TWO-PHASE SAMPLING STRATEGIES FOR DESIGN-BASED MAPPING OF CONTINUOUS SPATIAL POPULATIONS IN ENVIRONMENTAL SURVEYS

BY LORENZO FATTORINI*, SARA FRANCESCHI*, MARZIA MARCHESELLI*, CATERINA PISANI* AND LUCA PRATELLI**

University of Siena* and Naval Academy**

The estimation of a surface throughout a continuum of points in a study area is addressed by means of two-phase sampling strategies. To this aim, a family of two-phase inverse distance weighting interpolators is introduced and their design-based asymptotic properties are derived when the surface remains fixed and the number of sample points approaches infinity. In particular, conditions ensuring asymptotic unbiasedness and consistency are derived and are proven to hold for some of the most widely applied environmental sampling schemes. Furthermore, a computationally simple mean squared error estimator is proposed. Finally, a simulation study is performed to assess the theoretical results. The proposed strategy is adopted to provide the map of basal area in a forested region of Casentino Valley (Central Italy).

1. Introduction.

Comprehensive information about spatial phenomena is needed for effectual management of natural, social and economic resources and therefore mapping their patterns is crucial. In the framework of environmental surveys, for instance, accurate and updated maps of species distribution and diversity measures are essential in zoology and vegetation sciences, of soil composition or pollution and mineral concentration in soil sciences, of presence and amount of forest attributes in forest sciences. Generally, the aim is the mapping of a surface, that is the value of the survey variable at any point of the study region, which is a continuum. Obviously, the value of the survey variable cannot be recorded at any point but only for a sample of points and the aim is the estimation of the value of the survey variable at any point.

Surface estimation is usually addressed in a model-based approach: the surface is supposed to be generated by a super-population probability model and uncertainty arises from the assumed model (Gregoire, 1998). Kriging prediction is probably the most widely applied model-based technique, in which, in the most general case, a deterministic trend component plus a

Keywords and phrases: design consistency, two-phase sampling, spatial interpolation.
zero-mean intrinsically stationary random process is supposed to generate the value of the survey variable at any point (e.g. Cressie, 1993).

On the other hand, if surface estimation is addressed in a design-based approach, the surface is considered as a fixed attribute, thus avoiding any model assumption, and uncertainty only stems from the sampling scheme adopted to select the points. It is worth noting that, when estimating the value at a single point, either the point is sampled and the value of the survey variable at the point is known or the point is unsampled and no information about it is available to perform estimation. In these cases, the use of an assisting model is unavoidable. Design-based surface mapping has been recently approached by Fattorini et al. (2018a), who have exploited the Tobler’s first law of geography (Tobler, 1970) as assisting model, i.e. points that are close in space tend to have more similar values of the survey variable than points that are far apart. In particular, Fattorini et al. (2018a) adopt the so-called inverse distance weighting (IDW) interpolator, in which the value at an unsampled point is estimated by a weighted sum of the sampled values with weights inversely decreasing with distances to the point. In this sense the IDW interpolator is similar to kriging exploiting appropriate correlogram models, that also results in interpolators with weights decreasing with distances from the interpolated points. Moreover, the authors give conditions ensuring unbiasedness and consistency of the IDW interpolator as the surface remains fixed and the number of sample points approaches infinity.

In many environmental studies, especially when the size of the study region is large, sample surveys are performed in two-phases. Usually, the first phase of sampling is performed on screen adopting satellite imagery or orthophotos available for the whole study region. A large sample of points is selected by means of suitable sampling schemes able to ensure an even spread of the sampled points throughout the study region. Some variables, such as length of the tree-rows or size of the woodlots spotted by the sampled points can be measured on screen without any field work (e.g. Baffetta et al., 2011). On the other hand, some variables such as volumes or basal areas of the trees lying within the plots centred at the sampled points require these points to be visited. Usually, the visit of the whole first-phase sample of points is too demanding in terms of time and resources. Then, as customary in large-scale forest inventories, a second-phase sample is selected from the finite population of the first-phase points and only the second-phase sample of points is visited on the ground. While the design-based features such as unbiasedness, precision and consistency of two-phase estimators of population totals and averages are deeply investigated in literature (Man-
dallaz, 2008; Fattorini, 2015; Fattorini et al., 2017), the two-phase surface estimation has never been approached from a design-based perspective.

This lack of methodology has unfortunately precluded mapping of survey variables in several investigations we have been involved. No design-based mapping has been performed to depict forest carbon amount throughout the Trentino administrative district (Northern Italy) after the two-phase 2001 survey performed by Centro di Ecologia Alpina and the subsequent design-based estimation of the total quantity of carbon (Fattorini et al., 2009). No design-based map is provided for the forest attributes surveyed during the 2000-2006 Italian Forest Inventory (Fattorini et al., 2006 and www.ifnc.it), and no design-based map is provided for the forest aboveground biomass after the two-phase inventory performed in the southwestern part of Molise Region (Central Italy) and the subsequent model-assisted estimation of the total biomass (Chirici et al., 2016). In some of these cases, mapping has been approached by model-based techniques, ignoring the sampling scheme. For example, Chirici et al. (2020) adopt a random-forest inputation technique for mapping forest attributes from the Italian Forest Inventory, while Chirici et al. (2016) adopt a linear regression model exploiting LiDAR metrics as co-variates for mapping aboveground biomass from southwestern Molise survey. However, because in environmental and forest surveys estimation of totals and averages is traditionally performed from a design-based perspective (e.g. Gregoire and Valentine, 2008), it would be consistent to provide maps and to estimate their precision from the same perspective.

In this paper, design-based surface estimation is addressed when two-phase sampling schemes are adopted by introducing a family of two-phase IDW interpolators. Their asymptotic properties are investigated and conditions ensuring asymptotic unbiasedness and consistency are derived when the surface remains fixed and the number of first- and second-phase sample points approaches infinity. Moreover, the resulting conditions are proven to hold for some sampling schemes widely applied in environmental and forest surveys. Finally, an asymptotically design-based conservative estimator of mean squared errors is proposed. The theoretical findings are confirmed by a simulation study performed on artificial and real populations. An application is considered regarding a two-phase inventory performed in a forested region of Central Italy, a case study in which, after design-based inference on species abundance and ecological diversity (Corona et al., 2019), any design-based mapping of forest attributes remains precluded until the results of this work. Technical details and proofs are reported in the Appendix.
2. Notation and Setting. Consider a study region \( A \) that is assumed to be a compact set of \( \mathbb{R}^2 \) and let \( y \) be a measurable bounded function defined on \( A \). Without loss of generality, we suppose \( y \) with values on \([0, L]\) and \( y(p) \) is the value or the density of the survey variable at any point \( p \in A \). As the study region is a continuum, the aim is, at least in principle, the estimation of \( y(p) \) for each \( p \in A \), that is the estimation of the \( y \)-surface over \( A \). For example, in forest studies, attributes (e.g. volume or basal area) are concentrated at tree locations and are absent elsewhere. It is therefore suitable to consider the amount or the density of a forest attribute at a prefixed spatial grain, i.e., within a squared or circular plot of prefixed size centred at the points of the study region. Therefore, for any point \( p \in A \) there exists the amount or the density of the forest attribute within the plot of prefixed size centred at \( p \), giving rise to a surface throughout \( A \). For the first time, estimation of surfaces will be addressed in a design-based setting by using two-phase sampling designs which are very common in environmental and forest surveys when the study region is large.

2.1. First-phase Sampling Strategy. Let \( P_1, \ldots, P_n \) be \( n \) random variables that represent the locations of a large sample of points selected from \( A \) by means of a fixed-size sampling design able to evenly cover the study region. For example, in forest surveys it is customary to partition the study area into \( n \) sets of equal size and then selecting one point in each set. If the whole set of sample points is visited and the values of the survey variable \( y(P_1), \ldots, y(P_n) \) are recorded, the estimation of \( y(p) \) can be performed in a design-based framework adopting the IDW interpolator

\[
\hat{y}(p) = I(C_p)y(p) + I(C_p^c)\frac{\sum_{i=1}^{n} \varphi(\|p - P_i\|)y(P_i)}{\sum_{h=1}^{n} \varphi(\|p - P_h\|)}
\]

where \( I(E) \) is the indicator function of the event \( E, C_p = \cup_{i=1}^{n}\{P_i = p\} \), \( \varphi \) is a positive, continuous and decreasing function on \((0, \infty)\) with \( \varphi(0) = 0 \) and \( \|\cdot\| \) is a norm in \( \mathbb{R}^2 \). In accordance with Cordy (1993), if the existence of the continuous probability density function of \( P = (P_1, \ldots, P_n) \) is assumed, \( \Pr(\cup_{i=1}^{n}\{P_i = p\}) = 0 \) and the IDW interpolator is almost everywhere equal to

\[
\hat{y}(p) = \sum_{i=1}^{n} w_i(p)y(P_i)
\]

that is a weighted sum of the values at the sample points, with weights

\[
w_i(p) = \frac{\varphi(\|p - P_i\|)}{\sum_{h=1}^{n} \varphi(\|p - P_h\|)}.
\]
inversely decreasing with distances and summing to 1.

The use of the IDW interpolator has been proposed by Fattorini et al. (2018a). It constitutes the continuous counterpart of the IDW interpolator adopted by Bruno et al. (2013) for discrete populations of spatial units, subsequently investigated from a design-based perspective by Fattorini et al. (2018b). Fattorini et al. (2018a) derive conditions ensuring unbiasedness and consistency of (2.1) when the surface remains fixed while the number of sampled points approaches infinity. Moreover, the authors prove these conditions to hold for some sampling schemes widely used in environmental surveys.

When the sampling intensity is high and the study region is large, visiting the whole sample of points is too demanding. Usually, these points are examined on screen, recording qualitative or quantitative attributes that can be determined without field work, while the ground mensuration of the interest attribute is performed only for a portion of points selected in a second phase of sampling. Accordingly, (2.1) is actually unknown and, conditioned on the first-phase points, it constitutes a finite population total that must be estimated from the second-phase sample.

2.2. Second-phase Sampling Strategy. Let \( S \) be the second-phase sample of \( m \) points selected from the first-phase sample \( P \) of \( n \) points by means of a suitable finite population sampling design. For example, in some environmental surveys the set of first-phase points is partitioned into \( m \) subsets of neighbouring points and a point is selected within each subset (Fattorini et al., 2017). Alternatively in forest inventories it is customary to stratify the set of first-phase points in accordance with some of their characteristics readily determined on screen and then to select a sample from each stratum (Fattorini et al., 2006, Fattorini, 2015).

Conditioning on \( P \), denote by

\[
\pi_i(P) = \mathbb{E}[I(i \in S)|P]
\]

for any \( i = 1, \ldots, n \) and by

\[
\pi_{ij}(P) = \mathbb{E}[I(i, j \in S)|P]
\]

for any \( j > i = 1, \ldots, n \) the first- and second-order inclusion probabilities induced by the second-phase sampling design. Then, it is natural to apply the well-known Horvitz-Thompson (HT) criterion to estimate (2.1) as a population total, giving rise to the two-phase HT interpolator
(2.4) 
\[ \hat{y}_{HT}(p, S) = \sum_{i \in S} w_i(p)y(P_i)\pi_i(P)^{-1} = \frac{\sum_{i \in S} \varphi(||p - P_i||)y(P_i)\pi_i(P)^{-1}}{\sum_{h=1}^{n} \varphi(||p - P_h||)}. \]

However, in order to improve to precision of (2.4), one can take into account that the total of \(w_i\)'s over the population \(P\) invariably gives 1, in such a way the weights can be exploited as auxiliary variable to calibrate (2.4) in accordance with the well-known ratio criterion (e.g. Särndal et al., 1992, section 5.6). Therefore, dividing (2.4) by the HT estimator of the weights sum, the two-phase ratio interpolator reduces to
(2.5) 
\[ \hat{y}_{R}(p, S) = \frac{\sum_{i \in S} w_i(p)y(P_i)\pi_i(P)^{-1}}{\sum_{i \in S} w_i(p)\pi_i(P)^{-1}} = \frac{\sum_{i \in S} \varphi(||p - P_i||)y(P_i)\pi_i(P)^{-1}}{\sum_{i \in S} \varphi(||p - P_i||)\pi_i(P)^{-1}}. \]

Finally, we can bypass the estimation of (2.1) directly applying the IDW interpolator to the second-phase sample \(S\), giving rise to the two-phase direct interpolator
(2.6) 
\[ \hat{y}_{D}(p, S) = \sum_{i \in S} w_i^*(p)y(P_i) = \frac{\sum_{i \in S} \varphi(||p - P_i||)y(P_i)}{\sum_{h \in S} \varphi(||p - P_h||)} \]

where for each \(i \in S\)
\[ w_i^*(p) = \frac{\varphi(||p - P_i||)}{\sum_{h \in S} \varphi(||p - P_h||)} \]

are the second-phase weights inversely decreasing with distances and summing to 1.

It should be pointed out that the direct and ratio interpolators are members of a more general family of two-phase IDW interpolators given by
(2.7) 
\[ \hat{y}(p, S) = \frac{\sum_{i=1}^{n} \varphi(||p - P_i||)y(P_i)G_i}{\sum_{h=1}^{n} \varphi(||p - P_h||)G_h} \]

where \(G_i\) is a dichotomous not negative random variable equal to 0 if the \(i\)-th first-phase point is not selected in the second phase and such that
\[ I(G_i > 0) \leq G_i I(G_i > 0) \leq \gamma_0 I(G_i > 0). \]

In particular, for the direct interpolator (2.6), \(G_i = I(i \in S)\) with \(\gamma_0 = 1\), while for the ratio interpolator (2.5), \(G_i = I(i \in S)\pi_i(P)^{-1}\) with \(\gamma_0 = \)
1/\min_{i=1,...,n} \pi_i(P)$. It should be also pointed out that (2.7) can be rewritten as a weighted sum of the first-phase sample observations with weights

\[ w_i(p) = \frac{\varphi(||p - P_i||)G_i}{\sum_{h \in S} \varphi(||p - P_h||)G_h} \]

inversely decreasing with distances and summing to 1. As such, the two-phase HT interpolator (2.4) does not belong to the family. Finally, it should be pointed out that if the first-order inclusion probabilities \( \pi_i(P) \) are equal for all \( i = 1, \ldots, n \), the interpolators (2.5) and (2.6) coincide.

Design-based expectation and variance of the two-phase interpolators cannot be expressed in closed form, giving no insights about their bias and precision. Therefore, conditions providing asymptotic design-based unbiasedness and consistency are needed in order to render statistically sound the mapping of environmental and forest attributes achieved from two-phase samples.

3. Asymptotic results. Asymptotic unbiasedness and consistency of the two-phase interpolators will be achieved supposing the first-phase sample size \( n \) to increase indefinitely \( (n \to \infty) \) in such a way that also the second-phase sample size \( m \) can increase indefinitely \( (m \to \infty) \). Therefore, even if actually there are unique first- and second-phase sample sizes \( n \) and \( m \), if asymptotic unbiasedness and consistency holds, then the distributions of interpolated values can be considered approximately centered and concentrated on the true values for \( n \) and \( m \) large enough, as it is likely to occur in large-scale environmental and forest surveys.

As to the consistency of the two-phase HT interpolator (2.4), while the first-phase samples increase in size, nothing ensures that these samples are nested. Therefore, conditions ensuring the convergence in probability of \( \hat{y}_{HT}(p, S) \) to the first-phase interpolator \( \hat{y}(p) \), involving the presence of a sequence of nested first-phase samples, are not satisfied (Isaki and Fuller, 1982). For this reason, the second phase HT interpolator will not be considered further in this study.

On the other hand, conditions providing asymptotic design-based unbiasedness and consistency of the two-phase IDW interpolators of type (2.7) are derived broadening the results by Fattorini et al. (2018a). Technical derivations of these results are moved into Appendix A, even if they constitute the key point of this paper. Consistency results can be summarized as follows. From Theorem A.1, point and uniform consistency is achieved at the cost of supposing smoothness of the surface under estimation at the point or onto the whole set \( \mathcal{A} \), respectively, as well as a spatial balance asymptotically achieved by the two-phase sampling scheme. As to the smoothness of
the surface, in environmental and forest studies there may be surfaces with many discontinuities, especially at small scale. Obviously, consistency does not hold where discontinuities are present. However, the consistency of the whole map is preserved if these discontinuities occur for sets of measure zero. It should be pointed out that this last requirement is quite realistic in environmental and forest studies because there are parts of the study region in which surfaces change smoothly, well approaching the theoretical condition of continuity and, even when surfaces change abruptly, that usually occurs along borders delineating sudden variations in the characteristic of the study region. Thus, these borders may be realistically approximated by curves on the study area, well approaching the theoretical condition of discontinuity over a region of zero measure. This situation is typical for forest attribute maps, where the attribute amount jumps to 0 along forest edges.

As to the asymptotical spatial balance of the two-phase sampling scheme, condition (A.10) requires that as the first- and second-phase sample sizes increase, the two-phase scheme is able to evenly spread the second-phase sample in such a way that any point \( p \) of the study region is likely to have neighbouring points sampled. Condition (A.10) can be reformulated in terms of first- and second-order inclusion probabilities of the second-scheme by means of Proposition A.3. Moreover, from the same proposition it follows that, when the first- and second-order inclusion probabilities of the second-phase design do not depend on the first-phase sample, the first-order inclusion probabilities are invariably greater than a given threshold and the second-order inclusion probabilities are less than or equal to the product of the corresponding first-order ones, consistency of the two-phase IDW interpolators is achieved any time the first-phase design satisfies conditions for the consistency of the first-phase IDW interpolator. In the next section, these requirements will be proven to be satisfied by the most familiar sampling schemes adopted in environmental and forest surveys.

4. **Consistency results under some two-phase sampling.** When the study region is a continuum, there is a wide variety of schemes to sample locations in the first-phase. Uniform random sampling (URS) is the most straightforward scheme consisting in randomly and independently selecting \( n \) points onto the study area. Despite its simplicity, URS may lead to uneven surveying of the study region while the achievement of spatially balanced samples is generally preferable in environmental and forest surveys in order to cover the whole region. To this purpose, when the study region can be partitioned into \( n \) regular polygons, systematic grid sampling (SGS) can be performed by selecting a point within a polygon and then repeating it in
the same position within the other polygons. SGS has constituted a sort of standard design of wide application in the first phase of most national forest inventories (Opsomer et al., 2007; Tomppo et al., 2010). It is well known that the performance of SGS strongly deteriorates in presence of spatial regularities. In order to overcome the drawback, tessellation stratified sampling (TSS), consisting of partitioning the study region into \( n \) spatial subsets of equal size and randomly and independently selecting a point in each subset, is becoming increasing popular (e.g. Fattorini et al., 2006). Therefore, TSS seems preferable to SGS, avoiding deterioration under spatial regularities and being more generally applicable, stated that it does not entail the regularity of the subsets partitioning the study region. Fattorini et al. (2018a) derived conditions ensuring design consistency of the IDW interpolator when the number of points approaches infinity and, in particular, they prove consistency to hold under URS, SGS and TSS. Owing to these consistency results and to the popularity of these schemes in environmental and forest surveys, we focus on two-phase sampling schemes where URS, SGS and TSS are implemented to select points in the first-phase.

Once the first-phase sample has been selected, we already pointed out that, as the survey variable cannot be recorded at each point, the sample can be regarded as a finite population from which a second-phase sample can be suitably selected.

Among familiar second-phase sampling schemes, the plainest is simple random sampling without replacement (SRSWOR). It is at once apparent that SRSWOR is independent from the locations of the first-phase points. Moreover, if a constant fraction \( 0 < \pi_0 < 1 \) of second-phase points is selected by SRSWOR from the first-phase sample \( P \) of size \( n \), then \( \pi_i = \pi_0 \) and \( \pi_{ij} = (\pi_0^2 n - \pi_0)/(n - 1) \) for each \( i \neq j \in P \). Because \( \pi_i \geq \pi_0 > 0 \) for all \( i \) and \( \pi_{ij} - \pi_i \pi_j \) is invariably negative for each \( i \neq j \), conditions (A.18) and (A.19) hold. Consequently, both the direct and ratio-type interpolator turn out to be pointwise consistent when URS, SGS or TSS is adopted in the first-phase and SRSWOR in the second one.

A slight modification of this scheme is common in forest studies when forest stands are scattered throughout the region and large voids are present. In this case, it is customary to stratify the set of first-phase point \( P \) on the basis of aerial image into the stratum \( P_{NF} \) of points whose plots are completely lying outside forest and the stratum \( P_F \) of points whose plots are completely or partially lying within forest. The first-phase points of \( P_{NF} \) are selected in the second-phase with certainty, because the amount of forest attributes within them are known to be 0 without field work, while the fraction \( \pi_0 \) of points to select in the second phase by SRSWOR is completely
devoted to the stratum $P_F$ (e.g. Chirici et al., 2016, Corona et al., 2019). Because in the second phase a stratified sampling is implemented with the two strata depending on $P$, i.e. the inclusion probabilities are functions of $P$, consistency results about this scheme would be more complex to achieve. However, since SRSWOR ensures consistency and stratified sampling turns out to be more efficient than SRSWOR, then a fortiori this scheme should ensure consistency.

When the study region can be partitioned into a grid of regular polygons or into spatial subsets of equal size and SGS or TSS is implemented in the first-phase, a widely applied second-phase sampling scheme is One-Per-Stratum Sampling (OPSS). Under OPSS, the regular polygons or the spatial subsets are clustered into contiguous blocks/strata of equal size and one polygon/subset is randomly selected from each block/stratum (e.g. Fattorini et al., 2017). Because each polygon/subset univocally corresponds to a first-phase point, the selection of second-phase points may be viewed as the selection of polygons/subsets from the partition. In particular, if the $n$ polygons/subsets are clustered into $m$ blocks/strata, then $\pi_i = m/n$ for each $i \in P$. Moreover, as to the second-order inclusion probability, $\pi_{ij} = 0$ if the points $i$ and $j$ belong to the same block/stratum and $\pi_{ij} = m^2/n^2$ otherwise. Therefore, as $\pi_{ij} - \pi_i \pi_j$ is invariably negative for each $i \neq j$, if the second-phase sampling fraction $m/n$ is greater than a fixed $\pi_0 > 0$, conditions (A.18) and (A.19) hold and both the direct and ratio-type interpolator turn out to be pointwise consistent.

5. Mean squared error estimation. Generally, for providing well-defined maps, the estimation of $y(p)$ must be performed for many spatial points, sometimes even millions when large study regions are considered. Accordingly, the estimation of the mean squared error of the two-phase IDW interpolators should not be computationally demanding. Therefore estimators based on time-consuming resampling procedures should be avoided. Owing to Tobler’s law (Tobler, 1970), which in turn justifies the use of IDW interpolation, the value of the the survey variable at the sampled point nearest to $p$ is likely to be a good (known) proxy for $y(p)$. Then, if $\tilde{y}(p) = y(P_{\text{near}(p)})$, where near($p$) is the index of the second-phase sampled point nearest to $p$, that is

$$||p - P_{\text{near}(p)}|| = \min_{i \in S} ||p - P_i||,$$

a straightforward estimator for the mean squared error of $\hat{y}(p, S)$ is

$$\hat{V}(p) = (\hat{y}(p, S) - \tilde{y}(p))^2.$$
The asymptotical features of (5.1) are proven in Appendix B, where we
give evidence that \( \hat{y}(p) \) is an asymptotically conservative estimator of the
mean squared error at any continuity point for \( y \) if the two-phase design
ensures an asymptotic spatial balance. Moreover, if \( y \) is supposed to be a
Lipschitz function in a neighborhood of \( p \), SGS or TSS is implemented in
the first phase and OPSS in the second-phase, it is also proved that the
negative part of the bias of \( \hat{V}(p) \) is bounded by a \( O(m^{-1}) \) term.

6. Simulation study. The performance of the two-phase direct inter-
polator \( \hat{y}_D \) and of the two-phase ratio-type IDW interpolator \( \hat{y}_R \) is empir-
ically investigated by means of a simulation study under some two-phase
sampling schemes commonly adopted in environmental and forest studies
and already discussed in section 4.

Three surfaces are considered. An artificial surface on the unit square
defined at any point \( p = (p_1, p_2) \in (0; 1) \) as
\[
y(p) = C \sin 3p_1 \sin^2 3p_2
\]
where the constant \( C \) ensures a maximum surface value of 10. The artifi-
cial surface is considered to mimic the idyllic, even if unrealistic, situation
in which a map is uniformly continuous as well as lipschitzian. Moreover,
in order to evaluate the performance of the proposed interpolators in less
stylized situations, two real surfaces are considered. The first real surface
arises from a completely forested study area of approximately 4.8 \( ha \) in Val
di Sella (Northeastern Italy, southern Alpine area). On July 2008 the wood
volume (\( m^3 \)) of each tree lying in the area and with diameter at breast height
greater than 17.5 cm was evaluated by means of the local double-entry vol-
ume tables currently used in the forest management practice (Corona et al.,
2014). The surface value at \( p \) is the total volume of those trees within a
 circular plot of radius 13 m centered at \( p \). The second real surface arises
from a forested area with large not-forest voids within a quadrat of 9 \( ha \) lo-
cated within the beech forest of Monte Cimino (Central Italy). A complete
field enumeration carried out in 2016 gave tree locations and wood volumes
(\( m^3 \)) for each tree within the quadrat with diameter at breast height
greater than 7 cm. Also in this case, the surface value at \( p \) is the total volume of
trees within a circular plot of radius 13 m centered at \( p \). An ortho-corrected
IRS LISS-III (Indian Remote-Sensing satellite Linear Imaging Self Scanning
Sensor 3) imagery was available for the study area, allowing to detect those
plots completely lying in not-forest voids for which the total wood volume
is 0. The mathematical properties of surfaces giving the amount of forest
attributes within plots of prefixed size centred at any point of the study
Fig 1. Maps of the artificial surface (left), Val di Sella surface (centre) and Monte Cimino surface (right). Real surfaces are rescaled for giving values in $(0,10)$ as in the artificial surface.

region are discussed by Fattorini et al. (2019b). They are piecewise constant functions with many discontinuities along plot borders and as such they are continuous and lipschitzian except for sets of measure 0. In order to make the results achieved from the artificial surface comparable with those achieved from the real surfaces, real surfaces are rescaled for giving a maximum value of 10. The three surfaces are reported in Figure 1.

As to the choice of the first-phase design, Fattorini et al. (2018a) showed that TSS and SGS outperform URS as the sample size increases. Therefore, stated that TSS does not suffer from efficiency loss in presence of spatial regularities, only TSS is considered in the first phase of the simulation. For each surface, $R = 10,000$ first-phase samples of size $n = 100, 400$ and $900$ are independently selected. In particular, for the artificial and Monte Cimino surfaces TSS is performed by partitioning the two squared regions into grids of $10 \times 10$, $20 \times 20$ and $30 \times 30$ quadrats of equal size, while for the Val di Sella surface, since a regular tessellation is not possible because of its irregular shape, the study area is tessellated into $n$ irregular subsets of equal size using the R package “spcosa” provided by Walvoort et al. (2010). Then a point is randomly selected within each quadrat/subset and the surface value is recorded.

As to the second phase, 10% and 20% of the first-phase points are selected for each surface. In the case of artificial surface and Val di Sella surface, second-phase sampling is performed according to SRSWOR and OPSS. In particular, in the case of the artificial surface, OPSS is performed by partitioning the grids of $10 \times 10$, $20 \times 20$ and $30 \times 30$ quadrats into rectangular blocks of $5 \times 2$ or $5 \times 1$ contiguous quadrats and one quadrat is randomly selected within each block. In the case of Val di Sella surface, OPSS is performed by partitioning the $n$ irregular subsets into blocks of 10 or 5 contiguous subsets and one subset is randomly selected within each
block. The partition is performed by a modification of the k-means algorithm minimizing the within-block sum of distances between the subset centroids but also ensuring the same number of subsets per block. Since SRSWOR and OPSS induce constant first-order inclusion probabilities, the two-phase direct and ratio-type IDW interpolators coincide.

In the case of Monte Cimino surface, where plots lying completely outside forests are possible and can be detected by remote-sensing imagery, second-phase sampling is performed according to SRSWOR and stratified sampling with forest and not-forest strata. In particular, while SRSWOR is performed selecting $m$ points out of $n$, stratified sampling is performed including in the second-phase sample all the first-phase points whose plots lay completely outside forest, for which no field work is required, and then selecting $m$ points among the remaining ones. Since the stratified sampling induces first-order inclusion probabilities depending on the first-phase points, direct and ratio-type IDW interpolators differ.

For each combination of surfaces, sample sizes and second-phase sampling schemes, direct and (if differing) ratio-type IDW interpolators are computed, together with the corresponding mean squared error estimates obtained using (5.1). According to Fattorini et al. (2018a), interpolation is performed by using $\varphi(||p - P_i||) = ||p - P_i||^{-3}$ as distance function. In particular, for the artificial and Monte Cimino surfaces, estimation is performed for each of the $N = 10,000$ centroids of the equally spaced $100 \times 100$ grid located within the two squared study areas, while for the Val di Sella surface, the study area is tessellated into $N = 10,000$ subsets and estimation is performed at each subset centroid.

Therefore, denote by $p_1, \ldots, p_N$ the centroids and by $y_j = y(p_j)$ the surface value at the $j$-th centroid and let $\hat{y}_{r,j}$ and $\hat{V}_{r,j}^{1/2}$ the estimate of $y_j$ achieved by means of the direct and (if differing) ratio-type IDW interpolator and the estimate of its root mean squared error, respectively, resulting from the $r$-th replication of the two-phase sampling at the $j$-th centroid ($r = 1, \ldots, R$, for $j = 1, \ldots, N$). Then, for each combination of surfaces, sample sizes and second-phase sampling schemes, the absolute bias (AB) and the root mean squared error (RMSE) of the direct and (if differing) ratio IDW interpolators of $y_j$ at $p_j$ are empirically computed for each centroid $p_j$ together with the absolute bias of the root mean squared error estimator (ABRMSE) as

$$AB_j = \left| \frac{1}{R} \sum_{r=1}^{R} \hat{y}_{r,j} - y_j \right|, \ j = 1, \ldots, N$$
$$RMSE_j = \left\{ \frac{1}{R} \sum_{r=1}^{R} (\hat{y}_{r,j} - y_j)^2 \right\}^{1/2}, \ j = 1, \ldots, N$$

$$ABRMSE_j = \left| \frac{1}{R} \sum_{r=1}^{R} \hat{V}_{r,j}^{1/2} - RMSE_j \right|, \ j = 1, \ldots, N.$$ 

Tables 1 and 2 report the minimum, the mean and the maximum values of AB, RMSE and ABRMSE of the direct/ratio-type IDW interpolator for the artificial and Val di Sella surfaces respectively. Table 3 reports the same indexes for the direct/ratio-type IDW interpolator under SRSWOR and for the direct interpolator and ratio-type interpolator under stratified sampling, when the two intepolators differ.

<table>
<thead>
<tr>
<th>Second-phase</th>
<th>n</th>
<th>m</th>
<th>AB</th>
<th>RMSE</th>
<th>ABRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
<td>Mean</td>
</tr>
<tr>
<td>SRSWOR</td>
<td>100</td>
<td>10</td>
<td>0.001</td>
<td>1.446</td>
<td>3.027</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>40</td>
<td>0.000</td>
<td>0.734</td>
<td>2.254</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>90</td>
<td>0.000</td>
<td>0.476</td>
<td>1.766</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20</td>
<td>0.000</td>
<td>1.023</td>
<td>2.637</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>80</td>
<td>0.000</td>
<td>0.490</td>
<td>1.827</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>180</td>
<td>0.000</td>
<td>0.316</td>
<td>1.375</td>
</tr>
<tr>
<td>OPSS</td>
<td>100</td>
<td>10</td>
<td>0.000</td>
<td>1.336</td>
<td>2.846</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>40</td>
<td>0.000</td>
<td>0.578</td>
<td>2.034</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>90</td>
<td>0.000</td>
<td>0.376</td>
<td>1.589</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20</td>
<td>0.000</td>
<td>0.959</td>
<td>2.474</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>80</td>
<td>0.000</td>
<td>0.421</td>
<td>1.681</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>180</td>
<td>0.000</td>
<td>0.272</td>
<td>1.277</td>
</tr>
</tbody>
</table>

From Table 1 it is at once apparent that for the artificial surface, both under SRSWOR and OPSS, a sharp decrease of averages and maxima of AB and RMSE of the direct/ratio-type IDW interpolator occurs as the first-phase sample size, and consequently the second-phase sample size, increases. For both the sampling fractions, there is a reduction of about 3 times in the averages of AB and RMSE when passing from $n = 100$ to $n = 900$ while a reduction of about 1.5-2 times occurs for the maxima. Regarding RMSE estimation, the reduction in the absolute bias of estimator (5.1) is of about 4-6 times in the averages and of 1.5-2 in the maxima. Moreover, as
Table 2
AB, RMSE and ABRMSE for the Val di Sella surface for the two-phase direct/ratio-type IDW interpolators under first-phase TSS and second-phase SRSWOR and OPSS.

<table>
<thead>
<tr>
<th>Second-phase</th>
<th>n m</th>
<th>AB</th>
<th>RMSE</th>
<th>ABRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
</tr>
<tr>
<td>SRSWOR</td>
<td>100</td>
<td>10</td>
<td>0.000</td>
<td>1.089</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.000</td>
<td>0.917</td>
<td>5.225</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>0.000</td>
<td>0.786</td>
<td>4.419</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.000</td>
<td>1.010</td>
<td>5.777</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.000</td>
<td>0.798</td>
<td>4.487</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>0.000</td>
<td>0.669</td>
<td>3.671</td>
</tr>
<tr>
<td>OPSS</td>
<td>100</td>
<td>10</td>
<td>0.000</td>
<td>1.089</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.000</td>
<td>0.903</td>
<td>5.199</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>0.000</td>
<td>0.751</td>
<td>4.259</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.000</td>
<td>1.008</td>
<td>5.791</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.000</td>
<td>0.772</td>
<td>4.391</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>0.000</td>
<td>0.648</td>
<td>3.617</td>
</tr>
</tbody>
</table>

expected, OPSS invariably outperforms SRSWOR for all the performance indexes. These results confirm the theoretical findings about consistency of the direct/ratio-type interpolator under continuous surfaces.

Taking the results achieved under the stylized artificial surface as benchmarks, analogous results are obtained when the real surfaces are considered, even if the rates of decrease of maxima and averages of AB, RMSE and ABRMSEE and the improvements of OPSS over SRSWOR are less pronounced, as expected since many discontinuities are present, especially for Monte Cimino surface, where roughness is more marked with respect to Val di Sella surface owing to the presence of forest voids. That is apparent also from the magnitude of these indexes, that they are relevantly larger (especially in the maxima) than those achieved from the artificial surface, with the greatest magnitudes that are achieved for Monte Cimino surface.

In particular, regarding Monte Cimino results in Table 3, where the direct and ratio-type IDW interpolators differ under forest/not-forest stratification, it is worth noting that ratio-type interpolator invariably outperforms direct interpolator regarding averages and maxima of all the performance indexes. Moreover, as expected, ratio-type interpolator under stratified sampling outperforms direct/ratio-type interpolator under SRSWOR.

Figures 2-4 provide a graphical overview of the simulation results for the direct/ratio-type IDW interpolator when 10% of the first-phase points are sampled in the second phase according to SRSWOR. The improvement of
Table 3

*AB, RMSE and ABRMSEE for the Monte Cimino surface for the two-phase direct and ratio-type IDW interpolators under first-phase TSS and second-phase SRSWOR and stratified sampling with forest/not-forest strata.*

<table>
<thead>
<tr>
<th>Second-phase</th>
<th>n</th>
<th>m</th>
<th>AB</th>
<th>RMSE</th>
<th>ABRMSEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Max</td>
</tr>
<tr>
<td>SRSWOR</td>
<td>100</td>
<td>10</td>
<td>0.000</td>
<td>1.261</td>
<td>7.098</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>40</td>
<td>0.000</td>
<td>1.093</td>
<td>6.679</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>90</td>
<td>0.000</td>
<td>0.957</td>
<td>6.151</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20</td>
<td>0.000</td>
<td>1.183</td>
<td>6.951</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>80</td>
<td>0.000</td>
<td>0.972</td>
<td>6.207</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>180</td>
<td>0.000</td>
<td>0.824</td>
<td>5.399</td>
</tr>
<tr>
<td>Stratified</td>
<td>100</td>
<td>20</td>
<td>0.000</td>
<td>1.308</td>
<td>8.090</td>
</tr>
<tr>
<td>(direct-type)</td>
<td>400</td>
<td>40</td>
<td>0.000</td>
<td>1.103</td>
<td>7.632</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>90</td>
<td>0.001</td>
<td>0.953</td>
<td>6.876</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>40</td>
<td>0.000</td>
<td>1.165</td>
<td>7.522</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>80</td>
<td>0.000</td>
<td>0.923</td>
<td>6.623</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>180</td>
<td>0.000</td>
<td>0.776</td>
<td>5.679</td>
</tr>
<tr>
<td>Stratified</td>
<td>100</td>
<td>20</td>
<td>0.000</td>
<td>1.208</td>
<td>7.373</td>
</tr>
<tr>
<td>(ratio-type)</td>
<td>400</td>
<td>40</td>
<td>0.000</td>
<td>1.013</td>
<td>6.777</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>90</td>
<td>0.000</td>
<td>0.880</td>
<td>6.154</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>40</td>
<td>0.000</td>
<td>1.140</td>
<td>7.082</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>80</td>
<td>0.000</td>
<td>0.913</td>
<td>6.216</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>180</td>
<td>0.000</td>
<td>0.770</td>
<td>5.369</td>
</tr>
</tbody>
</table>

The two-phase IDW interpolation is apparent because maps of AB, RMSE and ABRMSEE become whiter and whiter for all the surfaces as first- and second-phase sample sizes increase, even if the palest maps are achieved for the artificial surface. Similar and even more appealing patterns are obtained when 20% of the first-phase points is selected in the second phase or when OPSS or stratified sampling are considered in the second phase. The corresponding figures are not reported for sake of brevity.
Fig 2. Maps of AB, RMSE and ABRMSE for the artificial surface for the two-phase direct/ratio-type IDW interpolator under second-phase SRSWOR. First-phase sample size 100 (row 1), 400 (row 2), 900 (row 3) and second-phase sampling fraction equal to 10%.
Fig 3. Maps of AB, RMSE and ABRMSE for the Val di Sella surface for the two-phase direct/ratio-type IDW interpolator under second-phase SRSWOR. First-phase sample size 100 (row 1), 400 (row 2), 900 (row 3) and second-phase sampling fraction equal to 10%.
7. Case study. The study was carried out in an extensive forest area of 4,900 ha located in Casentino Valley, the Eastern part of the Tuscany Region (Central Italy). According to the European classification, the forest types that covered the study area are the Apennine-Corsican mountainous beech forests, the Thermophilous deciduous forests dominated by chestnut, the Subalpine and mountainous spruce and mountainous mixed spruce-silver fir forest, and the Turkey oak, Hungarian oak and Sessile oak forest. The climate is temperate-humid: mean annual temperature is about 10°C and total annual rainfall is higher than 1,000 mm, with an average of more than 55 mm in the summer months (June-August). The forest grows on sandy-loamy or loamy soils, rich in humus on the surface horizons. Soil depth varies. The slopes are generally steep or very steep. The area is characterized by forest exploitation, very intensive from the second half of the eighteenth century when beech forests were extensively clear cut leaving only some seed
trees (generally 30 per hectare). Stands were thus transformed into coppices, which were repeatedly utilized until the second half of the twentieth century.

Owing to the intensive management, a two-phase sample survey was performed in 2013 to analyse the effect of human activities on the ecological diversity of the forest. In the first-phase, the study area was covered by a regular grid of 1,225 quadrats of 200 m side and a point was randomly selected within each quadrat. Each point was considered to be the centre of a circular plot of 20 m radius. In the second phase, on the basis of aerial imagery, the 1,225 plots selected in the first-phase were partitioned into two strata: the stratum of 1,151 plots completely or partially located within forest and the stratum of 74 plots completely located outside forest. A second-phase sampling effort of 95 plots was judged suitable in terms of time and available resources. Therefore, the 74 first-phase plots completely located outside forest were selected in the second-phase with certainty, since the value of any forest attribute was known to be 0 without field work, and the sample of 95 second-phase plots was selected from the stratum of 1,151 forest plots by means of SRSWOR (sampling fraction of about 8%). Each of the 95 second-phase plots of 20 m radius was visited on the ground and the basal area \((m^2)\) of all the trees, taller than 1.30 m with stem diameter at breast height greater than 5 cm, lying within the plot was recorded together with the tree species.

An accurate design-based diversity analysis was performed by Corona et al. (2019) from sample information. The overall abundance was estimated to be of about 454 trees per ha with a total basal area of about 23 \(m^2\) per ha. A total of 27 different species were detected on field plots, that can be considered a reliable number of species actually present owing to the massive management of forests. As to the apportionment of abundance and basal area to species, results provide evidence that both the attributes are concentrated in a few dominant species. The concentration was emphasized by the Lorenz curves that highly deviated from the equal-apportionment line and provided values of the concentration indexes near to one. However, the presence of dominant species, which should entail a poor ecological diversity, was mitigated by the presence of several rare species that tended to increase diversity. This aspect was evidenced by the Shannon and the Gini-Simpson diversity index estimates.

In this framework it would be of interest to provide maps depicting the spatial distributions of attributes and species throughout the study region, just on the basis of the sampling scheme adopted to perform the survey, without exploiting models that were never used in the previous analysis by Corona et al. (2019). That was not possible in that paper owing to the
methodological gaps on mapping attributes from two-phase sampling.

Now, from the results achieved in this paper, the IDW ratio-type interpolator (2.5) has been adopted to estimate the surface giving the amount of basal area of trees lying with the plot of 20 m radius centred at any point of the study area. The ratio-type interpolator has been preferred to the direct interpolator owing to the simulation results of section 6. Estimation has been performed for any point of a grid of 110x110 points superimposed to the study area. The corresponding RMSEs estimates have been computed by means of (5.1). The estimated map of the basal area and of the corresponding RMSE estimates are reported in Figure 5. The resulting map of basal area evidences smaller values in the south-west part of the study area, characterized by very steep slopes and hence less depth soils with less productive stands, as well as in the whole east part where past forest exploitations were more intensive.

![Estimated surface and RMSE estimates](image)

**Estimated surface**

**Estimated RMSE**

**Fig 5. Maps of estimated surface of the total basal area within a plot of 20 m radius and the corresponding RMSE estimates obtained by adopting the two-phase ratio-type IDW interpolator.**

8. **Concluding remarks.** Accurate and updated wall-to-wall maps depicting the spatial pattern of ecological attributes throughout the study area is becoming a burning issue in environmental and forest surveys. In the recent conference on “A century of national forest inventories informing past, present and future” (Sundvolden NW, May, 20-23 2019, https://nibio.pameldingssystem.no/nfi100years) the increasing demand of maps for both large and small areas emerged as one of the main necessity in forest studies. Recently, Di Biase et al. (2018) provided an updated review of model-based and design-based approaches for mapping environmental resources. While model-based methodologies are leaded by the probabilistic structure
that is supposed to have generated the population, completely neglecting
the (sometimes complex) sampling schemes adopted to perform the survey,
we have addressed mapping in a design-based framework under two-phase
sampling, as it is customary in environmental and forest surveys at large
scale. It is worth noting that the use of two-phase schemes has been not
only highly recommended, but has been considered quite mandatory in the
FAO expert meeting on national forest inventories held in Rome in 2011
(FAO, 2011).

In this scenario, the results of Appendix A not only prove the statistical
rigor of two-phase mapping under some existing schemes of wide application
in environmental and forest surveys, but they also provide guidelines for
designing novel sampling schemes, possibly exploiting auxiliary information
available for the whole study region or collected at the selected points in the
first-phase of sampling (e.g. LiDAR metrics as proxies of forest attributes).

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

References.
Baffetta, F., Fattorini, L. and Corona, P. (2011). Estimation of small woodlot and
Bruno, F., Cocchi, D. and Vagheggini, A. (2013). Finite population properties of
Comparing echo-based and canopy height model-based metrics for enhancing estimation
174 1–9.
Chirici, G., Giannetti, F., McRoberts, R. E., Travaglini, D., Pecchi, M.,
Maselli, F., Chiesi, M. and Corona, P. (2020). Wall-to-wall spatial prediction of
growing stock volume based on Italian National Forest Inventory plots and remotely
sensed data. Int. J. Appl. Earth Obs Geoinformation 84 Online First.
Estimation of standing wood volume in forest compartments by exploiting airborne
J. Forest Res. 44 1303–1311.
Corona, P., Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C.,
by two-phase inventories. Environmetrics 30 2, e2502.