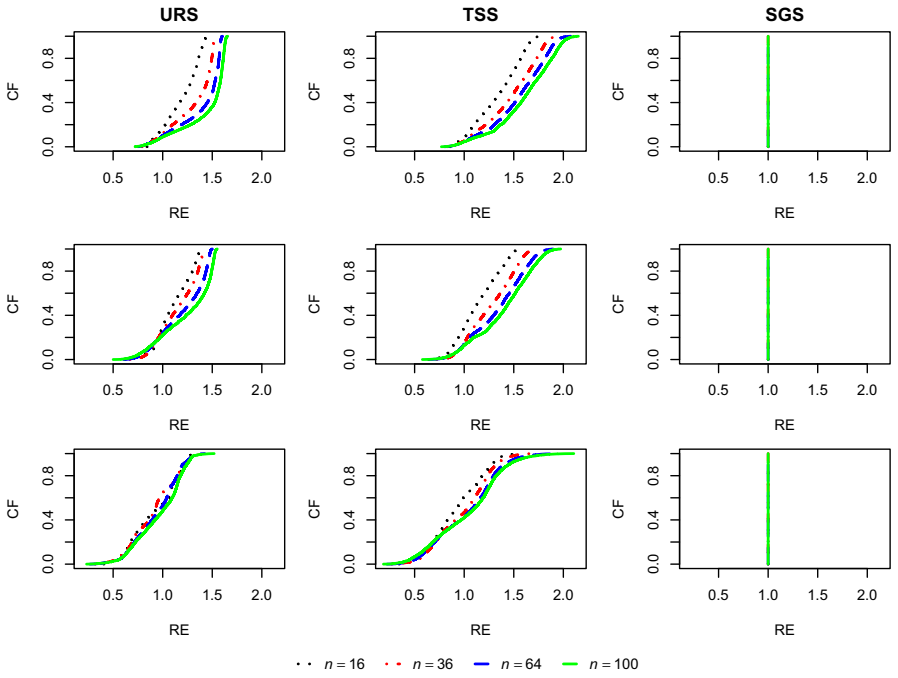

Fig. 1 Cumulative distribution functions (CF) of RE for each combination of surface (by row) and sampling scheme (by column) for continuous populations. Line types correspond to the different sample sizes

respectively. When the two interpolators do not coincide, even if the AB averages tends to be smaller for the NN interpolator, DD outperforms NN interpolator in terms of averages of RMSE and in terms of RE. The performance of the two interpolators tend to be more comparable under surface 3, which presents some discontinuities. Indeed, from Figs. 1 and 2, it is at once apparent that, for surface 1 and surface 2, under URS and TSS in case of continuous populations, and under SRSWOR and OPSS in case of populations of areas, the percentage of points where RE is smaller than 1 is rather low, while, for surface 3, the percentage is rather close to 50%.

When populations of units are considered, the percentage of simulation runs in which DD and NN interpolators coincide is mostly higher than 70% for all combinations of surfaces and sampling schemes (Table 3). However NN interpolator shows its superiority in terms of AB, RMSE and RE (Table 3; Fig. 3).

Therefore, in order to give some practical recommendations, when dealing with continuous populations and populations of areas, DD interpolator seems to be preferable to NN interpolator, while, with a population of units, NN interpolator should be adopted.

Finally, the DD interpolator is also implemented when α is chosen in order to minimize (9) giving rise to very similar results, not reported for the sake of brevity.

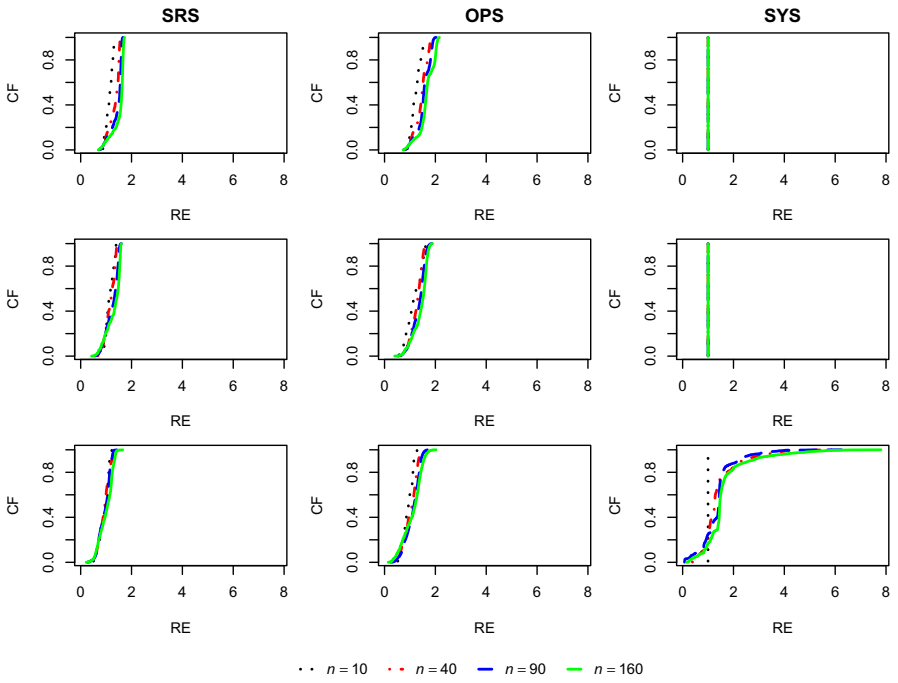

Fig. 2 Cumulative distribution functions (CF) of RE for each combination of surface (by row) and sampling scheme (by column) for populations of areas. Line types correspond to the different sample sizes

7 Case study

The proposed mapping strategy was applied for the estimation of the surface of the Shannon diversity index of tree diameter at breast height (DBH) in the experimental watershed of Bonis forest (139 ha) located in the mountain area of Sila Greca (Southern Italy) and mainly characterized by pinewoods originating from artificial reforestation. Mapping DBH Shannon diversity index is crucial for evaluating the re-naturalization processes on-going at various degrees across the watershed, which in turn are related to structural heterogeneity. In particular, for any location p of the watershed, the surface value, given by the DBH Shannon diversity index computed on a circular plot of radius 20 m centered at p, was of interest.

Data from a survey implemented in 2016, shared by the Department for Innovation in Biological, Agro-food and Forest Systems (University of Tuscia), were adopted. More precisely, plot sampling was performed by locating 36 circular plots of radius 20 m by means of URS (see Fig. 4). For each plot, DBHs were recorded and then grouped into five diameter classes (less than 17.5 cm, from 17.5 to 35 cm, from 35 to 52.5 cm, from 52.5 to 70 cm) and the Shannon diversity index was determined.

Surface estimation was performed at the centroids of 10,000 equal-sized polygons partitioning the study region by means of the IDW interpolator (3) with $\widehat{\alpha} = 3$ (see Fig. 5a). The selected smoothing parameter was obtained by the LOOCV procedure

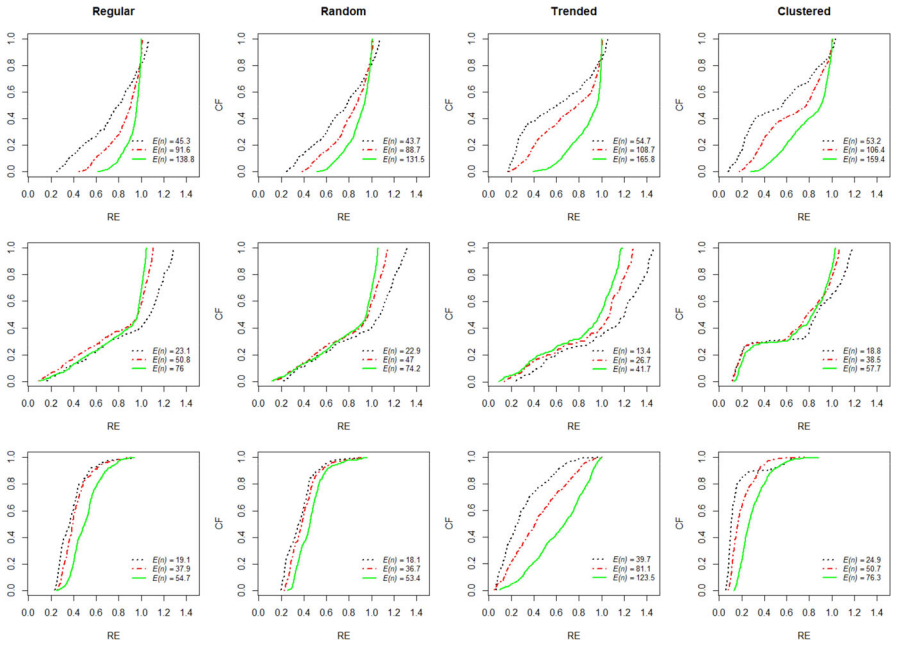

Fig. 3 Cumulative distribution functions (CF) of RE for each combination of surface (by row) and pattern (by column) for population of units. Line types correspond to the different sample sizes

described in Sect. 4. From the resulting surface, 1000 bootstrap samples of size 36 were selected according to URS to estimate RMSEs by means of (10).

Owing to the relationship between the re-naturalization process and the structual heterogeneity, Fig. 5a allows identifying areas characterized by more or less advanced re-naturalization. Furthermore, low values of uncertainties in a very large portion of the study region can be easily detected from Fig. 5b. Therefore, the estimated map can be reasonably considered as a very helpful tool for investigating re-naturalization processes and, more in general, for watershed management.

8 Conclusions

As pointed out by Maleika (2020) and Joseph and Kang (2011), a common choice for α is 2, which also constitutes the default value in widely applied GIS software and also smaller α-values are considered in the literature (see e.g., Bărbulescu et al. 2021). Often, the value of α is selected either by a visual inspection of the resulting map or by using a cross-validation approach. If the value is arbitrarily selected by the researcher, only values of $\alpha > 2$ should be considered as they guarantee the design-based consistency of the IDW interpolator (Fattorini et al. 2018a, b, 2019). In this paper, design-based consistency is proven to hold for $\alpha > 2$ also when α is obtained by optimizing any function of the sampled locations. In particular, when cross-validation techniques are adopted, minimization of the summary statistics quan-

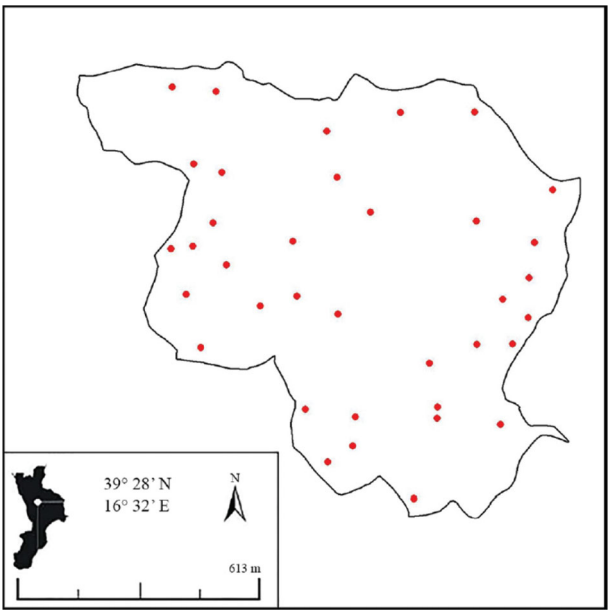

Fig. 4 Borders of the Bonis forest watershed in Sila Greca (Southern Italy) and the 36 sample plots centers

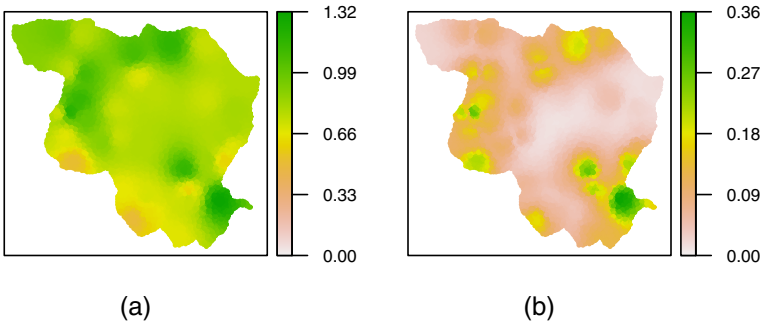

Fig. 5 Maps of the estimated diversity surface (**a**) and of the RMSE estimates (**b**) in the Bonis forest watershed

tifying the discrepancy between observed and predicted values should be performed only considering α values greater than 2. Indeed, these achieved consistency results add statistical rigor to extensively adopted cross-validation techniques for implementing IDW interpolation. Moreover, empirical results suggest that, for finite sample sizes, the performance of the DD interpolator seems to be superior to that of the NN interpolator when continuous populations and populations of areas are considered. However, their performance seems to be more comparable when discontinuities are present and no systematic designs are considered. Thus, with real populations, where discontinuities are present but the set of discontinuity points has measure zero, both interpolators are still consistent from a design-based perspective but probably the behavior of the NN

interpolator may become to be competitive. Finally, for the considered populations of units, under 3P sampling, the NN interpolator is undoubtedly preferrable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10651-023-00555-w.

Acknowledgements The authors acknowledge the support of NBFC to University of Siena, funded by the Italian Ministry of University and Research, PNRR, Missione 4 Componente 2, "Dalla ricerca all'impresa', Investimento 1.4, Project CN00000033.

Funding Open access funding provided by Università degli Studi di Siena within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix A

In order to derive the asymptotic design-based unbiasedness and consistency of the DD interpolator following the proofs by Fattorini et al. (2018a, b, 2019) it is enough to prove Proposition 1.

Proposition 1 *Suppose there exists* $\alpha_0 > 2$ *such that* $\widehat{\alpha} > \alpha_0$. *For any* $\delta, \delta' > 0$ *with* $\delta' < \delta$ *and for each* $p \in B$

$$E[\sum_{i=1}^{n} I(A_i(p,\delta))w_i(p,\widehat{\alpha})] \leq n\Big(\frac{\delta'}{\delta}\Big)^{\alpha_0} + P(B_n(\delta', p)) \tag{A1}$$

where $A_i(p,\delta) = Q_p^c \cap \{\|c(p) - P_i\| > \delta\}$ *and* $B_n(\delta', p) = \bigcap_{i=1}^{n}\{\|c(p) - P_i\| > \delta'\}$. *Moreover, it holds*

$$E[\sup_{p\in\mathcal{D}} \sum_{i=1}^{n} I(A_i(p,\delta))w_i(p,\widehat{\alpha})] \leq n\Big(\frac{\delta'}{\delta}\Big)^{\alpha_0} + P\Big(\bigcup_{p\in\mathcal{D}} B_n(\delta', p)\Big) \tag{A2}$$

where $\mathcal{D}$ *is a suitable countable subset of* B.

Proof Since $\widehat{\alpha} \geq \alpha_0$ and it holds

$$I(B_n^c(\delta', p) \cap A_i(p,\delta))w_i(p,\widehat{\alpha}) \leq I(B_n^c(\delta', p) \cap A_i(p,\delta))\frac{(\delta)^{-\widehat{\alpha}}}{\sum_{i=1}^{n}\|P_i - p\|^{-\widehat{\alpha}}} \leq \frac{(\delta)^{-\widehat{\alpha}}}{(\delta')^{-\widehat{\alpha}}},$$

for any $i = 1, \ldots, n$, then

$$I\big(B_n(\delta', p)\big) \sum_{i=1}^{n} I(A_i(p, \delta)) w_i(p, \widehat{\alpha}) \leq I\big(B_n(\delta', p)\big) \tag{A3}$$

and

$$I\big(B_n^c(\delta', p)\big) \sum_{i=1}^{n} I(A_i(p, \delta)) w_i(p, \widehat{\alpha}) \leq \frac{n(\delta)^{-\widehat{\alpha}}}{(\delta')^{-\widehat{\alpha}}} \leq n\big(\frac{\delta'}{\delta}\big)^{\alpha_0}. \tag{A4}$$

Adding (A3) and (A4) and taking expectation, inequality (A1) immediately follows. Similarly, inequality (A2) holds because

$$I\big(B_n(\delta', p)\big) \sum_{i=1}^{n} I(A_i(p, \delta)) w_i(p, \widehat{\alpha}) \leq I\big(\bigcup_{p \in \mathcal{D}} B_n(\delta', p)\big)$$

and

$$I\big(B_n^c(\delta', p)\big) \sum_{i=1}^{n} I(A_i(p, \delta)) w_i(p, \widehat{\alpha}) \leq \frac{n(\delta)^{-\widehat{\alpha}}}{(\delta')^{-\widehat{\alpha}}} \leq n\big(\frac{\delta'}{\delta}\big)^{\alpha_0}$$

which imply

$$\sup_{p \in \mathcal{D}} \sum_{i=1}^{n} I(A_i(p, \delta)) w_i(p, \widehat{\alpha}) \leq n\big(\frac{\delta'}{\delta}\big)^{\alpha_0} + I\big(\bigcup_{p \in \mathcal{D}} B_n(\delta', p)\big).$$

□

Appendix B

Proposition 2 *Suppose that, for a given sample size n, the sampling design ensures the existence of $\delta_n > 0$ such that*

$$\Pr\{\bigcap_{i=1}^{n} \{\| P_i - c(p)\| > \delta_n\}\} = 0 \tag{B5}$$

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

$$f(P_i) = f(p) + \langle a, P_i - c(p)\rangle + o(\|P_i - c(p)\|), \quad i = 1, \ldots, n. \tag{B6}$$

Then, there exists n_0 such that for any $n \geq n_0$ and for M large enough, it holds

$$\frac{E[\widehat{V}^*_{\widehat{\alpha}, M}(p)]}{E[\{\widehat{f_\alpha}(p) - f(p)\}^2]^{1/2}} \leq \sqrt{10}.$$

Proof Since $(P^*_{1,m}, \dots P^*_{n,m})$ for $m = 1, \dots, M$ are independent and identically distributed random vectors, owing to the strong law of large numbers, conditional to $P_1, \dots, P_n$, as M increases, $\widehat{V}^*_{\widehat{\alpha},M}(p)$ converges almost surely to

$$V^*_{\widehat{\alpha}}(p, P_1, \dots, P_n) = \left(E^*\left[\{\widehat{f}^*_{\widehat{\alpha},1}(p) - \widehat{f}_{\widehat{\alpha}}(p)\}^2\right]\right)^{1/2}$$

where E^* denotes expectation conditional to $P_1, \dots, P_n$.

Now consider the ratio

$$\begin{aligned}\hat{r}^*_{\widehat{\alpha}}(p) &= \frac{\{\widehat{V}^*_{\widehat{\alpha},M}(p)\}^2}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]} \\ &= \frac{\{\widehat{V}^*_{\widehat{\alpha},M}(p)\}^2 - \{V^*_{\widehat{\alpha}}(p, P_1, \dots, P_n)\}^2}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]} + \frac{\{V^*_{\widehat{\alpha}}(p, P_1, \dots, P_n)\}^2}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]}\end{aligned}$$

that, for M sufficiently large, is equivalent to

$$r^*_{\widehat{\alpha}}(p, P_1, \dots, P_n) = \frac{\{V^*_{\widehat{\alpha}}(p, P_1, \dots, P_n)\}^2}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]}$$

since $\widehat{V}^*_{\widehat{\alpha},M}(p)$ converges almost surely to $V^*_{\widehat{\alpha}}(p, P_1, \dots, P_n)$.

From the elementary inequality $(a+b)^2 \le 2(a^2+b^2)$ it follows that

$$r^*_{\widehat{\alpha}}(p, P_1, \dots, P_n) \le \frac{2E^*[\{\widehat{f}^*_{\widehat{\alpha},1}(p) - f(p)\}^2]}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]} + 2\frac{\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]}.$$

Since

$$\begin{aligned}E\{\sqrt{\hat{r}^*_{\widehat{\alpha}}(p)}\} &\le E\{\hat{r}^*_{\widehat{\alpha}}(p)\}^{1/2} \approx E\{r^*_{\widehat{\alpha}}(p, P_1, \dots, P_n)\}^{1/2} \\ &\le \Big(\frac{2E[\{\widehat{f}^*_{\widehat{\alpha},1}(p) - f(p)\}^2]}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]} + 2\Big)^{1/2}\end{aligned}$$

it is enough to prove that

$$\frac{E[\{\widehat{f}^*_{\widehat{\alpha},1}(p) - f(p)\}^2]}{E[\{\widehat{f}_{\widehat{\alpha}}(p) - f(p)\}^2]} \le 4$$

To this aim, note that

$$\widehat{f}^*_{\widehat{\alpha},1}(p) - f(p) = I(Q^{*c}_{p,1}) \sum_{i=1}^{n} \{\widehat{f}_{\widehat{\alpha}}(P^*_{i,1}) - f(p)\} w^*_{i,1}(p, \widehat{\alpha}) = X + Y,$$

where

$$X = I(Q^{*c}_{p,1}) \sum_{i=1}^{n} \{f(P^*_{i,1}) - f(p)\} w^*_{i,1}(p, \widehat{\alpha}),$$

$$Y = I(Q^{*c}_{p,1}) \sum_{i=1}^{n} \{\widehat{f_{\widehat{\alpha}}}(P^*_{i,1}) - f(P^*_{i,1})\} w^*_{i,1}(p, \widehat{\alpha}).$$

Since $\widehat{\alpha} = \phi(P_1, \ldots, P_n)$ and, for large n, $\widehat{\alpha} \approx \phi(P^*_{1,1}, \ldots, P^*_{n,1})$ it follows

$$E[X^2] \approx E[I(Q^c_p)[\sum_{i=1}^{n} \{f(P_i) - f(p)\} w_{i,1}(p, \widehat{\alpha})]^2] = E[\{\widehat{f_{\widehat{\alpha}}}(p) - f(p)\}^2].$$

Thanks to (B5) and (B6), the function f can be considered linear and, in this case,

$$E[Y^2] \approx E[\{\widehat{f_{\widehat{\alpha}}}(p) - f(p)\}^2].$$

Then, for large n, it holds

$$\frac{E[\{\widehat{f^*_{\widehat{\alpha},1}}(p) - f(p)\}^2]}{E[\{\widehat{f_{\widehat{\alpha}}}(p) - f(p)\}^2]} = \frac{E[(X+Y)^2]}{E[\{\widehat{f_{\widehat{\alpha}}}(p) - f(p)\}^2]} \leq \frac{2E[X^2] + 2E[Y^2]}{E[\{\widehat{f_{\widehat{\alpha}}}(p) - f(p)\}^2]} \approx 4.$$

The proposition is so proven. □

Remark When a random size sampling design is considered, Proposition 2 continues to hold under (B5) and (B6) if the expected value of the reciprocal of the sample size is sufficiently small. Indeed, let $\mathcal{N}$ the r.v. denoting the sample size. Then

$$P\Big(\mathcal{N} \leq n_0\Big) = P\Big(\frac{1}{\mathcal{N}} \geq \frac{1}{n_0}\Big) \leq E\Big(\frac{1}{\mathcal{N}}\Big) n_0$$

in such a way that the probability that $\mathcal{N}$ is greater than the threshold n_0 is large.

References

Bărbulescu A, Şerban C, Marina-Larisa I (2021) Computing the beta parameter in IDW interpolation by using a genetic algorithm. Water 13(6):863

Conti PL, Marella D, Mecatti F, Andreis F (2020) A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Bernoulli 26(2):1044–1069

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

Fattorini L, Franceschi S, Corona P (2020) Design-based mapping of tree attributes by 3p sampling. Biom J 62(7):1810–1825

Fattorini L, Marcheselli M, Pisani C, Pratelli L (2018a) Design-based maps for continuous spatial populations. Biometrika 105(2):419–429

Fattorini L, Marcheselli M, Pisani C, Pratelli L (2019) Design-based mapping for finite populations of marked points. Electron J Stat 13(1):2121–2149

Fattorini L, Marcheselli M, Pisani C, Pratelli L (2021) Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. Biometrics 78:1454
Fattorini L, Marcheselli M, Pratelli L (2018b) Design-based maps for finite populations of spatial units. J Am Stat Assoc 113(522):686–697
Franceschi S, Di Biase RM, Marcelli A, Fattorini L (2022) Some empirical results on nearest-neighbour pseudo-populations for resampling from spatial populations. Stats 5(2):385–400
Giraldo R, Delicado P, Mateu J (2011) Ordinary kriging for function-valued spatial data. Environ Ecol Stat 18(3):411–426
Gong G, Mattevada S, O'Bryant SE (2014) Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in texas. Environ Res 130:59–69
Grafström A, Tillé Y (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24(2):120–131
Gregoire T, Valentine H (2008) Sampling strategies for natural resources and the environment. CRC Press, Boca Raton
Hall P, Robinson AP (2009) Reducing variability of crossvalidation for smoothing-parameter choice. Biometrika 96(1):175–186
Ignaccolo R, Mateu J, Giraldo R (2014) Kriging with external drift for functional data for air quality monitoring. Stoch Environ Res Risk A 28(5):1171–1186
Jauslin R, Tillé Y (2020) Spatial spread sampling using weakly associated vectors. J Agric Biol Environ Stat 25(3):431–451
Joseph VR, Kang L (2011) Regression-based inverse distance weighting with applications to computer experiments. Technometrics 53(3):254–265
Kinnunen J, Maltamo M, Päivinen R (2007) Standing volume estimates of forests in Russia: how accurate is the published data? Forestry 80(1):53–64
Maleika W (2020) Inverse distance weighting method optimization in the process of digital terrain model creation based on data collected from a multibeam echosounder. Appl Geomat 12(4):397–407
Mashreghi Z, Haziza D, Léger C (2016) A survey of bootstrap methods in finite population sampling. Stat Surv 10:1–52
Montanari GE, Cicchitelli G (2014) Sampling theory and geostatistics: a way of reconciliation, contributions to sampling statistics. Springer, New York, pp 151–165
Noori MJ, Hassan HH, Mustafa YT (2014) Spatial estimation of rainfall distribution and its classification in Duhok governorate using GIS. J Water Resour Prot 6:75–82
Stevens DL Jr, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99(465):262–278
Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ geogr 46:234–240
Wu CY, Mossa J, Mao L, Almulla M (2019) Comparison of different spatial interpolation methods for historical hydrographic data of the lowermost Mississippi river. Ann GIS 25(2):133–151