UNIVERSITY OF
EASTERN FINLAND

Luonnontieteiden ja metsätieteiden tiedekunta
Faculty of Science and Forestry

PREDETERMINING DOMINANT TREE SPECIES TO IMPROVE SPECIES-SPECIFIC VOLUME PREDICTIONS YIELDED BY SPARSE AIRBORNE LASER SCANNING DATA

Janne Räty

MASTER'S THESIS
FOREST MENSURATION AND PLANNING

JOENSUU 2016

Räty, Janne. 2016. Predetermining dominant tree species to improve species-specific volume predictions yielded by sparse airborne laser scanning data. University of Eastern Finland, Faculty of Science and Forestry, School of Forest Sciences. Master's thesis in Forest Science specialization Forest Mensuration and Planning. 60 p.

ABSTRACT

New remote sensing forest inventory techniques developed during this century have become more and more common. Airborne laser scanning (ALS) has proved to be one of the most important remote sensing method that is also able to accomplish inventories for large areas cost efficiently. In Finland, the multisource method which utilizes ALS, aerial photography and field data together is operationally used. However, solely ALS-based area-based species-specific inventories have turned out to be a challenge. Increasing knowledge and the development of the methods has encouraged to study those methods more.

Here, the dominant species pre-classification method has been presented, and the hypothesis was to get accuracy improvements for plot-level volume predictions. The dominant species have been evaluated according to the field measurements but also ALS recognition has been studied. The study material consists of two different datasets and the first has been collected from northeastern (Kuhmo) and the second from southern Finland (Janakkala-Loppi). Three species classes were stratified: Scots pine, Norway spruce and deciduous. The species-specific volumes were predicted by means of Seemingly Unrelated regression (SUR) and compared to k-Most Similar Neighbor (K-MSN) method in the data of Kuhmo. The Kuhmo dataset was also tested to predict the dominant species by ALS using Linear Discriminant Analysis (LDA).

The results revealed that the pre-classification increased the accuracies of fitted SUR predictions. The improvements (RMSE) were 12.6–28.9 % and 20.9–36.9 % depending on the species for Kuhmo and Janakkala-Loppi, respectively. In comparison between parametric and non-parametric methods with Kuhmo data, the k-MSN got slightly better results. In case of predicted dominant species, the LDA predictions degraded the volume accuracies since the overall accuracy of classification was 76 % at best. Although the recognition of the species proved to be challenging with used dataset, the predictions implemented with fitted models (correct dominant species information) revealed that the pre-classifying strategy proposed here has real potential to improve species-specific volume models. According to the tests executed, it was noticed that the classification should be rationalized for each dataset individually to get the best advantage out of it.

Keywords: Airborne laser scanning (ALS), Light Detection And Ranging (LiDAR), Area-based approach, Seemingly Unrelated Regression (SUR), Species-specific volume model

Räty, Janne. 2016. Predetermining the dominant tree species to improve species-specific volume predictions yielded by sparse airborne laser scanning data. Itä-Suomen yliopisto, Luonnontieteiden ja metsätieteiden tiedekunta, Metsätieteiden osasto. Metsätieteen pro gradu -tutkielma, erikoistumisala metsänarviointi ja metsäsuunnittelu. 60 s.

TIIVISTELMÄ

Suomessa kuvioittainen arviointi on ollut vuosikymmenien ajan perinteinen tapa tuottaa tietoa operatiivisen metsätalouden tarpeisiin. Vuosituhannen alusta alkaen kaukokartoitusmenetelmät ovat kuitenkin kehittyneet nopeasti, ja ne ovatkin osittain korvaamassa perinteisiä menetelmiä. Lentolaserkeilaus on yksi kiinnostavimmista kaukokartoituksen menetelmistä, ja sen potentiaali tuottaa tarkkoja ennusteita kustannustehokkaasti on havaittu useissa tutkimuksissa. Käytännön metsätaloudessa käytetäänkin jo kaukokartoituspohjaista inventointimenetelmää, jossa yhdistetään lentolaserkeilauksen, ilmakuvien ja maastomittausten parhaita puolia. Pelkän lentolaserkeilauksen käyttö aluetason puulajikohtaisten tilavuuksien ennustamisessa on kuitenkin osoittautunut haasteelliseksi.

Tämän tutkimuksen tarkoituksena on esitellä ja testata koealakohtaista pääpuulajin esiluokittelua, jonka tarkoituksena on tuoda lisätarkkuutta laserkeilauspohjaisiin puulajikohtaisiin tilavuusmalleihin. Tutkimusaineistona on käytetty vertailun vuoksi kahta eri aineistoa, joista toinen on kerätty Suomen koillisosasta (Kuhmo) ja toinen eteläisestä Suomesta (Janakkala-Loppi). Molemmissa aineistossa eroteltiin kolme pääpuulajiryhmää: mänty, kuusi ja lehtipuut. Lajikohtaisten tilavuusmallien muodostamisessa käytettiin parametrista Seemingly Unrelated Regression -menetelmää. Kuhmon aineiston osalta tutkimuksessa esitetään myös ei-parametrisella k-MSN-menetelmällä tuotetut ennusteet. Kuhmon aineistoon ennustettiin myös pääpuulajiluokitus ALS-muuttujista Linear Discriminant Analysis -menetelmää käyttäen, jolloin myös käytännön ALS-pohjainen ennustustarkkuus kyseisessä aineistossa saatiin selville.

Pääpuulajin esiluokitus paransi SUR-tilavuusmallien sovitusten tarkkuutta (RMSE) riippuen puulajista 12.6–28.9 % Kuhmon aineistossa ja 20.9–36.9 % Janakkala-Lopin aineistossa. Ennustusmenetelmiä verrattaessa ei-parametrinen menetelmä tuotti lähes poikkeuksetta hieman tarkemmat tulokset. Laserkeilausaineistosta ennustettua pääpuulajia käytettäessä tilavuusennusteiden tarkkuudet heikkenivät, koska luokituksen oikeinluokitusprosentiksi saatiin parhaimmillaan ainoastaan 76 %. Vaikkakin laserkeilausaineiston mukaan suoritetun pääpuulajien luokittelun tulos jäi alhaiseksi, voidaan todeta, että tutkimuksessa esitetty esiluokittelu on varsin käyttökelpoinen menetelmä tavoitellessa lisätarkkuutta puulajikohtaisiin aluetason tilavuusennusteisiin. Mallien maastoperusteisia pääpuulajiluokitusvaihtoehtoja testattaessa havaittiin, että luokituksen rakennetta suunniteltaessa aineiston puustosuhteisiin tulee kiinnittää ehdottomasti huomiota.

Avainsanat: Lentolaserkeilaus (ALS), Light Detection and Ranging (LiDAR), Aluepohjainen laserkeilausinventointi, Seemingly Unrelated Regression (SUR), Puulajikohtainen tilavuusmalli

FOREWORDS

This master's thesis is based on the previous study (hereafter: the original study) that has been published in the journal of Forest Ecosystems with the following specifications:

Räty, J., Vauhkonen, J., Maltamo, M. & Tokola T. (2016) On the potential to predetermine dominant tree species based on sparse-density airborne laser scanning data for improving subsequent predictions of species-specific timber volumes. Forest Ecosystems.

The material of this master's thesis consists of two entities. The both data entities included field measurements and airborne laser scanning data. The first dataset has been collected from the area of Kuhmo. I would like to thank Arbonaut, Ltd., especially Dr. Jussi Peuhkurinen for allowing the use of that data collected earlier for other purposes. The second data set was earlier collected from the near area of Janakkala. About this data, I would like to thank UPM kymmene Oy for collecting the field measurements and Blom Kartta Oy for collecting the ALS data. Furthermore, I would like to thank Matti Maltamo for allowing the use of that data.

The original study, in which I was working as a research assistant, was carried out during the summer of 2015 in Faculty of Science and Forestry of University of Eastern Finland, exactly in school of Forest Sciences. The study was a contribution to the Forest Big Data work package of the Data to Intelligence (D2I) program coordinated by DIGILE, Ltd., and financed by the Finnish Funding Agency for Innovation (Tekes) and its business and research partners. This master's thesis was funded by the project of Multi-scale Geospatial Analysis of Forest Ecosystems, of which I would like to thank Professor Matti Maltamo.

The greatest thanks I would like to address to Dr. Jari Vauhkonen for excellent and encouraging supervision during the project in the summer of 2015 and the period of working this master's thesis. I also want to thank him for implementing the k-MSN imputations and refining the English language. Moreover, I would like to express my gratitude to Prof. Matti Maltamo and Prof. Timo Tokola for organizing and taking part in this process with different ways between the summer of 2015 and the moment when I managed to finish this master's thesis.

CONTENTS

1 INTRODUCTION

1.1 Study background

The importance of the forests resources for Finland is enormous since 86 % of area consists of forestry land according to the 11th national forest inventory (Peltola 2014). Due to the amount of the forest resources, strategic and operational planning are essential to get advantage of the resources and simultaneously taking into account the sustainability according to ecological, economic and social aspects. The national forest inventory is one example of strategic forest planning which aims to get comprehensive information of Finnish forests, such as data of growth, biomass, carbon balance and cutting possibilities (Holopainen et al. 2013). The operational forest planning aims to offer as precise information as possible for forest owners of their property. Forest owners need the unbiased data of their forests to make supported decisions for timing the silvicultural operations. Furthermore, forest inventory data are also needed in assessing the ability of soil to produce forest biomass and acquiring the exact information of quality of forests, which would be important for wood procurement (Holopainen et al. 2013; Vauhkonen et al. 2014b). The traditional stand-wise inventories have still implemented, but the remote sensing, both active and passive methods, has taken revolutionary footsteps to develop novel methods to attain more accurate and efficient inventory processes during the 21th century. This paper also aims to present a method related to active remote sensing.

The research of Airborne Laser Scanning (ALS) in forestry applications began with promising results of correlation between field measurements and height metrics extracted from ALS point cloud in the late 90s (e.g. Næsset 1997). The first approaches focused on plot-level forest attributes whereas the second fundamental approach, focusing on individual tree-level, was proposed a little bit later by Hyyppä & Inkinen (1999). After that, methods with Light Detection and Ranging (LiDAR) data collected by small-footprint airborne laser scanners have developed rapidly. At first, the data acquisition costs were high and it reduced the development of the new procedure. Nowadays, ALS is a common method in Finnish forest inventories, and it is supposed to keep on replacing partly the traditional forest inventory methods. At least, the most of the large area inventories are done in contribution with ALS data. Furthermore, The National Land Survey of Finland and Finnish Forest Centre are working co-operatively to implement project which aims to get national coverage of ALS data in Finland. According to the plans, the whole Finland should has been inventoried by 2019 (Maanmittauslaitos…2015).

At the beginning of the development of the ALS techniques, Area-Based Approach (hereafter ABA; Næsset 1997, 2002), was the mainstream method for forest inventories. Nonetheless, the Individual Tree Detection approach (hereafter ITD; Hyyppä & Inkinen 1999) was also studied although it was soon noticed to be more arduous and expensive in light of existing knowledge. Methods have improved a lot and nowadays the main, thus the most cost-effective, method in practical forest inventories have been the area-based approach. Although stand-wise ALS-based and ALS-aided inventories have produced sufficient results (e.g. total stand volumes), recognition of the species-specific data solely from sparse pulse ALS data has been noticed to be challenging even if the recent studies have given promising results for to facilitate those challenges (Vauhkonen et al. 2014c). Most of the studies considering ALS-based researches are located in Europe, especially in Scandinavia, but studies have also been published from areas of North America. The boreal forests which have only few significant tree species, are absolutely adequate locations to implement and develop ALS-based tree recognition.

Traditionally, the recognition task of the tree species has been processed by using aerial photography (Packalén & Maltamo 2006, 2007, 2008) or satellite images (Wallerman & Holmgren 2007). The spectral data derived from aerial or satellite imaginary have proved to give important information for the species-specific forest inventories because the reflected light of the electromagnetic spectrum differs remarkably according to the main boreal tree species, exactly between coniferous and deciduous (Vauhkonen et al. 2014c). ALS data and such images have usually been combined to get more accurate results for species-specific predictions. Many studies have proposed that aerial photography can improve accuracy of tree species characterization, thus species-specific predictions, considering current separating methods of ALS (e.g. Vauhkonen et al. 2012; Ørka et al. 2013). Regarding ALS data in recognition of tree species, there are two attribute classes extracted from ALS, which have been noticed to be able to give essential information about tree species-specific features. Consequently, species-specific information is mainly based on both structural information extracted from ALS data and intensities of returning laser echoes (Vauhkonen et al. 2014c).

In practice, forest assessments using ABA techniques have been executed with non-parametric Nearest Neighbor (hereafter NN) methods to search forest attributes for source grid cell from given feature space, and then attributes can be estimated for a stand and a whole farm. NN imputations are usually thought to be more efficient compared to the parametric regression-

based methods (Holopainen et al. 2013). However, the availability of comprehensive training data is essential when using k-NN methods because the predictions for unknown attributes of cells are obtained from the cells of the observed neighborhood as averages in terms of the detected distance (Holopainen et al. 2013). Especially, if sufficient reference data is not available, some parametric regression methods are also competitive for predicting forest attributes (e.g. Maltamo et al. 2009b, 2012). In this study, regression-based parametric method for plot-level species-specific volume estimations is presented but non-parametric predictions have also presented for dataset from the original study. The second dataset has been analyzed only with regression-based method.

Promising results of classification of the dominant species by ALS data in the previous studies (Ørka et al. 2013; Vauhkonen et al. 2014b) were the encouraging reason for researching this ALS method more. For example, Vauhkonen et al. (2014b) used sparse (< 2 pulses/m^2) to observe differences in ALS derived intensity features of different plots dominated by certain tree species. According to this, dominant species could be predicted accurately. Furthermore, previous studies have also used some kind of pre-acquired dominant tree proportions with ALS data as a predictor to produce more accurate predictions by RMSE, compared to aerial images, for species-specific basal areas in urban environments (Pippuri et al. 2013). Those observations encouraged to construct solely ALS-based volume models for the most significant tree species existing in Finland and to include the dominant tree pre-classification variable in those models. Here, the term of solely ALS-based model means that all the predictors used in models are ALS-based variables but field measurements have nevertheless been used in regression modeling and non-parametric imputation. In practice, it is difficult, almost impossible, to totally avoid field measurements in practical forest inventories. The hypothesis was that the plot-level dominant species information would be able to improve the prediction accuracies of forest attributes at least when the classification is correct.

In the beginning of this master's thesis, an overview of airborne laser scanning applications used in forest inventories has been presented. The purpose of that chapter is to give introductory information to clarify the subject matter of the subsequent chapters. The rest of this study will follow standard structure of a scientific research paper. The material and methods of this study have been presented. After that, results are introduced with visual plot diagrams. In the end, the

results have been analyzed and compared to the previous studies which have been in association with the topics of ALS-based species-specific forest inventories.

This thesis is deeply based on the previous study which has been recently published in the journal of Forest Ecosystems (see Forewords). The aim of the original study was to produce solely ALS-based species-specific timber volume models in a strongly pine-dominated study area. One of the issues was to test *a priori* classification information of dominant tree species to acquire more accuracy for species-specific models. The models have been produced with two different methods, SUR and k-MSN, which are compared. In this master thesis, the whole study has been presented as an extensive edition with some new and broader formatting and analysis of the results. Also supplementary ALS and field data have been analyzed and modeled to give comparative material beside the earlier results. The k-MSN part of the original study has been left as a comparison and the methodological emphasis has shifted on Linear Discriminant Analysis and Seemingly Unrelated Regression. Owing to the strong relation to the earlier study, this master's thesis attempts to offer broader aspect for questions presented in the original study instead of trying to present totally new study objectives.

1.2 Research objectives

The previous studies have proposed results of species-specific volume predictions, but pre-classifying of the dominant tree species is a novel idea that is not utilized earlier in the same way. Furthermore, the most species-specific studies are emphasized on individual tree detection methods and the predictions without tree delineation are quite uncommon. The individual tree methods have been observed to be more accurate in common but the development of the area based methods would be advantageous for practical forest management in which sparse airborne laser scanning data have been operationally used in contribution with other inventory data sources.

According to the background presented, the subsequent objectives can be stated: (1) the main objective is to attain improvements for the accuracies of the SUR-based species-specific plot-level volume predictions that are based on ALS data and observed dominant tree species, (2) two different datasets by tree species compositions are evaluated and according to the evaluation the possible guidelines for subsequent dominant tree classifications will be presented and

(3) a comparison between non-parametric and regression based estimation methods are examined. Furthermore, the discrimination of dominant tree species of sample plots by extracted ALS features was also one of the issues in the original study. Consequently, a solely ALS-based approach to yield species-specific volume predictions can be presented and evaluated.

2 AN OVERVIEW OF AIRBORNE LASER SCANNING

2.1 History and theory

Airborne laser scanning has matured into one of the most researched fields in the sector of forest mensuration. ALS method is often related to Light Detection and Ranging (LiDAR). Virtually, the ALS is utilizing the LiDAR, and it also uses positioning system (e.g. assisted Global Positioning System) to give very precise three-dimensional x-, y- and z-coordinates for an airplane processing the laser scanning. Thus, also the locations of target objects, such as echo returns from canopy, can be calculated. Considering term of ALS, it is originally from Europe whereas LiDAR has been developed in the United States (Holopainen et al. 2013). In forestry applications, the principle of ALS is to produce three dimensional point data of the vegetation beneath the airplane. The very first studies considering ALS were implemented in 1964 when airborne profiling LiDAR system was used to measure forest canopies (Rempel & Parker 1964). The revolutionary development can be observed during 1990s when GPS and Inertial Navigation Systems (INS) were integrated and become more available for public applications. By below 20 years, ALS has become one of the most important forest inventory method, and the traditional forest inventory methods, as field measurements and aerial images, are giving more space for modern ALS methods. The ALS methods and equipment are evolving all the time and due to that, the forest inventories are becoming more efficient considering both time and costs.

The remote sensing can be divided in two sections: active and passive. The passive remote sensing is based on methods that do not use external machines to produce emissions, and instruments can utilize the natural radiation which is emitted by the object of interest. In forestry applications, passive remote sensing technologies have traditionally been used, for example, as form of aerial imagines (Packalén & Maltamo 2007). Whereas in active remote sensing, the

instruments are emitting light (often near-infrared) beams (e.g. laser pulses) towards to the object. To take the advantage of emitting laser pulse, the returning echoes should be captured with external receiver. Thus, ALS is classified as an active remote sensing. Due to the physical principles of ALS, the main ranging equipment needed in the scanning process are: the emitting laser unit to send laser beams and the electro-optical receiver to catch echoes. Moreover, significant part of the ALS system consists of opto-mechanical scanner and unit for controlling and processing data (Wehr & Lohr 1999). Control and processing unit includes aided positioning system which consists of GPS/GNSS (GNSS is a global fused positioning system, Global Navigation Satellite System) and INS systems of which latter is measuring the orientation of an airplane (Holopainen et al. 2013).

Implementing ALS to produce three-dimensional point cloud, the time, between the moment of emitting the pulse and capturing the pulse echo, has to be measured to take the advantage of the process. To determine the height of the underlying object, the speed of the airplane and the elapsed time between transmitting and receiving the laser beam are required (Wehr & Lohr 1999). Since precise position and angle of the laser transmitter are known, the height of the reflection point can be reported. However, the canopy of forest does not form a solid surface and inevitably part of the transmitted laser pulses tend to divide, which will cause that the receiving unit will capture many returning echoes. However, the most common situation is that only one echo will be captured (Holopainen et al. 2013). For example, the first echo may be returned from the top branches of tree, the second from middle branches and the third from the ground. According to this notice, it is possible to produce forest characteristics that are describing structure of vegetation above the ground. For example, it is possible to process point clouds according to the echoes returned from the ground to produce Digital Terrain Model (DTM) or conversely the first-echoes are suitable for producing Canopy Height Model (CHM).

Nowadays, full-waveform ALS methods have become more available to produce full-wave recordings of the laser energy instead of only having individual echo points between the ground and a canopy (Roncat et al. 2014; Vauhkonen et al. 2014c). Full-waveform methods are able to produce more accurate information of the forest and presumably it will be one of the most interesting ALS techniques, together with multispectral ALS acquisitions, in the future. However, there is also need for some research in the current, economically more efficient, field of

small-footprint discrete-return methods which have been noticed to be able to yield even more useful and accurate information of forests.

The structural features (e.g. height and percentiles) of the canopy are the most important attributes which are extracted from ALS data. Furthermore, ALS techniques have rapidly improved and modern ALS equipment can also recognize laser echo intensity information which is especially advantageous in the ALS-based forest inventories that aim to yield species-specific predictions for forest attributes (Korpela et al. 2010). Further about species recognition in section 2.3.

2.2 Basic ALS inventory techniques

Considering airborne laser scanning, there are two different mainstream methods for predicting forest attributes. The most used method is called Area Based Approach (ABA), and it is proposed by Næsset (1997, 2002). In the ABA method, strong statistical correlation is required between forest inventory plot data and attributes extracted from ALS data. The other fundamental technique used in forestry applications is Individual Tree Detection (ITD) (Hyyppä & Inkinen 1999). According to this method, in order to recognize individual trees, the ALS-based surface models for canopy covers are often exploited. In practice, the locations of single trees are generally determined according to the local maximum points of canopy height models.

Most often, forest inventories are implemented using ABA in Finland nowadays. This method has showed its potential to model forest attributes with adequate accuracy for practical forestry, especially using passive remote sensing in contribution with ALS. Additionally, ABA method is almost always more cost efficient than ITD method when inventories are implemented in large forest areas. The main reason for cost efficiency of ABA is the number of pulses per spatial unit (Maltamo et al. 2009a). On the other hand, the ABA method requires always high quality forest mensuration data acquired from well-organized plot design. In the ABA method it is possible to use ALS pulse densities between 0.5–2.0 measurements m^{-2} whereas ITD mainly requires over 2.0 measurements m^{-2} (Holopainen et al. 2013). Considering the accuracies of ALS metrics on varying pulse densities, the difference is not directly noticed because standards for plot size and pulse density have not been chosen in the recent studies (Vauhkonen et al.

2014a). However, it has been noticed that remarkable degradation in estimations cannot be noticed if nominal pulse densities per square meter are reduced even to 0.06 pulses per plot (Maltamo et al. 2006; Gobakken & Næsset 2008). However, observations have been obtained in artificial and theoretical circumstances and the decrease of resolution in real data acquisition will probably have a more significant effect on data quality. In addition of pulse density, plot design is worth organizing carefully. The operating principle in ABA method is implemented by grid cells which cover whole inventory area and size of the precisely located plots of inventory area are matched with grid cell size. The predictions for other grid cells are estimated by using ALS-based metrics and observed inventory data of reference plot cells. Hence, plot selection for forest field training data is beneficial to be fitted according to the pre-information of inventory area (Maltamo et al. 2011). Maltamo et al. (2011) have proven that in the case of volumes, ALS data as *a priori* information in plot selection strategies can produce the most accurate results compared to random sampling or selection according to geographical location, especially when number of plots were kept under 150.

Implementing the ABA method for wall-to-wall forest inventory, different ALS variables have to be extracted and selected for to produce adequate independent variables that are able to form desirable regression models for forest attributes of interest. Alternatively, non-parametric k-NN methods are often used in predicting forest stand characteristics (Maltamo et al. 2006). Non-parametric methods are more often used because the construction of regression models individually for every forest object has proven to be arduous (Holopainen et al. 2013). The key ALS extracted variable for forest attributes, such as volume, is the height of canopy. Other often used ALS-based attributes, also used in this study, are for example height percentiles and corresponding densities. Also vegetation ratio is often used to describe understory of a forest. Usually, vegetation ratio threshold has been set on, for example, 2 meters above the ground. Intensities of laser echoes have turned out to be adequate especially in distinguishing tree species and in those cases high or very high density ALS data is used most often. In this study, intensity variables were nevertheless used in distinguishing tree species although the sparse ALS data was implemented.

With ITD method, it is possible to execute forest inventories even without field measurements (Holopainen et al. 2013). The first presumption is that ALS data could be sufficient for predict-

ing all or part of the forest attributes of interest and, secondly, common models should be available for unknown attributes. For example, diameter of recognized tree is not properly possible to estimate according to the ALS point cloud. In this case, for example allometric models (Kalliovirta & Tokola 2005) and local regression models (Peuhkurinen et al. 2007) are used. However, modelling the breast-height diameter is not simple case because the vertical dimensions of single trees are not the only variables affecting diameter of tree: such as silvicultural history and stand density are also variables which have an effect on growth of diameter (Maltamo et al. 2007). Furthermore, the most challenging challenge is that ITD-based forest inventories meet often problems with determining locations of stems and all trees cannot be detected from the height surface models which are based on ALS data (Vauhkonen et al. 2014a). The previous studies have proposed some alternatives to prevent those problems by, for example, using pre-assigned selection filters (Heinzel et al. 2011) or more accurate ALS data such as full-waveform data instead of conventional ALS echo data (Reitberger et al. 2009). The problems with duplicating inaccuracies of allometric models (e.g diameter at-breast-height) can be avoided by using, for example, NN-methods or regression (Maltamo et al. 2009b; Vauhkonen et al. 2010) to produce volume models straight from ALS data.

2.3 Species-specific assessments by utilizing ALS metrics

In practice, yielding forest attributes for needs of compartment-wise forest management, such as volume of timber or basal area for optimal management decisions, it is necessary to be able to produce species-specific inventory data from forest (Vauhkonen et al. 2014a). Tree species recognition has been one of the biggest issues in implementing forest inventories by ALS data, and this study also attempts to test some ideas for to relieve subsequent classifications to attain more accurate results in the future. Traditionally, recognition has been implemented with collaboration of hyperspectral or multispectral images but recent studies have stated that even pure ALS data could have potential to recognize at least species of the boreal forest well enough (Korpela et al. 2010). For example, Holmgren & Persson (2004) have managed to classify over 560 sample plots of Norway spruces (*Picea abies* [L.] H. Karst.) and Scots pine (*Pinus sylvestris L.*) with overall plot-level success rate of 95 %. Studies have denoted that deciduous species may cause some problems in the ALS-based classification process (Ørka et al. 2007). From the point of view of multisource inventories, this result is not overly insuperable since the spectral data of aerial images are capable to distinguish deciduous from coniferous due to the clear differences in ability to reflect the light in the infrared (over 750 nm) and red-edge (680–720 nm) areas of the spectrum (Vauhkonen et al. 2014c).

Ørka et al. (2012) tested again classification with coniferous and deciduous forests using height percentiles to characterize structure of forest beneath the canopy cover, and they also used normalized intensity variables. In that study, quite high pulse density was used and the method for the identification was based on individual tree approach. The overall accuracy of that classification could reach 77 %. Thus, aforementioned studies focused on to use the individual tree lineation and ALS data with quite high densities although operationally lower pulse densities are often used in Finland – usually densities beneath 1 observations m^{-2} are preferred in area-based approaches. However, earlier studies have also researched ABA methods, without individual tree lineation, to predict the species-specific composition in plot-level and at least dominant tree species can be separated quite well but minor species proportions, for example under the dominant canopy, are more challenging to predict compared to the individual tree detection methods (Ørka et al. 2013). However, the recognition of main species in plot-level is less studied because the generalized structural and intensity data is not so obviously describing species-specific properties than the individual tree properties of ITD methods. Some studies have proven that the methods are also able to yield species-specific estimations with ABA methods (Wallenius et al. 2012), most of them have also utilized spectral data in contribution with ALS (Packalén & Maltamo 2007). Exactly, those ALS-assisted ABA methods are used operationally in forest management in Finland (Vauhkonen et al. 2014c; further Maltamo & Packalén 2014).

All in all, according to the recent studies the most advantageous elements extracted from ALS data for species recognition is difficult to choose between structural and intensity features (Vauhkonen et al. 2014). However studies, such as Törmä (2000) being one of the earliest, have noticed the potential of intensity values during the ALS-era in forestry. The intensities of returning echoes are describing mainly the ability of reflectance of laser pulses but moreover the intensity values are affected also, for instance, by the size of ALS footprint, the power of transmitted pulse or otherwise the size and the quality of target. It is also worth noticing that there are some differences between sensors, and it is possible to normalize sensors to produce a normalized intensity (Ørka et al. 2012). Of course, the laser beam will be scattering all the time when it hits the targets and so the intensity is depending also on this variable. According to that notice, the highest intensity values will be captured from tree species that have large leaf surfaces, for example Maple (*Acer platanoides*) (Korpela et al. 2010). Thus, the possible advantage of the intensity features should be individually considered in every operational case according to equipment employed and area measured.

The other, more advanced, species recognition method which is based on very high density laser pulse data is proposed by Vauhkonen et al. (2008, 2009). The principle of this method is to create structural three dimensional alpha shapes for individual trees. According to these triangulated point clouds, it is possible to derive classification features, such as computational volumes of trees. The method has proven to be capable to yield very accurate results, for example, the overall accuracy of 93 % considering species of pine, spruce and deciduous trees. This alpha shape-based method has also been tested in plot-level when sparse ALS data have been employed with encouraging results (Vauhkonen et al. 2012).

2.4 Accuracy needs of species-specific area-based volume models

The traditional field inventory method is practically implemented by means of angle count sampling field measurements and visual assessments carried out by forest professional. Thus, the traditional method can be found more subjective than the ALS-based approaches in implementing forest inventories. Moreover this discussion, studies have proven that the inclusion of ALS data in multisource stand-level inventory operations is able to give at least as accurate species-specific results as the traditional way (e.g. Wallenius et al. 2012; see also multisource inventory by Packalén & Maltamo 2007) and especially considering totals of the forest attributes, the modern ALS-based method tends often to give more accurate results (Holopainen et al. 2013). The accuracies of, for example volume models, are often assessed by means of RMSE and BIAS (Packalén & Maltamo 2007). As a reference, the proper and useful predictions of pine stand volume should not achieve relative RMSE over 30 % and in mixed pine dominated forests relative RMSE should be at most about 20–40 % (Uuttera et al. 2002). Species-specific models will easily be more inaccurate even with the traditional stand-wise inventory methods, for example Haara & Korhonen (2004) observed relative RMSEs of 29 %, 43 %, 65 % and 25 % for pine, spruce, deciduous and stand total, respectively. According to the previous study of ALS-based ABA control inventories by Wallenius et al. (2012), the relative RMSEs of species-specific volumes have been 33 %, 63 %, 69 %, 15 % for pine, spruce, deciduous and total, respectively. Regarding to this study, species-specific models for minor species are not accurate enough but the total volume results can be found adequate for the practical forest management. It should be remembered that study area in the latter study had strong and the first had clear dominance of Scots pine, which is able to explain the inaccuracy of the minor species.

3 MATERIAL AND METHODS

3.1 Study areas

The first part of the data studied were originally collected for crown base height assessments (Korhonen 2012). Two test areas within a geographical distance of 30 km were established in Kuhmo, northeastern Finland. With respect to tree species proportions, the area is very homogenous and strongly dominated by Scots pine trees. The other species to be distinguished are Norway spruce and a group of deciduous trees, consisting mainly of birches (*Betula* spp. L.), form minor proportions that typically occur below the dominant canopy. Altogether 265 field sample plots with co-located ALS and field data were studied.

Circular sample plots with radii of 9 m were used in the field data collection. Every tree with a diameter at breast height (D_{bh}) > 5 cm was measured for the D_{bh} and crown base height. Trees with a D_{bh} corresponding to the basal area-weighted median tree of each species occurring on a plot were determined in the field and measured for tree height. The D_{bh} and height of these trees were used as the median tree diameter and height (D_{gM} and H_{gM}, respectively) of the corresponding species per plot, and the maxima of the values were used as the D_{gM} and H_{gM} of the entire plot. Plot basal area (G) was calculated by summing from the D_{bh} measurements. The missing tree heights were predicted by calibrating the prediction models for the parameters of Näslund's (1936) height curve presented by Siipilehto (1999) using the species-specific D_{gM} and H_{gM} estimates. The volumes of the individual trees were predicted by models of Laasasenaho (1982), employing the D_{bh}, height, and tree species as predictors. The models for birch were used for all deciduous trees. Central characteristics of the field measurements aggregated for the field plots are shown in Table 1.

Table 1. Species-specific volume characteristics of the 265 sample plots in Kuhmo.

	Mean	**Min**	**Max**	**Sd**
Total volume, m³	131.5	6.3	434.9	85.3
Pine volume, m³	87.2	0.0	295.6	66.4
Spruce volume, m³	28.6	0.0	401.6	53.7
Deciduous volume, m³	15.8	0.0	178.1	24.3

The second part of the data studied were acquired for Metsälaser 2 -project by UPM kymmene Oy during the summers of 2007 and 2008. The field data was collected from two separate areas,

Janakkala and Loppi, in southern Finland within a geographical distance of about 25 km. The study area was noticeably more heterogeneous by tree species compositions than the data of Kuhmo. The proportions of deciduous and spruces were stronger that is supposed to give interesting comparison with the strongly pine dominated data. However, this area was also strongly dominated by coniferous species. In this study data, there were also distinguished tree species classes for the study: pine, spruce and deciduous species. After combining the ALS data and the field data, altogether 434 field plots were studied.

The computational methods for calculating plot-level volumes and other characteristics of the second dataset were described by Kotamaa and Villikka (2008). Central characteristics of the joined data of Janakkala and Loppi are presented in Table 2.

Table 2. Species-specific volume characteristics of the 434 sample plots located in Janakkala-Loppi.

	Mean	**Min**	**Max**	**Sd**
Total volume, m^3	205.8	24.2	672.5	113.8
Pine volume, m^3	77.0	0	536.3	89.6
Spruce volume, m^3	105.4	0	672.5	132.3
Deciduous volume, m^3	23.3	0	254.6	40.2

3.2 ALS data acquisitions and the extracted features

The ALS data for areas of Kuhmo were acquired on September 4–7, 2011. Leica ALS50-II scanner was operated from an altitude of 2000 m using a field-of-view of 30°, a scanning rate of 52 Hz, and a pulse frequency of 58.9 Hz. These scanning parameters resulted in a nominal measurement density of 0.52 observations m^{-2}. The ALS data for area of Janakkala-Loppi were acquired during the summer of 2007 using an Optech ALTM3100 laser scanning system. The data acquisition was operated from an altitude of 2400 m using a field-of-view of 30° and a scanning frequency of 30 Hz. In this case, the nominal measurement density was 0.62 measurements m^{-2}. Owing to the data acquisition period, the leaf-on data have been used in study

The predictor features extracted for the study were mainly based on the earlier studies (e.g., Vauhkonen et al. 2014b). However, since a prediction of the crown base height of a tree has been found to be an useful indicator of its species based on tree-level studies (Holmgren &

Persson 2004; Holmgren et al. 2008) and the quality ALS-based CBH data were available for Kuhmo data, the CBH was supposed to be a potential independent variable to improve species-specific ALS-based discrimination. The area-based estimate of crown base height were implemented to distinguish plots dominated by various species. The CBH was earlier predicted by extracting connected alpha shape components from the lowest parts of the point cloud according to the method of Maltamo et al. (2010), which is a variant of a tree-level method described by Vauhkonen (2010).

The other ALS features considered were the mean and standard deviation of the intensity values and the proportion of the different echoes (Vauhkonen et al. 2014b). Following Ørka et al. (2012) and Vauhkonen et al. (2014b), for example, the intensity features were calculated separately based on all, only, or first-of-many echoes. However, the intensity variables were not available from Janakkala-Loppi ALS acquisition, thus only the structural variables have been used in the models considering that dataset. The most common structural ALS-based predictor variables (Næsset 2002), i.e., the maximum, the mean and the standard deviation of the height values, proportion of echoes above 2 m vegetation threshold, various height percentiles (5^{th}, 10^{th}…95^{th}) and the corresponding proportional densities of the ALS-based canopy height distribution were calculated according to Korhonen et al. (2008) for Kuhmo data. Principally, the same common variables were also available in data for Janakkala-Loppi (Kotamaa & Villikka 2008). All the structural ALS features were calculated according to the first echoes in all of the cases.

3.3 Methods

3.3.1 Methodological overview

There are two different data sets used in this study, which may easily cause confusion in implemented methods between data sets. Table 3 will clarify the meaning of the implemented methods in both datasets and the purposes of the stages are also presented. The data of Kuhmo was first used for all the experiments.. The Janakkala-Loppi data is used entirely as a supplemental data to verify the results obtained in the Kuhmo data. It was noticeably probable that the strong pine dominance has an effect on the accuracies of the volume predictions. Hence, the Janakkala-Loppi data was tested as a more heterogeneous area according to its species compositions, and it is supposed to give advantageous information for subsequent dominant tree classification

structures. This hypothesis was tested by re-fitting the SUR models. According to the results with Kuhmo data, it was very reasonable to leave Janakkala-Loppi out of the Linear Discriminant Analysis and final predictions because the dominant species structure is more complicated and the intensity values of ALS were not available.

Table 3. An overview of the analysis presented in this study. All of the analysis have been done with the original Kuhmo dataset whereas the data of Janakkala-Loppi have only been analyzed in the model fitting stage. The number codes in brackets: 1 – the fitting stage; 2 – the solely ALS-based prediction stage.

Analysis	Dataset	Purpose
SUR (1)	Kuhmo & Janakkala-Loppi	To predict volumes according to the ALS features and observed dominant species information; To compare predictions between datasets and verify the operability of predetermination method
LDA	Kuhmo	To determine the plot-level dominant tree species by means of ALS
SUR (2)	Kuhmo	To evaluate accuracies of solely ALS-based predictions
k-MSN (1,2)	Kuhmo	To compare volume predictions with SUR method

3.3.2 Pre-classification of the dominant species by field and ALS data

The species proportions were determined as the percentages of each species from the total plot basal area. The dominant species were subsequently determined based on these proportions. Several alternatives to determine the exact percentage values for the dominant species were tested, however, to analyze operationally feasible possibilities to derive this information by ALS (Table 4). First, the species with the highest percentage were set as the dominant species of the plot, yielding three dominant species classes (pine, spruce, and birch dominated). Second, the dominant species were determined using a threshold of 75%: whether a species had a proportion higher or equal to this level, it was set as the dominant species of the plot. Whether no species reached this threshold, the plot was labeled as "mixed". For example, this classification yielded the dominant species classes of pine, spruce, deciduous, and mixed. The rest of the classes were determined adding separate true pine class since the study area of the original study (Kuhmo) was noticed to be strongly dominated by pine. Those plots were selected using a threshold of 95 % and tested along the aforementioned two alternatives. In this master´s thesis, this idea was also tested in the supplement Janakkala-Loppi data in which the dominance of any species was not such strong. However, the inclusion of true pine class was reasonable to

test because areas were coniferous dominated as well. The definition alternatives for the dominant species are listed in Table 4.

Table 4. The different definitions used for the dominant tree species in this study.

Abbreviation	Definition for the dominant species	Classes[1]
Sp_{max}	Highest species-specific proportion of G per plot.	P, S, D
Sp_{max+95}	Highest species-specific proportion of G per plot + separately labeled plots with G ≥ 95 % of pine.	P_{95}, P, S, D
Sp_{75}	Species-specific proportion of G ≥ 75 %; plots with a lower dominant proportion pooled in a separate class.	P_{75}, S_{75}, D_{75}, M
Sp_{75+95}	Species-specific proportion of G ≥ 75 %; plots with a lower dominant proportion pooled in a separate class + separately labeled plots with G ≥ 95 % of pine.	P_{95}, P_{75}, S_{75}, D_{75}, M

[1] Dominated by pine (P), spruce (S), deciduous trees (D), or the aforementioned species with the proportion given in the subscript; or mixed (M).

The extracted ALS features were subsequently used for yielding classifications for aforementioned strategies to stratify the dominant species. The original study included only the Kuhmo data and the attempts to classify the dominant tree species according to the ALS variables were only implemented in that data. The supplement data of Janakkala-Loppi for this study were regarded as a complicated situation by tree compositions, which supported, with the results achieved in the original study, the speculation that the predicted classification would be redundant. The ALS-based predictors were first graphically assessed with respect to their abilities to discriminate between species and invariance with respect to tree size quantified in terms of the D_{gM} and H_{gM} characteristics.

3.3.3 A linear discriminant analysis

A linear discriminant analysis (LDA; a generalization of the method introduced by Fisher (1936)) implemented in the *MASS* package (Venables & Ripley 2002) of R (R Core Team 2013) was used to classify the data by tree species for the final prediction stage of the Kuhmo data. For producing the categorical classification for plots according to the ALS features, Linear Discriminant analysis was used. The principle of LDA is to form linear combination which maximize the ratio of the between-class to within-class variance based on the data of the original feature vectors (Venables & Ripley 2002). To clarify, the main effort is to determine the linear line which is able to maximize the variance between the classes in analysis, i.e., the data classes projected for the linear line are located as far away from the line as possible. The LDA was run

with a leave-one-out cross validation, in which the priors of the LDA were adjusted to give an equal probability for each species. The predictors used in the LDA were manually selected according to the graphical assessments. First, the discriminant functions were fitted with one predictor variable at the time. The variables resulting to best accuracies were added with a second variable and the accuracies of these combinations were further ordered. The procedure was repeated until the number of predictors was 4.

3.3.4 Modelling the species-specific volumes

Prior to the modeling, the predictors based on the ALS data were evaluated with respect to their relationships with the species-specific volumes in a similar manner than described in the previous section for LDA predictors. Finally, two modeling strategies, a parametric regression based approach and a non-parametric nearest neighbor, were tested in the workflow of the original study. This master´s thesis intends to emphasize the focus on the foremost, regression based, method although the principle and the results of a non-parametric method are also briefly presented.

3.3.5 Seemingly Unrelated Regression (SUR)

The species-specific volumes were predicted as a simultaneously fitted system of equations based on the Seemingly Unrelated Regression (SUR) modeling implemented using the *systemfit* package (Henningsen & Hamann 2007) of R (R Core Team 2013). The main idea of SUR (Zellner 1962) is to take into account the interactions between residual structures (disturbance terms) of different linear regression equations, and results of every regression model will have an effect on equations in SUR modelling (Henningsen & Hamann 2007). The coefficients of the SUR model were based on generalized least squares (GLS) estimation. A presumption for the GLS method is that the matrices which are constructed from the regression models should be correlated but unequal. The one alternative for GLS would has been the OLS (Ordinary Least Squares) for equation-by-equation models but due to the correlations between the explanatory variables, it was reasonable to employ the GLS estimator.

In the SUR modelling, the dominant tree species were taken into account by introducing a categorical predictor variable with levels corresponding to the dominant tree species classifier considered. Constructing the SUR model groups were implemented by examining every single

model individually (section 3.3.4). The ALS features were added as further predictors, with the categorical variable, of the model based on the coefficient of determination (R^2) values. Individual predictors were added attempting to maximize the R^2. However, a new predictor was included only if it affected the model significantly according to a p-value of the Student's t-test. The selection of the independent variables for species-specific models were manually managed due to the quite slight set of alternatives to test. It should be noticed that with larger ALS datasets this method will not presumably prove to be an adequate working technique. The significances of the ALS variables and categorical variables proved to alter due to the theory of SUR method when the models were joined for a SUR group. This notice was motivated to accomplish some tests for coefficients which are presented in the end of the results section.

3.3.6 K-Most Similar Neighbor (k-MSN)

The nearest neighbor (NN) approach, used for volume predictions, is based on an average of k-NN observations in terms of the ALS features. The NNs were determined according to the Most Similar Neighbor (MSN) distance metric (Moeur & Stage 1995). The k-MSN approach uses a canonical correlation analysis to produce a weighting matrix for suitable nearest neighbors from the feature space, i.e., from the training data.

The k-MSN imputation was implemented using the *yaImpute* package (Crookston & Finley 2007) of R (R Core Team 2013). In practice, the dominant species information was taken into account by restricting candidates in the feature space including only plots which had the same dominant tree species than the target plot. Taking into account this restriction, up to 1–10 NNs were selected from an initial neighborhood. The total and species-specific volumes were predicted simultaneously as arithmetic averages of the restricted k-NNs.

3.3.7 Accuracy assessment and tests

The accuracies of the predictions were originally assessed separately at the stages of model fitting and prediction. Due to the decision to ignore the prediction stage of the Janakkala-Loppi data, only the fitted models are evaluated in that material. In the case of prediction stage, the dominant species predicted according to the LDA were used to replace those observed dominant tree information which were captured in the field and used for fitting the SUR models.

The accuracy of the species-specific volume predictions was assessed by means of the root mean squared error (RMSE, Eq. 1) and mean difference (BIAS, Eq. 2) between the observed and estimated values.

$$RMSE = \sqrt{\frac{\sum(p-r)^2}{n}} \quad (1)$$

$$BIAS = \frac{\sum(p-r)}{n} \quad (2)$$

where p is the observed value based on field measurements, r is the predicted value, and n is the number sample plots.

The accuracies of the species classifications (LDA) were assessed by means of the overall accuracy and kappa (κ) scores. The overall accuracy gives the number of correctly classified dominant tree cases as a proportion of all observations. The κ coefficient (Eq. 3) can be interpreted as a proportion of chance-expected disagreements which do not occur (Cohen 1960). In this case, it describes how much better the results of LDA classification are compared to the corresponding material which is classified by chance. The κ coefficient was obtained as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (3)$$

where p_o is proportion of correctly classified observations and p_e is probability of correct classification by chance.

After actual accuracy assessment, it proved to be beneficial to check the significances of the categorical, so-called dummy, variables whether some of the coefficients are redundant. Furthermore, the numerical evidence for the operability of the dominant tree classifications would be important to present plausible outcomes of the method. The tests were implemented by means of the Wald-test of *CAR* package (Fox & Weisberg 2011) of R (R Core Team 2013**)**. The significances in SUR groups were assessed by using χ^2 for the Wald test.

The tests were carried out for every categorical variable of the considered classification strategy so that, at first, the whole variable of model group was ignored. It was implemented by setting coefficients in every equation as zero in restriction, i.e., in null hypothesis. The p-values showed whether the coefficient would be worth removing while the risk level of 5 % was set as a threshold value. Also stepwise test procedure was implemented for individual coefficients of categorical variables (Further description in Results). The procedure was executed to reveal redundant variables in a single species-specific equations to simplify subsequent equations and to notice possible congruence between two different datasets considered.

4 RESULTS

4.1 Relationships between ALS features and species-specific attributes

The CBH predicted by ALS for sample plots of Kuhmo had RMSEs of 1.58 m and 1.47 m and biases of -0.93 m and 0.07 m, when evaluated against the arithmetic and basal-area weighted means of the field measurements, respectively. These accuracies suggest that the area-based prediction of the CBH is a reliable estimate of this measure particularly with respect to the largest trees. The results are on the same accuracy level as in the earlier studies (e.g. Maltamo et al. 2010).

The CBH was however not an appropriate indicator of the tree species proportion (Figure 2). Instead, other ALS features produced a better discrimination between the dominant species considered. For example, considering data of Kuhmo, the features based on the proportions and intensities of the different echoes (Figure 2) indicated a difference in the leveling between pine and spruce dominated plots. This difference was also invariant to the size according to the D_{gM} measure. Although, the actual classification was not implemented for Janakkala-Loppi data, it was also interesting to compare relations between structurally different datasets. Thus, a set of used variables are presented in Figure 1. As we can see, the corresponding variables between datasets are giving such similar results although the ability to distinguish was better in the strongly pine dominated data. Generally, for deciduous dominated plots it was difficult to find ALS features which could separate them from the other species groups.

At first, the height metrics with density metrics were supposed to have a main role in describing volumes of the plots. The height and the density metrics had a quasi-linear relationship between the total and main species volumes, as illustrated in Figure 3 using a product of a height percentile and the ratio of echoes reflected above ground to all echoes, i.e., the canopy cover. However, the volumes of the minor species were not favorably related to these metrics (Figure 3). Concordant results were also noticed in the data of Janakkala-Loppi (Figure 4). However, variances between classes were greater compared to Figure 3. This could be explained by the much larger dataset that covers a vast variety of different plots, which also offers a better presentation for deciduous plots (c.f. Figures 3 & 4). The computational procedure of coefficients in multi-independent variable regression model is imitating the metrics idea presented in Figures 3 and 4. According to those notices, moreover with that the main species were generally well related to the produced metrics, it was natural to regard both density and height metrics as potential candidates for the species-specific volume models.

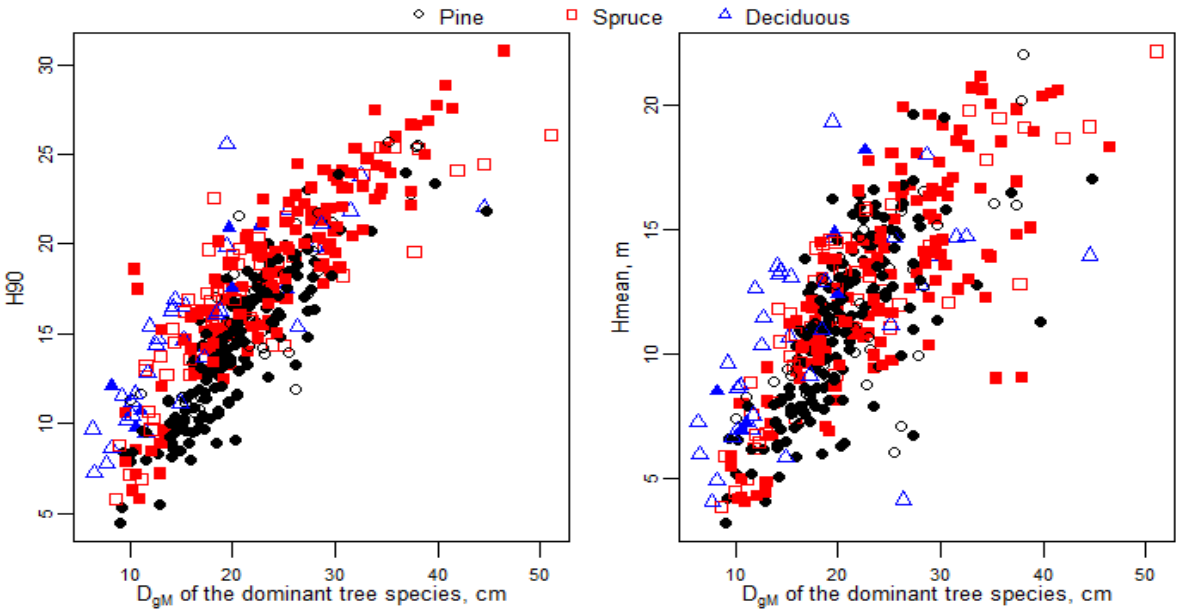

Fig. 1. A pair of ALS-based variables illustrating the species-specific differences in the data of Janakkala-Loppi. The field-measured D_{gM} is used in the x-axes to assess the invariance of the features to size. *H90* – the 90th height percentile, *Hmean* – the average of the height of the first-returns occurred in each plot. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

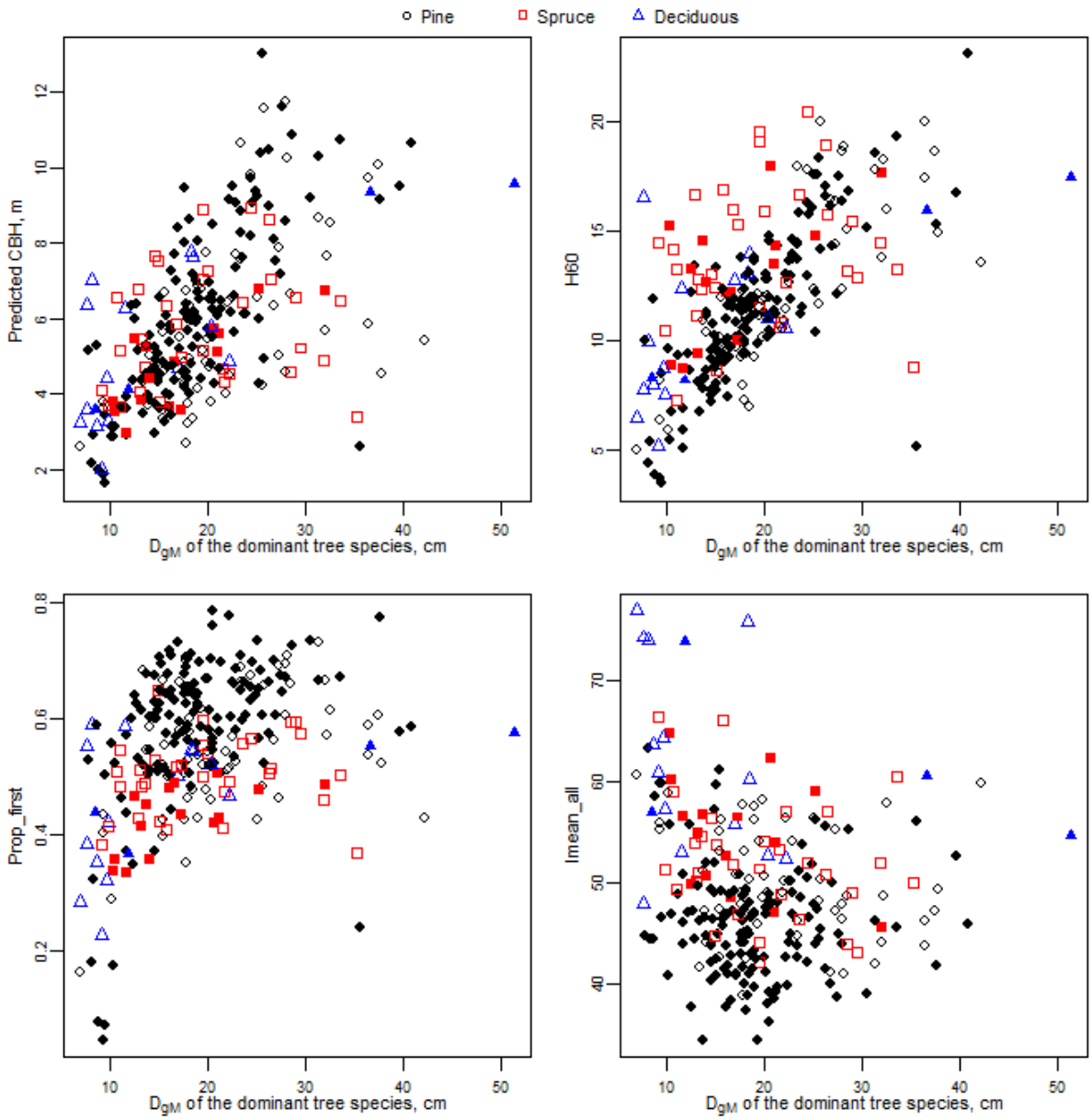

Fig. 2. Species-specific differences in selected ALS features of Kuhmo data, when the field-measured DgM is used in the x-axes to assess the invariance of the features to the size. Predicted *CBH* – crown base height, *H60* – the 60th height percentile, *Prop_first* – the proportion of the first-of-many returns to all returns above 2 m vegetation threshold, *Imean_all* – mean intensity value of all returns above the vegetation threshold. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

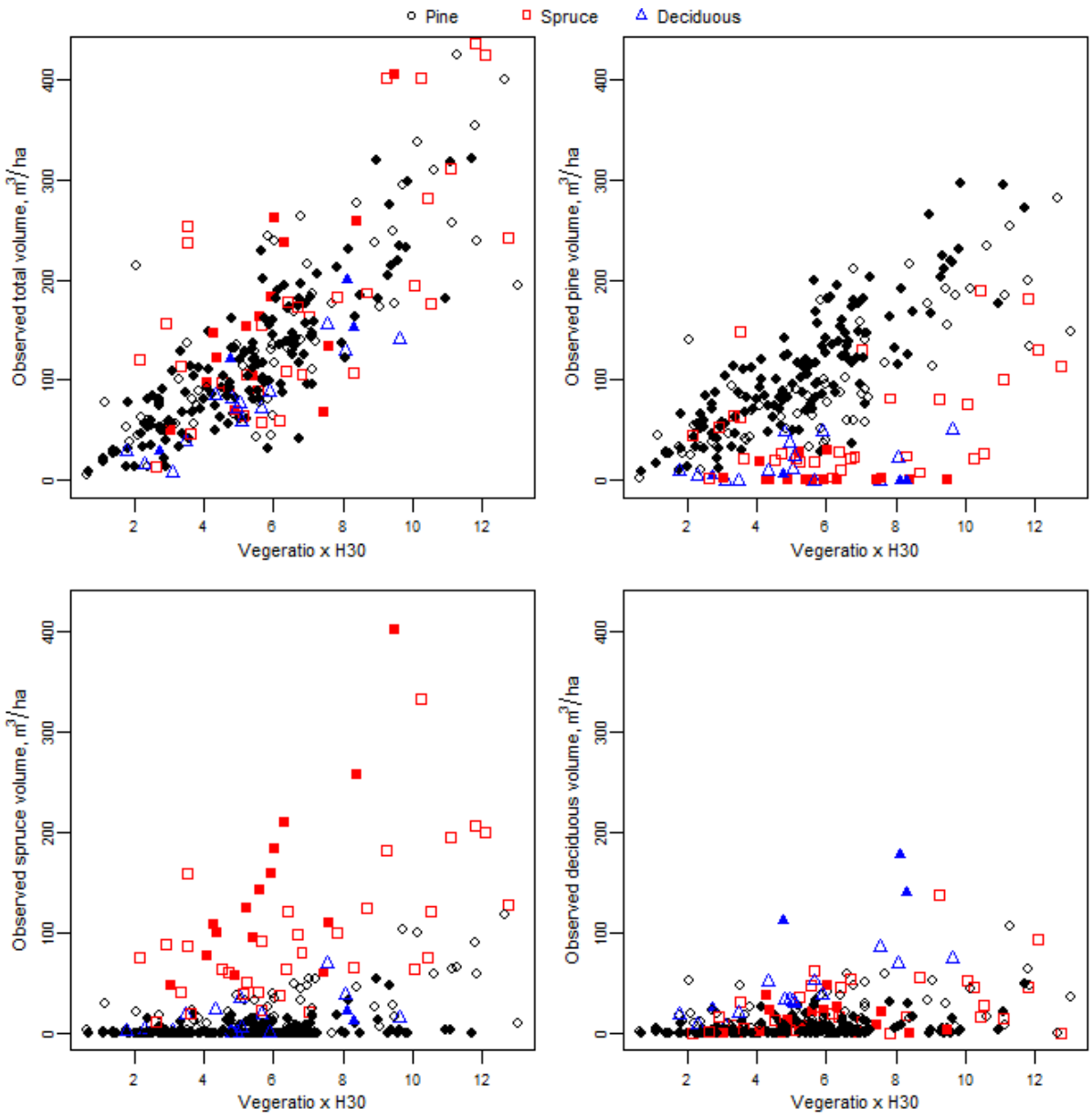

Fig. 3. Relationships between the species-specific volumes and the ratio of echoes above the 2 m vegetation threshold to all echoes (Vegeratio) × the 30th height percentile (H30) in Kuhmo data. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

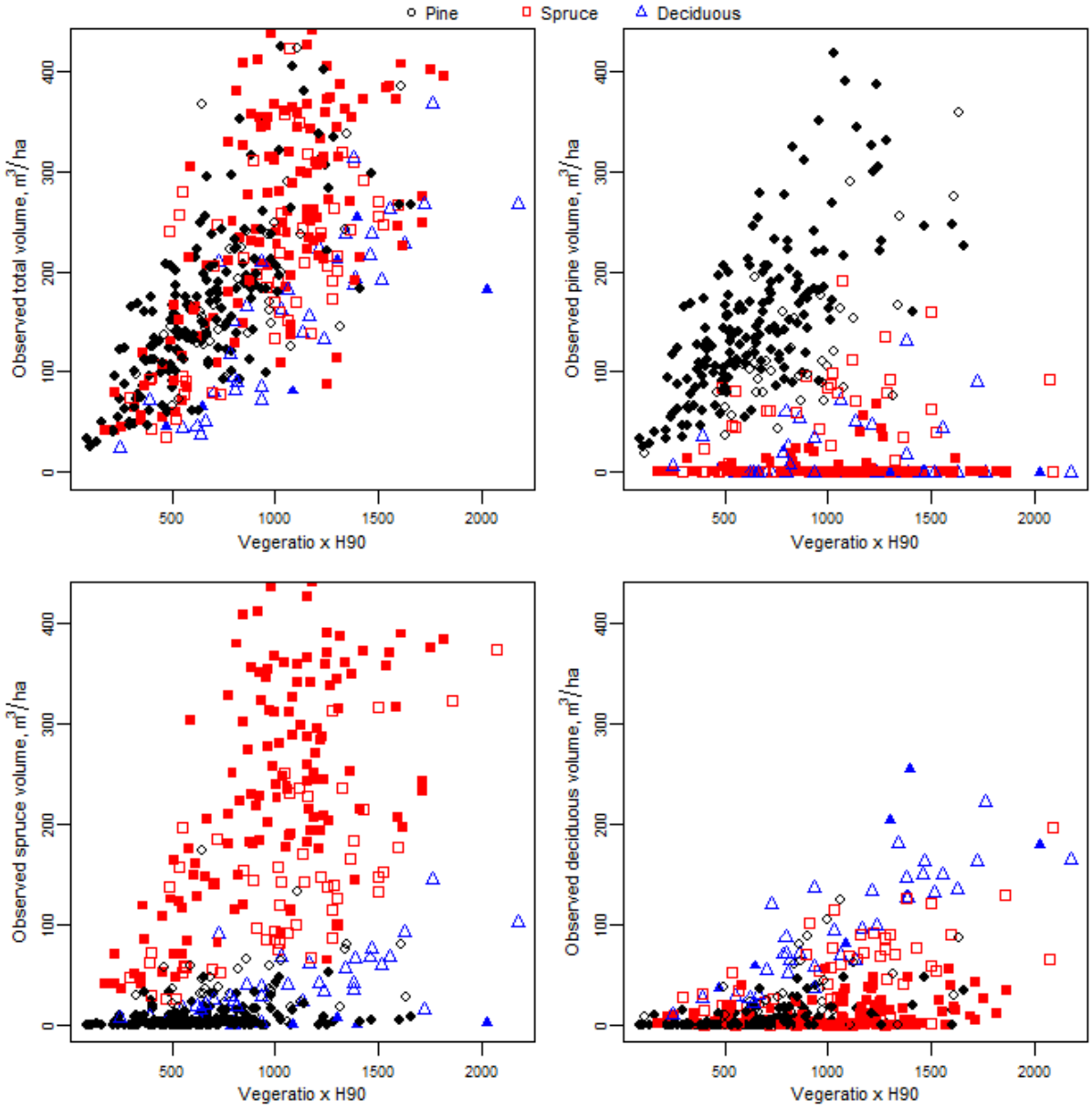

Fig. 4. Relationships between the species-specific volumes and the ratio of echoes above the 2 m vegetation threshold to all echoes (Vegeratio) × the 90th height percentile (H90) in Janakkala-Loppi data. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

4.2 Models for species-specific volumes

Before modeling the species-specific volumes with SUR, the predictor variables were systematically tested considering the goodness of the predictor features. Although the final composition of the predictor variables slightly varied depending on the species, usually the ratio of the echoes reflected above ground (2 m) to all echoes combined with a height percentile gave the best alternatives for volume models according to the coefficient of determination (R^2). This is reasonable since the first describes the density of the forest and together with the latter they form the components of the approximation of growing stock volume. However, for sample plots

of Kuhmo data which were dominated by the deciduous trees, the predictors describing intensity of returned ALS echoes were more appropriate. Considering Janakkala-Loppi in which the intensities were not available, the combination of height percentile and density feature proved to be successful.

The species-specific models employed in SUR, were typically composed of two ALS features and the dominant species information. All variables were most often significant according to the t-test for the model coefficients. The most essential results are presented in Tables 5 and 6 for Kuhmo data and, thus, the corresponding results for Janakkala-Loppi in Tables 7 and 8. However, producing the SUR composition, the significances tended to vary from the individual regressions. Due to that observation, the tests for the dominant species variables are presented in the last section of this chapter (Section 4.5). All of the models were fitted using the plots dominated by pine as the reference level. In practice, that means that applying the models without the species-specific coefficients, they will yield the species-specific volumes assuming that the dominant species of the plot is pine. Similar to the results mentioned earlier in this study, the structure of the model system differed depending on the dominant species in question.

Table 5. The SUR_1 model based on the Sp_{max+95} strategy to stratify the dominant species (Kuhmo).

Predictor[1]	**Vtotal**	**Vpine**	**Vspruce**	**Vdecid**
Intercept	-83.1778 ***	-72.1165 ***	-11.8327	-11.8591 *
Species				
P_{95}	-17.0412 *	12.47493 *	-13.6919 **	-14.5252 ***
S	21.61634 *	-89.1431 ***	99.7402 ***	7.486791 *
D	-41.1525 **	-87.2537 ***	-0.87776	44.85359 ***
ALS				
Vegeratio	167.1603 ***	113.0326 ***	54.32393 ***	-
H_{30}	12.73679 ***	-	-	-
H_{40}	-	10.46561 ***	-	-
H_{95}	-	-	-0.1767	-
H_{mean}	-	-	-	2.173142 ***
$I_{mean, first}$	-	-	-	0.175805 .

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

[1] Species: pine with $G \geq 95$ % (P_{95}), spruce (S) or deciduous trees (D). The I, D, and H refer to intensity, density, and height metrics; Vegeratio is the ratio of echoes above 2 m height to all echoes; and Prop_first is the proportion of first echoes to all echoes. The subscript indicates which descriptive statistic or percentile value was used and whether it was applied to a proportion of the echoes (sd=standard deviation, first=first-of-many echoes).

Table 6. The SUR_1 model based on the Sp_{75+95} strategy to stratify the dominant species (Kuhmo). For the abbreviations used, please refer to Table 5.

Predictor	**Vtotal**	**Vpine**	**Vspruce**	**Vdecid**
Intercept	-99.86995 ***	-63.060416 ***	-32.585827 **	-15.382497 **
Species				
P_{95}	-12.04051	0.247694	-3.370746	-8.13935 **
S	30.36624 *	-108.00624 ***	129.469296 ***	5.270839
D	-14.26998	-117.48788 ***	-0.555964	100.272105 ***
M	9.99804	-42.385164 ***	35.570611 ***	15.234861 ***
ALS				
Vegeratio	147.73339 ***	87.519629 ***	56.8353 ***	-
H_{30}	15.05592 ***	13.061476 ***	-	-
H_{sd}	-	-	2.170459 *	-
H_{mean}	-	-	-	1.632179 ***
$I_{sd,\ first}$	-	-	-	0.413249 *

Table 7. The SUR_2 model based on the Sp_{max+95} strategy to stratify the dominant species (Janakkala-Loppi). For the abbreviations used, please refer to Table 5.

Predictor	**Vtotal**	**Vpine**	**Vspruce**	**Vdecid**
Intercept	-138.995661 ***	75.126877***	-170.79225 ***	-43.9928872 ***
Species				
P_{95}	-4.347771	23.08388 **	-22.815650 *	-6.4725531
S	35.292819 ***	-133.981972 ***	173.967805 ***	-4.0344517
D	-42.77166 ***	-137.084530 ***	21.693204 .	73.4335707 ***
ALS				
Vegeratio	1.480844 ***	0.830904 ***	-	0.6522456 ***
H_{90}	-	-		1.7371693 ***
H_{70}	-	-	-0.689629	-
H_5	-	2.940213 ***		-
H_{mean}	21.446159 ***	-	17.909888 ***	-

Table 8. The SUR_2 model based on the Sp_{75+95} strategy to stratify the dominant species (Janakkala-Loppi). For the abbreviations used, please refer to Table 5.

Predictor	**Vtotal**	**Vpine**	**Vspruce**	**Vdecid**
Intercept	-139.50374 ***	68.145675 ***	-151.228556 ***	-55.347138 ***
Species				
P_{95}	-5.128199	12.799513	-17.497792 .	-0.929856
S	42.744521 ***	-156.033676 ***	213.227368 ***	-11.583733 **
D	-67.782674 **	-164.337313 ***	-12.595459	108.811071 ***
M	1.838347	-99.562928 ***	67.749118 ***	34.580047 ***
ALS				
Vegeratio	1.289625 ***	0.563688 ***	-	0.720063 ***
H_{90}	-	-	-	1.952479 ***
H_{70}	-	-	-2.983202 ***	-
H_{20}	-	5.148116 ***	-	-
H_{sd}	-	-	-	-
H_{mean}	22.251249 ***	-	18.332841 ***	-

Besides the SUR models the k-MSN imputation was also provided for Kuhmo data. Those results are presented and compared in the Table 9 according to the most essential stratifying strategies. For further comparison, that table also includes the accuracies of the SUR models for Janakkala-Loppi. Also the corresponding results are presented in a graphical form in the Figures 5, 6 and 7 which are presenting the observed total and species-specific volumes versus the corresponding predicted values for SUR_1 of Kuhmo, k-MSN of Kuhmo and SUR_2 of Janakkala-Loppi, respectively. The k-MSN predictions were generally more accurate than those obtained by SUR except when including the dominant species information (Sp_{max+95}) to the model of deciduous predictions. However, the comparison is originally based on the k-MSN applied with k=5, which has produced the most accurate predictions with this method in the original study.

Using both methods and both training datasets, the predictions regarding the total volume were well in line with the observed values (Figures 5, 6 and 7). As a difference to the Kuhmo, the predictions for pine volumes in pine dominated plots were worse in line than the predictions for spruce in spruce dominated plots in Janakkala-Loppi (Fig. 7). In all cases, the predictions of the minor species had considerably lower accuracies. Due to the coefficient structure of the SUR model of Kuhmo (Table 5), the predictions could not show values between 50 and 100 m^3/ha of the spruce volume (Figure 5). Hence, the predictions also saturated at certain values

(150 m^3/ha for spruce) whereas the true observed volumes were considerably higher (e.g. 400 m^3/ha for spruce). Considering the accuracy of the k-MSN, the models were better in line with observed values, also in the models of spruce and deciduous. The aforementioned saturating problem noticed in the SUR of Kuhmo (spruce and also deciduous), was also perceived in the SUR models of Janakkala-Loppi but the problems were observed in the model constructed for pine and deciduous. The coefficient structure (Table 7) for pine and deciduous forced to saturate the predictions at certain values. However, the pile was slightly skewer and wider than in the predictions of spruce in Kuhmo data (cf. Figures 5 & 7). Presumably, the problems in modeling are related to the predictors' ability to predict but also to the mean volumes (Table 1 & 2) and the proportions of the plots dominated certain species. In Kuhmo with Sp_{max+95} strategy data, there was the proportion of pine dominated plots of 79 %. Whereas the proportions for Janakkala-Loppi were 47 % and 44% for pine and spruce, respectively. The SUR compositions had challenges to model species-specific models if some species had strong dominance according to either on high volumes rates or amount of plots dominated by certain species. For example, the Kuhmo had very strong pine dominance, which had an effect on ability to predict plots dominated by spruce and deciduous (Figure 5). Whereas in Janakkala-Loppi, the plot proportions between coniferous species were quite equal but the spruce dominated plots had considerably higher observed volume values and wider projection scale (Figure 7). However, in the data of Janakkala-Loppi slight systematic inaccuracy was observed in the residual structure of species-specific model of spruce in which the volume predictions were overestimated on low observation values and underestimated on high observation values. To sum up, it seems to be challenging to get field data from forest which would have the most optimal species and age compositions for species-specific modeling with SUR.

The inclusion of the main species improved both the prediction types (SUR & k-MSN) considerably (Table 9). In addition, the achieved relative improvement was greater in the more heterogeneous data of Janakkala-Loppi. Using Sp_{max+95} as the information of the dominant species in data of Kuhmo, the RMSEs of pine, spruce, deciduous, and total volumes improved by 28.9 %, 25.4 %, 12.6 %, and 1.9 %, respectively, using SUR_1, whereas the corresponding species-specific figures for k-MSN were 16.4 %, 13.3 %, and 13.6 % for pine, spruce and deciduous, respectively. For comparison, the corresponding values for the RMSEs of Janakkala-Loppi were 36.4 %, 36.9 %, 20.9 % and 9.4% for pine, spruce, deciduous and total, respectively. However, using the k-MSN method with the species restriction degraded the accuracy of the total volume by 2.4%. In the case of k-MSN, the species-specific improvement was particularly

due to removing close-to-zero observations from the plots dominated by certain species which employed the dominant species restriction for the neighborhood. This restriction however reduced the number of potential nearest neighbors for some plots and therefore had a degrading effect on certain accuracy levels.

Table 9. RMSEs (m^3/ha) of the MSN/SUR_1/SUR_2 predictions for Kuhmo (MSN & SUR_1) and Janakkala–Loppi (SUR_2) with different strategies to stratify the dominant species when evaluated in the training data.

	Dominant species information		
Tree species	**-**	**Sp_{max+95}**	**Sp_{75+95}**
Pine	42.6 / 52.9 / 84.4	35.6 / 37.6 / 53.7	39.0 / 41.9 / 54.8
Spruce	36.0 / 47.6 / 108.1	31.2 / 35.5 / 68.2	33.2 / 39.3 / 66.1
Deciduous trees	21.3 / 21.5 / 34.9	18.9 / 18.8 / 27.6	18.4 / 17.1 / 27.6
Total	50.2 / 53.5 / 59.6	51.4 / 52.5 / 54.0	52.2 / 52.2 / 54.8

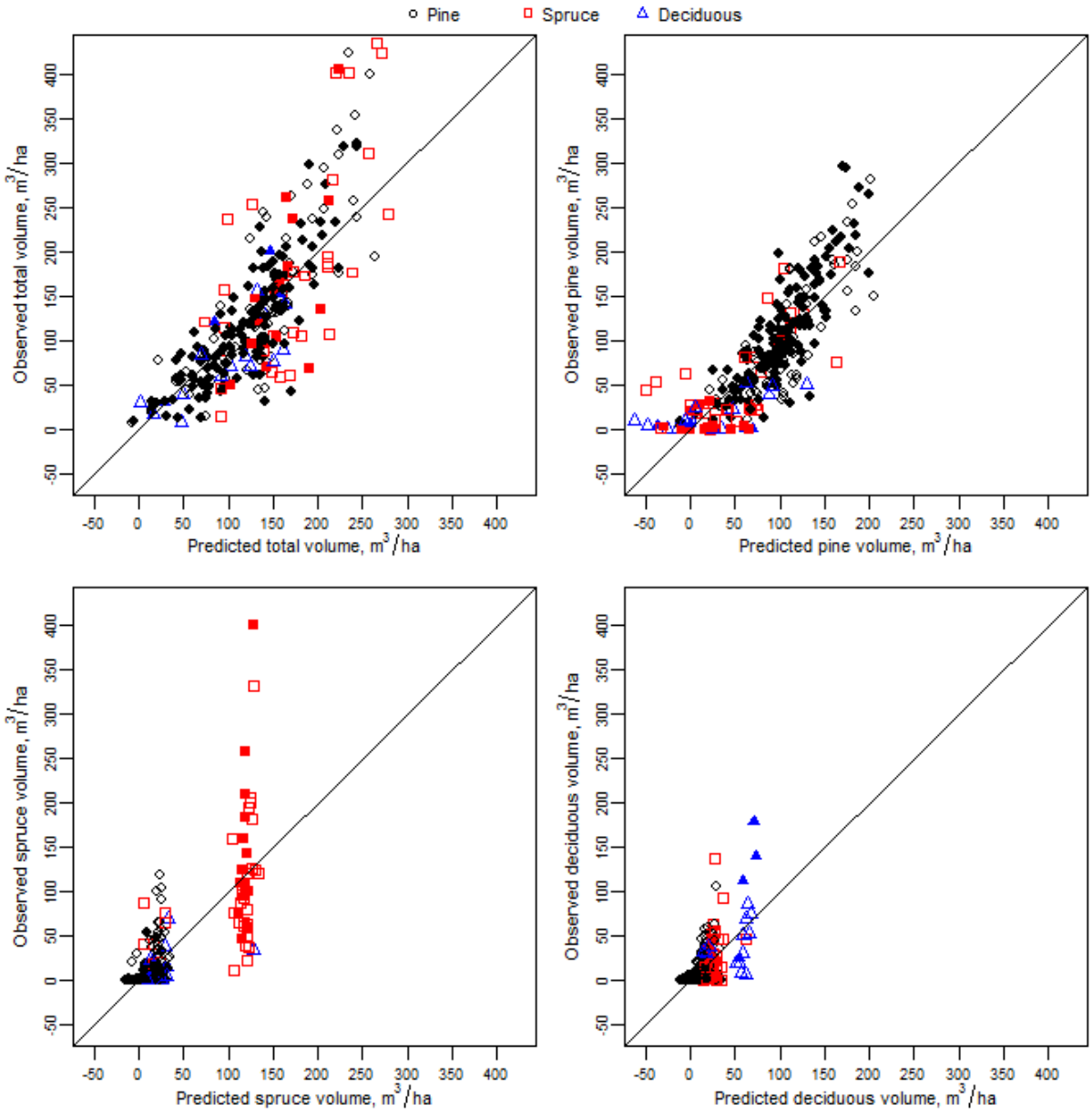

Fig. 5. Predicted vs. observed species-specific volumes in the Kuhmo training data based on the SUR_1 model structured in the Table 5. Thus Sp_{max+95} strategy to stratify the dominant species has been implemented. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

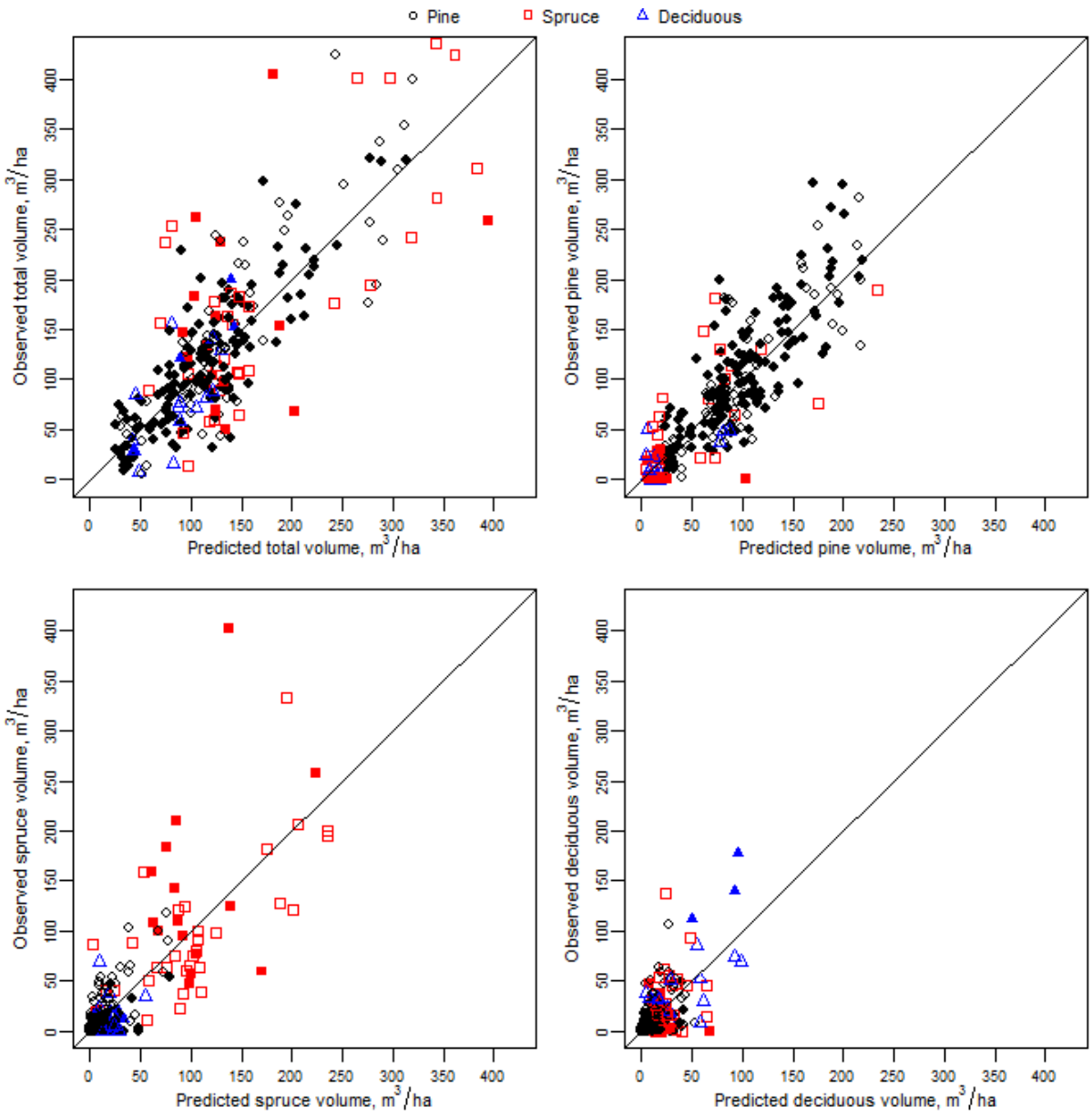

Fig. 6. Predicted vs. observed species-specific volumes in the Kuhmo training data based on the k-MSN imputation using k=5 and a neighborhood restricted by Sp_{max+95} (Table 4). The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

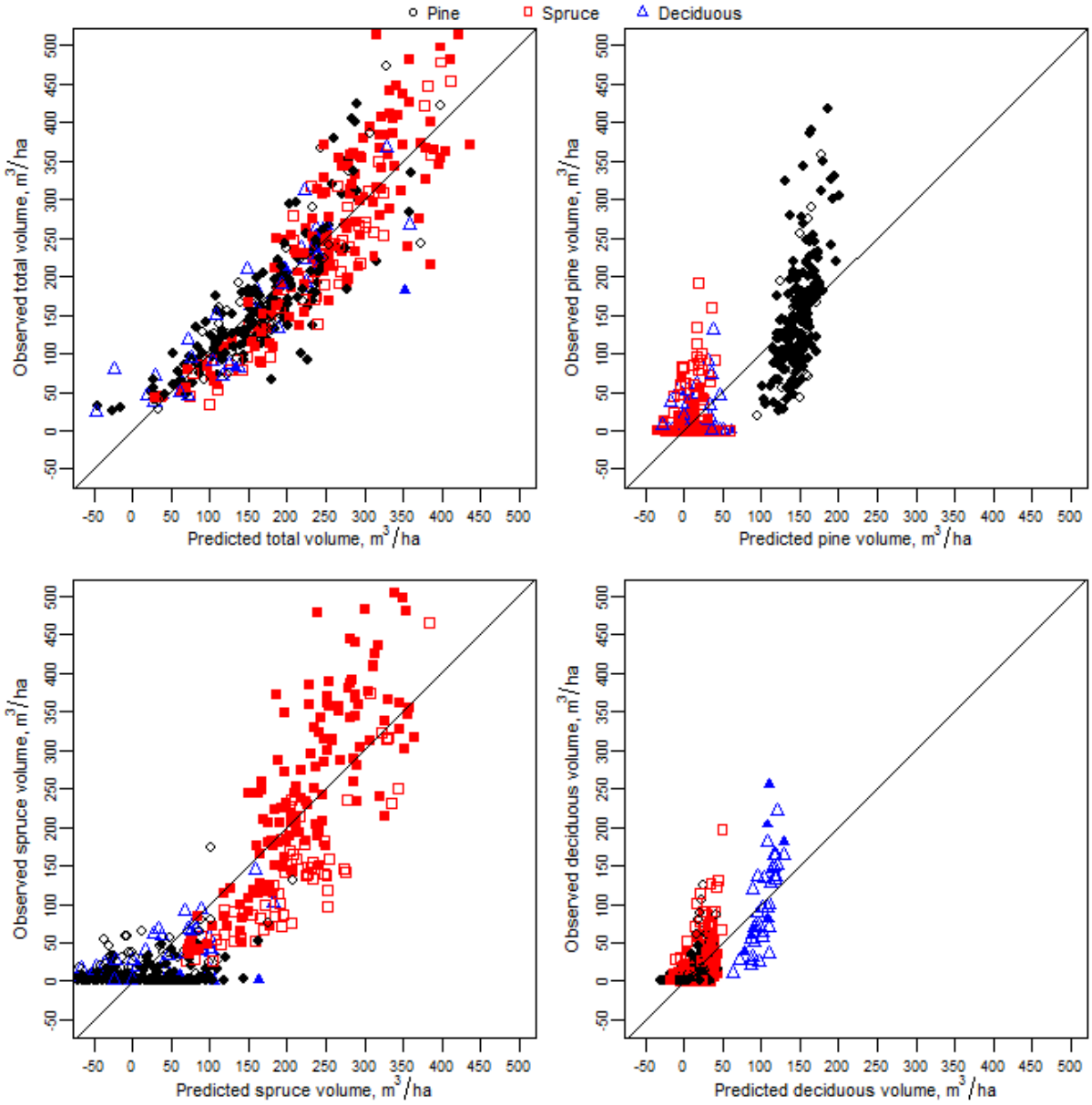

Fig. 7. Predicted vs. observed species-specific volumes in the Janakkala-Loppi training data based on the SUR_2 model structured in Table 7. Thus Sp_{max+95} strategy to stratify the dominant species has been used. The solid symbols have been used if the basal area proportion of the dominant tree species is ≥ 75 %.

4.3 Classification of the dominant species

The accuracies of the attempts to predict the dominant species using LDA in Kuhmo dataset are summarized in Table 10. The species stratification with only three classes (Sp_{max}) was the most simple to predict and these predictions also yielded the best results: overall accuracies of 73.6 % and 76.2 % and kappa coefficients of 0.40 and 0.48 using 3 or 4 predictors, respectively. When the plots were classified according to the ≥ 75 % species proportion, the most problematic case was naturally the "mixed" class including plots of lower dominance of the various species.

Excluding this class, the overall accuracy and the kappa of the classifier were 87.3 % and 0.56 (n=158), respectively, using 3 predictors. Considering only coniferous species with ≥ 75 % basal area proportions, the accuracy of the corresponding classification according to the overall accuracy was 91.6 %. This result ensures the same observation that has also been noticed in earlier studies: true pine and spruce areas can be distinguished with considerable accuracy by means of ALS (e.g. Holmgren & Persson 2004; for spruce among deciduous: Ørka 2007)

The inclusion of the pine plots with ≥ 95 % species proportion also complicated the classification and lowered the success rates. However, for the final volume predictions it was thought that it would be probably beneficial to distinguish true pine plots. Instead of increasing the number of classes in LDA, however, it was found equally accurate to select these plots manually based on thresholding of the predictor variables. Selecting the plots with ≥ 95 % species proportion manually was implemented and tested using the classification of Sp_{max} and Sp_{75} afterwards the LDA classification. In both cases, selecting the plots which had a standard deviation of the intensity values of all pulses < 30, a proportion of first pulses < 0.6 and a density in the 10th height percentile < 0.2 increased the overall accuracy by about 5–9 %-points, depending on the used model, compared to including a class with the pine plots with ≥ 95 % species proportion in the LDA. Applying these rules mainly resulted in confusion between plots with less pine (≥ 75 %) and decreased the overall accuracies although the dominance of the pine had been described better. It was pleasant to notice that the confusion between other tree species in manual thresholding did not just occur. Applying that idea while taking into account the strong pine dominance of the teaching area, the classified true pine plots will get more exact predictions for pine volumes. The poor result in LDA with ≥ 95 % species could be related to the priors applied with LDA, which had been set equal among the classes being classified. For these reasons, the manually composed classifications of Sp_{max+95} and Sp_{75+95} are presented in Table 10 and used later in this study.

Table 10. The achieved results of linear discriminant analysis with the dataset of Kuhmo. For the abbreviations used, please refer to Tables 4 and 5.

Classifier	Number of explanatory variables	Explanatory variables	Overall accuracy (%)	Kappa coefficient
Sp_{max}	3	$I_{mean, all}$ + Prop_first +D_{40}	73.6	0.40
	4	$I_{mean, all}$ + Prop_first + D_{40} + H_{60}	76.2	0.48
Sp_{max+95}	3	$I_{mean, all}$ + Prop_first +D_{40}	55.5	0.34
	4	$I_{mean, all}$ + Prop_first + D_{40} + H_{60}	57.7	0.39
Sp_{75}	3	$I_{mean, all}$ + Prop_first + D_{30}	58.5	0.34
	4	$I_{mean, all}$ + Prop_first + D_{40} + H_{70}	59.6	0.35
Sp_{75+95}	3	$I_{mean, all}$ + Prop_first + D_{30}	46.8	0.30
	4	$I_{mean, all}$ + Prop_first + D_{40} + H_{70}	45.7	0.28

4.4 Prediction accuracies

To observe the level of accuracies which are obtainable in a practical prediction, the dominant species predicted by LDA were combined with the fits of SUR_1 compositions. Moreover, the k-MSN predictions of the original study have also been presented for a comparison. The dominant species classification predictions which included 24–54 % of classification errors (Table 10) degraded the accuracies obtained earlier (Figure 8). The k-MSN method provided considerably more accurate results than the SUR models in species-specific cases. Nevertheless, when the dominant species had to be predicted with the aforementioned error levels, the insufficient classification was resulted in the predictions of dominant tree species by decreasing the accuracies of the volume models of dominant species. In conclusion and due to the LDA predictions, almost all the final prediction accuracies (Table 11 and Table 12) were worse compared to the corresponding predictions which have implemented without the information of the dominant tree species (Table 9).

The most accurate results based on SUR_1 were obtained using the model structured in Table 5 and LDA model with four explanatory variables to predict Sp_{max} and Sp_{max+95} (Table 11). Considering the RMSEs of the total volumes, there weren't considerable differences whether the dominant species was either predicted or observed (cf. Table 9, Table 11). However, the RMSE values of spruce were in particular considerably poorer and the predictions of pine plots in particular were biased (Table 11). The predictions of spruce plots were also seemingly biased especially when the LDA was run with four explanatory variables. Moreover, the vast gap in the RMSEs of the SUR_1 between the spruce predictions predicted with LDA of four or three independent variables could be related to the height percentile (H_{60}) of the Sp_{max} strategy. As we can see in Figure 2, H_{60} appears to be a quite adequate variable for distinguishing plots dominated by spruce.

The most accurate k-MSN predictions for the species-specific volumes were obtained using four explanatory variables to predict Sp_{max+95} (Table 12). Compared to the k-MSN predictions using the field-observed dominant tree information, predicting the dominant tree species degraded the RMSEs 25 %, 13 %, 15 % and 0.1 % for pine, spruce, deciduous and total volumes, respectively. The definition of the dominant species had generally less importance in the k-MSN predictions than those based on the SUR models (cf. values in Tables 9 & 12).

Table 11. RMSEs and (BIASes) of the species-specific volumes based on the SUR_1 models, when the dominant species were predicted by LDA. For the abbreviations used, please refer to Table 2.

Species information	Number of explanatory variables	m³/ha			
		Pine	Spruce	Deciduous	Total
Sp_{max}	3	52.2 (11.5)	60.0 (-8.3)	22.8 (-3.4)	56.8 (-0.5)
Sp_{max}	4	53.0 (14.2)	49.3 (-11.4)	23.2 (-3.8)	54.1 (-1.3)
Sp_{max+95}	3	51.7 (11.3)	61.1 (-8.1)	22.2 (-2.9)	58.2 (-0.2)
Sp_{max+95}	4	53.1 (13.9)	49.8 (-11.1)	22.6 (-3.1)	55.4 (-0.8)
Sp_{75}	3	56.7 (13.9)	61.3 (-8.1)	31.9 (-6.0)	54.9 (-0.6)
Sp_{75}	4	54.4 (11.5)	55.8 (-5.1)	31.8 (-5.8)	54.4 (0.2)
Sp_{75+95}	3	55.5 (12.7)	60.4 (-7.1)	31.8 (-5.1)	55.2 (0.2)
Sp_{75+95}	4	55.7 (12.8)	59.9 (-7.8)	31.3 (-4.6)	54.9 (0.0)

Table 12. RMSEs and (BIASes) of the species-specific volumes based on k-MSN, when the dominant species were predicted by LDA. For the abbreviations used, please refer to Table 2.

Species information	Number of explanatory variables	m³/ha			
		Pine	Spruce	Deciduous	Total
Sp_{max}	3	44.9 (0.9)	38.1 (-1.9)	22.2 (-0.6)	53.8 (-1.7)
Sp_{max}	4	44.8 (0.5)	35.3 (-0.9)	22.0 (-0.4)	52.0 (-0.8)
Sp_{max+95}	3	45.3 (1.6)	38.1 (-1.9)	22.2 (-0.8)	51.5 (-1.1)
Sp_{max+95}	4	44.5 (1.2)	35.2 (-1.3)	21.8 (-0.7)	51.5 (-0.8)
Sp_{75}	3	47.1 (-1.4)	38.4 (-2.8)	22.8 (-0.6)	54.6 (-4.8)
Sp_{75}	4	46.8 (-0.5)	38.2 (-2.2)	22.5 (-0.6)	51.9 (-3.3)
Sp_{75+95}	3	45.9 (-2.4)	39.6 (-3.0)	22.6 (-0.7)	53.5 (-6.0)
Sp_{75+95}	4	47.1 (-1.6)	37.7 (-2.4)	22.9 (-0.3)	53.0 (-4.3)

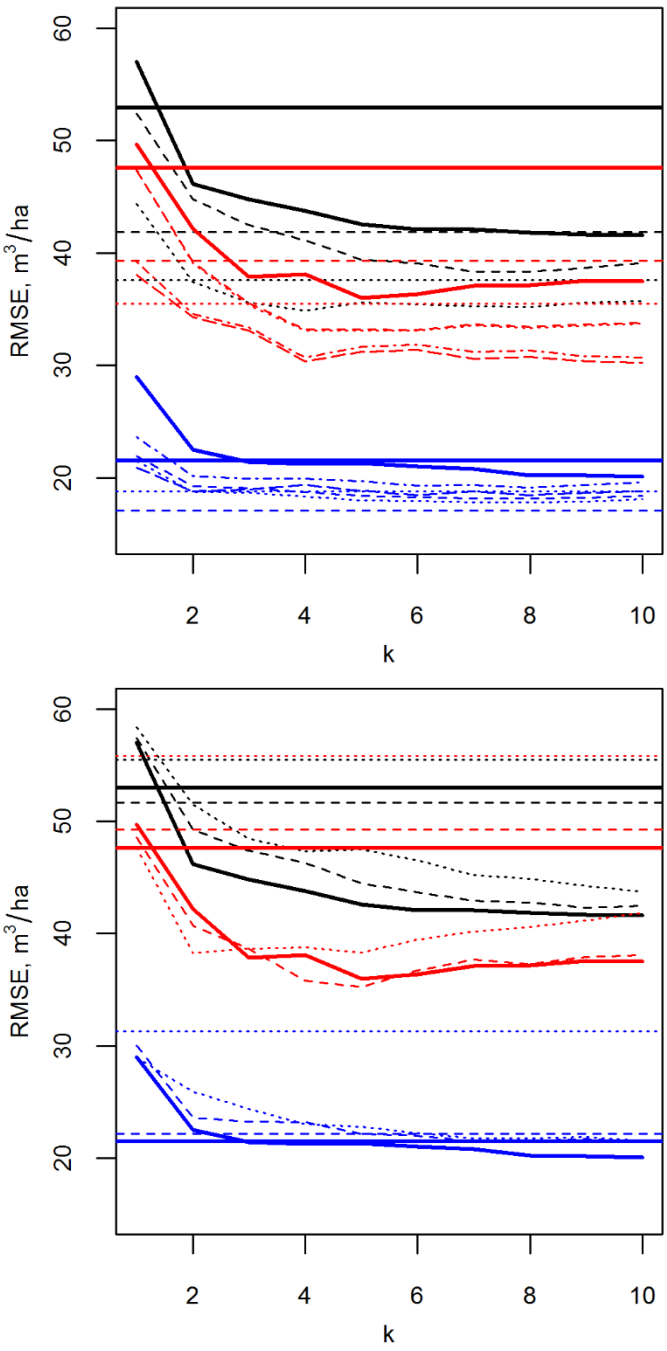

Figure 8. Comparison of the RMSEs obtained by the different approaches in the model fitting (above) and prediction (below). The broken and solid lines indicate the accuracies with and without dominant species information, respectively. The horizontal lines give the accuracies of the best SUR models and the other lines those of the k-MSN predictions with k=1–10. For the color codes used please refer to the previous figures. (Figure © Jari Vauhkonen)

4.5 Significance of the coefficients in the fitted models

During the process of SUR modeling, it was noticed that all of the categorical variables were not essential for the final fitted model group. Examining the results of the fitted SUR models in Tables 5, 6, 7 and 8 reveals that all of the coefficients have not deserved the significance according to the implemented t-test at the risk level of 5 %. To take into account the interactions between the categorical variables of various equations included in the SUR composition, the χ^2 was used for the Wald test.

Considering the data of Kuhmo, separating the plots with G ≥ 95 % of pine significantly affected the model predictions which were based on Sp_{max+95} (Table 5) whereas this information was insignificant using the Sp_{75+95} strategy to stratify the species (Table 6). The importance of the pine class of G ≥ 95 % was thus tested for the two most important fitted SUR groups (Sp_{max+95} and Sp_{75+95}). According to the test, it was proven that the true pine class is redundant in the case of stratifying the dominant species by Sp_{75+95} and thresholding the risk level at 5 % (Table 13). Generally, the importance of the true pine class was more significant in the Kuhmo data, which could be interpreted to be related simply to the proportions of the tree species, exactly to the strong mature pine dominance. Presumably, considerable changes in the RMSEs of the species-specific models won´t be observed when the redundant classes would be removed. However, it is reasonable to keep the structure of the volume models as simple as possible. The class of the deciduous trees had significant role in every case tested. The deciduous trees mainly had the role of the minor species in the both datasets used in this study.

Considering a chance to observe the redundant categorical variables in the single models of the SUR composition, it was noticed straight from the p-values of the t-test (see asterisks in the Tables 5–8) that there would be possibilities to simplify the models. This kind of simplifying would be more complicated because therefore the classification should be individually implemented for every species-specific models. However, stepwise test procedures was implemented to the species-specific models by the way of trying to observe the most insignificant variables of SUR group and thus setting the coefficients to zero. The first stage of the system was to decide the first, the most redundant, variable according to the p-value of the t-test and then the second variable was chosen by the Wald-test. Actually, the coefficients were added to the zero hypothesis according to the significance of the t-test while the Wald-test was used in checking the significance of the whole zero hypothesis on every step. This procedure was repeated until

the determined level of risk (5 %) was exceeded. The testing was operated above-mentioned way due to the comparability and efficiency. Setting the coefficients to zero one-by-one according to the t-test, would had forced to create vast amount of different dominant tree vectors. It is more reasonable to assess the significance according to the one p-value than many of them.

The results of the aforementioned test procedure were presented in the Table 14. In conclusion, it could be noticed that the stratifying strategy that includes less classes (Sp_{max+95}) is also including less non-significant categorical variables in both cases. The first variable to be removed in Kuhmo data showed to be deciduous class in the model of spruce volume. The same variable was also noticed to be redundant in the data of Janakkala-Loppi. It could be explained by the shortage of the mature spruces in the plots dominated by e.g. birches. Thus, significant difference between pine dominated and birch dominated plots could not have been observed and it is quite indifferent for spruce volume if the plot was dominated by deciduous or pine (pine is the reference level of the model). In practice, it is often noticed that understory of spruce are quite common in pine and birch dominated stands. In the case of Janakkala-Loppi, the 95-class is mentioned more often than in the data of Kuhmo, which is also noticed in the results of the Table 13. Generally, the results proved that the species proportions of the training data are affecting strongly to the significances.

Table 13. The results of testing significance of the two categorical variables of interest in the SUR fittings presented before (Tables 3–6). For the principle of the dominant species codes please refer to Table 4.

	The p-value of the Wald-test(χ^2)	
Kuhmo	**Pine G ≥ 95 %**	**Deciduous**
Sp_{max+95}	1.90E-07	2.20E-16
Sp_{75+95}	0.06791	2.20E-16
Janakkala-Loppi		
Sp_{max+95}	0.006731	2.20E-16
Sp_{75+95}	0.4544	2.20E-16

Table 14. Assessing significances of the categorical variables when adding the variables for restriction stage-by-stage until the zero hypothesis was rejected (p-value of the Wald test (χ^2) at risk level of 5 %). The letter code[1] indicates the equation considered in the SUR group while the number code[2] is representing the variable in question. The p-values are presented in brackets.

	Kuhmo		Janakkala-Loppi	
Number of the stage	**Sp_{max+95}**	**Sp_{75+95}**	**Sp_{max+95}**	**Sp_{75+95}**
1	S3 (0.9249)	S3 (0.9775)	T95 (0.58163)	D95 (0.833)
2	S1 (0.457)	P95 (0.9991)	D2 (0.3993)	T0 (0.954)
3	D1 (0.1379)	S95 (0.9649)	D95 (0.3346)	S3 (0.9532)
4	(0.01322)	T3 (0.954)	S3 (0.1659)	T95 (0.9086)
5	-	D2 (0.8593)	(7.03E-04)	P95 (0.5356)
6	-	T0 (0.6038)	-	S95 (0.6571)
7	-	(0.01549)	-	(0.04945)

[1]Equations for the volume models: S –spruce, P – pine, D – deciduous, T – total
[2]Variables: 1 – pine (maximum G), 2 – spruce (max. G or ≥75 %), 3 – deciduous (max G or ≥75 %), 95 – pine (G≥95 %), 0 – "mixed" (species ≤75 %)

5 DISCUSSION

Finally, the improvement of the pre-classifying can be observed in the predictions of the fitted models. The acquired improvement of the RMSEs in the species-specific volumes (Sp_{max+95}) were at best about 20.9–36.9 % using the SUR model in the data which didn´t have a clear dominance of certain species. In the case of strong pine dominance, the RMSEs decreased as well but the improvement was smaller being 12.6–28.9 % depending on the species-specific model. The k-MSN results of the original study was also presented for a comparison and noticed to be slightly better than SUR in the case of strongly pine dominated data. All in all, the accuracies of the fitted models were slightly more accurate with the earlier studies of predicting species-specific attributes without individual tree lineation. For example, Vauhkonen et al. (2012) had ALS-only plot-level RMSEs of 65 m³/ha (43 %), 76 m³/ha (114 %), 70 m³/ha (133 %), 48 m³/ha (161 %) for total, pine, spruce and deciduous in unbalanced field-data, respectively. Whereas the corresponding results of Janakkala-Loppi were 54 m³/ha (26 %), 54 m³/ha (70 %), 68 m³/ha (65 %) and 28 m³/ha (119%) for total, pine, spruce and deciduous, respectively. The Kuhmo data gave more accurate species-specific results, which can be explained by the homogenous species structure. The corresponding figures for Kuhmo were 53 m³/ha (40 %), 38 m³/ha (43 %), 36 m³/ha (124 %) and 19 m³/ha (16 %) for total, pine, spruce and deciduous, respectively. In this case, also the intensity variables were used which have proven to be

appropriate in distinguishing tree species (e.g. Törmä 2000; Ørka et al. 2012). However, in the SUR models of Kuhmo data set, the intensity variables were only used in the model for deciduous volumes since the power of intensity variables to predict volumes was not really excellent. The advantage of the intensity variables were absolutely important when implementing dominant species classification by LDA.

The pre-classification of dominant species to attain more accuracy for forest attribute models has also been used before in the researches of forestry field with encouraging results. Nevertheless, the idea of this study has some unique specifications which have not been seen in previous studies. For example, Maltamo et al. (2015) tested to improve k-NN-based species-specific models by stratifying the reference data according to the main tree species and development stages. Aerial photographs, field data and ALS data were used in the stratification. They observed slight improvisations in the models that were stratified according to the ALS forest structure or photo-based stratifying which was corrected with field observations. Moreover, Maltamo et al. (2006) have also proven that the field-based class variables can give noticeable improvement for the stand and plot predictions based on ALS. All in all, the first aforementioned study is absolutely not comparable to this study because, at first, the forest structure in Norway may be considerably different and, secondly, the stratification has emphasized on aerial photographs. The latter study is comparable better to this study although that study only includes predictions for total volumes. Also the earlier study of Pippuri et al. (2013), proposes the potential of pre-classifying idea by using various species proportions as a predictor in regression and k-NN methods. The data of the study was mainly collected from urban forest environments. Absolutely the environment differs from this study although the idea of stratifying the dominant trees was quite similar. However, the structure of the forest are not equal with this: here the coniferous dominated forests are used whereas Pippuri et al. (2013) have focused on deciduous, i.e. hardwood, species.

The discrimination of the tree species was turned out to be a bottleneck in this study. Finally, replacing the dominant species information by LDA predictions in models, the accuracies of the volume predictions were clearly degraded (Table 9 and 11). However, the previous studies have proven that distinguishing tree species solely from ALS data can be challenging. For example, Ørka et al. (2007) tested classification according to the intensity variables implementing the LDA, as well as in this study, with the overall accuracy at most 74.1 %. As difference, it

should be noticed that most studies, such as the aforementioned, have used the denser resolution of the ALS data, which is one subject to give more accuracy to the results. The best overall accuracy of LDA classification was 76.2 % in this study. Since the classification problem would have been much more complicated in the case of Janakkala-Loppi and moreover the intensity variables were not available, it was really reasonable to leave it out of the LDA. For example, the Figure 1 is demonstrating well the overlapping of the species in aforementioned data.

Although the classification of the species by attributes extracted from ALS was noted to be challenging, the potential of the SUR method in modelling the species-specific attributes must be taken to further examination. This study offers the results of SUR modelling with two considerably different training data-set. Also the k-MSN predictions are provided from the original study. Considering the accuracies of the models of Kuhmo (Table 9), the accuracies of species-specific volume models improved by 5.6 %, 13.8 % and 2.1 % for pine, spruce and total when using Sp_{max+95} strategy to stratify the dominant species and the k-MSN method instead of the SUR_1. The RMSE of deciduous degraded 0.5 % due to the lack of adequate neighbor plots which would have been dominated by deciduous species. Implementing the SUR method, it was noticed that some non-dominant species, according to the total amount of plots in area or mean volume inside all plots, tended to construct model structure which was not able to predict properly on high or low volume values and the predictions saturated on the narrow range of values (see Figures 5 & 7). Presumably, reason is related to the power of the variables to describe those "minor" species. In spite of those results, the potential of SUR-method must not be rejected. According to the previous studies, the k-NN methods are able to give better results than regression based methods (e.g. Pippuri et al. 2013; Maltamo et al. 2009b) but the ability of regression-based methods will be noticed in special cases, such as studies which do not have vast training data to predict the targets. This kind of problems was encountered in Kuhmo data, where the amount of deciduous plots were too small for having a quality neighborhood in NN imputation. Also the possibility to calibrate the reference data by stand-wise data are kept an advantage of the method although the accuracies could not reach the same level as non-parametric models are able to achieve (Maltamo et al. 2012). For instance, geophysical properties of each stand are giving an effect on the growth, and it can be taken into account with regression-based methods.

Implementing predictions by means of k-NN imputations, the amount of the predictor variables are often kept high without a doubt of any overfitting problems (Maltamo et al. 2009b). In this study, the models of SUR composition were constructed of tree independent variables of which one was the categorical variable illustrating the dominant tree species. In both modeling cases (Kuhmo & Janakkala-Loppi), the variables were searched manually using graphical assessments, coefficient of determination and p-values of t-test. Here, this method can be considered as a reasonable alternative due to the small amount of the potential variable candidates. Thus, the best alternatives have been efficiently found, and it would be unlikely to get better accuracy out of the models by searching the variables automatically. Also the manual selection of LDA variables from the set of ALS variables in data of Kuhmo, can be explained by the quite small amount of candidates. The discriminating variables were more efficient to search by graphical assessments. The biggest surprise was the discrimination ability of ALS-based crown base height that turned out to be unusable in this training data of Kuhmo even if the previous studies (Holmgren et al. 2008, Holmgren & Persson 2004) have given some promising results. The reason for the difference in the distinguishing ability between dominant species with CBH could be explained by the diameter distribution of the dominant species which emphasized on small diameters, exactly for spruce and pine. The diameter distribution has also caused some confusion with other variables, which can be seen in the Figure 2. For example, the ability of Imean_all and Prop_first to distinguish the ≥ 75 % plots of pine and spruce in the levels of D_{gM} of over 25 cm was better compared to the plots of small diameters.

Considering only the fitted models of this study (Table 5–8) the outcomes between different study areas can be discussed. In conclusion, the inclusion of the tree species classifiers in SUR gave better improvement in the more heterogeneous data with respect to the strongly pine dominated data (Table 9). The classifier of four separate classes (Sp_{max+95}) was proven to be better than the classifiers with five different classes in both datasets. However, it was noticed that the RMSEs of Janakkala-Loppi was considerably high and the suitability of the pre-classification should be examined. In the Figure 1, the amount of spruce dominated plots are quite equal with the pine dominated plots and the confusion between those plots seems to be really probable using only the ALS variables describing the structure of the canopy and understory. Thus, the lack of intensity data may have caused problems in species-specific modeling although the intensity variables were not outstanding good for predicting species-specific volumes in Kuhmo data. Nonetheless, it would have been interesting to clarify the explanatory power of the inten-

sity variables in species-specific volumes when the more complicated forest structure was examined. The importance of intensity variables in distinguishing pine and spruce has been noticed in the outcomes of numerous studies (e.g Holmgren & Persson 2004; Vauhkonen et al. 2009). Examining the species-specific volume models of Kuhmo (Tables 5 & 6), the intensity variables were not widely used, anyhow. The reason for this could also be explained by the lack of plots dominated by spruce. Earlier studies have proposed that the intensity variables are useful precisely in the discrimination between spruce and pine at least when employing high ALS densities (e.g. Holmgren & Persson 2004). Although the same strategies to stratify the dominant species in training data were used in both datasets to get better ability to compare them, the better performance could have been achieved by calibrating the classes according to the more heterogeneous forest composition in the case of Janakkala-Loppi. The place to develop the fitted SUR models would has been in the ≥ 95 % class. In Janakkala-Loppi, the separate class for spruce, for example G ≥ 95 %, would has been reasonable due to the discrimination observed in Figure 1.

One of the objectives of this study was to evaluate differences between the used datasets. As the results proved, the more complicated species compositions will have an effect on the predictions (Table 9). To sum up the main differences between the results, the RMSEs of the species-specific volumes in Janakkala-Loppi were noticeably higher than in the Kuhmo dataset. However, the relative improvement of employing the dominant tree species information was more important, although species-specific predictions were more inaccurate than corresponding results in Kuhmo dataset. The reason for the inaccuracies could be explained by species compositions because the differences between specifications of the ALS data acquisitions were quite insignificant. The data of Janakkala-Loppi was lacking of the ALS intensity values, but, instead of volume models, the intensities values have especially been proven to be advantageous in classification of the tree species and most advantageous they are in the individual tree approaches. That is why, the operationally used forest inventory methods in Finland, proposed by Packalén & Maltamo (2006), are incorporating aerial photographs with sparse ALS intensity and structure data to derive species-specific information. To give some guidelines for dominant species classification structures, the species compositions seem to have an important role. As the implemented tests proved (Table 13 & 14), the same classification strategy of the dominant species as in the Kuhmo data, did not work effectively in the Janakkala-Loppi data set. Nevertheless, the inclusion of the supplementary dataset ensured the hypothesis that the classification of the dominant species will improve species-specific volume predictions. Although, the more

heterogeneous species compositions have an effect on the accuracies of the predictions, and RMSE values are quite far away from the values that are sufficient for practical forestry management needs.

For the subsequent classification attempts, the tests for categorical variables describing the various dominant species were presented in section 4.5. Also the tests were employed to produce some numerical evidence of usefulness and significance of the dominant tree classifications strategies which are tightly bounded to the basic theory and objectives of this and the original study. The results of testing variables with χ^2 for Wald-test showed and ensured the assumptions discussed above. According to the Table 13, the true pine class showed to be redundant in the cases of stratifying strategies which were implemented with five dominant species strata. In cases of Sp_{max+95}, the whole SUR composition had significant results of the zero hypothesis when the true pine variable was evaluated at the risk level of 5 %. However, considering the single equations (Table 14) in the SUR compositions of Janakkala-Loppi, significances for total and deciduous volumes were unsuitable when the true pine class was evaluated. The testing results did not reveal the common guidelines for subsequent classifications but anyhow the importance of knowing the species compositions and development classes should be taken into account in implementing subsequent classification. The most problematic will be the true class with high basal area proportions. On the other hand, it could give slight accuracy improvement for predictions in simple cases of simple species composition, like in the final predictions using k-MSN in Kuhmo data. However, considering the fitted models, only the RMSE of the deciduous model get improvements in case of Kuhmo. For Janakkala-Loppi, the very slight improvement was achieved adding the ≥ 95 % class for Sp_{max}-classifier. The inclusion of the number of classes can be related to the problem of degrading accuracy observed by (Heinzel & Koch 2011). For example, the data of Kuhmo was quite simple by species compositions thus the pine ≥ 95 % did not have enough power to describe different phenomenon than the basic pine class had already described. The step-wise removal test of the single coefficients in fitted SUR models revealed some resemblances between datasets used. At first, the deciduous class for spruce model was observed to be redundant in all cases tested. That describes that the volume of spruce are behaving in a similar way in the deciduous ($G \geq 75$ %) and pine ($G \geq 75$ %) forests which is reasonable according to the canopy and understory structure of those forests. The significance of the true pine class of Sp_{max+95} strategy for data of Kuhmo could be related to the result of Table 14 in which the pine class as an intersect proved to be a redundant in two different single models due to the better describing power of the true pine class. The testing results of individual

coefficients have been evaluated to find some similarities between datasets and, thus, to help in combining some guidelines for classification structure to maximize the advantage of the pre-classifying idea studied. To sum up, the testing results in Table 13, it was proven again that increasing the amount of classes will show up in increasing amount of redundant categorical variables in SUR modeling.

Evaluating the other possibilities to improve the accuracies of the models presented in this study, the target of interest will move on over the technology and methodology used in the acquisition of both the field measurements and the ALS data. Carefully planned plot selection strategies have been proven to be important for yielding quality predictions for forest attributes (Maltamo et. al 2011). In many cases, the plot selecting strategies are difficult to compare because field inventories are implemented, naturally, once for the area of interest. The field inventories are implemented by cluster sampling in Kuhmo and for Janakkala-Loppi the method was unknown. However, the improvement of the possible better selection strategies can only be discussed but the results of the abovementioned study are worth considering when managing the selection systems. In this study, the more precise examination can be focused on the features of implemented ALS. As mentioned before, the intensity features have proven to have an important role in implementing the species-specific classification which should be more accurate for practical operations than the results presented in Table 10. In this study, the intensity values were more commonly used in the LDA classification than in the SUR modelling. The previous studies have proposed the normalization of intensity values for to improve tree species inventories. The studies have had an encouraging outcome to improve species recognition by normalization (e.g. Ørka et al. 2012; Korpela et al. 2010). The intensity data used were not normalized which should be one reason for such weak results in LDA. However, the other ALS features are also having an effect on the accuracy of the LDA and the SUR models, too. For example, the specifications of the flight and the ALS system operated in which, for example, density of measurements and footprint area are having an effect on the quality of ALS data. Korpela et al. (2010) have assessed two different sensor types, and they noticed that smaller footprint will probably enhance a signal-to-noise ratio. This study is based on the area-based method which attempts to predict the forest attributes by means of the sparse scanning density. Here, the nominal measurement densities of 0.52 and 0.62 measurements per m^{-2} were used for Kuhmo and Janakkala-Loppi, respectively. ALS acquisition of Janakkala-Loppi had higher flying altitude which tends to enlarge the footprint.

By the recent techniques of producing the species-species forest characteristics, it is seldom possible to produce solely ALS-based models which have satisfactory prediction accuracy for practical forest management (Uuttera et al. 2002). Hence, the supporting data have been used and studied in numerous studies. The most used way of getting more accuracy to the species-specific models is to include variables extracted from aerial images to the independent variables of the models. This method has been presented by Packalén & Maltamo (2006, 2007). However, the use of solely ALS data-based forest inventories is reasonable and attractive because extracting variables from the aerial images have been noticed to be challenging due to the varying quality of the photographs. The similarity of aerial imagines (e.g. light conditions and features of data acquisition) is demanded and the acquisition date with ALS data should be quite near with passive remote sensing (Packalén & Maltamo 2007). The potential of the use of aerial imagines has also been studied in other previous studies dealing with distinguishing species or producing the species-specific models (e.g. Vauhkonen et al. 2012).

Development during the last decades in the sector of ALS-based forest inventory techniques has been considerable rapid. It can be supposed that the development will continue more in the future due to the some potential methods which have been turned out to be too expensive nowadays. The one potential alternative to improve species-specific predictions would be the full-wave laser sensors which are not yet studied a lot in Scandinavian circumstances (Vauhkonen et al. 2014c). Moreover, the multispectral ALS sensors are also worth studying for the applications of species-specific forest inventories in the near future. The point densities have often been kept relative sparse due to the better economic efficiency (Maltamo et al. 2009a). The development of sensors will cause that the measurement densities are becoming denser, which is supposed to bring more exact information of the forest structure. The advantages of the denser ALS data have approved in previous studies, for example, the ITD methods of very high pulse density have produced considerable accurate results in tree species classification (c.f. Vauhkonen et al. 2008 and Vauhkonen et al. 2010). Also some interesting possibilities to improve species-specific models have been observed in the acquisition of ALS data. The bottleneck of species recognition, also in this study, has been the discrimination power between deciduous and coniferous, especially in mixed forests. The usage of leaf-off data in area-based forest inventories has been studied by Villikka et al. (2012) with encouraging results. The period of collection leaf-off data has recently been narrow because the leaf-off data have mainly been collected straight after when the snow coverage has melted in spring. The time period between melting

of snow and buds tend to vary slightly every year (Villikka et al. 2012). In the future, the opportunities will be changed since the climate change are modifying seasons in Finland that the coverage of snow will come later and maybe melt earlier. Villikka et al. (2012) proved that the leaf-off data can improve species-specific models and recognition of species solely from ALS data. However, they noticed that merging leaf-off and traditional leaf-on data increased bias values.

Principally, the main aim of this master's thesis was to present the possibilities to exploit the field-based pre-classifying method to improve the ALS-based species-specific volume predictions. The correct stratification of dominant tree species resulted in the improvements of 1.9–28.9 % and 9.4–36.4 % for the SUR groups fitted for Kuhmo and Janakkala-Loppi, respectively. In comparison with the results of earlier studies, Vauhkonen et al. (2010) have reached improvements of 9–47 and 33–50 percentage points by using balanced field data together with spectral imagines and ALS data. Also, for example Packalén et al. (2015), have managed to develop the method to improve plot-level volume predictions with edge-tree correction by 11–17 % according to the size of the plots. Moreover, Latifi et al. (2010) and Packalén et al. (2012) have proposed optimizing methods for NN-methods to get quite corresponding accuracy improvements. Although some of the previous studies have proposed numerous methods to reach at least corresponding improvements for species-specific volume predictions, the results of this study are able to give an appropriate alternative to increase accuracy of volume models when quality dominant tree information and ALS data are available. In the future, the ALS techniques and methods will develop, which will probably be seen in the better discrimination power between tree species by ALS data when the full advantage of the presented pre-classifying method could be reached. It is also probable that the individual tree detection methods will become more and more common in practical forestry. However, the area-based approaches are quite novel method in operational use of Finland, and it will be a mainstream method during near years. Hence, every applications to improve accuracies of those methods are surely welcome.

6 CONCLUSIONS

The results indicated that accuracies of the ALS-based species-specific plot-level volume predictions can be improved by including correctly predetermined dominant tree species information in both regression based and non-parametric imputations. The improvisations of RMSEs by using this method were at best 12.6–28.9 % and 20.9–36.9 % for SUR_1 of Kuhmo and SUR_2 of Janakkala-Loppi, respectively, depending on the species. The non-parametric k-MSN method implemented in Kuhmo data was slightly better by RMSE compared to the regression-based method. Consequently, the presented method gave encouraging results in both datasets which were seemingly differently distributed by tree species compositions. However, the comparison of two training datasets revealed that the species compositions have an important role in planning the dominant tree species classification structures. The predetermination of the dominant tree species by sparse ALS data proved to be a bottleneck to apply the method in solely ALS-based area-level species-specific volume predictions.

REFERENCES

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 96:37–46.

Crookston, N.L. & Finley, A.O. 2007. yaImpute: An R package for k-NN imputation. J Stat Softw 23:1–16.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179–188.

Fox, J. & Weisberg, S. 2011. An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage.

Gobakken, T. & Næsset, E. 2008. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner. Can J For Res 28:1095–1109.

Haara, A. & Korhonen, K. T. 2004. Kuvioittaisen arvioinnin luotettavuus. Metsätieteen aikakauskirja, 4(2004), 489–508.

Heinzel, J., Weinacker, H. & Koch, B. 2011. Prior-knowledge-based single-tree extraction. Int J Remote Sens 32:4999–5020.

Henningsen, A. & Hamann, J.D. 2007. Systemfit: A package for estimating systems of simultaneous equations in R. J Stat Softw 23:1–40.

Holmgren, J. & Persson, Å. 2004. Identifying species of individual trees using airborne laser scanner. Remote Sens Environ 90:415–423.

Holmgren, J., Persson, Å. & Söderman, U. 2008. Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images. International Journal of Remote Sensing, 29(5), 1537–1552.

Holopainen, M., Hyyppä, J. & Vastaranta, M. 2013. Laserkeilaus metsävarojen hallinnassa (In Finnish for "Laser scanning in the management of forest resources". University of Helsinki Department of Forest Sciences Publications 5:1–75.

Hyppä, J. & Inkinen, M. 1999. Detecting and estimating attributes for single trees using laser scanner. The Photogammetric Journal of Finland 16: 27–42.

Kalliovirta, J. & Tokola, T. 2005. Functions for estimating stem diameter and tree age using tree height, crown width and existing stand database information. Silva Fennica 39(2):227–248.

Korhonen, M. 2012. Puuston latvusrajan ennustaminen harvapulssisesta laserkeilausaineistosta mäntyvaltaisella alueella ja latvusrajan mittauksen tehostaminen (In Finnish for "Predicting crown base height of the tree stock using sparse airborne laser scanning data in a pine-dominated area and streamlining the reference measurements of the crown base height"). M.Sc. thesis, University of Eastern Finland

Korpela, I., Ørka, H.E., Maltamo, M., Tokola, T. & Hyyppä, J. 2010 Tree species classification using airborne LiDAR – effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. Silva Fennica 44(2):319–339.

Kotamaa, E. & Villikka, M. 2008. Report about processing the data for Metsälaser 2 project. University of Joensuu, 27 p.

Laasasenaho, J. 1982. Taper curve volume functions for pine, spruce and birch. Comm Inst For Fenn 108, 74 p.

Latifi, H., Nothdurft, A. & Koch, B. 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. Forestry 83:395–407.

Maanmittauslaitos. 2015. Pitkän aikavälin laserkeilaussuunnitelma kattaa vuoteen 2019 asti. Available: http://www.maanmittauslaitos.fi/kartat/laserkeilausaineistot/laserkeilausindeksit/laserkeilaussuunnitelma-2014-2019 . [Cited 23.11.2015] (In Finnish)

Maltamo, M., & Packalen, P. 2014. Species-specific management inventory in Finland. In: Maltamo, M., Næsset, E., Vauhkonen, J. (eds) Forestry applications of airborne laser scanning - concepts and case studies. Managing Forest Ecosystems 27. Springer, Dordrecht. pp 241–252.

Maltamo, M., Eerikäinen, K., Packalén, P. & Hyyppä, J. 2006. Estimation of stem volume using laser scanning based canopy height metrics. Forestry 79:217–229.

Maltamo, M., Mehtätalo, L., Vauhkonen, J. & Packalén, P. 2012. Predicting and calibrating tree attributes by means of airborne laser scanning and field measurements. Can J For Res 42:1896–1907.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., & Kangas, J. 2006. Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. Canadian Journal of Forest Research, 36(2), 426-436.

Maltamo, M., Bollandsås, O. M., Næsset, E., Gobakken, T. & Packalén, P. 2011. Different plot selection strategies for field training data in ALS-assisted forest inventory. Forestry, 84(1), 23-31.

Maltamo, M., Ørka, H. O., Bollandsås, O. M., Gobakken, T. & Næsset, E. 2015. Using preclassification to improve the accuracy of species-specific forest attribute estimates from airborne laser scanner data and aerial images. Scandinavian Journal of Forest Research, 30(4), 336–345.

Maltamo, M., Packalén, P., Peuhkurinen, J., Suvanto, A., Pesonen, A. & Hyyppä, J. 2007. Experiences and possibilities of ALS based forest inventory in Finland. In: Proceedings of the ISPRS workshop laser scanning 2007 and SilviLaser 2007, Espoo, Finland, 12–14 Sept 2007, IAPRS, vol XXXVI, Part 3/w52,2007,pp 270–279.

Maltamo, M., Packalén, P., Suvanto, A., Korhonen, K.T., Mehtätalo, L. & Hyvönen, P. 2009a. Combining ALS and NFI training data for forest management planning: a case study in Kuortane, Western Finland. Eur J Forest Res 128:305–317.

Maltamo, M., Pehkurinen, J., Malinen, J., Vauhkonen, J., Packalén, P. & Tokola, T. 2009b. Predicting tree attributes and quality characteristics of Scotch pine using airborne laser scanning data. Silva Fenn 43:507–521.

Maltamo, M., Bollandsås, O.M., Vauhkonen, J., Breidenbach, J., Gobakken, T. & Næsset, E. 2010. Comparing different methods for prediction of mean crown height in Norway spruce stands using airborne laser scanner data. Forestry 83:257–268

Moeur, M, Stage, A.R. (1995) Most similar neighbor: An improved sampling inference procedure for natural resource planning. For Sci 41:337–359

Næsset, E. 1997. Estimating timber volume of forest stands using airborne scanning data. Remote Sensing of Environment 61(2): 246–253.

Næsset, E. 2002. Predicting forest stand characteristics with airborne laser using a practical two-stage procedure and field data. Remote sensing of Environment 80: 88–99.

Ørka, H.O., Næsset, E. & Bollandsås, O.M. 2007. Utilizing airborne laser intensity for tree species classification. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2007, 36.Part 3: W52.

Ørka, H.O., Gobakken, T., Næsset, E., Ene, L. & Lien, V. 2012. Simultaneously acquired airborne laser scanning and multispectral imaginery for individual tree species identification. Can J Remote Sens 38:125–138.

Ørka, H.O., Dalponte, M., Gobakken, T., Næsset, E. & Ene, L.T. 2013. Characterizing forest species composition using multiple remote sensing data sources and inventory approaches. Scand J For Res 28:677–688.

Packalén, P. & Maltamo, M. 2006. Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. For Sci 52:611–622.

Packalén, P. & Maltamo, M. 2007. The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sens Environ 109:328–341.

Packalén, P. & Maltamo, M. 2008. Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. Can J For Res 38:1750–1760.

Packalén, P., Temesgen, H. & Maltamo, M. 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. Can J Remote Sens 38:557–569.

Peltola, A. 2014. Metsätilastollinen vuosikirja 2014. SVT Maa-, metsä-ja kalatalous.

Peuhkurinen, J. Maltamo, M., Malinen, J., Pitkänen, J. & Packalén, P. 2007. Preharvest measurement of marked stands using airborne laser scanning. Forest Science 53(6):653–661.

Pippuri, I., Maltamo, M., Packalen, P. & Mäkitalo, J. 2013. Predicting species-specific basal areas in urban forests using airborne laser scanning and existing stand register data. Eur J For Res 132:999–1012.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Reitberger, J., Schnörr, C., Krzystek, P. & Stilla, U. 2009. 3D segmentation of single trees exploiting full waveform LiDAR data. ISPRS J Photogramm Remote Sens 64:561–574.

Rempel, R.C. & Parker, A.K. An information note on an airborne laser terrain profiler for micro-relief studies. In: Proceedings of the 3rd symposium on remote sensing of environment. University of Michigan Institute of Science and Technology, pp 321–337.

Roncat, A., Morsdorf, F., Briese, C., Wagner, W. & Pfeifer, N. 2014 Laser pulse interaction with forest canopy: Geometric and radiometric issues. In: Maltamo, M., Næsset, E. & Vauhkonen, J. (eds) Forestry applications of airborne laser scanning - concepts and case studies. Managing Forest Ecosystems 27. Springer, Dordrecht, pp 19–41.

Siipilehto, J. 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. Silva Fenn 33:281–301.

Törmä, M. 2000. Estimation of tree species proportions of forest stands using laser scanning. Int Arch Photogramm Remote Sens XXXIII:1524–1531.

Uuttera, J., Hiltunen, J., Rissanen, P., Anttila, P. & Hyvönen, P. 2002. Uudet kuvioittaisen arvioinnin menetelmät-arvio soveltuvuudesta yksityismaiden metsäsuunnitteluun. Metsätieteen aikakauskirja 3/2002:523–531.

Vauhkonen, J. 2010. Estimating crown base height for Scots pine by means of the 3D geometry of airborne laser scanning data. Int J Remote Sens 31:1213–1226

Vauhkonen, J., Tokola, T., Maltamo, M. & Packalén, P. 2008. Effects of pulse density on predicting characteristics of individual trees of Scandinavian commercial species using alpha shape metrics based on airborne laser scanning data. Can J Remote Sens 34:441–459.

Vauhkonen, J., Tokola, T., Packalén, P. & Maltamo, M. 2009. Identification of Scandinavian commercial species of individual trees from airborne laser scanning data using alpha shape metrics. For Sci 55:37–47.

Vauhkonen, J., Korpela, I., Maltamo, M. & Tokola, T. 2010. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. Remote Sens Environ 114:1263–1276.

Vauhkonen, J., Seppänen, A., Packalén, P. & Tokola, T. 2012. Improving species-specific plot volume estimates based on airborne laser scanning and image data using alpha shape metrics and balanced field data. Remote Sens Environ 124:534–541.

Vauhkonen, J., Maltamo, M., McRoberts, R.E. & Næsset, E. 2014a. Introduction to Forestry Applications of Airborne Laser Scanning. In: Maltamo, M., Næsset, E., Vauhkonen, J. (eds) Forestry applications of airborne laser scanning - concepts and case studies. Managing Forest Ecosystems 27. Springer, Dordrecht, pp 1–16.

Vauhkonen, J., Packalen, P., Malinen, J., Pitkänen, J. & Maltamo, M. 2014b. Airborne laser scanning-based decision support for wood procurement planning. Scand J For Res 29(sup1):132–143.

Vauhkonen, J., Ørka, H.O, Holmgren, J., Dalponte, M., Heinzel, J. & Koch, B. 2014c. Tree species recognition based on airborne laser scanning and complementary data sources. In: Maltamo, M., Næsset, E. & Vauhkonen, J. (eds) Forestry applications of airborne laser scanning - concepts and case studies. Managing Forest Ecosystems 27. Springer, Dordrecht, pp 135–156.

Venables, W.N. & Ripley, B.D. 2002. Modern applied statistics with S. Springer

Wallenius, T., Laamanen, R., Peuhkurinen, J., Mehtätalo, L. & Kangas, A. 2012. Analysing the agreement between an airborne laser scanning based forest inventory and a control inventory – A case study in the state owned forests in Finland. Silva Fenn, 46, 111-129.

Wallerman, J. & Holmgren, J. 2007. Estimating field-plot data of forest stands using airborne laser scanning and SPOT HRG data. Remote Sens Environ 110:501–508.

Wehr, A. & Lohr, U. 1999. Airborne laser scanning – an introduction and overview. Photogrammetry & Remote Sensing 54:68–82.

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 Sensing. 2010; 2(6):1481–1495.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American statistical Association, 57(298), 348–368.