UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO

DIPARTIMENTO PER L'INNOVAZIONE NEI SISTEMI BIOLOGICI, AGROALIMENTARI E FORESTALI

Corso di Dottorato di Ricerca in Ecologia Forestale

XXVI ciclo

Statistical analyses and techniques for forest volume estimation and stand structural classification using airborne laser scanner data

SSD AGR/05

Chiara Torresan

Coordinator
Prof. Paolo De Angelis

Advisor
Dr. Gianfranco Scrinzi

Co-advisor
Prof. Piermaria Corona

June 2014

“The scenes in our life resemble pictures in a rough mosaic;
they are ineffective from close up,
and have to be viewed from a distance if they are to seem beautiful...."
Arthur Schopenhauer

Abstract

Forests represent an important environmental and economic resource and an accurate knowledge of the amount of timber volume and their structural features allows efficient forest management strategies on which decision making processes depend.

Airborne laser scanning (ALS) is widely used for the retrieval of detailed information on height and cover of forests at different levels, whether for individual trees, plots, stands or on a nationwide forest scale. Coupling aerial laser scanner data with inventory data provides information on timber volume and forest structural characterization, and improves the precision of forest inventory estimates.

This thesis attempted to provide specific solutions to some problems and objectives encountered by the Forest Service of Trento Province (Northern East of Italy) pertaining to both forest volume estimation at the forest compartment level and area of classification of structure of forest stands.

Volume estimation at stratum level can be obtained from field plots located within, on the contrary volume estimation in portions of strata may be problematic due to the small number or even the absence of plots in those portions. But if a Canopy Height Model (CHM) is available for the whole area under planning, height at pixel level can be used as auxiliary information. A ratio model presuming a proportional relationship between heights and volumes at pixel level was adopted to lead to an estimation. From this model, the volume within any portion was estimated as the proportionality factor estimate multiplied by the total of heights in that portion. This volume estimator was considered from the model-based, design-based and hybrid perspectives, and variances and their estimator were derived under the three approaches. Volume estimator and the model-based, design-based, and hybrid variance estimators were checked by a simulation study performed on a real forest in Val di Sella (South-Eastern of Trento Province), considered as a fixed population in which plots were randomly allocated at each simulation run. Furthermore, volume estimator and the hybrid variance estimator were applied to infer the volume of four forest compartments in the public forest estate of San Martino di Castrozza (North-Eastern of Trento province). The simulation and the case study showed that the bias of the volume estimator derived from the ratio model adopted depends on the similarity of the whole stratum and the portion of stratum in which estimation is performed. In presence of negligible bias, the results from simulation and case study are satisfactory in terms of relative root mean square error and hybrid variance estimates, respectively.

The Forest Service of Trento Province presently uses a structural classification system based on the percentage of basal area in small (or understory), medium (or mid-story), and large (or overstory) trees. The classification in forest structural types is one of the primary goals along with species composition, for forest managers to identify the strata in the process of forest inventory based on stratified sampling. Moreover, the forest structural classification is used during the decision making process.

This dissertation examined the use of CHM-derived metrics to predict one of the twelve forest structure classes of the classification system utilized by the Forest Service. Two approaches to predict size-based forest classifications were evaluated: in the first, supervised classification with both linear discriminant analysis (LDA) and random forest (RF) was attempted; in the second, basal areas of small, medium, and large trees from CHM-derived variables by *k*-nearest neighbor imputation (*k*-NN) and parametric regression were predicted, and then classified observations based on their predicted basal areas. Leave-one-out cross-validation was used to evaluate the ability to predict forest structure classes from CHM data and in the case of prediction-based classification approach the performances in predicting basal area were evaluated. The strategies proved moderately successful, being the LDA the techniques with best performance. In general, the prediction of basal areas of understory and overstory trees was the most successful.

An additional study was carried out to assess whether metrics extracted this time from lidar raw point cloud can be exploited to discriminate different forest types through machine learning techniques such as classification trees. This study suggests that the proportion of basal area in diameter size-classes is a valuable parameter to distinguish between different forest structural clusters through clustering analysis. The five airborne laser scanner metrics selected for this study were able to confirm that at least one of the five forest structure patterns identified by the clustering analysis (pole-stage, young, adult, mature, and old forests) is different from the other patterns. The classification tree model built into this study provided moderately satisfactory results in term of classification performance. The binary tree representing the model is clearly interpretable and easily usable by forest technicians in operational contexts.

This dissertation contributed to the body of knowledge that is related to the application of statistical analyses and techniques for forest resources inventory using airborne laser scanner data. Nevertheless, more work is suggested. Future research emphasis could be on how reflectivity from a laser of different wavelengths could be used to discriminate between different types of forest cover. Moreover fusion between lidar data and other image sources could be investigated.

Keywords: lidar, forest inventory, forest management, standing wood volume estimation, basal area, forest structure classification, statistical methods

Acknowledgments

I would like to thank Dr. Gianfranco Scrinzi for his advice coming from his strong expertise in forest management planning. Thanks to him I learned to doubt everything and to "dissect" the data to ask appropriate questions.

I would like to thank Prof. Piermaria Corona for his accurate and valuable suggestions and for giving me the chance to participate in workshops dedicated to the application of airborne laser scanning in forestry.

I would like to thank Antonio Floris, Fabrizio Clementel, Giacomo Colle, and Sandro Gecele for their willingness to exchange points of view, and for always offering support when needed.

Thanks to Warren Cohen, Yang Zhiqiang, Harold Zald, and Jacob Strunk for their welcome to the Laboratory for Applications of Remote Sensing in Ecology in Corvallis and their kindness during my two-month stay. I would especially like to thank Jacob and Harold because it is through their guidance in writing a paper that I learned the importance of answering basic research questions. Moreover they pushed me to challenge the validity of my own research, until the methods, statistical analyses and results were defensible.

I would like to thank all statisticians with whom I worked for my PhD research activities. Thanks to Sara Franceschi and to Jacob Strunk one more time, for help in the use of R software. Thanks to Joan Valls for time, knowledge, competence and support in learning statistical methodological procedures to conduct research using R software, for providing statistical advice, analytical critiques, and important revisions. Thanks to Prof. Lorenzo Fattorini for instilling the tenacity and perseverance of defending my own thesis.

Thanks to Maria Flora Mangano for helping me learn the rules and secrets about the communication of scientific research.

My most heartfelt to Dr. Amy Schoener for help in revising the English language of this dissertation.

I would also like to thank all people I met at the Forest Monitoring and Management Research Unit in Trento with whom I shared space and time, especially Fabio Pastorella for providing some of the data used in a study carried out within the PhD.

Last but not least, I must express my thanks to my partner Mirco Maistro for his continuous support during these three years. It has been a busy period handling both research activities and growing family, Tina and Emiliano. The future is uncertain but with my partner and his music by my side, I feel on safe ground.

Table of contents

1 Introduction

In this chapter an introduction to the issues of the dissertation is provided. An overview of airborne laser scanning and its products is given. After that, applications of airborne laser scanning to assess forest standing volume and to characterize forest stand structure are delineated. Research problems and aims of this dissertation are illustrated, and finally a short description of dissertation structure is described.

1.1 An overview on airborne laser scanning, its products, and accuracy of retrievable information

Airborne near-infrared lidar is an active remote-sensing technique that provides three-dimensional high-precision measurements of targets in the form of a point cloud (x, y, z, intensity of the backscattered power), based on laser-ranging measurements supported by the position and orientation information derived with use of a Differential Global Positioning System (DGPS) device and an Inertial Measurement Unit (IMU) mounted on an airplane or an helicopter [Yu *et al.*, 2010]. As an active sensing system, the airborne laser scanning (ALS) uses a laser beam as the sensing carrier which is considered as two optical beams, the emitted laser pulse and the received portion of that pulse. One advantage of ALS is that the laser scanners are not dependent on the sun as a source of illumination, and the interpretation of laser scanner data is not hampered by shadows caused by clouds or neighboring objects. This means that laser scanner pulses may travel unimpeded back and forth along the same path through small openings in a forest canopy, providing information about the tree crowns and the forest floor [Wagner *et al.*, 2004].

All airborne laser systems measure the distance between the sensor and the illuminated spot on the ground (Figure 1.1), applying specific ranging principles.

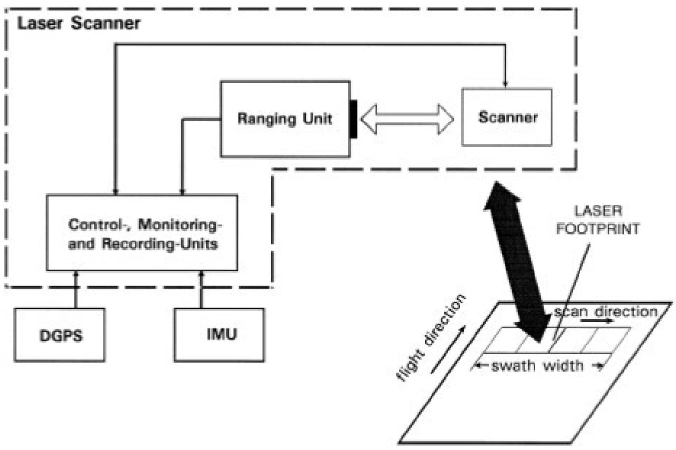

Figure 1.1 - A typical ALS system after Wehr and Lohr [1999]

In range measurements with laser, two major ranging principles are applied: the pulsed ranging principle, and ranging by measuring the phase difference between the transmitted and the received signal backscattered from the object surface. The former principle is applied when pulsed lasers (also known as discrete return systems) are used: in this case the travelling time (or time of flight) between the emitted and received pulse is recorded, and the range is calculated as the following:

$$R = \frac{1}{2} c t_L \qquad [1.1]$$

where c is the speed of light, t_L is travelling time of a light pulse.

With lasers that continuously emit light and then digitize the full waveform of the backscattered pulse (called continuous wave lasers or full waveform lasers), the phase difference principle is applied, and the range is calculated as the following:

$$R = \frac{1}{4\pi} \frac{c}{f} \phi \qquad [1.2]$$

where c is the speed of light, f is the frequency of signal emission, and Φ is the phase difference between the received and the transmitted signal.

In the very early lidar discrete return system, the laser scanner recorded at least the round-trip time of the emitted and received pulse. By 2000, commercial systems became capable of recording the time of multiple discrete returns (or echoes) for each transmitted pulse and the reflected intensity for each return. Multiple return lidar systems are very useful in forestry applications: when the laser beam from a multi-return system interacts with a tree canopy, then the first return is usually assumed to arrive from the top of the tree (or where the transmitted laser beam first interacts with the tree canopy). The last return may interact with the ground underneath the tree, although the ability to map the ground is largely dependent on the density of the vegetated canopy. Intermediate returns, perhaps 2^{nd}, 3^{rd}, and 4^{th}, are expected to be caused by tree branches and understory vegetation between the top of the canopy and the ground. The multiple return lidar systems produce an emitted pulse with a dimension (footprint) generally between 10 and 30 cm in diameter from the ground [Næsset, 2002]. For this reason they are called small footprint systems. Nowadays, there is a trend to design laser scanners with laser beams that have smaller and smaller beam divergence: with these sensors the number of multiple returns per emitted pulse will decrease, due to the fact that a smaller surface patch is illuminated [Wagner *et al.*, 2004]. This aspect entails that the information per beam decreases, and classification of returns is only possible in relation to the neighboring echoes. On the ground, the density of pulses may vary from many to few points per square meter.

In the case of full waveform lidar systems, the laser pulses that are returning to the sensor after being back-scattered by targets are sampled at a very high temporal resolution (Figure 1.2): for example, to achieve 10 cm vertical resolution the returning pulse echo has to be sampled at 1.5 GHz.

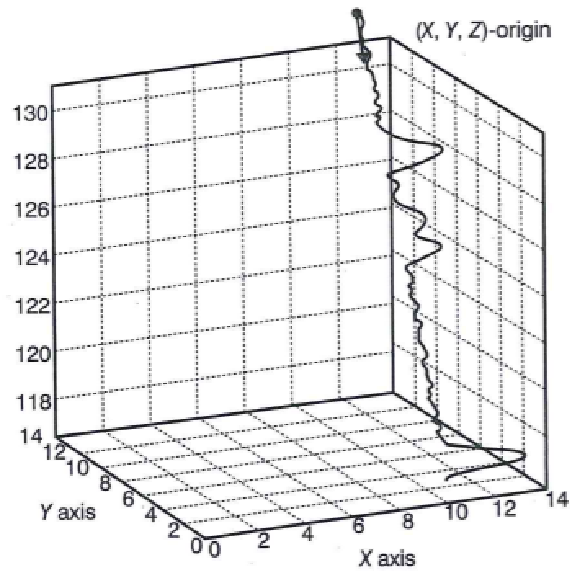

Figure 1.2 - An example of a 128 sample 8-bit laser scanner waveform, after Maas [2010]

In full waveform lidar systems the digitizer is able to digitize the complete echo waveform and delivers a full intensity profile of the echo of a laser pulse (Figure 1.3).

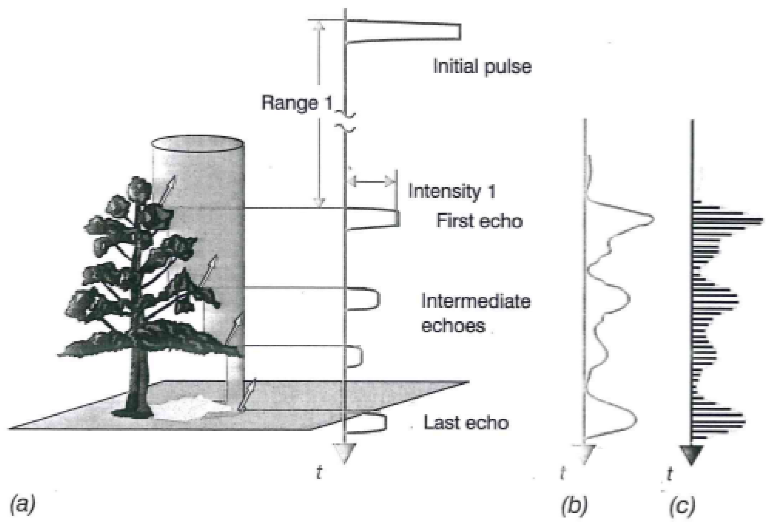

Figure 1.3 - The multiple echoes and full waveform: (a) discrete pulses, (b) waveform, and (c) digitized waveform, after Beraldin et al. [2010]

Total amount of data to be recorded with this system drastically increases: assuming 40 m tree height and 10 cm waveform resolution, 400 samples have to be stored for each laser pulse. This requires a large memory capacity. The first commercial full-waveform lidar system was introduced in 2004 [Hug *et al.*, 2004]: to record full waveform lidar data, the main commercial manufacturers have added digitization terminals to their systems and increased the storage media capacity [Pirotti, 2011]. In the first applications, the footprint ranged between 10 and 20 m (so they were called large footprint systems), today most commercial systems are small-footprint, typically 0.3 - 1 m diameter at 1000 m altitude of flight [Bretar *et al.*, 2008].

Rapid technological advances have made airborne laser scanning one of the most used technique for the retrieval of detailed information on forests at different levels, ranging in scale from the individual tree, to the plot, to the stand and, finally, to the nationwide forests (when "wall-to-wall" lidar data are available). ALS data has proven superior to the other main optical data used in forest measurements [Hyyppä and Hyyppä, 1999].

By basic processing, lidar returns from the ground are separated from those from above ground targets (classification process). The airborne lidar back-scattered signals can be used in their native form (raw data) or they can be spatially interpolated to produce digital models. Returns from the ground are used to produce Digital Terrain Models (DTM), while above ground returns are used to produce Digital Surface Models (DSM). In areas covered by trees the algebraic subtraction of DTM from DSM corresponds to the Canopy Height Model (CHM), which provides a measure of the height of the upper canopy. The combination of an accurate DTM derived from lidar points and the vertical distributions of lidar points above the estimated surface make it possible to make inferences about vegetation height and canopy cover [Strunk *et al.*, 2012]. These in turn can be used to predict forest attributes like biomass and volume. It should not, however, be assumed that lidar directly provides information about every individual tree (*e.g.* a tree list) or that lidar-derived height and cover are the same as field measured height and cover [Strunk *et al.*, 2012].

Despite the fact that ALS data can convey a rich summary of forest features due to its ability to determine the heights of forests, it is worth mentioning that the accuracy of retrievable information is highly dependent on both the extrinsic specifications of the ALS survey as well as the ground plot quality data.

For the former aspect, we must consider that the following may affect lidar data for forest measurements:

- mission specification (pulse rate, scanning pattern, flying height, airspeed, pulse diameter, etc.);
- time of year (leaf-off, leaf-on, snow free, etc.);
- lidar sensor and data processing;
- experience of lidar vendor.

ALS data most commonly used for forestry applications under alpine, temperate and Mediterranean environments are not produced by dedicated flights. Rather, forester technicians generally exploit the canopy height model available at low cost [Clementel *et al.*, 2012] or even at no cost from ALS surveys carried out for other non-forestry purposes (topographic measurements, DTM production, etc.). These surveys are frequently characterized by a relatively low point density per square meter and, additionally, are usually carried out in winter to minimize the noise by vegetation and to ensure a high number of pulses reaching the ground [Kraus and Pfeifer, 1998]. In summer, penetration rates less than 25-30% are normally obtained under forest stands [Kilian et al., 1996]. In the case of deciduous species (*i.e.* without green foliage during winter), such as those most frequent in temperate and Mediterranean forest environments, penetration rates can significantly be raised by winter surveys [Ackermann et al., 1994]. Thus, commercial ALS flights for DTM production are almost always made between November and March under temperate and Mediterranean environments and even under alpine environments, at least in those areas not permanently covered by snow during that period. On the contrary, the most favorable season to obtain ALS data specifically suitable for forestry applications under temperate and Mediterranean environments is during summer, since in winter only the woody part of the prevalently deciduous canopies generates lidar-significant returns. For example, Clementel *et al.* [2012] studied the correlation between metrics extracted from a CHM built with airborne laser scanner data collected in summer 2008 summer season and those extracted from a CHM built with airborne laser scanner data collected in winter 2006 in Trento province. They applied map algebra-focal statistical functions to the CHMs (square matrix of 31 m side) before extracting lidar metrics to reduce the effect of the unsystematic spatial displacement between homologous pixel of summer and winter CHM. The results showed a strong correlation (R = 0.98) between summer mean height and winter mean height, but with a significant underestimation of the mean height by the winter CHM. Therefore foresters and researchers in forestry field should use airborne laser scanner data coming from summer flight. Moreover when airborne laser scanning campaigns are being planned, knowing they will utilize such data, they should be asked to be involved in the definition of the terms of contract.

As regards to the second aspect, Reutebuch *et al.* [2000] suggests that to maximize quality, ground-plot data should be:

- collected in plots the geographic position of which reached an accuracy circa equal to the lidar horizontal accuracy (~ ± 1 m);
- collected in plots large enough to minimize plot edge effect, but small enough to characterize tree size differences within plots (~ 400-800 m^2 circular plot);
- collected in plots matching in time the lidar sampling (generally within 1-2 years of lidar acquisition);

- matched in what is measured by the lidar and on the plot: all stems that make up a significant portion of the above ground canopy (generally down to a 7-10 cm diameter at breast height lower limit, including all species).

In support of the arguments of Reutebuch *et al.* [2000], Dalponte *et al.* [2011] investigated the role of positioning error of ground plots and the optimal number of field measurements needed for area-based prediction of stem volume. For their purposes, Dalponte *et al.* [2011] used lidar data collected in 2007-2008 winter season over an area of more than 6000 km^2 in Trento Province (North East of Italy). An ordinary least squares method was used to estimated the volume of 799 plots using 13 predictor variables extracted from airborne laser scanner data. Their results showed that the error due to the position of the plots surveyed by means of a Global Positioning System (GPS) receiver (positioning error of about 10 m) did not significantly influence the prediction accuracy and that the reduction of the number of training samples does not compromise the performance of the prediction model. A sub-sample of only 53 training plots selected using a new proposed method based on genetic algorithms, instead of 534 as in the original dataset, was enough to obtain equally accurate results.

1.2 Area-based and individual tree detection approaches

There are two main approaches to derive forest information from laser scanner data: the area-based approach (ABA), and the individual tree detection approach (ITDA) [Vastaranta *et al.*, 2011].

In the ABA, plot level data are related to ALS data aggregated at plot level to estimate stand attributes [Næsset, 2002]. ABA relates CHM raster pixels or point ALS data to measured plot characteristics to build parametric (*e.g.* regression) or non-parametric models to predict the forest attributes of interest. In ABA, volume, biomass, diameter at breast height (dbh), basal area, and other parameters measured at plot level, ranging from one hundred up to one thousand square meters, are related to airborne laser scanner data building parametric models. Models built by regression have been shown to explain the majority of the variation in forest stand height, volume, biomass, and basal area [*e.g.* Næsset, 1997; Hudak *et al.*, 2006; Hollaus *et al.*, 2006; Garcia *et al.*, 2010].

With the methods used in the ITDA [Hyyppä and Inkinen, 1999], individual trees are segmented from the laser point cloud or from the CHM [Vastaranta *et al.*, 2011], and tree level attributes are either determined straight from the point cloud or estimated based on various other ALS features that are extracted for the tree segments [*e.g.* Hyyppä and Inkinen, 1999; Persson *et al.*, 2002; Leckie *et al.*, 2003; Popescu *et al.*, 2003; Maltamo *et al.*, 2009; Vauhkonen *et al.*, 2010]. Data about each individual tree can then be aggregated to any scale required to create stand-level or ecosystem-level estimates [Akay *et al.*, 2009].

ABA and ITDA are not mutually exclusive estimates, and several authors demonstrated how they can be combined [*e.g.* Lindberg *et al.*, 2010; Vastaranta *et al.*, 2012].

For the purposes of this thesis, we used an area-based approach to derive forest information from discrete return lidar data. We chose this approach due to the fact that many studies demonstrated that lidar data may provide spatial detail for planning and the optimization of forest management activities [White *et al.*, 2013; Corona *et al.*, 2012; Woods et al, 2011].

In contrast, an analysis of literature on the performance of single-tree detection algorithms highlighted that there is a non-homogeneous trend in tree detection accuracy. Furthermore, the weaknesses and strengths of ITDA methods depend on the types of forest systems in which they are applied, in that forest structure strongly affects the performance of algorithms [Vauhkonen *et al.*, 2011]. Vauhkonen *et al.* [2011] tested the accuracy of six algorithms developed in Finland, Germany, Norway and Sweden for tree detection by means of airborne laser scanner data in four sites located in Brazil, Germany, Norway, and Sweden. According to their results, the number of trees estimated by the algorithms in proportion to the number of trees actually measured in the field (*i.e.* detection rate) varied between test sites and ranged from 45.2% (Norwegian algorithm) to 100.7% (German algorithm). On average, the number of trees found by the algorithms corresponded to only 65% of the number of trees measured in field. A considerable difference in the performance of the algorithms applied in the test site located in Germany has emerged in the detection of coniferous and deciduous plots (best detection rate in the case of conifers), while lidar data point density had no clear effect on tree detection. In 2006, an international comparison of individual tree detection and extraction using ALS was carried out within the “Tree Extraction” project organized by EuroSDR (European Spatial Data Research) and ISPRS (International Society for Photogrammetry and Remote Sensing) to evaluate the quality, accuracy, and feasibility of automatic tree extraction methods based on laser scanner data [Kaartinen *et al.*, 2012]. Twelve partners from United States of America, Canada, Norway, Sweden, Finland, Germany, Austria, Switzerland, Italy, Poland and Taiwan participated in the test. The partners were requested to extract trees using the given ALS datasets. The results showed that the extraction method is the main factor affecting achieved accuracy (the percentage of detected trees varied between 25% to 102%) while the laser point density (2 to 8 points per square meter) is a factor that strongly affects marginal crown delineation accuracy [Kaartinen *et al.*, 2012]. It is evident that the accuracy of detection rate will influence the accuracy on volume or biomass estimation. Another weak point of the ITDA is that it entails the development of a height-based dendrometric approach versus a diameter-based dendrometric approach as implemented by Abramo *et al.* [2007]. Abramo *et al.* [2007] used the height of detected trees by an ITDA to compute the stem volume of each tree. Practically speaking, the height of each tree extracted from lidar data was used to calculate its diameter using the third-degree polynomial relationship derived from the height-diameter function of each investigated species. This kind of approach does not consider that, when height-diameter function is "overturned", the curve is increasing and upwardly concave, so it tends to reach an asymptote. Thus, reduced variations in height can lead to large differences in diameter size and consequently in volume [Abramo et al., 2007]. As a consequence, the methodology of overturning the height-diameter curve proposed by the authors does not work and does not allow for extrapolation outside of the field where it was originally applied.

This literature review and our preliminary analyses, led to the choice of an area-based approach as the basis of the research presented in this dissertation. For this reason, the description of the state-of-the-art ALS-assisted assessment of forest standing volume and forest stand structure in Italy refers only to the research carried out under an ABA. Moreover, research that integrated lidar data with satellite data were not considered.

1.3 ALS-assisted assessment of forest standing volume in Italy

Since foresters realized the potential of airborne laser scanner data utilization, most research in Italian forestry field became dedicated to the estimation of forest standing volume. The strategy most frequently utilized consist of developing models which can predict forest volume or biomass using ALS-derived statistics calculated in the lidar heights distribution (termed “lidar metrics”). Such metrics can be computed in a lidar point cloud or in CHM.

Corona *et al.* [2007; 2008] have examined the correlation between volume measured in circular plots of 314 m^2 and the sum of the CHM heights (raised to a power) within plots in a temperate broadleaf forest in the Riserva Naturale Stale Bosco della Fontana. The relationship between volume and height has proved to be heteroscedastic and linear through the origin, with a correlation coefficient equal to 0.78. Starting from this finding, Corona and Fattorini [2008] have subsequently proposed a design-based approach for the ABA ALS-assisted estimation of forest standing volume: adopting the height of the upper canopy at pixel level as auxiliary information, a ratio estimator of the total volume may be derived, together with the unbiased estimator of its sampling variance and the corresponding confidence interval. With reference to the above-mentioned broadleaf forest, these authors have shown that the use of ALS data as auxiliary information can give rise to a confidence interval for the estimation approximately between half and two-thirds smaller than that obtained by the direct sampling of field plots. Barbati *et al.* [2009] have tested the same estimation approach in a coastal Mediterranean pine forest in Tuscany region, finding a correlation coefficient between the volume measured in circular plots of 1256 m^2 and the sum of the CHM heights within the plots equal to 0.88.

A study by Fusco *et al.* [2008] on mesophilous coppice stands in Lombardy region confirms the relationship between the woody biomass and the sum of CHM heights, with a correlation coefficient equal to 0.87.

Extensive ABA research for assessing forest standing volume under alpine environments in Trento province has been carried out by Floris *et al.* [2010]: these authors developed regression models between the mean height of summer CHM and standing volume measured in sample plots (20 m of radius) in spruce forests of Paneveggio in Trento province to estimate the volume per surface unit. They found that the mean height of CHM pixels with height greater than 2 m can reliably used as predictor variable to estimate forest volume. Floris *et al.* [2010] model for the prediction of timber volume was able to explain the 88% of variability when externally validated in 60% of observations that contributed to the model development.

Also, under alpine conditions, Tonolli *et al.* [2011] have analyzed the relationship between the standing volume measured on plots with sizes ranging from 400 m^2 to 3600 m^2 and metrics obtained from point ALS data: the most explicative metrics resulted in the mean height of the first return and, with much less importance, the median of the heights of the second return. These authors also found that the relationship is significantly influenced by the plot size, with correlation coefficients increasing from 0.7 to 0.8 in the considered plot size range.

Clementel *et al.* [2012] developed a regression model for the Foresta di Cadino (Trento province) to estimate the volume per hectare using as predictor variable the mean height extracted from the winter CHM of 365 circular plots with radius of 20 m. The resulting determination coefficient was equal to 0.58, the root mean square error equal to 136.34 m^3/ha and the relative standard error equal to 26.87% (model's accuracy at plot level).

The results illustrated above demonstrate that methodologies used according an ABA have reached maturity to be used for mapping forest canopy height, volume and biomass under complex environments throughout large areas.

1.4 ALS-assisted assessment of forest stand structure in Italy

The potential of airborne laser scanner data also refers to applications for the assessment of forest stand structure.

First exploitations regarded the production of thematic maps whose goal was characterizing the forests according to indices extracted from the canopy height model built from laser scanner data.

Barilotti *et al.* [2005] described the canopy cover of coniferous-dominated and mixed forests in Friuli Venezia Giulia region (North-East of Italy) by the Laser Penetration Index (LPI), which is the capability of the laser pulses to reach the ground, and hence it is able to express the vegetation cover. Barilotti *et al.* [2005] observed that LPI is inversely proportional to tree height and stand density, and inversely proportional to the Leaf Area Index (LAI). The correlation between LAI and LPI was highly significant (R = 0.94), evidencing the potential of airborne laser scanner data to be used as input variables of models used in ecology which needed values of leaf area index.

ALS data have been used for forest canopy gap detection, which is relevant to study ecological stand dynamics, to exclude gaps from forested areas to be sampled for forest inventory purposes, and consequently to increase the sampling efficiency. For this purpose, Barbati *et al.* [2009] have applied the hot-spot analysis to the CHM based on the calculation of the local statistic variable Gi* of Getis-Ord [Getis and Ord, 1992] in a coastal Mediterranean pine forest of Tuscany region. To identify the forest gaps Bottalico *et al.* [2009] applied the approach proposed by Koukoulas and Blackburn [2004] to the CHM of the same study area but the results were less successful in the smaller gaps identification compared to those obtained applying the hot spot analysis by Barbati et al [2009].

Also Floris *et al.* [2009] carried out analysis to evaluate forest gaps in a wider study context aimed at producing thematic maps about forest gaps, canopy cover, and canopy roughness. They applied map algebra focal operators (matrix 15 m x 15 m wide) to the CHM of Foresta di Cadino (Trento province). They first performed a binary map (0,1) attributing 1 if the value of the central pixels of the moving window was higher than 5 m (a so called tree-pixel) and then considering no gap each pixel that was surrounded in the neighboring matrix by at least 20% of tree-pixels. The authors produce a canopy cover map attributing to a pixel of CHM the normalized value obtained from the ratio between the tree-pixel and the total number of pixels explored by the matrix. The canopy roughness map was performed assigning to the central pixel of the window the value of standard deviation of the heights in the matrix. These kinds of maps, if appropriately combined, can indicate stratification for forest inventory purposes.

A few authors used lidar data to classify forest structure. Chirici *et al.* [2013] used ALS metrics and spectral signatures from Indian Remote Sensing (IRS) imageries, extracted for a set of plots in Sicily region (South Italy), as predictor variables to *a posteriori* classify forest fuel types observed and identified by photo interpretation and fieldwork. We will not consider research integrating lidar data with satellite data in this review.

Torresan *et al.* [2014] used lidar-derived CHM metrics to predict forest structure types according to the amount of basal area present in understory, mid-story, and overstory trees. The results presented by the authors were obtained through research activities carried out during this PhD work and will be appropriately described in this dissertation.

1.5 Research problems and aims of the research activities

Researches activities carried out during the PhD were aimed at providing specific solutions to forest management problems and desires encountered by the Forest Service of Trento Province (in Italian language named Provincia Autonoma di Trento, PAT). Incidentally, we want to point out that when province is written in capital (*i.e.* Province) we are referring to the administrative province, while when written in lower case (*i.e.* province) we are referring to the territory of province.

Trento Province owns and manages 72% of forests presented in the province [De Natale and Pignatti, 2011]: timber sales represent an item which is linked to the Trento Province balance finances, and, therefore, accurate knowledge about the amount of forest resources is very important to make a financial plan. Information about state and changes of forest stands is fundamental for environmental and wood production assessment at various levels of forest management planning [*e.g.* Schreuder *et al.*, 1993]. At the level of forest estate, information is usually acquired from field inventories which provide estimates for the overall area, and for the strata by which the area is partitioned.

Since 2010, the Forest Service of the Trento Province has changed its inventory system for forest management purposes from full callipering on forest compartment basis to stratified sampling. The Forest Service achieves the sampling process through two phases: the stratification and the out-and-out sampling. In the stratification process the forest is subjectively subdivided into homogeneous sub-populations (the strata). Sampling is done in a number of plots and their location in each stratum is defined. Successively, field inventories allow data collection to update figures at strata and forest estate levels for the estimation of key attributes (standing wood volume, in particular timber volume). But forest compartment is the basic unit for forest management activities (tending of stand, felling, etc.), so the growing stock estimation is needed at forest compartment level. Forest compartments are forest management units, extended 20 ha on average, delineated for operative and harvesting purposes, whose boundaries, as far as possible, are easily detectable in the field (forest roads, streams, power lines, etc.). Considering that the intersection between strata and forest compartments produces the so called portions, at the moment, the volume of a forest compartment is computed as a weighted sum on the portion surfaces of the average volume of each stratum intersecting the forest compartment. This kind of procedure may produce inaccurate estimates due to the small number (or even the absence) of plots within a portion. The scarcity of sampling information within small portions of the survey area is a well-known problem of sample surveys. It is referred to as the small area estimation problem [*e.g.* Rao, 2005]. Usually the lack of sampling information is recovered by the use of auxiliary information available from other sources [Pfefferman, 2002], in particular from remote sensing. Considering that Forest Service of Trento Province has at its disposal a Canopy Height Model, and considering that CHM represents an effective information source to support forest volume estimation (refer to § 1.3), an objective of the dissertation was to estimate the volume at forest compartment level using the height from CHM as auxiliary information adopting a model supposing a proportional relationship between height and volume at pixel level for driving estimation. In particular we wanted consider the estimation from three points of view: the model-based perspective, the design-based perspective, and the hybrid-perspective and evaluate which is the inference approach that produces the more accurate estimation level. All these aspects are tackled and discussed in Chapter 3.

Another specific need of the Forest Service is the prediction of the structure of forests owned by the Province for mapping on a large scale. The Forest Service of Trento Province presently uses a structural classification system, both in inventory and forest management decision making phases, based on the distribution of basal area in three diameter size-classes. The classification in forest structural types (12 in total) is one of the primary references for forest managers to identify the strata in the stratification process. During the decision making process, forest managers use the forest structural classification system to identify the potential transition of a specific structural type toward another one, and consequently to decide which kind of silvicultural intervention has to be implemented by estimating the potential volume to fell in each tree diameter size-class [Ancel *et al.*, 1999; Scrinzi *et al.*, 2011].

Lidar, as an active remote sensing technology has been used in many studies outside Italy to describe and estimate forest structure in three-dimensions using the raw point cloud to compute a wide range of lidar metrics to predict and map many aspects of forest composition and structure, including basal area [Næsset, 1997; Means *et al.*, 1999; Dubayak *et al.*, 2000; Means *et al.*, 2000; Næsset, 2002; Andersen *et al.*, 2003; Zimble *et al.*, 2003; Riaño *et al.*, 2004; Maltamo *et al.*, 2005; Heurich and Thoma, 2008; Ferraz *et al.*, 2012]. There has been only limited empirical exploration of the potential use of only digital elevation models (DTM and DSM) in forest characterization and classification [Suárez *et al.*, 2012] because it has been typically believed that these two elevation models have limited applications for forestry purposes that require characterization of forest structure. But there are times that the original raw lidar point cloud may be not available. In fact, despite declining costs and increasing coverage, lidar data are still not widely available in all locations for forestry applications, so the ability to use lidar-derived DTMs and DSMs has the potential to improve forest classification and mapping in areas where raw lidar data are otherwise not available.

This dissertation tries to assess whether lidar-derived predictor variables extracted from CHMs can support the classification of the forests in the province of Trento, a domain that is still relatively unexplored and deserving of further attention. Within this general purpose, the specific goal was to compare the performance of supervised classification techniques toward specific methods of basal area prediction in the process of structural classification of the mountain forests. All these aspects are engaged and discussed in the Chapter 4.

Considering that the number of forest types in the forest classification system used by the Forest Service of Trento Province is rather high, we wanted to classify the forests of Trento province on the basis of basal area distribution of understory, mid-story, and overstory trees through an unsupervised procedure. Afterwards we wanted to evaluate if lidar-derived predictor variables extracted from raw data were able to distinguish between the discriminated types by means of learning machine techniques. In particular we wanted to evaluate the performance of classification trees in distinguishing forest stands. The Chapter 4 is dedicated to the description of the processes carried out to achieve these purposes.

To sum up, this thesis attempts to develop statistical analysis and techniques for:

- forest volume estimation at forest compartment level using metrics extracted from the CHM as auxiliary information for the local calibration of forest inventory;
- classification of structure of forest stands using the CHM and lidar raw data to support forest inventory process (*e.g.* stratification) and forest management strategies (*e.g.* definition of silvicultural interventions, identification of habitat requiring protection, identification of habitat with high level of heterogeneity).

1.6 Structure of the thesis

The thesis is organized into six chapters.

The present chapter, after briefly describing the airborne laser scanning technology, highlighted the stat-of-art of ALS-assisted assessment of forest standing volume and stand structure in Italy, pointed out the background and the motivations for this thesis, as well as the objectives.

In the Chapter 2 we provide the results of a study about the comparison of forest volume estimation by the model based on CHM-derived metrics developed by Floris *et al.* [2010] and by double-entry stem volume equations (based on full callipering and height-diameter functions) at forest compartment level.

Chapters 3 is focused on the strategies for the volume estimation at forest compartment level based on a ratio model presuming a proportional relationship between height from CHM and volume at pixel level. The volume estimator and the variance estimators are considered from the model-based, design-based and hybrid inferences.

Chapter 4 is dedicated to the comparison among statistical techniques for the classification of forest stand structure using CHM-derived metrics. In detail we examined the use of lidar-derived CHM metrics to predict forest structure classes according to the amount of basal area present in understory, mid-story, and overstory trees. We evaluated two approaches to predict diameter size-based forest classifications: in the first, we attempted supervised classification with both linear discriminant analysis (LDA) and random forest (RF); in the second, we predicted basal areas of lower, mid, and upper canopy trees from CHM-derived variables by *k*-nearest neighbor imputation (*k*-NN) and parametric regression, and then classified observations based on their predicted basal areas.

In Chapter 5 the evaluation of the performance of classification trees for discriminating forest stand structure using lidar point cloud-derived metrics is described.

In the last chapter of the dissertation conclusions on the proposed statistical analysis and techniques for volume estimation and stand structural classification are given. Furthermore, limitations, and future developments of the dissertation are discussed.

1.7 References

Abramo E., Barilotti A., Sepic F., 2007. *Dalla dendrometria diametrica alla dendrometria ipsometrica: stima del volume degli alberi da rilievi laser-scanning.* Foresta@ 4(4): 373-385. doi: 10.3832/efor0481-0040373.

Ackermann F., Englich M., Kilian J., 1994. *Die Laser-Profil-Befliegung Gammertingen*. ZfV 119(5), pp. 264-277.

Akay, A., Oðuz, H., Karas, I., Aruga, K., 2009. *Using lidar technology in forestry activities.* Environmental Monitoring and Assessment, 151: 117-124. doi: 10.1007/s10661-008-0254-1.

Ancel, P., Wilhelm, M.E., Lacombe, E., 1999. *Peuplements forestiers du massif Vosgien; typologie et sylvicultures.* CNPF. pp. 51.

Andersen, H.E., Foster, J.R., Reutebuch, E., 2003. *Estimating forest structure parameters on Fort Lewis Military Reservation using Airborne Laser Scanner (LIDAR) data.* Proceedings of the Second International Precision Forestry Symposium, Seattle, Washington, June 15-17, 2003, pp. 45-53.

Barbati, A., Chirici, G., Corona, P., Montaghi, A., Travaglini, D., 2009. *Area-based assessment of forest standing volume by field measurements and airborne laser scanner data.* International Journal of Remote Sensing. Vol. 30, No. 19, 10 October 2009, pp. 5177–5194. doi: 10.1080/01431160903023017.

Barilotti, A., Turco, S., Napolitano, R., Bressanc, E., 2005. *La tecnologia lidar per lo studio della biomassa negli ecosistemi forestali.* Proceedings of 15th Meeting of the Italian Society of Ecology, 12-14 September 2005, Torino.

Beraldin, J.-A., Blais, F., Lohr, U., 2010. *Laser Scanning Technology*. In: Vosselman, G., Maas, H.-G. Editors. Airborne and terrestrial laser scanning. Whittles publishing, CRC Press, Taylor & Francis Group, 2010. ISBN 978-1904445-87-6.

Bottalico, F., Montaghi, A., Travaglini, D., 2009. *Identificazione dei gaps nella copertura forestale con dati lidar.* Atti 13a Conferenza Nazionale ASITA, Bari, 1-4 Dicembre 2009.

Bretar, F., Chauve, A., Mallet, C., Jutzi, B, 2008. *Managing full waveform lidar data: a challenging task for the forthcoming years*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B1. Beijing 2008, pp. 415-420.

Chirici, G., Scotti, R., Montaghi, A., Barbati, A., Cartisano, R., Lopez, G., Marchetti, M., McRoberts, R.E., Olsonn, H., Corona, P., 2013. *Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery.* International Journal of Applied Earth Observation and Geoinformation, 25: 87-97. doi: 10.1016/j.jag.2013.04.006.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., Torresan, C., 2012. *Estimating forest timber volume by means of "low-cost" lidar data*. Italian Journal of Remote Sensing, 44(1): 125-140. doi:10.5721/ItJRS201244110.

Corona, P., Fattorini, L., 2008. *Area-based lidar-assisted estimation of forest standing volume.* Canadian Journal of Forest Research, 38(11): 2911-2916. doi: 10.1139/X08-122.

Corona, P., Lamonaca, A., Chirici, G., Travaglini, D., Marchetti, M., Minari, E., Montaghi, A., 2008. *Estimation of growing stock of broadleaved forests by airborne laser scanning*. pp 39-44. In: Gianelle, D., Travaglini, D., Mason, F., Minari, E., Chirici, G., Chemini, C. (eds.). Canopy analysis and dynamics of a floodplain forest. Rapporti Scientifici, 3. Centro Nazionale per lo studio e la conservazione della Biodiversità Forestale – Bosco della Fontana. Cierre Grafica Editore, Verona, pp. 96.

Corona, P., Cartisano, R., Salvati, R., Chirici, G., Floris, A., Di Martino, P., Marchetti, M., Scrinzi, G., Clementel, F., Travaglini, D., Torresan., C., 2012. *Airborne Laser Scanning to support forest resource management under alpine, temperate and Mediterranean environments in Italy*. European Journal of Remote Sensing, 45: 27-37, doi: 10.5721/EuJRS20124503.

Corona, P., Lamonaca, A., Chirici, G., Travaglini, D., Marchetti, M., Minari, E., Montaghi, A., 2007. *Sviluppo di una procedura innovativa per la cubatura del volume della massa legnosa di fustaie di latifoglie mediante lidar*. In: Atti della 11a Conferenza Nazionale ASITA, Torino, 6–9 Novembre, Vol. 1, pp. 767–772.

Dalponte, M., Martinez, C., Rodeghiero, M., Gianelle, D., 2011. *The role of ground reference data collection in the prediction of stem volume with lidar data in mountain areas*. ISPRS Journal of Photogrammetry and Remote Sensing, 66: 787–797.

De Natale, F., Pignatti, G., 2011. *Estensione e composizione dei boschi italiani.* In: Gasparini, P., Tabacchi, G. (a cura di) 2011 - L'Inventario Nazionale delle Foreste e dei serbatoi forestali di Carbonio INFC 2005. Secondo inventario forestale nazionale italiano. Metodi e risultati. Ministero delle Politiche Agricole, Alimentari e Forestali, Corpo Forestale dello Stato; Consiglio per la Ricerca e Sperimentazione in Agricoltura, Unità di Ricerca per la Pianificazione ed il Monitoraggio e la Pianificazione Forestale. Edagricole, Milano, pp. 99-111.

Dubayah, R.O, Knox, R.G, Hofton, M.A., Blair, J.B., Drake, J.B., 2000. *Land surface characterization using lidar remote sensing*. In, Spatial information for land use management, Hill, M., Aspinall, R.(Eds.), International Publishers Direct, Singapore (2000).

Ferraz, A., Bretar, F., Jacquemound, S., Gonçalves, G., Pereira, L., Tomé, M., Soares, P., 2012. *3-D mapping of multy-layered Mediterranea forest using ALS data.* Remote Sensing of Environment, 121: 210-223. doi: 10.1016/j.rse.2012.01.020.

Floris, A., Clementel, F., Farruggia, C., Scrinzi, G., 2009. *Il lidar nella stratificazione tematica dei soprassuoli forestali: applicazioni in Trentino.* Atti 13a Conferenza Nazionale ASITA, Bari, 1-4 dicembre 2009.

Floris, A., Clementel, F., Farruggia, C., Scrinzi, G., 2010. *Stima su base lidar delle provvigioni legnose forestali: uno studio per la Foresta di Paneveggio*. Rivista Italiana di Telerilevamento, 42(3): 15-32.

Fusco, S., Pflugmacher, D., Kirschbaum, A., Cohen, W., Chiatante, D., Montagnoli, A., 2008. *Uso di dati lidar per stima della biomassa forestale in un bosco misto di latifoglie: un caso studio in Valsassina (LC).* Atti 12a Conferenza Nazionale ASITA, L'Aquila, 21-21 Ottobre 2008.

García, M., Riaño, D., Chuvieco, E., Danson, F.M., 2010. *Estimating biomass carbon stocks for a Mediterranean forest in central Spain using lidar height and intensity data.* Remote Sensing of Environment, 114: 816-830. doi: 10.1016/j.rse.2009.11.021.

Getis, A., Ord, K., 1992. *The analysis of spatial association by use of distance statistics.* Geographical Analysis, 24(3): 189–206. doi: 10.1111/j.1538-4632.1992.tb00261.x.

Heurich, M., Thoma, F., 2008. *Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural European beech (Fagus sylvatica) and Norway spruce (Picea abies) forests.* Forestry, 81(5): 645-661. doi: 10.1093/forestry/cpn038.

Hollaus, M., Wagner, W., Eberhöfer, C., Karel, W., 2006. *Accuracy of large-scale canopy heights derived from lidar data under operational constraints in a complex alpine environment.* ISPRS Journal of Photogrammetry and Remote Sensing, 60(5): 323-338. doi: 10.1016/j.isprsjprs.2006.05.002.

Hudak, A.T., Crookston, N.L., Evans, J.S., Falkowski, M.J., Smith, A.M.S., Gessler, P.E., Morgan, P., 2006. *Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data*. Canadian Journal of Remote Sensing, 32: 126-138. doi: 10.5589/m06-007.

Hug, C., Ullrich, A., Grimm, A., 2004. *A Waveform-Digitizing LIDAR Terrain and Vegetation Mapping System*. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. 36 (Part 8/W2). Freiburg, Germany, pp. 24-29.

Hyyppä, H., Hyyppä, J., 1999. *Comparing the Accuracy of Laser Scanner with Other Optical Remote sensing data sources for stand attribute retrieval*. The photogrammetric journal of Finland, 01/1999, 16: 5-15.

Hyyppä, J., Inkinen, M., 1999. *Detecting and estimating attributes for single trees using laser scanner*. Laser Radar Technology and Applications, IV, 3707:57-69.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopaine, M., Heipke, C., Hirschmugl, M., Morsdorf, F., Næsset, E., Pitkänen, J., Popescu, S., Solberg, S., Bernd, M.W., Wu, J.C., 2012. *An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning*. Remote Sens. 2012, 4(4): 950-974. doi:10.3390/rs4040950.

Kilian, J., Haala, N., Englich, M., 1996. *Capture and evaluation of airborne laser scanner data.* International Archive of Photogrammetry and Remote Sensing. Volume 31, Part B3, pp. 383-388.

Koukoulas, S., Blackburn, G.A., 2004. *Quantifying the spatial properties of forest canopy gaps using lidar imagery and GIS.* International Journal of Remote Sensing, 25(15): 3049-3071. doi: http://dx.doi.org/10.1080/01431160310001657786.

Kraus, K., Pfeifer, N., 1998. *Determination of terrain models in wooded areas with airborne laser scanner data.* ISPRS Journal of Photogrammetry and Remote Sensing, Volume 53, Issue 4, August 1998, pp. 193-203. doi: 10.1016/S0924-2716(98)00009-4.

Leckie, D., Gougeon, F., Hill, D., Quinn, R., Armstrong, L., Shreenan, R., 2003. *Combined high-density lidar and multispectral imagery for individual tree crown analysis*. Canadian Journal of Remote Sensing, 29(5): 1–17.

Lindberg, E., Holmgren, J., Olofsson, K., Wallerman, J., Olsson, H., 2010. *Estimation of tree lists from airborne laser scanning by combining single-tree and area-based methods.* International Journal of Remote Sensing, 31: 1175-1192. doi: 10.1080/01431160903380649.

Maas, H.-G., 2010. *Forestry Applications*. In: Vosselman, G., Maas, H.-G. Editors. Airborne and terrestrial laser scanning. Whittles publishing, CRC Press, Taylor & Francis Group, 2010. ISBN 978-1904445-87-6.

Maltamo, M., Packalén, P., Yu, X., Eerikäinen, K., Hyyppä, J., Pitkänen, J., 2005. *Identifying and quantifying structural characteristics of heterogeneous boreal forests using laser scanner data*. Forest Ecology and Management, 216: 41-50. doi: 10.1016/j.foreco.2005.05.034.

Maltamo, M., Peuhkurinen, J., Malinen, J., Vauhkonen, J., Packalén, P., Tokola, T., 2009. *Predicting tree attributes and quality characteristics of Scots Pine using airborne laser scanning data*. Silva Fennica 43(3): 507-521.

Means, J.E., Acker, S.A., Fitt, B.J., Renslow, M., Emerson, L., Hendrix, C.J., 2000. *Predicting forest stand characteristics with airborne scanning lidar.* Photogrammetric Engineering & Remote Sensing, 66(11): 1367-1371.

Means J.E., Acker S.A., Harding D.J, Blair J.B., Lefsky M.A., Cohen W.B., Harmon M.E., McKee W.A., 1999. *Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon.* Remote Sensing of Environment, 67: 298-308. doi: 10.1016/S0034-4257(98)00091-1.

Næsset E., 1997. *Estimating timber volume of forest stands using airborne laser scanner data.* Remote Sensing of Environment, 61: 246-253. doi: 10.1016/S0034-4257(97)00041-2.

Næsset, E., 2002. *Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data.* Remote Sensing of Environment, 80: 88-99. doi: 10.1016/S0034-4257(01)00290-5.

Persson, Å., Holmgren, J., Söderman, U., 2002. *Detecting and measuring individual trees using an airborne laser scanner*. Photogrammetric Engineering & Remote Sensing, 68: 925-932.

Pfefferman, D., 2002. *Small area estimation – New developments and directions*. Int. Stat. Rev, 70(1): 125-143. doi:10.1111/j.1751-5823.2002.tb00352.x.

Pirotti, F., 2011. *Analysis of full-waveform lidar data for forestry applications: a review of investigations and methods*. iForest 4: 100-106 [online 2011-06-01] URL: http://www.sisef.it/iforest/show.php?id=562.

Popescu, S.C., Wynne, R.H., Nelson, R.F., 2003. *Measuring individual tree crown diameter with lidar and assessing its influence on estimating forest volume and biomass*. Canadian Journal of Remote Sensing, 29: 564-577.

Rao, J.N.K., 2005. *Small Area Estimation*. Wiley. New York. pp. 344.

Reutebuch, S.E., Andersen, H.E., McGaughey, R.J., 2005. *Light Detection and Ranging (LIDAR): An Emerging Tool for Multiple Resource Inventory*. Journal of Forestry, 103(6): 286-292.

Riaño D., Valladares F., Condésand S., Chuvieco E., 2004. *Estimation of leaf area index and covered ground from airborne laser scanner (Lidar) in two contrasting forests.* Agricultural and Forest Meteorology, 124: 269-275. doi: 10.1016/j.agrformet.2004.02.005.

Schreuder, H., Gregoire, T., Wood, G., 1993. *Sampling methods for multiresources forest inventory*. Wiley, New York. pp. 464.

Scrinzi, G., Clementel, F., Colle, G., Farruggia, C., Floris, A., Torresan, C., 2011. *L'inventario dendromemtrico nella nuova pianificazione forestale aziendale trentina (NPFAT).* Provincia Autonoma di Trento - Servizio Foreste e Fauna, Consiglio per la Ricerca e la Sperimentazione in Agricoltura, pp. 105. ISBN 978-88-7702-322-3.

Strunk, J.L., 2012. *Estimation and Modeling of Selected Forest Metrics with Lidar and Landsat*. Thesis for the degree of Doctor of Philosophy in Forest Resources presented on June 14, 2012, pp. 145.

Suárez, J., Fonweban, J., Gardiner, B., 2012. *Supporting precision forestry in Great Britain.* Earthzine (http://www.earthzine.org/2012/06/22/supporting-precision-forestry-in-great- britain/).

Tonolli, S., Dalponte, M., Vescovo, L., Rodeghiero, M., Bruzzone, L., Gianelle, D., 2011. *Mapping and modeling forest tree volume using forest inventory and airborne laser scanning*. Eurepean Journal of Forest Research, 130: 569-577. doi:10.1007/s10342-010-0445-5.

Torresan, C., Strunk, J., Zald, H.S.J., Zhiqiang, Y., Cohen, W.B., 2014. *Comparing statistical techniques to classify the structure of mountain forest stands using CHM-derived metrics in Trento province (Italy)*. European Journal of Remote Sensing - 2014, 47: 75-94. doi: 10.5721/EuJRS20144706.

Vastaranta, M., Kankare, V., Holopainen, M., Yu, X., Hyyppä, J., Hyyppä, H., 2012. *Combination of individual tree detection and area-based approach in imputation of forest variables using airborne laser data.* ISPRS Journal of Photogrammetry and Remote Sensing, 67: 73-79. doi: 10.1016/j.isprsjprs.2011.10.006.

Vastaranta, M., Holopainen, M., Yu, X., Haapanen, R., Melkas, T., Hyyppä, J., Hyyppä, H., 2011. *Individual tree detection and area-based approach in retrieval of forest inventory characteristics from low-pulse airborne laser scanning data*. The Photogrammetric Journal of Finland, Vol. 22, No. 2, 2011.

Vauhkonen, J., Ene, L., Gupta, S., Heinzel, J., Holgren, J., Pitkänen, J., Solberg, S., Wang, Y., Weinacker, H., Hauglin, K.M., Lien, V., Packalén, P., Gobakken, T., Koch, B., Næsset E., Yokola, T., Maltamo, M., 2012. *Comparative testing of single-tree detection algorithms under different types of forest*. Forestry, Vol. 85, No. 1, 2012. doi:10.1093/forestry/cpr051.

Wagner, W., Ullrich, A., Melzer, T., Briese, C., Kraus, K., 2004. *From Single-Pulse to Full-Waveform Airborne Laser Scanners: Potential and Practical Challenges*. Talk: International Society for Photogrammetry and Remote Sensing XXth Congress, Istanbul; 2004-07-12 - 2004-07-23; in: "*Proceedings*", O. Altan (ed.); Vol XXXV, Part B/3 (2004), ISSN: 1682-1750; 201 - 206.IAPRS, 35, PartB, pp. 201 ff.

White, J.C., Wulder, M.A., Varhola, A., Vastaranta, M., Coops, N.C., Cook, B.C., Pitt, D., Woods, M., 2013. *A best practices guide for generating forest inventory attributes froma airborne laser scanning data using an area-based approach (Version 2.0)*. Canadian Forest Service, Canadia Wood Fibre Centre, Information Report FI-X-010.

Woods, M., Pitt, D., Penner, M., Lim, K., Nesbitt, D., Etheridge, D., Treitz, P., 2011. *Operational implementation of a lidar inventory in boreal Ontario*. The Forestry Chronicle 87(4): 512-528.

Yu, X., Hyyppä, J., Holopainen, M., Vastaranta, M, 2010. *Comparison of area-based and individual tree-based methods for predicting plot-level forest attributes*. Remote Sens. 2(6): 1481-1495. doi: 10.3390/rs2061481.

Zimble, D.A., Evans, D.L., Carlson, G.C., Parker, R.C., Grado, S.C., Gerard, P.D., 2003. *Characterizing vertical forest structure using small-footprint airborne lidar.* Remote Sensing of Environment, 87: 171-182. doi: 10.1016/S0034-4257(03)00139-1.

2 Full validation of a linear model based on CHM-derived metrics for timber volume estimation

For timber volume estimation, models developed using airborne laser scanner data are already a concrete alternative to the traditional estimation systems. This chapter illustrates a comparison between values of timber volume of two forest compartments in the Paneveggio forest estimated through double-entry stem volume equations and through a linear regression model based on a CHM-derived metric, developed by Floris et al. [2010].

2.1 Introduction

Since 2007, the Forest Geomatics research laboratory of Forest Monitoring and Management Research Unit (MPF) of Agricultural Research Council (CRA) focuses its research on the use of airborne laser scanner data for forestry application on mountainous forests of Alps (North Italy).

Most of studies of Geomatics lab of MPF Unit supported the process established by the Forest Service of Provincia Autonoma di Trento aimed at developing new and innovative forest inventory procedures based on the use of remote sensing data, in particular airborne laser scanner data.

The first studies were oriented to develop ALS data processing techniques to compute indices from the canopy height model useful for the thematic classification of forest stands [Floris *et al.*, 2009] and addressed to support forest technicians in the stratification process for forest inventory purposes.

At a later stage, research was concentrated in the assessment of volume of forests owned by public institutions of Trento province (*i.e.* municipalities, forest agencies, Trento Province) and managed for commercial purposes. Volume estimation models were carried out by regressing dendrometric response variables, *i.e.* volume referred to different tree diameter size, against metrics extracted from the CHM processed from airborne laser scanner data collected trough campaigns on Trento province territory [Floris *et al.*, 2010; Clementel *et al.*, 2010; Clementel *et al.*, 2012].

In most cases, Geomatics lab of MPF Unit used lidar data collected for other non-forestry purposes, used area-based methodological approaches, and developed regression models for an easy operational use, with a low number of explicative variables.

The study here described aimed at full validating the statistical linear model for timber volume estimation developed by Floris *et al.* [2010], herein called "Model A". In particular, timber volume values estimated with Model A were compared with those computed by the application of full callipering and double-entry stem volume equations (ground truth) in two forest compartments located in Paneveggio forest (Trento province).

In the materials and methods paragraph, after describing the study area, how the linear Model A was developed and how it was used in this study to estimate the volume of two forest compartments is illustrated. In the results section values of volume in two forest compartments of Paneveggio forest obtained from the application of stem volume double-entry equations (diameter from a full-callipering and height from a height-diameter function specifically built for each forest compartment) and those obtained from the volume model developed by Floris *et al.* [2010] are compared.

2.2 Materials and methods

2.2.1 Study area

The study area is situated in the Eastern side of the province of Trento and cover 370 ha of "Foresta demaniale di Paneveggio" (namely Paneveggio forest), 2800 ha wide, which is part of the natural park called "Parco Naturale Paneveggio - Pale di San Martino" (Figure 2.1).

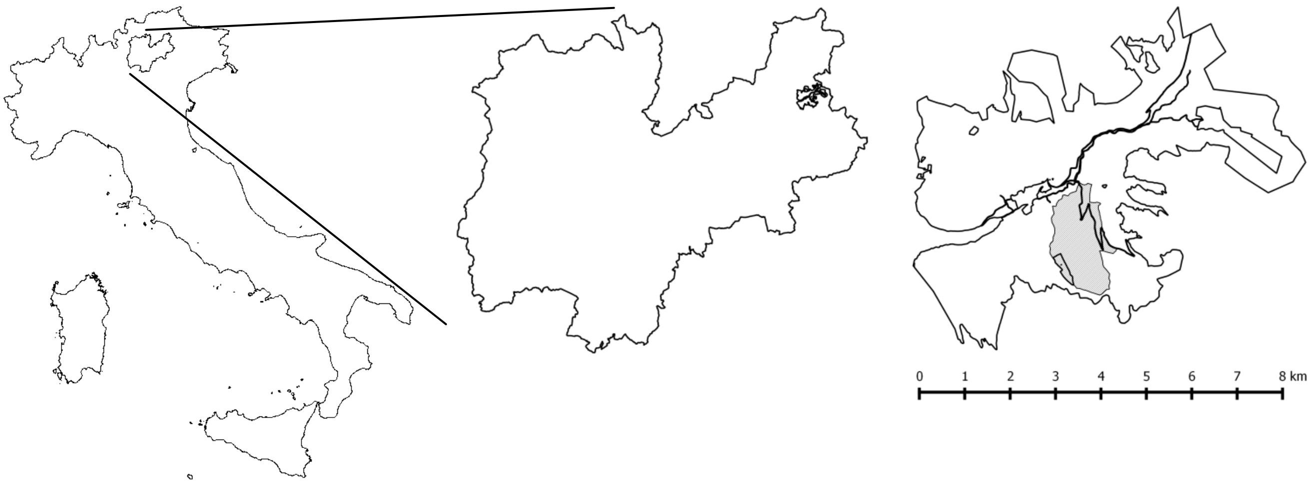

Figure 2.1 - The location of study area with respect to Paneveggio forest (right side), Trento province (center) and to Italy (left side)

Paneveggio forest is a typical sub-alpine Norway Spruce forest, located between 1500 and 2100 m a.s.l.. The dominant specie is Norway spruce (*Picea abies*), 95% of species composition, which grows together with Larch (*Larix decidua*), 3.3%, and Swiss stone pine (*Pinus cembra*), 1.2%. This forest is famous because the wood of Norway spruce growing in that area is a very valuable resonance wood for instrument making.

Approximately 58.1% of forested area is managed forest for commercial purposes, and the rest 41.9% is unmanaged protection forest.

This is a very fertile forest: the average growing stock of the productive Paneveggio Forest is 416 m^3/ha, which is very high compared to the 286 m^3/ha representing the average value of the Norway spruce forests in Italy [De Natale e Pignatti, 2011].

The forest is owned by the Provincia Autonoma di Trento and is managed by the "Agenzia provinciale delle foreste demaniali" of the same Province. The incomes from the sale of timber represent an important budget item for the Trento Province.

2.2.2 Model A development

Dendrometric data to develop the linear Model A for volume estimation, based on airborne laser scanner predictor variables, were collected on 39 circular sample plots randomly located in the study area. The plots had a radius of 13 m (531 m^2 of area) or 20 m (1257 m^2 of area), depending on the growth stage of the stand. They were surveyed in 2007 summer season.

The center of the circular plots was georeferenced with a Thales MobileMapper CE GPS or a Trimble Geo-XT GPS receiver, according to the protocol adopted in the Italian National Forest Inventory [Scrinzi and Floris, 2003]. The field coordinates were post-processed for differential correction and averaged.

In each plot, for each tree its species name was recorded and tree dbh was measured from a threshold of 2.5 cm. The volume of each tree was calculated by means of double-entry stem volume equations used by the Forest Service of Trento for forest management planning purposes [Scrinzi et al, 2010]. Two kinds of volume per hectare were estimated: the inventory volume (V_i, dbh threshold 2.5 cm) and the forest management volume (V_{fm}, dbh threshold 17.5 cm). In the present study, we only assess the performance of the model for the estimation of forest management volume, V_{fm}.

ALS data were acquired in July 2008 over around 400 ha of the study area (1500 m x 2800 m). Discrete return lidar data were collected using an Optech ALTM3033 laser system mounted on an APPLANIX 510 airborne. The lidar system recorded range and intensity of 4 returns per pulse, and achieved a nominal pulses density of 5 per m^2 (Table 2.1).

Table 2.1 - Lidar data 2008 summer acquisition parameters

Sensor	Optech ALTM3033
Acquisition date	July 2008
Flight altitude	1000 m above ground
Flight line sidelap	50%
Maximum off-nadir scan angle	Unknown
Returns/pulse	4
Density	5 pp/m^2
Pulse repetition	70 kHz
Laser wavelength	400-800 nm

The raw data were filtered and classified by the vendor using Terrascan software (Terrasolid Ltd.) which utilizes the Axelsson [1999] filtering algorithm based on Triangulated Irregular Network densification. Data delivered by the vendor included a DSM and a DTM as a regular grid with a spatial resolution of 1 m^2. The CHM was calculated as the difference between DSM and DTM. CHM values lower than 2 meters, including negative values, were set to 0.

On CHM, circular plots corresponding in position and size to ground plots were extracted. Five potential predictor variables were calculated on them, by means of map algebra ArcGIS tools:

- mean of all values (h);
- maximum (h_{max});
- standard deviation (std);
- mean of values above 2 m (h_a);
- canopy cover, which is the ratio between the number of pixels above 2 m and the total number of pixels (C).

A multiple regression model, having the general formulation of:

$$V_{fm} = ah + b\,h_{max} + c\,std + d\,h_a + eC \qquad [2.1]$$

was studied using stepwise forward technique. This kind of selection adds one predictor variable at once in the model, checks their p-value and chooses those with p-value less than a critical level of significance (α). No explanatory variables were left in the model with a partial p-value greater than 0.05. The stepwise procedure selected only mean (h) of all height values as explicative variable with a level of significance higher than the critical level.

For the validation of the model, authors used a cross-validation, exactly a repeated random sub-sampling validation. This method randomly splits the dataset into training and validation data. For each such split, the model is fit to the training data, and predictive accuracy is assessed using the validation data. Splitting is repeated many time and the performance measures (*e.g.* determination coefficient, R^2, adjusted determination coefficient, *adjusted* R^2, root mean square error, RMSE, and relative standard error, RSE) are then averaged over the splits. The authors used 60% of observations as training set and the remainder 40% of observations as validation set, and repeated the cross validation 20 times. Statistical analysis was performed using the *General Linear Model* module of STATISTICA 8 software (Statsoft).

The main parameters of the selected model and of the performance measures obtained through the external validation are reported in Table 2.2. The model for forest management volume explained 88.31% of variability. The negative value of intercept of forest management volume function originated from not considering the volume of trees with dbh lower than 17.5 cm while CHM picks up (*i.e.* intercepts) also the height of these trees. External validation confirmed the high performance of the model: the determination coefficient reached the value of 88.4% (versus 88.3%) and the RMSE reached the value of 91.1 m^3/ha (versus 88.3). More details on "Model A" development can be found in Floris *et al.* [2010].

Table 2.2 - Results from the regression analysis and timber volume model validation

Final model		
Variable estimated	V_{fm}	
Type of volume	Forest management volume	
Measurement unit	m^3/ha	
Model	$V_{fm} = a + bh$	
Parameters	*a*	-19.8870
	b	38.68690
Statistical significance (p-value)	*a*	>0.05
	b	<0.01
0.95 confidence interval	*a*	-99.8954-60.12142
	b	33.9990-43.37487
Determination coefficient	R^2	0.883130
Adjusted determination coefficient	*Adjusted* R^2	0.879972
Root mean square error (m^3/ha)	RMSE	88.27
Relative standard error	RSE (%)	14.8
Cross validation (training on 60% of observations, testing on 40% of observations)		
Variable estimated by the model		V_{fm}
Determination coefficient	R^2	0.884
Adjusted determination coefficient	*Adjusted* R^2	0.878
Root mean square error (m^3/ha)	RMSE	91.1
Relative standard error	RSE (%)	15.3

2.2.3 Field data collection for volume computation at forest compartment level

For the purposes of the present study, the forest compartments identified by the numbers 118 and 122 in the forest management plan of Paneveggio forest were chosen to fully validate the timber volume Model A (Figure 2.2). Forest compartments are management units of an area subjected to a forest management plan delineated for operative and harvesting purposes. The choice fell in those specific forest compartments because no plots used to develop Model A were located inside them. Moreover no cuttings were made and no tree smashes occurred between 2008 (year of airborne laser scanner campaign) and 2010 (year of field surveys to collect data for validating the model). The changes in growing stock were attributable only to woody growth.

The first step of survey was the field-validation of forest compartment boundaries that were created by GIS (Geographical Information System) tools from prior maps produced for forest inventory purposes. This operation was made to obtain the exact position of forest compartments and to guarantee accurate values of forest compartment surface. The geographical coordinates of forest compartments border were recorded by a GPS receiver Trimble Pro-XR following the blue lines marked on the ground on trees barks, stones, boulders, artifacts, etc. which correspond to the limit of the forest compartment. Coordinates were afterwards differentially corrected.

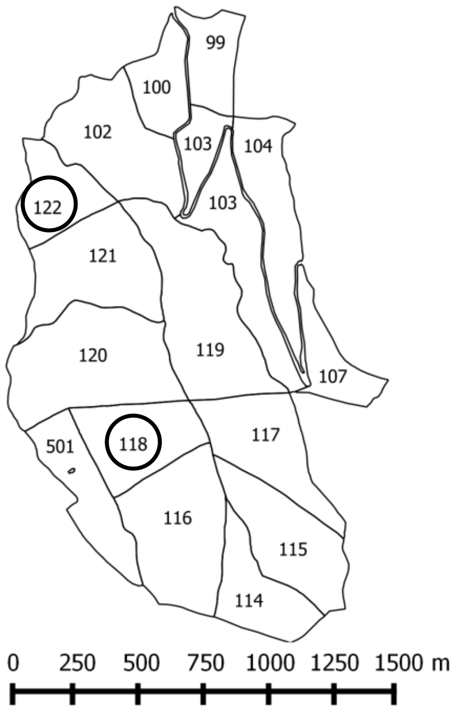

Figure 2.2 - Forest compartments 118 and 122 selected for full validating the timber volume model

During September and October 2010, the forest technicians of Agenzia provinciale delle foreste demaniali measured with calipers all trees with dbh equal or bigger than 17.5 cm in both forest compartments, and recorded tree species.

In each forest compartment the height of 44 trees with dbh equal or bigger than 17.5 cm randomly selected was measured by a Vertex-II ultrasonic hypsometer to build a height-diameter model from which to obtain the height of each tree needed in the double-entry tables to calculate the volume. The height-diameter model was built according to the Curtis [1967] equation:

$$h = 1.3 + \frac{d^2}{a + bd + cd^2} \qquad [2.2]$$

where h is the estimated height of the tree, and d is the dbh.

The volume of each tree was calculated by means of double-entry tree volume equations [Scrinzi et al, 2010] selected according to the tree species, entering in the table with the measured dbh and the height estimated by the height-diameter function.

The inventory volume was computed as the sum of the volume of trees with dbh bigger than 17.5 cm divided by the surface of the forest compartment and multiplied by 10000 to compute the value at hectare.

2.2.4 Application of Model A at forest compartment level

Model A was used to estimate the forest management volume in both forest compartments.

The explicative variable needed in Model A was extracted from the CHM intersecting the real borders of the forest compartments, and then used as input variables in the model to estimate V_{fm} (m^3/ha).

The timber volume of each forest compartment was calculated multiplying V_{fm} by the "true" forest compartment surface (re-measured by GPS). Considering that two years passed between the year of lidar data surveys (2008) with which the Model A was developed and the year of field surveys to collect the dendrometric data (2010) for full validating Model A, the volume estimated by the Model A was adjusted (in the sense updated) adding a value corresponding to 2% of annual increment according to that reported in the Paneveggio forest management plan. Considering the volume calculated by ground surveys (diameter from the full-callipering, tree height from the height-diameter functions, and volume from the application of stem volume double-entry equations) as "ground truth", the relative values estimated by the linear regression model were calculated.

2.3 Results

2.3.1 Field-validation of forest compartments boundaries and height-diameter function for two forest compartments

The field-validation of forest compartments boundaries pointed out the existence of a difference between the spatial location of borders according to the Paneveggio forest management plan and the borders on the ground (Figure 2.3). Moreover field-validation made it clear that the real value of forest compartment surfaces differed from the value recorded in the Paneveggio forest management plan: for forest compartment 118, 12.02 ha versus 12.39 ha, and for forest compartment 122 9.28 ha versus 9.05 ha.

In total, the surveyed forest surface was 21.3 ha and the trees callipered were 7.347 (4102 in the forest compartment 118 and 3245 in the forest compartment 122).

The main statistic parameters of the height-diameter function built for forest compartment 118 and 122 are reported in Table 2.3, while the corresponding graphs about the tree height-diameter relationship are in Figure 2.4.

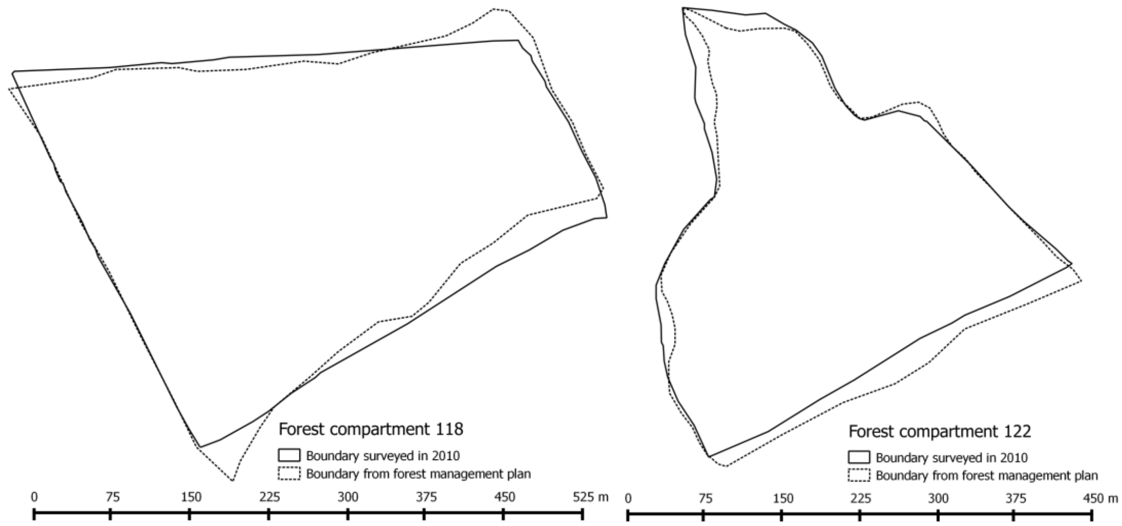

Figure 2.3 - Differences between boundaries surveyed by GPS (dot line) and those recorded in the Paneveggio forest management plan (solid line)

Table 2.3 - Main statistical parameters of the height-diameter model according to the Curtis [1967] equation for forest compartment 118 and 122

Statistic parameters	**Forest compartment 118**	**Forest compartment 122**
Dependent variable	h	h
Measure unit	m	m
Model equation	$h = 1.3 + \frac{d^2}{a + bd + cd^2}$	$h = 1.3 + \frac{d^2}{a + bd + cd^2}$
a	14.973687	5.83341
b	0.150295	0.86600
c	0.026404	0.01233
p-value of a	0.2232	0.5305
p-value of b	0.8032	0.0888
p-value of c	0.0004	0.0450
R^2	0.6874	0.8729
R^2 adjusted	0.6721	0.8667

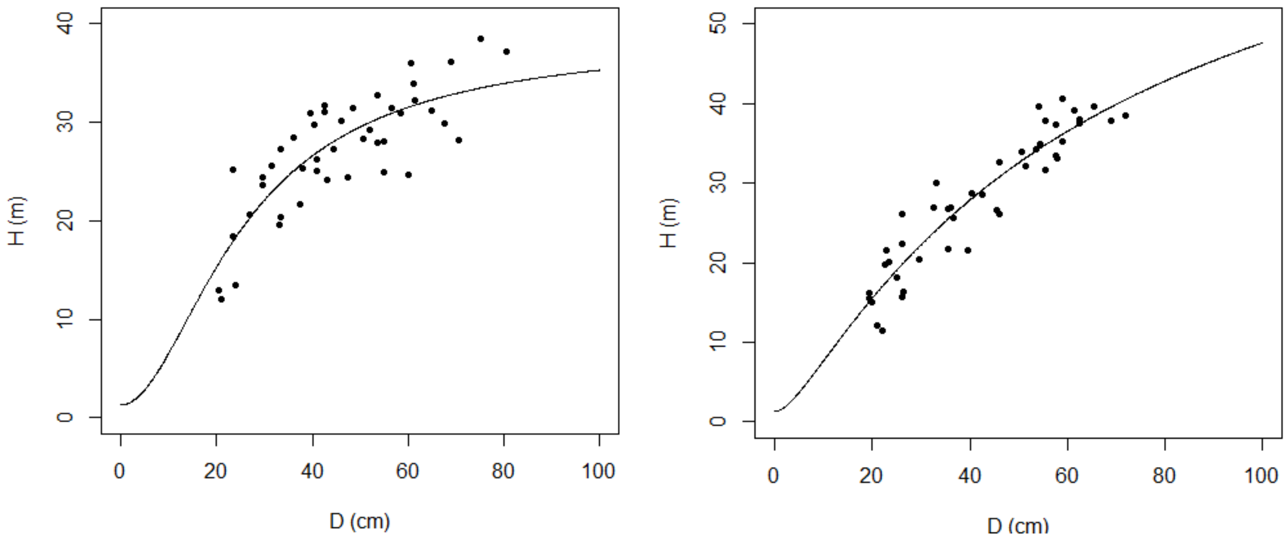

Figure 2.4 - Tree height-diameter relationship for forest compartment 118 (left side) and 122 (right side)

2.3.2 Full validation of Model A

Values of forest management volume for each forest compartment estimated by the application of stem volume double-entry equations and by Model A are in Table 2.4.

Table 2.4 - Comparison between forest management volume values obtained with stem volume double-entry equations and with Model A

Volume estimation approach	Forest compartment 118		Forest compartment 122		Both
	Volume (m^3)	Relative value (%)	Volume (m^3)	Relative value (%)	
Double-entry table (dbh from full callipering and height from height-diameter function)	7379.10	100	5101.34	100	100
Lidar linear Model A ($V_{fm} = -19.8870 + 38.6869h$)	6763.96	-	5142.92	-	-
Lidar linear Model A with two year of increment in growing stocks	7037.23	95.37	5350.69	104.89	99.26

In the case of forest compartment identified by the number 118, the relative value of volume estimated by Model A was 95.37% (-4.63%), while in the case of forest compartment 122 the relative value was 104.89% (+4.89%).

2.4 Discussion

The results obtained in the Floris *et al.* [2010] study have already indicated that the forest management volume model could successfully been applied for sub-alpine Norway Spruce forests in Trento province. The derived coefficient of determination (R^2 = 0.88) was comparable with those derived by Hollaus *et al.* [2007] in Austrian mountainous forests with very similar species composition (R^2 = 0.82) and with those derived from boreal forests (R^2 = 0.83-0.97) reported in Næsset [2004]. Moreover, Hollaus *et al.* [2007] summarized that the RMSE of the differences between estimated and ground reference values (dbh bigger than 10 cm) was 103.9 m^3/ha (radius sample plot size of 13 m), corresponding to a relative error of 24.6%.

The study confirms the high performance of the model for timber volume estimation developed by Floris *et al.* [2010] when applied at forest compartment level. The values of volume estimated in this study referred only to two forest compartments and cannot be considered to rigorously determine the error distribution. But, by the evaluations carried out in the present study, we can reasonably conclude that the performance of Model A, based on CHM-derived metric, is comparable with those obtained from field surveys activities. The big advantage is that the use of models of timber volume estimation is not time consuming as it is a full-callipering. The big advantage is that the use of models of timber volume estimation is not as time consuming as is a full-callipering. Moreover, for the operational application of the model it has to be considered that the variable derived from CHM is very easy to extract using common GIS applications, including those freely available.

Furthermore, the use of Model A can be proposed for estimations at forest compartment level with the added advantage of providing information about the spatial distribution of the growing stock within the area occupied by the forest compartment. In fact, by the spatialization of the results obtained from the model application, thematic maps about timber volume can be produced, which are important tools to support silvicultural interventions and to organize stem cutting activities.

2.5 References

Axelsson, P., 1999. *Processing of laser scanner data - algorithms and applications*. ISPRS J. Photogramm. Remote Sens. 54(2): 138-147. doi: 10.1016/S0924-2716(99)00008-8.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., Torresan, C., 2010. *Stima operativa di parametri dendrometrici forestali con riprese lidar invernali a bassa risoluzione*. Atti 14° Conferenza Nazionale ASITA, Brescia, 9-12 novembre 2010, pp. 625-630.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., Torresan, C., 2012. *Estimating forest timber volume by means of "low-cost" lidar data*. Italian Journal of Remote Sensing - 2012, 44(1): 125-140. doi: 10.5721/ItJRS201244110.

De Natale, F., Pignatti, G., 2011. Estensione e composizione dei boschi italiani. In: Gasparini, P., Tabacchi, G. (a cura di) 2011 - L'Inventario Nazionale delle Foreste e dei serbatoi forestali di Carbonio INFC 2005. Secondo inventario forestale nazionale italiano. Metodi e risultati. Ministero delle Politiche Agricole, Alimentari e Forestali, Corpo Forestale dello Stato; Consiglio per la Ricerca e Sperimentazione in Agricoltura, Unità di Ricerca per la Pianificazione ed il Monitoraggio e la Pianificazione Forestale. Edagricole, Milano, pp. 99-111.

Floris, A., Clementel, F., Farruggia, C., Scrinzi, G., 2009. *Il lidar nella stratificazione tematica dei soprassuoli forestali: applicazioni in Trentino*. Atti 13° Conferenza Nazionale ASITA, Bari, 1-4 dicembre 2009, pp. 1063-1070.

Floris, A., Clementel, F., Farruggia, C., Scrinzi, G., 2009. *Stima su base lidar delle provvigioni legnose forestali: uno studio per la Foresta di Paneveggio*. Italian Journal of Remote Sensing - 2010, 42(3): 15-32.

Hollaus, M., Wagner, W., Maier, B., Schadauer, K., 2007. *Airborne laser scanning of forest stem volume in a mountainous environment*. Sensors, 2007, 7, 1559-1577.

Næsset, E., 2004. *Practical large-scale forest stand inventory using a small-footprint airborne scanning laser*. Scandinavian Journal of Forest Research. 19: 164-179.

Scrinzi, G., 1989. *Precisione degli inventari assestamentali per cavallettamento totale: un'analisi di trent'anni d'esperienza applicativa in Trentino*. Annali dell'Istituto Sper. per l'Assestamento For. e per l'Alpicoltura, Vol. XI, Trento, pp. 143-186.

Scrinzi, G., Floris, A., 2003. *Procedure di individuazione, materializzazione e ritrovamento dei punti campionamento (dell'Inventario Nazionale delle Foreste e dei Serbatoi forestali del Carbonio – INFC, Corpo Forestale dello Stato)*. Isafa, Trento. pp 48.

Scrinzi, G., Galvagni, D., Marzullo, L., 2010. *I nuovi modelli dendometrici per la stima delle masse assestamentali in provincia di Trento*. Provincia Autonoma di Trento, Servizio Foreste e Fauna. pp. 95.

3 Volume estimation in forest compartments by exploiting CHM information: model-based, design-based and hybrid inferences

Forest compartments, which include portions of different strata, are the basic unit of management activities in forest planning context of Trento province, consequently the knowledge about growing stock at that level is very important. Nowadays the Forest Service of Trento province perform the timber volume estimation at stratum level from field plots located within, but the volume estimation at forest compartment may be not accurate. In this chapter model-based, design-based, and hybrid approaches to infer the volume at forest compartment level using CHM-derived metrics as auxiliary information are describe. A simulation study and an application to a public forest estate of the proposed approaches are used to test the volume estimation accuracy of inferences.

3.1 Introduction

Traditional ground-based forest inventory methods are designed to provide point estimates of inventory parameters for relatively large areas to a desired level of precision [Reutebuch *et al.*, 2005]: these relatively large areas can be the forest estate or the strata partitioning the forest estate.

Usually, inventory strata are established as large portions of the estate roughly homogeneous as concerns environmental, vegetation and silvicultural conditions. On the other hand, forest compartments are the management units within the estate delineated for operative and harvesting purposes. Most often forest compartments include portions (even very small) of different strata. Standing wood volume (henceforth referred to as volume) is a key attribute to be estimated for productive forest compartments [Koivuniemi and Korhonen, 2006].

Commonly a generic compartment is composed of portions of inventory strata constituting a larger area which is surveyed by means of field plots of fixed size. In this case, volume estimation within each stratum can be performed from the plots located within the stratum. On the other hand, volume estimation within a forest compartment may be problematic due to the small number (or even the absence) of plots within the portions of strata, called portion (see *e.g.* Figure 3.1).

The scarcity of sampling information within small portions of the survey area is referred to as the small area estimation problem [*e.g.* Rao, 2005]. Usually the lack of sampling information is recovered by the use of auxiliary information available from other sources [Pfefferman, 2002].

Airborne laser scanning technology is increasingly applied in forest surveys [*e.g* Gregoire *et al.* 2011, Næsset, 2011] due to its ability to directly measure the three-dimensional structure (*i.e.*, terrain, vegetation, and infrastructure) of imaged areas and to separate biospatial data (measurements of aboveground vegetation) from geospatial data (measurements of the terrain surface) using active remote sensing technologies [Reutebuch *et al.*, 2005]. The CHM obtained from the ALS data is the expression of the canopy height within pixels of few square meters.

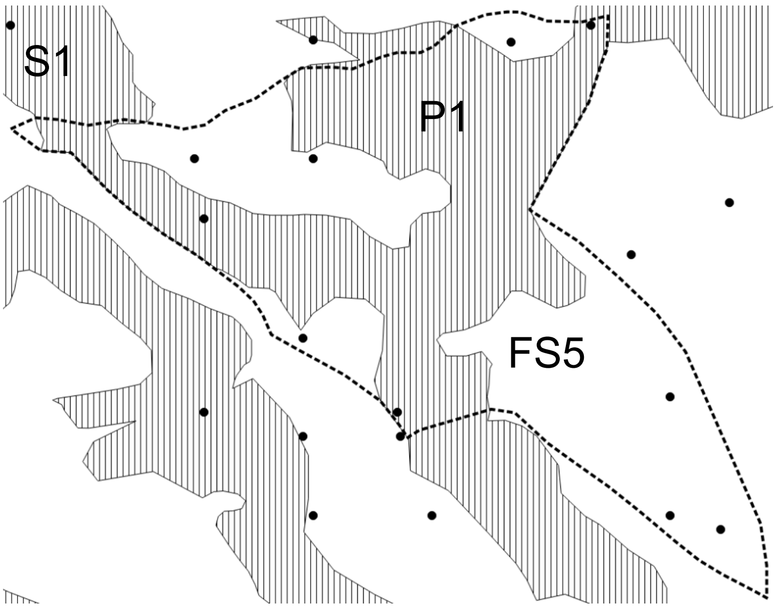

Figure 3.1 - The strata S1 intersecting the forest compartment FS5 (dot line) generates the portion P1 (area with vertical lines) in which just two sample plots fall in

Trento Province committed ALS surveys over all its territory for topographical purposes and hydrogeological investigations and canopy height model data are released to the general public for free: these data represent an effective information source to support forest estimation [*e.g.* Corona and Fattorini, 2008].

We used the height from CHM (henceforth referred to as height) as auxiliary information to perform estimation within forest compartments. For this purpose a model supposing a proportional relationship (ratio model) between height and volume at pixel level was adopted for driving estimation. The use of ratio model with uncorrelated errors and variances proportional to the covariate values is customary in forest surveys [*e.g.* Schreuder *et al.*, 1993, section 6.3.2]. One of the appeals of ratio model is that it preserves the same proportional relation volume/height at any spatial grain, *i.e.* for any agglomeration of pixels. This feature is of relevance for two reasons. First, it is possible to estimate the proportionality factor at plots level, within which volumes are measurable, rather than at pixel level, within which volumes are not. Second, it is possible to obtain a volume estimate within any portion of the survey area multiplying the proportionality factor estimate by the known height within that portion.

This estimation criterion was adopted in this study to estimate the volume of a generic compartment composed of small portions of strata. We suppose that strata are sampled by using field plots of fixed size, randomly allocated within each stratum as is customary in forest inventories [*e.g.* Mandallaz, 2008]. Estimation is considered from three points of view. The model-based perspective, in which the ratio model is viewed as the stochastic mechanism generating the population of pixel volumes from pixel heights, conditioning on the sampling outcome (in this framework, the random positions of plots). The design-based perspective, in which the ratio model is simply used as an assisting device to provide an adequate fitting of the population scatter heights-volumes. In this case, in accordance with the paradigms of the so-called model-assisted inference [Särndal *et al.*, 1992, p. 239], volumes are viewed as fixed characteristics and uncertainty only stems from the random allocation of plots. The hybrid perspective, in which both the uncertainty arising from the model and that arising from the random allocation of plots are considered. The three approaches give rise to the same volume estimator. On the other hand, because variability under different approaches is generated from different sources (*i.e.* the model, the sampling scheme and both), the variances under the three approaches have different formulations which, in turn, give rise to different variance estimators and confidence intervals.

Within the main goal of this dissertation concerning the development of statistical techniques for forest volume estimation at forest compartment level using lidar data as auxiliary information for the local calibration of forest inventory, we wanted to check the performance of the volume estimator and to compare the performance of the three approaches in estimating the accuracy. Irrespective of the approach adopted, in real world there is no model which generates volumes from heights. Rather, these values and any population or sub-population totals are fixed (even if unknown) characteristics of the survey area. Accordingly, the volume estimator is evaluated by a simulation study in a complete design-based setting, *i.e.* on the basis of its ability to guess the real quantity under estimation. In the same way, even if arising from different philosophical point of views, the variance estimators of the different approaches are evaluated on the basis of their ability to guess the actual variability of the volume estimator, *i.e.* the variability arising from the random selection of plots. Similarly, confidence intervals are evaluated on the basis of the proportion of times they capture the real quantity under estimation. As pointed out by Särndal *et al.* [1992, p. 21], the design-based properties of an estimator are real, not modeled or assumed.

The study presented here is focused on the variability entailed by sampling, by model or by both of them.

The study deliberately avoids consideration of other sources of uncertainty, such as measuring aspects [*e.g.* Schereuder *et al.*, 1993, chapter 7] and missing data, which occur when some plots cannot be reached by forest technicians [*e.g.* Fattorini *et al.*, 2013].

3.2 Forest volume estimation strategies

3.2.1 Preliminary notations

Consider a study region partitioned into K strata and suppose that a number of fixed-size plots is allocated at random within each stratum. Let us assume to be interested in the estimating the total volume within a forest compartment composed of the portions of the K strata constituting the study area as in Figure 3.2.

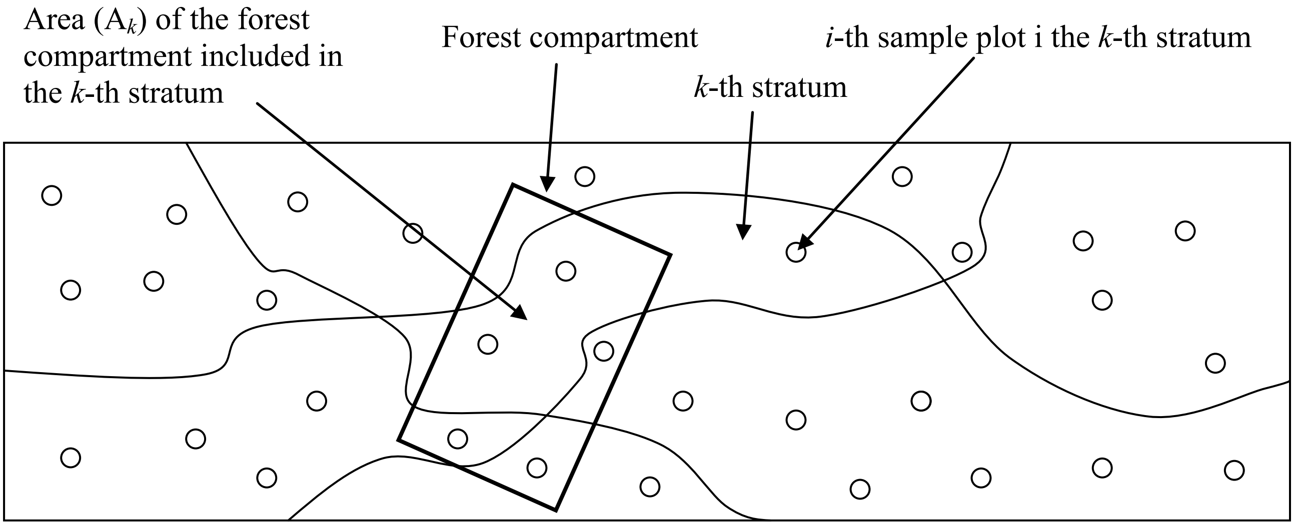

Figure 3.2 - Exemplification of a rectangular forest compartment within a forest estate partitioned into strata

Following stratification, the survey proceeds as a series of independent surveys performed within each stratum, in such a way that the overall estimate and the overall variance estimate for the entire forest compartment can be obtained as the sum of the independent within-strata estimates. Thus, only the estimation within a stratum will be considered henceforth. Accordingly, denote by U the population of pixels partitioning a stratum, by y_j the volume within pixel $j \in \mathsf{U}$, by h_j the height in the same pixel. Denote by

$$Y = \sum_{j \in \mathsf{U}} y_j \qquad [3.1]$$

and

$$X = \sum_{j \in \mathsf{U}} x_j \qquad [3.2]$$

the corresponding totals in the stratum. While the x_j s are directly available from CHM, the y_j s are only virtual, because the volume within a pixel is not practically quantifiable. Finally, denote by $\mathsf{A} \subset \mathsf{U}$ the set of pixels composing the portion of the forest compartment included in the stratum and by

$$Y_{\mathsf{A}} = \sum_{j \in \mathsf{A}} y_j \qquad [3.3]$$

and

$$X_{\mathsf{A}} = \sum_{j \in \mathsf{A}} x_j \qquad [3.4]$$

the total volume in A. While X_{A} is known, Y_{A} is an unknown quantity and constitutes the object parameter under estimation.

As already pointed out, volumes cannot be recorded at pixel level but only for a sufficiently larger spatial unit by summing the volumes of the trees lying within that unit. Because *n* plots are randomly and independently located onto the stratum, the volumes within plots are available. Because each plot *i* comprises a set of pixels, say $\mathsf{P}_i \subset \mathsf{U}$, the total height and the total volume observed within the plot are denoted by x_{P_i} and y_{P_i}, respectively (*i = 1, ... , n*). Inference on Y_{A} will be based on these values.

3.2.2 The ratio model and the volume estimator

Assume that, in accordance with some dendrometric considerations such as the so called Eichhorn's rule [*e.g.* Skovsgaard and Vanclay, 2008], conditioned on the x_j s, the y_j s are random variables generated by the ratio model

$$y_j = \beta x_j + \varepsilon_j , \; j \in \mathsf{U} \qquad [3.5]$$

where ε_j represents the error component with $\mathrm{E}_m(\varepsilon_j \mid x_j) = 0$, variance $\mathrm{V}_m(\varepsilon_j \mid x_j) = \sigma^2 x_j$ and covariances $\mathrm{E}_m(\varepsilon_j, \varepsilon_h \mid x_j, x_h) = 0$ for any $j \neq h \in \mathsf{U}$, and the subfix m denotes expectation with respect to the model. Consider any set $\mathsf{P} \subset \mathsf{U}$ of (not necessarily) adjacent pixels in the stratum and denote by y_{P} the total volume within this set. From [3.5] it follows that

$$y_{\mathsf{P}} = \sum_{j \in \mathsf{P}} y_j = \sum_{j \in \mathsf{P}} (\beta x_j + \varepsilon_j) = \beta \sum_{j \in \mathsf{P}} x_j + \sum_{j \in \mathsf{P}} \varepsilon_j = \beta x_{\mathsf{P}} + \varepsilon_{\mathsf{P}} \qquad [3.6]$$

where x_{P} denotes the total height within P, while ε_{P} denotes the total error in P. In practice, if model [3.5] is true, the model relating volume to height at a greater spatial grain P has the same structure of [3.5]. Indeed, ε_{P} is an error component with $\mathrm{E}_m(\varepsilon_{\mathsf{P}} \mid x_{\mathsf{P}}) = 0$, variance $\mathrm{V}_m(\varepsilon_{\mathsf{P}} \mid x_{\mathsf{P}}) = \sigma^2 x_{\mathsf{P}}$ and covariance $\mathrm{E}_m(\varepsilon_{\mathsf{P}}, \varepsilon_{\mathsf{Q}} \mid x_{\mathsf{P}}, x_{\mathsf{Q}}) = 0$, providing that P and Q are disjoint sets of pixels. Moreover, owing to the central limit theorem, the error term ε_{P}, being the sum of independent and identically distributed random variables, turns out to be normally distributed for a sufficiently large P.

Because the ratio model [3.5] when true entails relation [3.6] for any spatial unit larger than a single pixel, from [3.6] it follows that

$$Y_{\mathsf{A}} = \beta X_{\mathsf{A}} + \varepsilon_{\mathsf{A}} \qquad [3.7]$$

where $\mathrm{E}_m(\varepsilon_{\mathsf{A}} \mid X_{\mathsf{A}}) = 0$ and $\mathrm{V}_m(\varepsilon_{\mathsf{A}} \mid X_{\mathsf{A}}) = \sigma^2 X_{\mathsf{A}}$. In other words, under [3.5] the object quantity Y_{A} is an asymptotically normal random variable with expectation $\mathrm{E}_m(Y_{\mathsf{A}} \mid X_{\mathsf{A}}) = \beta X_{\mathsf{A}}$ and variance $\mathrm{V}_m(Y_{\mathsf{A}} \mid X_{\mathsf{A}}) = \sigma^2 X_{\mathsf{A}}$. Because X_{A} is known, the model-unbiased prediction of Y_{A} reduces to the model-unbiased estimation of β.

If the *n* plots were purposively selected from the stratum, the values $x_{\mathsf{P}_1}, \ldots, x_{\mathsf{P}n}$ are *n* constants. Thus, owing to [3.6], $y_{\mathsf{P}_1}, \ldots, y_{\mathsf{P}_n}$ are *n* normal random variables with mean $\mathrm{E}_m(y_{\mathsf{P}_i} \mid x_{\mathsf{P}_i}) = \beta x_{\mathsf{P}_i}$ and variance $\mathrm{V}_m(y_{\mathsf{P}_i} \mid x_{\mathsf{P}_i}) = \sigma^2 x_{\mathsf{P}_i}$. Moreover, if the *n* plots constituted disjoint sets of pixels, *i.e.* if $\mathsf{P}_i \cap \mathsf{P}_h = \varnothing$ for each $h > i = 1, \ldots, n$, the y_{P_i} s are independent. Under these conditions, inference on the model parameters β and σ^2 can be performed from the likelihood function

$$L(y_{\mathsf{P}_1}, \ldots, y_{\mathsf{P}n} \mid x_{\mathsf{P}_1}, \ldots, x_{\mathsf{P}n}, \beta, \sigma^2) = const \times \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_{\mathsf{P}_i} - \beta x_{\mathsf{P}_i})^2}{x_{\mathsf{P}_i}} \right\} \qquad [3.8]$$

Actually, the *n* plots are placed onto the stratum in accordance with a probabilistic sampling scheme. Hence the x_{P_i} s are random variables. As argued by Rubin [1976], in these cases estimation should be performed from the likelihood of the y_{P_i} s conditional to the observed values of the x_{P_i} s. The procedure is likely to be analytically difficult in most situations. Fortunately, in this case the allocation of plots is performed completely at random. Hence the sampled sites do not depend on the y_{P_i} s. Thus, sampling is ignorable, *i.e.* the procedure claimed by Rubin [1976] gives rise to [3.8] as if the x_{P_i} s were fixed.

From [3.8], the maximum likelihood estimator of β turns out to be

$$\hat{b}_n = \frac{\bar{y}_n}{\bar{x}_n} \qquad [3.9]$$

where $\bar{x}_n$ and $\bar{y}_n$ denote the arithmetic means of the x_{P_i} s and y_{P_i} s, respectively. Under a correct model specification, [3.9] is BLUE (Best Linear Unbiased Estimator) and is a normally distributed, model-unbiased estimator of β, with variance $\mathrm{V}_m(\hat{b}_n) = \sigma^2 / (n\bar{x}_n)$. Accordingly, from relation [3.5] an obvious model-dependent predictor of the random variable Y_{A} is given by

$$\hat{Y}_{\mathsf{A}} = \hat{b}_n X_{\mathsf{A}} \qquad [3.10]$$

3.2.3 The model based inference

In the model-based inference the properties of regression estimators are subjected to common modeling assumptions in which we suppose the data arise from a super population model, and we attempt to make inference about the super-population model parameters [Gregoire, 1998].

On the basis of [3.5], $\hat{Y}_{\mathsf{A}}$ constitutes a normal, model-unbiased predictor of Y_{A}, in the sense that $\mathrm{E}_m(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = 0$. The model-dependent mean squared prediction error is

$$\mathrm{V}_m(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = \mathrm{E}_m\left\{(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}})^2\right\} = \sigma^2\left(\frac{X_{\mathsf{A}}^2}{n\bar{x}_n} - 2\frac{X_{\mathsf{A}}}{n\bar{x}_n}\sum_{i=1}^{n}\sum_{j\in \mathsf{P}_i\cap \mathsf{A}} x_j + X_{\mathsf{A}}\right) \quad [3.11]$$

Moreover, because the unbiased estimator of σ^2 is given by

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(y_{\mathsf{P}_i} - \hat{b}_n x_{\mathsf{P}_i})^2}{x_{\mathsf{P}_i}} \quad [3.12]$$

a model-unbiased estimator of [3.11] is given by

$$\hat{V}_m = \hat{\sigma}_c^2\left(\frac{X_{\mathsf{A}}^2}{n\bar{x}_n} - 2\frac{X_{\mathsf{A}}}{n\bar{x}_n}\sum_{i=1}^{n}\sum_{j\in \mathsf{P}_i\cap \mathsf{A}} x_j + X_{\mathsf{A}}\right) \quad [3.13]$$

Finally, the interval $\hat{Y}_{\mathsf{A}} \pm 2\hat{V}_m^{1/2}$ has an asymptotic $(n \to \infty)$ probability of containing Y_{A} approximately equal to 0.95. Expression [3.13] highlights the critical dependence of the estimation on a correct formulation of the error terms.

Because the estimators in [3.9], [3.10] and [3.13] are derived from [3.8], they hold if the n plots are disjoint sets of pixels. Because plot centers are randomly placed onto the stratum, it may happen that some plots intersect each other. In this case, an amount of covariance is entailed among the y_{P_i} s and [3.8] is not the actual likelihood. While this fact has no effect on the unbiasedness of [3.9] and [3.10], it precludes the validity of [3.11] and the unbiasedness of the corresponding variance estimator $\hat{V}_m$. However, since plots are usually much smaller than a stratum, plot intersections are rare and should not adversely affect the validity of $\hat{V}_m$.

3.2.4 The design-based inference

Design-based inference refers to inference about the estimation of parameters such as the mean or the total of a realized finite population [Strunk, 2012]. In our case there exists the population of the x_j s and the population of the y_j s ($j \in \mathsf{U}$) which determine the population scatter $\{(x_j, y_j), j \in \mathsf{U}\}$. The population values (and the population scatter) are fixed characteristics of the stratum as well as their totals on the whole stratum, X and Y, and the subtotals on a portion of the stratum, such as X_{A} and Y_{A}.

By this perspective, the ratio model [3.5] may just be viewed as an artificial device presumed to provide a suitable fitting of the population scatter from which to built an estimation criterion [Särndal *et al.* 1992, chapter 6].

From [3.5], the weighted least-squares, BLUE estimate of β performed on the population scatter is $b = Y / X$. Thus, adopting bx_j as a convenient proxy for y_j (the BLUE proxy, if we trust in the model), bX_{A} is a convenient proxy for Y_{A}. However, b is an unknown characteristic of the stratum to be estimated from sample data. Hence, the design-based estimation of Y_{A} reduces to the design-based estimation of b.

Owing to the random placement of plots within the stratum, the n pairs $(x_{\mathrm{P}_i}, y_{\mathrm{P}_i})$ constitute independent and identically distributed two-dimensional random variables. Moreover, if edge corrections are performed during the plot survey to ensure an equal inclusion probability, say π, for each tree in the stratum, then the quantities $\hat{\bar{X}}_n = \pi^{-1}\bar{x}_{\mathrm{n}}$ and $\hat{\bar{Y}}_n = \pi^{-1}\bar{y}_{\mathrm{n}}$ are the arithmetic means of the Horvitz-Thompson estimates of X and Y achieved in each plot. Accordingly, $\hat{\bar{X}}_n$ and $\hat{\bar{Y}}_n$ are design-unbiased, consistent and asymptotically $(n \to \infty)$ normal estimators of the stratum totals X and Y, with design-based variances C_X^2 / n and C_Y^2 / n and covariance C_{XY} / n where the constants C_X^2, C_Y^2 and C_{XY} are mainly determined by the plot size as well as by the spatial distribution of trees throughout the stratum [Gregoire and Valentine 2008, chapter 7]. It is worth noting that edge effects are negligible if the stratum size is much greater than the plot size [Shiver and Borders, 1996, p. 90]. Otherwise, edge effects can be removed by several solutions usually referred to as edge corrections [Gregoire and Valentine 2008, section 7.5].

From these results, which only stem from the random placement of plots onto the stratum, and from the convergence theorem usually referred to as the delta method [*e.g.* Mardia *et al.*, 1979, Theorem 2.9.2], the ratio $\hat{\bar{Y}}_n / \hat{\bar{X}}_n$, which coincides with the estimator $\hat{b}_n$ given in [3.10], constitutes an asymptotically design-unbiased, consistent and asymptotically normal estimator of b. From a design-based point of view, the estimator $\hat{b}_n$, which in a model-based framework estimates the slope β of the presumed super population model [3.5], actually estimates the slope b of the weighted least-square line passing through the origin which fits the population scatter. Accordingly, the estimator $\hat{Y}_{\mathsf{A}}$ given by [3.10] is a design-consistent and asymptotically normal estimator of the quantity bX_{A}. Thus $\hat{Y}_{\mathsf{A}}$ is a design-biased estimator of Y_{A}, with an asymptotic design-based bias

$$\mathrm{E}_d(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = bX_{\mathsf{A}} - Y_{\mathsf{A}} \quad [3.14]$$

where the subfix d denotes expectation with respect to the random placement of plots onto the stratum. Denote by $b_{\mathsf{A}} = Y_{\mathsf{A}} / X_{\mathsf{A}}$ the weighted least-square, BLUE estimate of β performed on the scatter determined by the pixels lying in A, $\{(x_j, y_j), j \in \mathsf{A}\}$. From the least-squares properties, for each $j \in \mathsf{A}$, y_j can be decomposed as $y_j = b_{\mathsf{A}} x_j + e_{\mathsf{A}j}$ where the $e_{\mathsf{A}j}$ s are the zero sum regression residuals. Accordingly, expression [3.14] can be rewritten as

$$\mathrm{E}_d(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = \sum_{j \in \mathsf{A}} (bx_j - y_j) = \sum_{j \in \mathsf{A}} (bx_j - b_{\mathsf{A}} x_j - e_j) = (b - b_{\mathsf{A}}) X_{\mathsf{A}} \quad [3.15]$$

From [3.15], $\hat{Y}_{\mathsf{A}}$ is an asymptotic design-unbiased and consistent estimator of Y_{A} only if b and b_{A} are equal. Obviously, this situation will never occur in practice. Thus, $\hat{Y}_{\mathsf{A}}$ will always be biased. However, when the scatter from the entire stratum and the scatter from pixels lying in A are similar, both the weighted least-square lines passing through the origin have similar slopes ($b \approx b_{\mathsf{A}}$) and the bias will be negligible. Similar point scatters are possible when the portion of A in the stratum does not greatly differ from the whole stratum. It is worth noting that there is no way to check these conditions at pixel level, because volumes are recorded at plot levels. However, the proportional relationship, if present at pixel level, is likely to be reproduced at plot level. Thus, care must be taken in using the estimator [3.10] when the scatter obtained from the (few) plots fallen in A differs relevantly from the scatter obtained from the plots fallen in the remaining part of the stratum.

If there is no motivation to suspect large bias, variance estimation can be performed in order to evaluate the design-based variance of [3.10]. From the delta method [*e.g.* Mardia *et al.* 1979, Theorem 2.9.2], the asymptotic design-based variance of $\hat{Y}_{\mathsf{A}}$ turns out to be

$$\mathrm{V}_d(\hat{Y}_{\mathsf{A}}) = \frac{X_{\mathsf{A}}^2}{n} \left(\frac{C_Y^2}{X^2} - 2Y \frac{C_{XY}}{X^3} + Y^2 \frac{C_X^2}{X^4} \right) \quad [3.16]$$

Accordingly, the estimation of [3.16] can be performed by

$$\hat{V}_d = \frac{X_{\mathsf{A}}^2}{n} \left(\frac{s_Y^2}{\bar{x}_{\mathsf{P}}^2} - 2\bar{y}_{\mathsf{P}} \frac{s_{XY}}{\bar{x}_{\mathsf{P}}^3} + \bar{y}_{\mathsf{P}}^2 \frac{s_X^2}{\bar{x}_{\mathsf{P}}^4} \right) \quad [3.17]$$

where s_X^2 and s_Y^2 denote the variances of the x_{P_i} s and y_{P_i} s respectively, and s_{XY} denotes their covariance. Alternatively, stated that the pairs $(x_{\mathsf{P}_i}, y_{\mathsf{P}_i})$ are independent realizations of the same bivariate random vector, we are legitimate to adopt the jackknife variance estimator

$$\hat{V}_{d(jack)} = \frac{X_{\mathsf{A}}^2}{n(n-1)} \sum_{i=1}^{n} (\hat{b}_{n-1}^{(i)} - \hat{b}_{n(jack)})^2 \qquad [3.18]$$

where $\hat{b}_{n-1}^{(i)}$ represents the i-th pseudo-value and $\hat{b}_{n(jack)}$ represents the arithmetic mean of the pseudo-values. Both [3.17] and [3.18] are consistent estimators of [3.16]. Moreover, as pointed out by Shao and Tu [1995, Chapter 2], the two estimators are asymptotically equivalent. Both the variance estimators give rise to confidence intervals, say $\hat{Y}_{\mathsf{A}} \pm 2\hat{V}_d^{1/2}$ or $\hat{Y}_{\mathsf{A}} \pm 2\hat{V}_{d(jack)}^{1/2}$, with asymptotic coverage of about 95%. In presence of large bias, the estimation of variance is of little importance. If an estimator is highly biased, it is a poor consolation to achieve a small variance.

While bias does not depends on the validity of ratio model, the validity of the model may impacts on the accuracy of the strategy. Indeed, it can be proved that $\hat{Y}_{\mathsf{A}}$ guesses Y_{A} without error if a perfect proportional relationship occurs. Thus, $\hat{Y}_{\mathsf{A}}$ is likely to be accurate under close proportional relationships.

3.2.5 The hybrid inference

Suppose that the ratio model [3.5] generates the population values and consider, at the same time, that uncertainty also stems from the random selection of plots onto the stratum. Under the ratio model, the two sets of values $\{y_j, j \in \mathsf{U}\}$ and $\{y_j, j \in \mathsf{A}\}$ can be viewed as two random samples from [3.5], in such a way that $\mathrm{E}_m(b) = \mathrm{E}_m(b_{\mathsf{A}}) = \beta$. Hence, from [3.15] the asymptotic $(n \to \infty)$ bias of [3.10] turns out to be

$$\mathrm{E}_{md}(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = \mathrm{E}_m\{\mathrm{E}_d(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}})\} = \mathrm{E}_m\{(b - b_{\mathsf{A}})X_{\mathsf{A}}\} = 0 \qquad [3.19]$$

where the subfix md denotes expectation with respect to both model and design.

Shao and Steel [1999] provide a general expression for the mean squared prediction error which holds under any hybrid approach. Using this formulation, the mean-squared prediction error of $\hat{Y}_{\mathsf{A}}$ can be written as

$$\mathrm{V}_{md}(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = \mathrm{E}_m\{\mathrm{V}_d(\hat{Y}_{\mathsf{A}})\} + \mathrm{V}_m\{\mathrm{E}_d(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}})\} \qquad [3.20]$$

Thus, inserting the asymptotic design-based bias [3.15] and variance [11] of $\hat{Y}_{\mathsf{A}}$ in the previous expression, the asymptotic mean-squared prediction error turns out to be

$$\mathrm{V}_{md}(\hat{Y}_{\mathsf{A}} - Y_{\mathsf{A}}) = \mathrm{E}_m\left\{\frac{X_{\mathsf{A}}^2}{n}\left(\frac{C_Y^2}{X^2} - 2Y\frac{C_{XY}}{X^3} + Y^2\frac{C_X^2}{X^4}\right)\right\} + \mathrm{V}_m\{(b - b_{\mathsf{A}})X_{\mathsf{A}}\} \qquad [3.21]$$

From [3.5] the second term of [3.21] gives rise to

$$\mathrm{V}_m\{(b - b_{\mathsf{A}})X_{\mathsf{A}}\} = X_{\mathsf{A}}^2\mathrm{V}_m(b - b_{\mathsf{A}}) = \sigma^2(1 - f_{\mathsf{A}})X_{\mathsf{A}} \qquad [3.22]$$

where $f_{\mathsf{A}} = X_{\mathsf{A}} / X$. While the first term of [3.21] can be estimated by one of the design-based criteria of type [3.17] and [3.18], an estimator for [3.22] is obtained by replacing σ^2 with its model-unbiased estimator $\hat{\sigma}_c^2$. Accordingly, an unbiased variance estimator under the hybrid approach is given by

$$\hat{V}_{md} = \hat{V}_d + \hat{\sigma}_c^2(1 - f_{\mathsf{A}})X_{\mathsf{A}} \qquad [3.23]$$

Practically speaking, in the hybrid approach the estimation of the mean-squared prediction error necessitates a further component to be added to the design-based variance estimate. This component takes into account the bias entailed by the unavoidable difference between b and b_{A}. In this sense, the hybrid approach would provide a more cautionary evaluation of the accuracy than the design-based one.

3.3 The simulation study

3.3.1 Study site and field data collection

Because the estimation of volume within a forest compartment proceeds by a series of independent surveys performed within each stratum, the performance of the previous strategies was checked considering a sole stratum. To this purpose, a real study area of size about 4.8 ha settled in Val di Sella (North-Eastern Italy, Southern Alpine area) at an average elevation of 950 m a.s.l. was considered as the stratum of interest.

Even if small, the stratum was characterized by a relatively high variation in forest characteristics. The forest was composed mainly by silver fir (*Abies alba*) and beech (*Fagus sylvatica*), together with Norway spruce (*Picea abies*), larch (*Larix decidua*) and other minor species of conifers and hardwoods.

In summer 2008, the site was completely surveyed on the ground, recording the geographical location and the diameter at breast height of each tree having dbh greater than 17.5 cm. The volume of a single tree was reckoned by means of the local double-entry volume tables, currently used in the forest management practice in Trento province. For the purposes of the simulation study, three patches (each corresponding to about 10% of the whole stratum) were artificially delineated on the edge of the stratum and referred to as A1, A2 and A3 (Figure 3.3).

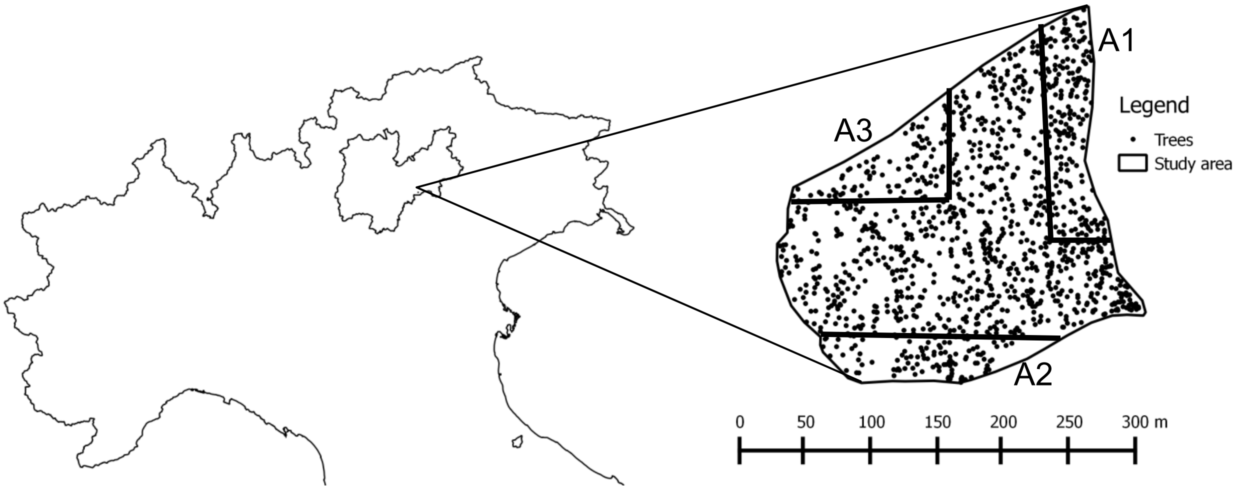

Figure 3.3 - Location of study area in Trento province and the detail about the stratum considered in the simulated case study

These patches represented the portions of as many forest compartments in the stratum and their volumes, say Y_1, Y_2 and Y_3, were the quantities under estimation.

The volume in the stratum was Y = 2625.7 m^3 while the volumes in the three portions were $Y_1 = 265.1$ m^3, $Y_2 = 218.6$ m^3 and $Y_3 = 251.1$ m^3.

The slope of the weighted least-squares line in the stratum was b = 0.00246 while the slopes in the three portions were b_1 = 0.00249, b_2 = 0.00237 and b_3 = 0.00213.

3.3.2 Lidar data

In July 2008 the Forest Service of Trento Province commissioned Blom CGR S.p.A. acquiring ALS data in Val di Sella. Discrete return lidar data were collected using an Optech ALTM3033 laser system mounted in APPLANIX 510. The lidar system recorded range and intensity of 4 returns per pulse, and achieved a nominal pulses density of 5 per m^2 (Table 3.1).

Table 3.1 - Lidar data 2008 summer acquisition parameters

Sensor	Optech ALTM3033
Acquisition date	July 2008
Flight altitude	1000 m above ground
Flight line sidelap	50%
Maximum off-nadir scan angle	Not known
Returns/pulse	4
Density	5 pulses m^{-2}
Pulse repetition	70 kHz
Laser wavelength	Not known

The raw data were filtered and classified by the vendor using Terrascan software (Terrasolid Ltd.) which utilizes the Axelsson [1999] filtering algorithm based on Triangulated Irregular Network densification. Data delivered by the vendor included a DSM and a DTM as a regular grid with a spatial resolution of 1 m^2. We calculated the CHM as the DSM minus the DTM by means of spatial analyst tools in ArcGis 9.x software (ESRI). Canopy height values greater than 2 m were assumed to be vegetation hits, this means that values less than 2 m were set to 0.

The CHM data were released as open data for general public. The height in the whole stratum was X = 1068565.5 m, while the heights in the three portions, say X_1, X_2 and X_3, were X_1 = 106369.5 m, X_2 = 92042.3 m and X_3 = 118042.5 m.

3.3.3 Simulation procedure

For each portion, the simulation was based on 10000 runs. At each run, n = 25, 50, 100 plots of radius 10 m were randomly and independently selected within the stratum. The plot size was that customarily adopted in most Italian inventories. Neglecting overlapping plots, the fraction of surveyed area corresponding to n = 25, 50, 100 was 0.16, 0.32 and 0.64 respectively. The sum of heights and volumes within each plot were determined and the estimate of the total volume within the portion was computed from [3.10], together with the estimates of the model-based, design-based and hybrid variances by means of the estimators [3.13], [3.17] and [3.18], and [3.23], respectively. Both the estimators [3.17] and [3.18] were used in [3.23]. For each variance estimate, the estimate of relative standard error (RSEE) was obtained by dividing the square root of the variance estimate by the volume estimate. For each variance estimate, the 0.95 confidence interval (95CI) was obtained by summing and subtracting two times the square root of the variance estimate to the volume estimate.

Because the plot size (about 314 m^2) was quite small compared with the stratum size, no edge correction was performed. Thus, a certain amount of bias is attributable to edge effects. Because the simulation was performed on a fixed population, with plots which were newly randomly allocated at each simulation run, simulation results referred to the design-based properties of the model-based, design-based and hybrid estimators.

From the resulting Monte Carlo distributions, the relative bias (RB) and the relative root mean squared error (RRMSE) of the estimator [3.10] were empirically evaluated for each portion, together with the expectations of the four estimators of the RSE (ERSEE) and the coverage of the corresponding 0.95 confidence intervals (C95), *i.e.* the percentage of times the intervals included the true volume under estimation.

The description of simulation procedure is illustrated in Figure 3.4.

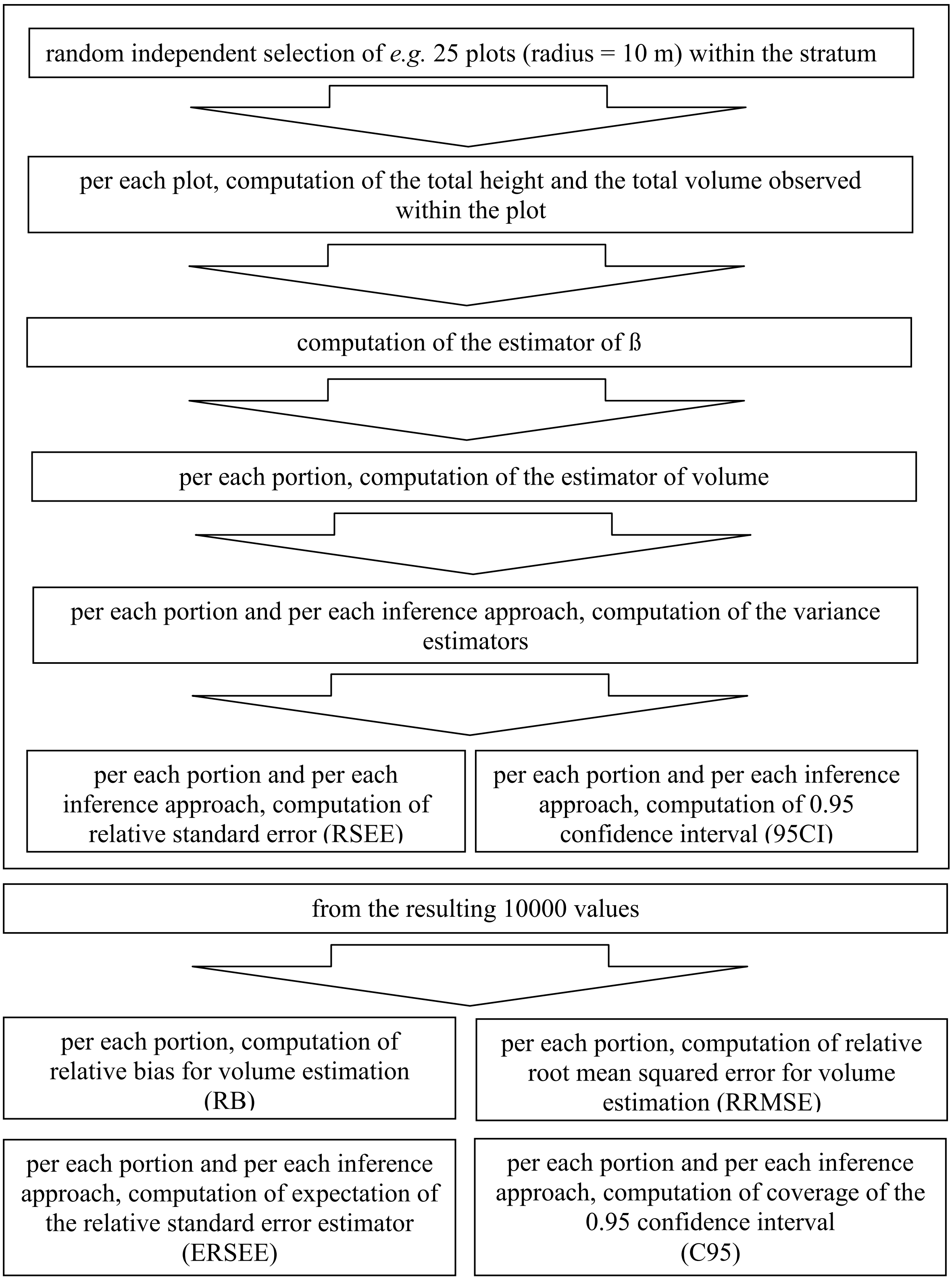

Figure 3.4 - Scheme of the simulation procedure (case of random independent selection of 25 plots)

3.3.4 Results from the simulation study and discussion

The results from the simulated study are reported in Table 3.2 each portion, the table reports the percentage values of RB, RRMSE, ERSEE and C95. The slope of ß of the weighted least-square line fitting the population scatter is 0.00246.

Table 3.2 - Results from the simulated study: percentage value of relative bias (RB), and relative root mean squared error (RRMSE) for volume estimation within the forest compartment portions A1, A2 and A3. Expectation of the relative standard error estimator (ERSEE), and coverage of the 0.95 confidence intervals (C95). ERSEE and C95 are achieved under model-based, design-based, and hybrid approaches

				Model-based		**Design-based (delta-method)**		**Design-based (jackknife-method)**		**Hybrid (delta method)**		**Hybrid (jackknife-method**	
Compartment portion	**N° plots per stratum**	**RB**	**RRMSE**	**ERSEE**	**C95**	**ERSEE**	**C95**	**ERSEE**	**C95**	**ERSE E**	**C95**	**ERSEE**	**C95**
A1 (b1 = 0.00249)	25	-0.2	7.4	10.9	99.2	7.5	94.1	7.3	93.6	11.4	99.4	11.3	99.4
	50	0.1	5.2	9.7	99.9	5.2	94.9	5.2	94.6	10.1	100	10.1	100
	100	0	3.7	9.0	100	3.7	94.9	3.7	94.8	9.4	100	9.4	100
A2 (b2 = 0.00237)	25	4.7	8.6	11.5	98.7	7.5	89.9	7.3	89.3	12.0	99.2	11.9	99.1
	50	5	7.0	10.3	99.8	5.2	85.3	5.2	84.9	10.7	99.9	10.7	99.9
	100	4.9	5.9	9.7	100	3.7	75.6	3.7	75.3	10.1	100	10.0	100
A3 (b3 = 0.00213)	25	16.9	16.2	9.5	69.5	7.5	49.6	7.3	48.3	11.0	81.5	11.0	81.0
	50	17.2	15.6	8.0	56.4	5.2	19.0	5.2	18.6	9.7	77.8	9.7	77.5
	100	17.1	15.1	7.2	42.4	3.7	2.0	3.7	2.0	8.9	77.6	8.9	77.5

The analysis of Table 3.2 leads to some considerations. Bias is nearly negligible in A1, because b_1 and b are similar (with a difference of 1.2%). Bias increases to about 5% in A2, owing to a moderate difference between b_2 and b (3.6%), reaching the level of 17% in A3, owing to the marked difference between b_3 and b (13%). These results confirm what theoretically proven by [3.15]. Bias is primarily due to the difference between b and b_A. Thus, a negligible bias can be achieved even in absence of a proportional relationship. This fact is apparent from Figure 3.5, which shows the scatter of volumes *vs.* sums of heights observed in 1000 plots randomly located on the study area.

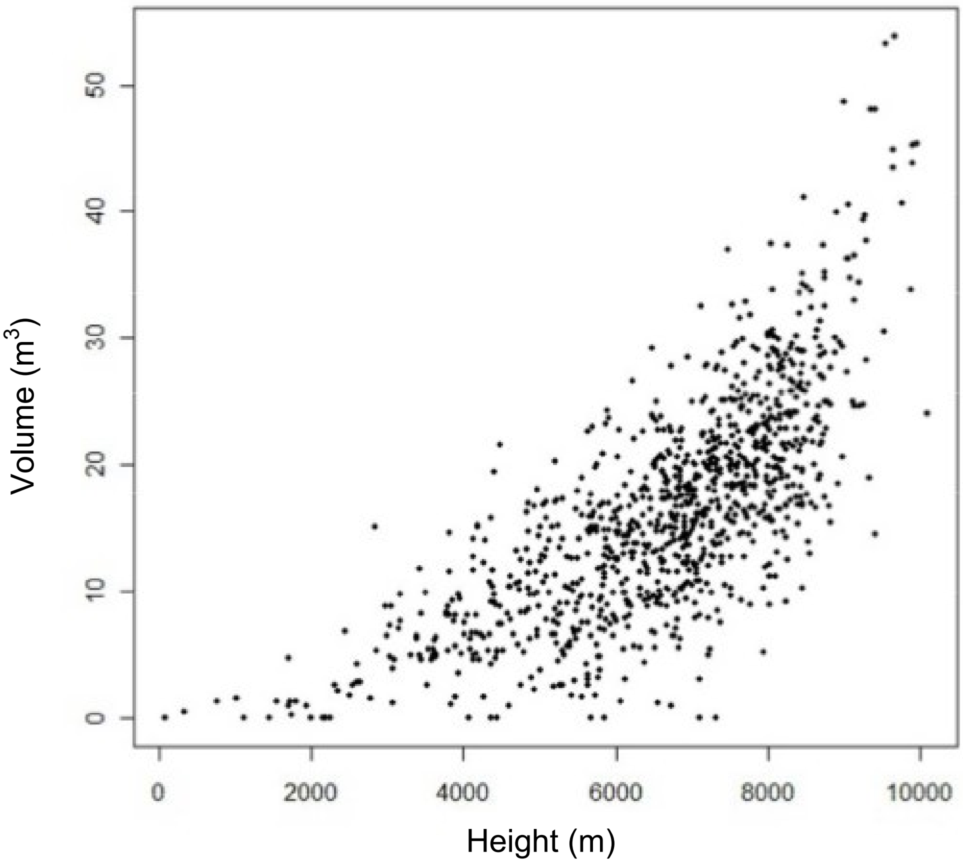

Figure 3.5 - Scatter of volumes vs. sums of heights observed in 1000 plots randomly located onto the study area

Figure 3.5 shows that proportionality fails owing to the presence of a negative intercept. Indeed, volumes are 0 in those plots containing no trees with dbh > 17.5 cm, while a positive height may be extracted owing to the presence of trees with smaller dbh. Notwithstanding this, a small bias is reached in A1.

The best performance is achieved in A1 and A2, with RRMSE values invariably smaller than 9% while in A3 the RRMSE values increases to about 15%, owing to the presence of a massive bias which inflates the mean squared errors.

When the bias is negligible (A1), the design-based estimators of variance (delta method and jackknife) perform well in estimating the RRMSE as well as in providing confidence intervals with coverage near to the nominal level. As theoretically argued, the two variance estimators give similar results, even for moderate values of *n*. On the other hand, the presence of non-negligible bias (A2, A3) inflates the RRMSE over the level due to variance. Thus, the design-based variance estimators underestimate the RRMSE. Moreover, non-negligible bias skews the design-based confidence intervals, providing coverage smaller than 95%. If the bias is moderate (3-5% as in A2) the drawback can be amended by using the model-dependent or the hybrid estimators of variance. Because both estimators turn out to be conservative, they can cover the increase of RRMSE due to the bias and provide confidence intervals with coverage greater than the nominal value. In presence of an heavy bias (17% in A3), even the use of conservative variance estimators does not suffice to take into account the massive increase of RRMSE, providing confidence interval with coverage much smaller than the nominal level.

Practically speaking, in presence of a moderate bias (smaller than 5%), which is likely to occur if the strata are homogeneous, the use of the estimator [3.10], joined with the use of the hybrid estimator of variance, provides a conservative evaluation of accuracy and reliable confidence intervals.

3.4 The case study

3.4.1 Study site and field data collection

The study site about the case study is a public forest estate located in San Martino di Castrozza (North-Eastern Italy, Southern Alpine area). The forest (604.5 ha) included typical sub-alpine *Picea abies* (sporadically mixed with *Larix decidua* and *Pinus cembra*) stands. The compartments selected for this application were those with identification numbers 5 (22.7 ha), 10 (16.6 ha), 24 (15.6 ha) and 27 (28.4 ha).

A field timber inventory (dbh > 17.5 cm) was carried out in August 2009. The forest area was partitioned into 8 homogeneous strata. 314 plots were allocated to strata in proportion to their extents and were randomly selected within (Figure 3.6). Plot size varied with strata in accordance with their characteristics. Details about strata are reported in Table 3.3.

Each forest compartment considered in the case study included portions of several strata (Figure 3.7).

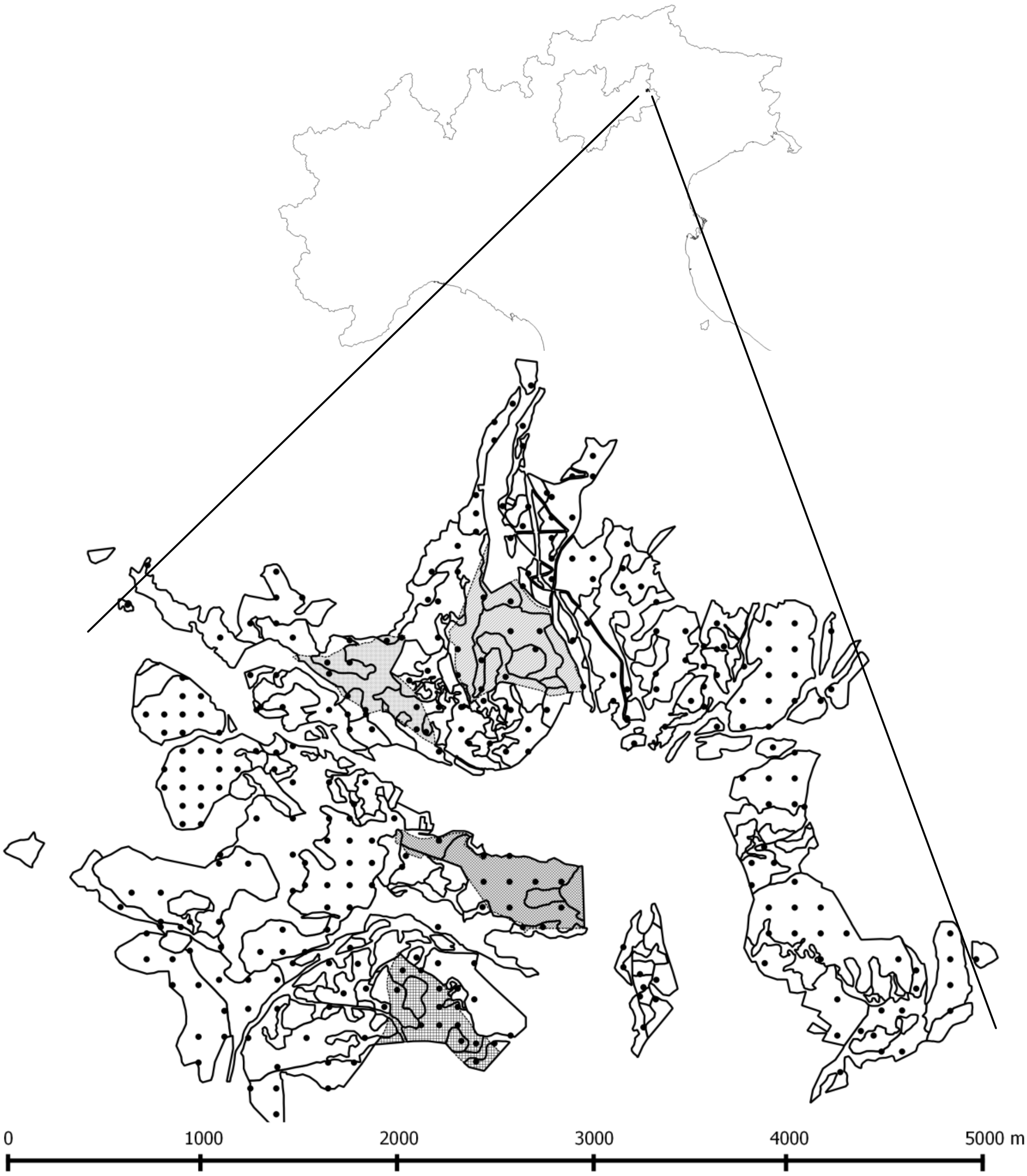

Figure 3.6 - Location of the case study area in Trento Province (North-Eastern Italy) with the detail of the public forest estate in San Martino di Castrozza and four forest compartments selected to apply the strategy

Table 3.3 - Characteristics of the strata partitioning the public forest estate in San Martino di Castrozza

Stratum	Area (ha)	Description	Plot radius (m)	N° plots	Height (m)
1	145.2	Young stands of Norway spruce	9.5	37	15976300
2	66.7	Mature stands of Norway spruce	11.5	46	10211700
3	84.7	Mature stands of Norway spruce and larch	16.5	32	8206140
4	160.4	Mature stands of Norway spruce and larch	15.0	83	19060600
5	56.5	Multi-layered stands of Norway spruce and larch with predominance of large trees	15.5	57	5084050
6	20.5	Two layered stands with a regular structure with large trees coexisting with small trees	19.0	9	1919760
7	38.2	Multi-layered stands of Norway spruce and larch	13.0	26	3068800
8	32.3	Multi-layered stands of Norway spruce and larch	16.5	24	2524310

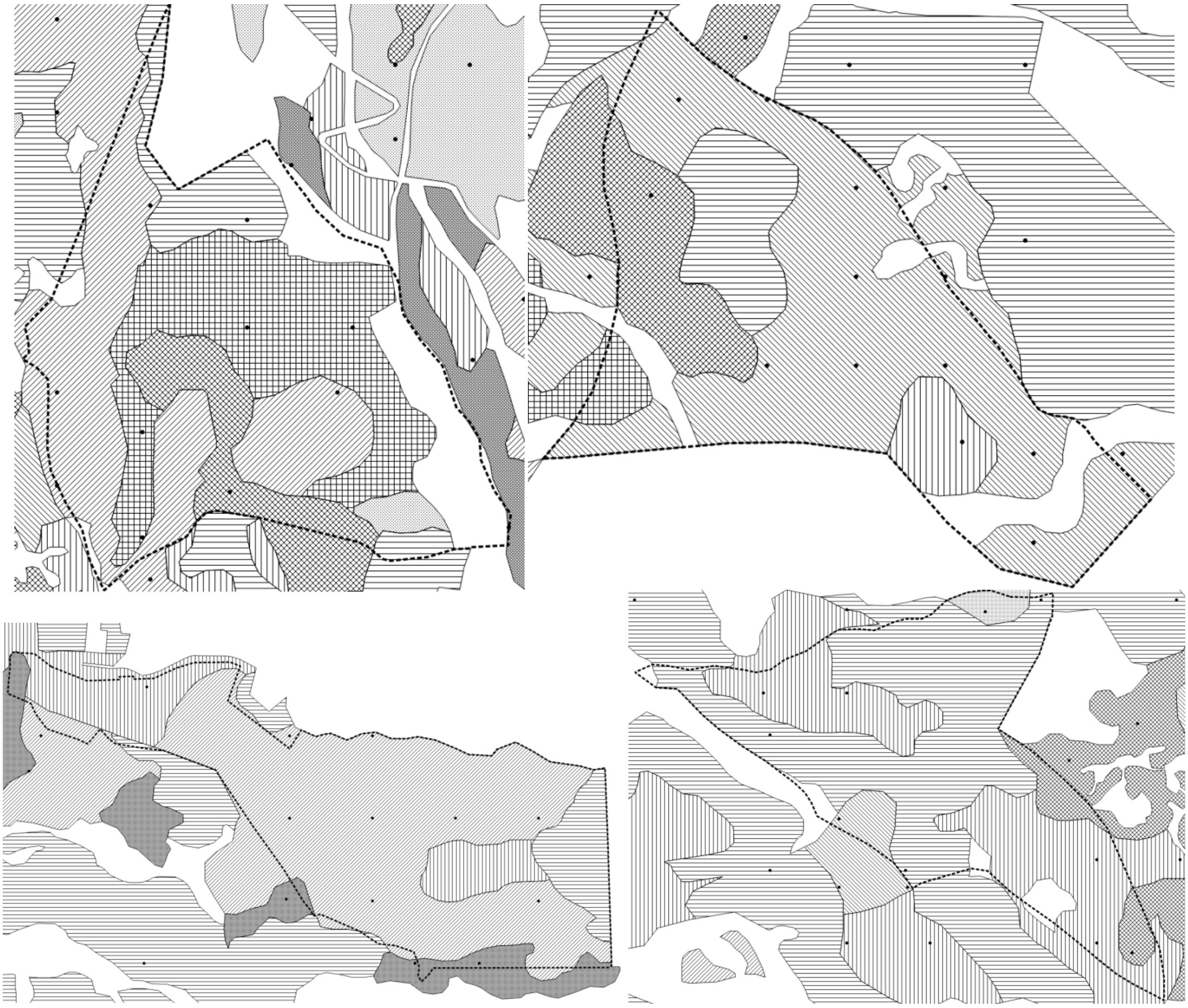

Figure 3.7 - From upper left to lower right, forest compartments 5, 10, 24, and 27 partitioned by strata and plots of field timber inventory in black dots

3.4.2 Lidar data

During the 2006-2007 autumn and winter seasons, the Province of Trento commissioned Blom CGR S.p.A. to conduct an airborne laser scanning campaign over its entire territory, aimed at producing a DTM to be used for land planning, hydro geological purposes, topographic measurements, etc. As known, the surveys made for these goals are frequently characterized by a relatively low density of points per square meter and, above all, they are usually made in winter (between November and March in temperate and Mediterranean environments and even under alpine environments, at least in those areas not permanently covered by snow in that period) to minimize the noise by vegetation, since the aim is to achieve a high penetration rate through the vegetation canopy [Corona *et al.*, 2012].

Discrete return lidar data for the whole surface of Trento province were acquired using an Optech ALTM3100C laser system mounted in Partenavia P68. The lidar system recorded range and intensity of 2 returns per pulse, and achieved a nominal pulses density of 1.28 per m^2 (Table 3.4).

Table 3.4 - Lidar winter 2006 and 2007 data acquisition parameters

Sensor	Optech ALTM3100C
Acquisition date	Autumn and winter 2006, 2007
Flight altitude	1500 m above ground
Flight line sidelap	50%
Maximum off-nadir scan angle	25°
Returns/pulse	2
Density	1.28 pulses m^{-2}
Pulse repetition	33 kHz
Laser wavelength	800 nm

The raw data were filtered and classified by the vendor using Terrascan software (Terrasolid Ltd.) which utilizes the Axelsson [1999] filtering algorithm based on Triangulated Irregular Network densification. Data delivered by the vendor included a DSM and a DTM as a regular grid with a spatial resolution of 1 m^2. We calculated the CHM as the DSM minus the DTM by means of spatial analyst tools in ArcGis 9.x software (ESRI). Canopy height values greater than 2 m were assumed to be vegetation hits, this means that values less than 2 m were set to 0.

3.4.3 The application of hybrid approach

The proposed strategy - the volume estimator [3.10] plus the hybrid variance estimator [3.23] - was applied to estimate the timber volume within four forest compartments in the public forest estate in San Martino di Castrozza.

The low number of plots in each portion precluded a direct estimation of the timber volumes within. Thus, heights were used as auxiliary information. The estimator [3.10] was adopted in each portion, and the volume estimate for the entire compartment was computed as the sum of the estimates from the single portions. The relative standard error estimates were obtained using the conservative hybrid procedure with [3.23].

The description of the steps for the application of hybrid approach is illustrated in Figure 3.8.

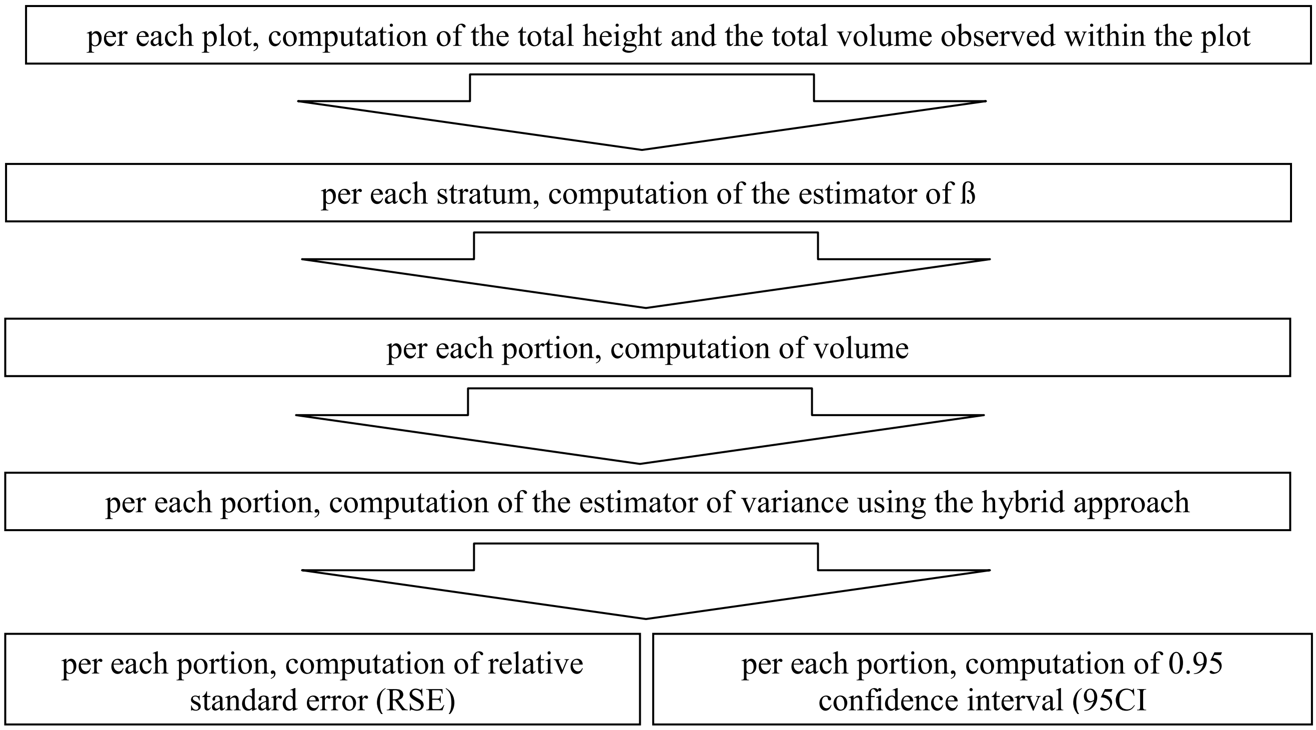

Figure 3.8 - Scheme of the application of hybrid approach in the case study

3.4.4 Results from the case study

Estimation results about forest compartments 5, 10, 24, and 27 are reported in Table 3.5, Table 3.6, Table 3.7, and Table 3.8 respectively. The code of the stratum, the number of plots per portion included in the stratum, the volume estimate per portion (m^3), the hybrid variance estimate per portion (m^6) with the relative standard error (in parenthesis), and the corresponding 0.95 confidence interval are reported respectively from the first (left side) to the fifth (right side) column.

Table 3.5 - Volume estimation within the forest compartment 5

Stratum	N° plots	Volume (m^3)	Hybrid variance (m^6) (RSE)	95CI
1	0	873	10045.64 (11.48%)	673-1072
2	1	1870	14192.16 (6.37%)	1632-2108
4	7	8947	326029.96 (6.38%)	7805-10089
7	1	366	5798.49 (20.80%)	213-519
Total		**12056**	**356066.26 (4.95%)**	**10863-13249**

Table 3.6 - Volume estimation within the forest compartment 10

Stratum	N° plots	Volume (m^3)	Hybrid variance (m^6) (RSE)	95CI
1	0	1004	12880.80 (11.30%)	777-1231
2	1	570	1998.09 (7.84%)	481-659
3	0	165	1139.68 (20.51%)	97-233
5	9	4247	200118.21 (10.53%)	3352-5142
6	1	721	11901.22 (15.14%)	503-939
Total		**6706**	**228038.04 (7.12%)**	**5751-7661**

Table 3.7 - Volume estimation within the forest compartment 24

Stratum	N° plots	Volume (m^3)	Hybrid variance (m^6) (RSE)	95CI
1	3	4651	230091.66 (10.31%)	3692-5610
2	4	3401	42195.01 (6.04%)	2990-3812
5	1	373	5043.95 (19.02%)	231-515
6	1	292	2513.89 (17.14%)	192-392
8	1	186	1206.07 (18.66%)	117-255
Total		**8904**	**281050.58 (5.95%)**	**7844-9964**

Table 3.8 - Volume estimation within the forest compartment 27

Stratum	N° plots	Volume (m^3)	Hybrid variance (m^6) (RSE)	95CI
1	1	1191	17522.03 (11.11%)	926-1456
3	4	2800	64811.83 (9.09%)	2291-3309
4	3	4775	105329.15 (6.80%)	4126-5424
6	1	545	7223.72 (15.60%)	375-715
8	0	199	1319.34 (18.23%)	126-272
Total		**9511**	**196206.10 (4.66%)**	**8625-10397**

Provided that large amounts of bias did not occur, accuracy turned out to be satisfactory for the estimation of the volumes within the entire forest compartments, which constituted the quantities of actual interest. Notwithstanding the relative standard error estimates were obtained using the conservative hybrid procedure, they resulted always smaller than 7.5%. Estimation on the portions of single strata was less satisfactory, with relative standard errors which raised the value of 20% for very small portions.

3.5 Discussion

A proportional relationship volume/height at pixel level is presumed in accordance with the ratio model to perform volume estimation in forest compartments. Because ratio model when true preserves the proportionality parameter at any spatial grain, it is possible to estimate the parameters from sample plots. Regarding the validity of the ratio model, it is likely to fail at pixel and small-plot level owing to the presence of a negative intercept. If plots are of 5, 10 m radius (as customary in forest surveys), it is indeed possible that no tree with dbh greater than the established threshold is present within the plots. In these cases, volumes are set to 0 while heights may be positive owing to the presence of trees with smaller dbh. This feature is even exacerbated within pixels of few square meters. This feature is apparent from Figure 3.5 (regarding plots of 10 m radius) and is more marked in smaller plots of 5 m radius. On the other hand, as the plots size increases, the negative intercept tends to disappear and the relationship tends to become proportional with stable proportionality coefficients. The change of relationships at different spatial grains is a well-known problem of spatial statistics, often referred to as the modifiable areal unit problem (MAUP).

Notwithstanding its possible failures, the ratio model is just adopted as a working device involving simplifying assumptions to derive a volume estimator from height and sample data. The actual bias of the volume estimator derived from the model does not depend on the validity of the model. Rather, bias depends on the similarity of the whole stratum and the portion of stratum in which estimation is performed. That is theoretically proven by [3.15] and empirically confirmed by the simulation. On the other hand, the validity of the model affects the accuracy of the volume estimator. In presence of negligible bias, the results from simulation and case study are satisfactory in terms of RRMSEs and hybrid variance estimates, respectively.

In conclusion, the proposed strategy may represent a viable way to perform volume estimation at compartment level. The sole shortcoming is the presence of an unavoidable, undetectable bias which, if non negligible, may lead to unreliable estimates and confidence intervals. Unfortunately, when we are forced to solve the scarcity of sample information by auxiliary sources and model assumptions, the resulting estimators are invariably affected by a structural bias which some statisticians accept to mask adopting model-based approaches.

3.6 References

Axelsson, P., 1999. *Processing of laser scanner data - algorithms and applications*. ISPRS J. Photogramm. Remote Sens., 54(2): 138-147. doi: 10.1016/S0924-2716(99)00008-8.

Corona, P., Cartisano, R., Salvati, R., Chirici, G., Floris, A., Di Martino, P., Marchetti, M., Scrinzi, G., Clementel, F., Travaglini, D., Torresan., C., 2012. *Airborne Laser Scanning to support forest resource management under alpine, temperate and Mediterranean environments in Italy*. European Journal of Remote Sensing, 45: 27-37, doi: 10.5721/EuJRS20124503.

Corona, P., Fattorini, L., 2008. *Area-based lidar-assisted estimation of forest standing volume.* Canadian Journal of Forest Research, 38(11): 2911-2916. doi: 10.1139/X08-122.

Fattorini, L., Franceschi, S., Maffei, D., 2013. *Design-based treatment of unit nonresponse in environmental surveys using calibration weighting*. Biometrical J. Volume 55, Issue 6: 823-972, doi: 10.1002/bimj.201100262.

Gregoire, T.G., 1998. *Design-based and model-based inference in survey sampling: appreciating the difference*. Can. J. Forest. Res. 28(10): 1429-1447.

Gregoire, T.G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., Holm, S., 2011. *Model-assisted estimation of biomass in a lidar sample survey in Hedmark County, Norway*. Can. J. For. Res. 41(1): 83-95. doi: 10.1139/X10-195.

Gregoire, T.G., Valentine, H.T., 2008. *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall, Boca Raton, UK. pp. 474.

Koivuniemi, J., Korhonen, K.T., 2006. *Inventory by compartments*. In Forest Inventory, Methodology and Applications. Edited by Kangas, A. and Maltamo, M. Springer, Dordrecht, NL. pp. 271-278.

Mandallaz, D., 2008. *Sampling Techniques for Forest Inventories*. Chapman & Hall, Boca Raton, UK. pp. 272.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. *Multivariate Analysis*. Academic Press. London. pp. 521.

Næsset, E., 2011. *Estimating above-ground biomass in young forests with airborne laser scanning*. Int. J. Remote Sens. 32(2): 473-501. doi: 10.1080/01431160903474970.

Pfefferman, D., 2002. *Small area estimation – New developments and directions*. Int. Stat. Rev, 70(1): 125-143. doi:10.1111/j.1751-5823.2002.tb00352.x.

Rao, J. N. K., 2005. *Small Area Estimation*. Wiley. New York. 344 pp.

Reutebuch, S.E., Andersen, H.E., McGaughey, R.J., 2005. *Light Detection and Ranging (LIDAR): An Emerging Tool for Multiple Resource Inventory*. Journal of Forestry, 103(6): 286-292.

Rubin, D.B., 1976. *Inference and missing data*. Biometrika 63(3): 581-592. doi:10.1093/biomet/63.3.581.

Särndal, C.E., Swensson, B., Wretman, J., 1992. *Model-Assisted Survey Sampling*. Springer-Verlag, New York. pp. 696.

Schreuder, H.T., Gregoire, T.G., Wood, G.B., 1993. *Sampling methods for multiresource forest inventory*. Wiley, New York. pp. 464.

Shao, J., Steel, P., 1999. *Variance estimation for survey data with composite imputation and nonnegligible sampling fractions*. J. Am. Stat. Assoc. 94(445): 254-265. doi:10.1080/01621459.1999.10473841.

Shao, J., Tu, D., 1995. The jackknife and the bootstrap. Springer-Verlag, New York. pp. 516.

Shiver, B.D., Borders, B.E., 1996. *Sampling Techniques for Forest Resource Inventoriy*. Wiley, New York. pp. 365.

Skovsgaard, J.P., Vanclay, J.K., 2008. *Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands*. Forestry 81(1): 13-31. doi:10.1093/forestry/cpm041.

Strunk, J.L., 2012. *Estimation and Modeling of Selected Forest Metrics with Lidar and Landsat*. Thesis for the degree of Doctor of Philosophy in Forest Resources presented on June 14, 2012, pp. 145.

4 Classification of forest stand structure using CHM-derived metrics: a comparison among statistical techniques

In this chapter the use of lidar-derived CHM metrics to predict forest structure classes according to the amount of basal area present in understory, mid-story, and overstory trees are examined. We evaluated two approaches to predict size-based forest classifications: in the first, we attempted supervised classification with linear discriminant analysis and random forest; in the second, we predicted basal areas of lower, mid, and upper canopy trees from CHM-derived variables by k-nearest neighbor imputation and parametric regression, and then classified observations based on their predicted basal areas. The strategies proved moderately successful.

4.1 Introduction

Forest structure analysis and characterization is part of most forest planning processes for sustainable forest management and is of special interest from a silvicultural point of view. Knowledge about current forest structural conditions is crucial to understanding the specific silvicultural prescriptions necessary to meet management objectives for forest attributes such as biomass, wildlife habitat, and species diversity.

The distribution of basal area in size-classes allows forest managers to quantify the current structure and condition of forest stands and predict how this may change over time, and consequently to make informed management-decisions [Ancel *et al.*, 1999; Scrinzi *et al.*, 2011].

Forest structure directly affects the capacity of stands to provide market products (*i.e.* wood) and no-market products (*i.e.* ecosystems services, including soil conservation and carbon sequestration), moreover forest structure contributes to the conservation of biological diversity [*e.g.* Staudhammer and LeMay, 2001].

Lidar, as an active remote sensing technology, has been used in many studies to describe and estimate forest structure in three-dimensions: lidar data for forestry applications involves processing the raw point cloud to compute a wide range of vegetation metrics to predict and map many aspects of forest composition and structure, including basal area [Næsset, 1997; Means *et al.*, 1999; Dubayak *et al.*, 2000; Means *et al.*, 2000; Næsset, 2002; Andersen *et al.*, 2003; Zimble *et al.*, 2003; Riaño *et al.*, 2004; Maltamo *et al.*, 2005; Barilotti *et al.*, 2007; Travaglini *et al.*, 2007; Heurich and Thoma, 2008; Ferraz *et al.*, 2012; Alberti *et al.*, 2013].

There are times that the original raw lidar point cloud may be not available, in particular when lidar data are collected for other non-forestry purposes, and only the Digital Terrain Model and Digital Surface Model are at one's disposal. It has been typically believed that these two elevation models have limited applications for forestry purposes that require characterization of forest structure. Consequently, there has been limited empirical exploration of the potential use of only digital elevation models in forest characterization and classification [Suárez *et al.*, 2012]. Despite declining costs and increasing coverage, lidar data are still not widely available in all locations for forestry applications, so the ability to use lidar-derived DTMs and DSMs has the potential to improve forest classification and mapping in areas where raw lidar data are otherwise not available.

The present study was carried out because we were interested in assessing whether lidar-derived predictor variables extracted from CHMs can support the classification of the forests in the province of Trento. Within this general purpose, the specific goal of this research was to compare the performance of supervised classification techniques toward specific methods of basal area prediction in classifying mountain forests, a theme in development in the Italian forestry research based on the exploitation of lidar data [Corona *et al.*, 2012].

4.2 Material

4.2.1 Study area

The province of Trento (6212 Km^2) is situated in the North-East of Italy on the Southern side of the Alps chain (referred to the left side of Figure 2.1).

Its climate is cool, temperate and mild continental. The average annual temperature is 11.5 °C and the average annual precipitation is 883 mm, whit May and October the months with the heaviest amount of rain, respectively 94 mm and 110 mm (CEA, 2004).

The territory is almost entirely mountainous, and 60% of its surface is covered by forests (De Natale and Pignatti, 2011). Approximately, 72% of the forest surface is owned by public institutions and is subject to a forest management plan with broad objectives, such as maintaining productive function of the forest, management of the services provided by the forest (protection, tourism and recreation, carbon dioxide fixation, etc), and improvement and conservation of biodiversity in terms of species, habitat and landscape.

The area of Trento province, according to its geographical and climatic gradients, can be divided into three zones: endalpica, mesalpica and esalpica [Odasso, 2002]. The endalpica zone includes upland areas with higher elevation and landlocked valleys. There are environments with harsh continental climate, particularly favorable to forest communities dominated by boreal conifers like Norway spruce (*Picea abies*) or cembro pine (*Pinus cembra*). The mesalpica zone includes mountains with relatively lower elevations, generally found on plateaus and in valleys, typically with East-West orientation and with average elevation around 1000 m a.s.l.. Cool climate, from sub-continental to sub-oceanic, characterizes this zone, where forests are dominated by mesophile tree species like silver fir (*Abies alba*) and beech (*Fagus sylvatica*). The esalpica zone is concentrated in a central strip orientated North-South in the Trento province territory, with elevation below 1000 m a.s.l., characterized by incursions of species with sub-Mediterranean or steppe character, dominated by the forests composed of thermophile broad leaved trees (*Ostrya carpinifolia*, *Carpinus betulus*, *Fraxinus ornus*, *Quercus pubescens*, *Quercus petrae*, etc.).

4.2.2 The forest structural classification system used by the Forest Service of Trento Province

The Forest Service of Trento Province is nowadays using a structural classification system based on the distribution of basal area in specific tree size-classes both in inventory and forest management decision making phases. The classification in forest structural types is one of the primary references, along with species composition, for forest managers to identify the strata in forest inventories based on stratified sampling. During the decision making process, forest managers use the forest structural classification to identify the potential evolution of a stand, *i.e.* the potential transition of a specific structural type toward another one, and consequently decide which kind of silvicultural interventions to implement in estimating the potential volume to fell in each tree size-class.

The forest structural classification system recently adopted by the Forest Service of Trento Province is based on the percentage of basal area in three classes of trees [Scrinzi *et al.*, 2011], defined as small trees (17.5 cm to 27.4 cm dbh), medium trees (27.5 to 47.4 cm dbh), and large trees (above 47.5 cm dbh).

When the percentage of basal area per hectare is greater than 15% in all classes, then the structure is classified as irregular, which could be dominated by small trees (I1), medium trees (I2) or large trees (I3). If the percentage of basal area per hectare is greater than 15% in two classes, then the stand is classified as regular dominated by small (R21 and R31), medium (R12 and R32) or large trees (R13 and R23). When the percentage of basal area per hectare is greater than 15% in only one class, then the stand is classified as regular widely dominated by small (R11), medium (R22) or large trees (R33). For example, if in a forest stand the basal area of small trees is 25%, the basal area of medium trees is 30% and the basal area of large trees is 45% then the stand is classified as irregular dominated by large trees, and identified by the code I3 where "I" stands for irregular and 3 is the number that distinguishes large trees.

Based on these combinations, the structure of forest stands will classified as one of twelve forest structural types, including three irregular types and nine regular types (Figure 4.1).

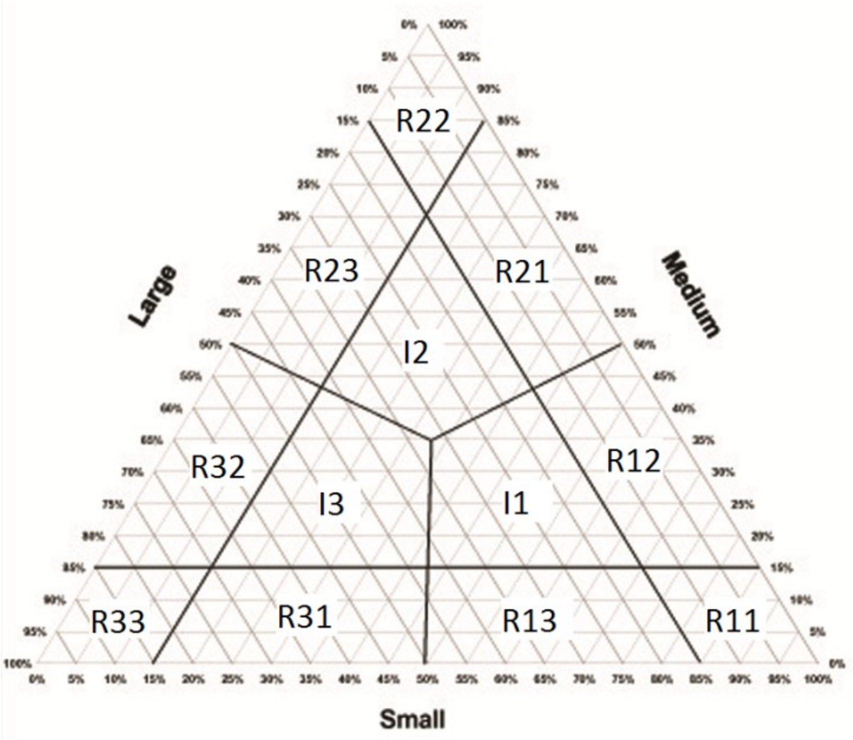

Figure 4.1 - Forest structural classification system used by the Forest Service of Trento Province where R = regular, I = irregular, 1 = small trees, 2 = medium trees, 3 = large trees

4.2.3 Field data collection and response variables

The number of field plots necessary to collect the data about the distribution of basal area in small, medium and large tree size-classes was established using a two step sampling design to ensure adequate representation of all structural types present in Trento province.

In the first step of the sampling, all forest compartments callipered for forest planning purposes from 1990 to 2006 and dominated by a single species (80% of species compositions) were selected. In the second step, the 1843 compartments selected through the first phase were reduced to 90, maintaining the proportional representation of different forest structural type. Finally, ground-truth points were randomly placed within the forest compartments ensuring the presence of two points per compartment, one being the point to be surveyed, and the other, as a reserve to cover those cases in which conditions in the field did not meet the target forest type.

The field surveys were carried out from June to September 2007. The center of each circular plot was georeferenced with a Thales MobileMapper CE Global Positioning System (GPS) receiver recording at minimum 200 satellite positions on the ground surface. Afterwards the geographical coordinates of plot centers were differentially corrected and averaged thus obtaining a plot location accuracy of less than 3 m. In this study, different plot sizes (531, 1257, 1964 and 2827 m^2) were employed to ensure the minimum level of tree density in each plot. The dbh and height were measured for all trees in each plot. The diameter of each tree (dbh $\geq$ 7.5 cm) was measured from two orthogonal axes using a timber calipers at breast height (1.3 m) and then averaged. The height of each tree was measured by a Vertex-II ultrasonic hypsometer.

The dbh was measured to calculate the field-based proportion of basal area on each plot in the three size-classes (small, medium and large) needed to classify the forest structure: both the proportion of basal area, as continuous variable, and forest structure type, as categorical variable, were used as response variable.

The heights and diameters were used to build a height-diameter model: in fact, although we already had field measured tree height, we modeled the tree height to obtain a diameter-height function working for all trees and for all different forest structural type. In this way it was possible to identify specific height thresholds in which the small, medium and large trees could be placed, and consequently threshold the prediction variables. The model was built according to the Curtis [1967] equation (refer to equation [2.2]). The main statistic parameters of the height-diameter function are reported in Table 4.1.

Table 4.1 - Main statistic parameters of the height-diameter model according to the Curtis [1967] equation

Statistic parameter	**Model**
Dependent variable	*H*
Measure unit	M
Model equation	$h = 1.3 + \dfrac{d^2}{a + bd + cd^2}$
a	42.31
b	-0.60
c	0.03
a significance level (p-value)	$p<0.001$
b significance level (p-value)	$p<0.05$
c significance level (p-value)	$p<0.001$
R^2	0.51

By means of the height-diameter function we estimated that the height of small trees ranged from 8.5 to 15.6 m, the height of medium trees ranged from the 15.6 to 26.5 m and large trees were greater than 26.5 m.

The 90 plots surveyed were classified into nine forest structure types: I2 (6 plots), I3 (3 plots), R11 (9 plots), R12 (7 plots), R21 (14 plots), R23 (22 plots), R31 (1 plot), R32 (25 plots), and R33 (3 plots).

4.2.4 Lidar sampling and lidar predictor variables

For the description of lidar data characteristics and processing, referrer to § 3.5.2.

From the available CHM for each of the 90 plots, grid cells, corresponding in shape and size at those of each field plot, were extracted using ArcGIS. The *maptools* package in R [Lewin-Koh and Bivand, 2012] was used to compute the predictor variables from the height probability distributions (thirteen predictor variables), and from the relative frequency distributions of vegetation heights (four predictor variables) of each plot: in total, seventeen candidate predictor variables were generated for modeling (Table 4.2).

Table 4.2 - Seventeen candidate predictor variables

HMAX	Maximum of height pixels intersecting plot
HMEAN	Mean of height pixels intersecting plot
HKURT	Kurtosis of the height pixels distribution intersecting the plot
HSKEW	Skewness of the height pixels distribution intersecting the plot
HVAR	Variance of height pixels intersecting plot
HCV	Coefficient of variation of height pixels intersecting plot
HSTD	Standard deviation of height pixels intersecting plot
H05PCT	Height at which 5% of pixels intersecting plot fall below
H10PCT	Height at which 10% of pixels intersecting plot fall below
H25PCT	Height at which 25% of pixels intersecting plot fall below
H50PCT	Height at which 50% of pixels fall below (median of height pixels intersecting plot)
H75PCT	Height at which 75% of pixels intersecting plot fall below
H90PCT	Height at which 90% of pixels intersecting plot fall below
COVSMALL	Cover of small trees (ratio of the number of pixels intersecting plot with height >=8.5 m and <15.6 m over the total number of pixels intersecting plot)
COVMEDIUM	Cover of medium trees (ratio of the number of pixels intersecting plot with height >=15.6 m and <26.5 m over the total number of pixels intersecting plot)
COVLARGE	Cover of large trees (ratio of the number of pixels intersecting plot with height >=26.5 over the total number of pixels intersecting plot)
COV2	Cover of pre-inventory trees (ratio of the number of pixels intersecting plot with height >=2.0 m and <8.5 m over the total number of pixels intersecting plot)

4.3 Methods

4.3.1 Supervised classification approach

4.3.1.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is a supervised classification approach. In LDA, a linear combination of auxiliary variables is identified which maximizes separation between categorical response groups (Reimann *et al.*, 2008):

$$w_1 = a_1 x_1 + a_2 x_2 + ... + a_k x_k = \sum_{i=1}^{k} a_i x_i \qquad [4.1]$$

The weight a_{1j} is chosen to maximize the separation between groups. A classification rule is developed in combination with the weighting function in [2]. Using maximum likelihood theory, classifications are assigned to ranges of values. *E.g.*, if $w_1 \in (l_1, l_2]$ for an observation where l_1 and l_2 represent limits for a given class, then the observation is assigned to that class. If $w_1 \notin (l_1, l_2]$ then the observation is assigned to an alternate class (Figure 4.2).

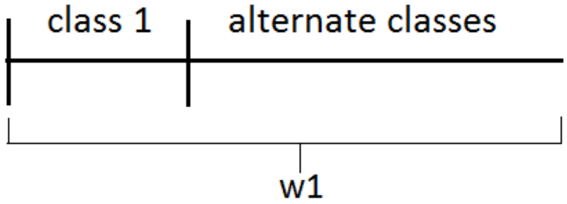

Figure 4.2 - Schematic describing assignment of LDA values (predictions) to response classes

LDA was performed in R using the *MASS* package [Venables and Ripley, 2002], which implements a Bayesian decision theory approach. For LDA the covariates are assumed to have a common multivariate normal distribution.

4.3.1.2 Random forest

Random forest (RF) is a classification technique, based on the use of classification and regression trees, developed by Breiman [2001].

Classification trees build rules for assigning current observations into classes using numerical and/or categorical predictor variables by recursive binary partitioning into regions that are increasingly homogeneous with respect to the class variable. The homogeneous regions are called nodes [Cutler *et al.*, 2007]. RF fits many classification trees to a data set, and then combines the predictions from all the trees. The algorithm begins with the selection of many (*e.g.* 500) bootstrap samples from the data. Observations in the original data set that do not occur in a bootstrap samples are called out-of-bag observations. A classification tree is fit to each bootstrap sample, and at each node, a small number of randomly selected variables are available for the binary partitioning. The trees are fully grown and each is used to predict the out-of-bag observations. The predicted class of an observation is calculated by majority vote of the out-of-bag predictions for that observation, with ties split randomly.

Breiman [2001] called this procedure random forest because the base constituents of the ensemble are tree-structured predictors, and because each of these trees is constructed using an injection of randomness.

The RF classification was carried out using the *randomForest* package in R [Liaw and Wiener, 2002]. The RF algorithm in R works as follows:

- grow a specified number of trees (n_{tree}) by the bootstrap method from the original data;
- for each of the bootstrap samples, grow an unpruned classification so that at each node randomly sample a number of variables (m_{try}) of the predictors as candidates at each split;
- predict new data by aggregating the predictions of the n_{tree} trees.

The *randomForest* package optionally produces the measure of the importance of the predictor variables, and a measure of the internal structure of the data, that give information about the proximity of different data points to one another.

4.3.2 Prediction-based classification approach

4.3.2.1 K-nearest neighbor

k-NN imputation predicts the value of the variable of interest as a weighted average of values of nearest neighboring observations [Maltamo, 2003]. The neighbors are defined with some similarity (or distance) measure in the predicting variables, selecting the neighbors from the observations whose variables were previously measured.

For nearest neighbor search and imputation different methods can be used: in this study the most similar neighbor (MSN) and the RF methods were considered. In the MSN method, the nearness is defined using the weighted Euclidean distance [Crookston and Finley, 2008] based on canonical correlation analysis between the independent variables and the dependent variables. In the RF method, the observations are considered similar if they tend to end up in the same terminal nodes in a suitably constructed collection of classification and regression trees. The distance measure is one minus the fraction of trees in which two separate observations fall in the same terminal node.

The *yaImpute* package in R [Crookston and Finley, 2008] was used for *k*-NN imputation by the MSN and RF methods. In this package, any k number of reference observations can be selected to impute a target observation. The prediction accuracy usually increases for some number of k greater than one (*e.g.* 3 or 5), then gradually declines for larger values of k [Muinonen *et al.*, 2001].

4.3.2.2 Parametric regression

The second strategy used to predict basal areas for small, medium and large trees used parametric linear and nonlinear models.

Parametric regression analysis is performed to evaluate the dependence of a response variable on one or several predictors. The function is specified explicitly and can be linear and non linear in the parameters. In this study, a linear regression model with the selected explanatory variables was developed by using ordinary least squares (OLS). There are extensive examples in which OLS has been successfully applied to predict basal area using lidar metrics [*e.g.* Hudak *et al.*, 2006; Lefsky *et al.*, 1999].

A non-linear regression model with multiple explanatory variables was also developed by using the non-linear least squares (NLS).

4.3.3 Selection of predictor variables

To avoid the risk of over fitting in case of a high number of predictor variables, and to be able to compare the results obtained with the different statistical techniques for the structure classification and prediction considered in this study, starting from our pool of seventeen candidate predictor variables, the best subsets regression modeling approach was employed to support the choice of the best predictor variables for predicting the total basal area per hectare (small, medium and large trees). We used the *regsubsets()* function available in the *leaps* package of R [Lumley, 2004], which selects the best regression subsets through exhaustive search. The *regsubsets()* requires the user to set a maximum number of variables in a subset model (in our case set to 4). The model statistic used to determine best subsets was the Bayesian Information Criterion (BIC) of Schwarz [1978]. The predictor variables selected by this approach were HCV, COVSMALL, COVMEDIUM, COVLARGE.

4.3.4 Performance statistics

Performance statistics were computed both for the classification accuracy of our multiple approaches, and for the prediction accuracy from modeling basal area with CHM derivatives.

We used leave-one-out (LOO) cross-validation to evaluate our ability to predict forest structure classes from CHM data. The results of LOO were categorized in confusion matrices, and evaluated by the overall, producer's, and user's accuracies. The overall model accuracy represents the percentage of plots correctly classified with respect to the total number of sample plots. The producer's accuracy refers to the probability that a certain forest structural type of a stand on the ground is classified as such. The user's accuracy refers to the probability that a stand labeled as a certain forest structural type is really this type.

We looked at the prediction performances of basal area models using the RMSE. According to Kendall and Buckland [1975], "*in general, the mean square error of a set of values is the arithmetic* *mean of the squares of their differences from some given value, namely their second moment about that value. When the mean square is regarded as an estimator of certain parental variance components the sum of the squares about the observed mean is usually divided by the number of degrees of freedom, not the number of observations*”.

When considering the basal area per hectare of small, medium, large and all trees as the response variables, the RMSE was calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(G_i - \hat{G}_i)^2}{n-p}} \qquad [4.2]$$

where G_i is the observed basal area per hectare of sample plot i, $\hat{G}_i$ is the basal area of sample plot i estimated from the predicted distribution, $n - p$ is the degrees of freedom. To facilitate comparisons between the performances of basal area prediction techniques, we also reported a scaled RMSE obtained by dividing the RMSE by the standard deviation of the training dataset. For all basal area prediction techniques considered, we also computed the coefficient of determination, which is the proportion of total variation explained by the model.

4.4 Results

The results of the analyses within the supervised classification approach are shown with confusion matrices in Table 4.3 (LDA technique) and Table 4.4 (RF technique).

Table 4.3 - Classification result by LDA

		Truth forest structural type										
		I2	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**N° classified plots**	**User's accuracy**
Classified forest structural types by LDA	**I2**	**1**	0	0	1	1	0	0	0	0	3	33%
	I3	0	**0**	0	0	0	0	0	0	0	0	-
	R11	0	0	**6**	3	2	1	0	0	0	12	50%
	R12	0	0	0	**0**	0	0	0	0	0	0	-
	R21	3	1	2	1	**6**	3	0	1	0	17	35%
	R23	2	2	1	1	4	**7**	1	8	1	27	26%
	R31	0	0	0	0	0	0	**0**	0	0	0	-
	R32	0	0	0	1	1	9	0	**16**	1	28	57%
	R33	0	0	0	0	0	2	0	0	**1**	3	33%
N° ground truth plots		6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		17%	0%	67%	0%	43%	32%	0%	64%	33%		
Overall accuracy		41%										

Table 4.4 - Classification result by RF (n_{tree} = 200 and m_{try} = 4)

		Truth forest structural type										
		I2	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**No classified plots**	**User's accuracy**
Classified forest structural types by RF	**I2**	**0**	0	1	1	1	1	0	0	0	4	0%
	I3	0	**0**	0	0	0	0	0	1	0	1	0%
	R11	1	0	**4**	2	3	1	0	0	0	11	36%
	R12	1	0	0	**1**	2	0	0	0	0	4	25%
	R21	2	0	4	2	**3**	3	0	1	0	15	20%
	R23	1	2	0	0	4	**8**	1	6	3	25	32%
	R31	0	0	0	0	0	0	**0**	0	0	0	-
	R32	1	1	0	1	1	9	0	**17**	0	30	57%
	R33	0	0	0	0	0	0	0	0	**0**	0	-
No ground truth plots		6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		0%	0%	44%	14%	21%	36%	0%	68%	0%		
Overall accuracy		37%										

The overall accuracy of LDA was 41%. The user's accuracy has not exceeded 57% while the producer's accuracy has not exceeded the 67%. The forest structure types that had especially good producer's accuracy were the R11 and R32, while the I3, R12 and the R31 reached low levels of producer's accuracy.

In the case of RF classification, a stable overall accuracy (37%) was obtained setting the number of classification trees per response variable equal to 200 and choosing the number of predictor variables to use in each split equal to 4: using lower or greater than 200 trees resulted in overall declining accuracy. Also in this case, the user's and producer's accuracy has not respectively exceeded 57% and 68%. The worst producer's accuracy was in the classification of plots characterized by irregular structure (I2 and I3) and regular structure dominated by large trees (R31 and R33), while the better produce's accuracy was in classifying the R32 forest structure type.

The results of the analyses within the prediction-based classification approach are shown with confusion matrices in Table 4.5 and Table 4.6 (*k*-NN with RF and MSN methods respectively) and Table 4.7 and Table 4.8 (multiple linear regression and multiple non linear regression respectively).

Between the techniques used in the prediction-based classification approach, the *k*-NN with the RF method provided better results with respect to the multiple linear regression and non linear regression. The overall accuracy for *k*-NN imputation by the RF method was 37% (Table 4.5). The maximum level of user's accuracy was 62% while the maximum level of producer's accuracy was 56%. The irregular structures (I2 and I3) and the regular stands dominated by large trees (R31 and R33) had poor producer's accuracy, while the R11 and R32 forest type had the higher accuracy (56% and 52% respectively).

Table 4.5- Classification result by k-NN imputation using RF method

		Truth forest structural types										
		I2	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**No classified plots**	**User's accuracy**
Classified forest structural types by *k*-NN (RF)	**I2**	**0**	1	1	1	2	0	0	2	0	7	0%
	I3	2	**0**	0	1	1	1	0	1	0	6	0%
	R11	2	0	**5**	1	4	1	0	0	0	13	38%
	R12	0	0	1	**2**	2	0	0	0	0	5	40%
	R21	1	0	2	2	**4**	3	1	0	0	13	31%
	R23	1	2	0	0	1	**9**	0	9	2	24	38%
	R31	0	0	0	0	0	0	**0**	0	0	0	-
	R32	0	0	0	0	0	7	0	**13**	1	21	62%
	R33	0	0	0	0	0	1	0	0	**0**	1	-
No ground truth plots		6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		0%	0%	56%	29%	29%	41%	0%	52%	0%		
Overall accuracy		37%										

The overall accuracy for *k*-NN imputation by the MSN method was 34%. The maximum value of producer's accuracy was 40%, while the user's accuracy did not pass 46% (R23). The forest structure types that had especially good producer's accuracy were the R23 and R32, while the I3, R31 and the R31 reached low levels of producer's accuracy (Table 4.6).

Table 4.6- Classification result by k-NN imputation using MSN method

		Truth forest structural types										
		I2	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**No classified plots**	**User's accuracy**
Classified forest structural types by *k*-NN (MSN)	**I2**	**1**	0	0	0	2	0	0	2	0	5	20%
	I3	0	**0**	0	1	1	0	0	0	0	2	0%
	R11	0	0	**2**	1	3	1	1	0	0	8	25%
	R12	0	0	1	**2**	1	1	0	0	0	5	40%
	R21	3	2	4	2	**4**	0	0	3	0	18	22%
	R23	0	1	1	1	1	**12**	0	8	2	26	46%
	R31	0	0	1	0	0	0	**0**	0	0	1	0%
	R32	2	0	0	0	2	7	0	**10**	1	22	45%
	R33	0	0	0	0	0	1	0	2	**0**	3	0%
No ground truth plots		6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		17%	0%	22%	29%	29%	55%	0%	40%	0%		
Overall accuracy		34%										

The overall accuracy in the classification by the multiple linear regression was 28%. (Table 4.7) The maximum level of producer's accuracy was 83% while the maximum level of user's accuracy was 54%: this prediction-based classification technique shows poor level of producer's accuracy in the classification of I1, I3, R11, R21, R31 and R33, and good level in the classification of I2.

Table 4.7 - Classification result by multiple linear regression

		Truth forest structural types											
		I1	**I2**	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**No classified plots**	**User's accuracy**
Classified forest structural types by OLS	**I1**	**0**	0	0	0	1	4	0	0	0	0	5	0%
	I2	0	**5**	3	2	2	8	11	1	6	0	38	13%
	I3	0	0	**0**	0	0	0	0	0	0	0	0	-
	R11	0	0	0	**0**	0	0	0	0	0	0	0	-
	R12	0	0	0	7	**4**	2	1	0	0	0	14	29%
	R21	0	0	0	0	0	**0**	0	0	0	0	0	-
	R23	0	0	0	0	0	0	**2**	0	5	0	7	29%
	R31	0	0	0	0	0	0	0	**0**	0	0	0	-
	R32	0	1	0	0	0	0	8	0	**14**	3	26	54%
	R33	0	0	0	0	0	0	0	0	0	**0**	0	-
No ground truth plots		**0**	6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		-	83%	0%	0%	57%	0%	9%	0%	56%	0%		
Overall accuracy		28%											

The overall accuracy of the classification based on the results of non linear regression was 26% (Table 4.8). The table shows that the user's accuracy has not exceeded 57% while the producer's accuracy has not passed 67%. The forest structure types that had especially good producer's accuracy were R11 and R32, while the I3, R12 and R31 reached low levels of producer's accuracy.

Table 4.8 - Classification result by non-linear regression

		Truth forest structural type											
		I1	**I2**	**I3**	**R11**	**R12**	**R21**	**R23**	**R31**	**R32**	**R33**	**No classified plots**	**User's accuracy**
Classified forest structural types by NLS	**I1**	**0**	0	0	0	2	2	0	0	0	0	4	0%
	I2	0	**4**	2	2	2	9	11	0	6	0	36	11%
	I3	0	0	**0**	0	0	0	0	1	0	0	1	0%
	R11	0	0	0	**0**	0	0	0	0	0	0	0	-
	R12	0	1	1	6	**3**	3	1	0	0	0	15	20%
	R21	0	0	0	0	0	**0**	0	0	0	0	0	-
	R23	0	0	0	0	0	0	**2**	0	5	0	7	29%
	R31	0	0	0	0	0	0	0	**0**	0	0	0	-
	R32	0	1	0	1	0	0	8	0	**14**	3	27	52%
	R33	0	0	0	0	0	0	0	0	0	**0**	0	-
No ground truth plots		0	6	3	9	7	14	22	1	25	3	**90**	
Producer's accuracy		-	67%	0%	0%	43%	0%	9%	0%	56%	0%		
Overall accuracy		26%											

We can note a common trend across all techniques (used both in supervised classification and prediction-based classification) consisting in high levels of producer's accuracy of the forest structural type R32. Supervised classification techniques are poor in classifying the irregular structure, *i.e.* multi-stored stands, vice versa the linear and non-linear regression techniques provided the higher levels of producers' accuracy in these types of forest structure.

RF *k*-NN, MSN *k*-NN, OLS and NLS had comparable scaled RMSE values in prediction the basal area per hectare for small, medium and large trees and of all trees (Table 4.9). In all cases the error was greater in the prediction of basal area per hectare of medium trees (around 10 m^2/ha), and lower in the prediction of basal area per hectare of small trees (around 6.5 m^2/ha).

The technique that provided the best performance in basal area prediction was the *k*-NN with the MSN method: the R^2 to predict small, medium, large and all trees was 0.71, 0.48, 0.69 and 0.56 respectively.

Table 4.9 - Root Mean Square Error (RMSE), scaled Root Mean Square Error (scaled RMSE) and determination coefficient (R^2) for the prediction basal areas techniques for small, medium, large trees and all trees

	RMSE (m^2/ha)				Scaled RMSE				R^2			
	***k*-NN (RF)**	**K-NN (MSN)**	**OLS**	**NLS**	***k* -NN (RF)**	***k* -NN (MSN)**	**OLS**	**NLS**	***k* -NN (RF)**	***k* -NN (MSN)**	**OLS**	**NLS**
Small	7.03	6.29	6.91	6.48	0.70	0.63	0.69	0.65	0.68	0.71	0.54	0.62
Medium	11.06	11.09	9.49	9.49	1.06	1.06	0.91	0.91	0.49	0.48	0.20	0.20
Large	8.74	8.37	8.56	8.43	0.67	0.64	0.66	0.65	0.68	0.69	0.58	0.58
All	11.16	10.65	8.82	8.86	0.93	0.89	0.73	0.74	0.54	0.56	0.48	0.47

4.5 Discussion

To date, it often happens that the Italian forest technicians can capitalize on the CHM as a product of ALS surveys made for other non-forestry purposes at low [Clementel *et al.*, 2012] or even no cost [Corona *et al.*, 2012]. This availability has meant that the technicians, who deal with forest management planning, familiarize themselves with this product. For these reasons, it was considered essential to evaluate if this kind of limited lidar dataset could have application to forest structural prediction, and consequently for mapping forest structural types.

The results of this study suggest that the lidar-derived predictor variables extracted from CHMs may be used just for preliminary analyses within, for example, forest inventory processes that require the recognition of forest structural types with the advantage of reducing time-consuming and expensive fieldworks activities. In fact, the techniques used both in supervised classification approach and prediction-based classification approach did not produced satisfactory levels of classification accuracy. In this context, the non parametric techniques produced more accurate classifications with respect to the parametric techniques, and in particular the linear discriminant analysis provided the best results.

It is difficult to compare our results to other related works, because in other studies lidar metrics were derived from the normalized cloud points. It seemed important to evaluate our results instead in relation to those obtained from researches with similar goals, where classification approaches were applied.

Zhang *et al.* [2011] obtained a higher level of overall accuracy (91.4%) performing the linear discriminant analysis to classify five forest types in the Strzelecki Ranges (southeast Victoria, Australia) using eighteen lidar-derived predictor variables from the normalized returns. The level of accuracy obtained in our study is not comparable to the Zhang *et al.* [2011] results, but probably the high number of forest structure types we wanted to classified (eleven *vs.* five) is a limiting factor of classification accuracy, and it must take in consideration that the number of observations is not equitably distributed among the forest structural types.

Chirici *et al.* [2013], to classify nine forest fuel types in the Mediterranean province of Palermo and Catania (Italy) previously observed and identified by photo interpretation, applied the RF technique using thirty-one ALS-based metrics calculated from the normalized height returns. The overall accuracy obtained through this technique was 45%, very similar to that obtained in this study.

Hudak *et al.* [2008], imputing plot-level basal area of eleven conifer species in Moscow Mountain and St. Joe Woodlands in North-Central Idaho (USA) with MSN and RF imputation methods, obtained similar scaled RMSD to the present work (if the basal area of all trees is considered).

Considering that the results we obtained are very similar to those mentioned studies, we can infer that the potential to use limited lidar data (*i.e.* CHM) to predict forest structure is confirmed: variables extracted from the CHM have the same potential as those extracted from lidar point cloud data sets in predicting forest structure types. Furthermore, forests of Trento province cover a wide range of structural conditions which are not typical of this own geographic area, are independent of site fertility and of climate conditions, this means that the findings from this study can be extrapolated and applied to other forests located outside the province of Trento.

The sample design of this study could have potentially influenced the statistical scope of inference limiting the validity of the results: the unbalanced number of cases per forest structural type, due to the fact that in some cases the real forest structure type did not match that selected by sampling, could have influenced the model classification accuracy. In fact, if the within forest type variability (internal variability) is relatively high with respect to the between variability (external variability), then probably the difference among the forest structural types is the result of the internal variability. The availability of more data, acquired by future surveys carried out during the operations of forest plans revision, will allow retesting the techniques considered in this work. In addition, further investigations will consider families of structure that group together the twelve types of the structural classification system: one option could be to take in consideration the mono-stored forest stands (R11, R22, and R33), the bi-stored forest stands (R12, R13, R21, R23, R31, and R32), and the multi-stored forest stands (I1, I2, and I3).

The methodology used to calculate the canopy cover of small, medium and large trees through a diameter-height function working for all trees of the different forest structural types, does not represent a limitation of this study. The uncertainty when assigning a tree to a diameter class based only in its height by the diameter-height function subsists even on plot level, but this method of cover estimation does not pose a limitation in our analysis. In fact, if we consider, for example, the cover of small trees, at a first analysis the cover of this category of trees seems include the cover of medium and large trees, but each cover stratum includes only the pixels which height value falls in the range of tree height modeled by the diameter-height function and in this way we avoid the risk of double or triple count. Hudak *et al.* [2008], instead of calling this predictor variable cover, called these variables STRATUMn. They considered seven strata where for example STRATUM4 is the percentage of vegetation returns >10 m and <= 20 m in height. The height interval can be chosen arbitrarily or can be define in some objective way, as in our study, but the cover estimation does not represent a problem in the analysis we would carry out. Moreover, considering that the CHM takes into account the outer canopy surface, it is more likely that the CHM "blanket" represents the tallest part of the crown. Lidar data coming from ALS campaigns that record more than first and last return per pulse, allow more insight into what happens inside each stratum. For this reason forest technicians should be involved in the process of flight protocol definition.

4.6 References

Alberti, G., Boscutti, F., Pirotti, F., Bertacco, C., De Simon, G., Sigura, M., Cazorzi, F., Bonfanti, P., 2013. *A lidar-based approach for a multi-purpose characterization of Alpine forests: an Italian case study.* iForest, 6: 156-168, doi: 10.3832/ifor0876-006.

Ancel, P., Wilhelm, M.E., Lacombe, E., 1999. *Peuplements forestiers du massif Vosgien; typologie et sylvicultures*. CNPF. pp. 51.

Andersen, H.E., Foster, J.R., Reutebuch E., 2003. *Estimating forest structure parameters on Fort Lewis Military Reservation using Airborne Laser Scanner (LIDAR) data*. Proceedings of the Second International Precision Forestry Symposium, Seattle, Washington, June 15-17, 2003, pp. 45-53.

Barilotti, A., Sepic, F., Abramo, E., Crosilla, F., 2007. *Assessing the 3D structure of the single crowns in mixed alpine forests*. In: Stilla, U., Mayer, H., Rottensteiner, F., Heipke, C., Hinz, S. Editor(s) PIA07. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVI-3/W49A.

Breiman, L., 2001. *Random Forests*. Machine Learning, 45(1): 5-32. doi 10.1023/A:1010933404324.

CEA, 2004. *Il clima de Trentino. Distribuzione spaziale delle principali variabili climatiche*. Centro di Ecologia Alpina. Report n. 33. Trento 2004.

Chirici, G., Scotti, R., Montaghi, A., Barbati, A., Cartisano, R., Lopez, G., Marchetti, M., McRoberts, R.E., Olsonn, H., Corona, P., 2013. *Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery.* International Journal of Applied Earth Observation and Geoinformation, 25: 87-97. doi: 10.1016/j.jag.2013.04.006.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., Torresan, C., 2012. *Estimating forest timber volume by means of "low-cost" lidar data*. Italian Journal of Remote Sensing, 44(1): 125-140. doi:10.5721/ItJRS201244110.

Corona, P., Cartisano, R., Salvati, R., Chirici, G., Floris, A., Di Martino, P., Marchetti, M., Scrinzi, G., Clementel, F., Travaglini, D., Torresan, C., 2012. *Airborne Laser Scanning to support forest resource management under alpine, temperate and Mediterranean environments in Italy*. European Journal of Remote Sensing, 45: 27-37. doi: 10.5721/EuJRS20124503.

Crookston, N.L., Finley, A.O.,2008. *yaImpute: An R Package for kNN Imputation*. Journal of Statistical Software. 23(10): 1-16.

Curtis, R., 1967. *Height-diameter and height-diameter-age equations for second-growth Douglas-fir*. Forest Science, 13: 365-375.

Cutler, R.D., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.H., Gibson, J., Lawler, J.J., 2007. *Random forests for classification in ecology*. Ecology, 88(11): 2783-2792.

De Natale, F., Pignatti, G., 2011. *Estensione e composizione dei boschi italiani.* In: Gasparini, P., Tabacchi, G. (a cura di) 2011 - L'Inventario Nazionale delle Foreste e dei serbatoi forestali di Carbonio INFC 2005. Secondo inventario forestale nazionale italiano. Metodi e risultati. Ministero delle Politiche Agricole, Alimentari e Forestali, Corpo Forestale dello Stato; Consiglio per la Ricerca e Sperimentazione in Agricoltura, Unità di Ricerca per la Pianificazione ed il Monitoraggio e la Pianificazione Forestale. Edagricole, Milano, pp. 99-111.

Dubayah, R.O, Knox, R.G, Hofton, M.A., Blair, J.B., Drake, J.B., 2000. *Land surface characterization using lidar remote sensing*. In, Spatial information for land use management, Hill, M., Aspinall, R.(Eds.), International Publishers Direct, Singapore (2000).

Ferraz, A., Bretar, F., Jacquemound, S., Gonçalves, G., Pereira, L., Tomé, M., Soares, P., 2012. *3-D mapping of multy-layered Mediterranea forest using ALS data.* Remote Sensing of Environment, 121: 210-223. doi: 10.1016/j.rse.2012.01.020.

Heurich, M., Thoma, F., 2008. *Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural European beech (Fagus sylvatica) and Norway spruce (Picea abies) forests*. Forestry, 81(5): 645-661. doi: 10.1093/forestry/cpn038.

Hudak, A.T., Crookston, N.L., Evans, J.S., Falkowski, M.J., Smith, A.M.S., Gessler, P.E., Morgan, P., 2006. *Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data*. Canadian Journal of Remote Sensing, 32: 126-138. doi: 10.5589/m06-007.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., Falkowski, M.J., 2008. *Nearest neighbor imputation of species-level, plot-scale forest structure attributes from lidar data*. Remote Sensing of Environment, 112: 2232-2245. doi: 10.1016/j.rse.2007.10.009.

Kendall, M.G, Buckland, W.R., 1975. *A Dictionary of Statistics Terms*. Third edition. Longman Group Limited. pp. 166. ISBN 0 582 46304 1.

Lefsky, M.A., Harding, D., Cohen, W.B., Parker, G., Shugart, H.H., 1999. S*urface Lidar Remote Sensing of Basal Area and Biomass in Deciduous Forests of Eastern Maryland, USA*. Remote Sensing of Environment, 67: 83-98.

Lewin-Koh, N.J., Bivand, R., 2012. *maptools: Tools for reading and handling spatial objects*. R package version 0.8-14.

Liaw, A., Wiener, M., 2002. *Classification and regression by randomForest*. R News, 2(3): 18-22, ISSN 1609-3631.

Lumley, T., 2004. *Leaps: regression subset selection. R package version 2.7. Uses Fortran code by Alan Miller*.

Maltamo, M., Malinen, J., Kangas, A., Härkönen, S., Pasanen, A.M., 2003. *Most similar neighbor based stand variable estimation for use in inventory by compartments in Finland*. Forestry, 76(4): 449-464. doi: 10.1093/forestry/76.4.449.

Maltamo, M., Packalén, P., Yu, X., Eerikäinen, K., Hyyppä, J., Pitkänen, J., 2005. *Identifing and quintifying structural characteristics of heterogeneous borela forests using laser scanner data*. Forest Ecology and Management, 216: 41-50. doi: 10.1016/j.foreco.2005.05.034.

Means, J.E., Acker, S.A., Fitt, B.J., Renslow, M., Emerson, L., Hendrix, C.J., 2000. *Predicting forest stand characteristics with airborne scanning lidar*. Photogrammetric Engineering & Remote Sensing, 66(11): 1367-1371.

Means, J.E., Acker, S.A., Harding, D.J, Blair, J.B., Lefsky, M.A., Cohen, W.B., Harmon, M.E., McKee, W.A., 1999. *Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon*. Remote Sensing of Environment, 67: 298-308. doi: 10.1016/S0034-4257(98)00091-1.

Muinonen, E., Maltamo, M., Hyppanen, H., Vainikainen, V., 2001. *Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information*. Remote Sensing of Environment, 78: 223-228.

Næsset, E., 1997. *Estimating timber volume of forest stands using airborne laser scanner data*. Remote Sensing of Environment, 61: 246-253. doi: 10.1016/S0034-4257(97)00041-2.

Næsset, E., 2002. *Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data*. Remote Sensing of Environment, 80: 88-99. doi: 10.1016/S0034-4257(01)00290-5.

Odasso, M., 2002. *I tipi forestali del Trentino*. Report del Centro di Ecologia Alpina, 25, pp. 192.

Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2008. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. WILEY, 2008, pp. 362. ISBN 047098581X.

Riaño, D., Valladares, F., Condésand, S., Chuvieco, E., 2004. *Estimation of leaf area index and covered ground from airborne laser scanner (Lidar) in two contrasting forests*. Agricultural and Forest Meteorology, 124: 269-275. doi: 10.1016/j.agrformet.2004.02.005.

Schwarz, G., 1978. *Estimating the dimension of a model*. Annals of Statistics, 6: 461-464, 1978.

Scrinzi, G., Clementel, F., Colle, G., Farruggia, C., Floris, A., Torresan, C., 2011. *L'inventario dendromemtrico nella nuova pianificazione forestale aziendale trentina (NPFAT)*. Provincia Autonoma di Trento – Servizio Foreste e Fauna, Consiglio per la Ricerca e la Sperimentazione in Agricoltura, pp. 105. ISBN 978-88-7702-322-3.

Staudhammer, C.L., LeMay, V.M., 2001. *Introduction and evaluation of possible indices of stand structural diversity*. Canadian Journal of Forest Research. 31: 1105-1115. doi: 10.1139/cjfr-31-7-1105.

Suárez, J., Fonweban, J., Gardiner, B., 2012. *Supporting precision forestry in Great Britain*. Earthzine (http://www.earthzine.org/2012/06/22/supporting-precision-forestry-in-great-britain/).

Travaglini, D., Bottalico, F., Brundu, P., Chirici, G., Minari, E., 2007. *Testing neighborhood-based indices for monitoring spatial forest stand structure*. In: Gianelle, D., Travaglini, D., Mason, F., Minari, E., Chirici, G., Chemini, C. (Eds.), 2007. Canopy Analysis and Dynamics of a Floodplain Forest. Rapporti Scientifici, 3. Centro Nazionale per lo studio e la conservazione della Biodiversita' Forestale – Bosco della Fontana. Cierre Grafica Editore, Verona, pp. 96.

Venables, W.N., Ripley, B.D., 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Zhang, Z., Liu, X., Peterson, J., Wright, W., 2011. *Statistical analysis of airborne lidar data for forest classification in the Strzelecki Ranges, Victoria, Australia*. In: MODSIM 2011: Sustaining Our Future: Understanding and Living with Uncertainty, 12-16 Dec 2011, Perth, Australia.

Zimble, D.A., Evans, D.L., Carlson, G.C., Parker, R.C., Grado, S.C., Gerard, P.D., 2003. *Characterizing vertical forest structure using small-footprint airborne lidar*. Remote Sensing of Environment, 87:171-182. doi: 10.1016/S0034-4257(03)00139-1.

5 Classification trees to predict forest structural clusters through lidar metrics extracted from discrete return laser scanner data

In this chapter a study is reported which assesses whether metrics extracted from lidar raw point cloud can be exploited to discriminate different forest types by means of classification trees. A bivariate analysis by means of Pearson and Spearman statistical tests has been developed to find associations between lidar metrics, the proportion of basal area into three diameter size-classes (small, medium, and large trees) and forest structural indices of 243 random distributed plots surveyed from 2007 to 2012 within Trento province. An unsupervised clustering approach was adopted to determine forest structural patterns on the basis of basal area, using a k-means procedure combined with a previous hierarchical classification algorithm. A classification tree model to predict forest structural patterns originating from the cluster analysis has been developed and validated.

5.1 Introduction

Structural characteristics of forest stands influence many aspects of special interest in forest management: biodiversity, possibility to host specific volatile compounds, ecosystem processes, regeneration and gap dynamics, hydrology, and timber production.

Forest managers usually utilize structural classification for management purposes.

Basal area distribution into three diameter classes (small, medium, and large trees) is the criterion used by the Forest Service of Trento Province to classify the structure of forest stands (refer to § 4.2.2).

Remote sensing is a valuable source of information in mapping and monitoring forest features, and machine learning techniques, such as regression trees, artificial neural networks and support vector machines, have been used exploiting satellite imagery [*e.g.* Franklin *et al.*, 2002; Moghaddam *et al.*, 2002; Taškova *et al.*, 2006] and satellite imagery integrated with airborne laser scanner data [*e.g.* Lefsky *et al.*, 1999; Hudak *et al.*, 2002; Džeroski *et al.*, 2006].

In the use of remote sensing data for forestry applications, a typical machine learning task is to develop a predictive model which employs a set of remote sensing observations whose aim is to predict the value of forest conditions or properties usable for other cases. The data input to the machine learning system consists of information extracted from remote sensing data sources, while the output of the system is a predictive model that describes the forest property.

In this study we study the problem of forest structure prediction from lidar point cloud-derived metrics using classification trees as machine learning technique and investigate the performance of the developed tree model.

5.2 Materials

5.2.1 Study area

The study area for this research problem and its related research activities is the province of Trento, whose characteristics have been described in paragraph 4.2.1 (refer to that section for details).

5.2.2 Field data collection

The dendrometric data used in this study comprise a collection of data coming from four different sets.

The first data set is composed by the data already described in the paragraph 4.2.3, therefore for each detail refer to this section.

The second data set amounts to data collected during research activities whose aim was to investigate stand structure indices as tools to support forest management [Pastorella and Paletto, 2013]. Stratified sampling was chosen as method of sampling from the original population of forests of Trento province. By means of the stratification technique four strata were identified based on forest species composition: Norway spruce forests pure and mixed with silver fir, beech forests, Scots pine forests, and European larch forests. Afterwards only the areas with a minimum surface of 10 ha were selected, and finally a simple random sampling was applied to place 60 plots within the strata. In this way 30 plots were positioned in Norway spruce forests pure and mixed with silver fir, 12 plots in beech forests, 12 plots in Scots pine forests, and 6 plots in European larch forests. Field surveys were carried out during the maximum leaf-on growing season (June-September 2012), and due to the specific geomorphological condition, only 24 plots were surveyed in the Norway spruce forests pure and mixed with silver fir. The center of each circular plot was georeferenced with a Garmin eMap GPS receiver recording at minimum 100 satellite positions on the ground surface and accepting an estimated accuracy of 5 m. The randomly located sample plots were circular with a radius of 13 m (531 m^2 in surface). For all trees with dbh greater than or equal to 4.5 cm, the specie was recorded, the angle and the distance from the center of the plot, and the dbh were measured. The dbh allowed calculation of the field-based proportion of basal area on each plot in the three size-classes (small, medium and large) needed to classify the forest structure.

The third data set is made up by data collected in three specific forest areas of Trento province: Foresta Demaniale di Paneveggio (North East of Trento province), Padergnone municipality (central part of Trento province), and Val di Sella (South East of Trento province) where there are Norway spruce forests (pure and mixed with silver fir) and European beech forests, the most representative forest types from the specie composition and structure type point of view. In total 110 plots were randomly distributed over those three areas. The field surveys were carried out during the 2007 summer seasons. The randomly located sample plots were circular with a radius of 13 m (531 m^2 in surface) and 20 m (1257 m^2 in surface) depending on the stand density, using the smaller radius in full stocking stands. The nominal coordinates of plot center were reached using a navigation assisted by a GPS receiver according to the protocol adopted in the National Forest Inventory [Scrinzi and Floris, 2003]. With the parameters suggested by this protocol, positioning uncertainty (RMSE - 1σ) was estimated around 3 to 4 meters. The center of each circular plot was georeferenced with Thales MobileMapper CE and Trimble Geo-XT GPS receivers recording at minimum 170 satellite positions on the ground surface. Afterwards the geographical coordinates of plot centers were differentially corrected and averaged.

In each plot, tree dbh was measured (as average of two orthogonal axes) from a threshold of 2.5 cm using timber calipers. The dbh distribution was used to calculate the field-based proportion of basal area on each plot in the three diameter size-classes (small, medium and large) used as discriminant variable of forest structures.

The fourth data set amounted to field data collected within a post-doc research framework conducted by Andrea Lamonaca, aimed at applying techniques for the analysis of high spatial resolution satellite images for the structural monitoring of forest stands. Field data were surveyed during June and July of 2008 in a 4.8 ha wide forest located in Val di Sella. A full callipering was made within the forest, and stem position was measured by Haglöf® Vertex III system and a compass. The coordinates of the boundaries of the callipered forest, and the position of the points from which the surveyor conducted the survey were recorded by Trimble® ProXH GPS receiver. The geographic coordinates were corrected in post processing using Trimble® GPS Pathfinder® Office, a package of Global Navigation Satellite System post processing tools. Reference data for 2669 trees with dbh larger than 1 cm have been collected: the dbh of 1015 European beeches, and 101 other broad-leaved species, 808 fir trees, 674 Norway spruces, and 71 other conifer species were measured. The data collected were used to simulate circular plots with a radius of 20 m. To do a script in R was developed which using the following steps:

- produce in a random way within the forest callipered area a specified number of values of coordinates corresponding to the center of the plots;

- for each tree:
  - compute the distance between its position and the center of the plot;
  - if the distance is less than 20 m, compute the basal area of each tree and classifies it as the basal area of a small, medium or large tree according to the dbh value;
- for each plot:
  - sum up the values of basal area of small, medium and large trees, compute the total basal area of the plot, and compute the percentage of basal area of each tree size class;
  - save the results in a data frame (coordinates East and North of plot center, basal area of small, medium, and large trees, and the corresponding percentage).

In Figure 5.1 the location of plots corresponding to the four data sets is illustrated.

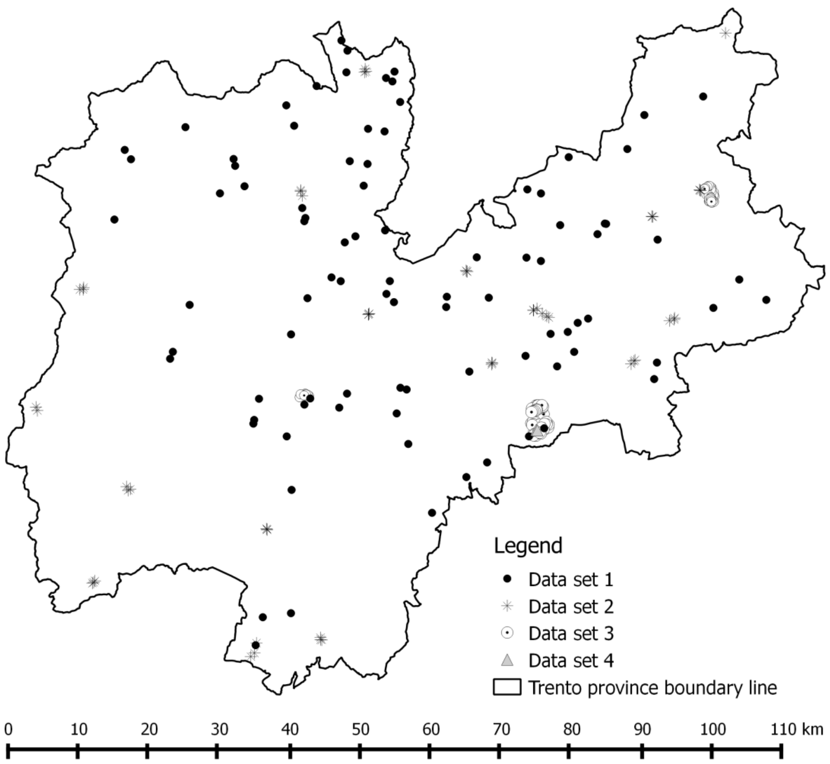

Figure 5.1 - Plot locations over the territory of Trento province differentiated in symbol according to the data set they belong to

5.2.3 Forest structural indices from field data

The forest stand structure was evaluated throughout variability and heterogeneity indices computed from the basal area values referred to each plot. The indices selected for our study were Shannon-Weaver [Shannon and Weaver, 1949], Gini [Ceriani and Verme, 2012], and an *ad-hoc* elaborated forest structural index, that we called Scrinzi index.

Shannon-Weaver index has been widely used in ecological studies and has proven to be a good indicator of the structural complexity of the forest [Kuuluvainen *et al.*, 1996].

Shannon-Weaver index based on the proportion of basal area in each diameter size-class [Liang *et al.*, 2007] was used to evaluate the tree-size diversity, as following

$$SW = -\sum_{i=1}^{k} p_i \ln p_i \qquad [5.1]$$

where i is the tree diameter class, k is the number of diameter classes (3 in our case corresponding to the classes of small, medium and large trees), and p_i is the proportion of basal area in each tree size class, *e.g.* if the percentages of basal area of small, medium, and large trees are 25%, 35%, and 40% respectively then the proportions are 0.25, 0.35, and 0.40 respectively. *SW* increases with the evenness of the distribution of basal area by diameter class. This means the higher the value, the higher the diversity (heterogeneity). In computing *SW*, adding 0.23 m^2/ha to p_i before natural logarithm transforming was necessary, due to basal area values = 0 at some plots lacking trees of some specific size. The value 0.23 m^2/ha is the second minimum value of basal area detectable from all three diameter classes [Agresti, 2013]. We normalized Shannon-Weaver index dividing it by the natural logarithm of the number of diameter size-classes [Ramezani, 2012]: in this way its value range between 0 (minimum value of forest structural complexity) and 1 (maximum value of forest structural complexity).

The Gini index was used to explain the heterogeneity of forest structure. It has been calculated as following:

$$G = \sum_{i=1}^{k} p_i (1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2 \qquad [5.2]$$

For our purposes, we normalized the Gini index by dividing its value by the maximum value that the index can achieve (0.66), in this way the value of Gini index ranged from 0 to 1 [Costantini, 2007]. The more the value of G tends to 0, the more the structure is homogeneous, meaning that heterogeneity is low. Similarly, the more the value of G tends to 1, the more the structure is heterogeneous.

An index was invented to evaluate the heterogeneity of the structure with respect to the repartition of the basal area in three diameter classes. We called that the Scrinzi index:

$$S = 1 - \frac{p_{\max} - p_{\min}}{\# classeswithp_i > 0.1} \qquad [5.3]$$

where p_{max} and p_{min} are the maximum and minimum value respectively of proportion of basal area in three diameter classes, and *# classes with* $p_i > 0.1$ is the number of diameter size-classes with the proportion of basal area bigger than 10%. *S* index describes how the basal area is distributed in the diameter size-classes. When S is equal to 0 this means that basal area is distributed in the diameter size-classes. When S is equal to 0 the basal area is concentrated in one class, and hence the forest stand has an homogeneous structure, while when S is equal to 1 the basal area is equally distributed in all diameter size-classes and so the forest stand has a maximal heterogeneous structure.

5.2.4 ALS data acquisition and delivering

The data used for this analysis were those acquired during the lidar campaign commissioned by the Province of Trento during the 2006-2007 autumn and winter seasons: the description of lidar data acquisition parameters has already been provided in § 3.5.2, so for any details refer to that section.

In 2013, the Province of Trento made available for the research purposes of this dissertation the raw lidar data throughout all its territory. Data were delivered in ASCII file format with extension .all, the textual version of the standard, publicly available lidar data exchange format called LAS, which extension is .las. The raw data are the set of all recorded points which have not undergone any transformation process. For each point, the GPS time, the return number (code 1 for first and code 9 for last), the East and North coordinates, the ellipsoidal height, the intensity, and the flight line number were recorded. Lidar data were provided in tiles of 2x2 km^2 without overlap.

The data set presented some limitations: return number and the point classification were not assigned according to the ASPR (American Society for Photogrammetry & Remote Sensing) LAS specifications. In particular in phase of data analysis, we discovered that some first returns were not first and vice versa some last were not last returns, and that code 9 was used for the classification of last returns. The former aspect represented a big limitation to the possibility to implement specific computation of lidar point cloud metrics based on the distinction between first and last returns. The second aspect has resulted in remarkable amount of data processing.

5.2.5 Classification, and normalization of lidar point clouds

Raw lidar data processing took a long time and implied the use of R, TerraScan (Terrasolid), and LASTOOLS (rapidlasso GmbH) software. The main steps of data processing, made for the only tiles in which the plots fell, are described in the following points:

- conversion of the .all file extension in .txt file extension;
- change the return number code 9 into 2 in the field about return number of .txt files: for this step a script in R was developed;
- conversion of lidar data from the ASCII format (file in .txt extension) into the binary LAS representations: *txt2las* command of LASTOOLS was used;

- setting the parameters of Axelsson [1999] filtering algorithm in TerraScan by numerous attempts: details about the values of parameters are provided below;
- run a macro in TerraScan to classify the points of the tiles according to the values of the parameters selected in the previous step;
- computation of the height of each LAS point above the ground: this was made using the *lasheight* command in LASTOOLS, which assumes that ground points have already been classified (classification == 2) so they can be identified and used to construct a ground Triangulated Irregular Network (TIN).

It seems relevant to describe in detail the classification process which was the most significant part of lidar data processing. Lidar point clouds need to be classified in order to extract the points belonging to ground, to building or to vegetated areas. Axelsson [1999] developed a filtering algorithm based on TIN densification that is now used in TerraScan software. The algorithm derives a TIN from the neighboring minima lidar points as a first approximation of the bare earth. Throughout iterative steps, the TIN is modified by adding other laser points that meet certain distance and orientation criteria in relation to the triangles that contains them [Sithole and Vosselman, 2003]. The vegetation and the buildings are classified using a cost function based on the second derivatives on the elevation differences [Axelsson, 1999]. The model assumes that the buildings consist of connected planar surfaces, so along a scan line the points belonging to a change of direction of the roof will cause a non-zero value of the second derivatives and will be classified as breakpoints, in the opposite case (zero value of the second derivatives) points will be classified as belonging to a straight line segment, vegetation instead is modeled as points with randomly distributed second derivatives.

TerraScan classification procedure provides different routines to perform the classification [Soininen, 2012]: the Axelsson [1999] algorithm is implemented in *Ground* routine. The routine performed in TerraScan and the values of parameters setting are shown below:

a. *by class* routine: by this all points in a given class are changed to another class. The setting was:

   From class All categories

   To class Default

b. *low points* routine: by this, points which are lower than a group of points in the vicinity are classified as low points. This routine will basically compare the elevation of each point (consider the center) with every other point within a given distance. If the center point is clearly lower than any other point, it will be classified. The setting was:

   From class Default

To class Low point

Search Groups of points

Max count 3

More than 0.5 m

Within 5.00 m

c. *ground* routine: by this, points are classified as ground points by iteratively building a triangulated surface model. The routine starts by selecting some local low points that are confident hits on the ground: the initial point selection is controlled with the *Max building size* parameter. If maximum building size is for example 5 m, the routine assumes that any 5 by 5 m area will have at least one hit on the ground and that the lowest point is a ground hit. The application builds an initial triangulated surface model from the selected low points. Triangles in this initial model are mostly below the ground with only the vertices touching ground. The routine then starts modeling the model by iteratively adding new laser points to it. Each added point makes the model following the ground surface more closely. Iteration parameters determine how close a point must be to a triangle plane for being accepted as ground point and added to the model. *Iteration angle* is the maximum angle between a point, its projection on triangle plane and the closest triangle vertex. *Iteration distance* parameter makes sure that the iteration does not make big jumps upwards when triangles are large (Figure 5.2). The smaller the *Iteration angle* (small undulations in terrain or hits on low vegetation), the less eager the routine is to follow changes in the point cloud. The use of a small angle (close to 4.0) is suggested in flat terrain while a bigger angle (close to 10.0) is suggested in hilly terrain.

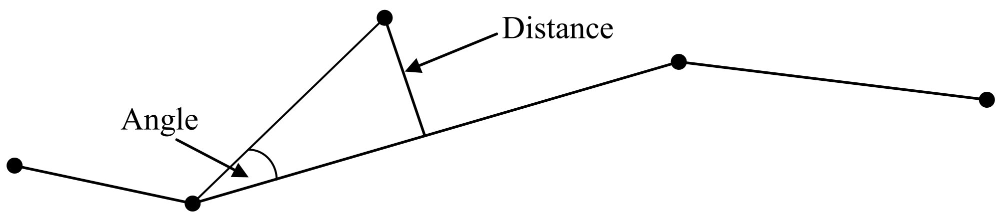

Figure 5.2 - Iteration parameters (angle and distance) determine whether a point should accepted as ground point and added to the triangulated surface model

The setting was:

From class Default

To class Ground

Select Default

Max building size 5 m

Terrain angle 80°

Iteration angle 8°

Iteration distance 2 m

d. *below surface* routine*:* points are classified in this routine which are lower than neighboring points in the source class. This routine can be run after ground classification to locate points which are below the true ground surface. The algorithm for each point, called central point, in the source class, finds up to 25 closest neighboring source points. It fits a plane equation to the neighboring points. If the central point is above the plane or less than *Z tolerance* below, it will not be classified. It computes standard deviation of the elevation differences from the neighboring points to the fitted plane. If the central point is more than *Limit* times standard deviation below the plane, it classifies it into the target class.

From classes Ground

To class Problems

Limit 8 * std. deviation

Z tolerance 0.1 m

e. *air points* routine: points are classified which are clearly higher than the median elevation of surrounding points. It can be used to classify noise up in the air. For each point, this routine will find all the neighboring points within a given search radius. It will compute the median elevation of the points and the standard deviation of the elevations. A point will be classified if it is more than *Limit* times standard deviation above the median elevation [Soininen, 2012].

The setting was:

From classes Default *To class* Problems

Search radius 5 m

Require 10 points

Limit 5 * std. deviation

When the classification process was finished, lidar returns were normalized using the *lasheight* command of LASTOOLS, which calculates the height of each point with respect to the ground surface: this tool performs a TIN using the lidar points classified as ground and then subtracts ground elevations from lidar elevations. Practically speaking, lidar heights are lidar point elevations minus ground elevations for the same positions. In this way we could have lidar height data ready for the next step, consisting in the computation of lidar point cloud statistics.

5.2.6 Computation of lidar point cloud metrics

Once the lidar data were normalized, the point data were processed to extract lidar metrics at the plot level: the plot level variables are statistics calculated on the lidar points falling within the spatial border corresponding to the boundary of each field plot. We were specifically interested in the 3D lidar point cloud data that represent measurements of forest canopy structure.

The statistics were computed on all return lidar heights above 2 m using freely-available FUSION software [McGaughey, 2012].

Lidar metrics considered in our analysis are described in Table 5.1.

Table 5.1 - Lidar metrics used in the analysis

Name of metric	Description
hmean	Mean of lidar heights intersecting plot
hmode	Mode of lidar heights intersecting plot
hstd	Standard deviation of lidar heights intersecting plot
hcv	Coefficient of variation of lidar heights intersecting plot
hiqdist	Interquartile distance of lidar heights intersecting plot
hske	Skewness of lidar heights intersecting plot
hkur	Kurtosis of lidar heights intersecting plot
haad	Average absolute deviation of lidar heights intersecting plot (=absolute difference between the mean height value and the height of the considered point)
hp01	Height at which 1% of lidar heights intersecting plot fall below
hp05	Height at which 5% of lidar heights intersecting plot fall below
hp50	Height at which 50% of lidar heights intersecting plot fall below (median of lidar height intersecting plot)
hp95	Height at which 95% of lidar heights intersecting plot fall below
hp99	Height at which 99% of lidar heights intersecting plot fall below
hp90hp10	Difference between height at which 90% of lidar heights intersecting plot fall below and height at which 10% of lidar heights intersecting plot fall below
crr	Canopy-relief ratio (ratio of the difference between the mean and the minimum of lidar heights intersecting plot over the difference between the maximum and the minimum of lidar heights intersecting plot)
vegcover	Canopy cover (ratio of the number of returns intersecting plot with height >=2 over the total number of returns intersecting plot)
percretabmean	Percentage of all returns above the mean of lidar heights intersecting plot
percretabmode	Percentage of all returns above the mode of lidar heights intersecting plot

Canopy relief ratio is a quantitative descriptor of the relative shape of the canopy from altimetry observation [Parker and Russ, 2004; Pike and Wilson, 1971], defined as mean height returns minus the minimum height divided by the maximum height minus the minimum height. This ratio reflects the degree to which canopy surfaces are in the upper (crr > 0.5) or in the lower (crr < 0.5) portions of the height range.

Within the lidar metrics that FUSION extracts, there is the total number of returns: this value allowed calculation of the number of lidar points per square meter in each plot and we use this value as threshold to accept a plot as suitable or not for analysis. For our study, we decided to consider only the plots with a density of at least 0.5 points/m^2: in this way 243 plots (out of the 256 available) were selected and used for the purposes of this study.

5.3 Methods

5.3.1 Overview of the methodological strategy

To achieve the goal of the study, first of all a bivariate analysis was performed to evaluate the relationship between lidar variables, the proportion of basal area of understory, mid-story, and overstory trees, and the heterogeneity indices. For this purpose, Pearson and Spearman correlation tests were applied using Bonferroni method to correct p-values for multiple testing.

To further investigate differences in forest structure, an unsupervised cluster analysis was performed to determine different patterns of forests based on the proportion of basal area in three classes, using a hierarchical classification algorithm based on a Ward [1963] criterion to assess the number of clusters to subsequently apply a k-means algorithm to improve the clustering. A comparison among the identified clusters was performed using the Kruskal-Wallis test.

To evaluate the ability of lidar metrics as predictors to explain different forest structures, *i.e.* the unsupervised forest structure patterns, classification trees were used.

The results of this supervised technique were then internally validated with the misclassification error to evaluate its global performance. Internal validation is important to obtain a sample estimate of the performance of the model using the same observations that have driven the model. However this estimation may be influenced by overfitting. In any case, internal validation can offer an upper limit to the expected performance in other settings. To overcome this, an external validation, which is the performance on a portion of observations used to develop the model, was accomplished splitting the general sample into two sub-samples, typically named the training set (75% of observations) and the validation set (25% of observations). A key characteristic is that data for model development and evaluation are both random samples from the same underlying population. The misclassification error computed when predicting the testing data using the results obtained by the model driven by the training data is a much fairer and reliable estimation of the true misclassification error.

The results of validation were categorized in a contingency table, reporting errors of commission and errors of omissions. Errors of commission are sometimes also called "false positives." They refer to the percentage of observations incorrectly classified. Error of omission are called also "false negative". They refer to percentage of observations that are missed in classification in the specific group.

All statistical analyses were performed using *stats* and *rpart* R packages.

An examination of the theoretical aspects of statistical tests used, cluster analysis and classification trees are reported in the sections below.

5.3.2 Bivariate analysis by means of Pearson and Spearman statistical tests

A correlation coefficient is a symmetric, scale-invariant measure of association between two random variables [Daalgard, 2008]. It ranges from -1 to +1, where the extremes indicate perfect correlation and 0 means no correlation. The value is negative when large values of one variable are associated with small values of the other (inverse relationship) and positive if both variables tend to be large or small simultaneously (proportional relationship).

The correlation coefficient is the statistical test whose probability distribution allows determination of the significance level of that coefficient. A specific value of the correlation coefficient has a bigger significance with a large number of observations (and hence of degrees of freedom). Increasing the number of observations, a correlation never will improve, simply it will boost the confidence interval we have in the statistics of the sample being a good estimator of the parameters of the population.

The correlation coefficients can be calculated with parametric and non-parametric methods. Spearman's rank correlation coefficient (or simply Spearman correlation coefficient) and Pearson product-moment correlation coefficient (or simply Pearson correlation coefficient) are non-parametric and parametric coefficient respectively.

The Spearman correlation coefficient is a non-parametric (distribution-free) rank statistic proposed by Charles Spearman. Spearman's coefficient is not a measure of the linear relationship between two variables: it assesses how well an arbitrary monotonic function can describe a relationship between two variables, without making any assumptions about the frequency distribution of the variables. It does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level [Hauke and Kossowski, 2011].

The Spearman correlation coefficient can be used with small samples. The minimum acceptable number of observations in this test is seven, under this value the dispersion could blind a true correlation or suggest one false.

This test classifies by rank the values of each variable in ascending order: the rank 1 is assigned to the smallest value, while the *n* rank is assigned to the biggest value (*n* is the number of observations). The true values of the variables are then distributed in ranks which become the basic data used in the test.

Spearman correlation coefficient is calculated as following [Fowler and Cohen, 1993]:

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} \qquad [5.4]$$

where d is the arithmetic difference between the ranks of two variables, and n is the number of observations.

It is worth pointing that if there are a large number of observations, it is almost impossible to avoid that some data having identical value (rank ties or value duplicated). Consequently their ranks will be equal or *ex-aequo* and they need a proper treatment. Identical values are assigned a rank equal to the median of the ranks they would have to assumed if they were not equal.

If the calculated values of r_s is bigger than its tabulated values at a specific levels of significance, than the hypothesis that the variables are correlated is demonstrated.

The Pearson correlation coefficient is a measure of the linear dependence between two variables. It is used with samples consisting of more than 30 observations randomly extracted, and when the normal distribution of the values of two variables is assumed, *i.e.* their joint probability distribution function is normal bivariate.

It is defined as the covariance of the two variables divided by the product of their standard deviation:

$$r_p = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{\sqrt{[n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2][n\sum_{i=1}^{n} y^2 - (\sum_{i=1}^{n} y)^2]}} \qquad [5.5]$$

where x and y are the values of two variables, and n is the number of observations.

The statistical significance of r_p can be directly read from the tables for the probability distribution of r_p, as it is done for the r_s.

If in the bivariate analysis we compare all available variables, we ought to correct for multiple testing. In fact, performing many tests will increase the probability of finding one of them to be significant and hence the p-value tends to be exaggerated. The Bonferroni correction is a common adjustment method, which is based on the fact that the probability of observing at least one of n events is less than the sum of the probabilities for each event. Thus, by multiplying the p-value by the number of comparisons, namely 18, we obtain a conservative test where the probability of a significant result is less than or equal to the formal significance level [Dalgaard, 2008]. If the product results a value bigger than 1, than the adjustment procedure sets the adjusted p-value to 1.

The *cor.test()* function of *stats* library [R Core Team, 2013] was used for testing association between paired predictor and response variables. This function provides an indication of whether the correlation is significantly different from 0, providing the p-value, the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

5.3.3 Cluster analysis

"*Cluster analysis is a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual*" [Everitt *et al.*, 2001].

Clustering is considered to be an "unsupervised" learning model because the observations in the data set do not contain a target (or response) variable to provide guidance as to correct response. Clusters are identified in such a way that the observations are as much as possible homogeneous inside the group and as much as possible inhomogeneous between groups. There are two types of clustering algorithms: partitive and hierarchical. In the former, to belong to a group is defined using the dissimilarity from one representative point (centroid) of the entire cluster. In the second, a hierarchy of partitions characterized by an increasing or decreasing number of groups is built. Hierarchical clustering has two forms: agglomerative (or aggregative) and divisive (or disaggregative). The agglomerative form is a bottom up approach, in the sense that each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The divisive form is a top down approach: in this case all observations start in one cluster, and splits are performed progressively as one moves down the hierarchy.

In the present study, as a first step of analysis, an agglomerative hierarchical algorithm was used to assess the number of clusters, subsequently a partitive clustering based on *k*-means algorithm was applied to improve the clustering.

In the hierarchical clustering, in order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of a measure of distance between pairs of observations, and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Some commonly used distance for hierarchical clustering are Euclidean, squared Euclidean, Manhattan, maximum, and Mahalanobis.

Most common and basic linkage criteria between two sets of observations are single linkage, complete linkage, average linkage, centroid linkage, and Ward.

In this study the Ward criterion was selected as linkage method. It is based on the optimal value of an objective function which is the error sum of squares [Murtagh and Legendre, 2011] which, in practice, minimizes the variance of the merge cluster.

The *hclust()* function of *stats* library was used for the first step of cluster analysis (agglomerative hierarchical clustering). This function recursively implements Ward linkage by a Lance-Williams algorithm [Cormack, 1971] which updates cluster distances at each step. In fact, at each step, it is necessary to optimize the objective function: the recursive formula simplifies finding the optimal pair of clusters to merge. By the Ward criterion in the *hclust()* function compact and spherical clusters are found [R Core Team, 2013].

By plotting the result obtained by the clustering process, a dendrogram is produced: it is a branching diagram that represents the relationship of similarity among group, in which each branch is called a clade and the terminal end of each clade is called a leaf, chunk or segment. The arrangement of the clades tells us which leaves are most similar to each other. The height of the branch points indicate how similar or different they are from each other: the greater the height, the greater the difference. The horizontal orientation of dendrogram is irrelevant.

The dendrogram was used to decide an optimal number of clusters to partition our data set, being k this number.

When this choice was made, a partitive clustering was carried out using the k-means method which is performed in *kmeans()* function of *stats* library in R. This function aims to partition the observations into k groups such that the sum of squares from observations to the assigned cluster centers is minimized. At the minimum, all cluster centers are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster center). The algorithm of Hartigan and Wong [1979] is used by default in *kmeans()* function.

5.3.4 Kruskal-Wallis test

Once the cluster analysis was performed, the non parametric Kruskal-Wallis test was used for testing whether the previously identified clusters originate from the same distribution. When the Kruskal-Wallis test leads to significant results, then at least one of the groups is different from the other groups. The test does not identify where the differences occur or how many differences actually occur. The Kruskal-Wallis test is most commonly used when there is one categorical variable and one measurement variable. The Kruskal-Wallis test does not make assumptions about normality, being a non-parametric method.

The Kruskal-Wallis test does not test the null hypothesis that the populations have identical means. It also does not test the null hypothesis that the populations have equal medians. It does assume that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes, the Kruskal-Wallis test may give inaccurate results [Fagerland and Sandvik, 2009].

Like most non-parametric tests, it is performed on ranked data, so the measurement observations are converted to their ranks in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on. Tied observations get average ranks; thus if there were four identical values occupying the fifth, sixth, seventh and eighth smallest places, all would get a rank of 6.5.

The *kruskal.test()* function of *stats* library was used to perform the Kruskal-Wallis rank sum test.

5.3.5 Classification and regression trees (CART)

Classification and regression trees, also known as recursive partitioning regression [Breiman *et al.*, 1984], are widely used in remote sensing applications [*e.g.* Friedl *et al.*, 1999; Hansen *et al.*, 2000; Huang and Townshend, 2003; Prasad *et al.*, 2006].

Trees subdivide the space spanned by predictor variables into regions for which the value of the response variable are approximately equal, and then estimate the response variable by a constant in each of these regions.

In the example reported in Figure 5.3, the first partition occurs when the predictor variable X_1 is equal to t_1 constant. The first region ($X_1 <= t_1$) is therefore partitioned with X_2 equal to t_2, and the second region ($X_1 > t_1$) with $X_1 = t_3$, and so successively. The result is the partition of the space in five regions.

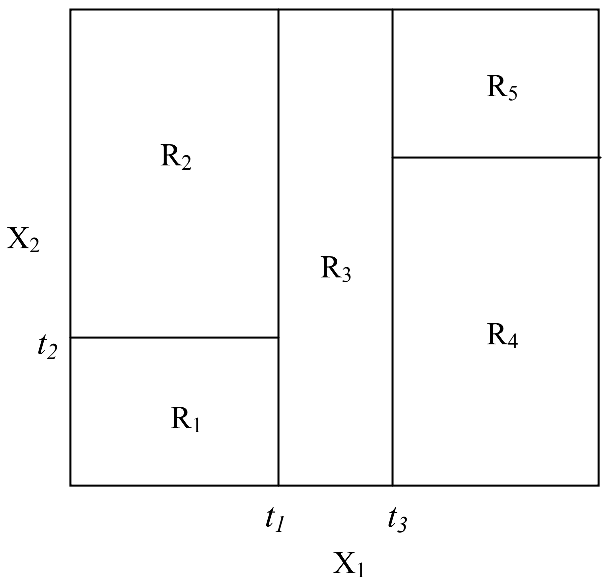

Figure 5.3 - Example of regions identification using recursive binary partitions

The prediction of response variable Y is done with a different constant in each region, which can be denoted as:

$$\hat{f}(X) = \sum_{m=1}^{n} c_m I\{(X_1, X_2) \in R_m\} \qquad [5.6]$$

where

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m}^{n} y_i \qquad [5.7]$$

with $N_m = \#\{x_i \in R_m\}$, n the number of regions, m the m-th region, and c the constant value in the region R_m.

The model is represented with a binary tree as in the Figure 5.4.

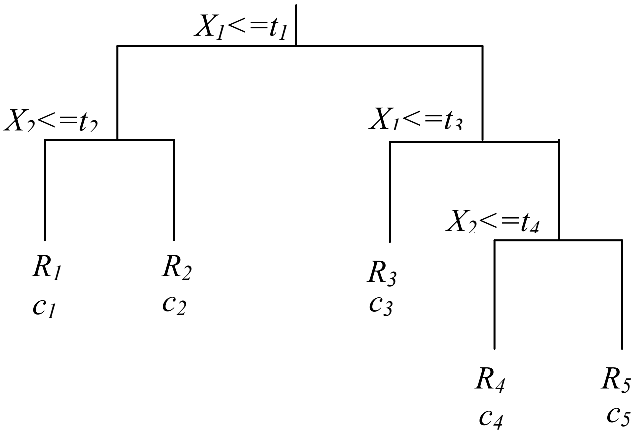

Figure 5.4 - Binary tree representing the partition in Figure 5.3

The tree is called a classification tree if the response variable is qualitative and a regression tree if the response variable is quantitative. In the case of classification trees the categorical response variable establishes a classification of observations on a number of levels. Partitions are identified with the particular goal of classifying observations into nodes, so that the levels (or classes) observed in each of the node are ideally the same, which corresponds to a condition of purity. For a categorical response variable with L levels, 2^{L-1}-1 different partitions can be considered. The purity of a node can be measured in different ways. Let n_{ik} the number of observations with class k in the terminal node i. Let p_{ik} be the observed proportion in the terminal node i of individuals with class k. The quantity D_i measures the purity of node i. The most used options to evaluate the purity of a node are the deviance, computed as following:

$$D_i = -2\sum_k n_{ik} \log(p_{ik}) \qquad [5.8]$$

the entropy (or information index), calculated as following:

$$D_i = -2\sum_k p_{ik} \log(p_{ik}) \qquad [5.9]$$

and the Gini index computed as following:

$$D_i = \sum_k p_{ik}(1 - p_{ik}) = 1 - \sum_k p_{ik}^2 \qquad [5.10]$$

D_i reaches the minimum value when all the observations in a node have the same class.

The criterion used to define the regions is based on searching for that partition that minimize the total purity index of the tree, computed as

$$\sum_i p_i D_i \qquad [5.11]$$

where p_i is the proportion of observations of the sample in the terminal node i. In practice the purity index of the tree is the sum over all terminal nodes of the purity of a node multiplied by the proportion of the observations that reach that node.

The *rpart()* function of *rpart* library [Therneau *et al.*, 2014] was used to built classification model and to predict the forest structural clusters identified by the cluster analysis considering the significant lidar metrics as predictors. In *rpart()* it is possible to control the maximum size of a tree, and so prune it, by setting a cost complexity parameter (cp). This parameter allows us to fix value at which a split must decrease the overall lack of fit. We set the complexity parameters at cp = 0.01.

5.4 Results

5.4.1 Statistical distribution of dendrometric data and forest structural indices

In Table 5.2 the statistical distribution of response variables about the 243 plots under investigation are reported.

Table 5.2 - Statistical distribution of percentage of basal area of small, medium and large trees and forest structural indices

	Minimum	Maximum	Mean	Standard deviation
Percentage of basal area of small trees	0.00	100	21.57	21.87
Percentage of basal area of medium trees	0.00	100	46.04	20.41
Percentage of basal area of large trees	0.00	97.53	32.39	25.41
Shannon-Weaver index	0.04	0.99	0.73	0.17
Gini index	0.00	0.98	0.72	0.19
Scrinzi index	0.00	0.95	0.73	0.18

5.4.2 The measure of association between lidar predictor variables and response variable

The result from the bivariate analysis between lidar variables and the proportion of basal area of understory, mid-story, and overstory trees are reported in Table 5.3. The mean lidar height intersecting plots is negatively correlated with the percentage of basal area of small trees: this means that large values of mean height are associated with small values of the proportion of basal area of understory trees. Conversely large values of mean height are associated with large values of the proportion of basal area of overstory trees.

Table 5.3 - Pearson and Spearman correlation coefficients (r_P and r_S respectively) with the associated Bonferroni corrected p-values (p-value$_{BC}$) between lidar metrics and proportion of basal area of understory (%G small trees), mid-story (%G medium trees), and overstory trees (%G large trees). Values in bold indicate p values <0.05

		%G small trees	%G medium trees	%G large trees
hmean	r_p (p-value$_{BC}$)	**-0.54 (0.0000)**	-0.13 (0.7205)	**0.57 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.64 (0.0000)**	-0.12 (0.9494)	**0.56 (0.0000)**
hmode	r_p (p-value$_{BC}$)	**-0.49 (0.0000)**	-0.09 (1.0000)	**0.49 (0.0000**
	r_s (p-value$_{BC}$)	**-0.63 (0.0000)**	-0.10 (1.0000)	**0.50 (0.0000**
hstd	r_p (p-value$_{BC}$)	**-0.68 (0.0000)**	**-0.21 (0.0164)**	**0.76 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.67 (0.0000)**	**-0.27 (0.0004)**	**0.77 (0.0000)**
hcv	r_p (p-value$_{BC}$)	**-0.28 (0.0002)**	-0.06 (1.0000)	**0.29(0.0001)**
	r_s (p-value$_{BC}$)	-0.15 (0.3445)	-0.08 (1.0000)	**0.33 (0.0000)**
hiqdist	r_p (p-value$_{BC}$)	**-0.60 (0.0000)**	**-0.20 (0.0364)**	**0.67 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.59 (0.0000)**	**-0.22 (0.0098)**	**0.70 (0.0000)**
hske	r_p (p-value$_{BC}$)	-0.02 (1.0000)	-0.06 (1.0000)	0.07 (1.0000)
	r_s (p-value$_{BC}$)	0.17 (0.1354)	-0.07 (1.0000)	0.08 (1.0000)
hkur	r_p (p-value$_{BC}$)	**0.30 (0.0000)**	-0.02 (1.0000)	**-0.24 (0.0034)**
	r_s (p-value$_{BC}$)	0.13 (0.8619)	0.06 (1.0000)	**-0.25 (0.0014)**
haad	r_p (p-value$_{BC}$)	**-0.66 (0.0000)**	**-0.21 (0.0205)**	**0.73 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.64 (0.0000)**	**-0.25 (0.0014)**	**0.75 (0.0000)**
hp01	r_p (p-value$_{BC}$)	0.04 (1.0000)	-0.06 (1.0000)	0.02 (1.0000)
	r_s (p-value$_{BC}$)	-0.04 (1.0000)	-0.06 (1.0000)	-0.02 (1.0000)
hp05	r_p (p-value$_{BC}$)	-0.15 (0.2981)	-0.04 (1.0000)	0.17 (0.1559)
	r_s (p-value$_{BC}$)	-0.22 (0.0121)	-0.02 (1.0000)	**0.10 (0.0000)**
hp50	r_p (p-value$_{BC}$)	**-0.54 (0.0000)**	-0.12 (1.0000)	**0.56 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.66 (0.0000)**	-0.12 (1.0000)	**0.55 (0.0000)**
hp95	r_p (p-value$_{BC}$)	**-0.66 (0.0000)**	**-0.20 (0.0387)**	**0.73 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.72 (0.0000)**	**-0.22 (0.0127)**	**0.72 (0.0000)**
hp99	r_p (p-value$_{BC}$)	**-0.66 (0.0000)**	**-0.20 (0.0280)**	**0.73 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.70 (0.0000)**	**-0.24 (0.0033)**	**0.73 (0.0000)**
hp90hp10	r_p (p-value$_{BC}$)	**-0.66 (0.0000)**	**-0.19 (0.0453)**	**0.72 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.62 (0.0000)**	**-0.25 (0.0016)**	**0.74 (0.0000)**
crr	r_p (p-value$_{BC}$)	0.00 (1.0000)	0.03 (1.0000)	-0.02 (1.0000)
	r_s (p-value$_{BC}$)	-0.18 (0.0892)	0.03 (1.0000)	-0.05 (1.0000)
vegcover	r_p (p-value$_{BC}$)	-0.10 (1.000)	-0.07 (1.0000)	0.14 (0.5272)
	r_s (p-value$_{BC}$)	-0.13 (0.6921)	-0.08 (1.0000)	0.16 (0.1902)
percretabmean	r_p (p-value$_{BC}$)	-0.11 (1.0000)	-0.03 (1.0000)	0.12 (1.000)
	r_s (p-value$_{BC}$)	-0.24 (0.022)	-0.02 (1.0000)	0.15 (0.2827)
percretabmode	r_p (p-value$_{BC}$)	-0.02 (1.0000)	-0.05 (1.0000)	0.05 (1.0000)
	r_s (p-value$_{BC}$)	0.06 (1.0000)	-0.04 (1.0000)	0.06 (1.0000)

The highest percentile (95th and 99th) and the difference between the 90th and the 10th percentile are significantly correlated with the proportion of basal area of understory, mid-story, and overstory trees. No variables related to canopy cover (vegetation cover, percentage of all return above the mean or the mode of lidar heights intersecting plots) were significantly correlated with the proportion of basal area of small, medium, and large trees.

The results from the bivariate analysis between lidar variables and Shannon-Weaver, Gini and Scrinzi indices are reported in Table 5.4 (values in bold indicate *p*-values < 0.05).

Table 5.4 – Pearson and Spearman correlation coefficients (r_P and r_S respectively) with the associated Bonferroni corrected p-values (p-value$_{BC}$) between lidar metrics and forest structural indices

		Shannon-Weaver index	Gini index	Scrinzi index
hmean	r_p (p-value$_{BC}$)	0.00 (1.0000)	0.02 (1.0000)	-0.06 (1.0000)
	r_s (p-value$_{BC}$)	0.01 (1.0000)	0.01 (1.0000)	-0.06 (1.0000)
hmode	r_p (p-value$_{BC}$)	-0.03 (1.0000)	-0.01 (1.0000)	-0.09 (1.0000)
	r_s (p-value$_{BC}$)	-0.02 (1.0000)	0.00 (1.0000)	-0.08 (1.0000)
hstd	r_p (p-value$_{BC}$)	**0.28 (0.0002)**	**0.24 (0.0022)**	0.17 (0.1661)
	r_s (p-value$_{BC}$)	**0.24 (0.0031)**	**0.21 (0.0227)**	0.18 (0.0710)
hcv	r_p (p-value$_{BC}$)	**0.32 (0.0000)**	**0.27 (0.0004)**	**0.27 (0.0004)**
	r_s (p-value$_{BC}$)	**0.34 (0.0000)**	**0.29 (0.0001)**	**0.36 (0.0000)**
hiqdist	r_p (p-value$_{BC}$)	**0.29 (0.0001)**	**0.24 (0.0021)**	0.19 (0.0568)
	r_s (p-value$_{BC}$)	**0.28 (0.0002)**	**0.23 (0.0050)**	**0.23 (0.0063)**
hske	r_p (p-value$_{BC}$)	**0.29 (0.0001)**	**0.25 (0.0013)**	**0.29 (0.0001)**
	r_s (p-value$_{BC}$)	**0.33 (0.0000)**	**0.28 (0.0002)**	**0.35 (0.0000)**
hkur	r_p (p-value$_{BC}$)	**-0.24 (0.0026)**	-0.19 (0.0526)	**0.20 (0.0351)**
	r_s (p-value$_{BC}$)	**-0.26 (0.0008)**	**-0.20 (0.0338)**	**0.25 (0.0012)**
haad	r_p (p-value$_{BC}$)	**0.29 (0.0001)**	**0.25 (0.0020)**	0.17 (0.1139)
	r_s (p-value$_{BC}$)	**0.26 (0.0010)**	**0.22 (0.0117)**	0.20 (0.0277)
hp01	r_p (p-value$_{BC}$)	-0.13 (0.6917)	-0.11 (1.0000)	0.10 (1.0000)
	r_s (p-value$_{BC}$)	-0.18(0.0913)	-0.16 (0.2172)	**0.20 (0.0239)**
hp05	r_p (p-value$_{BC}$)	-0.16 (0.2589)	-0.12 (1.0000)	0.14 (0.5374)
	r_s (p-value$_{BC}$)	-0.15 (0.3312)	-0.13 (0.9044)	**0.20 (0.0399)**
hp50	r_p (p-value$_{BC}$)	-0.01 (1.0000)	0.00 (1.0000)	0.08 (1.0000)
	r_s (p-value$_{BC}$)	-0.01 (1.0000)	0.00 (1.0000)	0.08 (1.0000)
hp95	r_p (p-value$_{BC}$)	0.15 (0.3909)	0.14 (0.5482)	-0.06 (1.0000)
	r_s (p-value$_{BC}$)	0.12(1.0000)	0.11 (1.0000)	-0.05 (1.0000)
hp99	r_p (p-value$_{BC}$)	**0.20 (0.0381)**	0.18 (0.0763)	-0.10 (1.0000)
	r_s (p-value$_{BC}$)	0.16 (0.2319)	0.14 (0.4509)	-0.09 (1.0000)
hp90hp10	r_p (p-value$_{BC}$)	**0.29 (0.0001)**	**0.25 (0.0013)**	-0.18 (0.0768)
	r_s (p-value$_{BC}$)	**0.26 (0.0008)**	**0.22 (0.0108)**	**-0.21 (0.0177)**
crr	r_p (p-value$_{BC}$)	**-0.32 (0.0000)**	**-0.28 (0.0002)**	**0.31 (0.0000)**
	r_s (p-value$_{BC}$)	**-0.34 (0.0000)**	**-0.29 (0.0001)**	**0.37 (0.0000)**
vegcover	r_p (p-value$_{BC}$)	0.06 (1.0000)	0.06 (1.0000)	-0.06 (1.0000)
	r_s (p-value$_{BC}$)	0.07 (1.0000)	0.06 (1.0000)	-0.07 (1.0000)
percretabmean	r_p (p-value$_{BC}$)	-0.04 (1.0000)	-0.03 (1.0000)	0.05 (1.0000)
	r_s (p-value$_{BC}$)	-0.03 (1.0000)	-0.03 (1.0000)	0.05 (1.0000)
percretabmode	r_p (p-value$_{BC}$)	0.12 (1.0000)	0.11 (1.0000)	-0.14 (1.0000)
	r_s (p-value$_{BC}$)	0.10 (1.0000)	0.07 (1.0000)	-0.11 (1.0000)

When the Spearman correlation coefficient is considered, out of 18 lidar variables, there were 8 significant correlations with the *SW* index, and *G* index (hstd, hcv, hiqdist, hske, hkur, haad, hp90hp10, and crr), and 7 significant correlations with *S* index (hcv, hiqdist, hske, hku, haad, hp01, hp05, hp90hp10, and crr). Only *S* index resulted in significant correlations with percentiles and, in particular, with the lowest percentiles (hp01, and hp05). This means that the *S* index is able to provide a higher representation of the component of small trees with respect to the other indices.

The coefficient of variation was positively correlated with *S* index, which means that if the dispersion of lidar heights around the mean value is large, then *S* index assumes large values (it tends to 1), and so the forest structure of the stand is considered heterogeneous.

The canopy relief ratio is significantly negatively correlated with both *SW* and G indices: this means that forests with canopy surfaces in the upper portions of the height range are associated with small values of *SW* and *G* indices, and consequently we can say that high values of crr are correlated with forest structure homogeneous. Conversely forest stands with canopy surfaces in the lower portions of the height range (small values of crr) are associated with heterogeneous or unevenness (large values of *SW* and *G* indices) in forest structure.

Moreover, we can say that large values of standard deviation of height are associated with large values of *SW* and *G* indices which are representative of a heterogeneous forest structure.

5.4.3 The characteristics of the forest structure patterns

The results coming from the agglomerative hierarchical clustering allowed identification of 5 clusters of forest structure. Consequently the application of *k*-means algorithm with $k = 5$ in the partitive clustering provides an optimization of this classification. According to the results, we classified our plots in pole-stage forests, young forests, adult forests, mature forests, and old forests. Figure 5.5 shows the dendrogram about the forest structure patterns identified by the cluster analysis after applying *k*-means algorithm. Most parts of the forests in Trento province are old forests (30.0%), while the lower part is represented by pole-stage forests (7.0%). Young forests represent almost 15% of all forest patterns, and adult and mature forests are equally distributed (24.3 %).

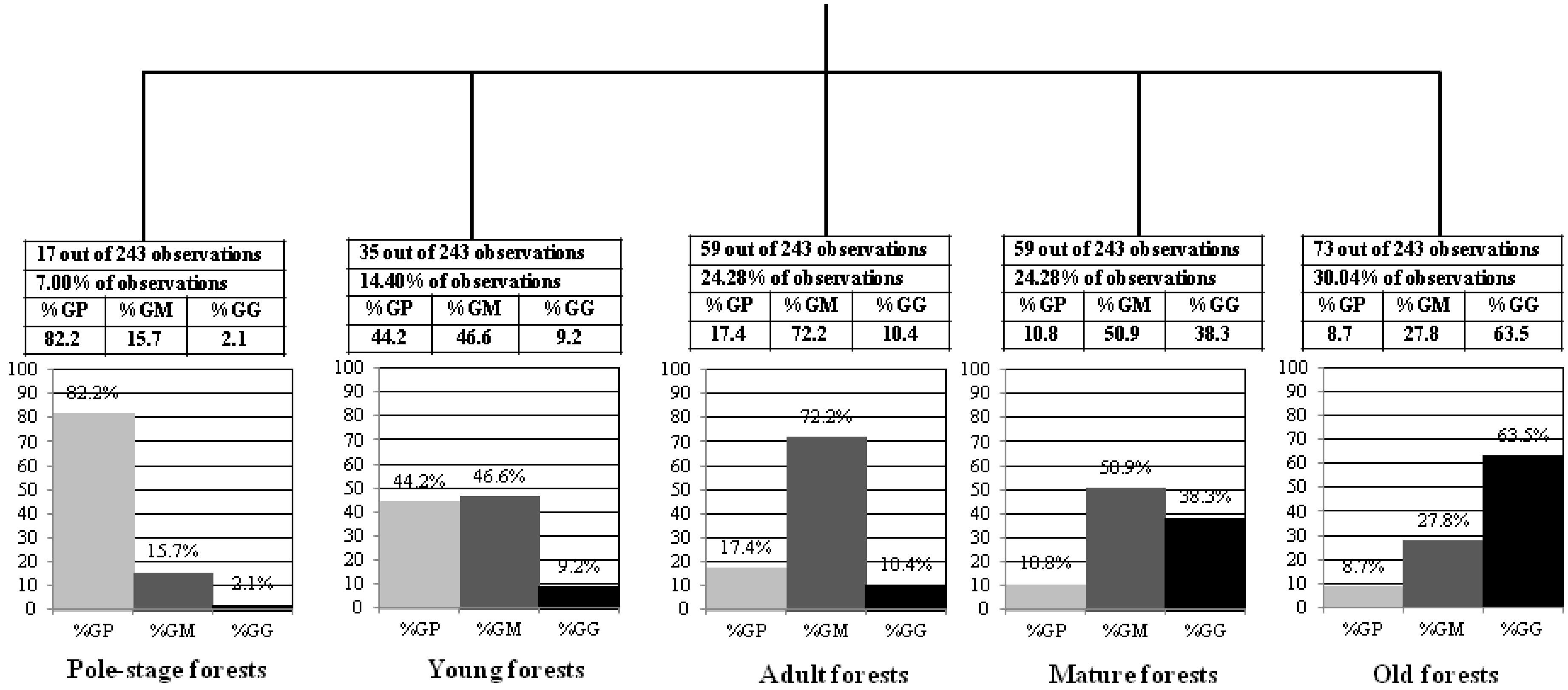

Figure 5.5 – Dendrogram resulting from the cluster analysis with details about the five forest structure patterns

The results from the non-parametric Kruskal-Wallis tests together with the average values of lidar metrics for the five forest structure patterns are reported in Table 5.5 and also illustrated through bar charts in Figure 5.6.

Pole-stage forests have mean height of 13 m, standard deviation of height around 2.6 m, height at the 99th percentile of 17.8 m. Most trees are small, usually one cohort is present, tree diameters are not large yet. New individuals are excluded trough light or underground competition. Canopy in this forest pattern is continuous closed.

Young forests are taller than pole-stage forests (mean height of 14 m), standard deviation of height is around 4 m, and height at the 99th percentile around 22 m. In this forest pattern initiation of new cohorts as older cohort occupies less than full growing space. Overstory may be medium or small trees.

In adult forests medium trees are prevalent, the mean height is around 16.3 m, standard deviation is 4.6 m and height at the 99th percentile is around 25 m. Two cohorts are present through establishment after periodic harvest events. Assortment of tree sizes and canopy strata present but large trees are still lacking.

Mature forests have mean height of 16 m, with standard deviation of height around 6 m, height at the 99th percentile of 30 m. Two or more cohorts present, with medium trees best represented. This pattern corresponds to multi-aged forests with assortment of tree sizes, including large trees.

Old forests have mean height of 21 m, standard deviation of height around 7 m, height at the 99th percentile of 33.5 m. They are single stratum of medium to large trees of one or more cohorts. Understory consists of small trees.

Out of 18, 11 lidar variables (hmean, hmode, hstd, hcv, hiqdist, hkur, haad, hp50, hp95, hp99, and hp90hp10) confirmed that at least one of the five patterns is different from the other patterns.

Unfortunately canopy cover predictor variable was not useful to identify characteristics of canopy closure, due to the fact that Kruskal-Wallis test lead to at least one of the five patterns being different from the other groups.

Table 5.5 – Average values of lidar metrics for the 5 forest structure patterns and p-values from the Krustal-Wallis test

	All observations	**Pole-stage forests (G1)**	**Young forests (G2)**	**Adult forests (G3)**	**Mature forests (G4)**	**Old forests (G5)**	
	Mean (St. dev.)	Mean (St. dev.)	Mean (St. dev.)	Mean (St. dev.)	Mean (St. dev.)	Mean (St. dev.)	p-value
hmean	17.92 (5.25)	12.99 (3.08)	14.15 (4.01)	16.34 (4.41)	19.12 (4.3)	21.18 (5.07)	**0.0000**
hmode	18.91 (6.66)	13.85 (3.25)	14.59 (5.15)	16.87 (5.82)	20.9 (5.54)	22.18 (7.05)	**0.0000**
hstd	5.27 (1.67)	2.58 (0.85)	3.77 (1.14)	4.61 (1.18)	5.82 (1.01)	6.69 (1.1)	**0.0000**
hcv	0.3 (0.09)	0.2 (0.06)	0.28 (0.1)	0.3 (0.09)	0.32 (0.08)	0.33 (0.09)	**0.0000**
hiqdist	7.12 (2.85)	3.18 (1.22)	4.71 (1.92)	6.11 (2.07)	7.95 (2)	9.33 (2.37)	**0.0000**
hske	-0.49 (0.6)	-0.82 (0.53)	-0.43 (0.88)	-0.51 (0.58)	-0.51 (0.51)	-0.42 (0.53)	0.1925
hkur	3.52 (1.57)	4.89 (2.39)	4.32 (2.14)	3.48 (1.23)	3.22 (1.33)	3.1 (1.08)	**0.0003**
haad	4.21 (1.44)	1.99 (0.7)	2.92 (0.98)	3.66 (1.02)	4.67 (0.93)	5.41 (1.02)	**0.0000**
hp01	4.66 (2.43)	5.87 (2.51)	4.46 (2.19)	4.58 (2.51)	4.4 (2.35)	4.76 (2.51)	0.1024
hp05	8.43 (4.04)	8.34 (3.15)	7.56 (3.44)	7.95 (3.67)	8.49 (4.03)	9.2 (4.68)	0.4848
hp50	18.4 (5.61)	13.34 (3.13)	14.32 (4.4)	16.77 (4.72)	19.73 (4.55)	21.76 (5.52)	**0.0000**
hp95	25.63 (6.62)	16.58 (3.69)	19.82 (4.58)	23.08 (4.99)	27.55 (4.54)	31.02 (4.89)	**0.0000**
hp99	27.88 (6.94)	17.82 (4.05)	22.04 (5.19)	25.2 (5.19)	29.91 (4.59)	33.54 (4.95)	**0.0000**
hp90hp10	13.49 (4.79)	6.00 (2.68)	9.24 (3.13)	11.78 (3.53)	15.03 (3.15)	17.42 (3.29)	**0.0000**
crr	0.56 (0.1)	0.61 (0.09)	0.53 (0.13)	0.56 (0.1)	0.55 (0.08)	0.55 (0.1)	0.2007
vegcover	0.68 (0.13)	0.67 (0.17)	0.66 (0.17)	0.66 (0.12)	0.7 (0.11)	0.69 (0.12)	0.2780
percretabmean	36.57 (7.6)	37.47 (10.08)	34.87 (10.24)	35.49 (6.79)	37.9 (5.9)	36.97 (7.28)	0.2196
percretabmode	31.25 (11.9)	27.38 (6.38)	31.95 (12.52)	31.73 (13.58)	29.59 (10.81)	32.77 (11.89)	0.4815

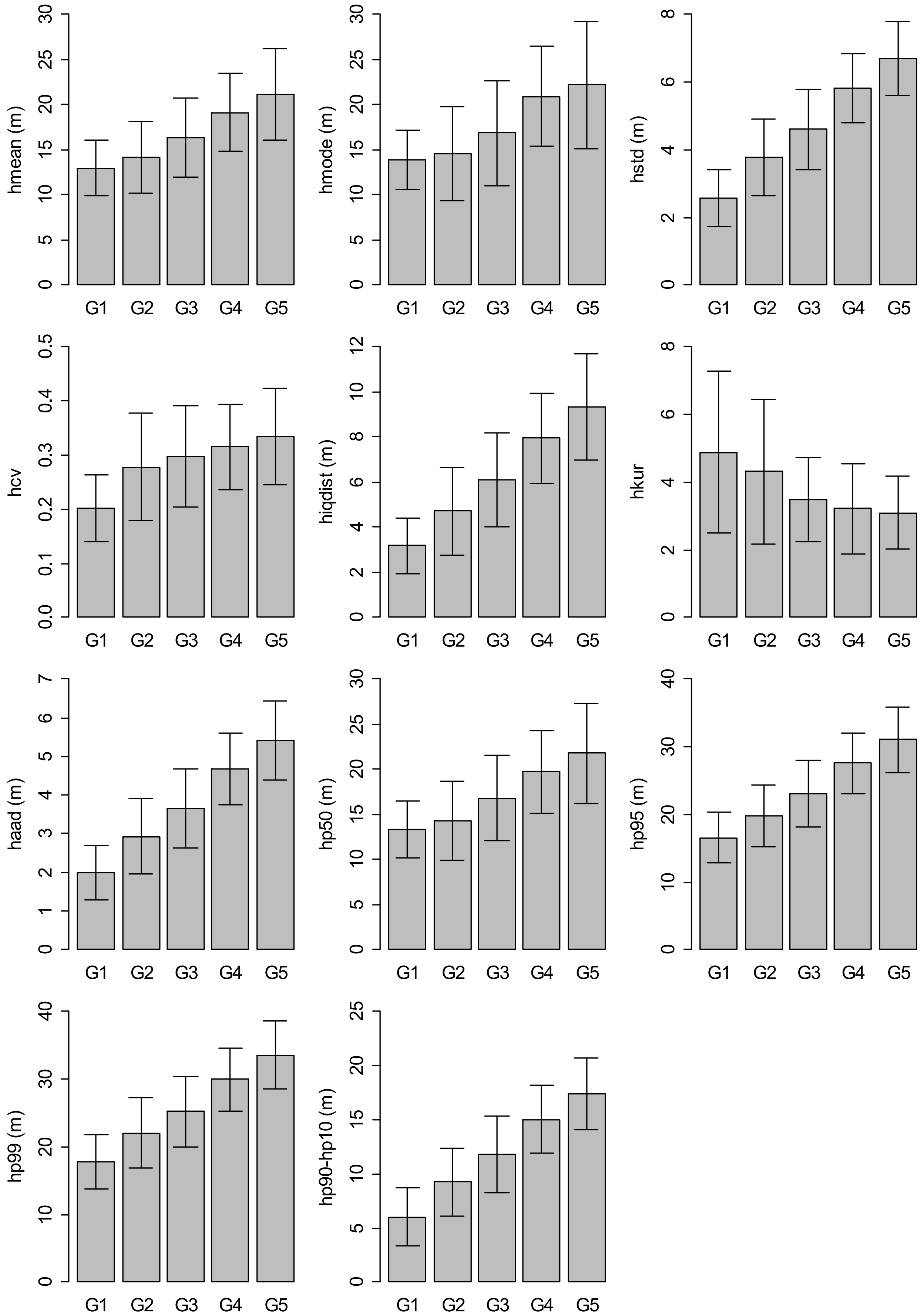

Figure 5.6 - Average values with standard deviation of 11 significant lidar metrics for 5 forest structure types

5.4.4 Classification trees to predict forest structure

Five ALS metrics were selected by the classification tree to predict the forest structural types: standard deviation and mode of canopy heights, height at which 95% and 99% of canopy heights fall below, difference between height at which 90% and 10% of canopy heights fall below.

The structure of classification tree to predict forest patterns is showed in Figure 5.7: the values in the rectangles show the number and percentage of observations in the node.

The tree model provided classification rules to discriminate between the five forest patterns which are synthesized in the following lines.

A stand is a pole-stage forest (G1):

- if standard deviation < 2.3 m.

A stand is a young forest (G2):

- if 2.3 m < standard deviation < 4.3 m and height at 95th percentile < 15.0 m;
- if 2.3 m < standard deviation < 4.3 and height at 99th percentile < 22.4 m.

A stand is an adult forest (G3):

- if 2.3 m < standard deviation < 5.1 m and mode >= 20.3 m;
- if 2.3 m < standard deviation < 5.1 m and mode < 20.3 m and difference between 90th and 10th percentile >= 12.2 m;
- if 2.3 m < standard deviation < 4.3 m and 95th percentile > 15.0 m and 99th percentile < 22.4 m;
- if 5.1 m < standard deviation < 7.0 m and 95th percentile < 24.7 m and difference between 90th and 10th percentile >= 16.0 m.

A stand is mature forest (G4):

- if 2.3 m < standard deviation < 4.3 m and mode < is 20.3 m and difference between 90th and 10th percentile < 12.2 m;
- if 5.1 m <= standard deviation < 7.0 m and 95th percentile < 24.8 m;
- if 5.1 m <= standard deviation < 7.0 m and 95th percentile < 24.8 m and difference between 90^{th} and 10^{th} percentile < 16.0 m.

A stand is an old forest (G5):

- if standard deviation >= 5.1 m and height at 95th percentile >= 29.0 m;
- if standard deviation >= 7.0 m and height at 95th percentile < 29.0 m.

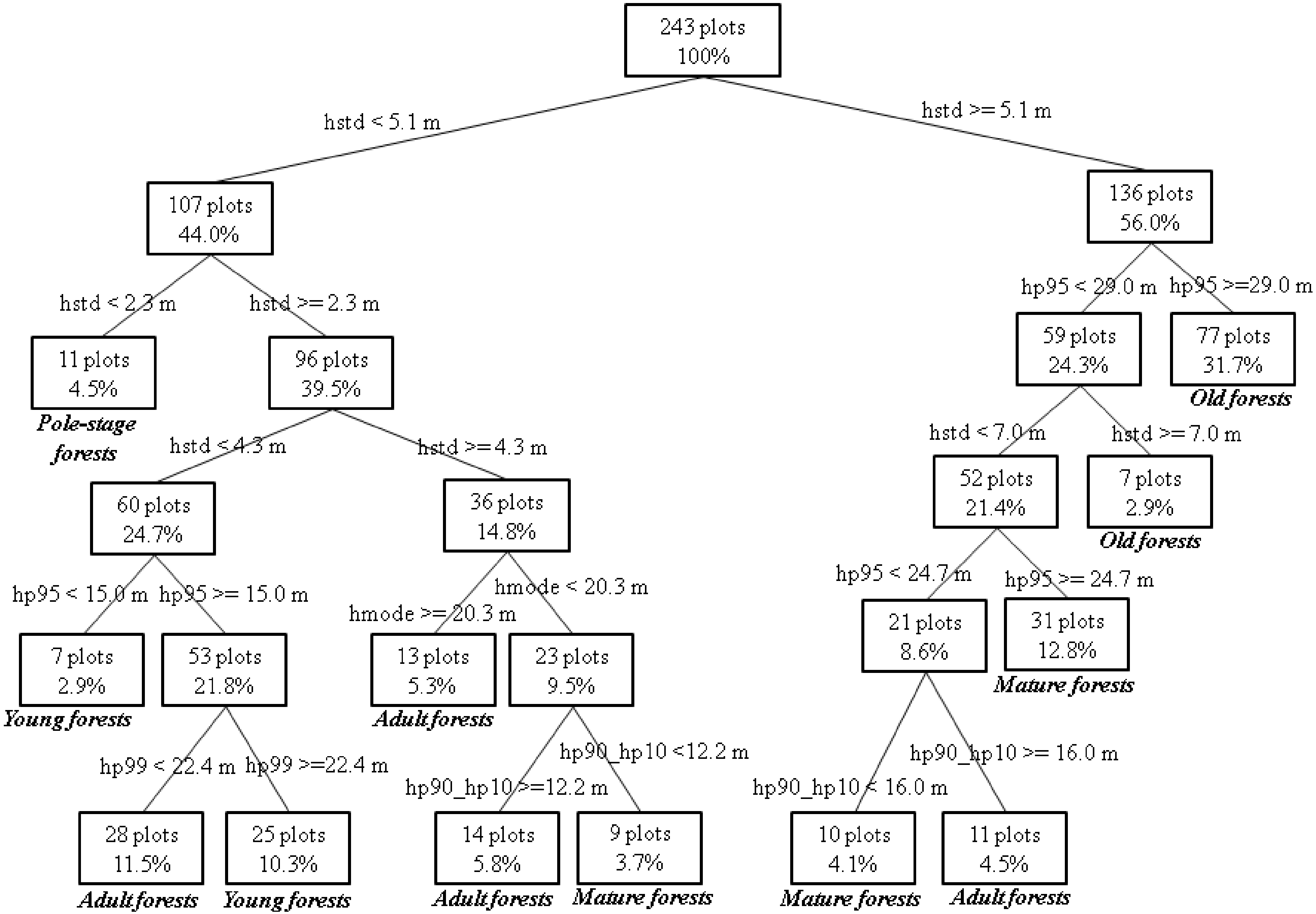

Figure 5.7 - Classification tree predicting forest patterns

Table 5.6 reports the results of the internal validation of tree model. The basal error or the error for the null model, calculated as difference between 1 and the forest structure pattern most represented, reached 70.0%. The misclassification error reached 36.2%. Omission error ranged between 18.2% (pole-stage forests) and 44.0% (mature forests), while commission error between 19.2% (old forests) and 52.5% (mature forests). There are no cases in which old forests have been classified as pole-stage or young forests and vice versa.

Table 5.6 – Contingency table resulting from the classification of five forest structural pattern by the classification tree (internal validation on all observations)

		Predicted							
		Pole-stage forests	Young forests	Adult forests	Mature forests	Old forests	Total	User accuracy	Commission error
Observed	Pole-stage forests	**9**	2	5	1	0	17	52.9%	47.1%
	Young forests	2	**19**	9	5	0	35	54.3%	45.7%
	Adult forests	0	8	**40**	7	4	59	67.8%	32.2%
	Mature forests	0	3	7	**28**	21	59	47.5%	52.5%
	Old forests	0	0	5	9	**59**	73	80.8%	19.2%
Total		11	32	66	50	84	**243**		
Producer accuracy		81.8%	59.4%	60.6%	56.0%	70.2%	Total accuracy		63.8%
Omission error		18.2%	40.6%	39.4%	44.0%	29.8%	Misclassification error		36.2%

Table 5.7 and Table 5.8 report the results of the validation of tree model in the training and validation data set respectively. The misclassification error reached 45.9% in the validation data set. No pole-stage and young forests were classified as old forests, while 2 out of 17 cases of adult forests were classified as old forests (11.8%), and 2 out of 12 cases of mature forests were classified as old forests (16.7%), 1 out of 51 cases of old forests was classified as young forest (2.0%).

Table 5.7 – Contingency table of external validation resulting from the classification of five forest structural pattern by the classification tree (training set made up with 75% of observations)

		Predicted							
		Pole-stage forests	Young forests	Adult forests	Mature forests	Old forests	Total	User accuracy	Commission error
Observed	Pole-stage forests	**7**	1	6	0	0	14	50.0%	50.0%
	Young forests	1	**20**	4	3	0	28	71.4%	28.6%
	Adult forests	0	7	**24**	5	6	42	57.1%	42.9%
	Mature forests	0	1	10	**29**	7	47	61.7%	38.3%
	Old forests	0	1	2	11	**37**	51	72.5%	27.5%
Total		8	30	46	48	50	182		
Producer accuracy		87.5%	66.7%	52.2%	60.4%	74.0%	Total accuracy		64.3%
Omission error		12.5%	33.3%	47.8%	39.6%	26.0%	Misclassification error		35.7%

Table 5.8 – Contingency table of external validation resulting from the classification of five forest structural pattern by the classification tree (validation set made up with 25% of observations)

		Predicted					**Total**	**User accuracy**	**Commission error**
		Pole-stage forests	**Young forests**	**Adult forests**	**Mature forests**	**Old forests**			
Observed	**Pole-stage forests**	**1**	2	0	0	0	3	33.3%	66.7%
	Young forests	0	**4**	2	1	0	7	57.1%	42.9%
	Adult forests	0	1	**7**	7	2	17	41.2%	58.8%
	Mature forests	0	0	2	**8**	2	12	66.7%	33.3%
	Old forests	0	1	0	8	**13**	22	59.1%	40.9%
Total		1	8	11	24	17	61		
Producer accuracy		100.0%	50.0%	63.6%	33.3%	76.5%	**Total accuracy**		54.1%
Omission error		0.0%	50.0%	36.4%	66.7%	23.5%	**Misclassification error**		45.9%

5.5 Discussions

A proper understanding of forest structure is one of the keys to the sustainable management of forests. The description and mapping of forest structure are important aspects for forest inventory purposes (*e.g.* in the stratification process of a forest estate), for individualization of silvicultural interventions needed to achieve a diversified landscape, or to drive a specific forest structure toward desired conditions.

Forest technicians often deal with processes of discrimination between different types of forest structures for forests management purposes. Manual activities, such as photo interpretation or fieldwork, traditionally used for forest structure interpretation are time-consuming and expensive. The results of this study suggested that airborne laser scanner data may be a useful information source to predict forest structure patterns.

This study suggests that the proportion of basal area in diameter size-classes is a valuable parameter to distinguish different forest structural clusters and that lidar variables selected for this study were able to confirm that at least one of the five forest structure patterns identified by the clustering analysis is different from the other patterns.

We were not surprised that among ALS-derived metrics canopy cover variables were not selected as significant variable in expressing differences among groups. In fact, as we already explained raw lidar data provided by the vendor and used for this study showed some limitations: return number and the point classification were not assigned according to ASPR LAS specifications, these aspects resulted in non-reliable values of canopy cover.

The classification tree model built in this study provided moderately satisfactory results in term of classification performance. A misclassification error around 46% in the validation set is reasonably acceptable when the model is used for preliminary analysis, but still not trusted for decision making purposes. It is worth highlighting that the model does not produce rough errors, in the sense that no pole-stage or young forest were classified as old forests or vice versa.

The representation of the classification tree model obtained through this study is clearly and easily interpretable; forest technicians can use classification trees in an operational context.

5.6 References

Costantini, P., 2007. *Esercitazione di Statistica 1 del 9 novembre 2007*. Università di Cassino.

Ramezani, H., 2012. *A note on the normalized definition of Shannon's diversity index in landscape pattern analysis*. Environmental and Natural Resources, Vol. 2, No 4, 2012, pp. 54-60. ISSN 1927-0488, E-ISSN 1927-0496. Published by Canadian Center of Science and Education. doi: 10.5539/enrr.v2n4p54

Agresti, A., 2013. *Categorical Data Analysis*. 3rd Edition, pp. 744. ISBN: 978-0-470-46363-5.

Axelsson, P., 1999. *Processing of laser scanner data - algorithms and applications*. ISPRS J. Photogramm. Remote Sens. 54(2): 138-147. doi: 10.1016/S0924-2716(99)00008-8.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.I., 1984. *Classification and regression trees*. Belmont, CA: Wadsworth International Group, 1984.

Ceriani, L., Verme, P., 2012. *The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini*. The Journal of Economic Inequality. September 2012, Volume 10, Issue 3, pp 419-420.

Cormack, R.M., 1971. *A Review of Classification*. Journal of the Royal Statistical Society, Series A, 134(3): 321-367.

Dalgaard, P., 2008. *Introductory Statistics with R*. Springer Ed.. 2nd ed. 2008, XVI, pp. 364.

Džeroski, S., Kobler, A., Gjorgjioski, V., Panov, P., 2006. *Using decision trees to predict forest stand height and canopy cover from Landsat and lidar data*. In: Tochtermann, K., Scharl, A. (Eds.), Managing Environmental Knowledge: EnviroInfo 2006: Proceedings of the 20th International Conference on Informatics for Environmental Protection. Aachen: Shaker Verlag, Graz, Austria, pp. 125–133.

Everitt, B.S., Landau, S., Leese, M., 2001. *Cluster Analysis*. 4th Edition, Oxford University Press, Inc., New York; Arnold, London. ISBN 0340761199

Fagerland, M.W., Sandvik, L., 2009. *The Wilcoxon-Mann-Whitney test under scrutiny*. Statist. Med. 28 (2009), pp. 1487-1497.

Fowler, J., Cohen, L., 1993. *Statistica per ornitologi e naturalisti*. Edizioni F. Muzzio, Padova.

Franklin, S., Wulder, M., 2002. *Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas*. Progress in Physical Geography 26(2): 173-205.

Friedl, M.A., Brodley, C.E., Strahler, A.H., 1999. *Maximizing land cover classification accuracies produced by decision trees at continental to global scales*. IEEE Transactions on Geoscience and Remote Sensing, 37, 969-977.

Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., 2000. *Global land cover classification at 1 km spatial resolution using a classification tree approach*. International Journal of Remote Sensing 21, 1331-1364.

Hartigan, J. A.,Wong, M. A., 1979. *A K-means clustering algorithm*. Applied Statistics, 28: 100-108.

Hauke, J., Zkossowski T., 2011. *Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data*. Quaestiones Geographicae 30(2), 2011, pp. 87-93.

Huang, C., Townshend, J., 2003. *A stepwise regression tree for nonlinear approximation: Applications to estimating subpixel land cover*. International Journal of Remote Sensing, 24(1), pp. 75-90.

Hudak, A., Lefsky, M., Cohen, W., Berterretche, M., 2002. *Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height*. Remote Sensing of Environment 82(2), 397-416.

Kuuluvainen, T., Penttinen, A., Leionen, K., Nygren M., 1996. *Statistical opportunities for comparing stand structural heterogeneity in managed and primeval forests: an example from boreal Spruce forest in Southern Finland*. Silva Fenn. 30: 315-328.

Lefsky, M.A., Cohen, W.B., Hudak, A., Acker, S.A., Ohmann, J.L., 1999. *Integration of LIDAR, Landsat ETM+ and forest inventory data for regional forest mapping*. Proceedings of the ISPRS Workshop Mapping Surface structure and topography by airborne and spaceborne lasers: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXII-3/W14.

Liang, J., Buongiorno, J., Monserud, R.A., Kruger, E.L., Zhou., M., 2007. *Effects of Diversity of Tree Species and Size on Forest Basal Area Growth, Recruitment, and Mortality*. Forest Ecology and Management, 243: 116-127.

McGaughey, R.J. 2012. FUSION/LDV: Software for LIDAR Data Analysis and Visualization, Version 3.01. USFS.

Moghaddam, M., Dungan, J., Acker, S., 2002. *Forest variable estimation from fusion of SAR and multispectral optical data*. IEEE Transactions on Geoscience and Remote Sensing 40 (10), 2176–2187.

Murtagh, F., Legendre, P., 2011. *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*. In: *CoRR*, abs/1111.6285 (2011). http://arxiv.org/pdf/1111.6285v2.pdf

Parker, G.G., Russ, E.M., 2004. *The canopy surface and stand development assessing forest canopy structure and complexity with near-surface altimetry*. Forest Ecology and Management, 189, pp. 307-315.

Pastorella, F., Paletto, A., 2013. *Stand structure indices as tools to support forest management: an application in Trentino forests (Italy)*. Journal of Forest Science, 59, 2013(4): 159-168.

Pike, R.J., Wilson, S.E., 1971. *Elevation-relief ratio, hypsometric integral, and geomorphic area-altitude analysis*. Bulletin of the Geological Society of America, 82, pp. 1079-1084.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. *Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction*. Ecosystems (2006)9: 181-199. doi: 10.1007/s10021-005-0054-1.

R Core Team, 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Scrinzi, G., Floris, A., 2003. *Procedure di individuazione, materializzazione e ritrovamento dei punti campionamento (dell'Inventario Nazionale delle Foreste e dei Serbatoi forestali del Carbonio – INFC, Corpo Forestale dello Stato)*. Isafa, Trento. pp. 48.

Shannon, C.E., Weaver, W., 1949. *The mathematical theory of communication.* The University of Illinois Press, Urbana, pp. 117.

Sithole, G., Vosselman, G., 2003. *Comparison of filtering algorithms*. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 34, part 3/W13, 71-78, Dresden, Germany.

Soininen, A., 2012. *TerraScan User's Guide*. © 1998-2012, Terrasolid.

Taškova, K., Panov, P., Kobler, A., Džeroski, S., Stojanova, D., 2006. *Predicting forest stand properties from satellite images with different data mining techniques*. Proceedings of the 9th International Multiconference Information Society IS 2006, 9th– 14th October 2006, Ljubljana, Slovenia, pp. 259–262.

Therneau, T., Atkinson, B., Ripley, B., 2014. *rpart: Recursive Partitioning. R package version 4.1-5*. http://cran.r-project.org/web/packages/rpart/rpart.pdf

Ward, J.H., 1963. *Hierarchical Grouping to Optimize an Objective Function*. Journal of the American Statistical Association, 58(301), 236-244. doi:10.1080/01621459.1963.10500845.

6 Conclusion

In this chapter conclusions on the proposed statistical analysis and techniques for volume estimation and stand structural classification are given. Furthermore, limitations, and future developments of the dissertation are discussed.

6.1 Contributions

The research performed in this dissertation contributed to the body of knowledge that is related to the application of statistical analyses and techniques to the field of forest resources inventory using airborne laser scanner data.

Forest inventory is a monitoring and assessment application which responds to the demand for information regarding to the amount of forest resources (*e.g.* standing volume, biomass), and structural attributes in a specific area [Corona, 2010]. Data needed to produce the solicited information are typically acquired using sampling methods through which only a part of the population (the sample) is surveyed, and inferences regarding the entire population of a sub-population are based on this sample [Corona, 2010].

Coupling aerial laser scanner data and inventory data allows us to study the relationship between remote sensing data and attributes of forests. Thus we can build models by means of which to estimate values for larger area of interest. Although it is not an issue that has been discussed in detail in this dissertation, the research carried out to compare the amount of timber volume in two forest compartments estimated through double-entry stem volume equations and through a linear regression model based on a CHM-derived metric contributed indirectly to field of mapping as a result of the model's application, which is an important extension of forest inventory work in support of forest management [Corona, 2010].

The research carried out within this dissertation contributed to the calculation of standing volume estimation at forest compartment level through an inference strategy which may represent a workable way to infer volume by integrating field plot data and airborne laser scanner data. The proposed approach allows coupling remotely sensed data with sample inventory data to improve the precision of volume inventory estimates, which in Trento province are made at stratum level. The ratio model adopted to lead the estimation presumes a proportional relationship between height from Canopy Height Model and volume at pixel level for the entire forested area under planning. It is adopted as a working device involving simplifying assumptions to derive a volume estimator from height and sample data, which estimate the volume within any portion of the survey area as the proportionality factor estimate multiplied by the total of heights in that portion. The volume estimator was considered from the model-based, design-based and hybrid perspectives. Variances and their estimators were derived under the three approaches together with the corresponding confidence intervals. The hybrid perspective of standing wood volume inference in four forest compartments of a case study reached an uncertainty of estimate comparable to that achieved with full measuring made by forest technicians in Trento forest according to Scrinzi [1989] thirty year field study surveys. When CHM is available, the proposed strategy may represent a strong and highly effective support to forest managers improving the accuracy of volume estimation at forest compartment level.

Nowadays, most forest inventories have evolved towards multipurpose resources surveys and for this reason they have broadened their goals to include additional variables that are not directly related to timber assessment and wood harvesting [Corona, 2010], such as forest structural attributes. At the present, in fact, forest structure analysis and characterization is part of most forest planning processes for sustainable forest management and is of special interest from a silvicultural point of view. In most cases classification systems, based on various criteria, are used both in inventory and forest management decision making phases by forest services. This is the case for the Trento Province Forest Service which based its classification of forest structure on the distribution of basal area in understory, mid-story, and overstory trees (12 forest types). This dissertation contributed to an understanding of whether lidar-derived CHM metrics could be successfully used to predict forest structural classes through supervised classification techniques (linear discriminant analysis, and random forest) and techniques of basal area prediction (*k*-nearest neighbor imputation and parametric regression). The research carried out contributed to suggest that the canopy structure metrics computed from the CHM (coefficient of variation, cover of understory, mid-story and overstory trees) can be moderately useful for basal area imputation and, hence, in the classification process. An important discovery was that the CHM for forest structure characterization is comparable to that of raw lidar data. Considering that some time lidar point cloud data may not be available, in particular when ALS data are collected for geo-morphological or other non-forestry purposes, the possibility to using the CHM to achieve the same performance with lidar raw data is a great advance. Linear discriminant analysis, random forest, and *k*-nearest neighbor imputation, which outperformed traditional linear and non-linear regression analyses, may be used to validate the preliminary classification made by the foresters through a visual estimate during the forest inventory operations. Linear discriminant analysis provided the best levels of accuracy among the techniques considered, so we suggest the use of this technique with caution when a high number of predictor variables is considered. In fact, linear discriminant analysis is too flexible in situations with hundreds of highly correlated predictor variables, and it is too rigid in situations where the class boundaries in predictor space are complex and nonlinear [Hastie *et al.*, 1995]. On the contrary, random forest algorithm can handle thousands of input variables without variable deletion and runs efficiently on large data sets. It does not over fit, and there is no need for variable pre-selection [Strobl *et al.*, 2007]. Moreover, random forest is a flexible classification algorithm that directly provides measures of variable importance (related to the relevance of each variable in the classification process), that has the ability to model complex interactions among predictor variables and include an algorithm to imputing missing values.

The study was used to determine whether metrics extracted from lidar point cloud could be exploited to discriminate between different forest types by means of learning machine technique, such as classification trees. A more simplified classification system of forest structure was suggested. The Forest Service of Trento Province considers 12 different forest structure types. In fact, the unsupervised cluster analysis based on the proportion of basal area in understory, mid-story and overstory trees conducted in a representative sample of forest stands of Trento province allowed discrimination of five structural patterns of forests: pole-stage forests, young forests, adult forests, mature forests, and old forests. The five forest types identified by the clustering seem to be more representative of the complexity of the forests in Trento province. The dissertation contributed proof that the learning machine techniques, specifically classification trees, represent valuable tools in developing predictive models employing a set of airborne laser scanner variables to predict forest structure conditions. Values of internal and external misclassification error confirmed that the classification tree model proposed in this dissertation can support forest managers, to do the following: 1) compare various type of stands and ensure the interpretation of the differences; 2) prestratify forested areas; 3) direct the development of a forest stand toward a given target by applying specific silvicultural interventions in order to foster the dynamics in a stand. The classification tree model obtained through this study is clearly and easily interpretable, and so forest technicians can use classification trees in an operational context.

The potential of aerial laser scanning for supporting operational forest management has demonstrated by this dissertation, both for standing volume estimation and for the analysis and classification of the structure of forest stands.

6.2 Limitations

Some limitations can be identified in the characteristics of lidar data used for the research.

Research at the provincial scale concerning the classification of forest stand structure using the CHM and lidar raw data could be done thanks to the availability of lidar data for the whole territory of Trento province. Unfortunately there data were collected during the winter, and it is now known that the most favorable season to obtain ALS data specifically suitable for forestry applications under temperate and Mediterranean environments is during summer. The significant underestimation of the mean height by the winter lidar data may have influenced the performance of supervised classification techniques applied in the studies. This lidar data property limited the possibility exploring the relationship between ecological coverage metrics, *e.g.* vertical and angular canopy covers, and canopy cover metrics extracted from lidar data, in particular those from the lidar point cloud. In fact, explorative studies accomplished in this direction, provided no significant correlation between vertical and angular canopy covers and cover index computed considering the distinction between single canopy points, and first canopy.

In the research activities carried out to study the use of lidar-derived CHM metrics to predict forest structure classes according to the amount of basal area present in understory, mid-story, and overstory trees examined, a limitation was due to the unequal distribution of observations within the different forest structural types identified through the classification system used by the Forest Service of Trento Province. In particular the low number of cases in some specific forest type represented a "bad starting point" for the application of non-parametric techniques.

An additional limitation can be identified in the difference between the date when ground-plot data were collected and the date of lidar data acquisition. In some case, Reutebuch *et al.* [2000] suggestion of respecting time-gaps of 1-2 was not possible. Lidar data obtained in a sampling mode, *e.g.* strips, instead of wall-to-wall acquisition, could be a solution to that problem allowing most frequent airborne laser scanning.

6.3 Future research

The potential of lidar as forest inventory tool has been demonstrated, but certain aspects have remained undeveloped. For example, by future research should emphasize whether the intensity component of the laser return signal can be used as a source of information for forestry may be placed. In particular, specific research could be developed to study how reflectivity from a laser (with different wavelength) could be used to discriminate between different types of forest cover. Moreover, fusion between lidar data and other image sources could be investigated. Within this topic the simultaneous use of lidar and multispectral images, or the combination of three-dimensional information obtained from lidar and standard photogrammetric techniques, or the combination of lidar with Synthetic Aperture Radar (SAR) data can be considered. Combining the three-dimensional lidar information whit two-dimensional spectral information for forestry inventory and mapping is becoming a research topic and it has significant potential. Other machine learning algorithms than those used in this dissertation (*e.g.* vector machines, neural networks, boosted regression trees, among others) could be taken in consideration combined with lidar data and other remotely sensed data.

6.4 References

Corona, P., 2010. *Integration of forest mapping and inventory to support forest management*. iForest 3: 59-64 [online: 2010-05-17] URL: http://www.sisef.it/iforest/show.php?id=531

Hastie, T., Buja, A., Tibshirani, R., 1995. *Penalized discriminant analysis.* The Annals of Statistics, Vol. 23, No. 1, 73-102.

Scrinzi, G., 1989. *Precisione degli inventari assestamentali per cavallettamento totale: un'analisi di trenat'anni d'esperienza applicativa in Trentino*. Annali dell'Istituto Sper. per l'Assestamento For. e per l'Alpicoltura, Vol. XI, Trento, pp. 143-186.

Strobl, C., Boulesteix, A.L., Augustin, T., 2007. *Unbiased split selection for classification trees based on the Gini index*. Computational Statistics & Data Analysis 52(1): 483–501.

Reutebuch, S.E., Andersen, H.E., McGaughey, R.J., 2005. *Light Detection and Ranging (LIDAR): An Emerging Tool for Multiple Resource Inventory*. Journal of Forestry, 103(6): 286-292.