Università degli Studi della Tuscia di Viterbo
Dipartimento per la INNOVAZIONE NEI SISTEMI BIOLOGICI, AGROALIMENTARI E FORESTALI

Corso di Dottorato di Ricerca in
SCIENZE, TECNOLOGIE E BIOTECNOLOGIE PER LA SOSTENIBILITÀ XXXII Ciclo
Curriculum Ecologia Forestale e Tecnologie Ambientali

A PROPOSAL FOR DESIGN-BASED MAPPING OF FOREST RESOURCES

(SSD SECS-S/01)

Tesi di dottorato di:
Dott. ssa Rosa Maria Di Biase

Coordinatore del corso
Prof. Andrea Vannini

Tutore
Prof. Lorenzo Fattorini

Co-tutore
Prof. Piermaria Corona

A.A. 2019/20

To Matteo

Time zones keep us apart
but you are always
on my side

Acknowledgments

I would like to express my sincere gratitude to Professor Lorenzo Fattorini and Professor Piermaria Corona for their guidance during these three years of research and writing of this thesis.
Besides my advisors, I thank Professor Timothy Gregoire, who helped me during my stay at Yale University. Also I gratefully acknowledge the funding received towards my Ph.D. from the Research Centre for Forestry and Wood (Arezzo).
Last but not least, I would like to thank my family and my friends for their constant support.

Short abstract

This Ph.D. thesis is aimed at the reconstruction of spatial maps under a model-assisted framework. In particular, starting from the asymptotic unbiasedness and consistency conditions achieved for maps of finite populations of spatial units when exploiting the sole spatial information, the IDW interpolator is extended to obtain spatially explicit estimators when exploiting auxiliary variables, even if the theoretical properties of this novel proposal and the way to estimate its precision are still under study.

Key words: spatial interpolation, spatial maps, regression, design-based inference, model-based inference.

Extended abstract

A proposal for design-based mapping of forest resources

1 Introduction

Wall-to-wall maps of forest attributes are essential information sources for forest management (e.g. Corona, 2016). Knowledge of the distribution of such attributes is crucial in biodiversity conservation, water regulation and mitigation of climate change impacts (e.g. Köhl et al., 2006). Moreover, forest maps geographically depicting forest resource locations can be used to achieve a better stratification of the area, improving future inventory estimates (McRoberts and Tomppo, 2007).

Unfortunately, the distribution of forest attributes is an unknown parameter and sampling procedures are needed for its estimation. The study area is sometimes partitioned into a finite population of spatial units and the survey variable is recorded only for a sample of units and estimated in the non-sampled ones using an estimation criterion. Then, these spatially explicit estimates are used to create a map of the variable distribution in the whole region.

Most methodologies applied for mapping forest variables rely on model-based inference because pure design-based methods are usually avoided for constructing wall-to-wall maps, owing to the difficulty to estimate non-sampled values without any assumption. Indeed, the only way to recover the information in the unsampled units is to rely on model-based methods, such as kriging (Cressie, 1993) and k-nearest neighbour (k-nn) (McRoberts et al., 2007) predictors. Despite their popularity, very few comparative studies on the performance of these model-based techniques in forest studies have been conducted.

Working in a model-assisted framework seems the only way to produce wall-to-wall maps of forest attributes of design-based nature. However, when estimating the value of the survey variable at a single location, there is no way to correct the error provided by the assisting model and, consequently, the resulting map is destined to be design-biased. Accordingly, as pointed out by Fattorini et al. (2018a), any map arising from a model-assisted criterion can achieve statistical soundness only if it is proven to be design-based

asymptotically unbiased and consistent (DBAU&C). Fattorini et al. (2018a) derived those DBAU&C conditions for finite populations of spatial units when the sole spatial information is available. The authors used Tobler's first law of geography (Tobler, 1970) as the criterion leading to the inverse distance weighting (IDW) interpolator (Shepard, 1968) for the estimation of the values of the survey variable at the unsampled locations.

This novel approach of Fattorini et al. (2018a) completely neglects the auxiliary sources of information available for the area under examination, such as data provided by satellites and aircraft-based sensors. When those proxies are highly correlated to the interest variable, they should improve the estimation obtained with the sole spatial information.

For these reasons, auxiliary information will be exploited in a model-assisted framework. This novel approach is presented in section 2, whereas sections 3 and 4 contain a simulation study and a case study, respectively. Lastly, section 5 contains final remarks and future developments.

2 Materials and methods

Consider a study area $\mathcal{A}$ of size A partitioned into N spatial units $\mathrm{a}_1,\ldots,\mathrm{a}_N$ and let d_{jh} be the Euclidean distance between each couple of units; y_j denotes the amount of interest variable within the unit a_j and $f_j = y_j / |a_j|$ is its density. Let U denote both the set of spatial units partitioning $\mathcal{A}$ and the set of indexes $\{1,\ldots,N\}$ identifying the spatial units. The aim is the estimation of y_j or, equivalently, f_j, for each pixel of the population on the basis of a sample S and exploiting auxiliary variables.

Therefore, let y_j^0 be the amount of a proxy for y_j that, at the moment, is supposed to be known for each spatial unit and let f_j^0 be the corresponding density. Moreover, we considered the relationship $y_j = y_j^0 + \varepsilon_j$, where ε_j is the error occurred in predicting y_j by means of y_j^0. Equivalently, $f_j = f_j^0 + e_j$, where $e_j = \varepsilon_j / |a_j|$ is the intensity of the prediction error for the unit j. Then, Tobler's first law of geography can be exploited to estimate the e_j's by means of the IDW interpolator. Practically speaking, for each

$j \in \mathsf{S}$, the intensity of the prediction error $e_j = f_j - f_j^0$ is known, in such a way that for each $j \in \mathsf{U}$ the intensity of the prediction errors can be interpolated by $\hat{e}_j = Z_j e_j + (1 - Z_j)\sum_{i \in \mathsf{U}} w_{ij} e_i$, where Z_j is a random variable equal to 1 if $j \in \mathsf{S}$ and 0 otherwise, $w_{ij} = \dfrac{Z_i \varphi(d_{ij})}{\sum_{l \in \mathsf{U}} Z_l \varphi(d_{lj})}$ is the weight attached to the error intensity of unit *i* to estimate the error intensity of unit *j* and $\varphi : [0, \infty) \to \mathbb{R}^+$ is a not increasing function with $\varphi(0) = 0$ and $\lim_{d \to 0^+} \varphi(d) = \infty$. Accordingly, the spatial interpolator of f_j based on proxy information is given by $\hat{f}_j = f_j^0 + \hat{e}_j$.

For achieving consistency of $\hat{f}_j$, the following assumption is introduced: there exists a Riemann integrable function φ from $\mathcal{A}$ onto $[-L, L]$ which gives the intensity of the prediction errors at any point of the study area. This assumption, which recalls the first DBAU&C condition determined by Fattorini et al. (2018a) for design-based maps when the sole spatial information is available, is quite reasonable. Indeed, jumps and irregularities in the density of the survey variable are absorbed by presumably similar jumps and irregularities of the proxy. It follows that $\hat{f}_j$ can be proven to be design-based asymptotically unbiased and consistent under the same conditions defined by Fattorini et al. (2018a), i.e. Riemann integrability, shape regularities, asymptotical spatial balance and suitable weighting function.

Furthermore, let suppose that some covariates are known for each spatial unit of the population. Accordingly, let $\mathbf{x}_j = \left[x_{1j}, \ldots, x_{Kj}\right]^{\mathrm{T}}$ be the amounts of *K* auxiliary variables $X_1, \ldots, X_K$ within a_j and let $\mathbf{g}_j = \left[g_{1j}, \ldots, g_{Kj}\right]^{\mathrm{T}}$ be the corresponding densities $g_{kj} = x_{kj} \big/ \left|a_j\right|$. Then, as in the generalized regression (GREG) estimation criterion (Särndal et al., 1992), the quantity $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j$ is adopted as a proxy for y_j, in such a way that $y_j = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j + \varepsilon_j$, where ε_j is the error occurred in predicting y_j by means of $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j$. Equivalently, $f_j = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}_j + e_j$.

If $\boldsymbol{\beta}$ was a vector of known parameters, then $y_j^0 = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$ and it would be possible to estimate the values of y_j as previously described. However, the $\boldsymbol{\beta}$'s are unknown parameters and they need to be estimated from the sample. Using the ordinary least squares (OLS) method, the estimate of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^0 = \left[\sum_{j\in\mathsf{U}} \frac{\mathbf{g}_j\mathbf{g}_j^{\mathrm{T}}}{\sigma_j^2}\right]^{-1} \sum_{j\in\mathsf{U}} \frac{f_j\mathbf{g}_j}{\sigma_j^2}$.

The ordinary least squares estimate $\boldsymbol{\beta}^0$ is unknown, because it requires the knowledge of all the population densities and all the variances. However, the population variances should have a structure such that they disappear in the computation of $\boldsymbol{\beta}^0$ without being estimated (Särndal et al., 1992) and the densities f_j are known for each $j \in \mathsf{S}$. Consequently, $\boldsymbol{\beta}^0$ can be estimated from the sample using the well-known Horvitz-Thompson (HT) criterion, in such a way that the GREG estimator of $\boldsymbol{\beta}$ (Särndal et al., 1992) can be calculated as

$$\hat{\boldsymbol{\beta}} = \left[\sum_{j\in\mathsf{S}} \frac{\mathbf{g}_j\mathbf{g}_j^{\mathrm{T}}}{\pi_j\sigma_j^2}\right]^{-1} \sum_{j\in\mathsf{S}} \frac{f_j\mathbf{g}_j}{\pi_j\sigma_j^2}$$

where the σ_j^2s vanish in the computation of $\hat{\boldsymbol{\beta}}$ (Särndal et al., 1992).

At this point, the intensity of the prediction error can be estimated by $\tilde{e}_j = f_j - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j$ for any $j \in \mathsf{S}$. Then, for each $j \in \mathsf{U}$, the prediction errors can be obtained using the IDW interpolator as $\hat{\tilde{e}}_j = Z_j\tilde{e}_j + (1-Z_j)\sum_{i\in\mathsf{U}} w_{ij}\tilde{e}_i$, where Z_j and w_{ij} are as above. Finally, the spatial interpolator of f_j based on GREG auxiliary information is given by $\hat{\tilde{f}}_j = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j + \hat{\tilde{e}}_j$. Its DBAU&C conditions should follow from the unbiasedness of $\hat{\boldsymbol{\beta}}$ and its bias and variance can be estimated using a modification of the bootstrap method.

Following this novel approach, the estimated $\hat{\tilde{f}}_j$ values are used as the pseudo-population from which B bootstrap samples are selected. For each $b = \{1,\ldots,B\}$, a sample S_b^* is selected from the estimated map $\hat{\tilde{f}}_1,\ldots,\hat{\tilde{f}}_N$ by means of the same sampling scheme adopted for the selection of the original sample S. Then, from the b-bootstrap sample S_b^*, the estimated map $\hat{\tilde{f}}_{b1}^*,\ldots,\hat{\tilde{f}}_{bN}^*$ is achieved by repeating the IDW

interpolation. Finally, the B bootstrapped maps are used to achieve the bootstrap estimates of MSEs

$$\hat{\mathrm{MSE}}_{\mathrm{boot}}^{(1)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{f}}_{b1}^{*} - \hat{\bar{f}}_{1}\right)^2 \ , \ldots , \ \hat{\mathrm{MSE}}_{\mathrm{boot}}^{(N)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{f}}_{bN}^{*} - \hat{\bar{f}}_{N}\right)^2$$

3 Simulation study

This simulation study was performed on a real population of beeches located in the area of Monte Cimino, Italy. It is an area of 9 ha, partitioned into 100 pixels of 30×30 m^2, censed several times through the years. Accordingly, the real distribution of the interest variable, i.e. the basal area G, and the real spatial map of its distribution are known. For each pixel of the population, data obtained from Landsat-8 acquisitions were available, so that it was possible to use them as auxiliary information for the estimation of G. However, in order to obtain the best possible correlation, some pixels were discarded, considering an area of just 79 pixels and applying the following regression model

$$G = b_0 + b_1 \mathrm{NDVI} + b_2 \cos(\mathrm{EXP}) + e$$

where NDVI and cos(EXP) are two Landsat-8 indexes.

This regression model explains very little of the response variability (adjusted R-squared $\cong 0.3$) and for this reason a new simulation study was conducted on the area of Monte Cimino, using all 100 original pixels and the same regression model. However, the values of the auxiliary variables NDVI and cos(EXP) were manipulated, so as to obtain a higher correlation and a better explanation of the response variability (adjusted R-squared $\cong 0.9$).

Lastly, in order to test out the performance of the IDW interpolator and the bootstrap MSE estimator with an increasing number of sample cells, the population of Monte Cimino was repartitioned into 400 square cells of 15×15 m^2, dividing each one of the original 100 pixels into 4 square sub-pixels. Moreover, the amounts of interest and auxiliary variables in each sub-pixel were distributed in such a way that the sum of G in each 4 sub-pixels of pixel j, $j = 1,\ldots,100$, was equal to the amount of interest variable in j. Also in this case, the same regression model was applied, with R-squared of 0.9.

The simulation study was performed as follows. First, the sample selection by means of one-per-stratum-stratified (OPSS) sampling was repeated $R = 10000$ times and the values of the interest variable in each pixel $j = 1,\ldots,N$ and for every replication

$r = 1, \ldots, R$ were estimated accordingly to the IDW interpolator, using $\alpha = 3, 6$. Second, the estimated values $\hat{y}_{j,r}$ were used as the pseudo-populations from which $B = 10000$ bootstrap samples were selected by means of OPSS as well. Finally, for each pixel, the following set of indexes was calculated:

- $\mathrm{E}\left(\hat{y}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\hat{y}_{j,r}$
- $\mathrm{B}_j = \mathrm{E}\left(\hat{y}_j\right) - y_j$
- $\mathrm{MSE}\left(\hat{y}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_{j,r} - y_j\right)^2$
- $\mathrm{RMSE}\left(\hat{y}_j\right) = \sqrt{\mathrm{MSE}\left(\hat{y}_j\right)}$
- $\mathrm{E}\left(\mathrm{R\hat{M}SE}_{\mathrm{boot}}^{(j)}\right) = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\mathrm{M\hat{S}E}_{\mathrm{boot}}^{(j,r)}}$, where

$$\mathrm{M\hat{S}E}_{\mathrm{boot}}^{(j,r)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{y}_{b,j,r}^{*} - \hat{y}_j\right)^2$$

- $\mathrm{D}_j = \mathrm{E}\left(\mathrm{R\hat{M}SE}_{\mathrm{boot}}^{(j)}\right) - \mathrm{RMSE}\left(\hat{y}_j\right)$

The maps of bias and precision have been reconstructed using the absolute values of B_j and D_j, referred to as absolute bias (AB) and absolute difference (AD).

In the population of Monte Cimino partitioned into 79 pixels, high values of AB, AD and RMSE were obtained for both distance functions, due to the very low correlation. The bootstrap RMSE estimator underestimates the real RMSE in almost every pixel. However, increasing the value of α from 3 to 6, the bootstrap RMSE estimator performs slightly better, with similar value of minima, maxima and mean, but a greater number of pixels with a positive D_j.

For Monte Cimino partitioned into 100 pixels, minima, maxima and mean values of AB, RMSE were less high, due to both the increase of the sample size and a better correlation. However, the bootstrap RMSE estimator underestimates the real RMSE as well and there are no relevant differences when using a distance function or the other.

There are no relevant differences among minima, maxima and means of AB, RMSE and AD when the value of α increases from 3 to 6 in the population of Monte Cimino partitioned into 400 pixels and, once again, the bootstrap RMSE estimator is not a conservative estimator most cases. It should be noted that the values of AB, AD and RMSE are a lot smaller than those achieved for the other two populations. These better

results are due to the increasing number of units in the sample and the decreasing size of the units as well, that better approach the consistency conditions.

4 An application to real case

The described methodology was adopted to provide the map of the living wood volume in the forest complex of Rincine, Italy. The area was partitioned into 5449 square cells of $23 \times 23\,m^2$, grouped into 50 strata. The total wood volume of the population sampled by means of OPSS was partitioned into three classes: small trees (Y1), medium trees (Y2) and large trees (Y3). For each cell, several Airborne Laser Scanning (ALS) indexes were available, but only 15 ALS height and density metrics were used as independent variables.

Instead of the ordinary least squares, the regression coefficients were estimated by means of the least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996), using Y1, Y2 and Y3 as three different dependent variables. It follows that three different sets of variables were selected from sample data, one for each size class.

This estimation process generated three values (one per class) of the wood volume estimates for each cell of the population, so that the total volume, i.e. the sum of the volumes of the three size classes, was obtained as a sum of these three values.

The resulting values (four for each cell) were used as the pseudo-population from which 20000 bootstrap samples were selected, so as to estimate the MSE, and consequently the RMSE, associated to each map cell and to each size class. To better appreciate the improvement achieved with the exploitation of the auxiliary variables, the spatial map of the total volume when using ALS data has been compared with the spatial map of the total volume when using the sole spatial information. As a result, the exploitation of auxiliary sources of information reduces the *smoothing* associated to IDW interpolator when using the sole spatial information.

Although the pulse density of the available ALS data was relatively low, this approach seems promising for estimating wood volume exploiting ALS data and relatively satisfying levels of accuracy have been reached when predicting wood volume from small, medium and large trees. Moreover, almost all the selected ALS metrics represent the upper part of the canopy in leaf-on conditions and these facts improve the capabilities for identifying trees at the top of the canopy, but makes the assessment of

lower layers (classes Y1 and Y2) worse than larger and higher trees (class Y3). The possible integration of leaf-on and leaf-off scans as well as the exploitation of ALS data characterized by high pulse density might improve the accuracy of the obtained models (Sumnall et al., 2016).

5 Conclusion and future perspectives

Wall-to-wall maps of the spatial distribution of forest resources are important instruments for monitoring forest health. Moreover, the introduction of remote sensing imagery in the last decades has eased the process of making maps, providing easily accessible and affordable sources of auxiliary information sources.

Consequently, the development of procedures exploiting auxiliary variables to predict and map forest attributes is a crucial issue in forest monitoring.

Thus, objective and reliable information are essential to obtain precise and useful maps. Since pure design-based methods cannot be used and model-based ones involve severe assumptions, working in a model-assisted framework seems the sole way to produce reliable maps, at least asymptotically unbiased and consistent.

The procedure here proposed is promising for estimating forest attributes exploiting auxiliary data, even if further analysis and studies on its properties and on the DBAU&C conditions are needed.

6 References

Corona P (2016) Consolidating new paradigms in large-scale monitoring and assessment of forest ecosystems. *Environ. Res.* **144**: 8-14.

Cressie N (1993) *Statistics for spatial data.* New York (USA): Wiley.

Fattorini L, Marcheselli M, Pratelli L (2018a) Design-based maps for finite populations of spatial units. *J. Am. Stat. Assoc.* **113**: 686-697.

Köhl M, Magnussen SS, Marchetti M (2006) *Sampling methods, remote sensing and GIS multiresource forest inventory*, Heidelberg (Germany): Springer.

McRoberts RE, Tomppo EO (2007) Remote sensing support for national forest inventories. *Remote Sens. Environ.***110**: 412-419.

McRoberts RE, Tomppo EO, Finley AO, Heikkinen J (2007) Estimating areal means and variances of forest attributes using the k-nearest neighbours technique and satellite imagery. *Remote Sens. Environ*. **111**:466-480.

Särndal CE, Swensson B, Wretman J (1992) *Model assisted survey sampling*, New York (USA): Springer-Verlag.

Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In Proceedings of the *23rd ACM National Conference*, New York (USA), August 27-29, pp. 517-524.

Sumnall MJ, Hill RA, Hinsley SA (2016) Comparison of small-footprint discrete return and full waveform airborne lidar data for estimating multiple forest variables. *Remote Sens. Environ.* **173**: 214-223.

Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**: 267-288.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**: 234-240.

Contents

Chapter 1

Introduction

1.1 The importance of forest mapping

Many human activities directly benefit from forest services, including timber, energy and non-woody forest products (Holmgren and Persson, 2002; Corona, 2016). Indeed, forests occupy a central role in a wide range of environmental issues related to biodiversity conservation, water regulation and soil protection, mitigation of climate change impacts and the provision of many ecosystem services (e.g. Köhl et al., 2006; Maselli et al., 2005).

Forest management and assessment are relevant for forest industry and environmental stakeholders (McRoberts and Tomppo, 2007), representing the main requirements for the compliance of international agreements such as the FAO Global Forest Resource Assessment (FRA) and the Kyoto protocol (McRoberts and Tomppo, 2007). At the same time, forest management decisions must rely on objective, reliable and geo-referenced information (Corona et al., 2011), such as that provided by forest inventories and remote sensing techniques (Köhl et al., 2006). Forest inventory should be based on statistically sound estimation of forest attributes in a given area (Corona, 2010). Sampling procedures for large-scale forest inventories, such as National Forest Inventories (NFI), were pioneered in the 20th century in North America and Scandinavia (Corona, 2000) with the sole aim to estimate harvestable timber amount across a determined forest stand. Sweden, for instance, had its first NFI in 1923, when the country feared the beginning of a wood shortage (Holmgren and Persson, 2002). Gradually, the interest around NFI estimation changed and the aim moved towards multipurpose surveys (Lund, 1998; Corona et al., 2002). In this transformation, variables not directly related to timber assessment and communities constituted by non-traditional objects (e.g. urban forests, woodlots and tree rows) were included within targeted ones (Kleinn, 2002). However, an expected drawback of multipurpose forest inventories is the increase in costs and time, unavoidable when working with an increasing number of variables (McRoberts and Tomppo, 2007). Fortunately, the

increasing availability of remote sensing tools, techniques and information, i.e. data acquired from satellites and aircraft-based sensors (Schowengerdt, 2006) at no or low cost, can be used as ancillary data to improve the estimation of forest attributes, without increasing the sampling effort and costs (e.g. Opsomer et al., 2007; Mattioli et al., 2012). In particular, Airborne Laser Scanning (ALS) is the main source of auxiliary data in forest inventories in several countries (Maltamo et al., 2014), also providing accurate wall-to-wall maps of forest structure over large areas (Noordermeer et al., 2019).

A further step in the scenario of enlarging NFI goals is to perform spatially-explicit estimation in order to generate forest maps geographically depicting forest resource locations (McRoberts and Tomppo, 2007; Corona, 2010), with the additional advantage of using the resulting map to achieve a better stratification of the area, so as to improve future inventory estimates. It should be noted that the use of satellite imagery for stratification increases the precision of estimators of forest area, but it is not as effective for forest attributes such as biomass and growing stock volume (McRoberts et al., 2006). For these attributes, ALS data are more effective as a source of stratification information (McRoberts et al., 2018).

1.2 Spatial map reconstruction

Wall-to-wall maps of forest attributes are essential information sources for making forest management decisions (e.g. Corona, 2016), since the knowledge of the distribution of such attributes, e.g. basal area and stem number, can be useful in many ecological applications, as those reported above. This distribution is obviously an unknown parameter and the available resources are commonly too scarce to allow for a complete census of the area under examination. Thus, sampling procedures are needed for its estimation.

Usually, the survey region is partitioned into a finite population of spatial units, in such a way that the survey variable is recorded only for a sample of units and estimated in the non-sampled ones using an estimation criterion. Those spatial units can be a collection of irregular patches (Figure 1.1a) determined by some characteristics of the region or a set of regular polygons, such as pixels (Figure 1.1b), the latter being more common in forest inventories (Opsomer et al., 2007; Fattorini, 2015). Once the survey variable has been estimated in each spatial unit, these spatially explicit estimates are used for creating the map of the survey variable distribution in the whole region.

(a) (b)

Figure 1.1 A survey region partitioned into a collection of irregular patches (a) and the same region partitioned into pixels (b).

Similarly, sometimes the survey variable is defined on the continuum of the whole survey area (Figure 1.2), but also in this case the survey variable is recorded only for a sample of points and estimated in the non-sampled ones.

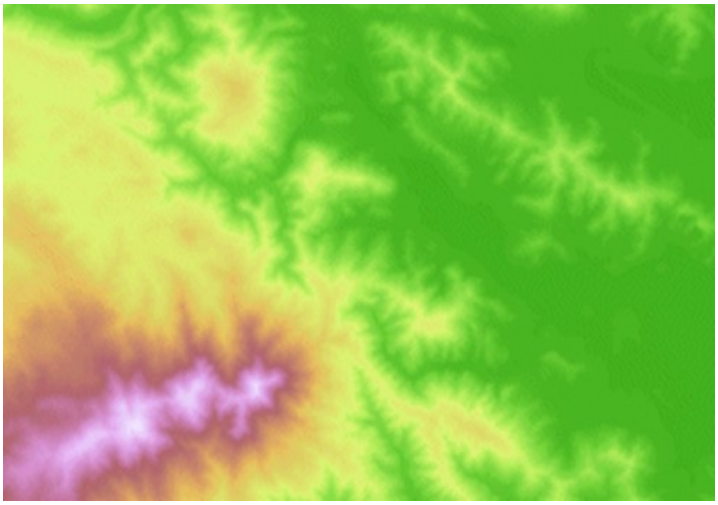


Figure 1.2. A survey variable defined on the continuum of the whole region of interest.

Most statistical methodologies applied for wall-to-wall mapping rely on model-based inference. This trend reflects the skeptical approach of geostatisticians to design-based methods, as pointed out by de Gruijter and ter Braak (1990), but the contraposition between the two approaches is a deeply rooted issue. Indeed, the differences between model-based and design-based inference are well delineated in both statistical literature

(e.g. Thompson, 2002) and forestry applications (e.g. Gregoire and Valentine, 2008). Särndal et al. (1992) described the design-based approach as *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*. On the other hand, model-based inference views the population as realization of a random process, called super-population, therefore making assumptions about the mechanism generating the super-population and it views the sampled units as fixed, i.e. purposively selected. When model assumptions hold, it is possible to achieve optimal predictors and their related model-based properties (Corona et al., 2014).

However, the choice of using a model in the process of making maps does not exclusively rely on the geostatisticians' disbelief in design-based inference. Indeed, the crucial reason of this choice is the impossibility to estimate non-sampled values without any assumptions. From a design-based point of view, either a unit – or a point in the continuous framework – is sampled and there is no need for estimation, or it is unsampled and there are no information to perform the estimation. Thus, the use of a model seems the sole way to recover the information in the non-sampled units or points. For these reasons, the most common techniques adopted to perform spatial interpolation are of model-based nature, such as kriging (Cressie, 1993) – and its numerous modifications – and k-nearest neighbour (k-NN) (McRoberts et al., 2007) predictors. These methods are described in Chapter 2, together with a novel approach proposed by Fattorini et al. (2018a), in which an assisting model is used to retrieve the information in the unsampled units when working in a design-based framework. Following this approach, in Chapter 3 a method for forest map reconstruction in a design-based model-assisted framework is presented. This method exploits auxiliary sources of information, such as remote sensing data, whereas the sole spatial information is used in Fattorini et al. (2018a). A simulation study and a real case application are also present and the results are discussed in Chapter 4 and Chapter 5, respectively. Lastly, Chapter 6 contains final remarks and future developments.

Chapter 2

State-of-the-art

2.1 Model-based methods

Model-based methodologies can be classified into three major groups: parametric, semi-parametric and non-parametric. Contrary to parametric methods, non-parametric approaches do not require to fully specify the distribution of the survey variables; a mixed approach using fully specified distributions subsequently corrected by data-driven procedures is referred to as semi-parametric.

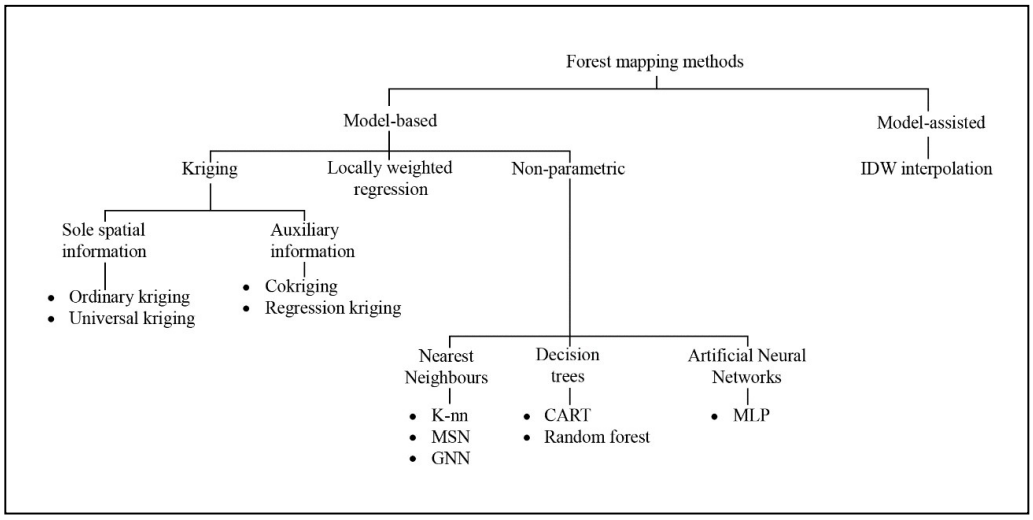


Figure 2.1. Classification of literature regarding statistical methods applied for forest mapping.

Due to the vastness of model-based methodologies adopted for mapping, only few methods, chosen on the basis of their popularity in forest studies, are here presented. The selected methods (Figure 2.1) include kriging and its variations as parametric methods, locally weighted regression as a semi-parametric method and nearest neighbor techniques, decision trees and artificial neural networks as non-parametric methods. Despite their popularity, only a few comparative studies have been conducted on the performance of these model-based techniques for forest attribute prediction. Moreover, all these works involve severe assumptions, which means that the results are only valid

in quite unrealistic situations. Indeed, Corona et al. (2014) proved that, from a design-based perspective, i.e. by holding the populations fixed during the simulation runs and selecting plots by means of tessellation stratified sampling (TSS), all model-based methods turned out to be biased with non-negligible bias, especially when the correlation between the interest and the auxiliary variable was weak.

2.1.1 Kriging

In geostatistics, kriging denotes a set of procedures for spatial prediction, i.e. the estimation of the value of the interest variable at non-sampled locations. Matheron (1963) named the method after the proposal of the mining engineer Danie Gerhardus Krige, even though the formulation of spatial prediction did not come from Krige's work (Cressie, 1993).

Kriging is *a minimum-mean-squared error method of spatial prediction* (Cressie, 1993, p. 106) predicting the value of the interest variable at non-sampled locations as linear combinations of the values collected at the sampled locations. However, it should be noted that, even though kriging is a linear predictor (e.g. Journel and Huijbregts, 1978), nonlinear spatial prediction methods are now part of the "kriging family" (Cressie, 1990) as well. Methods of this family are summarized by Schabenberger and Pierce (2002).

The kriging predictor is based on the assumption that an intrinsically stationary process has generated the super-population and it can also be used for estimating the values of a finite population of spatial units, although it was originally developed for the interpolation of continuous surfaces. Its simplest version is the *ordinary kriging* predictor, but it is rarely used in forest mapping because forest inventory data rarely show a "pure" spatial autocorrelation trend, which is the main assumption of this technique. Freeman and Moisen (2007) used ordinary kriging with the aim of improving the point prediction accuracy of the nationwide forest biomass map, but they found out that neither the field biomass nor the residual biomass were good candidates for ordinary kriging.

A generalization of ordinary kriging widely applied in forest studies is *universal kriging* (Matheron, 1969). The relevant feature of this method is that the mean is "driven" by an ancillary data available for the whole study region and in addition to the sampled locations. Obviously, when no auxiliary data are available, universal kriging reduces to

ordinary kriging. Lochhead et al. (2018) recently tested this technique to provide wall-to-wall information on forest attributes and even though universal kriging was very accurate, its implementation for a macroscale was computationally complex.

Regression kriging (Odeh et al., 1995) is a method combining a regression model with kriging, quite similar to universal kriging. Indeed, many authors (e.g. Deutsch and Journel, 1992) use the term "universal" when the deterministic trend is supposed to be a function of the coordinates only, whereas the term "regression" is used when the trend is supposed to be a function of some auxiliary variables. In regression kriging, trend and residuals are usually estimated separately and summed afterwards (Odeh et al., 1995). An example of regression kriging implementation is provided by Brus et al. (2012), where a statistical mapping of forest tree species across Europe was performed. In this paper, the ICP Forests plots were extended with the NFI plot data of eighteen countries, using a soil map, a biogeographical map and bioindicators derived from temperature and precipitation data were used as predictors. The estimated overall accuracy was 43% with the sole ICP Forests plots and 57% in areas with NFI plots, but this gain was mainly attributable to the much denser plot data, less to the prediction method.

An alternative approach for exploiting ancillary data is constituted by *cokriging* methods, where both auxiliary and interest variables are presumed to be intrinsically stationary random processes. An interesting comparison between ordinary kriging, universal kriging, cokriging and regression kriging was published by Meng et al. (2009) for mapping the basal area using Landsat ETM+ images as auxiliary data. In this case, regression kriging resulted in the smallest errors and the highest R-squared indicating the best geostatistical method for spatial predictions.

2.1.2 Locally weighted regression

Locally weighted regression (LWR), also referred to as *loess model*, was introduced by Cleveland and Davlin (1988) and its principal assumption is that the data are generated by a semi-parametric model. The use of LWR is particularly recommended when the absence of stationarity in the relationship between interest and auxiliary variables reduces the efficiency of kriging methods (e.g. Brunsdon et al., 1996). This technique has been often used to study spatial trends or to derive insights on single tree growth trends. For example, Wang et al. (2005) successfully implemented LWR to obtain a net primary production regression model and including spatial non-stationary in the

parameters estimated for forest co-systems in China, obtaining better results compared to least squares regression.

2.1.3 Nearest neighbour techniques

Non-parametric models have been proposed as an alternative to parametric prediction in order to account for spatial heterogeneity. Among non-parametric approaches, the popularity of nearest neighbour (NN) techniques has quickly raised through researchers, especially in forest applications (McRoberts et al., 2010b; Baffetta et al., 2012). Main advantages of NN methodologies include: i) prediction in both univariate and multivariate cases and ii) no assumptions on the distribution of the variables (McRoberts et al., 2010a).

Practically speaking, NN methods predict unsampled values using a linear combination of observations that are nearest or most similar to the location to be predicted in the space of the auxiliary variables.

The most general NN technique is the *k-NN* method, which includes most NN techniques adopted in forest studies (McRoberts, 2012). However, there are no theoretical studies on the model-based properties of k-NN (Corona et al., 2014) and the only theoretical attempt (McRoberts et al., 2007) involves severe, restrictive assumptions. k-NN is probably the most used technique in forest mapping from NFI plots (McRoberts and Tomppo, 2007). For example, Reese et al. (2003) used k-NN to provide a synoptic coverage of forest areas in Sweden.

Most similar neighbour (MSN) (Moeur and Stage, 1995) and *gradient nearest neighbour* (GNN) (Ohmann and Gregory, 2002) are two k-NN techniques that only use a single neighbor. This neighbor is chosen using canonical correlation analysis (Hotelling, 1936) for MSN and direct gradient analysis (Gauch, 1982) and stepwise canonical correspondence analysis (ter Braak, 1986) in the case of GNN.

2.1.4 Decision trees

A decision tree (Carbonell et al., 1983) is a non-parametric method that recursively partitions the space of auxiliary variables into classes, which are determined from the sampled values (McRoberts et al., 2010b). Decision trees are widely used for prediction of forest attributes (Helmer et al., 2010) because they can accommodate non-linear responses, continuous and categorical auxiliary variables, missing data and collinear

variables (Urban, 2002). On the other hand, it can be difficult to select appropriate model parameters (Brosofske et al., 2014).

Classification and Regression trees (CARTs) (Breiman et al., 1984) are single decision tree models predicting categorical (classification tree) or continuous (regression tree) variables. An interesting case study was provided by Torresan et al. (2016), where metrics extracted from a LiDAR sensor were exploited to predict different forest structure types by means of classification trees, obtaining moderately satisfactory results.

Among the methods based on the construction of a multitude of decision trees, *random forest* (Breiman, 2001) is probably the most used. The random forest algorithm uses CARTs constructed from bootstrap samples and each fully grown tree is used to predict the out-of-bag data, i.e. the data not in the bootstrap sample. Random forest was used by Hudak et al. (2008) for imputing basal area and tree density, obtaining overall good results.

Other decision tree methodologies, such as bagging (Breiman, 1996) and boosting (Schapire et al., 1998), have not been here considered because they are not common in forest applications (Brosofske et al., 2014).

2.1.5 Artificial neural networks

Artificial neural networks (ANNs) are prediction techniques inspired by the biological neural network of animal brains and its processing information system (McRoberts et al., 2010b). Typically, an ANN is formed by a collection of single processing units linked by an assemblage of weighted connections (Agatonovic-Kustrin and Beresford, 2000) and a learning rule (Baret, 1995). An ANN gathers its knowledge by detecting patterns and relationships in the training data and, as learning proceeds, the weighted connections are iteratively adjusted; once all patterns and relationships in the training data are learned, the ANN can be used to predict unknown values (Carvalho, 2001). A shortcoming of this group of algorithms is their “black box” nature. Indeed, as Cartwright (2008) pointed out, ANNs are less suitable in situations in which the knowledge of the relationships underlying the predictions are of primary interest, because no information on these relationships are provided. Görgens et al. (2015) compared the performance of ANN and random forest for predicting stand volume of

fast-growing forest plantations. In this case, the random forest algorithm produced the lowest RMSE.

Among the several types of ANNs, *Multi Layer Perceptron* (MLP) is the most widely applied in remote sensing studies (Carvalho, 2001). This algorithm has at least three layers of artificial neurons that, with the exception of the input ones, are activated by non-linear functions. Mas et al. (2004) analyzed deforestation processes in order to predict the spatial distribution of tropical deforestation, using a MLR to estimate the propensity to deforestation as a function of the explanatory variables. The model performance was quite high and able to classify correctly 69% of the grid cells.

2.2 Design-based methods

As stated in the Introduction (Chapter 1), pure design-based methods cannot be used for constructing wall-to-wall maps, owing to the impossibility to estimate non-sampled values without any assumptions. Specifically, from a design-based point of view, either a location is sampled and then there is no need for estimation, or it is unsampled and then there is no information to perform the estimation. Thus, the sole way to recover information in the unsampled locations is to use an assisting model (Fattorini et al., 2018a).

While the use of an assisting model is a widely adopted and effective way to estimate totals and averages of finite populations, that is not true for estimating single population values. Indeed, when estimating totals and averages the assisting model is used to predict each population value, but the total of errors performed by the model is estimated from the sample and then adopted to correct the total or average achieved from those predictions in accordance with the criterion provided by the difference estimator (Särndal et al., 1992). In this way, we achieve design-unbiasedness of the resulting estimators, or approximate unbiasedness up to the first term of approximation, as well as exact or approximate design-based variance expressions and suitable variance estimators, as happens for the well-known generalized regression and ratio estimators. On the other hand, when estimating the value at a single location, there is no way to correct the errors invariably provided by the model prediction. Thus, any model-assisted estimator at a single location and the subsequent map are destined to be design biased. Accordingly, as pointed out by Fattorini et al. (2018a), any map arising from a model-assisted criterion can achieve statistical soundness only if it is proven to be design-based

asymptotically unbiased and consistent (DBAU&C). These issues have precluded the use of model-assisted inference in forest mapping.

Two recent works of Fattorini et al. (2018a; 2018b) deal with the condition ensuring DBAU&C for model-assisted maps for populations partitioned into spatial units and continuous surfaces, respectively. In particular, the authors used an inverse distance weighting (IDW) interpolator (Shepard, 1968) to estimate unsampled locations when the sole spatial information is available. The assisting model, i.e. the criterion leading to the IDW interpolator, is the simple Tobler's first law of geography (Tobler, 1970), asserting that a spatial location is more similar to the nearby locations than to those further apart. In accordance with this principle, the prediction of the survey variable at the unsampled units is obtained as a weighted sum of the sampled values with weights decreasing with the distance to the point to be predicted. This approach is described in detail in the next Chapter (Section 3.2).

Chapter 3

Design-based model-assisted mapping

3.1 Statement of the problem

Consider a study region $\mathcal{A}$ of size A, which is a connected and compact set of $\mathbb{R}^2$. Moreover, suppose that $\mathcal{A}$ is partitioned into N spatial units $\mathrm{a}_1,\ldots,\mathrm{a}_N$ of sizes $|a_1|,\ldots,|a_N|$. In the following, with a slight abuse of notation, U will denote both the set of spatial units partitioning $\mathcal{A}$ and the set of indexes $\{1,\ldots,N\}$ which identify the spatial units. For each unit, let $\mathrm{c}_j \in \mathrm{a}_j$ be the point identifying the spatial location of the unit (e.g. the centroid of the unit), in such a way that for each couple of units $\mathrm{a}_j, \mathrm{a}_h$ $(h > j = 1,\ldots,N)$ it is possible to determine the distance between the units, say $d_{jh} = \|\mathbf{c}_j - \mathbf{c}_h\|$, where $\|\ \|$ denotes a norm in $\mathbb{R}^2$. Furthermore, for each unit, let y_j be the amount of the survey variable Y within the unit a_j and $f_j = y_j / |a_j|$ its density.

Now suppose that we are interested in the reconstruction of the whole map of the spatial population, i.e. we want to estimate the value y_j for each $j \in \mathsf{U}$ on the basis of a sample S of n units selected from U by means of a sampling scheme inducing invariably positive first order inclusion probabilities π_j and second order inclusion probabilities $\pi_{jh}\,(h > j \in \mathsf{U})$. Moreover, let $X_1,\ldots,X_K$ be a set of K auxiliary variables whose values $\mathbf{x}_j = [x_{j1},\ldots,x_{jK}]^{\mathrm{T}}$ are known for each unit of the population.

Because in most situations the $|a_j|$'s are known for all the population units, it is theoretically equivalent to estimate y_j for each $j \in \mathsf{U}$ directly by means of a sample estimator $\hat{y}_j$ or to estimate its density f_j by means of a sample estimator $\hat{f}_j$ and then achieve the estimator of y_j as $\hat{y}_j = |a_j| \hat{f}_j$. While the two estimation criteria are equivalent for a finite population of N units, density estimation is more suitable when working in an asymptotic scenario in which the number of units partitioning the study

area increases in such a way that their size may tend to zero. In this case, there is no sense in considering the difference $\hat{y}_j - y_j$ for proving any sort of consistency. Obviously these problems are bypassed if we estimate densities, because they not necessarily approach zero as unit sizes decrease. For this reason, henceforth we deal with density estimation.

3.2 Design-based mapping from spatial information

Exploiting the Tobler's (1970) first law of geography in a model-assisted framework, Bruno et al. (2013) proposed to estimate the y_j's using the inverse distance weighting (IDW) interpolator (Shepard, 1968), in which the values of the unsampled units are estimated by a weighted sum of the sample values with weights inversely decreasing with distances to the units under estimation. Practically speaking, the IDW interpolator can be expressed as

$$\hat{f}_j = Z_j f_j + \left(1 - Z_j\right) \sum_{i \in \mathsf{U}} w_{ij} f_i \tag{1}$$

where $Z_j = \begin{cases} 1 & j \in \mathsf{S} \\ 0 & j \notin \mathsf{S} \end{cases}$ and $w_{ij} = \dfrac{Z_i \phi\left(d_{ij}\right)}{\sum_{l \in \mathsf{U}} Z_l \phi\left(d_{lj}\right)}$ is the weight attached to the density of unit *i* to estimate the density of unit *j*. The function $\phi : [0, \infty) \to \mathbb{R}^+$, with $\phi(0) = 0$, is a non-increasing function and $\lim_{d \to 0^+} \phi(d) = \infty$. It is worth noting that $\hat{f}_j$ is equal to the true density f_j for each sampled unit *j* and the sole random variables are the Z_j's, which describe the sampling outcome.

3.3 Asymptotic results

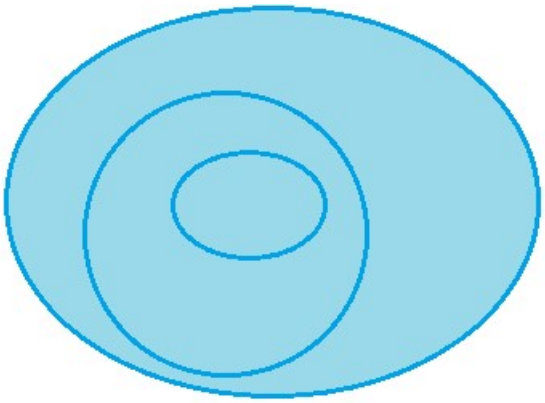
As any other model-assisted estimator of single population values, the IDW predictor is obviously biased. Therefore, the only way to render statistically sound a design-based model-assisted map is to determine the conditions ensuring design-based asymptotic unbiasedness and consistency (DBAU&C). Fattorini et al. (2018a) determined the conditions ensuring DBAU&C for design-based maps when the sole spatial information is available. Following the authors' approach, let consider a sequence $\{\mathsf{U}_k\}$ of partitions of $\mathcal{A}$. Each partition U_k is constituted of N_k spatial units $\mathrm{a}_1^{(k)}, \ldots, \mathrm{a}_{N_k}^{(k)}$ of measures

$\left|a_1^{(k)}\right|,\ldots,\left|a_{N_k}^{(k)}\right|$, whose centroids are identified with $c_1^{(k)},\ldots,c_{N_k}^{(k)}$ and $d_{jh}^{(k)} = \left\|\mathbf{c}_j^{(k)} - \mathbf{c}_h^{(k)}\right\|$ is the distance between the units. Furthermore, let $y_j^{(k)}$ be the amount of the survey variable Y within the unit $\mathrm{a}_j^{(k)}$ and $f_j^{(k)} = y_j^{(k)} \Big/ \left|a_j^{(k)}\right|$ its density. Suppose that $\mathcal{A}$ is partitioned into an increasing number of spatial units, i.e. $N_k \to \infty$, so that all the units decrease in size as k increases in such a way that $\sup_{j\in \mathsf{U}_k} \mathrm{diam}\left(\mathrm{a}_j^{(k)}\right) \to 0$. Finally, suppose a sequence of designs to select a sample S_k of size n_k from U_k. For each $h > j \in \mathsf{U}_k$, $\pi_j^{(k)}$ and $\pi_{jh}^{(k)}$ denote the first- and second-order inclusion probabilities induced by these designs and denote by $Z_j^{(k)}$ the indicator variables that equals 1 if unit j is selected from the k population and 0 otherwise. In this way, it is possible to determine the asymptotic design-based behavior of the IDW interpolator (1) as the partition of the study area becomes thinner and thinner. From (1) it follows that the IDW interpolator of $f_j^{(k)}$ can be calculated as

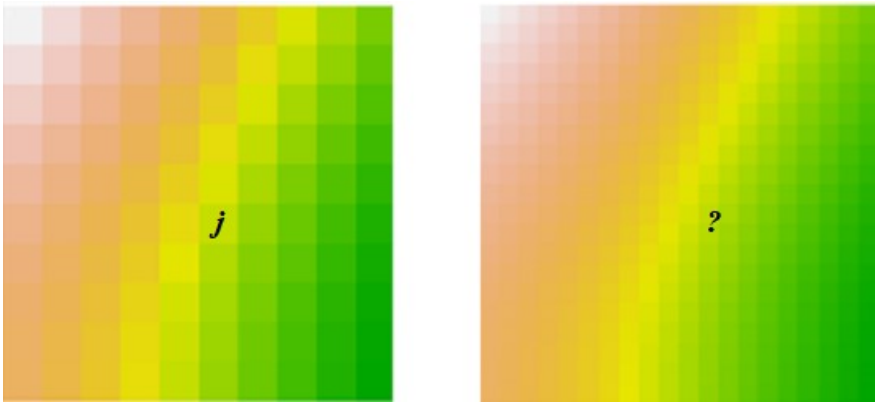
$$\hat{f}_j^{(k)} = Z_j^{(k)} f_j^{(k)} + \left(1 - Z_j^{(k)}\right) \sum_{i\in \mathsf{U}_k} w_{ij}^{(k)} f_i^{(k)}$$

where $w_{ij}^{(k)} = \dfrac{Z_i^{(k)} \phi\left(d_{ij}^{(k)}\right)}{\sum_{l\in \mathsf{U}_k} Z_l^{(k)} \phi\left(d_{lj}^{(k)}\right)}$.

It is worth noting that, contrary to the finite population asymptotic scenario usually adopted in design-based inference (e.g. Isaki and Fuller, 1982) in which the populations are nested, in this case (Figure 3.1a), the unit j of the kth population is lost in the subsequent populations (Figure 3.1b).



(a)



(b)

Figure 3.1. Nested populations as common in the finite population asymptotic scenario (a) and thinner partitioning of the area in such a way that unit *j* of the *k*th population is lost in the subsequent populations (b).

In order to handle this problem, two sequences of functions from $\mathcal{A}$ onto $\mathbb{R}^+$ are introduced:

$$f_k(\mathbf{x}) = \sum_{j \in \mathsf{U}_k} f_j^{(k)} I\left(\mathbf{x} \in \mathrm{a}_j^{(k)}\right),\ \mathbf{x} \in \mathcal{A},$$

in which the densities of the *k*th population are substituted by a piecewise constant function in $\mathcal{A}$, and

$$\hat{f}_k(\mathbf{x}) = \sum_{j \in \mathsf{U}_k} \hat{f}_j^{(k)} I\left(\mathbf{x} \in \mathrm{a}_j^{(k)}\right),\ \mathbf{x} \in \mathcal{A},$$

where the density estimators are substituted by a piecewise constant random process with parameter $\mathbf{x}$ onto $\mathcal{A}$. With this notation, the IDW interpolator is defined to be point-wise design consistent at $\mathbf{x} \in \mathcal{A}$ if $p\lim_{k\to\infty} \left|\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x})\right| = 0$ and it is uniformly

design consistent onto $\mathcal{A}$ if $p\lim_{k\to\infty}\left\|\hat{f}_k - f_k\right\|_\infty = 0$, where $\|\ \|_\infty$ is the uniform norm. It is worth noting that the point-wise design consistency implies the convergence in probability of $\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x}) \to 0$ at the single point $\mathbf{x} \in \mathcal{A}$, whereas the uniformly design consistency requires the convergence in probability of $\sup_{\mathbf{x}\in\mathcal{A}}\left|\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x})\right| \to 0$.

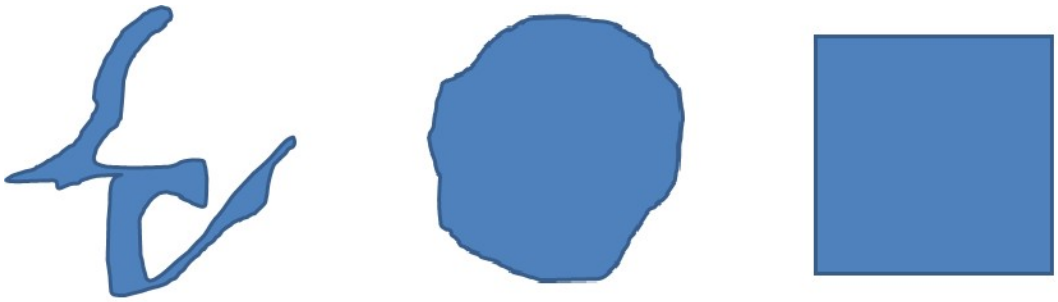
At this point, we can introduce the four DBAU&C conditions determined by Fattorini et al. (2018a). The design consistency of the IDW interpolator (1) is achieved assuming that there exists a Riemann integrable function f from $\mathcal{A}$ onto $[0, L]$, which gives the density of the survey variable *Y* at any point of the study area. This assumption, referred to as Riemann integrability (RI) assumption, implies that, since the $f_j^{(k)}$'s are uniformly bounded by *L*, the $\hat{f}_j^{(k)}$'s are also bounded by *L*, in such a way that point-wise or uniform design consistency of the IDW interpolator also entails point-wise or uniform design asymptotic unbiasedness.

(a) (b)

Figure 3.2. A real forest satisfies the first DBAU&C condition of discontinuity over region of measure zero if the density changes smoothly, i.e. continuously, throughout space (a) or if it changes abruptly along borders, such as those that divide forest from non-forest (b).

It should be pointed out that the RI assumption is reasonably valid in many real situations. Indeed, in real world the density of a variable changes smoothly throughout space (continuity) (Figure 3.2a) or abruptly, usually along borders delineating variations in the characteristics of the study region (Figure 3.2b). The borders may be realistically approximated by curves well approaching the theoretical condition of discontinuity over a region of zero measure.

In order to provide sufficient conditions for point-wise and uniform design consistency, some assumptions about the shapes of the units are needed. In particular, we assume the existence of a constant $u > 0$ such that $\sqrt{N_k} \sup_{j \in \mathsf{U}_k} \operatorname{diam}\left(\mathrm{a}_j^{(k)}\right) \leq u$. This assumption ensures that, when working with spatial units, their diameters approach 0 and there are no excessively elongated unit, in such a way that units can be covered by smaller and smaller balls. It follows that the use of pixels or regular polygons, such as hexagons, is an appropriate choice (Figure 3.3), whereas the use of irregular patches does not satisfy this second DBAU&C condition.



(a) (b) (c)

Figure 3.3. Examples of unit shapes that can be used to partition a study region. In particular, (a) does not satisfy the second DBAU&C condition, while (b) and (c) do.

We also assume that for any $\mathbf{x} \in \mathcal{A}$ and any $\varepsilon > 0$ there exists an integer $T > 1$ and an integer k_0 such that $\Pr\left\{Z_k\left(\mathbf{x}, uT/\sqrt{N_k}\right) = 0\right\} < \varepsilon \quad \forall k > k_0$, where $Z_k(\mathbf{x}, \delta)$ represents the number of units of U_k selected among the neighbors of the unit which contains $\mathbf{x}$. This condition states that the probability of selecting no units near the unit containing $\mathbf{x}$ becomes definitely smaller than ε for $k > k_0$. Practically speaking, this condition ensures a spatial balance asymptotically achieved by the sampling scheme, in such a

way that, for a sufficiently large *k*, the scheme is able to evenly spread out the sampled units onto the study region and any unit is likely to have neighboring units sampled. The spatial balance required for the sampling scheme is ensured under some widely applied schemes, such as systematic sampling (Breidt, 1995) (SYS), one-per-stratum-stratified (OPSS) sampling scheme (Breidt, 1995) and simple random sampling without replacement (SRSwoR). The OPSS sampling scheme has been used in both the simulation study (Chapter 4) and the case study (Chapter 5) and for this reason it will be described in detail in the next section.

There exist many sampling schemes explicitly constructed to work with spatial units that guarantee spatially balanced samples, such as the generalized random-tessellation stratified sampling (Stevens and Olsen, 2004) and the doubly balanced spatial sampling (Grafström and Tillé, 2013). These schemes are so complex that they cannot be proven to meet the DBAU&C conditions, albeit their effectiveness in providing spatial balance may induce to presume that they satisfy the DBAU&C conditions as well (Fattorini et al., 2018a). However, when sampling from regular grids, Fattorini et al. (2015) demonstrated that those more complex schemes provide similar results to those achieved by means of OPSS and SYS, which are significantly simpler to apply.

Finally, the last consistency condition concerns the distance function ϕ, requiring that $\lim_{d\to 0^+} d^2\phi(d) = \infty$. Negative powers of distances such as $\phi(d) = d^{-\alpha}$ meet this requirement for $\alpha > 2$.

Regarding the application of design-based maps in forest studies, it is worth noting that the conditions ensuring DBAU&C seem to match some real situations usually encountered. Indeed, beside the condition on the distance function, which may be ensured by the user, the existence of a bounded and RI density of the survey variable onto the study area is realistic when changes occur along borders. However, when frequent discontinuities occur within units, the precision of the map can be deteriorated. In these cases, Fattorini et al. (2018a) suggests to handle this problem by performing prediction at a sufficiently small grain and then aggregating the results in a greater-grain map. As the grain at which prediction is performed approaches 0, design consistency holds for the aggregated map. Moreover, the use of data recorded at pixel level, such as remote sensing data, satisfies the requirement of regularity in unit size and shape. Because ecologists prefer maps at a wider scale than the pixel scale, these aggregations

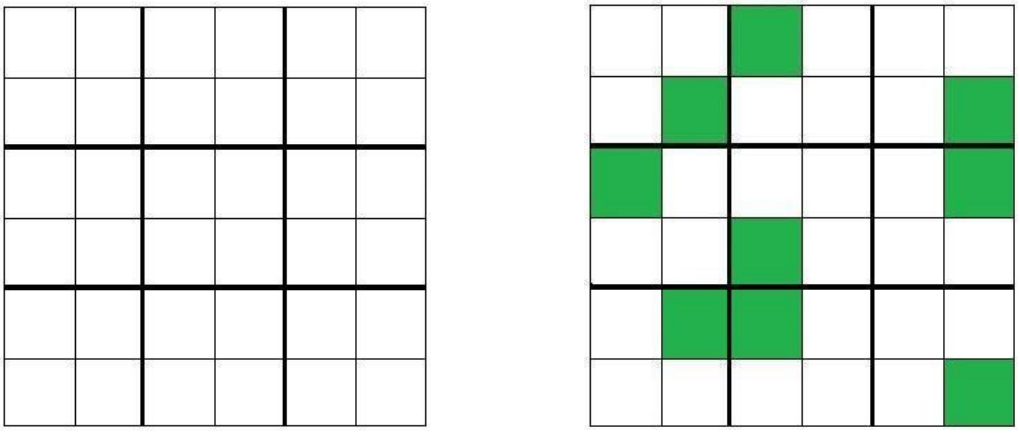
are likely to bypass the discontinuities at pixel level, ensuring the design consistency of the resulting aggregated maps. Lastly, many sampling designs applied for forest mapping satisfy the assumption of spatial balance.

3.4 One-per-Stratum-Stratified Sampling

The sample S can be selected by means of several sampling schemes inducing invariably positive first-order inclusion probabilities π_j and second-order inclusion probabilities π_{jh} $(h > j \in \mathrm{U})$. In this work, we chose to select the sample by means of the OPSS sampling scheme that constitutes a widely adopted design in forest inventories (e.g. the Italian NFI).

With this sampling procedure, the N spatial units of the population are grouped into n blocks of contiguous units and then a unit is selected in each block. The selection within each block can be performed by several sampling schemes, but we chose to select units within each block completely at random (Figure 3.4). In order to ensure an equal distribution of the sample into the area, which is one of the most important objective of spatial sampling, the blocks should have, at least approximately, the same number of units $N_l \approx N/n$ for each $l = 1,\ldots,n$.

The main advantage of this sampling scheme is the spreading out of the sampled units all over the area.



(a) (b)

Figure 3.4. Area partitioned into $N = 36$ units, which are grouped into $n = 9$ blocks (a), and random selection of one unit in each block (b) by means of OPSS. The 9 green units are the 9 elements of the sample S.

OPSS first-order inclusion probabilities are

$$\pi_j = \frac{1}{N_l} \text{ if } j \in \mathrm{U}_l$$

If all the blocks have the same number of units, OPSS first-order inclusion probabilities are equal.

The second-order inclusion probabilities are

$$\begin{cases} \pi_{jh} = 0 & \text{if } \; j,h \in \mathrm{U}_l \\ \pi_{jh} = \dfrac{1}{N_l N_k} & \text{if } \; j \in \mathrm{U}_l, h \in \mathrm{U}_k \end{cases}$$

Indeed, if units *j* and *h* are in the same block, they cannot be sampled together and therefore the second-order inclusion probability is equal to zero. On the other hand, if *j* and *h* are not in the same block, the second-order inclusion probability is given by the product of their first-order inclusion probabilities.

3.5 Design-based mapping exploiting auxiliary information

Even if the IDW interpolator is leaded by Tobler's first law of geography and as such it moves in the framework of model-assisted estimation, it exploits the sole information provided from space, i.e. unit locations. However, remote-sensing imagery usually provides a large set of auxiliary variables available for the whole study area such as elevation, slope and a huge sequence of spectral bands. These auxiliary data constitute a great opportunity to improve spatial interpolation. That can be readily done extending the difference estimation criterion adopted for inference on population totals and averages (Särndal et al., 1992), predicting the values of the interest variable as a suitable function of the auxiliary variables and then interpolating the prediction errors by means of the IDW interpolator.

Therefore, let y_j^0 be the amount of a proxy for y_j that, at the moment, is supposed to be known for each spatial unit and let f_j^0 be the corresponding density. With the purpose of performing spatial interpolation exploiting the proxy information, we considered the following relationship

$$y_j = y_j^0 + \varepsilon_j \tag{2}$$

where ε_j is the error occurred in predicting y_j by means of y_j^0. It is worth noting that (2) is simply a reckoning relation and does not involve any assumption. Dividing both sides of relations (2) by $|a_j|$ it follows that

$$f_j = f_j^0 + e_j \tag{3}$$

where $e_j = \varepsilon_j / |a_j|$ denotes the (positive or negative) intensity of the prediction error for the unit *j*. However, if y_j^0 actually provides a good proxy for y_j, it is quite natural to presume that the errors ε_j and the corresponding intensities e_j should be smoothed. Indeed, from (2) it follows that $\varepsilon_j = y_j - y_j^0$, in such a way that the errors are given by the difference between survey variable and proxy values. Then, if y_j^0 is a good proxy for y_j, the prediction errors ε_j are likely to be small for every *j*.

Therefore, Tobler's first law of geography can be exploited to estimate the e_j's by means of the IDW interpolator. Practically speaking, for each $j \in \mathsf{S}$, the intensity of the prediction error $e_j = f_j - f_j^0$ is known, in such a way that for each $j \in \mathsf{U}$ it can be interpolated by

$$\hat{e}_j = Z_j e_j + (1 - Z_j) \sum_{i \in \mathsf{U}} w_{ij} e_i \tag{4}$$

where Z_j is the usual random variable equal to 1 if $j \in \mathsf{S}$ and 0 otherwise, $w_{ij} = \dfrac{Z_i \varphi(d_{ij})}{\sum_{l \in \mathsf{U}} Z_l \varphi(d_{lj})}$ is the weight attached to the error intensity of unit *i* to estimate the error intensity of unit *j* and $\varphi : [0, \infty) \to \mathbb{R}^+$ is a not increasing function with $\varphi(0) = 0$ and $\lim_{d \to 0^+} \varphi(d) = \infty$. From (2) and (4), the spatial interpolator of f_j based on proxy information is given by

$$\hat{f}_j = f_j^0 + \hat{e}_j \tag{5}$$

It is worth noting that (5) is a genuine interpolator because (4) is a genuine interpolator. Indeed, when $j \in \mathsf{S}$, $\hat{e}_j = e_j$, in such a way that

$$\hat{f}_j = f_j^0 + \hat{e}_j = f_j^0 + e_j = f_j^0 + f_j - f_j^0 = f_j$$

Because the f_j^0's are constant, the bias and the variance of the IDW interpolator depend on the bias and the variance of the $\hat{e}_j$'s.

For achieving consistency of the interpolator (5) the following assumption is introduced: there exists a Riemann integrable function φ from $\mathcal{A}$ onto $[-L, L]$ which gives the intensity of the prediction error ε at any point of the study area. This assumption recalls the first DBAU&C condition of the IDW interpolator when the sole spatial information is used, the only difference being that φ assumes values in $[-L, L]$, instead of $[0, L]$. However, the results achieved for variables on $[0, L]$ are also valid for variables on $[-L, L]$.

As previously stated, the RI assumption of smoothness for the surface of the density of the interest variable is quite reasonable in many real situations, even when the density changes abruptly throughout the survey region. Nevertheless, this assumption of smoothness becomes even more realistic when considering error surfaces. Indeed, in these cases, jumps and irregularities in the density of the survey variable are absorbed by presumably similar jumps and irregularities of the proxy.

Consequently, the interpolator (5) can be proven to be design-based asymptotically unbiased and consistent under the same conditions defined for (1), i.e. Riemann integrability, shape regularities, asymptotical spatial balance and suitable weighting function, with the last three conditions ensured by the researcher.

3.6 Prediction based on generalized regression criterion

Let $\mathbf{x}_j = \left[x_{1j}, \ldots, x_{Kj}\right]^{\mathrm{T}}$ be the amounts of K auxiliary variables $X_1, \ldots, X_K$ within a_j which are known for each spatial unit of the population and let $\mathbf{g}_j = \left[g_{1j}, \ldots, g_{Kj}\right]^{\mathrm{T}}$ be the corresponding densities $g_{kj} = x_{kj} \big/ |a_j|$. Then, as in the generalized regression (GREG) estimation criterion (Särndal et al., 1992), the quantity $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$ is adopted as a proxy for y_j, in such a way that

$$y_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j + \varepsilon_j \tag{6}$$

where ε_j is the error occurred in predicting y_j by means of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$. If $\boldsymbol{\beta}$ is a vector of known parameters, then $y_j^0 = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$ and it would be possible to estimate the values of y_j for each units of the population using the IDW interpolator based on the difference criterion as in (5). However, the $\boldsymbol{\beta}$'s are unknown parameters and they need to be estimated from the sample. Once again, we divide both sides of the relation (6) by $|a_j|$, so that we obtain $f_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}_j + e_j$. The GREG model (Särndal et al., 1992) can also be written in vector notation as $\mathbf{f} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{G} + \mathbf{e}$, where $\mathbf{f} = [f_1,\ldots,f_N]^{\mathrm{T}}$ is the vector of population densities, $\mathbf{G} = [\mathbf{g}_1,\ldots,\mathbf{g}_N] = \begin{bmatrix} g_{11} & \cdots & g_{N1} \\ \cdots & \cdots & \cdots \\ g_{1K} & \cdots & g_{NK} \end{bmatrix}$ is the $K \times N$ matrix of densities of the auxiliary variables and $\mathbf{e} = [e_1,\ldots,e_N]^{\mathrm{T}}$, with expectations and variances presumed by the model and equal to $\mathrm{E}_m(\mathbf{e}) = \mathbf{0}$ and $V_m(\mathbf{e}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sigma_N^2 \end{bmatrix}$. It follows that the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^0 = \left[\sum_{j\in\mathrm{U}} \frac{\mathbf{g}_j\mathbf{g}_j^{\mathrm{T}}}{\sigma_j^2}\right]^{-1} \sum_{j\in\mathrm{U}} \frac{f_j\mathbf{g}_j}{\sigma_j^2}$.

It is evident that the ordinary least squares estimate $\boldsymbol{\beta}^0$ is unknown, because it requires the knowledge of all the population densities and all the variances. However, the population variances should have a structure such that they disappear in the computation of $\boldsymbol{\beta}^0$ without being estimated (Särndal et al., 1992) and the densities f_j are known for each $j \in \mathrm{S}$. Consequently, $\boldsymbol{\beta}^0$ can be estimated from the sample using the well-known Horvitz-Thompson (HT) criterion, in such a way that the GREG estimator of $\boldsymbol{\beta}$ can be calculated as

$$\hat{\boldsymbol{\beta}} = \left[\sum_{j\in\mathrm{S}} \frac{\mathbf{g}_j\mathbf{g}_j^{\mathrm{T}}}{\pi_j\sigma_j^2}\right]^{-1} \sum_{j\in\mathrm{S}} \frac{f_j\mathbf{g}_j}{\pi_j\sigma_j^2}$$

where the σ_j^2s vanish in the computation of $\hat{\boldsymbol{\beta}}$ (Särndal et al., 1992). It is worth noting that $\hat{\boldsymbol{\beta}}$ is obviously unbiased because it has been estimated with the HT criterion.

At this point, the intensity of the prediction error $e_j = f_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}_j$ can be estimated by $\tilde{e}_j = f_j - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j$ for any $j \in \mathsf{S}$. Then, for each $j \in \mathsf{U}$, the prediction errors can be interpolated by

$$\hat{\tilde{e}}_j = Z_j \tilde{e}_j + (1 - Z_j)\sum_{i \in \mathsf{U}} w_{ij}\tilde{e}_i \tag{7}$$

where Z_j and w_{ij} are as above. Then, the spatial interpolator of f_j based on GREG auxiliary information is given by

$$\hat{\tilde{f}}_j = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j + \hat{\tilde{e}}_j \tag{8}$$

It is worth noting that (8) is a genuine interpolator, in the sense that when $j \in \mathsf{S}$, $\tilde{e}_j = \hat{\tilde{e}}_j$ in such a way that

$$\hat{\tilde{f}}_j = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j + \hat{\tilde{e}}_j = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j + \tilde{e}_j = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j + f_j - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{g}_j = f_j$$

DBAU&C of the interpolator (8) should follow from the unbiasedness of $\hat{\boldsymbol{\beta}}$. This result is under study and its formal proof will constitute a future work.

3.7 LASSO regression

The regression coefficients $\boldsymbol{\beta}$ can also be estimated by means of other estimation criteria, not only with ordinary least squares (OLS) as shown in the previous section. Indeed, in the case study (Chapter 5), we adopted the least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996) for the estimation of $\boldsymbol{\beta}$. The LASSO is a procedure that, setting some of the regression coefficients to zero, performs both estimation and variable selection. Practically speaking, while OLS chooses the coefficients that minimize the residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n}\left(y_j - \beta_0 - \beta_1 x_{1j} - \ldots - \beta_K x_{Kj}\right)^2$$

the LASSO minimizes the quantity $RSS + \lambda\sum_{j=1}^{K}\left|\beta_j\right|$. Depending on the values of the parameter $\lambda > 0$, different values of the coefficients are obtained. The optimal λ is chosen by cross-validating a range of possible λ values and then selecting the one that minimizes the prediction error. It should be noted that if $\lambda = 0$, the coefficients are the

same as OLS regression; if $\lambda = \infty$, all coefficients are zero; if $0 < \lambda < \infty$, coefficients are between zero and those obtained with OLS.

3.8 Bootstrap MSE estimation

Bias and variance of (8) cannot be expressed in closed form and, accordingly, they cannot be estimated with any usual method. Therefore, a modification of the classical bootstrap methodology (Efron, 1979) has been used for their estimation, following an innovative technique whose consistency is still under study.

As to the bootstrap methods adopted in design-based inference, Mashreghi et al. (2016) have identified three classes:

- pseudo population bootstrap, in which a pseudo-population is first created from the sample and then a number of bootstrap samples is selected from it using the original sampling scheme;
- direct bootstrap, in which the bootstrap samples are selected from the original sample (or from a rescaled version of it) using a with replacement scheme differing from the original scheme;
- bootstrap weight methods, in which the sample remains fixed and bootstrap weights are generated instead of creating bootstrap samples.

Pseudo-population bootstrap methods seem the most convincing because they maintain the original spirit behind the classical Efron's bootstrap technique. However, the biggest issue with pseudo-population bootstrap is the reconstruction of the pseudo-population itself. Indeed, if the pseudo-population mimics the characteristics of the unknown population from which the sample is selected, then the bootstrap distribution generated from the bootstrap samples should mimic the distribution of any sample statistic.

Therefore, the novel, basic idea adopted in this study is to exploit the spatial map produced using the IDW interpolator as the pseudo-population from which a number of bootstrap samples is selected using the same spatial scheme adopted to achieve the original samples. Practically speaking, the estimated $\hat{\tilde{f}}_j$ values are used as the pseudo-population from which B bootstrap samples are selected. For each $b = \{1, \ldots, B\}$, a sample S_b^* is selected from the estimated map $\hat{\tilde{f}}_1, \ldots, \hat{\tilde{f}}_N$ by means of OPSS as well, using the same blocks of units adopted for the selection of the original sample S. Then,

from the *b*-bootstrap sample S_b^*, the estimated map $\hat{\tilde{f}}_{b1}^*,\ldots,\hat{\tilde{f}}_{bN}^*$ is achieved by repeating the IDW interpolation. Finally, the *B* bootstrapped maps are used to achieve the bootstrap estimates of MSEs

$$\hat{\mathrm{MSE}}_{\mathrm{boot}}^{(1)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\tilde{f}}_{b1}^* - \hat{\tilde{f}}_1\right)^2 \ , \ldots , \ \hat{\mathrm{MSE}}_{\mathrm{boot}}^{(N)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\tilde{f}}_{bN}^* - \hat{\tilde{f}}_N\right)^2$$

The square root of resulting MSE estimates, i.e. the root mean squared error (RMSE) estimates, were then calculated and used for representing the precision map.

Chapter 4

Simulation study

4.1 Monte Cimino experimental forest

To investigate the performance of the IDW interpolator, the exploitation of the auxiliary variables as well as the bootstrap MSE estimator, a simulation study was performed on a real population of beeches in Monte Cimino, Italy.

This beechwood, referred to as *Faggeta* by locals, is located in the Municipality of Soriano nel Cimino (Viterbo) and it extends on the north side of Monte Cimino, the highest peak (1053 m) of the Cimini Mountains, for about 60 ha between 950 and 1050 m asl.

Monte Cimino beechwood is an historical forest, the highest part of the infamous *Silva Ciminia* that, for a long time, impeded the Roman expansion in Etruria. Indeed, as Livy wrote (Ab Urbe Condita, Liber IX, 36, English translation by Spillan and Edmonds, 1849), *the Ciminian forest was in those days deemed as impassable and frightful as the German forests have been in latter times; not even any trader having ever attempted to pass it*.

Due to its historical importance, there is an enormous bibliography on the ecological evolution of the forest through the years (Lamonaca, 2005). In the last 150 years, the forest has been exploited for recreational purposes only (Ziaco et al., 2012), with no logging since 1949 (Lo Monaco, 1984). At present, the structure of the forest, which has trees higher than 50 m and older than 200 years (Ziaco et al., 2012), is so impressive that it is now part of the UNESCO World Heritage list, in the *Primeval forests of the Carpathian beech and other parts of Europe* category.

Within the borders of the *Faggeta*, it has been located an experimental forest, which covers a 320 m side squared area, for a total of 10.24 ha (Figure 4.1), and has been censed several times through the years. It follows that all the characteristics of the area and the trees within, such as tree locations, diameter at breast height (DBH), height, status (alive or dead), basal area, volume and many other, are known. In this study, we used the data collected in 2016.

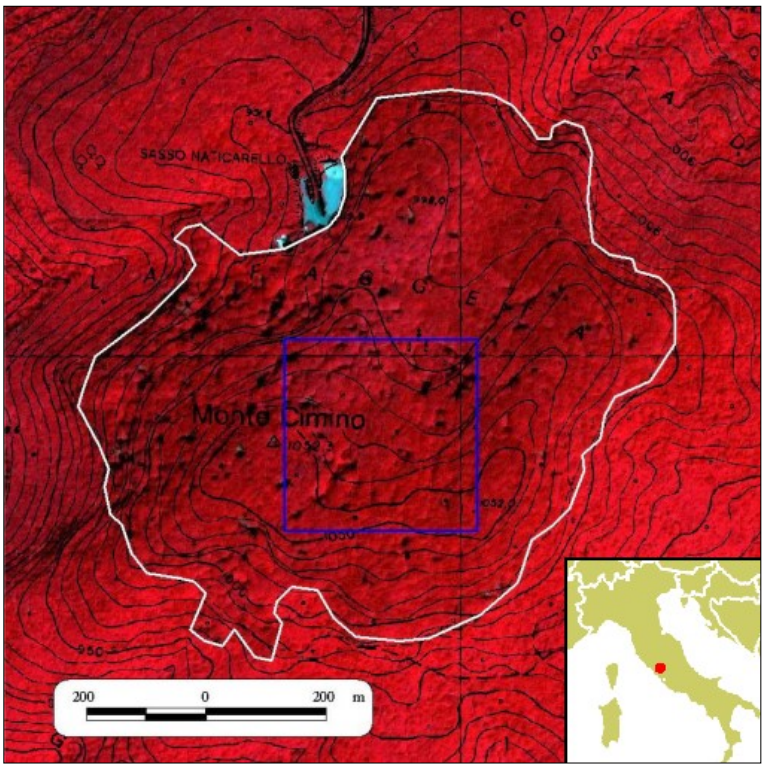


Figure 4.1. Map of the *Faggeta* of Monte Cimino (white border), with the experimental forest borders in blue (Lamonaca, 2005).

Moreover, Landsat-8 satellite images were available for the entire experimental forest and they were used as auxiliary information in the model-assisted estimation. This satellite was launched in February 2013 and carries a two-sensor payload (Irons et al., 2012), the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). OLI and TIRS images provide a seasonal coverage of the global landmass (Markham, 2011) with different spatial resolutions (Table 4.1). Actually, TIRS data are collected at 100 m spatial resolution and then they are resampled at 30 m to match the majority of OLI bands (Irons et al., 2012).

In order to create a perfect overlapping between the experimental forest and the Landsat-8 bands, the forest was partitioned into pixels, matching the Landsat-8 ones. Given that the Landsat-8 pixels have a 30 m side and the Cimino experimental forest has a 320 m side, it follows that a few Landsat-8 pixels were not completely included within the experimental forest borders. Consequently, they were excluded from the population, decreasing even more the surface of the experimental forest taken into consideration for this simulation study.

Table 4.1. OLI and TIRS spectral bands and relative wavelength (micrometers) and spatial resolutions (meters).

Band Name	Wavelength (μm)	Resolution (m)
Band 1 – Coastal Aerosol	0.433-0.453	30
Band 2 – Blue	0.450-0.515	30
Band 3 – Green	0.525-0.600	30
Band 4 – Red	0.630-0.680	30
Band 5 – Near Infrared (NIR)	0.845-0.885	30
Band 6 – Shortwave Infrared (SWIR) 1	1.560-1.660	30
Band 7 – Shortwave Infrared (SWIR) 2	2.100-2.300	30
Band 8 – Panchromatic	0.500-0.680	15
Band 9 – Cirrus	1.360-1.390	30
Band 10 – Thermal Infrared (TIRS) 1	10.6-11.2	100
Band 11 – Thermal Infrared (TIRS) 2	11.5-12.5	100

In particular, the study region covers 9 ha of the experimental forest and it was partitioned into 100 pixels of 30×30 m^2. Obviously, the real distribution of the basal area, i.e. the interest variable in this simulation study, is known for this portion of the experimental forest as well and the real spatial map of its distribution through the area is shown in Figure 4.2. For each pixel of the population, data from Landsat-8 acquisitions were exploited as auxiliary variables for the estimation of the basal area *G*.

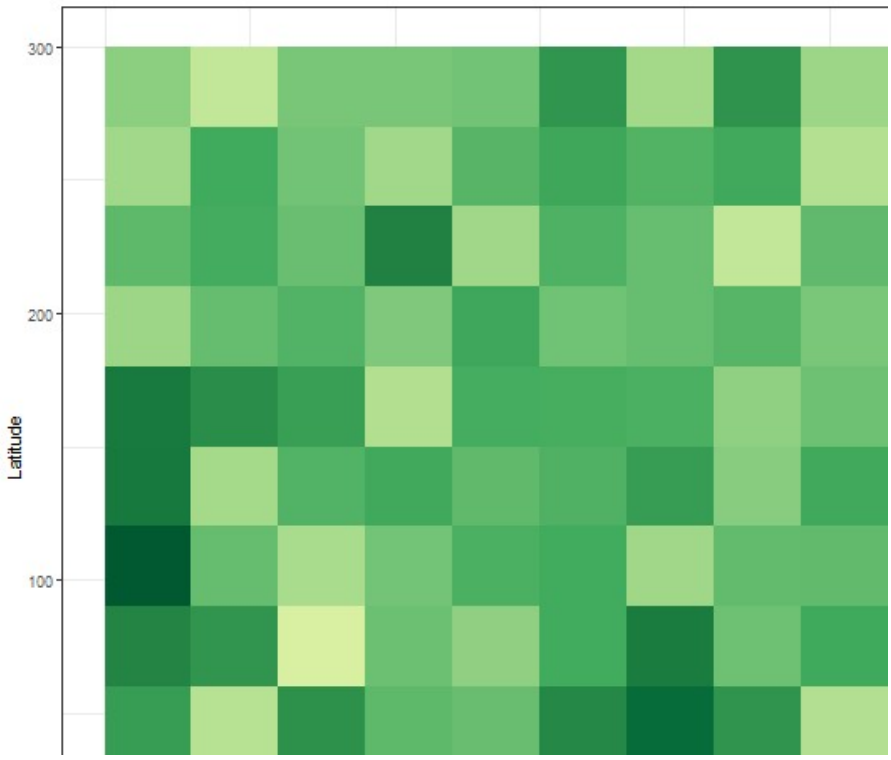


Figure 4.2. Spatial map of the real basal area distribution in the sub-population of Monte Cimino experimental forest considered for the simulation study.

4.2 Monte Cimino with 79 pixels

In order to obtain the best possible correlation between interest and auxiliary variables, a few pixels were discarded, reducing even more the region under investigation. Indeed, only 79 pixels out of 100 were considered (Figure 4.3).

Mimicking the multiple regression model in (6), the relationship between the interest variable G and the auxiliary Landsat-8 variables was modelled as follows

$$G = b_0 + b_1 \text{NDVI} + b_2 \cos(\text{EXP}) + e \tag{9}$$

where NDVI and cos(EXP) are two indexes calculated from Landsat-8 bands. Particularly, the Normalized Difference Vegetation Index (NDVI) is one of the most used index for the satellite assessment and monitoring of the vegetation cover (Huete and Liu, 1994), owing to its strong correlation with vegetation parameters such as green-leaf biomass and area (Justice et al., 1985). NDVI is defined as

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

where NIR and Red represent the near-infrared and visible red radiations, respectively. It should be noted that the regression model (9) explains very little of the response variability (adjusted R-squared $\cong 0.3$), but this population was used for the simulation study nevertheless. In particular, the samples were selected by means of OPSS, partitioning the population into 20 strata, with 19 strata of 4 pixels each and a stratum of 3 pixels. Then, a pixel per stratum was randomly selected.

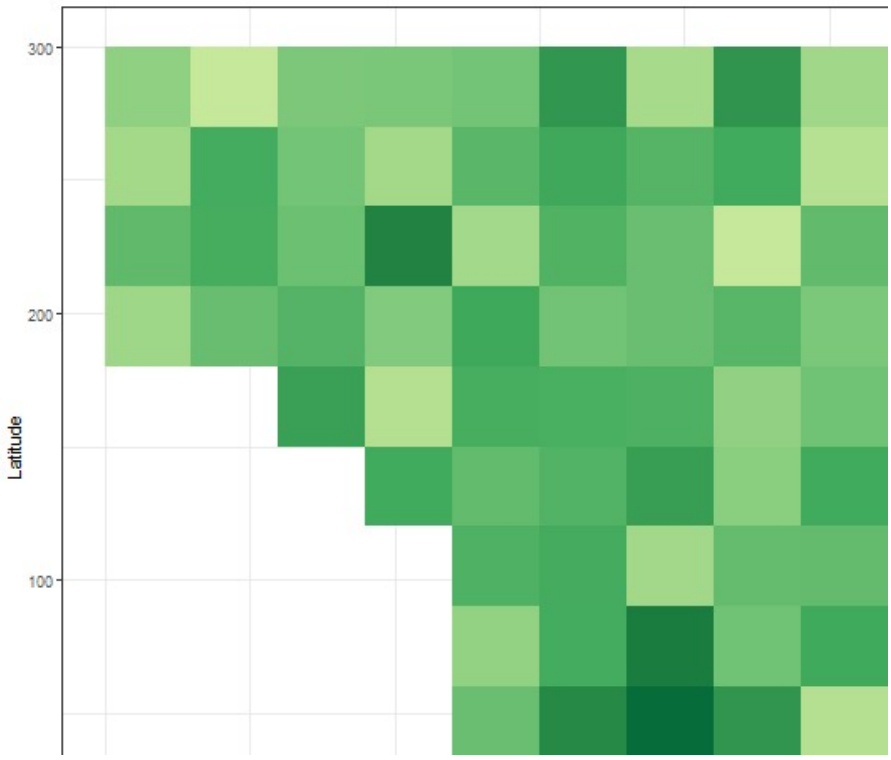


Figure 4.3. Spatial map of the real basal area distribution in the Monte Cimino population partitioned into 79 pixels.

4.3 Monte Cimino with 100 pixels

Due to the low correlation between interest and auxiliary variables in the previous population, a new simulation study was conducted on the area of Monte Cimino, using all 100 original pixels (Figure 4.2) and the same regression model (9). However, the values of the auxiliary variables NDVI and cos(EXP) were manipulated, so as to obtain a higher correlation and a better explanation of the response variability (adjusted R-squared $\cong 0.9$). In this case, the population was partitioned into $n = 25$ strata of 4 pixels each.

4.4 Monte Cimino with 400 pixels

In order to test out the performance of the described methodologies with an increasing number of sample cells, the Monte Cimino area was repartitioned into 400 square cells of 15×15 m^2, dividing each one of the original 100 pixels into 4 square sub-pixels. Those 400 pixels were grouped into $n = 100$ strata of 4 pixels each.

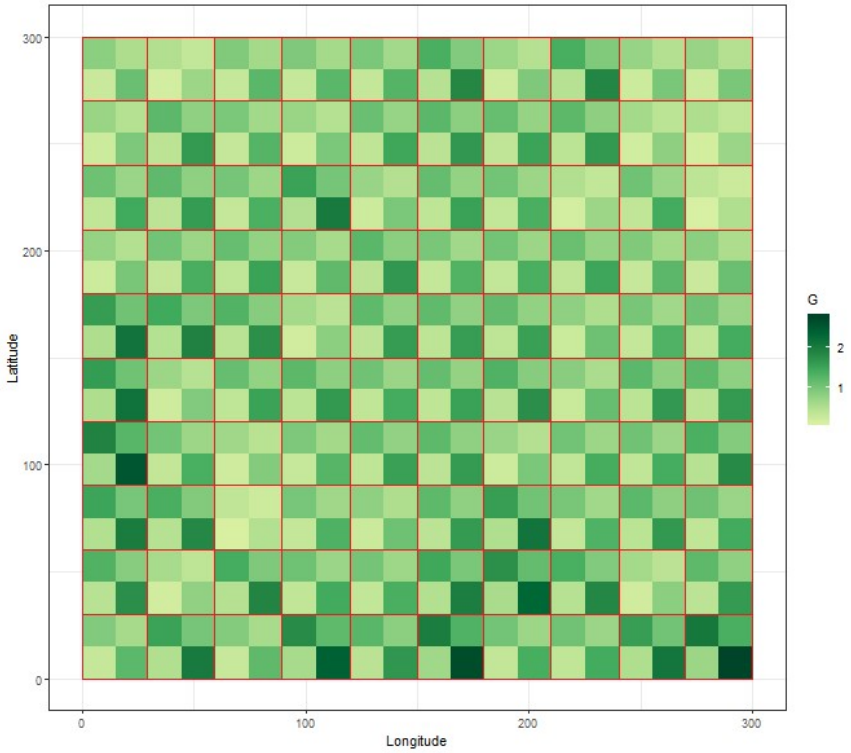


Figure 4.4. Spatial map of the basal area distribution in the Monte Cimino population partitioned into 400 pixels. The red lines represent the original 100 pixels.

Furthermore, the amount of interest variable in every sub-pixel (Figure 4.4) was distributed in such a way that the sum of G in each 4 sub-pixel of pixel j, $j = 1,\ldots,100$, was equal to the amount of interest variable in j. It should be noted that this was not an equal distribution. Indeed, 3 random numbers u, v, w were generated from the uniform distribution on $(0,1)$; a segment of length 1 was divided into 4 sub-segments at u, v, w; the total amount of G in pixel j was multiplied by the length of each sub-segment and assigned to each one of the sub-pixels of j. The values of NDVI and cos(EXP) were modified as well, in order to obtain an adjusted R-squared of 0.9 when applying the regression model (9) .

4.5 Simulation procedure

The simulation study, hereafter referred to as "simulation procedure", was performed as follows. First, the sample selection by means of OPSS was repeated $R = 10000$ times and the values of the interest variable in each pixel $j = 1,\ldots,N$ and for every replication $r = 1,\ldots,R$ were estimated accordingly to the IDW interpolator based on generalized regression criterion (Section 3.6) using two different values for the parameter α, $\alpha = 3,6$. Second, the estimated values $\hat{y}_{j,r}$ were used as the pseudo-populations from which $B = 10000$ bootstrap samples were selected. Finally, for each pixel, the following set of indexes was calculated:

- $\mathrm{E}\left(\hat{y}_j\right) = \dfrac{1}{R}\sum_{r=1}^{R}\hat{y}_{j,r}$
- $\mathrm{B}_j = \mathrm{E}\left(\hat{y}_j\right) - y_j$
- $\mathrm{MSE}\left(\hat{y}_j\right) = \dfrac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_{j,r} - y_j\right)^2$
- $\mathrm{RMSE}\left(\hat{y}_j\right) = \sqrt{\mathrm{MSE}\left(\hat{y}_j\right)}$
- $\mathrm{E}\left(\mathrm{RM\hat{S}E}_{\mathrm{boot}}^{(j)}\right) = \dfrac{1}{R}\sum_{r=1}^{R}\sqrt{\mathrm{M\hat{S}E}_{\mathrm{boot}}^{(j,r)}}$, where $\mathrm{M\hat{S}E}_{\mathrm{boot}}^{(j,r)} = \dfrac{1}{B}\sum_{b=1}^{B}\left(\hat{y}_{b,j,r}^{*} - \hat{y}_j\right)^2$
- $\mathrm{D}_j = \mathrm{E}\left(\mathrm{RM\hat{S}E}_{\mathrm{boot}}^{(j)}\right) - \mathrm{RMSE}\left(\hat{y}_j\right)$

4.6 Simulation results

The several descriptive indexes listed in the simulation procedure (Section 4.5) have been calculated for each combination of populations, specifically Monte Cimino with 79, 100 and 400 pixels, and distance function $\varphi\left(d\right) = d^{-\alpha}$, with $\alpha = 3$ or $\alpha = 6$. However, the resulting maps of bias and errors have been reconstructed using the absolute values of B_j and D_j, henceforth referred to as absolute bias (AB) and absolute difference (AD). Table 4.2 contains the minima, the maxima and the means of AB, RMSE and AD obtained from the simulation study, whereas the resulting spatial maps are reported in Figures 4.5-4.7.

The spatial maps of the distribution of AB and AD in the population of Monte Cimino partitioned into 79 pixels are reported in Figure 4.5. In particular, the maps obtained

when using d^{-3} are presented in the first row (Figures 4.5a, 4.5b) and those obtained with d^{-6} are in the second row (Figures 4.5c, 4.5d).

Table 4.2. Minima, maxima and means of the absolute value of B_j (AB) and D_j (AD) and of the root mean squared error (RMSE) achieved using two different distance functions applied to the three populations of Monte Cimino.

	AB			**RMSE**			**AD**		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**
Monte Cimino 79									
d^{-3}	0.0171	1.9154	0.5606	0.1982	2.3810	0.7946	0.0638	1.9550	0.5182
d^{-6}	0.0077	1.9973	0.5817	0.1701	2.5538	0.8808	0.0065	2.1683	0.5240
Monte Cimino 100									
d^{-3}	0.0067	0.7179	0.2388	0.1322	0.9621	0.3321	0.0026	0.6607	0.1997
d^{-6}	0.0099	0.7237	0.2411	0.1645	0.9789	0.3634	0.0232	0.6145	0.1912
Monte Cimino 400									
d^{-3}	0.0024	0.2520	0.1148	0.0401	0.2918	0.1490	0.0003	0.2211	0.0910
d^{-6}	0.0002	0.2763	0.1152	0.0487	0.3204	0.1670	0.0001	0.2517	0.0891

Moreover, pixels with a negative value of the bias B_j are enclosed into red squares (Figures 4.5a, 4.5c), while pixels with a negative value of D_j are enclosed into black squares (Figures 4.5b, 4.5d). In particular, pixel 79 (right upper corner) in Figure 4.5b is the only one with a positive bias D_j, whereas four pixels have a positive D_j in Figure 4.5d. In this case, high values of AB, AD and RMSE are obtained (as shown in Table 4.2) for both distance functions, due to the very low correlation among interest and auxiliary variables and the bootstrap RMSE estimator underestimates the real RMSE in almost every pixel. However, increasing the value of α from 3 to 6, the bootstrap

RMSE estimator performs slightly better, with similar value of minima, maxima and mean, but a greater number of pixels with a positive D_j.

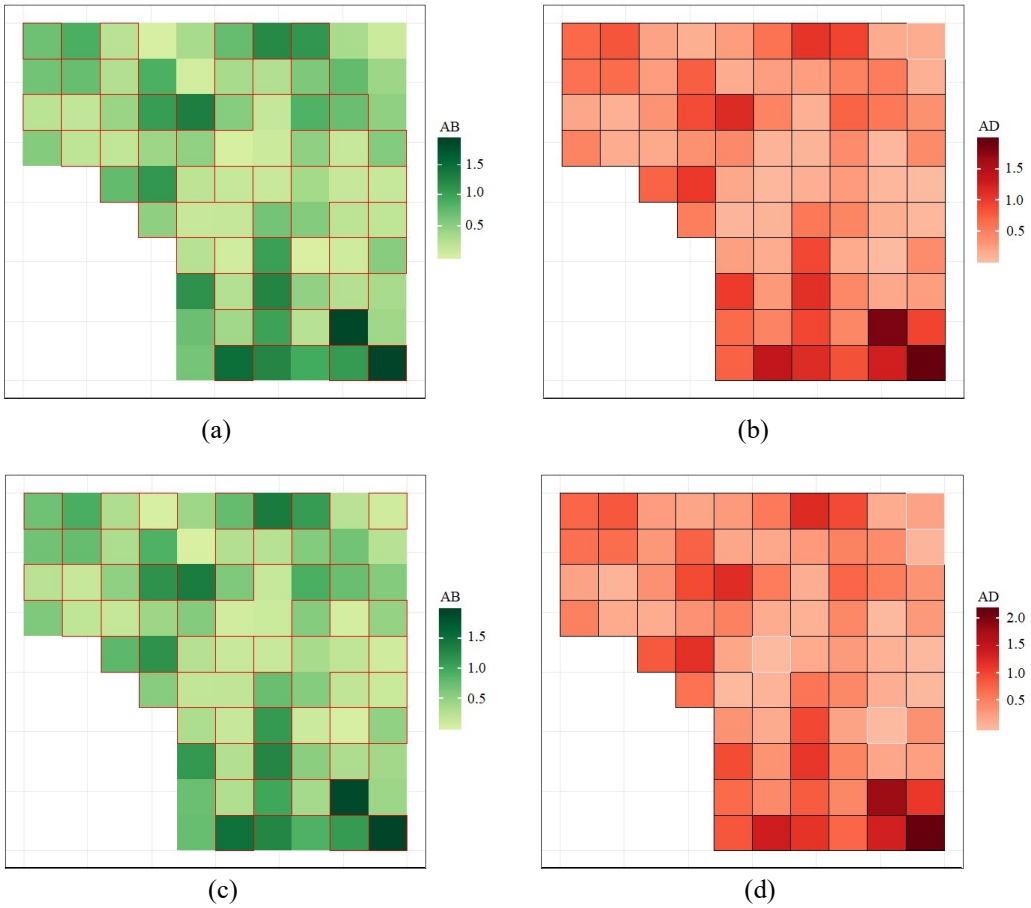


Figure 4.5. Spatial maps of AB (on the left), with negative values of the bias enclosed into red squares, and AD (on the right), with negative values of the RMSE bias enclosed into black squares, in the population of Monte Cimino partitioned into 79 pixels. First row figures were obtained using the distance function d^{-3}, whereas d^{-6} was used for (c) and (d).

In the case of the population of Monte Cimino partitioned into 100 pixels, the spatial maps of the distribution of AB and AD are reported in Figure 4.6.

Again, the maps obtained when using d^{-3} are presented in the first row (Figures 4.6a, 4.6b), those obtained with d^{-6} are in the second row (Figures 4.6c, 4.6d) and pixels with a negative value of the bias B_j are enclosed into red squares (Figures 4.6a, 4.6c), while pixels with a negative value of D_j are enclosed into black squares (Figures 4.6b,

4.6d). For both values of the parameter α, pixel 36 in Figure 4.6b and Figure 4.6d is the only one with a positive bias D_j.

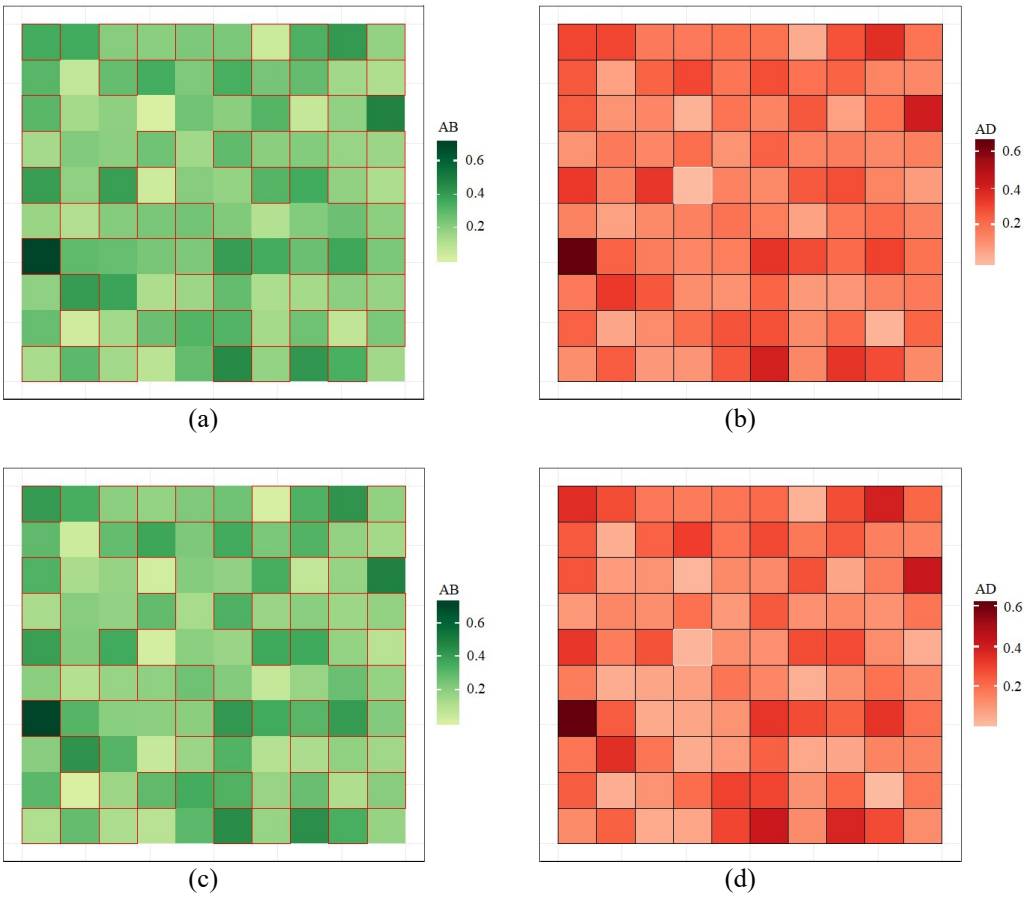


Figure 4.6. Spatial maps of AB (on the left), with negative values of the bias enclosed into red squares, and AD (on the right), with negative values of the RMSE bias enclosed into black squares, in the population of Monte Cimino partitioned into 100 pixels. First row figures were obtained using the distance function d^{-3}, whereas d^{-6} was used for (c) and (d).

Differently from the previous results, the minima, maxima and mean values of AB, RMSE and AD in Table 4.2 are less high, due to both the increase of the sample size and a better correlation among the variables. However, the bootstrap RMSE estimator underestimates the real RMSE as well and there are no relevant differences with the two distance functions.

Lastly, Figure 4.7 presents the resulting spatial maps of the distribution of AB and AD in the population of Monte Cimino partitioned into 400 pixels, for both distance functions. Once again, the maps obtained when using d^{-3} are presented in the first row

(Figures 4.7a, 4.7b), those obtained with d^{-6} are in the second row (Figures 4.7c, 4.7d); pixels with a negative value of the bias B_j are enclosed into red squares (Figures 4.7a, 4.7c) and pixels with a negative value of D_j are enclosed into black squares (Figures 4.7b, 4.7d).

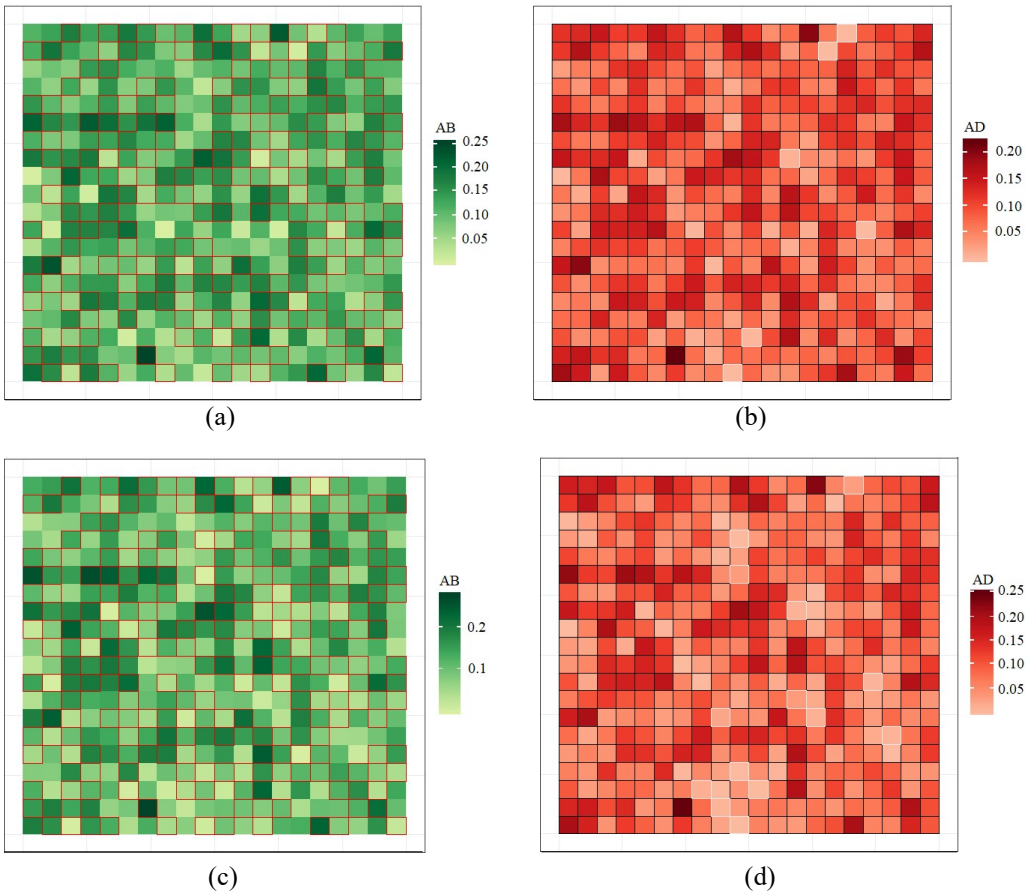


Figure 4.7. Spatial maps of AB (on the left), with negative values of the bias enclosed into red squares, and AD (on the right), with negative values of the RMSE bias enclosed into black squares, in the population of Monte Cimino partitioned into 400 pixels. First row figures were obtained using the distance function d^{-3}, whereas d^{-6} was used for (c) and (d).

Although Table 4.2 do not show relevant differences amongst minima, maxima and means of AB, RMSE and AD when the value of the parameter α increases from 3 to 6, many more pixels have a positive bias D_j. Indeed, Figure 4.7b has only 6 pixels with a positive D_j, instead of the 14 pixels in Figure 4.7d. However, the bootstrap RMSE estimator is not a conservative estimator in the majority of cases.

It should be noted that the values of AB, AD and RMSE (Table 4.2) are a lot smaller than those achieved for the other two populations. These better results are due to the increasing number of units in the sample and the decreasing size of the units as well.

Chapter 5

An application to real case

5.1 Rincine forest complex

The IDW interpolator described in Section 3.5 together with the LASSO regression (Section 3.7) was adopted to provide the map of the living wood volume in the forest complex of Rincine (Figure 5.1), in the eastern part of Tuscany in Central Italy (43.9 N, 11.6 E). The forest is under the management of a public body, the Union of Mountain Municipalities of Valdarno and Valdisieve. Forest stands, spreading over an area of approximately 290 ha, are dominated by plantations of exotic species, thermophilous deciduous stands and alpine conifer stands; prevailing tree species are Douglas fir, European black pine and Turkey oak forests.

5.2 Sampling procedure

The area was partitioned into 5449 square cells of $23 \times 23\,\mathrm{m}^2$, grouped into 49 strata of 109 squares each and one stratum of 108 squares, for a total of 50 strata. Following the OPSS sampling procedure, one square was randomly selected within each stratum, determining a sample of $n = 50$ units (Figure 5.1).

Field measurements on the selected square plots were carried out from June to November 2016. Tree species, tree DBH, tree height, four crown radii and tree status (live or dead) of all the trees with DBH>2.5 cm were collected. The total wood volume of the sampled population was partitioned into three classes, suitable for timber valuation according to local wood market: small trees (Y1, stem with DBH≤12 cm), medium trees (Y2, 12<DBH≤27 cm) and large trees (Y3, DBH>27 cm).

5.3 ALS data

For each cell, several Airborne Laser Scanning (ALS) indexes were available. In particular, ALS data were acquired in the first week of May 2015 by a LiDAR RIEGL LMS-Q680i sensor with a pulse density of 4.4 points/m^2. Common procedures for pre-processing ALS data include removal of outliers, ground/non-ground classification and computation of normalized height. First, air points that are clearly higher than the

median elevation of surrounding points and isolated points with few neighbors resulting from sensor errors or backscatter by flying objects were removed. Subsequently, ground points were classified on the basis of the adaptive triangular irregular network (TIN) model algorithm (Axelsson, 2000) and a raster digital terrain model (DTM) with a 1 m cell grid was then produced from TIN. The DTM was used to subtract the ground elevation from point elevation to obtain a normalized point cloud, thereby obtaining the normalized Crown Height Model (CHM). In accordance with recommendation when combining field sample plots and ALS data (e.g. Gobakken and Næsset, 2008), it is essential that the field plot size corresponds to the grid cell size. Hence, the relative height above ground of each echo was calculated and used to extract the canopy metrics for all the 5449 cells.

5.4 Estimation criterion

The estimation criterion previously described was applied to each of three size classes Y1, Y2 and Y3. Practically speaking, 15 ALS height and density metrics (Table 5.1) were selected from the ALS data and used as independent variables in the LASSO regression, with Y1, Y2 and Y3 used as three different dependent variables. It follows that three different sets of variables were selected from sample data, one for each size class.

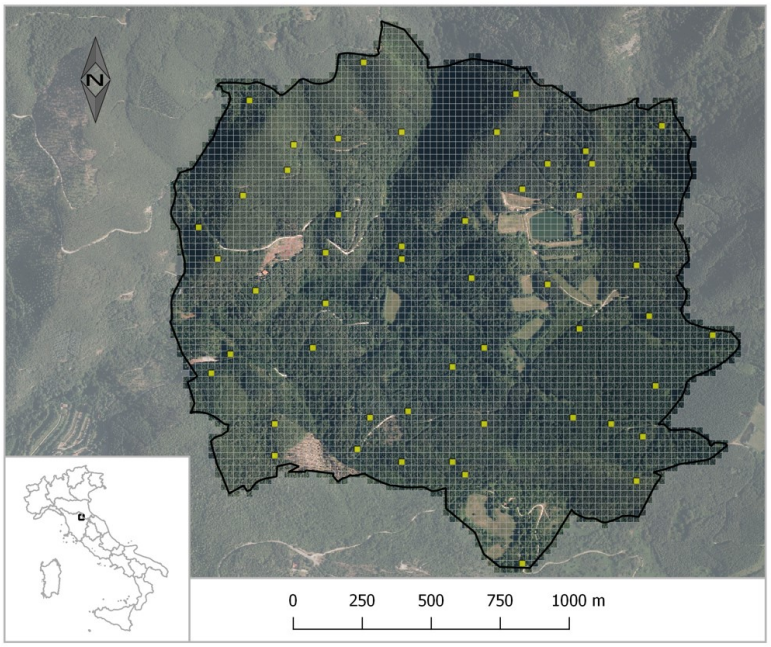


Figure 5.1. Rincine forest complex and spatial distribution of field sampling plots (yellow dots).

Table 5.1. Description of the considered ALS metrics. First four rows refer to height-based metrics, the last two describe density-based metrics.

ALS Metrics	Definition
Maximum (h_{max}) and mean (h_{mean}) heights	The maximum and mean heights above the ground of all first returns.
Coefficient of variation of heights (h_{cv})	Coefficient of variation of heights of all first returns.
Percentile heights (h_{P10}, h_{P20}, h_{P50}, h_{P75}, h_{P90})	The percentiles of the heights distributions of all first returns.
Mean heights within three layers (h_{mean1}, h_{mean2}, h_{mean3})	Mean of heights lower than 1/3 (h_{mean1}), between 1/3 and 2/3 (h_{mean2}), and higher than 2/3 (h_{mean3}) of h_{max}.
Percentage of points over the ground (*ogp*)	The number of first returns classified as no-ground over the total first returns.
Percentage of points within three layers (d_1, d_2, d_3)	Percentage of points in three layers lower than 1/3 (d_1), between 1/3 and 2/3 (d_2), and higher than 2/3 (d_3) of h_{max}.

This estimation process generated three values (one per class) of the wood volume estimate for each cell of the population, so that the total volume, i.e. the sum of the volumes of the three size classes, was obtained as a sum of these three values. Practically speaking, for each $j \in \mathrm{U}$, $\hat{y}_j = \hat{y}_{1j} + \hat{y}_{2j} + \hat{y}_{3j}$.

The resulting values (four for each cell) were used as the pseudo-population from which 20000 bootstrap samples were selected, so as to estimate the MSE, and, consequently, the RMSE, associated to each map cell and to each size class.

It is worth noting that the problem of forcing additivity on a set of classes, i.e. forcing the total estimate to be equal to the sum of the estimates of its component, which are consistent with each other (Cunia and Briggs, 1984b), is a common problem in forest literature, especially for biomass estimation (e.g., Chiyenda and Kozak, 1984). For

instance, Cunia and Briggs (1984a) analyzed three different methods, originally proposed by Cunia (1979), for forcing additivity on a set of tree biomass tables in a sample of 29 sugar maple trees. All three methods satisfy the additivity property, but either the total biomass estimate or the single class estimates are not the best possible estimates from a statistical point of view. Moreover, the estimates are not monotonically increasing in the range of diameters of interest, in such a way that the average biomass of 10 cm trees, for example, is higher than that of 11 cm trees. It is worth noting that these cumbersome procedures are necessary for providing a statistically sound estimator of the sampling variance. On the other hand, the problem has been here solved in a very easy way thanks to the bootstrap procedure that views the estimated map as pseudo-population and, consequently, allows for the straightforward estimation of sampling variances.

5.5 Results

The algorithm for the estimation of the wood volume in the area of Rincine for the three tree size classes was built using the ALS metrics selected by means of the LASSO regression. In particular, this regression determined four non-zero coefficients for small trees (Y1) and seven for both medium trees (Y2) and large trees (Y3). Statistical processing was performed in the R statistical computing environment (R Core Team, 2018). The selected ALS metrics are reported in Table 5.2.

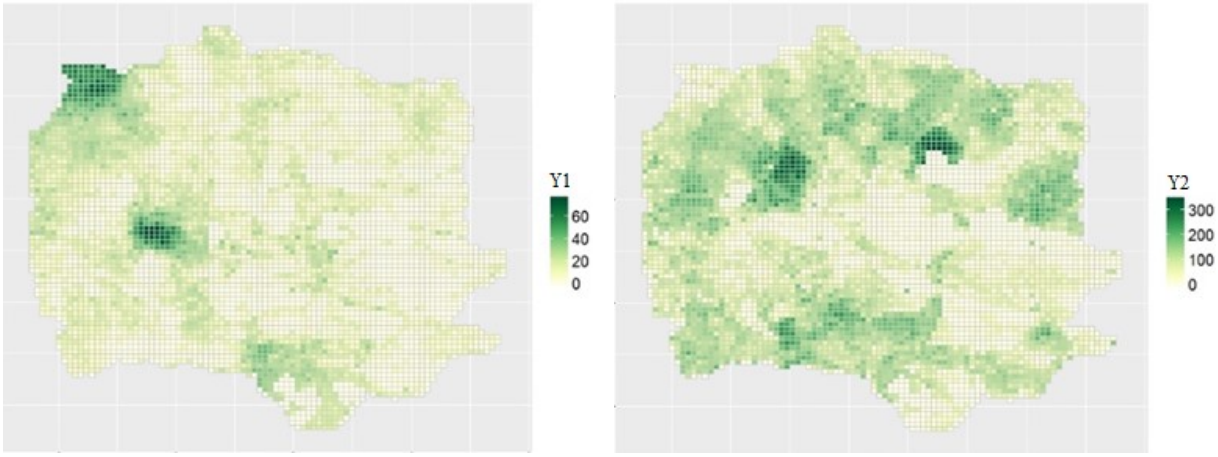
Once the regression coefficients have been estimated from the sample of $n = 50$ cells, the estimation of the wood volume for the three tree size classes in each of the 5449 spatial units was performed using the IDW interpolator. It should be noted that negative predictions were set to 0. Using this precaution, the estimated wood volume values range from 0 to 78 m^3ha^{-1} for Y1, from 0 to 358 m^3ha^{-1} for Y2 and from 0 to 1100 m^3ha^{-1} for Y3. The three resulting maps are reported in Figure 5.2.

Table 5.2. The ALS metrics included in each model are indicated by *; the number of times (expressed as a percentage of total repetitions) in which each ALS metric has been selected by the bootstrap procedure is also reported.

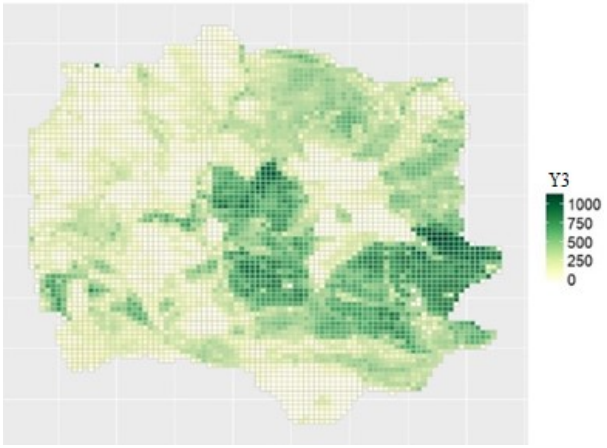
	Y1	**Y2**	**Y3**	**Whole model**
h_{max}	72%	91%	97%	4%
h_{mean}	9%	1%	*3%	20%
h_{cv}	*47%	*2%	2%	52%
h_{P10}	9%	-	*3%	28%
h_{P20}	32%	*11%	*16%	41%
h_{P50}	19%	9%	*3%	45%
h_{P75}	99%	3%	17%	97%
h_{P90}	37%	9%	12%	24%
h_{mean1}	*29%	*2%	*1%	46%
h_{mean2}	*-	82%	5%	2%
h_{mean3}	48%	*1%	*5%	45%
Ogp	56%	*4%	4%	68%
d_1	19%	*32%	16%	22%
d_2	*3%	30%	*100%	-
d_3	-	*99%	1%	-

Moreover, for each unit, the sum of the three estimated values (one per class) was used as total volume estimate and the resulting map is reported in Figure 5.3a. In order to better appreciate the improvement achieved thanks to the exploitation of the auxiliary variables, the spatial map of the total volume in the area of Rincine obtained using the sole spatial information for its estimation, as described by Fattorini et al. (2018a), is reported in Figure 5.3b.

The exploitation of auxiliary sources of information, such as ALS data, reduces the *smoothing* associated to IDW interpolator when using the sole spatial information, as it is evident from the comparison between Figure 5.3a and 5.3b. Indeed, Figure 5.3a is much more detailed as it shows the wood volume changes on a fine scale, while in Figure 5.3b the changes are more blurred and less detailed.

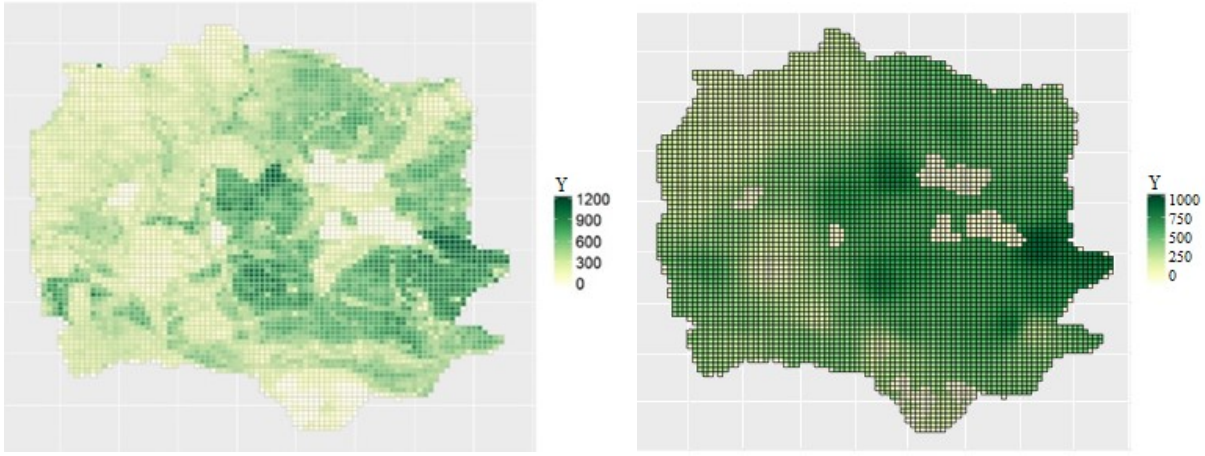


(a) (b)



(c)

Figure 5.2. Spatial maps of the wood volume in the area of Rincine for each tree size class: small trees Y1 (a), medium trees Y2 (b) and large trees Y3 (c).



(a) (b)

Figure 5.3. Spatial maps of the total wood volume in the area of Rincine exploiting ALS data (a) or using the sole spatial information (b).

As previously described, the resulting values (four for each cell) were used as the pseudo-population for the estimation of the bootstrap RMSE associated to each map cell and to each size class. The resulting RMSE maps are reported in Figure 5.4, while Table 5.2 reports the number of times that each ALS metric has been selected by the bootstrap procedure.

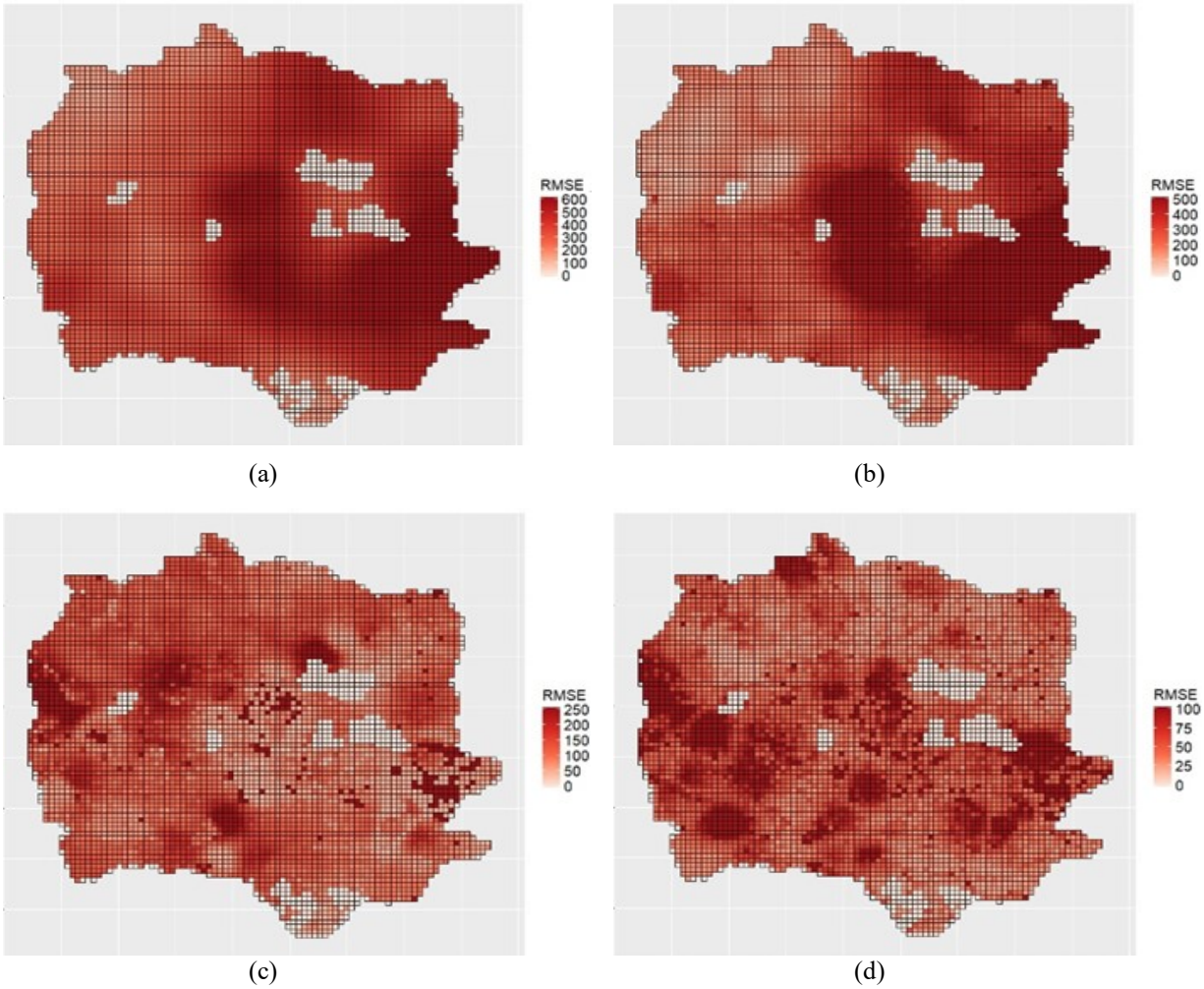


Figure 5.4. Spatial maps of the bootstrap RMSE estimates in the area of Rincine for small trees Y1 (a), medium trees Y2 (b), large trees Y3 (c) and total volume Y (d).

Although the pulse density of the available ALS data was relatively low (4.4 points/m^2), the proposed approach seems promising for estimating wood volume exploiting ALS data and relatively satisfying levels of accuracy have been reached when predicting wood volume from small, medium and large trees. Moreover, with the exception of metric d_1 (percentage of points in the lower forest layer), all the selected ALS metrics

represent the upper part of the canopy. On the one hand, these metrics allow the discrimination of different volume levels based on distributional shapes and height frequencies, thus providing a solid justification for inclusion in the predictive model. On the other hand, the acquisition of ALS data at the beginning of June, in leaf-on conditions, improves capabilities for identifying trees at the top of the canopy, but makes the assessment of lower layers (classes Y1 and Y2) worse than larger and higher trees (class Y3). The possible integration of leaf-on and leaf-off scans as well as the exploitation of ALS data characterized by high pulse density might improve the accuracy of the obtained models (Sumnall et al., 2016).

Chapter 6

Conclusions and future perspectives

Forest monitoring and assessment are relevant for forest industry, environmental stakeholders and many other forest services, including biodiversity conservation and mitigation of climate change impacts.

Thus, any decision related to the maintenance and the enhancement of forest health needs to be based on objective and reliable information. As a result, forest monitoring and assessment are rapidly evolving as new techniques and tools become available.

In this context, wall-to-wall maps are important instruments for depicting the location of forests and related attributes within a given region and, consequently, spatial map reconstruction is becoming a focal point in forest management and assessment (e.g. Brosofske et al., 2014). In principle, the spatial distribution of forest resources could be obtained measuring every tree in the region. However, this measuring process is usually too expensive and time consuming, especially in large areas, to allow for a complete census. Consequently, information is acquired by means of sampling procedures so as to obtain objective and useful maps. Moreover, the introduction of remote sensing imagery in the last decades has eased the process of making maps, providing easily accessible and affordable sources of auxiliary information. Therefore, the development of procedures exploiting auxiliary variables, such as ALS data, to predict and map forest attributes is a crucial issue in forest monitoring.

The most widely applied methodologies for forest mapping using the sole spatial information as well as those exploiting auxiliary variables are model-based methods, owing to the difficulty to estimate non-sampled values when working in a pure design-based framework. Despite their popularity, very few comparative studies on the performance of these model-based techniques in forest studies have been conducted and all these works involve severe assumptions, which means that the results are only valid in quite unrealistic situations.

It follows that both pure design-based and model-based methods are not the most suitable tools for forest mapping and working in a model-assisted framework seems the sole way to produce reliable maps of forest attributes. However, when estimating single

population values, there is no way to correct the error provided by the assisting model and the resulting map is destined to be design-biased. Accordingly, any map arising from a model-assisted method needs to be DBAU&C (Fattorini et al., 2018a).

Following this novel approach of Fattorini et al. (2018a), in which the sole spatial information is exploited, Tobler's first law of geography is used as the criterion leading to the IDW interpolator, also exploiting auxiliary sources of information for estimating the values of the survey variable at the unsampled locations.

The exploitation of auxiliary sources of information is a very common procedure because when those auxiliary variables are strongly correlated to the interest variable, they can improve the estimation obtained with the sole spatial information. However, model-based techniques which work directly on the space of auxiliary variables, such as k-NN, are not suitable when making maps. Indeed, DBAU&C of the IDW interpolator are derived supposing regularities in the sizes and shapes of units, which ultimately correspond to regularities in the arrangement of unit centers. Obviously, this feature cannot be realistically ensured for the values of auxiliary variables. Thus, notwithstanding its wide use for creating maps of forest attributes, it seems quite difficult to prove asymptotic unbiasedness and consistency of k-NN.

The model-assisted procedure proposed in this thesis seems promising for the estimation of forest attributes exploiting auxiliary data. Nevertheless, further analysis and studies on its properties and on the DBAU&C conditions are needed. Indeed, DBAU&C of the interpolator based on the generalized regression criterion should follow from the unbiasedness of $\hat{\boldsymbol{\beta}}$, but this result is still under study. Its formal proof, which constitutes a future work, is necessary to show that maps created with this method are DBAU&C.

Another important aspect to improve concerns the choice of the parameter α used in the distance function. Indeed, the values of α have been here decided *a priori*. However, it would be better to define an algorithm for its choice on the basis of sample data and including it in the bootstrap procedure. In this way, it would be possible to take into consideration the uncertainty arising from its estimation.

Finally, it should be noted that the simulation results are not as satisfying as expected, owing to the partitioning of Monte Cimino into pixels that are too big for the area, determining a small sample size. On the contrary, when the area was repartitioned into

smaller and more numerous pixels, the results became slightly better. Indeed, in order to satisfy the consistency conditions, the estimation should be performed at a sufficiently small spatial grain and the resulting values should be aggregate at a greater grain (Fattorini et al., 2018a).

With the exception of the proposals of Fattorini et al. (2018a, 2018b) on which the procedure here proposed is based, there are no other methods that move in the design-based model-assisted framework for forest mapping (e.g. Brosofske et al., 2014; Di Biase et al., 2019). Indeed, especially in recent years with the exploitation of remote sensing data, several model-based methodologies for forest map reconstruction have been proposed and their number is continuously growing, emphasizing the current relevance of forest mapping for forest monitoring and assessment.

Usually, spatial explicit estimates are used for wall-to-wall reconstruction of forest attributes, such as growing stock volume (e.g. Hu et al., 2020; Mura et al., 2018; Nilsson et al., 2017), forest cover (e.g. Matasci et al., 2018; Waser et al., 2015) and biomass (e.g. Puliti et al., 2020; Ferraz et al., 2016; Latifi and Koch, 2012), or of forest risks, such as fires (e.g. Rahimi et al., 2020; Gheshlaghi et al., 2019; Satir et al., 2015). For example, Chirici et al. (2020) compared the performance of k-NN, random forest, multiple linear regression and geographically weighted regression for the wall-to-wall prediction of growing stock volume in Central Italy, whereas Forkuor et al. (2020) combined satellite data with field data in a random forest regression to map above ground biomass in West Africa.

As for the auxiliary variables, ALS, LiDAR and satellite data are among the most used (e.g. Fernández-Landa et al., 2018; Persson et al., 2017; Waser et al., 2017, Mascaro et al., 2011; Boudreau et al., 2008; Houghton et al., 2007), with applications to boreal forests (e.g. Kangas et al., 2018) and Mediterranean areas (e.g. Condés and McRoberts, 2017; Bottalico et al., 2017; Maselli et al., 2014).

It is evident that *the matter is still controversial* (Corona et al., 2014, p. 30) and further investigations are needed. However, model-assisted methods seem to be the right path to follow for its resolution.

References

Agatonovic-Kustrin S, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **22**: 717-727.

Axelsson PE (2000) DEM generation from laser scanner data using adaptive TIN models. *Int. Arch. of Photogramm. and Remote Sens.* **33**: 110-117.

Baffetta F, Corona P, Fattorini L (2012) A matching procedure to improve k-nn estimation of forest attribute maps. *Forest Ecol. Manag.* **272**: 35-50.

Baret F (1995) Use of spectral reflectance variation to retrieve canopy biophysical characteristics. In Danson FM, Plummer SE (Eds) *Advances in environmental remote sensing*, Chichester (UK): Wiley.

Bottalico F, Chirici G, Giannini R, Mele S, Mura M, Puxeddu M, McRoberts RE, Valbuena R, Travaglini D (2017) Modeling Mediterranean forest structure using airborne laser scanning data. *Int. J. Appl. Earth Obs. Geoinformation.* **57**: 145-153.

Boudreau J, Nelson RF, Margolis HA, Beaudoin A, Guindon L, Kimes DS (2008) Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sens. Environ.* **112**: 3876-3890.

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

Breiman L (1996) Bagging predictors. *Machine Learning* **24**: 123-140.

Breiman L (2001) Random forests. *Machine Learning* **45**: 5-32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees*, Belmont (USA): Wadsworth.

Brosofske KD, Froese RE, Falkowski MJ, Banskota A (2014) A review of methods for mapping and prediction of inventory attributes for operational forest management. *Forest Science* **60**: 733-756.

Bruno F, Cocchi D, Vagheggini A (2013) Finite population properties of individual predictors based on spatial pattern. *Environ. Ecol. Stat.* **20**: 467-494.

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* **28**: 281-298.

Brus DJ, Hengeveld GM, Walvoort DJJ, Goedhart PW, Heidema AH, Nabuurs GJ, Gunia K (2012) Statistical mapping of tree species over Europe. *Eur. J. For Res.* **131**: 145-157.

Carbonell JG, Michalski RS, Mitchell TM (1983) An Overview of Machine Learning. In Michalski RS, Carbonell JG, Mitchell TM (Eds) *Machine Learning*, Berlin (Germany): Springer.

Cartwright HM (2008) Artificial neural networks in biology and chemistry: the evolution of a new analytical tool. *Methods Mol. Biol.* **458**: 1-13.

Carvalho LMT (2001) Mapping and monitoring forest remnants: A multiscale analysis of spatio-temporal data. *PhD thesis*, Wageningen University (Netherlands).

Chirici G, Giannetti G, McRoberts RE, Travaglini D, Pecchi M, Maselli F, Chiesi M, 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**: 101959.

Chiyenda SS, Kozak A (1984) Additivity of component biomass regression equations when the underlying model is linear. *Can. J. For Res.* **14**: 441-446.

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**: 596-610.

Condés S, McRoberts RE (2017) Updating national forest inventory estimates of growing stock volume using hybrid inference. *For. Ecol. Manag.* **400**: 48-57.

Corona P (2000) *Introduzione al rilevamento campionario delle risorse forestali*, Firenze (Italy): Edizioni CUSL.

Corona P (2010) Integration of forest mapping and inventory to support forest management. *iForest* **3**: 59-64.

Corona P (2016) Consolidating new paradigms in large-scale monitoring and assessment of forest ecosystems. *Environ. Res.* **144**: 8-14.

Corona P, Chirici G, Marchetti M (2002) Forest ecosystem inventory and monitoring as a framework for terrestrial natural renewable resource survey programmes. *Plant Biosyst.* **136**: 69-82.

Corona P, Chirici G, McRoberts RE, Winter S, Barbati A (2011) Contribution of large-scale forest inventories to biodiversity assessment and monitoring. *For. Ecol. Manag.* **262**: 2061-2069.

Corona P, Fattorini L, Franceschi S, Chirici G, Maselli F, Secondi L (2014) Mapping by spatial predictors exploiting remotely sensed and ground data: A comparative design-based perspective. *Remote sens. Environ.* **152**: 29-37.

Cressie N (1990) The origins of kriging. *Math. Geol.* **22**: 239-252.

Cressie N (1993) *Statistics for spatial data.* New York (USA): Wiley.

Cunia T (1979) On tree biomass tables and regression: some statistical comments. In Frayer WE (Ed) *Forest resource inventory workshop proceedings Vol. 2*, Fort Collins (USA), July 23-26, pp. 629-642.

Cunia T, Briggs RD (1984a) Forcing additivity of biomass tables: some empirical results. *Can. J. For Res.* **14**: 376-384.

Cunia T, Briggs RD (1984b) Forcing additivity of biomass tables: use of generalized least squares method. *Can. J. For Res.* **15**: 23-28.

de Gruijter JJ, ter Braak CJF (1990) Model free estimation from survey samples: A reappraisal of classical sampling theory. *Math. Geol.* **22**: 407-415.

Deutsch C, Journel A (1992) *Geostatistical Software Library and User's Guide*, New York (USA): Oxford University Press.

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.

Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**: 1-26.

Fattorini L (2015) Design-based methodological advances to support National Forest Inventories: a review of recent proposals. *iForest* **8**: 6-11.

Fattorini L, Corona P, Chirici G, Pagliarella MC (2015) Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use and cover estimation. *Environmetrics* **26**: 216-228.

Fattorini L, Marcheselli M, Pratelli L (2018a) Design-based maps for finite populations of spatial units. *J. Am. Stat. Assoc.* **113**: 686-697.

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

Fernández-Landa A, Fernández-Moya J, Tomé JL, Algeet-Abarquero N, Guillén-Climent ML, Vallejo R, Sandoval V, Marchamalo M (2018) High resolution forest inventory of pure and mixed stands at regional level combining National Forest

Inventory field plots, Landsat, and low density lidar. *Int. J. Remote Sens.* **39**: 4830-4844.

Ferraz A, Saatchi S, Mallet C, Jacquemoud S, Goncalves G, Silva Ca, Soares P, Tomé M, Pereira L (2016) Airborne Lidar estimation of aboveground forest biomass in the absence of field inventory. *Remote Sens.* **8**: 653.

Forkuor G, Zoungrana JB, Dimobe K, Ouattara B, Vadrevu K, Tondoh J (2020) Above-ground biomass mapping in West African dryland forest using Sentinel-1 and 2 datasets – A case study. *Remote Sens. Environ.* **236**: 111496.

Freeman EA, Moisen GG (2007) Evaluating Kriging as a Tool to Improve Moderate Resolution Maps of Forest Biomass. *Environ. Monit. Assess.* **128**: 395-410.

Gauch HG (1982) *Multivariate analysis in community ecology*, New York (USA): Cambridge University Press.

Gheshlaghi HA, Feizizadeh B, Blaschke T (2020) GIS-based forest fire risk mapping using the analytical network process and fuzzy logic. *J. Environ. Plan. Manag.* **63**: 481-499.

Gobakken T, Næsset E (2008) Assessing effect of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. *Can. J. For Res.* **38**: 1095–1109.

Görgens EB, Montaghi A, Rodriguez LCE (2015) A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. *Comput. Electron. Agr.* **116**: 221-227.

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

Gregoire TG, Valentine HT (2008) *Sampling strategies for natural resources and the environment*, New York (USA): Chapman & Hall.

Helmer EH, Ruzycki TS, Wunderle JM, Vogesser S, Ruefenacht B, Kwit C, Brandeis TJ, Ewert DN (2010) Mapping tropical dry forest height, foliage height profiles and disturbance type and age with a time series of cloud-cleared Landsat and ALI image mosaics to characterize avian habitat. *Remote Sens. Environ.* **114**: 2457–2473.

Holmgren P, Persson R (2002) Evolution and prospects of global forest assessments. *Unasylva* **53**: 3-9.

Hotelling H (1936) Relations between two sets of variates. *Biometrika* **28**: 321–377.

Houghton RA, Butman D, Bunn AG, Krankina ON, Schlesinger P, Stone TA (2007) Mapping Russian forest biomass with data from satellites and forest inventories. *Environ. Res. Lett.* **2**: 045032.

Hudak AT, Crookston NL, Evans JS, Hall DE, Falkowski MJ (2008) Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sens. Environ.* **112**: 2232-2245.

Hu Y, Xu X, Wu F, Sun Z, Xia H, Meng Q, Huang W, Zhou H, Gao J, Li W, Peng D, Xiao X (2020) Estimating forest stock volume in Hunan province, China, by integrating in situ plot data, Sentinel-2 images, and linear and machine learning regression models. *Remote Sens.* **12**: 186.

Huete AR, Liu H (1994) An error and sensitivity analysis of the atmospheric – and soil – correcting variants of the NDVI for de MODIS-EOS. *IEEE Trans. Geosci. Remote Sens.* **32**: 897-905.

Irons JR, Dwyer JL, Barsi JA (2012) The next Landsat satellite: the Landsat data continuity mission. *Remote Sens. Environ.* **122**: 11-21.

Isaki CT, Fuller WA (1982) Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.* **77**: 89-96.

Journel AG, Huijbregts CJ (1978) *Mining Geostatistics*, London (UK): Academic Press.

Justice CO, Townshend JRG, Holben BN, Tucker CJ (1985) Analysis of the phenology of global vegetation using meteorological satellite data. *Int. J. Remote Sens.* **6**: 1271-1318.

Kangas A, Astrup R, Breidenbach J, Fridman J, Gobakken T, Korhonen KT, Maltamo M, Nilsson M, Nord-Larsen T, Næsset E, Olsson H (2018) Remote sensing and forest inventories in Nordic countries – roadmap for the future. *Scand. J. For. Res.* **33**: 397-412.

Kleinn C (2002) New technologies and methodologies for national forest inventory. *Unasylva* **53**: 10-15.

Köhl M, Magnussen SS, Marchetti M (2006) *Sampling methods, remote sensing and GIS multiresource forest inventory*, Heidelberg (Germany): Springer.

Lamonaca A (2005) Sperimentazione di tecniche per l'analisi multilivello della diversità strutturale di soprassuoli forestali. *PhD thesis*, Univesity of Tuscia, Viterbo (Italy).

Latifi H, Koch B (2012) Evaluation of most similar neighbor and random forest methods for imputing forest inventory variables using data from target and auxiliary stands. *Int. J. Remote Sens.* **33**: 6668-6694.

Lo Monaco A (1984) Un po' di storia sulla faggeta del Monte Cimino (Viterbo). *Monti e boschi* **3**: 38-40.

Lochhead K, LeMay V, Bull G, Schwab O, Halperin J (2018) Multivariate estimation for accurate and logically consistent forest-attributes maps at macroscales. *Can. J. For Res.* **48**: 345-359.

Lund G (1998) A comparison of multipurpose resource inventories (MRIs) throughout the world. *EFI working paper* **14**, Joensuu (Finland).

Maltamo M, Næsset E, Vauhkonen J (2014) *Forestry applications of airborne laser scanning*, Dordrecht (Netherlands): Springer.

Markham B (2011) Landsat data continuity mission: overview and status. *10th annual JACIE workshop*, Boulder (USA), March 29-31.

Mas JF, Puig H, Palacio JL, Sosa-López A (2004) Modelling deforestation using GIS and artificial neural networks. *Environ. Modell. Softw.* **19**: 461-471.

Mascaro J, Detto M, Asner G, Muller-Landau H (2011) Evaluating uncertainty in mapping forest carbon with airborne LiDar. *Remote Sens. Environ.* **115**: 3770-3774.

Maselli F, Chirici G, Bottai L, Corona P, Marchetti M (2005) Estimation of Mediterranean forest attributes by the application of k-NN procedures to multitemporal Landsat ETM+ images. *Int. J. Remote Sens.* **26**: 3781-3796.

Maselli F, Chiesi M, Mura M, Marchetti M, Corona P, Chirici G (2014) Combination of optical and LiDAR satellite imagery with forest inventory data to improve wall-to-wall assessment of growing stock in Italy. *Int. J. Appl. Earth Obs. Geoinformation* **26**: 377-386.

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

Matasci G, Hermosilla T, Wulder MA, White JC, Coops NC, Hobart GW, Zald HSJ (2018) Large-area mapping of Canadian boreal forest cover, height, biomass and other structural attributes using Landsat composites and lidar plots. *Remote Sens. Environ.* **209**: 90-106.

Matheron G (1963) Principles of geostatistics. *Econ. Geol.* **58**: 1246-1266.

Matheron G (1969) *Le Krigeage Universel*, Fontainebleau (France): Cahiers du Centre de Morphologie Mathématique.

Mattioli W, Quatrini V, Di Paolo S, Di Santo D, Giuliarelli D, Angelini A, Portoghesi P, Corona P (2012) Experimenting the design-based k-NN approach for mapping and estimation under forest management planning. *iForest* **5**: 26-30.

McRoberts RE (2012) Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecol. Manag.* **272**: 3-12.

McRoberts RE, Tomppo EO (2007) Remote sensing support for national forest inventories. *Remote Sens. Environ.***110**: 412-419.

McRoberts RE, Holden GR, Nelson MD, Liknes GC, Gormanson DD (2006) Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Can. J. For Res*. **36**: 2968-2980.

McRoberts RE, Tomppo EO, Finley AO, Heikkinen J (2007) Estimating areal means and variances of forest attributes using the k-nearest neighbours technique and satellite imagery. *Remote Sens. Environ*. **111**: 466-480.

McRoberts RE, Tomppo EO, Næsset E (2010a) Advances and emerging issues in national forest inventories. *Scand. J. For Res.* **25**: 368-381.

McRoberts RE, Cohen WB, Næsset E, Stehman SV, Tomppo EO (2010b) Using remotely sensed data to construct and assess forest attribute maps and related spatial products. *Scand. J. For Res.* **25**: 340-367.

McRoberts RE, Chen Q, Gormanson DD, Walters BF (2018) The shelf-life of airborne laser scanning data for enhancing forest inventory inferences. *Remote Sens. Environ.* **206**: 254-259.

Meng Q, Cieszewski C, Madden M (2009) Large area forest inventory using Landsat ETM+: A geostatistical approach. *ISPRS J. Photogramm. Remote Sens.* **64**: 27-36.

Moeur M, Stage AR (1995) Most similar neighbour: an improved sampling inference procedure for natural resource planning. *Forest Science* **41**: 337-359.

Mura M, Bottalico F, Giannetti F, Bertani R, Giannini R, Mancini M, Orlandini S, Travaglini D, Chirici G (2018) Exploiting the capabilities of the Sentinel-2 multi spectral instrument for predicting growing stock volume in forest ecosystems. *Int. J. Appl. Earth Obs. Geoinformation* **66**: 126-134.

Nilsson M, Nordkvist K, Jonzen J, Lindgren N, Axensten P, Wallerman J, Egberth M, Larsson S, Nilsson L, Eriksson J, Olsson H (2017) A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. *Remote Sens. Environ.* **194**: 447-454.

Noordermeer L, Bollandsås OM, Ørka HO, Næsset E, Gobakken T (2019) Comparing the accuracies of forest attributes predicted from airborne laser scanning and digital aerial photogrammetry in operational forest inventories. *Remote Sens. Environ.* **226**: 26-37.

Odeh IOA, McBratney AB, Chittleborough DJ (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma* **67**: 215-226.

Ohmann JL, Gregory MJ (2002) Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, USA. *Can. J. For Res.* **32**: 725-741.

Opsomer JD, Breidt FJ, Moisen GG, Kauermann G (2007) Model-assisted estimation of forest resources with generalized additive models. *J. Am. Stat. Assoc.* **102**: 400-416.

Persson HJ, Olsson H, Soja MJ, Ulander LMH, Fransson JES (2017) Experiences from large-scale forest mapping of Sweden using TanDEM-X data. *Remote Sens.* **9**: 1253.

Puliti S, Hauglin M, Breidenbach J, Montesano P, Neigh C, Rahlf J, Solberg S, Klingenberg TF, Astrup R (2020) Modelling above-ground biomass stock over Norway using national forest inventory data with ArtictDEM and Sentinel-2 data. *Remote Sens. Environ.* **6**: 111501.

R Core Team (2018) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna (Austria).

Rahimi I, Azeez SN, Ahmed IH (2020) Mapping forest-fire potentiality using remote sensing and GIS, case study: Kurdistan region-Iraq. In Al-Quraishi A, Negm A (Eds) *Environmental Remote Sensing and GIS in Iraq*, Cham (Switzerland): Springer, pp. 499-516.

Reese H, Nilsson M, Granqvist Pahlén T, Hagner O, Joyce S, Tingelöf U, Egberth M, Olsson H (2003) Countrywide Estimates of Forest Variables Using Satellite Data and Field Data from the National Forest Inventory. *A Journal of the Human Environment* **32**:542-548.

Särndal CE, Swensson B, Wretman J (1992) *Model assisted survey sampling*, New York (USA): Springer-Verlag.

Satir O, Berberoglu S, Donmez C (2015) Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomat. Nat. Haz. Risk.* **7**: 1645-1658.

Schabenberger O, Pierce F (2002) *Contemporary statistical models for the plant and soil sciences*, Boca Raton (USA): CRC Press.

Schapire R, Freund Y, Bartlett P, Lee W (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Stat.* **26**: 1651-1686.

Schowengerdt RA (2006) *Remote Sensing: Models and Methods for Image Processing*, 3rd ed, Burlington (USA): Academic Press.

Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In Proceedings of the *23rd ACM National Conference*, New York (USA), August 27-29, pp. 517-524.

Spillan D, Edmonds C (1849) *History of Rome by Titus Livius, books IX to XXVI*, London (UK).

Stevens DJ, Olsen AR (2004) Spatially balanced sampling of natural resources. *J. Am. Stat. Assoc.* **99**: 262–278.

Sumnall MJ, Hill RA, Hinsley SA (2016) Comparison of small-footprint discrete return and full waveform airborne lidar data for estimating multiple forest variables. *Remote Sens. Environ.* **173**: 214-223.

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167-1179.

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

Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**: 267-288.

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**: 234-240.

Torresan C, Corona P, Scrinzi G, Valls Marsal J (2016) Using classification trees to predict forest structure types from LiDAR data. *Ann. For Res.* **59**: 281-298.

Urban DL (2002) Classification and regression trees. In McCune B, Grace JB (Eds) *Analysis of ecological communites*, Gleneden Beach (USA): MjMSofware Design, pp. 222-232.

Wang Q, Ni J, Tenhunen J (2005) Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. *Glob. Ecol. Biogeogr.* **14**: 379-393.

Waser LT, Fischer C, Wang Z, Ginzler C (2015) Wall-to-wall forest mapping based on digital surface models from image-based point clouds and NFI forest definition. *Forests* **6**: 4510-4528.

Waser LT, Ginzler C, Rehush N (2017) Wall-to-wall tree type mapping from countrywide airborne remote sensing surveys. *Remote Sens.* **9**: 766.

Ziaco E, Alessandrini A, Blasi S, Di Filippo A, Dennis S, Piovesan G (2012) Communicating old-growth forest through an educational trail. *Biodivers. Conserv.* **21**: 131-144.

List of figures

List of tables

List of papers

Chiarucci A, **Di Biase RM**, Fattorini L, Marcheselli M, Pisani C (2018) Joining the incompatible: exploiting purposive lists for the sample-based estimation of species richness. *Ann. Appl. Statist.* **12**: 1679-1699.

Corona P, **Di Biase RM**, Fattorini L, D'Amati M (2018) A Monte Carlo appraisal of the estimation of tree abundance and stand basal area in forest inventories based on terrestrial laser scanning. *Can. J. For Res.* **49**: 41-52.

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.

Fattorini L, **Di Biase RM**, Giuliarelli D, Marcheselli M, Pisani C, Corona P (2018) Mapping the diversity of forest attributes: a design-based approach. *Can. J. For Res.* **49**: 190-197.