UNIVERSITÀ
DI SIENA
1240

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA E STATISTICA

Alice Bartolini
Rosa Maria Di Biase
Lorenzo Fattorini
Sara Franceschi
Agnese Marcelli

Design-based mapping of plant species presence, association and richness by nearest-neighbor interpolation

n. 854 – Maggio 2021

Design-based mapping of plant species presence, association and richness by nearest-neighbor interpolation

by

A Bartolini[1], RM Di Biase[2], L Fattorini[1], S Franceschi[1], A Marcelli[3,4]

[1] Department of Economic and Statistics – University of Siena.

[2] CREA – Research Centre for Forestry and Wood, Arezzo.

[3] Department for Innovation in Biological, Agro-food and Forest Systems – University of Tuscia.

[4] Sustainable Agro-Ecosystems and Bioresources Department, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige (TN).

Abstract

The difference between potential and actual distribution of species is emphasized, pointing out the ecological importance of maps depicting the actual species presence on the study region. Owing to the impossibility of performing complete surveys over large areas, the presence/absence of species at a pre-fixed spatial grain is estimated for any location of the study region from the presences/absences recorded within plots centered at sample locations and having the same grain. Estimation is performed in a design-based framework by means of the well-known nearest-neighbor interpolator. Association maps and species richness maps are obtained as products and sum of the presence maps of single species. The design-based asymptotic unbiasedness and consistency of these maps are theoretically proven and pseudo-population bootstrap estimators of their precision are proposed and discussed. A simulation study is performed on a real community of 302 tree species settled in a 50-ha rectangle in the lowland tropical moist forest of Barro Colorado Island (BCI), central Panama, to check the finite-sample performance of the proposal. A case study for estimating the presence map and the association of holly oak and white violet in the Montagnola Senese (Central Italy) is reported. Technical details are contained in the appendices.

Keywords: species distribution, asymptotic unbiasedness, consistency, pseudo-population bootstrap, simulation study, case study.

Introduction

Accurate and updated wall-to-wall maps depicting the spatial distribution of plant species throughout the study region represent crucial information for many aspects of environmental research (e.g. , natural RESOURCES management and NATURE conservation planning (e.g., Corona et al., 2010; Franklin, 2010) also playing a basic role in determining economic values of environmental resources (e.g., Champ et al., 2017).

Because species presence occurs at individual locations, it is customary to consider presence at a prefixed spatial grain, i.e., within regular plots (e.g., circles or oriented quadrats) of a prefixed size that determines the spatial resolution of maps (Turner at al., 2001). Recording presence instead of single individuals is especially suitable when dealing with plants that often exhibit clonal reproduction (Palmer, 1990). In accordance with this approach, any location of the study region can be virtually considered as the center of a plot, in such a way that there exists a presence/absence of species at any location, giving rise to the "presence surface" of the species on the study region.

In most cases, the available resources make impossible to completely census the entire region. Usually, presence/absence is recorded only within those plots centered on a sample of locations and an estimation criterion is adopted to make inference on the spatial distribution of species throughout the whole study region. It is worth noting that the procedure lies in the framework of the so-called fixed-area plots (e.g., Gregoire and Valentine, 2008, chapter 7, p. 209), in which a perfect detection takes place within plots. That obviously involves accurate, expensive fieldworks performed by well-trained crews.

Generally, methods adopted to make inference on the spatial distribution of species from presence/absence data lie in the realm of model-dependent inference, i.e., the sampled locations are held fixed (as if they were purposively selected) and values at these locations are supposed to be random variables generated from a spatial process (super-population). Therefore, under model-dependent approaches, uncertainty stems from the super-population model that has been supposed to generate the presences/absences, conditional to the sampled locations.

Most species distribution models link partial presence/absence information with environmental covariates to predict the spatial distribution of species potential, i.e., the potential occurrence at any unsampled location of the study region. In literature, these models are also referred to as "predictive habitat distribution models" (Guisan and Zimmermann, 2000) and "spatially explicit habitat suitability models" (Rotenberry et al., 2006), while the resulting maps have been variously referred to as, among others, "ecological response surfaces" (Lenihan, 1993), "biogeographical models of species distributions" (Guisan et al., 2006), "spatial predictions of species distribution" (Austin, 2002), "predictive maps" (Franklin, 1995), "predictions of occurrence" (Rushton et al., 2004), "predictive distribution maps" (Rodriguéz et al., 2007). In practice, map values predict the likelihood of presence and have been variously interpreted as the probability of species presence or the potential species distribution (Scott et al., 2002), the potentially occupied locations (Guisan and

Zimmermann, 2000; Edwards et al., 2006) or the location suitability or quality (Hirzel and Guisan, 2002; Hirzel et al., 2002; Gibson et al., 2004).

However, our purpose is not to deal with the myriad of models adopted to predict species presence, such as geostatistical models, generalized linear and additive models, multivariate adaptive regression splines, and machine learning methods. On these topics, Franklin (2010) provides an accurate, excellent review. Rather, we here emphasize the necessity of distinguish between the potential distribution and the actual distribution of a species, that we have previously referred to as the presence surface. Indeed, for several reasons, a species may be found in unsuitable locations or may be absent from those suitable (e.g., van Horne, 1983). As Franklin (2010) points out, confounding occupancy with suitability may be an oversimplification. While predictive maps are useful for extrapolating purpose (e.g., to predict the species presence in unobserved locations), presence surfaces are useful to depict the actual species presence on the study region. In practice, in the latter case, for any location of the study region we have to estimate presence/absence (usually labelled as 1/0) of the species on the basis of the presences/absences recorded at sample locations.

In forest studies, Mc Roberts et al. (2010) propose to predict presence if the estimated presence probabilities - achieved by means of a logistic regression from remote sensing covariates - are greater than 0.5 and predict absence otherwise. However, even if frequently applied, the proposal just represents an arbitrary rule of thumb without any theoretical foundation that may justify its use. On the other hand, presence surfaces can be estimated by using the nearest neighbor (NN) interpolator, i.e., assigning at any location in the study region the value observed at the nearest sample location. Therefore, the NN interpolator has the appealing property that interpolated values have the same support of the survey variable, even when the support is dichotomous as in our case.

Owing to its simplicity, mapping by NN interpolation constitutes a widely extended practice in environmental surveys (e.g., Li and Heap, 2008). Unfortunately, despite its large use, NN interpolator has been invariably adopted just as a descriptive technique. By a model-dependent perspective, Cressie (1991, section 5.9) relegates NN interpolator in a class of techniques referred to as “non-stochastic methods of spatial prediction” for which no stochastic model is assumed and hence no uncertainty is associated. That is probably due to the widely spread opinion that inference in spatial mapping is hard to perform without referring to models.

In this note, we follow the alternative proposal by Fattorini et al. (2021) that approaches the NN interpolator from a design-based perspective, i.e., the surface to be mapped is viewed as constant and uncertainty stems from the probabilistic sampling scheme adopted to select locations. Differences between model-dependent and design-based inferences are well delineated in statistical literature (Smith, 1994, 2001; Gregoire, 1998; Thompson, 2002, chapter 10; Schreuder et al., 1993;

Little, 2004). As pointed out by Särndal et al. (1992, p. 21), the main appeal of the design-based approach is that “Design-based inference is objective, nobody can challenge that the sample was really selected according to the given sampling design. The probability distribution associated with the design is real, not modelled or assumed.”

Following the proposal by Fattorini et al. (2021), the presence/absence of a species at single unsampled locations is estimated by the presence/absence recorded at the nearest sample location and the asymptotic design-based properties of the NN interpolator are derived from the features of the surface to be interpolated as well as from some characteristics of the adopted sampling schemes as the number of sampled locations increases.

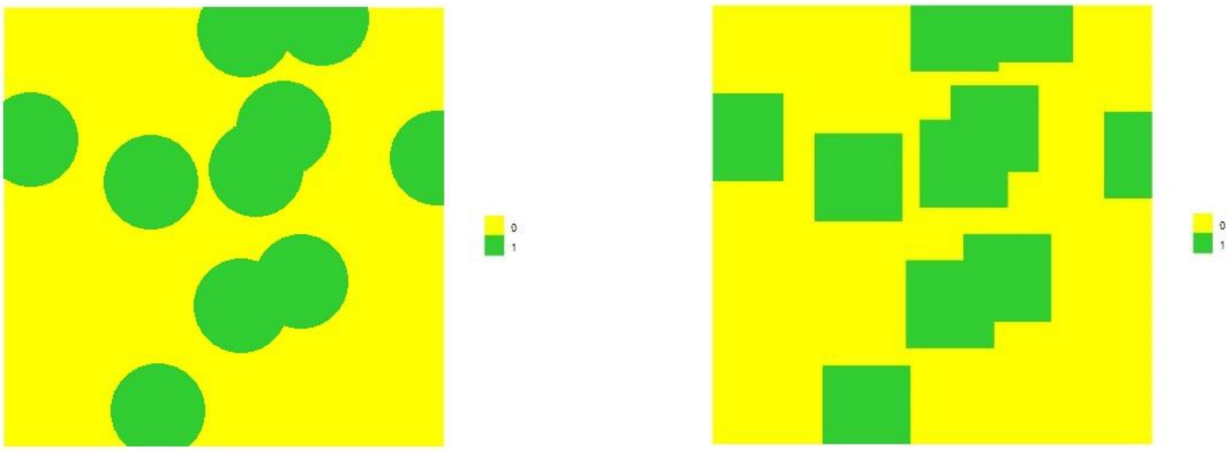
The paper is organized as follows. In section 2, the presence surface and related ones - i.e. species association surfaces and species richness surfaces - are introduced, emphasizing how these surfaces share suitable mathematical properties. They are proven to be piecewise Lipschitz functions almost everywhere. This feature is of relevant importance for ensuring design-based consistency of the NN interpolation that is derived in section 3 by exploiting the results by Fattorini et al. (2021). It is important to emphasize that, besides the Lipschitzian nature of these surfaces, consistency requires the capacity of the adopted sampling scheme to evenly spread sample locations in such a way that, as the number of sample locations increases, any non-sampled location is likely to have neighboring locations sampled. This feature is usually referred to as spatial balance in sampling literature. A bootstrap estimator of the precision indexes of the interpolated values is proposed once again from the results by Fattorini et al. (2021). In section 4, a simulation study is performed from a real stand of trees settled in a 50-ha rectangle in the lowland tropical moist forest of Barro Colorado Island (BCI), central Panama. In section 5, a case study is considered to estimate presence and association maps of a tree species and a grass species throughout the Montagnola Senese, a region protected as a Site of Community Importance in Central Italy. Concluding remarks are contained in section 6.

2. Preliminaries and notations

Denote by A the study region of size $|A|$ and let U be the population of N individuals of a plant species settled on A at locations $p_1, \ldots, p_N$. For any point $p \in A$, let $y(p)$ be the surface value that is equal to 1 if there exists at least one individual $j \in U$ such that $\|p_j - p\| \leq r$, and equal to 0 otherwise, where $\|\ \|$ denote a norm in R^2. The resulting surface

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

will be referred to as the presence surface of the species. In this framework, r determines the spatial grain of the surface, i.e., the size of plots in which presence/absence is recorded and the norm $\|\ \|$ determines the shape of these plots. For example, if $\|\ \|$ is the Euclidean norm, then plots are circles of radius r, while if $\| \|$ is the Chebyshev norm, then plots are oriented quadrats of side $2r$. Irrespective of their shape, these plots will be referred to as r-plots. Practically speaking, the presence surfaces is equal to 1 on the union the N r-plots centered at the individual locations and is equal to 0 otherwise. That is exemplified in Figure 1 that shows the presence surface in the case of Euclidean norm (a) and Chebyshev norm (b) of a species having 10 individuals settled on a squared study region. It is worth noting that in a design-based approach, the presence surface is a fixed, unknown characteristic of the species distribution on the study region that must be estimated from the presences/absences recorded at sampled locations, rather than a realization of a spatial model as assumed in model-dependent approaches.



(a) (b)

Figure 1. Presence surface (green) for a species with 10 individuals settled in a squared region under Euclidean norm (a) and Chebishev norm (b).

Regarding the mathematical properties of (1), it is at once apparent from Figure 1 that it is piecewise constant, jumping from 0 to 1 along borders of measure 0. Therefore, in mathematical terms, the presence surface (1) is a Lipschitz function almost everywhere. These (quite apparent) properties will play an important role in determining the properties of the NN interpolator. Obviously, the more the individuals of the species, the more the discontinuities of the presence surface, as shown in Figure 2 that depicts the presence surface of amarillon (*Lonchocarpus heptaphyllus*) resulting from a collection of 712 trees settled in BCI at the spatial grain of 13 m radius circular plots.

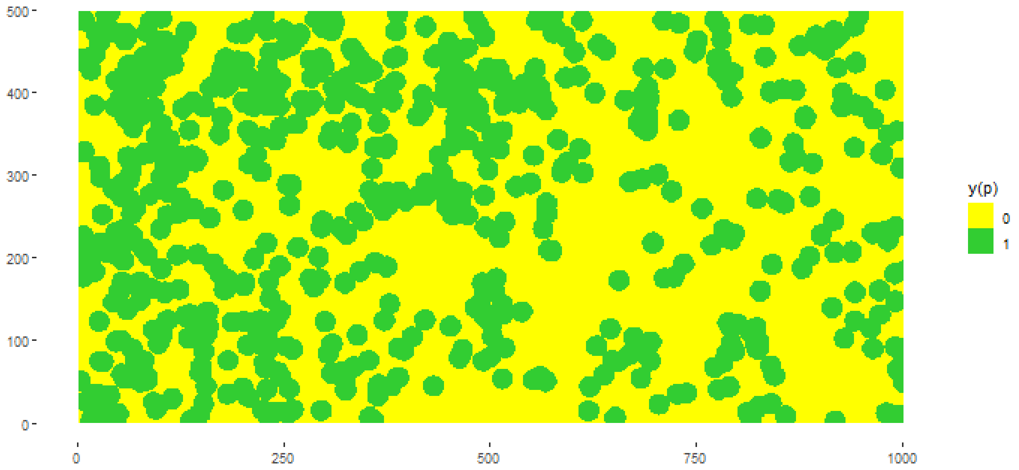


Figure 2. Presence surface (green) of amarillon resulting from 712 trees settled in BCI at the spatial grain of 13 m radius circular plots.

From the presence surface (1) other surfaces of relevant ecological importance can be achieved. To this purpose, denote by L the list of the K species that are present in the study region. Obviously, for each species $l \in L$ there is a presence surface $\{y_l(p), p \in A\}$ depicting its distribution on the study region. Therefore, for each pair of species $l, h \in L$ the product of their presence surfaces

$$y_{lh}(p) = y_l(p)y_h(p) \ \ ,p \in A \tag{2}$$

will be referred to as the association surface of species l and h. Practically speaking, the surface (2) is equal to 1 for any point whose r-plot contains both species and is equal to 0 otherwise. Therefore, (2) depicts the association of the two species throughout the study region.

Finally, from the presence surfaces of each species $l \in L$, their sum

$$y_L(p) = \sum_{l \in L} y_l(p) \ \ ,p \in A \tag{3}$$

gives the number of species that are present in the r-plot centered at p and as such it will be referred to as the richness surface. It is worth noting that surfaces (2) and (3), arising as the product and the sum of Lipschitz functions almost everywhere, are both Lipschitz functions almost everywhere.

3. Nearest-neighbor estimation of presence, association and richness surfaces.

As stated in the Introduction, it is usually unfeasible to completely census the entire study region. In most cases, the presence/absence of species is recorded at n sample locations $P_1, \dots, P_n$, i.e., a r-plot is centered at each sample location P_i $(i = 1, \dots, n)$ and the presence/absence of the species under study is recorded as $y(P_i) = 1$ if the species is present and $y(P_i) = 0$, otherwise. Then a criterion is necessary to estimate presence/absence at any unsampled location $p \in A$.

Most estimation criteria for mapping species distribution lie in the realm of model-dependent inference and provide probability of presence rather than actual presence. We here attempt to estimate realized presence surfaces in a design-based framework, in such a way that the properties of the resulting maps are only determined by the probabilistic sampling scheme adopted to select the sample locations $P_1, \dots, P_n$.

Following the idea by Fattorini et al. (2021), we adopt the NN interpolator in which the presence/absence at a non-sampled location $p \in A$ is estimated by

$$\hat{y}(p) = y\left(P_{NN(p)}\right) \tag{4}$$

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

Design-based expectation and variance of (4) cannot be expressed in closed forms, giving no insights about its bias and precision. Therefore, conditions ensuring design-based asymptotic unbiasedness and consistency are needed to obtain the statistical soundness of (4). Without entering the theoretical complexities involved in Fattorini et al. (2021) to achieve the asymptotic features of (4), it is sufficient to point out that a presence surface may show many discontinuities but only along r-plot borders or traits of borders, being constant elsewhere. Therefore, even if no consistency of (4) can hold where discontinuities are present, consistency holds under suitable sampling schemes at any continuity point of the surface, i.e., almost everywhere. Hence, consistency holds also for the whole map. In turn, regarding the sampling schemes needed for consistency, they should be able to achieve an asymptotic spatial balance, i.e., any location of the study area should have neighboring locations sampled for a sufficiently large sample size n.

In accordance with these results, under asymptotically balanced schemes, the estimated presence surface converges to the true presence surface at any continuity point p in such a way that the mean absolute error

$$MAE\{\hat{y}(p)\} = E[|\hat{y}(p) - y(p)|] \tag{5}$$

and the mean integrated absolute error

$$MIAE\left(\hat{f}\right) = \int_B AE\{\hat{y}(p)\}\, dp \tag{6}$$

both converge to 0.

In particular, Fattorini et al. (2021) prove that consistency occurs for those schemes widely applied in environmental surveys, such as uniform random sampling (URS), in which n locations are randomly and independently selected under the study region, tessellation stratified sampling (TSS), in which the study region is partitioned into n patches of equal size and one location is randomly selected within each patch, and systematic grid sampling (SGS), in which the study area is partitioned into n regular polygons, one location is randomly selected in one polygon and then repeated in the others (e.g., Barabesi et al., 2003).

Obviously, association and richness surfaces can be readily estimated from the estimated presence surfaces by means of

$$\hat{y}_{lh}(p) = \hat{y}_l(p)\hat{y}_h(p) \ \ , p \in A \tag{7}$$

and

$$\hat{y}_L(p) = \sum_{l \in L} \hat{y}_l(p) \ \ , p \in A \tag{8}$$

respectively. Consistency of (7) and (8) at any continuity point and for the whole maps readily follow from the consistency of (4).

Besides these findings, owing to the dichotomic nature of presence and association surfaces, more compelling results are achieved. At first, it can be proven that the bias is invariably positive if $y(p) = 0$ and negative if $y(p) = 1$, and in both cases its absolute value coincides with the error probability $Pr\{\hat{y}(p) \neq y(p)\}$. Moreover, the error probability completely determines precision because it also coincides with the mean absolute error (5) as well as with the mean squared error (See Appendix A for the proof). Subsequently, it can be proven that under URS, the error probability at any continuity point approaches zero at least as c^n with $c \in (0,1)$, while under TSS and SGS, the error probability is definitively equal to 0 for a sufficiently large sample size. Similar results hold also for the richness surface. However, because the richness surfaces is not dichotomic, bias and precision cannot be directly determined by the error probability (see Appendix B).

Regarding the estimation of map precision, Fattorini et al. (2021) propose to follow the pseudo-population bootstrap (PPB) approach, based on constructing a pseudo-population likely to resemble the true population from which bootstrap samples are selected using the same sampling scheme adopted in the survey. Therefore, the key problem under PPB is to reconstruct pseudo-populations able to mimic the characteristics of the unknown population, in such a way that the bootstrap distribution of any statistic can resemble the true distribution with indexes of precisions approaching the true ones (e.g., Quatemberg, 2015). Accordingly, in order to estimate the precision of the surfaces (4), (7) and (8), we pursue the idea of using the estimated maps as pseudo-populations from which bootstrap samples are selected using the same spatial scheme adopted to select the original sample. Because the estimated maps converge to the true maps, bootstrap distributions of nearest neighbour interpolator achieved by resampling from these maps should converge to the true distributions, also providing consistent estimators for their indexes of precision such as mean squared errors, root mean squared errors, relative root mean squared errors.

To this purpose, let $\hat{y}(A) = \{\hat{y}(p), p \in A\}$ be the estimated presence or association map based on the sample observations $y(P_1), \dots, y(P_n)$. Because in these cases mean square errors coincide with error probabilities, it seems suitable to use mean squared errors as indexes of precision to be estimated at any $p \in A$ by the bootstrap mean squared error

$$\widehat{mse}_B^*(p) = \frac{1}{B} \sum_{b=1}^{B} [\hat{y}_b^*(p) - \hat{y}(p)]^2 \tag{9}$$

where B is the number of bootstrap samples, $P_{1,b}^*, \dots, P_{n,b}^*$ are the locations selected in the b-th bootstrap resampling using the scheme adopted to select the original sample, $\hat{y}(P_{1,b}^*), \dots, \hat{y}(P_{n,b}^*)$ are the sample observations at these locations derived from the estimated map $\hat{y}(A)$, and $\hat{y}_b^*(p)$ is the bootstrapped value of the nearest neighbour interpolator at $p \in A$ based on $\hat{y}(P_{1,b}^*), \dots, \hat{y}(P_{n,b}^*)$, i.e.

$$\hat{y}_b^*(p) = \hat{y}\left(P_{NN(p),b}^*\right), \quad p \in A, b = 1, \dots, B \tag{10}$$

where $P_{NN(p),b}^* = argmin_{i=1,\dots n} \left\| P_{i,b}^* - p \right\|$.

The same bootstrap procedure can be performed, *mutatis mutandis*, for the estimates of association surfaces $\hat{y}_{lh}(A) = \{\hat{y}_{lh}(p), p \in A\}$ and richness surfaces $\hat{y}_L(A) = \{\hat{y}_L(p), p \in A\}$. However, because richness surfaces take integer values from 0 to K, if $y_L(p) > 0$ for each $p \in A$, it seems more meaningful to consider relative root mean squared errors as indexes of precision to be estimated by

$$\widehat{rrmse}_B^*(p) = \frac{\left\{\frac{1}{B} \sum_{b=1}^{B} \left[\hat{y}_{L,b}^*(p) - \hat{y}_L(p)\right]^2\right\}^{1/2}}{\hat{y}_L(p)} \qquad 11)$$

Unfortunately, because presence, association and richness surfaces are piecewise constant, the requirements necessary for proving the conservative nature of these bootstrap estimators do not hold (see Fattorini et al., 2021, Proposition 3). Indeed, as argued in the Appendix C, the bootstrap estimators of mean squared errors and relative root mean squared errors may be quite unstable especially near the borders where discontinuities occur or in the inner parts of presence/absence regions, where the true mean squared errors vanish.

4. Simulation study

To investigate the performance of the estimators (4), (7) and (8) and of their bootstrap estimators of precision (9) and (11), a simulation study was performed on some species of a real population. In particular, the study region A considered in the simulation was a rectangle of 50 ha located in the lowland tropical moist forest of Barro Colorado Island (BCI), central Panama, where a complete field enumeration of trees was carried out in 2010 to give location and species for each freestanding woody stem with at least 1 cm diameter at breast height. The field work mapped a population of $N = 221{,}758$ trees partitioned into $K = 302$ species (data available at https://repository.si.edu/handle/10088/20925).

Three different tree species with different degrees of presence were chosen for performing the simulation study. Cerillo tree (*Lacmellea panamensis*), caimito de mono (*Chrysophyllum argenteum*) and muskwood (*Guarea guidonia*) were chosen as species with low, medium and high degree of presence with 102, 775 and 1993 trees, respectively. For each species, the presence surface at p was the presence/absence of trees within a circular plot of radius 13 m centered at p (see Figures 3-5).

Sampling was performed selecting $n = 50{,}100{,}150{,}200$ locations by means of uniform random sampling (URS), tessellation stratified sampling (TSS) and systematic grid sampling (SGS). For implementing the last two schemes, the study area was partitioned in to 10x5, 10x10, 15x10 and 20x10 grids of equal sized rectangles and a location was randomly selected in each rectangle.

For each combination of species, sampling scheme and sample size, sampling was replicated $R = 10{,}000$ times. At each simulation run, interpolator (4) was computed onto a regular grid G of 200x100 locations on A and $B = 1{,}000$ bootstrap samples were independently selected using the same scheme adopted to select the original sample, in order to compute the bootstrap mean squared error estimator (9).

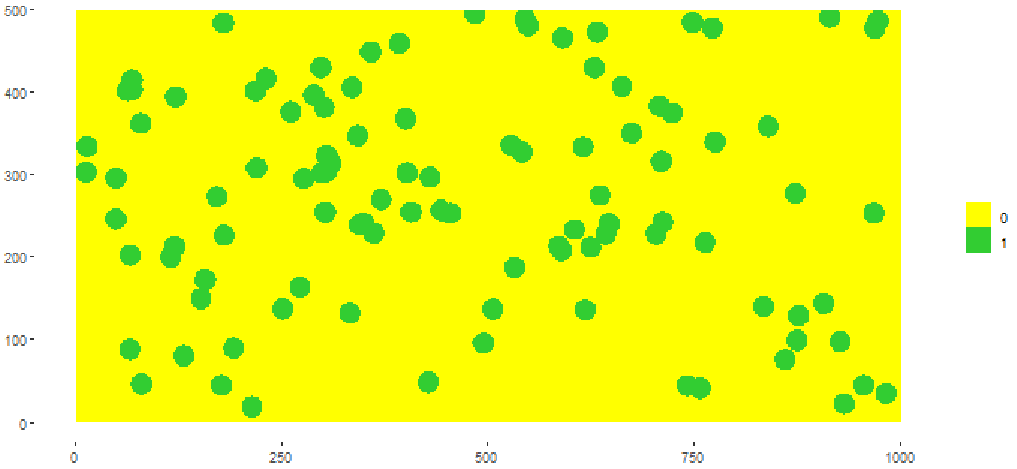


Figure 3. Presence surface (green) of cerillo tree resulting from 102 trees settled in BCI at the spatial grain of 13 m radius circular plots.

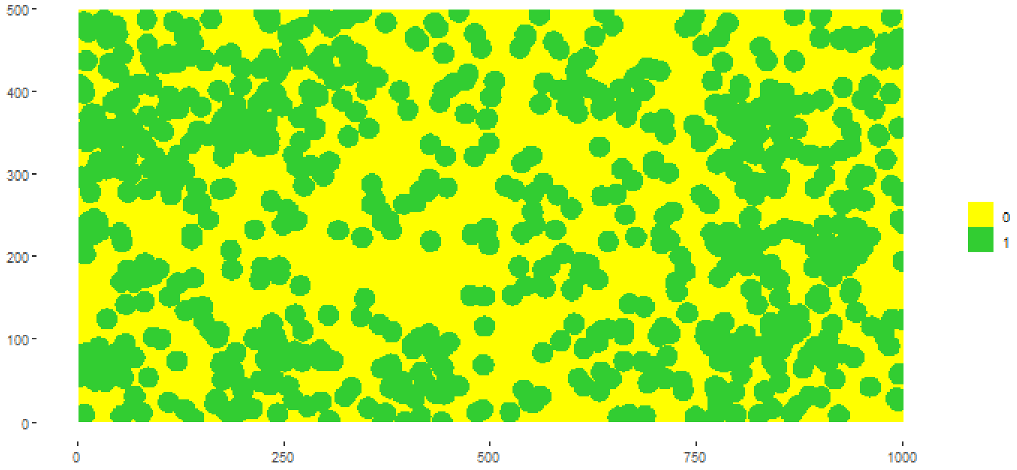


Figure 4. Presence surface (green) of caimito de mono resulting from 775 trees settled in BCI at the spatial grain of 13 m radius circular plots.

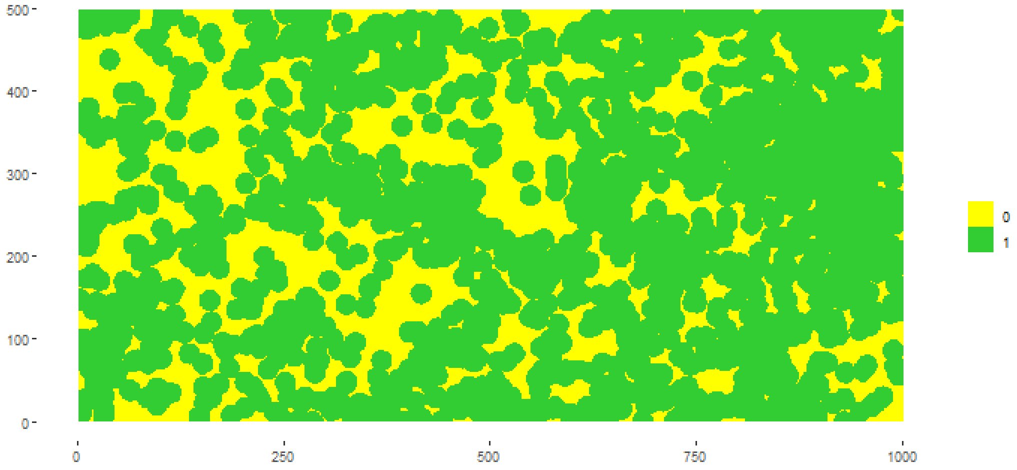


Figure 5. Presence surface (green) of muskwood resulting from 1993 trees settled in BCI at the spatial grain of 13 m radius circular plots.

Denote by $\hat{y}_r(p)$ the interpolator (4) and by $\widehat{mse}^*_{B,r}(p)$ the bootstrap mean squared error estimator (9) achieved at the r-th simulation run for each $p \in G$. Then, the expectation (E)

$$e(p) = \frac{1}{R}\sum_{r=1}^{R} \hat{y}_r(p) \tag{12}$$

the bias (BIAS)

$$b(p) = e(p) - y(p) \tag{13}$$

and the mean squared error (MSE)

$$mse(p) = \frac{1}{R}\sum_{r=1}^{R}[\hat{y}_r(p) - y(p)]^2 \tag{14}$$

were empirically determined from the Monte Carlo distributions of the estimates, together with the expectation of the bootstrap mean squared error estimator (9)

$$ebmse(p) = \frac{1}{R}\sum_{r=1}^{R}\widehat{mse}_{B,r}^{*}(p)$$

from which the ratio to the true MSE was achieved (BORAT)

$$borat(p) = \frac{ebmse(p)}{mse(p)} \tag{15}$$

for those $p \in G$ for which $mse(p) > 0$.

For any combination of species, sampling scheme and sample size, Tables 1-3 report the minima, averages and maxima of BIAS, MSE and BORAT, while Figures 6-8 show the corresponding spatial patterns of these quantities, together with the spatial pattern of the expectations (first column) under TSS. Spatial patterns achieved under URS and SGS are quite similar and are not reported for brevity.

Table 1. Values of minima, averages and maxima for BIAS, MSE and BORAT achieved from the simulation performed on the presence surface of cerillo tree.

Scheme	n	BIAS			MSE			BORAT		
		min	mean	max	min	mean	max	min	mean	max
URS	50	-0.95	0.00	0.26	0.01	0.17	0.95	0.04	0.97	4.05
	100	-0.91	0.00	0.37	0.00	0.16	0.91	0.05	1.29	18.33
	150	-0.89	0.00	0.44	0.00	0.15	0.89	0.07	1.81	44.50
	200	-0.88	0.00	0.49	0.00	0.15	0.88	0.09	2.35	77.00
TSS	50	-0.95	0.00	0.28	0.00	0.17	0.95	0.02	1.95	303.00
	100	-0.93	0.00	0.41	0.00	0.16	0.93	0.04	3.38	440.00
	150	-0.90	0.00	0.53	0.00	0.15	0.90	0.05	5.98	416.00
	200	-0.88	0.00	0.60	0.00	0.14	0.88	0.05	7.38	374.00
SGS	50	-0.97	0.00	0.32	0.00	0.17	0.97	0.02	1.26	257.00
	100	-0.94	0.00	0.48	0.00	0.16	0.94	0.02	1.55	115.86
	150	-0.90	0.00	0.60	0.00	0.15	0.90	0.02	2.66	563.00
	200	-0.87	0.00	0.61	0.00	0.14	0.87	0.04	1.59	63.52

Table 2. Values of minima, averages and maxima for BIAS, MSE and BORAT achieved from the simulation performed on the presence surface of caimito de mono.

Scheme	n	BIAS			MSE			BORAT		
		Min	mean	max	min	mean	max	min	mean	max
URS	50	-0.79	0.00	0.82	0.17	0.45	0.82	0.20	0.57	1.22
	100	-0.83	0.00	0.87	0.07	0.43	0.87	0.17	0.62	2.04
	150	-0.83	0.00	0.91	0.03	0.41	0.91	0.15	0.67	3.57
	200	-0.82	0.00	0.92	0.02	0.39	0.92	0.13	0.71	4.76
TSS	50	-0.80	0.00	0.87	0.11	0.45	0.87	0.11	0.51	1.50
	100	-0.86	0.00	0.92	0.04	0.42	0.92	0.13	0.60	2.97
	150	-0.85	0.00	0.92	0.01	0.40	0.92	0.08	0.65	9.01
	200	-0.88	0.00	0.98	0.00	0.38	0.98	0.06	0.88	387.00
SGS	50	-0.86	0.01	0.90	0.05	0.44	0.90	0.13	0.50	3.59
	100	-0.88	0.00	0.92	0.00	0.42	0.92	0.09	0.59	44.31
	150	-0.85	0.00	0.96	0.00	0.39	0.96	0.06	0.72	185.80
	200	-0.85	0.00	0.95	0.00	0.37	0.95	0.07	0.81	65.67

Table 3. Values of minima, averages and maxima for BIAS, MSE and BORAT achieved from the simulation performed on the presence surface of muskwood.

Scheme	n	BIAS			MSE			BORAT		
		Min	mean	max	min	mean	max	min	mean	max
URS	50	-0.59	0.00	0.97	0.01	0.33	0.97	0.03	0.73	2.76
	100	-0.67	0.00	0.99	0.00	0.32	0.99	0.02	0.82	6.26
	150	-0.71	0.00	0.99	0.00	0.30	0.99	0.02	0.90	18.67
	200	-0.73	0.00	0.99	0.00	0.29	0.99	0.01	0.97	39.00
TSS	50	-0.71	0.00	1.00	0.00	0.33	1.00	0.02	0.70	8.08
	100	-0.76	0.00	1.00	0.00	0.31	1.00	0.01	0.86	22.67
	150	-0.79	0.00	0.99	0.00	0.29	0.99	0.00	1.17	345.00
	200	-0.77	0.00	0.99	0.00	0.28	0.99	0.00	1.42	443.00
SGS	50	-0.72	0.00	1.00	0.00	0.33	1.00	0.01	0.74	52.75
	100	-0.79	0.00	0.99	0.00	0.30	0.99	0.00	0.85	37.27
	150	-0.84	0.00	0.99	0.00	0.29	0.99	0.00	1.08	172.50
	200	-0.79	0.00	0.99	0.00	0.28	0.99	0.00	1.03	38.41

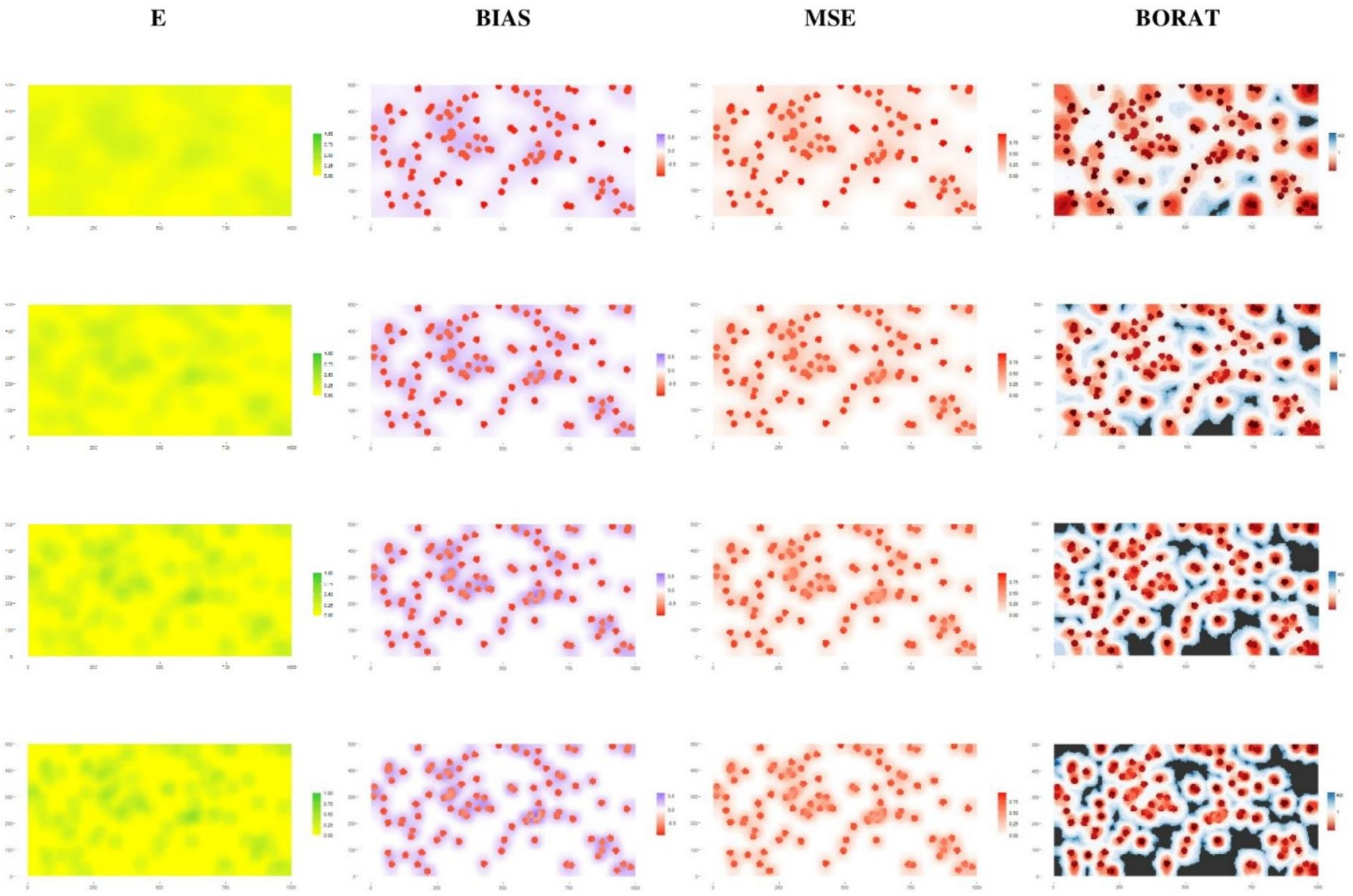


Figure 6. Maps of expectations (E), bias values (BIAS), mean squared errors (MSE) and bootstrap ratios (BORAT) achieved from the simulation performed on the presence surface of cerillo tree under TSS and $n = 50,100,150,200$ sample locations. Black patches denote locations where MSEs are equal to 0.

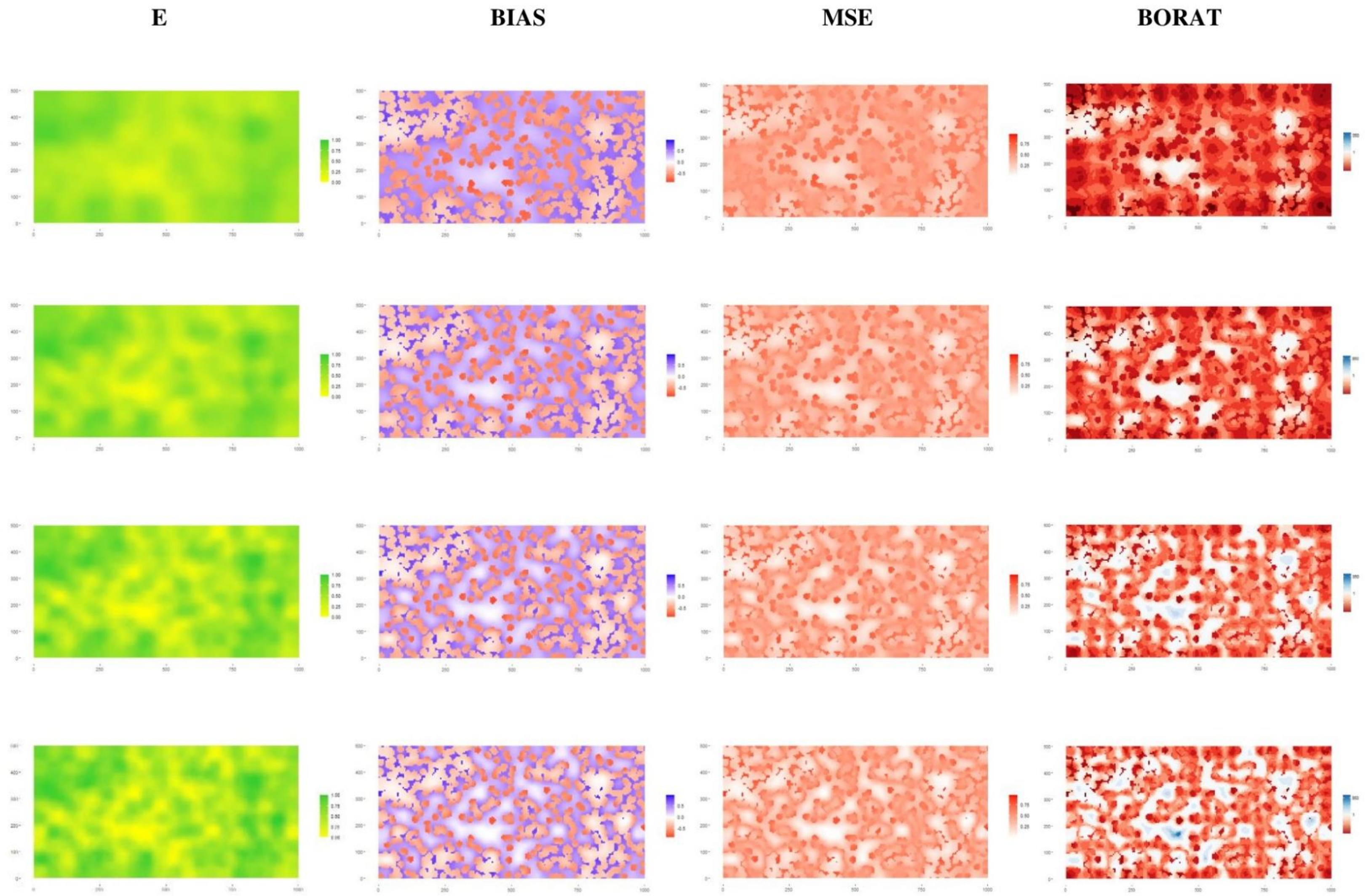


Figure 7. Maps of expectations (E), bias values (BIAS), mean squared errors (MSE) and bootstrap ratios (BORAT) achieved from the simulation performed on the presence surface of caimito de mono under TSS and $n = 50,100,150,200$ sample locations. Black patches denote locations where MSEs are equal to 0.

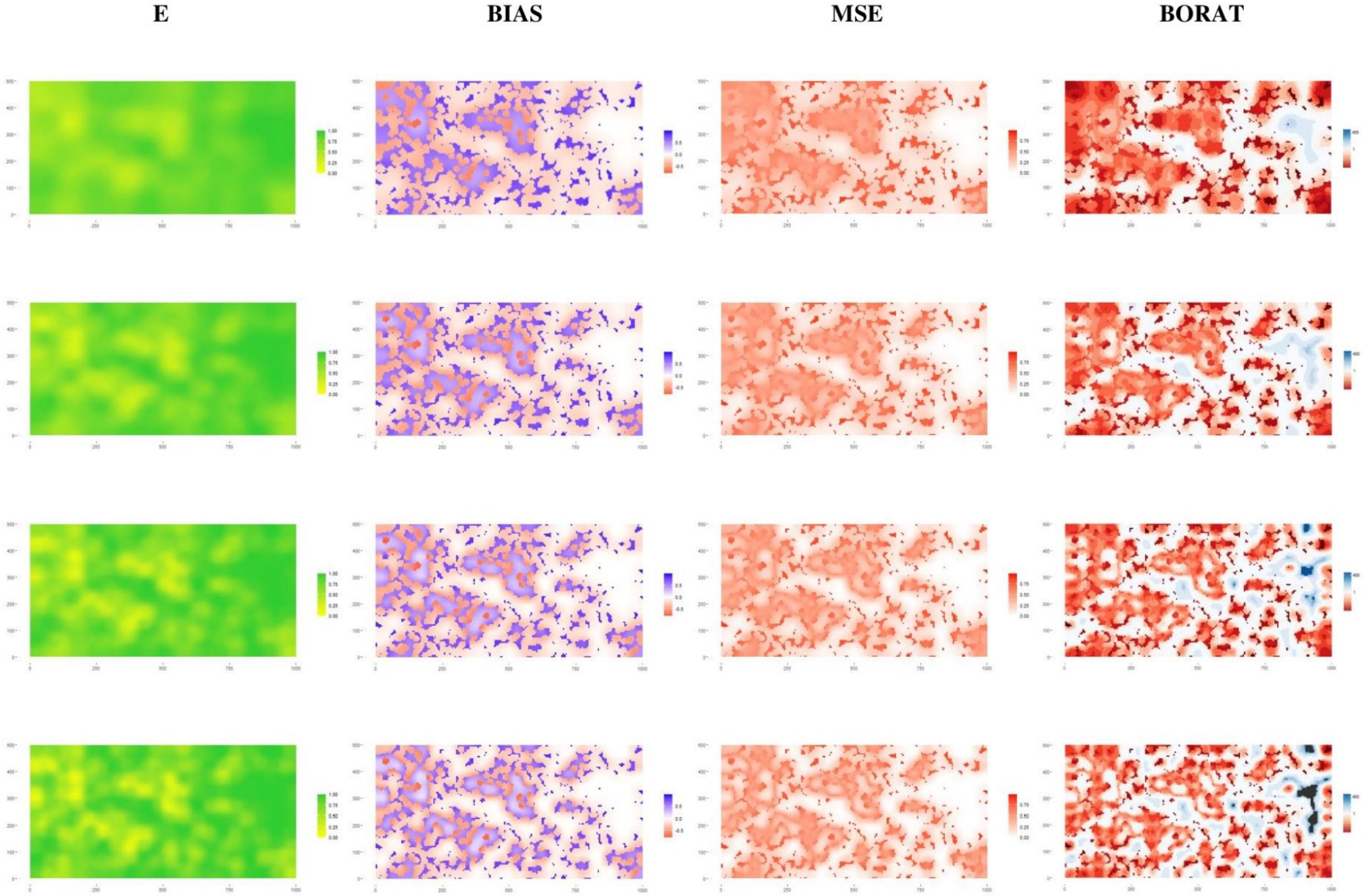


Figure 8. Maps of expectations (E), bias values (BIAS), mean squared errors (MSE) and bootstrap ratios (BORAT) achieved from the simulation performed on the presence surface of muskwood under TSS and $n = 50{,}100{,}150{,}200$ sample locations. Black patches denote locations where MSEs are equal to 0.

For the estimation of species association and species richness maps, the association surface of caimito de mono with muskwood was considered as the product of the presence surfaces of the two species and the richness surface was achieved as the sum of the presence surfaces of the $K = 302$ tree species settled on the study area (Figures 9-10).

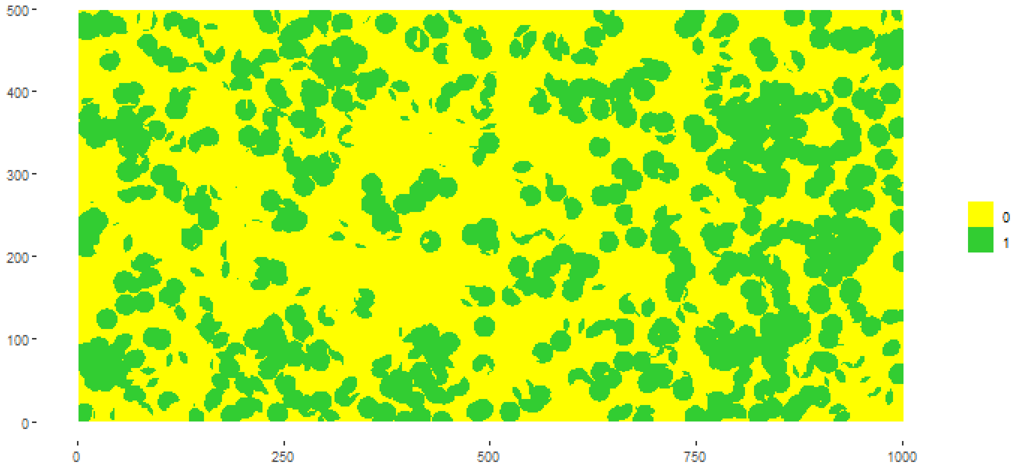


Figure 9. Association surface (green) of caimito de mono with muskwood resulting from 775 and 1993 trees, respectively, settled in BCI at the spatial grain of 13 m radius circular plots.

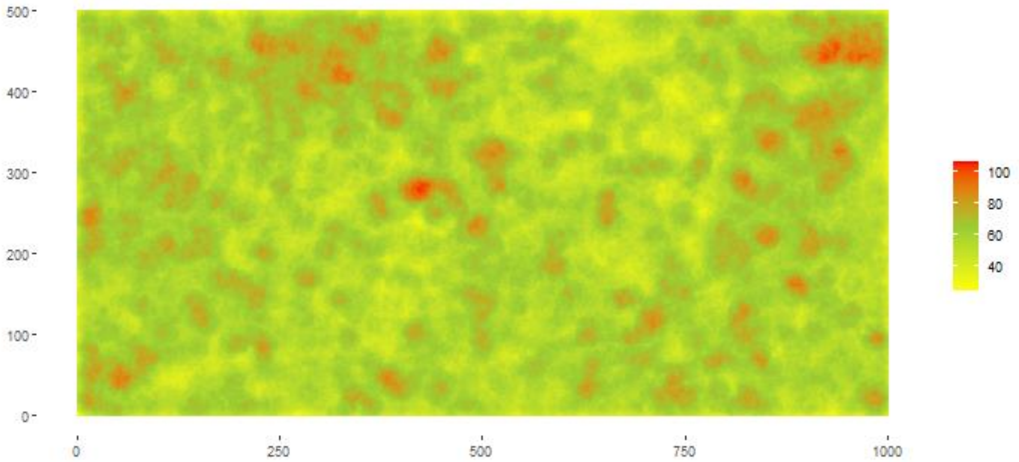


Figure 10. Richness surface achieved from the $K = 302$ tree species settled in BCI at the spatial grain of 13 m radius circular plots.

Regarding the estimation of the association surface of caimito de mono with muskwood, at each simulation run it was estimated for each $p \in G$ by means of equation (7) while the bootstrap mean squared error was computed by means of (9). Then, from the resulting Monte Carlo distributions, E, BIAS, MSE and BORAT were computed as in (12)-(15). For any combination of sampling scheme and sample size, Table 4 report the minima, averages and maxima of BIAS, MSE and BORAT, while Figures 11 shows the corresponding spatial patterns of these quantities, together with the spatial pattern of the expectations (first column) under TSS. Spatial patterns achieved under URS and SGS are quite similar and are not reported for brevity.

Table 4. Values of minima, averages and maxima for BIAS, MSE and BORAT achieved from the simulation performed on the association surface of caimito the mono with muskwood.

Scheme	n	BIAS			MSE			BORAT		
		min	mean	max	Min	mean	max	min	mean	max
URS	50	-0.87	0.00	0.76	0.12	0.43	0.87	0.19	0.59	1.40
	100	-0.89	0.00	0.86	0.05	0.41	0.89	0.14	0.65	2.15
	150	-0.91	0.00	0.89	0.02	0.39	0.91	0.13	0.70	3.60
	200	-0.91	0.00	0.91	0.01	0.38	0.91	0.12	0.75	4.46
TSS	50	-0.92	0.00	0.83	0.08	0.43	0.92	0.10	0.53	1.66
	100	-0.90	0.00	0.88	0.03	0.40	0.90	0.09	0.63	3.38
	150	-0.92	0.00	0.92	0.02	0.38	0.92	0.08	0.74	93.20
	200	-0.93	0.00	0.98	0.00	0.36	0.98	0.07	1.03	383.00
SGS	50	-0.92	0.00	0.86	0.02	0.42	0.92	0.11	0.52	5.35
	100	-0.91	0.00	0.90	0.00	0.40	0.91	0.07	0.67	42.62
	150	-0.95	0.00	0.96	0.00	0.38	0.96	0.08	0.86	315.33
	200	-0.94	0.00	0.95	0.00	0.36	0.95	0.04	0.89	67.21

Regarding the estimation of the richness surface of Figure 10, at each simulation run it was estimated for each $p \in G$ by means of equation (8). Moreover, because the richness surface was invariably positive, the bootstrap relative root mean squared errors were computed for each $p \in G$ by means of equation (11). Then, the expectation $e_L(p)$ was computed as in (12) and the relative bias (RBIAS)

$$rb_L(p) = \frac{e_L(p) - y_L(p)}{y_L(p)}$$

was adopted as an index of bias, while the relative root mean squared error (RRMSE)

$$rrmse_L(p) = \frac{\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{y}_{L,r}(p) - y_L(p)\right]^2\right\}^{1/2}}{y_L(p)}$$

was adopted as index of precision, together with the expectation of the bootstrap relative root mean squared error estimator (11)

$$ebrrmse_L(p) = \frac{1}{R}\sum_{r=1}^{R} \widehat{rrmse}^*_{B,r}(p)$$

from which the ratio to the true RRMSE was achieved (BORAT)

$$borat_L(p) = \frac{ebrrmse_L(p)}{rrmse_L(p)}$$

For any combination of sampling scheme and sample size, Table 5 report the minima, averages and maxima of RBIAS, RRMSE and BORAT, while Figures 12 shows the corresponding spatial patterns of these quantities, together with the spatial pattern of the expectations (first column) under TSS. Spatial patterns achieved under URS and SGS are quite similar and are not reported for brevity.

Table 5. Values of minima, averages and maxima for RBIAS, RRMSE and BORAT achieved from the simulation performed on the richness surface.

Scheme	n	RBIAS			RRMSE			BORAT		
		min	mean	max	min	mean	max	min	mean	max
URS	50	-0.41	0.03	1.04	0.10	0.20	1.12	0.13	0.76	1.13
	100	-0.38	0.02	1.07	0.09	0.19	1.14	0.13	0.79	1.22
	150	-0.35	0.02	1.09	0.08	0.18	1.16	0.13	0.82	1.29
	200	-0.32	0.02	1.07	0.07	0.18	1.15	0.12	0.83	1.31
TSS	50	-0.41	0.03	1.12	0.10	0.20	1.21	0.08	0.70	1.22
	100	-0.35	0.02	1.15	0.07	0.19	1.21	0.10	0.77	1.31
	150	-0.31	0.02	1.13	0.07	0.18	1.19	0.08	0.78	1.35
	200	-0.31	0.02	1.10	0.06	0.17	1.17	0.08	0.80	1.55
SGS	50	-0.41	0.03	1.10	0.08	0.20	1.16	0.06	0.67	1.37
	100	-0.33	0.03	1.15	0.07	0.19	1.20	0.10	0.73	1.81
	150	-0.31	0.02	1.15	0.06	0.17	1.21	0.08	0.76	1.86
	200	-0.30	0.02	1.11	0.05	0.16	1.18	0.09	0.79	1.84

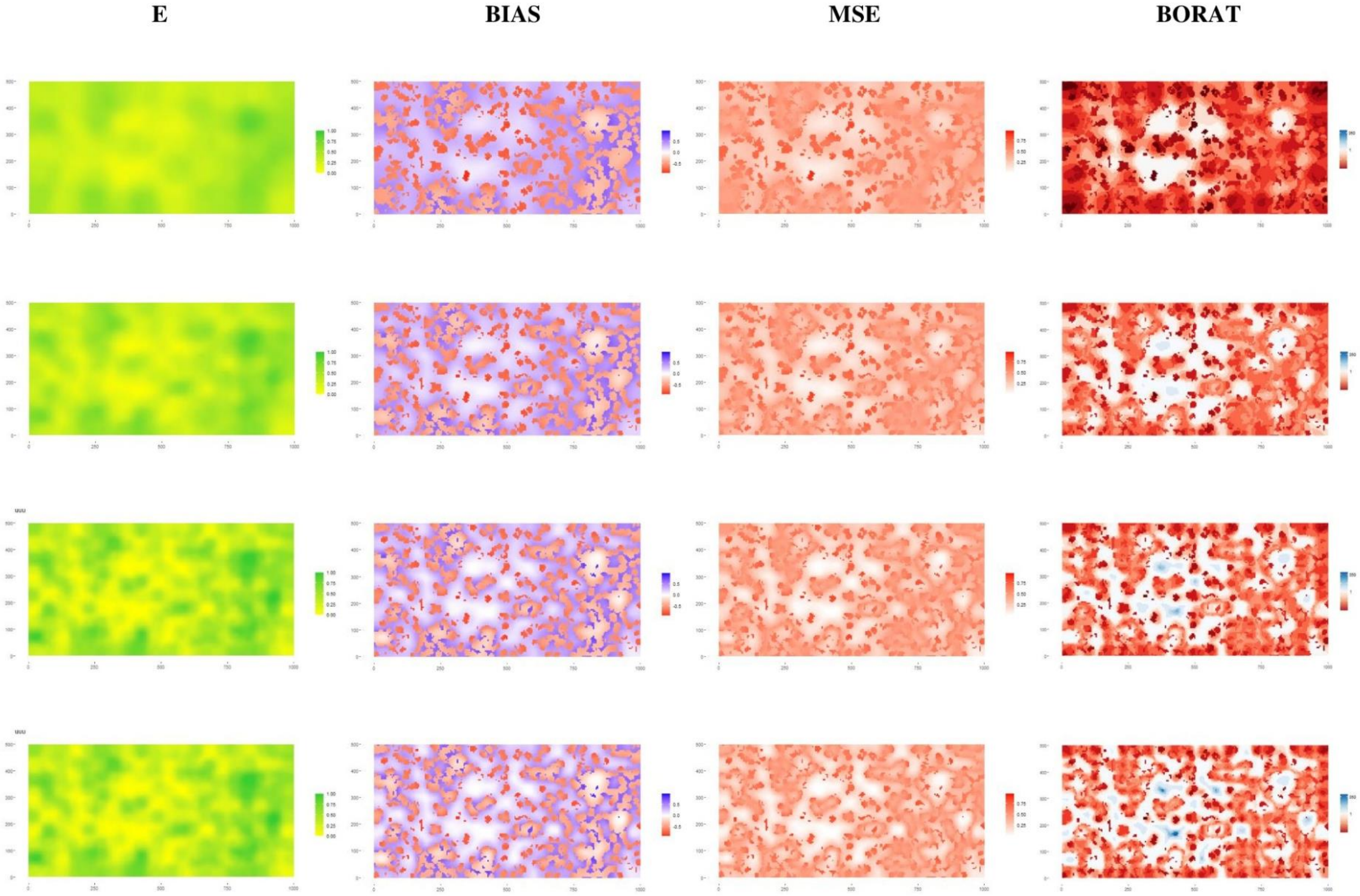


Figure 11. Maps of expectations (E), bias values (BIAS), mean squared errors (MSE) and bootstrap ratios (BORAT) achieved from the simulation performed on the association surface of caimito de mono with muskwood under TSS and $n = 50,100,150,200$ sample locations.

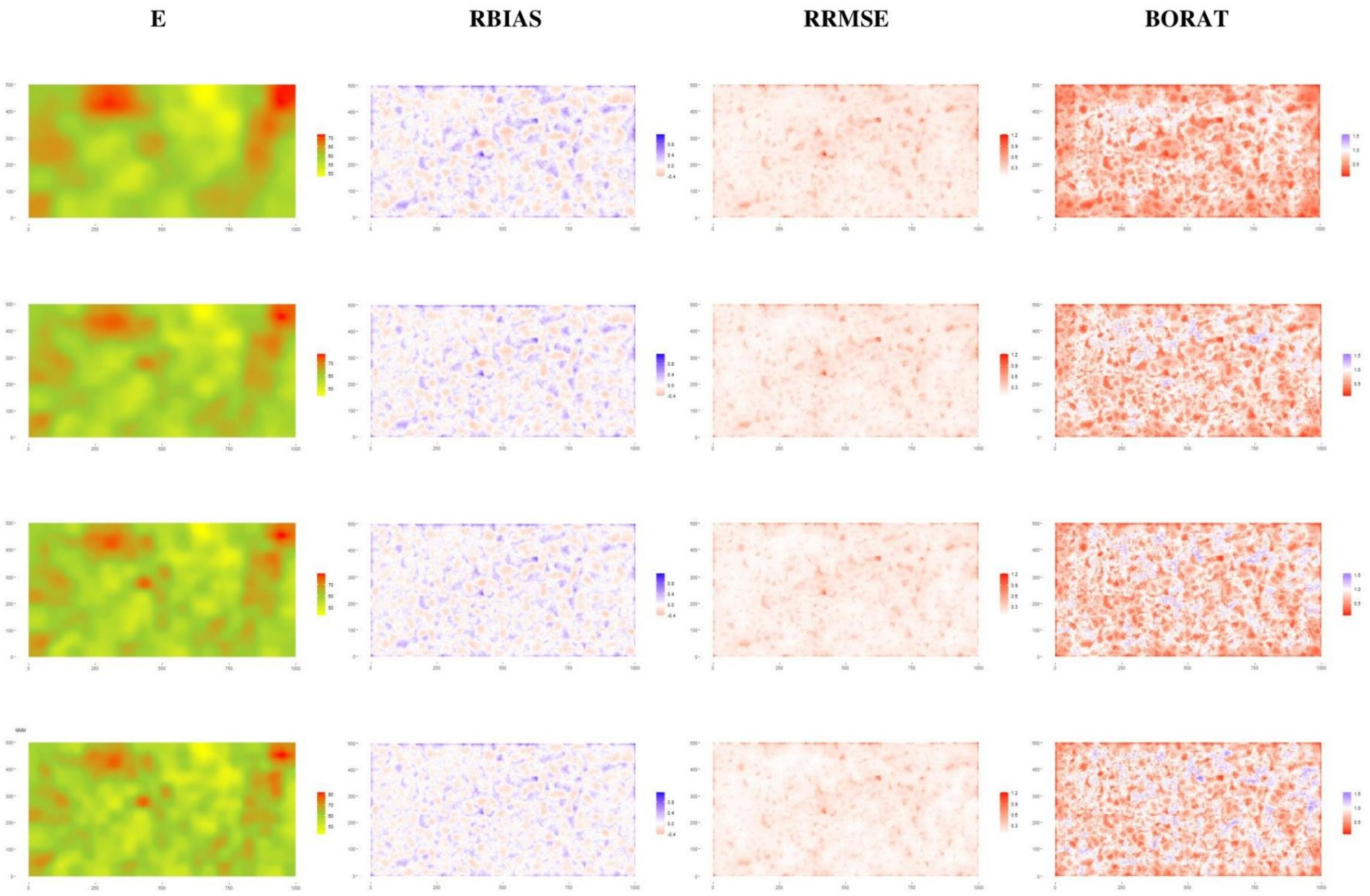


Figure 12. Maps of expectations (E), relative bias values (RBIAS), relative root mean squared errors (RRMSE) and bootstrap ratios (BORAT) achieved from the simulation performed on the richness surface under TSS and $n = 50,100,150,200$ sample locations.

Simulation results confirm the theoretical findings. Under TSS, the expected maps (first column of Figures 6-8, 11 and 12) approach the true maps (Figures 3-5, 9 and 10, respectively) as the number of sample locations increases, thus confirming the asymptotic unbiasedness and consistency of the NN interpolator. Similar results are achieved under URS and SGS. For the dichotomous surfaces, i.e., presence and association surfaces, asymptotic unbiasedness and consistency are also confirmed by the average values of MSEs (see Appendix A), that invariably decrease with the number of sample locations, even if in some cases the maxima of MSEs and the minima (negative) and maxima (positive) of bias values may show some increases as the number of sample locations increases. That is due to the greater number of locations near the discontinuity borders that obviously occur when sample locations become denser (see Tables 1-4). As to richness surfaces, consistency is confirmed by the convergence of the average maps (first column of Figure 12) to the true map (Figure 10) as well as by the averages of RRMSEs that decrease as the number of sample locations increases. Also in this case, some increases in the maxima of RRMSEs and in the minima and maxima of bias values occur for the same reasons argued before (see Table 5). Finally, regarding the bootstrap estimators of MSEs, the last column of Figures 3-5, 9 and 10, show their tendency to unbiasedness with white regions (those for which BORAT values are equal to 1) that become larger and larger. However, the high instability of these estimators, theoretically argued in Appendix C, is confirmed by the minima and maxima of BORATs, with minima near to 0 and maxima that reach some hundreds in several cases (see Tables 1-4). A better performance is achieved by the bootstrap RRMSEs, for which maxima are smaller than in the case of MSEs (see Table 5). As theoretically argued in Appendix C, that is due to the squared root that is present in (C.9) that mitigates the largest ratios, as well as to the smaller extents of inner zones (those far by discontinuity points) where estimation occurs with no error and BORAT denominators approach 0.

5. Case study

The NN interpolator was adopted to estimate the presence surfaces of holly oak (*Quercus ilex*) and white violet (*Viola alba Besser*) throughout the Montagnola Senese, a hilly area (up to 650 m) protected as a Site of Community Importance in Central Italy. Holly oak is the dominant tree species in the semi-natural forest that covers approximately 75% of the area, while white violet is a grass species of nature conservation importance.

The sample data used to perform estimation were collected from the last week of April to the first week of July 2007. The sampling design was that adopted in the 2000-2006 National Italian Forest Inventory, in which, in accordance with the TSS scheme, the Italian territory was covered by a grid of 1 km^2 square cells and a point was randomly selected within each cell (Fattorini et al., 2006). On the basis of this scheme, the study area was covered by $n = 106$ cells, and the points randomly selected within those cells during the inventory were adopted as sample points. Then, a quadrat plot with 10 m sides was centered at each sample location, and the presence of each species within plots was recorded (Chiarucci et al., 2008).

The NN interpolator was adopted to estimate the presence surfaces of both species, together with their association surface. The R function *polygrid* from the package GeoR (Ribeiro Jr and Diggle, 2001) was run to build a grid of 10,691 points within the study region where estimation is performed. Figures 13-15 report the estimated presence surfaces of the two species and their association surface together with the maps of their bootstrap mean squared errors achieved from the PPB procedure of section 3, performed at the 10,691 grid points.

Figure 13 shows the massive presence of holly oak throughout the study area, while Figure 14 shows the less wide presence of white violet. Regarding the association of the two species, Figure 15 shows high degree of association, explained by the fact that white violet is quite common in open forest stands as those present in the area, in which holly oak is common. Regarding the precision of maps, maps of bootstrap mean squared errors show that the greatest uncertainty occurs where changes from presence to absence occur.

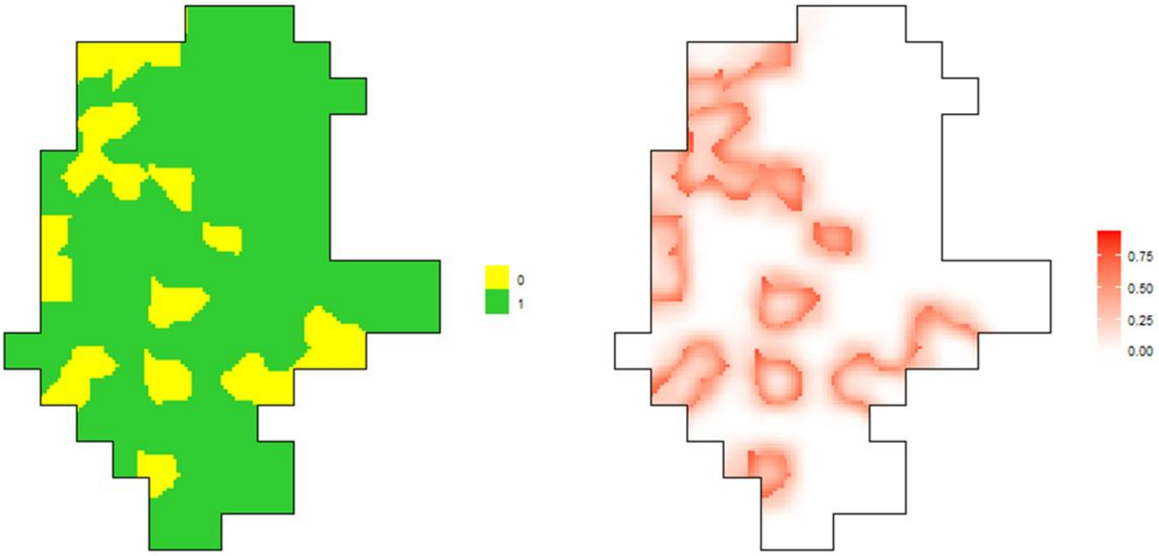


Figure 13. Estimate of presence surface of holly oak in the Montagnola Senese (left) and bootstrap mean squared errors (right).

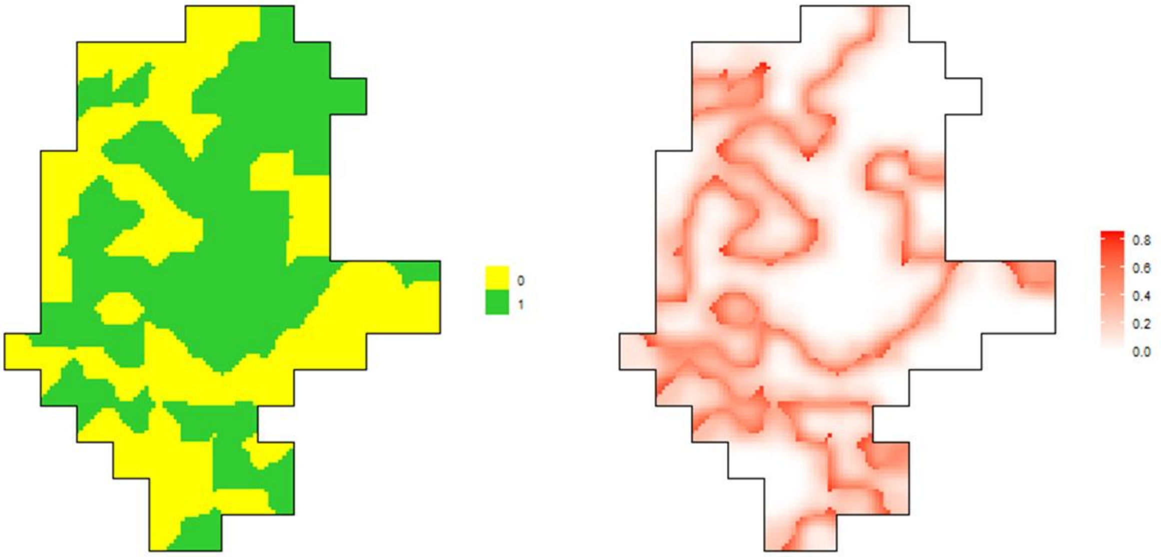


Figure 14. Estimate of presence surface of white violet in the Montagnola Senese (left) and bootstrap mean squared errors (right).

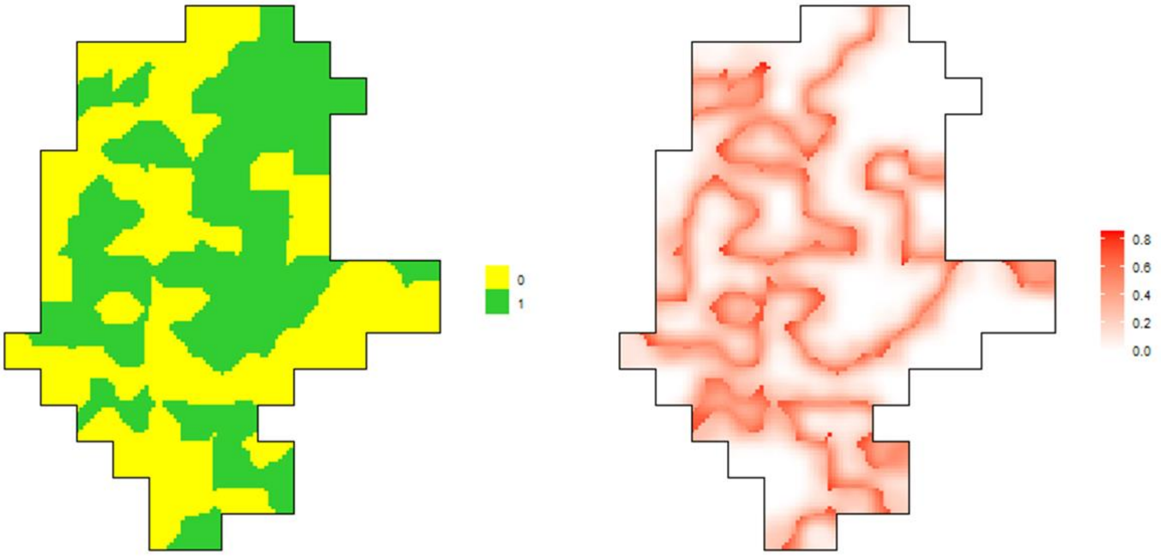


Figure 15. Estimate of association surface of holly oak and white violet in the Montagnola Senese (left) and bootstrap mean squared errors (right).

6. Final remarks

As pointed out by Fattorini et al. (2018, p.686) in their first attempt to perform mapping from a design-based perspective, the idea of making maps in this perspective is challenging because when estimating the value at a single location, either the location is sampled and there is no need for estimation, or the location is unsampled so that we have no information about it to perform estimation. Thus, the use of an assisting model to estimate unsampled values on the basis of some assumptions seems to be the sole way to fill the information vacancy. Obviously, as in any model-assisted, design-based framework, the assisting model is only adopted to determine the analytical form of the estimator, while its properties continue to be determined by the sampling scheme adopted to select locations (Di Biase et al. 2018).

In our scenario, it is quite obvious that the assisting model behind the NN interpolator is the simple, well-known Tobler's law of geography, that is, locations that are close in space tend to be more similar than locations that are far apart (Tobler, 1970). As any other model, the Tobler law may be useful but it is wrong (Box, 1979). Indeed, there may be several situations in which locations close in space may not be similar. For example, that may occur in some forest stands, under the presence of pairs of plants with different sizes at close proximity, in accordance with the so-called small-scale size diversity (e.g., Marcelli et al., 2019 and references therein). However, the great appeal of the design-based NN inference on species distribution is that even when the assisting Tobler's law is not suitable, consistency is ensured by the mathematical properties of presence, association and richness surfaces joined with the use of sampling schemes that provide asymptotic spatial balance, such as URS, TSS and SGS. Therefore, NN interpolation becomes perfect in any situation as the sampling intensity increases.

Another appealing characteristic of the design-based NN inference on species distribution is that in this case consistency does not necessitates the two basic assumptions that are behind the myriad of previous results achieved on this topic. As Chao and Colwell (2017, p.9) point out (see also Colwell et al., 2012), the presence/absence data adopted to make inference on species distribution, i.e. a matrix of 0's (absences) and 1's (presences) with as many columns as the sample locations and as many rows as the detected species, necessitate independence between rows and columns and identical distribution for data in the same row. Unfortunately, while the assumption of independence between columns and the equality in distribution within the same row may be ensured by the simple use of URS, i.e., the completely random and independent selection of sample locations (the requirement is only unsuitable because entails the use of URS while many other schemes could be more efficient), the independence between rows means independence between species detection in the same plot, an assumption that never holds in practice. Indeed, sampling species by plots constitutes a without-replacement selection that excludes independence among the sampled plants owing to the spatial association or exclusion of species that invariably occurs in any community. If this assumption was well delineated in its practical sense, it would alarm any ecologist and would sound like an oxymoron for any botanist familiar with the concept of inter-specific association. On the other hand, the assumption is introduced by means of equations such as

(1a) and (1b) in the paper by Chao and Colwell (2017), that are likely to sound obscure to any ecologist not well-trained in sampling. All these drawbacks are overcome by the use of design-based NN interpolation that does not necessitate the unrealistic independence among species detection at the same time allowing for the use of sampling schemes more efficient than URS.

Appendix A. Bias and precision quantification for dichotomous maps

Consider presence or association surfaces. Owing to their dichotomous nature, the expectation of the NN interpolator is given by

$$\mathrm{E}\{\hat{y}(p)\} = 0 \times \Pr\{\hat{y}(p) = 0\} + 1 \times \Pr\{\hat{y}(p) = 1\} = \Pr\{\hat{y}(p) = 1\}$$

Therefore, if $y(p) = 0$ the bias is given by

$$\mathrm{B}\{\hat{y}(p)\} = \mathrm{E}\{\hat{y}(p)\} - y(p) = \mathrm{E}\{\hat{y}(p)\} = \Pr\{\hat{y}(p) = 1\} = \Pr\{\hat{y}(p) \neq y(p)\}$$

while if $y(p) = 1$ the bias is given by

$$\mathrm{B}\{\hat{y}(p)\} = \mathrm{E}\{\hat{y}(p)\} - y(p) = \mathrm{E}\{\hat{y}(p)\} - 1 = \Pr\{\hat{y}(p) = 1\} - 1$$
$$= 1 - \Pr\{\hat{y}(p) = 0\} - 1 = -\Pr\{\hat{y}(p) = 0\} = -\Pr\{\hat{y}(p) \neq y(p)\}$$

in such a way that $|\mathrm{B}\{\hat{y}(p)\}| = \Pr\{\hat{y}(p) \neq y(p)\}$.

Moreover, owing to the dichotomous nature of presence and association surfaces, the absolute errors $|\hat{y}(p) - y(p)|$ and the squared errors $\{\hat{y}(p) - y(p)\}^2$ are also dichotomous, i.e. equal to 1 if an error occurs and equal to 0 otherwise. Therefore

$$\mathrm{E}\{|\hat{y}(p) - y(p)|\} = \mathrm{E}\{[\hat{y}(p) - y(p)]^2\} = \Pr\{\hat{y}(p) \neq y(p)\} \qquad (A.1)$$

Appendix B. Consistency results for presence, association and richness maps

For any $\delta > 0$ and for any location $p \in A$, denote by $B(p,\delta) = \{q : q \in A, \|p - q\| < \delta\}$ the δ-ball of p within A. Moreover, denote by $V(p,\delta) = \bigcap_{i=1}^{n}\{P_i \notin B(p,\delta)\}$ the event that a void, i.e. no sample location, occurs within the δ-ball of p.

Now consider presence or association surfaces. Owing to their dichotomous nature, the study area A is split into two sets, the set $D = \{p : p \in A, y(p) = 1\}$ where the surface is equal to 1 and its complement $D^c = \{p : p \in A, y(p) = 0\}$ where the surface is equal to 0. Then, if p is a continuity point of y, i.e. $p \in A \backslash \partial D$, let δ_p be the greatest value for which $B(p,\delta_p) \cap \partial D = \emptyset$. Obviously, if the event $V^c(p,\delta_p)$ occurs, the event $\{\hat{y}(p) = y(p)\}$ occurs, i.e. if at least one sample location falls within the δ_p-ball of p then the NN value is equal to $y(p)$. Therefore

$$\Pr\{\hat{y}(p) = y(p)\} \geq \Pr\{V^c(p,\delta_p)\} = 1 - \Pr\{V(p,\delta_p)\} \qquad (B.1)$$

Now, denote by a the size of A and by $a(p) \leq a$ the size of the δ_p-ball of p. Under URS, the probability that the i-th sample location falls outside the δ_p-ball of p is given by

$$\Pr\{P_i \notin B(p,\delta_p)\} = 1 - \frac{a(p)}{a} \quad , \qquad i = 1, \dots, n$$

in such a way that, owing to independence of sample locations under URS, the probability that no sample location falls within the δ_p-ball of p is given by

$$\Pr\{V(p,\delta_p)\} = \left\{1 - \frac{a(p)}{a}\right\}^n \qquad (B.2)$$

Then, substituting (B.2) into (B.1), under URS it holds that

$$\Pr\{\hat{y}(p) = y(p)\} \geq 1 - \left\{1 - \frac{a(p)}{a}\right\}^n \qquad (B.3)$$

or equivalently that

$$\Pr\{\hat{y}(p) \neq y(p)\} \leq \left\{1 - \frac{a(p)}{a}\right\}^n \qquad (B.4)$$

in such a way that $\lim_{n\to\infty} \Pr\{\hat{y}(p) \neq y(p)\} = 0$, i.e., the probability of missing the true value by NN interpolation approaches 0 as n increases. Therefore, because as proven in Appendix A, the error probability coincides with the absolute bias, the mean absolute error and the mean squared error, that proves the asymptotic unbiasedness and consistency of the NN interpolator of presence and association surface under URS at least at a c^n rate, with $c \in (0,1)$.

Similarly, if we consider richness surfaces, owing to their piecewise constant nature, the study area A is partitioned into K sets, $D_0, \dots, D_K$, where $D_k = \{p: p \in A, y_L(p) = k\}$ denotes the set where richness is equal to k ($k = 0,1, \dots, K$). Let $C = \cup_k \partial D_k$. If p is a continuity point of y_L, i.e. $p \in A \backslash C$, the greatest distance δ_p such that $B(p,\delta_p) \cap C = \emptyset$ defines the event $V^c(p,\delta_p)$ in such a way that if the event $V^c(p,\delta_p)$ occurs the event $\{\hat{y}_L(p) = y_L(p)\}$ occurs. Therefore, an inequality similar to (A.4) can be derived under URS also for the mean absolute errors of richness surfaces. In this case, if $y_L(p) = k$ and if $p \in A_k \backslash C$, it follows that

$$\mathrm{E}\{|\hat{y}_L(p) - y_L(p)|\} = \sum_{h \neq k = 1}^{K} |h - k| \Pr\{\hat{y}_L(p) = h\}$$

$$\leq \Pr\{\hat{y}_L(p) \neq y_L(p)\} \sum_{h \neq k = 1}^{K} |h - k| \leq const_k \times \left\{1 - \frac{a(p)}{a}\right\}^n$$

in such a way that $\lim_{n\to\infty} \mathrm{E}\{|\hat{y}_L(p) - y_L(p)|\} = 0$. The same conclusion holds for the mean squared error, with quadrats instead of absolute values. That prove the consistency under URS of the NN interpolator of richness surface at least a c^n rate, with $c \in (0,1)$.

Regarding consistency under TSS and SGS, for a sample size n, denote by $A_{1,n}, \dots, A_{n,n}$ the n patches of equal size $|A|/n$ partitioning A, and denote by $i(p)$ the label identifying the patch containing p. Suppose that as n increases the $A_{i,n}$s decrease in size in such a way that $\lim_{n\to\infty} \sup_{i=1,\dots,n} \operatorname{diam}(A_{i,n}) = 0$. Therefore, there exists a sample size n_0 such that, for each $n > n_0$ it holds that $A_{i(p),n} \subset B(p,\delta_p)$, in such a way that

$$\Pr\{\hat{y}(p) = y(p)\} \geq \Pr\{V^c(p,\delta_p)\}$$

$$\geq \Pr\{P_{i(p)} \in B(p,\delta_p)\} \geq \Pr\{P_{i(p)} \in A_{i(p),n}\} = 1$$

That obviously prove the consistency of the NN interpolator of presence and association surfaces at any continuity point, i.e. almost everywhere, and the same result holds, *mutatis mutandis*, for the richness surfaces.

Appendix C. Features of bootstrap estimators of precision indexes for presence, associations and richness maps.

Consider presence or association surfaces. Owing to their dichotomous nature, for each $p \in A$ the bootstrap means squared error (9) can be rewritten as

$$\widehat{mse}_B^*(p) = \frac{1}{B}\sum_{b=1}^{B}[\hat{y}_b^*(p) - \hat{y}(p)]^2 = \frac{1}{B}\sum_{b=1}^{B} I[\hat{y}_b^*(p) \neq \hat{y}(p)]$$

Accordingly, for B sufficiently large, owing to the strong law of large numbers and conditional to the original sample $P_1,\dots,P_n$, it holds that

$$\widehat{mse}_B^*(p) \sim E\{I[\hat{y}^*(p) \neq \hat{y}(p)] | P_1,\dots,P_n\}$$

where $\hat{y}^*(p)$ denotes the estimate of $y(p)$ occurred in a generic bootstrap resampling.

Now, in analogy with the notation adopted in Appendix B, denote by $\widehat{D} = \{p: p \in A, \hat{y}(p) = 1\}$ the set where the estimates are equal to 1 and by $\widehat{D}^c = \{p: p \in A, \hat{y}(p) = 0\}$ the set where the estimates are equal to 0. In practice, $\widehat{D}$ and $\widehat{D}^c$ are the sample counterparts of D and D^c, respectively. Obviously, $\widehat{D}$ can be rewritten as

$$\widehat{D} = \{p: p \in A, \hat{y}(p) = 1\} = \{p: p \in A, y(P_{NN(p)}) = 1\} = \{p: p \in A, P_{NN(p)} \in D\}$$

in such a way that

$$\widehat{mse}_B^*(p) \sim E\{I[\hat{y}^*(p) \neq \hat{y}(p)] | P_1,\dots,P_n\}$$

$$= I(P_{NN(p)} \in D) E\{I(P_{NN(p)}^* \in \widehat{D}^c) | P_1,\dots,P_n\} + I(P_{NN(p)} \in D^c) E\{I(P_{NN(p)}^* \in \widehat{D} | P_1,\dots,P_n)\}$$

$$= I(P_{NN(p)} \in D)\Pr\{P_{NN(p)}^* \in \widehat{D}^c | P_1,\dots,P_n\} + I(P_{NN(p)} \in D^c)\Pr\{P_{NN(p)}^* \in \widehat{D} | P_1,\dots,P_n\} \qquad (C.1)$$

where $P_{NN(p)}^*$ denotes the nearest neighbour of p occurred in a generic bootstrap resampling.

Then, if p is a continuity point of y, i.e. $p \in A \backslash \partial D$ and $p \in D$, i.e. $y(p) = 1$, from (C.1) and from the identity $I(P_{NN(p)} \in D) = 1 - I(P_{NN(p)} \in D^c)$ it follows that

$$\widehat{mse}_B^* \sim \Pr\{P_{NN(p)}^* \in \widehat{D}^c | P_1,\dots,P_n\} + (P_{NN(p)} \in D^c)[2\Pr\{P_{NN(p)}^* \in \widehat{D} | P_1,\dots,P_n\} - 1]$$

However, as stated in Appendix B, under URS, $\Pr\left(P_{NN(p)} \in D^c\right)$ quickly approaches 0 at a rate of at least c^n with $c \in (0,1)$, while under SGS and TSS, it is definitively equal to 0 for a sufficiently large n. Therefore, the random variable $I\left(P_{NN(p)} \in D^c\right)$ converges almost surely to 0, in such a way that

$$\widehat{mse}_B^*(p) \sim \Pr\left\{P_{NN(p)}^* \in \widehat{D}^c | P_1, \dots, P_n\right\}$$

Then, from (A.1), it holds that

$$\frac{\widehat{mse}_B^*(p)}{mse(p)} \sim \frac{\Pr\left\{P_{NN(p)}^* \in \widehat{D}^c | P_1, \dots, P_n\right\}}{\Pr\{\hat{y}(p) = 0\}} = \frac{\Pr\left\{P_{NN(p)}^* \in \widehat{D}^c | P_1, \dots, P_n\right\}}{\Pr\left\{P_{NN(p)} \in D^c\right\}} \qquad (C.2)$$

Similarly, but in a reversed way, if $p \in A \backslash \partial D$ and $p \in D^c$, i.e. $y(p) = 0$, it follows that

$$\frac{\widehat{mse}_B^*(p)}{mse(p)} \sim \frac{\Pr\left\{P_{NN(p)}^* \in \widehat{D} | P_1, \dots, P_n\right\}}{\Pr\left\{P_{NN(p)} \in D\right\}} \qquad (C.3)$$

The two relationships (C.2) and (C.3) can be rewritten in a unified way as

$$\frac{\widehat{mse}_B^*(p)}{mse(p)} \sim \frac{\Pr\{\hat{y}^*(p) \neq \hat{y}(p) | P_1, \dots, P_n\}}{\Pr\{\hat{y}(p) \neq y(p)\}} \qquad (C.4)$$

It should be noticed that the denominator of (C.4) may be 0 for n sufficiently large, as in the cases of TSS and SGS, and that (C.4) is the ratio of two quantities that approach 0 at rates at least of exponential nature and as such it may be very unstable especially for p near to ∂D. As a consequence, the conservative nature of $\widehat{mse}_B^*(p)$ cannot be proved in this case because the expectation

$$borat(p) = \frac{\mathrm{E}\{\widehat{mse}_B^*(p)\}}{mse(p)} \sim \frac{\mathrm{E}[\Pr\{\hat{y}^*(p) \neq \hat{y}(p) | P_1, \dots, P_n\}]}{\Pr\{\hat{y}(p) \neq y(p)\}}$$

does not admit any upper bound greater than one, as proven in Proposition 3 by Fattorini et al. (2021) in the case of surfaces that are differentiable at p with non-null derivatives.

Similarly, in the case of richness surfaces, for B sufficiently large, and owing to the strong law of large numbers, conditional to the original sample $P_1, \dots, P_n$, it holds that

$$\widehat{mse}_B^*(p) = \frac{1}{B} \sum_{b=1}^{B} \left[\hat{y}_{L,b}^*(p) - \hat{y}_L(p)\right]^2 \sim \mathrm{E}\{[\hat{y}_L^*(p) - \hat{y}_L(p)]^2 | P_1, \dots, P_n\}$$

where $\hat{y}_L^*(p)$ denotes the estimate of $y_L(p)$ occurred in a generic bootstrap resampling.

Then, in analogy with the notation adopted in Appendix B, denote by $\widehat{D}_k = \{p : p \in A, \hat{y}_L(p) = k\}$ the set where the richness estimates are equal to k, with $k = 0,1, \dots, K$. In practice, the $\widehat{D}_k$s are the sample counterparts of the D_k. Obviously, each $\widehat{D}_k$ can be rewritten as

$$\widehat{D}_k = \{p : p \in A, \hat{y}_L(p) = k\} = \left\{p : p \in A, y_L\left(P_{NN(p)}\right) = k\right\} = \left\{p : p \in A, P_{NN(p)} \in D_k\right\}$$

in such a way that

$$\widehat{mse}_B^*(p) \sim \mathrm{E}\{[\hat{y}_L^*(p) - \hat{y}_L(p)]^2 | P_1, \dots, P_n\}$$

$$= \sum_{k=0}^{K} I\big(P_{NN(p)} \in D_k\big) \sum_{h \neq k=0}^{K} (h-k)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\} \qquad (C.5)$$

Then, if p is a continuity point of y_L, i.e. $p \in A \backslash C$, and $p \in D_{k_0}$, i.e. $y_L(p) = k_0$, from (C.5) and from the identity that $I\big(P_{NN(p)} \in D_{k_0}\big) = 1 - I\big(P_{NN(p)} \in D^c_{k_0}\big)$ it follows that

$$\widehat{mse}^*_B(p) \sim \sum_{h \neq k_0=0}^{K} (h-k_0)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\}$$

$$-I\big(P_{NN(p)} \in D^c_{k_0}\big) \sum_{h \neq k_0=0}^{K} (h-k_0)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\} +$$

$$\sum_{k \neq k_0=0}^{K} I\big(P_{NN(p)} \in D_k\big) \sum_{h \neq k=0}^{K} (h-k)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\} \qquad (C.6)$$

Once again, as stated in Appendix B, under URS, $\Pr\big(P_{NN(p)} \in D^c_{k_0}\big)$ quickly approaches 0 at a rate of at least c^n with $c \in (0{,}1)$, while under SGS and TSS, it is definitively equal to 0 for a sufficiently large n. Therefore, the random variable $I\big(P_{NN(p)} \in D^c_{k_0}\big)$ converges almost surely to 0, and, *a fortiori*, each $I\big(P_{NN(p)} \in D_k\big)$ for $k \neq k_0$ converges almost surely to 0 Then, from (C.6) it holds that

$$\widehat{mse}^*_B(p) \sim \sum_{h \neq k_0=0}^{K} (h-k_0)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\}$$

in such a way that

$$\frac{\widehat{mse}^*_B(p)}{mse(p)} = \frac{\sum_{h \neq k_0=0}^{K} (h-k_0)^2 \Pr\big\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\big\}}{\sum_{h \neq k_0=0}^{K} (h-k_0)^2 \Pr\big\{P_{NN(p)} \in D_h\big\}} \qquad (C.7)$$

Also in this case, in accordance with Appendix B, (C.7) is the ratio of two quantities that approach 0 at rates at least of exponential nature and as such it may be very unstable especially for p near to ∂C. The same considerations also hold for the ratio

$$\frac{\widehat{rrmse}^*_B(p)}{rrmse(p)} = \frac{y_L(p)}{\hat{y}_L(p)} \sqrt{\frac{\widehat{mse}^*_B(p)}{mse(p)}} \qquad (C.8)$$

because the first ratio in the right side of (C.8) approaches 1 owing to the consistency of $\hat{y}_L(p)$, in such a way that (C.8) is asymptotically equivalent to the squared root of (C.7). Therefore, owing to the volatility of (C.8), as in the case of dichotomous surfaces, the expectation

$$borat(p) = \frac{\mathrm{E}\{\widehat{rrmse}^*_B(p)\}}{rrmse(p)} \sim E\left\{\sqrt{\frac{\widehat{mse}^*_B(p)}{mse(p)}}\right\} \qquad (C.9)$$

does not admit any upper bound greater than one. Anyway, BORAT values achieved with RRMSEs are likely to be more stable than those achieved with MSEs. Indeed, the squared root that is present in (C.9) mitigates the largest MSE ratios. Moreover, richness surfaces are likely to be more fragmented than dichotomous presence and association surfaces, with more discontinuity borders and smaller extents of inner zones (those far by discontinuity points) where estimation is precise, true MSEs approach 0 and MSEs ratios approach infinity.

Acknowledgements

The authors wish to thank Luca Pratelli from Naval Academy (Livorno, Italy) for its suggestions about the theoretical issues of the paper and Alessandro Chiarucci from Alma Mater Studiorum University of Bologna (Italy) for providing the data of the case study. The authors also acknowledge research support from the Research and Innovation Centre Fondazione Edmund Mach (San Michele all' Adige, TN). This paper was carried out in the frame of the subproject "Precision Forestry" (AgriDigit program) (DM 36503.7305.2018 of December 2, 2018) funded by the Italian Ministry of Agricultural, Food, and Forestry Policies (MiPAAF).

References

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157: 101–118.

Barabesi L (2003) A Monte Carlo integration approach to Horvitz-Thompson estimation in replicated environmental designs. Metron LXI: 355–374.

Box GEP (1979) Robustness in Statistics. London: Academic Press.

Champ PA, Boyle K, Brown TC (Eds) (2017) A Primer on Nonmarket Valuation (2nd ed.). New York: Springer.

Chao A, Colwell RK (2017) Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. SORT 41:3–54.

Chiarucci A, Bacaro G, Rocchini D (2008) Quantifying plant species diversity in a Natura 2000 network: Old ideas and new proposals. Biological Conservation 141: 2608–2615.

Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, Longino JT (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3–21.

Cressie N (1991) Statistics for Spatial Data. New York: Wiley.

Di Biase RM, Fattorini L, Marchi M (2018) Statistical inferential techniques for approaching forest mapping. A review of methods. Annals of Silvicultural Research 42: 46-58.

Edwards TC, Cutler DR, Zimmermann NE, Geiser L, Moisen GG (2006) Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling 199:132–141.

Fattorini L, Marcheselli M, Pisani C (2006) A three-phase sampling strategy for large-scale multiresources forest inventories. Journal of Agricoltural, Biological and Environmental Statistics 11: 296–316.

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

Fattorini L, Marcheselli M, Pisani C, Pratelli L (2021) Design-based properties of the nearest neighbour spatial interpolator and its bootstrap mean squared error estimator. Under review.

Franklin J (1995) Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19: 474–499.

Franklin J (2010) Mapping Species Distributions. Cambridge (UK): Cambridge University Press.

Gibson LA, Wilson BA, Cahill DM, Hill J (2004) Modelling habitat suitability of the swamp antechinus (*Antechinus minimus maritimus*) in the coastal heathlands of southern Victoria, Australia. Biological Conservation 117: 143–150.

Gregoire TG (1998) Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research 28: 1429–1447.

Gregoire TG, Valentine HT (2008). Sampling Strategies for Natural Resources and the Environment. Boca Raton (FL): Chapman & Hall/CRC.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186.

Guisan A, Lehmann A, Ferrier S, Austin M, Overton JMCC, Aspinall R, Hastie T (2006) Making better biogeographical predictions of species distributions. Journal of Applied Ecology 43: 386–392.

Hirzel A, Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modelling? Ecological Modelling 157: 331–341.

Hirzel A, Hausser J, Chessel D, Perrin N (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83: 2027–2036.

Lenihan JM (1993) Ecological response surfaces for North American boreal tree species and their use in forest classification. Journal of Vegetation Science 4: 667–680.

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

Little RJ (2004) To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association 99: 546–556.

Marcelli A, Corona P, Fattorini L (2019) Design-based estimation of mark variograms in forest ecosystem surveys. Spatial Statistics 30: 27–38.

Mc Roberts RE, Cohen WB, Naesset E, Stehman SV, Tomppo EO (2010) Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scandinavian Journal of Forest Research 25: 340–367.

Palmer MW (1990) The estimation of species richness by extrapolation. Ecology 71: 1195–1198.

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

Ribeiro Jr PJ, Diggle PJ (2001) geoR: A package for geostatistical analysis. R News 1: 15–18.

Rodriguéz JP, Brotons L, Bustmante J, Seoane J (2007) The application of predictive modelling of species distribution to biodiversity conservation. Diversity and Distributions 13: 243–251.

Rotenberry JT, Preston KL, Knick ST (2006) GIS-based niche modeling for mapping species habitat. Ecology 87: 1458–1464.

Rushton SP, Ormerod SJ, Kerby G (2004) New paradigms for modelling species distributions? Journal of Applied Ecology 41: 193–200.

Särndal CE, Swensson B, Wretman J (1992) Model Assisted Survey Sampling, New York: Springer.

Schreuder HT, Gregoire TG, Wood GB (1993) Sampling Methods for Multiresource Forest Inventory. New York: Wiley.

Scott JM, Heglund PJ, Morrison ML et al. (Eds) (2002) Predicting Species Occurrences: Issues of Accuracy and Scale. Covelo (CA): Island Press.

Smith TMF (1994) Sample surveys 1975-1990; an age of reconciliation? International Statistical Review 62: 5–34.

Smith TMF (2001) Biometrika Centenary: Sample Surveys. Biometrika 88: 67–13.

Thompson SK (2002) Sampling (2nd ed.). New York: Wiley.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit Region. Economic Geography 46: 234–240.

Turner MG, Gardner RH, O'Neill RV (2001) Landscape Ecology in Theory and Practice. New York: Springer-Verlag.

Van Horne B (1983) Density is a misleading indicator of habitat quality. Journal of Wildlife Management 47: 893–901.