AN ABSTRACT OF THE DISSERTATION OF

Bryce Frank for the degree of Doctor of Philosophy in Sustainable Forest Management presented on March 17, 2020.

Title: Aerial Laser Scanning for Forest Inventories: Estimation and Uncertainty at Multiple Scales

Abstract approved:

_____

Temesgen Hailemariam

As remote sensing data continues to proliferate, development and assessment of methods that generate predictions of forest attributes is needed to inform operational use. For several decades, lidar data collected from aerial platforms has informed assessments of forest resources, but many opportunities remain for the technology. We examine and develop methodology for three different problems facing forest management inventories.

In our first manuscript, we assess three different classification methods for identifying stands for commercial thinning operations in southwestern Oregon, USA using a set of fixed radius plots and a coincident aerial lidar acquisition. We also assess the impact of sample size using a simulation procedure. We found that random forests and a newly developed gradient boosting algorithm exhibited moderate performance for commercial thinning classification. We also observed that performance of all classification methods stabilized at sample sizes between 200 and

300, which may be an attainable sample size in some operational forest inventories. This study was motivated by the lack of literature that examined prediction of categorical variables such as management classes, rather than prediction of continuous forest structural variables, as a potential method for assisting forest management planning decisions. We anticipate a number of extensions to other management-oriented classifiers, such as pre-commercial thinning and various other silvicultural objectives, as future studies.

In the second manuscript we compared a segmentation-based method against an area-based approach method for generating stand-level predictions of forest attributes. Particularly, we leverage small area estimation methods to produce model-based mean squared errors for stand-level predictions. The analysis suggests that the segmentation-based method tends to produce lower mean squared errors in stands where sample sizes increased due to tree segmentation. Furthermore, models based on detected segments were less prone to extrapolation than models produced using the area-based method. However, area-based models generally produced lower mean squared errors for stands that did not contain sampled population units. This study was motivated by a lack of investigation into the use of tree-segmentation methods for producing stand-level predictions of forest attributes, which is a typical objective of many forest management inventories. We believe that this manuscript lays the foundation for continued assessment of alternative tree-segmentation methods with rigorous assessments of prediction error.

The final manuscript employed the multivariate Fay-Herriot model, a recent theoretical development in the small area estimation literature, for producing stand-

level predictions of forest attributes. We compared bivariate Fay-Herriot models to their univariate counterparts for five forest attributes and observed that, for at least one bivariate pairing, stand-level mean squared errors were reduced for both sampled and unsampled stands. Additionally, we identify the uniformity and strength of correlation among stand-level direct-estimators as an important indicator of the success of a bivariate model over a univariate model. We plan to conduct a future study that leverages the multivariate Fay-Herriot model for use in a time-series analysis to unify remote sensing and field data collected at disparate times.

Aerial Laser Scanning for Forest Inventories: Estimation and Uncertainty at Multiple Scales

by
Bryce Frank

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented: March 17, 2020
Commencement: June 2020

Doctor of Philosophy dissertation of Bryce Frank presented on March 17, 2020

APPROVED:

Major Professor, representing Sustainable Forest Management

Head of the Department of Forest Engineering, Resources and Management

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Bryce Frank, Author

ACKNOWLEDGEMENTS

Ever since I began my doctoral program at Oregon State University, I have thought about when I would finally be able to write this section of my dissertation. There are many people I must thank for getting me here.

For her endless support in several categories: emotional, mental, dietary, etc., I express my deep appreciation and love for my partner and best friend, Sara Finkle. It would have been an impossible endeavor without her. I must also acknowledge our cats, Kiki (may she rest in peace) and Chloe for accompanying me on my sleepless nights, and reminding me to be confident, as they always are.

For his infinite patience, technical assistance, and guidance, I must acknowledge Dr. Francisco Mauro. You have lain the foundation of my expertise in forest biometrics and have been there every step of the way. I am looking forward to continuing our research careers together in the coming years.

For his wisdom, advice, teaching and support, I deeply thank Dr. Temesgen Hailemariam. It seems not so long ago that I sent you an email on a whim asking to be your student. How quickly the time goes. I am blessed to have been given the opportunity to explore the subject of forest biometrics as freely as I have during my program.

I would like to give thanks to my academic committee members, Drs. Alix Gitelman, Rebecca Hutchinson, Leah Rathbun and Bogdan Strimbu for their guidance and advice throughout the course of my program.

I was fortunate to have a number of excellent lab mates during my journey, who were always ready to listen to my caffeine-charged ramblings: Al Pancoast, Ty Nietupski, Karin Kralicek, Joonghoon Shin, Jake Putney, Krishna Poudel, and Quinton Big Knife.

One must stand on the shoulders of giants, they say. I have several I would like to mention. Some I have met and some I have not. All the same, I express deep gratitude

to Drs. Isabel Molina, Jim Flewelling, Kim Iles, and Johannes Breidenbach for inspiring my work in many different ways.

Finally, although distant in space, we are always close at heart. Thank you to my family: Latonya, Nichelle, Brian, Diane, Jeremy and Tanya for always being at my side.

CONTRIBUTION OF AUTHORS

Drs. Temesgen Hailemariam, Francisco Mauro, Kevin Ford, and Vicente Monleon provided technical advice and thorough revisions. Dr. Martin Ritchie provided the data for the third manuscript.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX TABLES

1 Introduction

1.1 Forest Inventories and Design-Based Inference

If the reader were to conduct a survey of historic articles published in the field of forest biometrics, they might find questions that are somewhat trifling: how best to randomly select trees from a forest (Grosenbaugh 1958)? how should we determine the mean distance between two trees (Cottam and Curtis 1949)? is it fair to consider the cross-section of a stem a circle (Matern 1956)? The nature of the questions proposed by forest biometricians perhaps betrays their leading trait: an absolute refusal to consider any detail too small. Such a trait is necessary in a field where the scales at which information is required is often orders of magnitude larger than the available set of observations. Thus, the driving force of large-scale forest inventory research is primarily that of error control and mitigation, which is necessitated by a stark imbalance of available observations and the large scales at which predictions and estimates of forest attributes are required.

For many decades, forest inventories relied in large part on a particular form of statistical methodology referred to as "design-based inference" to provide estimates of forest attributes at large scales. Neyman (1934) is often credited with the establishment of design-based inference, who emphasized that the randomization process used to select a sample, i.e. a sampling design, is a superior form of inference to that of "purposive", or non-random, sample selection. As Gregoire (1998) elaborates, design-based inference does not rely on any assumptions about the population to produce point estimates and error estimates, and its validity is derived from the randomization process introduced in the mechanism used to select a sample from the population. The principles of design-based inference can be seen in all facets of forest inventory, from academic publications

that elaborate on and expand the methodology (e.g. Mandallaz 1991; Grafström et al. 2017), to publications for practitioners that relay its requirements and peculiarities (Bell and Dilworth 1990; Iles 2003). While design-based inference undoubtedly comes with a number of benefits and guarantees, it is incontrovertible that a number of alternative methods have arisen in recent decades that present attractive opportunities for the science of forest inventory.

1.2 Remote Sensing Assisted Forest Inventory

The diversification of available methods in forest inventory is coincident with the proliferation of remote sensing auxiliary information. The past several decades have seen the development of a new sub-field in forest inventory that explicitly considers the use of remote sensing data for predicting forest attributes at multiple scales. We refer to this sub-field as remote sensing assisted forest inventory (RS-FI). RS-FI is now accepted as an operational practice, and examples of its use abound (Maltamo and Packalen 2014; Næsset 2014; White et al. 2017). In turn, the way in which estimates for parameters of interest are obtained has shifted to those methods that explicitly consider the auxiliary information, such as remote sensing data. Datasets such as lidar, multispectral imagery, photogrammetric point clouds and radar have been used to provide auxiliary information to this end (Yu et al. 2015; White et al. 2016). Forest inventory as a practice itself is also evolving. Once a field primarily concerned with the quantification of the timber asset, forest inventory is faced with new challenges in providing information about a wider range of phenomena such as ecological structure, silvicultural variables, wildlife habitat, and fuel structure (Temesgen et al. 2007). The interplay between these two movements in

forest inventory systems has resulted in new conceptualizations, methods and approaches in the practice.

While many remote sensing datasets are available for use in RS-FI, lidar data acquired from aircraft has continued to demonstrate its dominance as an auxiliary data source for use in RS-FI, and is the primary auxiliary data source used in these studies. Aerial lidar systems record a large number of positions in 3-dimensional space by measuring the position of the sensor itself, the angle of the emission and the time required for a lidar pulse to emit and return to the sensor as the aircraft flies over the forest. The resultant dataset is a 3-dimensional point cloud that contains vertical and horizontal structural information about the forested area over which the dataset was acquired. Aerial lidar data has seen application for forest inventory problems dating back to the 1970s (see Nelson 2013), but is commonly regarded to come into operational practice in the early 2000s, particularly due to important developments in GPS technology and research efforts in Norway (Næsset 2014 pp. 222-223).

RS-FI that leverages lidar as an auxiliary information source is typically deconstructed into two frameworks including the individual tree crown (ITC) and area-based approach (ABA) methods. ITC methods seek to explicitly segment individual tree objects from the available auxiliary data, and generates predictions for the unobserved tree objects (Hyyppa et al. 2001). The ABA method tessellates the study area into a set of regular grid cells and develops predictions based on a model made from, typically, fixed-radius field plots with georeferenced locations that are paired with summaries of lidar data (Næsset 2002). While this delineation is somewhat useful for introductory purposes, and by-and-large represent the most common operational and research applications, it

should be noted that many examples that violate, combine, or otherwise modify these two frameworks are possible (e.g. Lindberg et al. 2010; Shin and Temesgen 2018), including the second and third manuscripts in this dissertation.

Perhaps the most fascinating feature of the field of RS-FI is the diverse set of perspectives held by those that develop the science, and the unique characteristics that research teams and practitioners across the world bring to the table. Clearly the demands on RS-FI vary widely by different forest structures, regulatory requirements, and traditions in the larger body of forest inventory, and solutions to these problems are therefore just as diverse. Take for example the case of Norway and Finland: two relatively close countries with relatively similar boreal forests (Brainerd and Rolstad 2002). However, as noted by Næsset (2014 p. 223), smaller scale Finnish forest inventories rely on more explicit tree species information, which led to the development and application of methods that provide these types of predictions, such as k-Nearest Neighbor methods, that acted as an alternative to the regression models typically used in the Norwegian case (Maltamo and Packalen 2014). Another example is that of regulatory requirements in Spain that require errors of predictions to be reported for forest areas that are frequently under 1,000 hectares (Comunidad de Madrid 2012). Such a requirement led to the application and development of small area estimation (SAE) for use in forest inventories, a body of methodology that is well-suited for providing measures of error at small scales (e.g. Mauro et al. 2016). Finally, in the Pacific Northwest of the United States there is a rich tradition of stand-level forest management and planning (Wilson and Puettmann 2007), that require predictions and measures of error at the scale of forest

management stands (Kangas et al. 2015 pp. 13-14), tightly related to the SAE methods relayed previously, and is a major of theme of this dissertation.

1.3 Breiman's Dichotomy

The variety of methods and approaches available for prediction within RS-FI can be cleaved along a major axis, that of non-parametric and parametric methods, and is reflective of a larger rift in predictive modeling in general. Breiman (2001) referred to this rift as "the two cultures" of modeling. On the one hand there are "data modelers" who posit hypothetical models for the data generating process and assess the appropriateness of these hypothetical models using observed data. Such an approach typically relies on parametric models, and forms the basis of "model-based" methods (see Gregoire 1998). On the other hand there are "algorithmic modelers", who regard the data generating process as unknown, and develop predictive algorithms that estimate the data generating process through "learning" procedures (Hastie et al. 2009 Ch. 2).

While it is beyond the scope of this dissertation to argue the superiority of one approach over the other, it is clear that much of current RS-FI falls along these delineations due to the numerous advantages that one my gain by adopting either the data modeling approach or the algorithmic modeling approach. Algorithmic modeling has shown a much greater flexibility in the types of response variables and predictors that can be accommodated than data modeling, especially in the realm of classification problems (e.g. Valbuena et al. 2016) and prediction of complex multivariate responses such as diameter distributions (e.g. Mauro et al. 2019; Räty et al. 2019). However, the uncertainty of predictions generated using algorithmic modeling approaches are typically assessed using cross-validation procedures, which do not provide clear in-roads for inference

about populations without further theoretical developments (see Mentch and Hooker 2017 as one such development). In contrast, while data modeling approaches are more rigid with respect to the configurations of predictors, response variables, etc., and rely on a number of assumptions, they have a rich history in model-based inference and small area estimation, which provide a more complete understanding of the sources of error for predictions of population units as well as subpopulation parameters which are analogous to the standard errors commonly produced using traditional, design-based, forest stand exams discussed previously.

1.4 Research Problems

Clearly, RS-FI has been adapted and modified in many different geographic and regulatory contexts to assist in forest management decision-making for the past several decades. However, many opportunities remain for improvements and contributions to this field of research. We examine three such opportunities in this dissertation.

At the time of authorship of the first manuscript, a large part of the focus of the RS-FI research community, particularly those using aerial lidar data, had been on the prediction of continuous forest inventory variables such as standing timber volume, basal area, and quadratic mean diameter. However, forest management planning, particularly in the Pacific Northwest, can rely on categorical variables, such as silvicultural development classes (Pippuri et al. 2016; Valbuena et al. 2016). For forestland managers such as the Bureau of Land Management (BLM) in Oregon, which manages tens of thousands of hectares, selecting stands that meet the structural requirements of commercial thinning is a large undertaking. Traditional methods for commercial thinning stand selection within the BLM for the study area we considered relied on a combination of field visits, records

of past management activity, examination of aerial photographs, and field data (Ford, personal communication, January 7, 2020). The objective of the study was to leverage a wall-to-wall aerial lidar dataset that covered 76,183 ha of Douglas-fir dominated forests and a set of 653 field plots in southwestern Oregon to predict thinning eligibility using an area-based approach. We assessed three common classification methods, including the non-parametric random forests (RF), the relatively newer method XGBoost (XGB), as well as the parametric logistic regression (LOG). XGB and RF produced reliable predictions of thinning eligibility at the scale of individual pixels, followed by LOG.

While the prediction of thinning eligibility using aerial lidar data had been investigated in the Nordic context previously (Pippuri et al. 2012; Korhonen et al. 2013), such an assessment had not been made for Douglas-fir stands in the Pacific Northwest. Furthermore, the two previously cited studies relied on professional expertise to categorize field samples into thin and no-thin categories, whereas our approach relied on commonly measured forest inventory variables (basal area, stem density, and volume), which provides a flexible alternative for already existing forest inventories that contain these variables. In addition to the assessment using all 653 field plots, we also assessed the performance of the classifiers at reduced sample sizes and noted stable performance at sample sizes of 200 to 300 for all three classification methods, which is a relevant result for forest inventories that have smaller sample sizes available.

Chapters 3 and 4 examine the use of small area estimation (SAE) for producing stand level predictions and associated measures of error. While the use of SAE in forest inventory is not explicitly new (section 1.2), many forest inventory problems that can benefit from an SAE framing have remained unexplored, and potentially useful

theoretical developments in the field have yet to be investigated for use in RS-FI. It is of the opinion of the first author that SAE methodology is one of the most complete and extensive theoretical frameworks available for RS-FI and can be used to formally assess the appropriateness of various methodologies proposed in the modern RS-FI literature. More specifically, model-based methods in SAE, specifically unit-level and area-level (i.e. Fay-Herriot) models provide formalized methods for producing predictions of stand-level forest attributes, such as stand-level mean basal area or stand-level mean timber volume, that are based on the general linear model. Additionally, SAE models based on the general linear model provide formalized methods for estimating the mean squared error of area-level predictions, and provide assessments of uncertainty analogous to that of design-based standard errors (section 1.1).

Chapter 3 compares the performance of unit-level models constructed using a semi-individual tree crown approach and unit-level models constructed using the area-based approach at the scale of individual forest stand predictions. This study was motivated by a lack of available literature that explicitly considered the task of stand-level prediction and measures of error when using tree segmentation approaches. While a plethora of literature is available that develops methods for the segmentation of trees from point clouds and other remote sensing data, very few studies explicitly consider their use for stand-level forest inventories. For tree segmentation approaches to receive serious consideration for use in these contexts, a rigorous assessment of their error properties at the scale of forest management stands was needed. The objectives of Chapter 3 were to formalize the use of tree segmentation methods within the unit-level model, compare point predictions and measures of error using models constructed from

detected tree segments and models constructed in the manner of the ABA, and to provide a rigorous assessment of the error properties of both model types for four different forest inventory variables. We elected to adopt the semi-individual tree crown approach (s-ITC) as the population framework for the tree segmentation models.

We conducted the assessment for a 2,580 ha study area in western Oregon that contained a sample of 78 16 m fixed-radius field plots. The results indicated that the s-ITC models exerted higher random effect variance than models constructed using the ABA, which implies that larger sample sizes are required to achieve the same level of error. In some cases, the segmentation method provided larger stand-specific sample sizes (i.e. where many trees were detected, such as in younger, more-dense stands) and lower errors were therefore achieved relative to the ABA. S-ITC models also demonstrated lower risk of extrapolation due to the larger sample sizes available to estimate model parameters. However, the uniformly higher random effect variances resulted in predictions for unsampled stands (i.e. synthetic predictions) that had higher mean squared errors than ABA models. These results suggest that the s-ITC models are better suited to estimation of stand-level parameters in younger stands where samples are already present as a way to augment the prediction of stand level parameters in an already existing forest inventories.

Chapter 4 assesses the use of a different SAE model referred to as the Fay-Herriot model. The Fay-Herriot model can be used to predict forest attributes at the scale of forest stands, but does so by using data collected at the scale of small areas, rather than at the scale of population units (hence the alternative term “area-level model”). Typically, this data consists of pairs of design-based estimates of stand level attributes and lidar

predictors aggregated to the stand level. This type of model is attractive for forest inventories where precise plot locations are unknown, as it does not require the assignment of covariates to population units. Rather, only the membership of field plots to the small area of interest is needed to estimate model parameters and produce predictions. While the Fay-Herriot model has seen some application in forest management inventories (Ver Planck et al. 2018; Breidenbach et al. 2018), these applications universally rely on univariate Fay-Herriot (UVFH) models. Recently, however, the method has been extended to the multivariate setting, and have shown promising gains in efficiency in other fields, such as in demographic surveys (Ubaidillah et al. 2019).

Chapter 4 sought to expand the use of bivariate Fay-Herriot (BVFH) models and address considerations particular to forest inventory, such as limited stand-specific sample sizes, and errors related to synthetic and large-area predictions, that are not typically considered in other research domains. We constructed UVFH models for five forest attributes, including mean stand volume, basal area, stem density, height and diameter. For each pair of attributes BVFH models were constructed and compared against their UVFH counterparts. We found that for each attribute, a bivariate pairing exists such that the BVFH model produced lower mean squared errors than the UVFH counterpart. Our analysis was consistent with previous literature that demonstrated BVFH models tend to provide lower mean squared errors when the attributes used in the model are strongly (positively or negatively) correlated with each other.

2 Manuscript I

Analysis of Classification Methods for Identifying Stands for Commercial Thinning using LiDAR

Bryce Frank, Francisco Mauro, Hailemariam Temesgen, and Kevin R. Ford

Canadian Journal of Remote Sensing

Canadian Aeronautics Institute

350 Terry Fox Drive, Suite 104

Kanata, ON K2K 2W5, Canada

Volume 45, Issue 5

Abstract

Commercial thinning (CT) is an important tool that meets a diverse set of forest management objectives, including the generation of intermediate revenue, promotion of regeneration tree growth, and the modification of vertical and horizontal fuel structure for wildfire mitigation.

Using a set of 653 fixed radius plots and a coincident LiDAR acquisition, we compared three different classification methods to predict CT eligibility for Douglas-fir (*Pseudotsuga menziesii*) stands in southwestern Oregon. We assessed logistic regression (LOG), random forests (RF), XGBoost (XGB) to classify areas eligible for CT operations based on three structural attributes, volume (VOL), basal area (BA) and Curtis' Relative Density index (CRD). We also assessed their predictive performance and reliability via cross-validation at different sample sizes.

We used the area under the receiver operating characteristic curve (AUC) as our primary performance measure. Estimated AUCs were 0.86, 0.77 and 0.68 for XGB, RF and LOG, respectively. We observed that classifier performance stabilized between sample sizes of 200 and 300 plots, which suggests that the development of a CT eligibility classifier is appropriate for operational applications of the method with similar sample sizes and large area attributes.

2.1 Introduction

The scheduling of commercial thinning (CT) operations in forest management planning is a dynamic process that is dictated by the structure of the forest, the operational feasibility and the objectives of forest land owners. CT harvesting operations are generally more complex and less cost effective than alternative harvesting methods due to the complexities introduced in felling and transportation of harvested material (Mizuniwa et al. 2016). In many forested areas, CT is a valuable option for forest land managers, and this is especially the case in the Pacific Northwest of the United States. CT generates an intermediate source of revenue, promotes the growth of regeneration trees (Bailey and Tappeiner 1998), modifies and promotes certain types of wildlife habitat (Suzuki and Hayes 2003), and alters the horizontal and vertical distribution of fuels for wildfire management objectives (Graham et al. 1999).

Frequently, forestland managers schedule CT projects after a particular stand has met a set of specified criteria, which are based primarily on the structure of the forest and the project's operational feasibility. The former can be considered a function of structural variables of the forest itself, such as its relative density, age, and volume per unit area. The latter can be considered a function of operational considerations, such as market conditions, geographic position of the stand relative to wood processing facilities and other non-structural variables. We focus our study on the former, predicting whether a stand that is structurally eligible for commercial thinning is a pre-requisite to determine whether it is an operationally feasible project.

Typically, forest stands are determined to be CT eligible by using field surveys. Implementing these field surveys can be a costly endeavour for large forested landscapes.

Given the increased operational complexity and reduced revenues for CT operations, optimization involved with their planning is needed. Remote sensing, specifically light detection and ranging data (LiDAR) collected via airborne laser scanning, provides wall-to-wall fine-scaled auxiliary information that can be used to predict the vertical and horizontal structure of forested landscapes via three-dimensional point clouds, including commercial thinning eligibility.

Remote sensing assisted forest inventories frequently employ the area-based approach (ABA) to develop a relationship between field plot measurements and LiDAR derived covariates (see e.g. Næsset 2002). A regular grid, typically with cells of equal size to that of the field plots, is cast over the study area. Plot-level observations are considered to be a sample from the population of grid cells. A relationship between the attribute of interest and the LiDAR covariates measured on the plots is estimated via some regression or classification method. In the regression case ABA has been shown to provide predictions at fine scales of many continuous variables used in forest operations planning including total volume, above-ground biomass, mean height, and other variables (e.g. Babcock et al. 2015; Shin et al. 2016). Predictions using the ABA can provide estimates for stand-level forest management that are comparable to those using only field-based ground surveys (Maltamo and Packalen 2014).

However, some forest management decisions, such as CT planning, rely on qualitative classes of forest structure such as thinning eligibility and silvicultural development classes. The issue of producing a CT eligibility model using remotely sensed data can be viewed as a supervised classification problem where a binary response for thinning eligibility, which is a function of forest attributes such as relative density,

basal area and volume, is produced from field measured plots with coincident LiDAR derived metrics as covariates. Supervised classification, along with LiDAR data, has been used to predict silvicultural development classes (Valbuena et al. 2016), land use and land cover classification (Pippuri et al. 2016), need for seedling stand tending (Korhonen et al. 2013), species distribution modeling (Farrell et al. 2013) and prediction of individual tree species (Dalponte et al. 2013). Thinning eligibility itself has been assessed previously in Pippuri et al. (2012), which focused on the development of novel LiDAR derived predictors that discriminate between stands that require first thinning and those that do not using linear discriminant analysis (LDA).

The previous studies have presented many situations in which supervised classification can be used for forest inventory applications with LiDAR data. Field sampling campaigns in forest inventory make up a large portion of the operational cost; therefore, it is a high priority for forest land managers to know the impact of sample size on the performance of classification methods. The focus of our study is two-fold. First, we assess the performance of three classifiers, random forests (RF), logistic regression (LOG) and a recently developed classification method, XGBoost (XGB) for predicting eligibility for commercial thinning in Douglas-fir stands in southwestern Oregon. Following this, we examine the performance of these classifiers in reduced sample size settings. In so doing we develop recommendations for management specific classifiers with regard to type of classification method and sample size, and highlight issues in classifier tuning and performance assessment for different sample sizes.

2.2 Materials and Methods

2.2.1 Study Area

A 1.5 million hectare LiDAR acquisition was coordinated by the Oregon Department of Geology and Mineral Industries (DOGAMI) during 2008 and 2009 in Coos-Bay and Curry counties in Southwestern Oregon (SWO). Within this acquisition are 98,104 ha of forested lands managed by the Bureau of Land Management (BLM). Figure 2.1 displays a contextual map of the study area. The dominant tree species is Douglas-fir (*Pseudotsuga menziesii*), but other species are present, including western hemlock (*Tsuga heterophylla*), Sitka spruce (*Picea sitchensis*), western redcedar (*Thuja plicata*), red alder (*Alnus rubra*), and bigleaf maple (*Acer macrophyllum*).

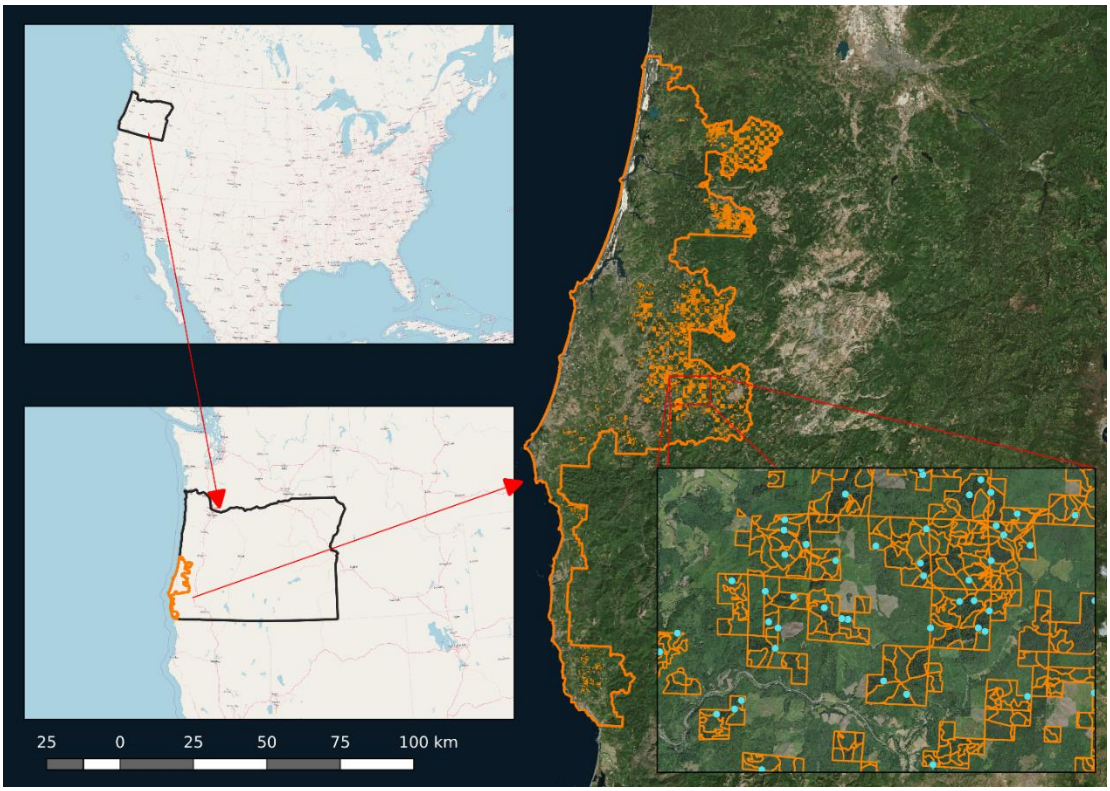

Figure 2.1: Map of Southwestern Oregon BLM administered commercial Douglas-fir stands indicated in orange and field plot positions indicated in blue.

In addition to the LiDAR acquisition, the BLM maintains the Forest Operations Inventory (FOI) spatial database. This database provides descriptions of forest stands for all land administered by the BLM in western Oregon. The FOI database contains information about forest stands segmented into vertical vegetation layers of description. For each stand, each vegetation layer is described by its age class and tree species composition. Using this auxiliary information, we constrained our analysis to only those stands that were described as pure Douglas-fir stands (i.e. Douglas-fir was the only species recorded as present in these stands across all layers within that stand) because the thinning decision boundaries developed for this application only apply to commercial Douglas-fir stands. After this selection, 6,238 stands were considered for the analysis with a total study area size of 76,183 ha and an average stand size of 12.2 ha.

2.2.2 Field Plots

Field plots were selected using stratified random sampling using a procedure similar to that described in Hawbaker et al. (2009) as part of a larger field campaign that was not in direct support of this study. Pixels were first allocated into 10 different bins based on 80th percentile LiDAR height such that an equivalent number of pixels were present in each height bin. Then, within each height bin, pixels were allocated into three bins based on standard deviation of LiDAR height such that an equivalent number of pixels were present in each standard deviation bin. Within each of the 30 bins, the locations of 30 pixels were selected using simple random sampling, and served as the positions for the centers of fixed radius plots. The field plots were then located for measurement, resulting in a sample size of 900 field plots. Under this sampling design all pixels (j), in the study area are assigned a sampling inclusion probability of

$$\pi_{i,j} = \frac{900}{N} \; \forall \; i \; and \; j, \tag{2.1}$$

regardless of the bin (j). Inclusion probabilities for pairs of units indexed by i and j, and k and l respectively, (i and k indicate bin membership, j and l indicate the unit within a bin and N indicates the total number of population units) are:

$$\pi_{(i,j),(k,l)} = \begin{cases} \left(\frac{900}{N}\right)^2 \; if \; i \neq k \\ \frac{900}{N}\frac{29}{(\frac{N}{30}-1)} \; if \; i = k \end{cases} \tag{2.2}$$

Given the population size of the study area, (i.e. $N \cong 1.8 \cdot 10^6$) the inclusion probabilities for pairs are very close to those obtained using simple random sampling (SRS), and their differences with SRS are negligible.

After constraining the study area using the FOI database (section 2.1) a total of 653 field plots exist within the study area. The field campaign was conducted between May 25, 2010 and May 10, 2011 (see Shin et al. 2016 for more details). Field crews located plot centers using a coarse acquisition code GPS. Phase observations were recorded at plot center for at least 10 minutes and those observations were post processed so that location errors are expected to be negligible. A nested plot configuration was used such that a 0.052 ha fixed radius plot encompassed a smaller 0.0081 ha fixed radius plot. On the larger plot, all trees greater than 13.97 cm diameter at breast height (DBH) were measured for tree position relative to plot center, total height (THT), species (SP), and DBH. On the smaller fixed radius plot, live trees less than 13.97 cm DBH were measured for DBH, THT, and SP. For all trees, standing volumes were estimated using the National Volume Estimator library (Wang 2019). Accuracies for estimated tree volumes

are reported to be between 0.5% and 1% for Douglas-fir (Flewelling and Raynes 1993), and are treated as known quantities for the purposes of this study.

2.2.3 LiDAR Acquisition and Processing

A discrete return LiDAR dataset was acquired between 2008 and 2009 during leaf on conditions using a Leica LiDAR ALS50 Phase II laser attached to a fixed wing aircraft. The average return density was 8.1 points / m^2, the average flying altitude was 900 meters above sea level and the field of view was 28 degrees ($\pm$ 14 degrees from nadir). See Shin et al. (2016) for a more detailed description of this acquisition. Within each grid cell a vector of predictors, $\boldsymbol{x}_i$, was computed using a grid cell size of 22.86 m by 22.86 m (0.052 ha) with FUSION software (McGaughey 2016). In total, 28 predictors were derived using FUSION. See Table A 1 for a list and description of considered LiDAR predictors. Field plots were segmented from the total LiDAR acquisition using the georeferenced positions and the same set of metrics was computed for these plot-level point clouds. We consider grid cells and the segmented field plot and point clouds to represent the same population for this study.

2.2.4 Thinning Decision Boundaries

For a given classification procedure, the training data, which are observations of continuous variables, must first be assigned into the classes of interest. In this case, each individual field plot must be assigned to either NO_THIN or THIN categories. This was achieved by applying management decision boundaries on attributes used in BLM forest practices for commercial Douglas-fir stands at the plot-level. These attributes included basal area (BA), merchantable volume (VOL), and Curtis Relative Density index (CRD, equation (1)). CRD is a relative density measure developed particularly for Douglas-fir

stands (Curtis 1982):

$$CRD_i = \frac{BA_i\ (\frac{m^2}{ha})}{\sqrt{QMD_i\ (cm)}} \tag{2.3}$$

where BA_i refers to the basal area of the ith plot, and QMD_i refers to the quadratic mean diameter of the ith plot. These boundary variables are values that assign a given plot into thinning eligible or non-eligible categories. We considered a given plot to be of class THIN if its CRD exceeded 127, its merchantable volume was between 245 $m^3/\ ha$ and 700 $m^3/\ ha$ and its basal area exceeded 32 $m^2/\ ha$. Otherwise, the plot was considered a NO_THIN.

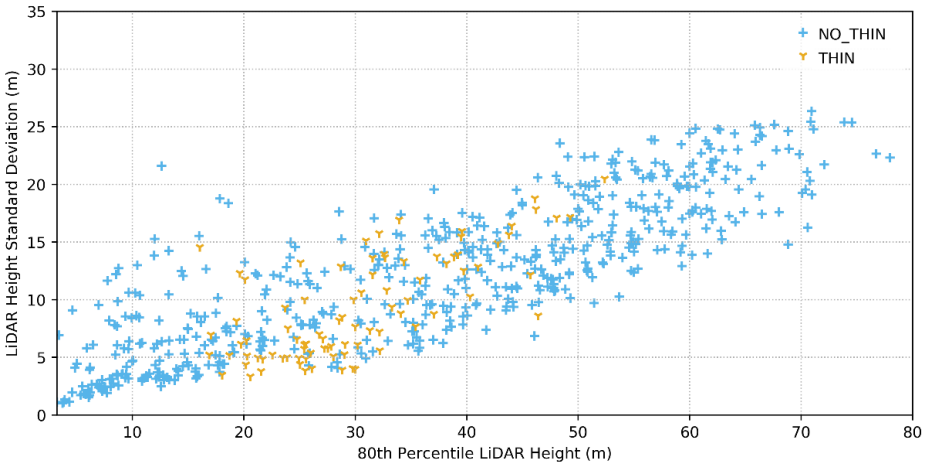

Figure 2.2: The distribution of THIN and NO_THIN field samples as they are distributed against 80th percentile LiDAR height and LiDAR height standard deviation.

Figure 2.2 displays the distribution of field plots with respect to 80th percentile LiDAR height and standard deviation of LiDAR height. It is clear that the data cloud is intermixed with regard to the thinning assignment, i.e. the plots do not demonstrate any linear separation with respect to the THIN or NO_THIN labels. This is a result of

labelling a plot as THIN or NO_THIN based on a rigid set of threshold criteria. For example, it is entirely possible for two plots to be very similar with respect to the LiDAR variables and yet have different THIN or NO_THIN labels. It is also apparent that the thinning eligible plots are less variable with respect to threshold attributes (Table 2.1).

Table 2.1: Summary statistics for field plot attributes according to thinning eligibility class.

	NO_THIN		**THIN**	
Attribute	Mean	Standard Deviation	Mean	Standard Deviation
Basal Area $(m^2/\,ha)$	65.4	48.2	59.7	24.0
Curtis' Relative Density	151.1	93.7	166.1	38.4
Volume $(m^3/\,ha)$	885.2	759.8	548.3	116.6
n	565		88	

2.2.5 Classifiers

2.2.5.1 Random Forests

Random forests (RF), introduced in Breiman (2001), is a non-parametric method for regression and classification problems. RF has been used in conjunction with aerial LiDAR data for forest classification problems. Valbuena et al. (2016) used a RF classifier to predict eight different forest development classes and found the performance comparable to other non-parametric methods, with a kappa value of 0.66. Fedrigo et al. (2018) used a RF classifier in conjunction with LiDAR-derived structural profiles to predict rainforest areas and found similar performance to that of linear unmixing models.

RF produces an ensemble of decision trees using full bootstrap samples with replacement, randomly selecting predictor variables at each terminal node and splitting recursively until some stopping condition is met, as defined by the user. This produces M_{RF} classification trees. After the ensemble has been trained, each tree provides a

prediction for a class, $\hat{C}_m(\boldsymbol{x}_i)$, given a vector of predictors $\boldsymbol{x}_i$ and these predictions are aggregated to produce a score,

$$\hat{y}_i = \frac{1}{M_{RF}} \sum_{m=1}^{M_{RF}} \hat{C}_m(\boldsymbol{x}_i) \qquad \hat{C}_m(\boldsymbol{x}_i) \in \{0, 1\} \tag{2.4}$$

that the *i*th population unit is THIN. We used the package *sklearn* along with the *RandomForestClassifier* implementation of random forests which provides several hyperparameters available for tuning the performance (Pedregosa et al. 2011).

Tuning the hyperparameters of a classification or regression procedure is essentially a way to control the bias-variance trade-off. Classifiers that are too sensitive to changes in covariates, i.e. they overfit the training data, have a large variance and, conversely, classifiers that are not sensitive enough to changes in covariates have a larger bias (Hastie et al. p. 221). The ideal classifier is one that minimizes a specified loss function, which can be decomposed into bias and variance components (Friedman 1997). Controlling the bias-variance trade-off, then, is a means to selecting the classifier configuration that minimizes a specified loss function among all possible candidates.

In random forests, the bias-variance trade-off is primarily controlled by adjusting behaviors of the M_{RF} trees, referred to as tree-specific hyperparameters. We focused on two tree-specific hyperparameters, *min_samples_leaf* and *max_features*, see Table 2.2 for descriptions of these hyperparameters. In both cases, deeper trees and more covariates, trees are able to learn complex patterns in the training data, but may risk overfitting and generalize poorly to new data.

2.2.5.2 XGBoost

XGBoost (XGB) is a recently developed ensemble learning method that is a variant of tree boosting (Chen and Guestrin 2016). XGB is relatively new, and has not seen as many particular applications to forest structure classification. However, some studies have used XGB for other forestry problems. Sandino et al. (2018) used XGB in conjunction with hyperspectral imagery attained via an unmanned aerial vehicle (UAV) to classify forested pixels by degree of pathogen infection. Moore and Lin (2019) used XGB to determine the drivers for wind damage attrition for radiata pine in New Zealand.

XGB is a particular example of a larger family of methods referred to as tree boosting. Applications of tree boosting in more general remote sensing contexts typically focus on a particular variant of tree boosting, referred to as gradient tree boosting (e.g. Freeman et al. 2015, Yang et al. 2018). The objective in gradient tree boosting is to define a function, $\hat{f}$, using a series of base learners defined in an iterative process, referred to as forward stagewise additive modelling (FSAM). In classification contexts, a classification tree is used as the base learner. An initial base learner is fit to the training data, and a vector of negative gradients (i.e. residuals), $\boldsymbol{g}_0$, is calculated that determines the "step direction". The iterations proceed by adding a new base learner that more closely fits the negative gradients from the previous stage, $\boldsymbol{g}_{m-1}$. The degree to which the new base learner adapts to the negative gradients, referred to as a "step length", is determined typically by a line search procedure (Hastie et al. 2009 p. 359). XGB is a variant of FSAM that modifies this iterative process such that both the step direction and step length are optimized at each iteration, without the use of a separate line search, referred to as Newton tree boosting (Nielsen 2016).

In this process, it is clear that overfitting may occur if too many iterations are conducted. An important topic in boosting is the method in which each iteration is penalized to prevent overfitting, called regularization. Similar to RF, an analyst is free to set a number of tree-specific hyperparameters toward this end. Gradient boosting, and XGB in particular, also rely on an important global hyperparameter (i.e. one that does not control tree behaviour) called the learning rate.

Most tree-specific hyperparameters in XGB mirror those in the RF case, however XGB provides a much larger number of tree-specific hyperparameters. See Table 2.2 for descriptions of the hyperparameters investigated in this study. Due to the vast quantity of hyperparameters available for adjustment in XGB, we focus our attention on adjustment of the *learning_rate* hyperparameter as well as the tree-specific *colsample_bytree* and *subsample* hyperparameters. We also adjusted the maximum number of trees, *n_estimators* and the maximum depth of each tree *max_depth*. Previous studies have suggested focusing on these hyperparameters to simplify the model search (Xia et al. 2017).

2.2.5.3 Logistic Regression

Logistic regression (LOG) is a commonly applied parametric classification procedure in remote sensing and other contexts. LOG is firmly rooted in the remote sensing and forest inventory literature as a popular method for classification problems. Senécal et al. (2018) used LOG to discriminate between canopy gaps and non-regenerative openings (openings not caused by tree damage) using a low-density aerial LiDAR dataset and automated canopy gap detection methods. Mund et al. (2015) used LOG to predict multi-layered and single-layer forest pixels in eastern Germany. They used high-density aerial LiDAR data

(> 25 pulses / m^2) and produced an overall classification accuracy of 90%.

As a specific case of generalized linear regression, logistic regression seeks to model the log-odds of the probability of an event occurrence, or membership to a class, as a linear combination of a vector of predictors (Hastie et al. 2009 p. 119). In the binary case the response, y_i, is assumed to be a Bernoulli random variable:

$$y_i \sim Bernoulli(p_i)$$
$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{2.5}$$

where p_i is the probability that the ith population unit is THIN, $\boldsymbol{\beta}$ is a vector of slope coefficients, and $\boldsymbol{x}_i$ is a vector of predictors for the ith population unit.

One concern in parametric modelling is the variable selection process. The purpose of variable selection is two-fold, the first is to improve predictive accuracy and the second is to improve interpretability of the fitted model. Variable selection is typically done in the remote sensing and LiDAR literature using best subsets selection via a stepwise selection procedure. Another option is to use LASSO, or least absolute shrinkage and detection operator. Introduced by Tibshirani (1996), LASSO is a method in regression analysis for automated variable selection and regularization and has a history of use in the LiDAR literature (e.g. Kankare et al. 2013, Takayama and Iwasaki 2016). A vector of scaled estimated regression parameters $\widehat{\boldsymbol{\beta}}^*$ exists in variable space. The error around this vector can be envisioned as an n-dimensional ellipsoid. LASSO performs variable selection by finding the points at which the error contours intersect with an n-dimensional hypercube placed at the origin, thus forcing certain coefficient errors to "shrink" to zero for some given hyperparameter c. We investigate the use of LASSO as it provides an automated way to control the bias-variance trade off in logistic regression

contexts, and therefore is conceptually similar to the automated variable selection processes employed in RF and XGB.

Table 2.2: Names, descriptions and types of each hyperparameter for the classification methods considered in this study.

Classifier	**Type**	**Hyperparameter**	**Description**
RF	Tree-Specific	min_samples_leaf	The minimum number of samples needed at a node for splitting. Primarily controls tree depth.
		max_features	The number of LiDAR covariates considered at a node to find the optimal covariate for splitting.
	Global	n_estimators	The number of total trees constructed, M_{RF}
XGB	Tree-Specific	max_depth	The maximum depth of an individual tree.
		subsample	The proportion of training data used to generate trees.
		colsample_bytree	The proportion of LiDAR covariates used to generate trees.
	Global	learning_rate	The shrinkage parameter applied to each new tree at each iteration.
		n_estimators	The number of trees (i.e. iterations) constructed.
LOG	Global	c	The strength of regularization, smaller values indicate stronger regularization.

2.2.6 Classifier Assessment

Binary classifiers are assessed in the remote sensing literature using myriad approaches contingent on the end-use of the classifier, the structure of the data and the validation procedure. One commonly used performance measure is the area under the receiver operating characteristic curve (ROC AUC) (e.g. Singh et al. 2015, Melin et al. 2015). When the response variable is the probability of class membership, the threshold probability value to assign a population unit to a class, called the discrimination

threshold, is an arbitrary selection. Therefore, performance measures that rely on a discretized response, such as overall accuracy and the kappa statistic, will only assess the performance of a classifier for a given threshold. The ROC captures this information for a given classifier by varying the discrimination threshold and computing the false positive and true positive rates at each threshold value using a set of testing data. From this curve the AUC is calculated and serves as the primary performance measure for a given classifier.

AUC is a random variable contingent on the selection of training data used to train a given classifier and thus must be estimated using a cross-validation procedure. In this case we will adopt a stratified k-fold cross-validation procedure. In stratified k-fold cross-validation the data are randomly assigned to k subsets, while attempting to maintain a distribution of classes that is identical to that of the original dataset across each of the k folds (Diamantidis et al. 2000). The classifier is trained on $k-1$ of the subsets and measures of performance are produced using the remaining subset. This process is repeated k times, iterating through each fold, to estimate the following quantity:

$$\widehat{AUC}_{KF} = \frac{1}{k}\sum_{i=1}^{k} \widehat{AUC}^{(i)} \tag{2.6}$$

where k is the number of folds and $\widehat{AUC}^{(i)}$ is the predicted AUC for the ith fold. In the case of $k = n$ we refer to this quantity as a leave one out estimate of AUC.

In some cases, we wished to compare classifiers in a threshold setting. This required selecting a threshold value to bin the continuous predictions produced by the

classification methods to a binary output. For each configuration and at each threshold value between 0 and 1, the kappa statistic was estimated:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.7}$$

$$\hat{\kappa}_{KF} = \frac{1}{k}\sum_{i=1}^{k} \kappa^{(i)} \tag{2.8}$$

where p_o is the overall accuracy and p_e is the probability of agreement. The threshold value that produced the largest value of $\hat{\kappa}_{KF}$ was selected as the optimal threshold to produce binary predictions. Confusion matrices and its derivatives such as overall accuracy, true positive rate and false negative rate were also produced as outputs to assist in the assessment of threshold classifiers.

2.3 Results

2.3.1 General Performance Assessment of Classifiers

LOG, RF, and XGB performances can be compared in the unthresholded case using $\widehat{AUC}_{KF}$. Figure 2.3 displays a summarized 10-fold ROC curve along with unsummarized ROC curves for each fold. The ideal classifier would reach the point such that the true positive rate is 1 and the false positive rate is 0 (i.e. the top left corner). A naive classification model would follow the diagonal line. XGB and RF clearly outperform LOG by this measure of performance, with $\widehat{AUC}_{KF}$ values of 0.86, 0.77 and 0.68 respectively. XGB and RF also and tend to have smaller variance across folds, i.e. the estimate of their performance is more precise. XGB and RF perform similarly. This result is consistent with the threshold models discussed previously. Table 2.3 reports the mean

and standard error of $\widehat{AUC}_{KF}$ for each classifier, while Figure 2.3 displays the behaviour of each fold at each threshold value, along with error envelopes.

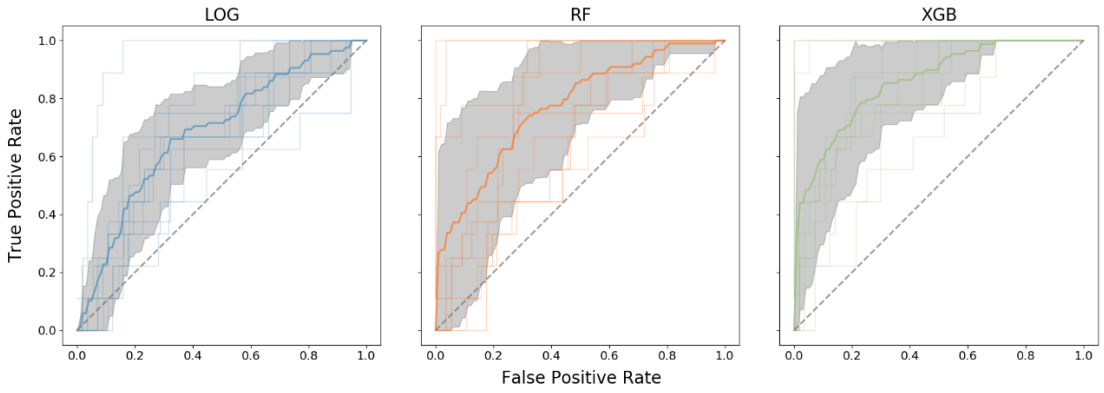

Figure 2.3: Summarized ROC curves across k=10 folds with unsummarized fold-specific ROC curves for each classification method. The dark colored line in each pane indicates the mean true positive rate across the 10 folds. The fainter colored lines indicate true positive rates for each individual fold. The gray envelope indicates the standard deviation of the true positive rate at a specified false positive rate. The dashed line indicates the behavior of a naïve classification, such that a given prediction is a Bernoulli random variable with success probability 0.5.

Table 2.3: AUC point estimates and standard errors for classification methods using a 10-fold cross validation.

Classifier	$\widehat{AUC}_{KF}$	**SE**($\widehat{AUC}_{KF}$)
LOG	0.68	0.10
RF	0.77	0.15
XGB	0.86	0.11

The threshold performances of the selected classification procedures, measured by $\hat{\kappa}_{KF}$, displayed in Figure 2.4 show RF and XGB with similar performance followed by LOG in terms of kappa. Our results indicate that LOG had a $\hat{\kappa}_{KF}$ of 0.24, which is considered poor performance, while RF and XGB had $\hat{\kappa}_{KF}$ values of 0.42 and 0.40 respectively, which is considered good performance (Landis and Koch 1977). Additionally, confusion matrices and derivative products such as overall accuracy and other measures are reported in Table A 6 using the final selected models from the full sample configuration search.

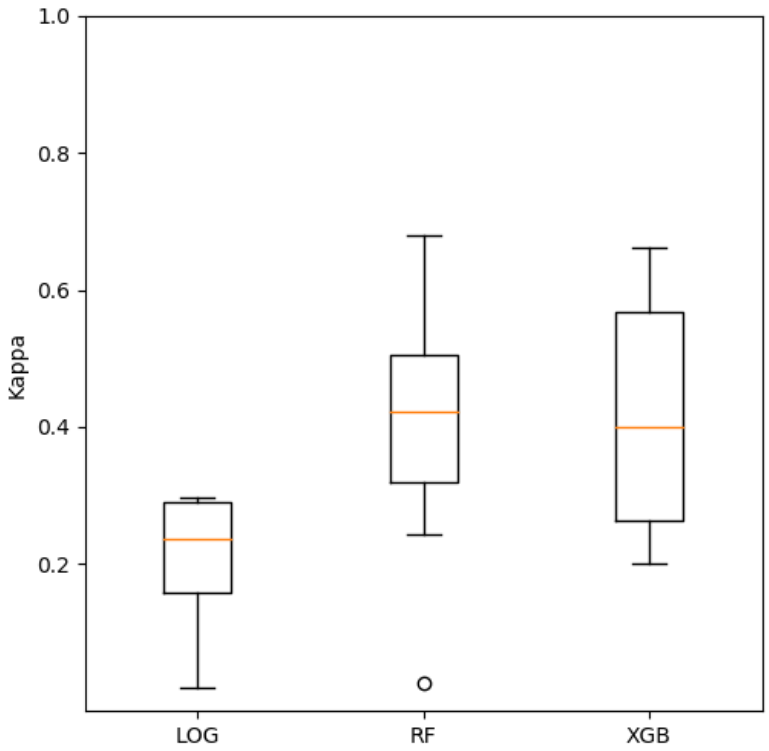

Figure 2.4: $\hat{\kappa}_{kf}$ of thresholded classification methods estimated using a 10-fold cross validation. Horizontal lines represent the median, box edges represent the upper and lower quartiles, and whiskers represent the range of the 10 folds.

2.3.2 Downsampling Performance Assessment

At each downsample size, and for each classification method, the configuration with the largest $\widehat{AUC}_{merge}$ was considered the best performing configuration. Figure 2.5 demonstrates the performance of the best configuration for each classification method at each sample size, measured by $\widehat{AUC}_{merge}$. Performance stabilizes for XGB and RF at $n = 200$ and LOG stabilizes at $n = 300$. Additionally, XGB and RF performance seem to increase at a much faster rate than LOG between $n = 30$ and $n = 100$. This figure also demonstrates the very similar performances of XGB and RF at $n = 200$. Additionally, large variances in classifier performance are evident at small sample sizes. This is exceptionally apparent for RF at sample sizes less than 300.

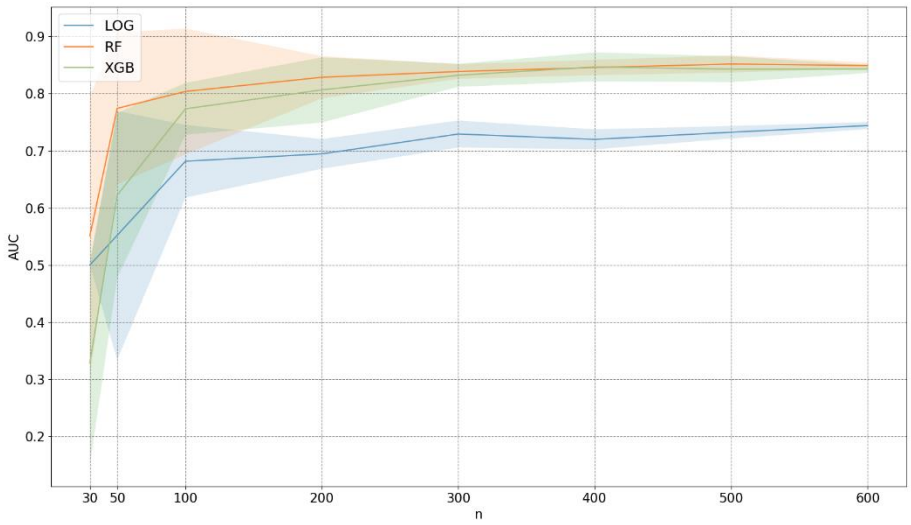

Figure 2.5: Values of $\widehat{AUC}_{merge}$ for the best configuration found for each classification method using reduced sample sizes, the standard error from the n_{pop} simulations describes the envelope around each method line.

One of the benefits of conducting the semi-exhaustive search procedure at each level of n is that we were able to track individual model specifications across the sample size gradient. This provides insight into the sensitivity of each hyperparameter configuration and their absolute and relative performances. RF configurations were found to be sensitive to the depth of each tree and the number of features considered for splitting each node, which we controlled via the *min_samples_leaf* and *max_features* hyperparameters respectively. Figure 2.6 shows the n-wise relative performance ranks for a subset of the models tested in the search procedure, confined to a region of that consisted of the highest performing configurations measured by $\widehat{AUC}_{merge}$. RF configurations trained on larger sample sizes benefited from increased tree depth, whereas forests trained on smaller sample sizes benefited from shallower trees and fewer features. This result can be framed with respect to the bias-variance trade-off. Reducing

max_features and *min_samples_leaf* effectively reduces the amount of information available to each base learner, which reduces the variance of the trained classifier. In the small sample size case these hyperparameters are controlling variance by reducing the amount of information available to each base learner in exchange for bias, i.e. systematic prediction error. In the large sample case, increased *max_features* does not produce high-performing configurations, with the best configurations using only two variables for splitting at each node. However, in the large sample case performance is very sensitive with regard to tree depth, favoring increased values of *min_samples_leaf*.

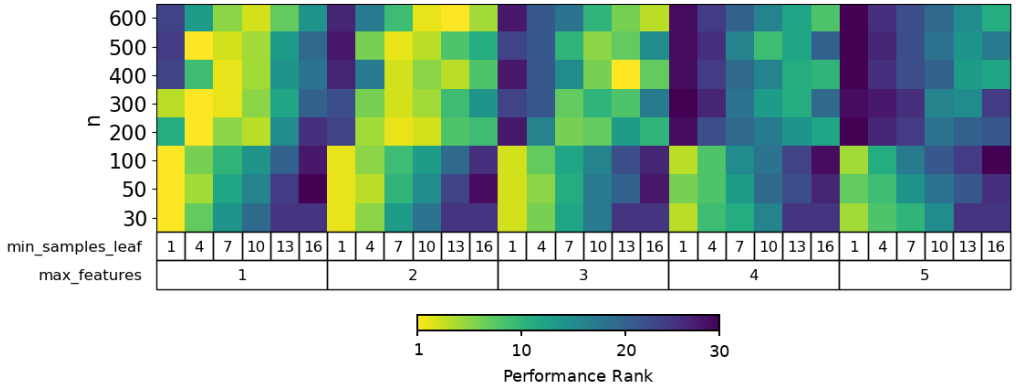

Figure 2.6: Relative classifier performance for a subset of 30 of the original RF configurations. Configurations are ranked by $\widehat{AUC}_{merge}$ row-wise. The lightest color in a given row indicates the best model for that level of n_d. To reduce the number of models displayed, only configurations of 1000 trees were included, this hyperparameter did not have a high impact on performance.

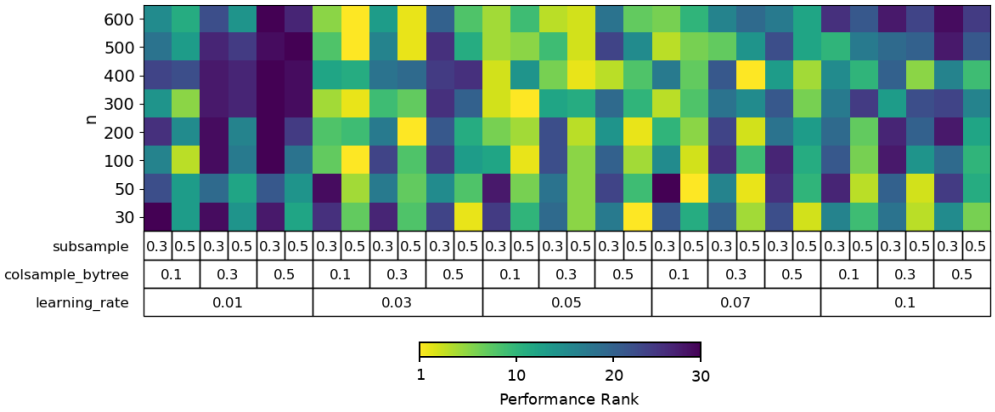

Figure 2.7: Relative classifier performance for a subset of the original XGB configurations. Configurations are ranked by $\widehat{AUC}_{merge}$ row-wise. The lightest color in a given row indicates the best model for that level of n_d. To reduce the number of models displayed, only configurations with a depth of 6 and 100 trees are included, these hyperparameters did not have a high impact on performance.

Discussion around XGB hyperparameter tuning, and gradient boosting in general, tends to focus most on the adjustment of the *learning_rate* hyperparameter (e.g. Xia et al. 2017). Our results confirm that *learning_rate* has a strong effect on classifier performance rank, with many of the best models for n = 200, ..., 600 aggregating in *learning_rate* = 0.03 and 0.05 configurations. For smaller sample sizes, it is clear that controlling the *subsample* hyperparameter increases the rank of those configurations (where an increase in rank refers to an increase in $\widehat{AUC}_{merge}$ relative to the other configrations). Figure 2.7 demonstrates these relative model performance ranks for XGB.

Smaller values for the LASSO regularization hyperparameter, c, increased the rank for LOG configurations at sample sizes greater than n = 400. At sample sizes less than 400 the best configurations tended to use $c = 1$, which implies no regularization.

2.3.2 Commercial Thinning Classification

To create a useful product for forest managers to select areas for commercial thinning operations two outputs were produced, a threshold binary map with predictions of

thinning eligibility across the study area, and a score map, showing the predicted score of thinning eligibility. We selected the XGB classifier to produce the maps because its performance exceeded that of the other selected models in $\widehat{AUC}_{KF}$. We applied thresholds to this model using the same procedure described in section 2.6 to produce the binary raster map and used its score predictions to produce the continuous map (Figure 2.8).

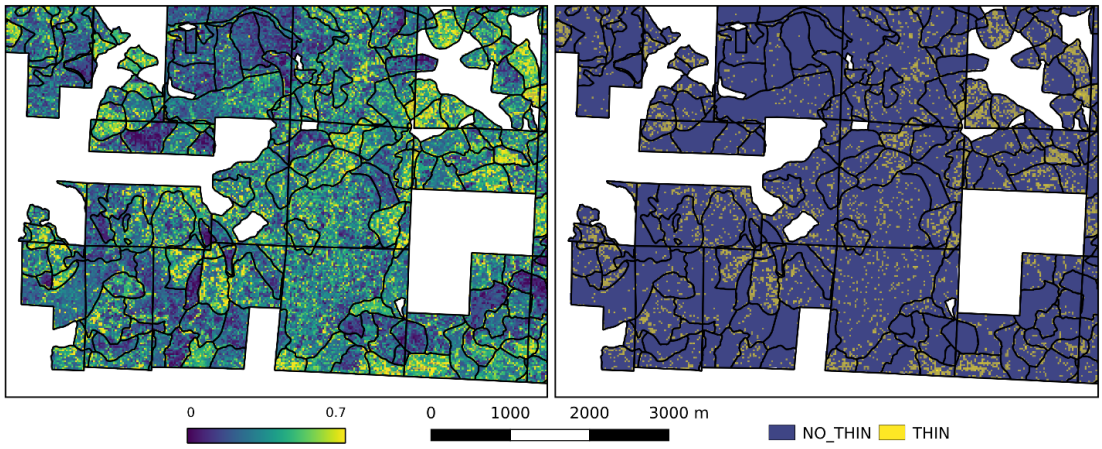

Figure 2.8: Raw score predictions (left) and thresholded predictions via maximum kappa (right) for THIN action generated from the XGB classifier.

In the threshold case the output can be used to provide point estimates of thinning eligible area for a given stand or arbitrary selection of pixels. In the case of continuous predictions, we consider the use of the score for THIN membership to serve as a proxy for confidence of membership because it allows for a more nuanced view of the model output that could be useful in deriving CT project extents that are not restricted to existing stand boundaries.

XGB is also capable of producing a variable importance measure, useful for investigating the influence of predictor variables on the model output. Variable importance for tree-based FSAM is derived in the same way that it is in RF (see e.g. Hudak et al. 2008), and represents the number of times that predictor was used in a

splitting node weighted by the improvement of the loss function. Figure 2.9 shows the variable importance output for the final XGB classifier. In general, it is clear that predictor variables that describe the level of penetration of returns into the canopy and correlate well with volume play a key role in discriminating CT eligible population units using this classifier. These include *elev_p_95*, *all_1st_cover_above2m* and *elev_p25*.

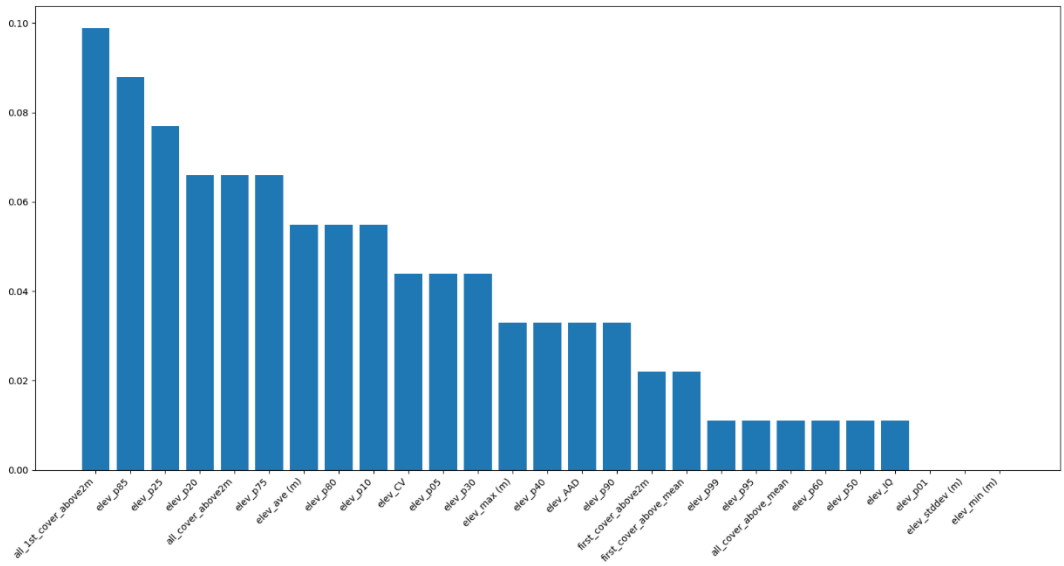

Figure 2.9: Variable importance measures for each predictor variable used in the final XGB model.

2.4 Discussion

2.4.1 Performance and Tuning of Classifiers

XGB had larger values of $\widehat{AUC}_{KF}$ than RF and LOG in the full sample analysis and the dowsampling procedure showed that XGB had more stable performance at smaller sample sizes, as indicated by the widths of the error envelopes in Figure 2.5. The classification methods, when thresholded to produce binary predictions, showed similar patterns. After thresholding the continuous outputs and assessing models via $\hat{\kappa}_{KF}$, the difference in performance between XGB and RF is no longer clear. We selected

hyperparameter configurations based on $\widehat{AUC}_{KF}$ for the full sample case and, in so doing, $\hat{\kappa}_{KF}$ may not be maximized for that particular configuration. If the optimal configuration is desired such that $\hat{\kappa}_{KF}$ is maximized, it should be done so directly in the exhaustive search procedure.

RF and XGB present a large number of hyperparameters for adjustment. It is clear that the optimal values for these hyperparameters is related to sample size (Figures 2.6 and 2.7). For RF, we found that *max_features* and *min_samples_leaf* played the most important role in determining classifier performance, measured by $\widehat{AUC}_{merge}$, while *n_estimators* had little impact. For XGB, it is clear that *learning_rate* provides the most direct way of controlling model performance followed by *colsample_bytree* and *subsample*.

One difficulty we encountered with the XGB method is the large number of hyperparameters available for tuning. We employed a computationally expensive exhaustive search procedure. Other more computationally efficient, but less thorough, methods are available. Xia et al. (2017) propose the use of Bayesian hyperparameter optimization introduced by Hutter et al. (2011), termed a sequential model-based global optimization (SMBO), in the context of tuning gradient boosting classifiers.

Our results indicate that LOG had uniformly the smallest values of $\widehat{AUC}_{KF}$, $\widehat{AUC}_{merge}$ and $\hat{\kappa}_{KF}$. These smaller values for LOG are likely due to the distribution of the class labels as they relate to the LiDAR covariates. Logistic regression performs well when class labels are linearly separable with respect to some linear combination of covariates (Ng and Jordan 2002). If the data cloud is tightly intermixed, as indicated in Figure 2.2 with respect to two LiDAR covariates, then the performance measures of

logistic regression will likely be smaller relative to methods such as RF and XGB that do not rely on linear separability. For other management-oriented classifiers, such as pre-commercial thinning, where the class labels tend to be located at the extremes of the data cloud, logistic regression may attain better performance.

2.4.2 Influence of Sample Size on Performance

The role of sample size in the classification case manifests primarily in two ways. First, sample size has a clear effect on the $\widehat{AUC}_{merge}$ of the classifiers we assessed, with results stabilizing between sample sizes of n = 200 and n = 300. This suggests that operational application of a commercial thinning classifier as presented is feasible for forest management inventories (FMI) that generally have smaller sample sizes than the data set available for SWO. Second, smaller sample sizes will imply larger standard errors of performance measures, and can also result in having too few plots in the minority class to compute k-fold cross-validation measures, this introduces difficulty when comparing the performance of classification configurations and introduces additional uncertainty into the model selection process. Additionally, the types of model validation that can be conducted are restricted. For example, in our study, we had to use $\widehat{AUC}_{merge}$ instead of $\widehat{AUC}_{KF}$ to provide stable performance measures for extremely small sample sizes because not enough minority class elements were available to construct a valid assessment in a traditional k-fold approach. A related but distinct issue in classification assessment is the role of imbalanced data, where imbalance refers to a large difference in the proportions of observed classes in the sample. This study presents an imbalanced data set, with a small portion of sampling units representing the positive, i.e. THIN, class (Table 2.1). Although we did not address this imbalance issue in our methodology, it is important to consider

the implications of handling imbalanced data in classification problems. The impact of imbalanced data is two-fold. First, it can prevent a given classifier from adequately "learning" the minority class. We attempted to correct for data imbalance using over sampling and the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002) but did not observe statistically significant improvements for any of the classification methods in the full sample case, and thus its results are not included. Second, imbalanced data presents issues in assessing classifier performance, leading to large variances in cross-validated measures. The impact of this issue, referred to as dataset shift, can be reduced by employing stratified k-fold cross validation (Moreno-Torres et al. 2012, Sanchez-Lopez et al. 2018) which was employed in our study.

Importantly, imbalanced data and its effects may be more severe than that observed in our study. In forest management inventories (FMI), sampling is often conducted using stratified random sampling, with sampling intensities assigned in each strata according to an auxiliary variable that correlates well with the response variable of interest, typically gross timber volume (Tomppo et al. 2001, 2014; McRoberts and Tomppo 2007). In these cases, thinning eligible stands will be assigned smaller probabilities of selection than other, more mature, stands. Furthermore, in the BLM sampling design for Southwestern Oregon, sampling was conducted using a methodology that is intended to produce a uniform spread of sample units in predictor space (see e.g. Hawbaker et al. 2009), therefore the proportion of CT eligible sample units we observed may be larger than in FMI that have smaller inclusion probabilities for CT eligible units. In the case of CT eligibility, and other response variables such as pre-commercial thinning eligibility, or presence or absence of wildlife habitat, imbalanced data is nearly

guaranteed. Further research is needed to examine the impact of imbalanced data on management-oriented classifiers where data imbalance is more extreme than that presented in this study.

2.4.3 Operational Feasibility

The operational feasibility of a LiDAR-based commercial thinning classifier must be contrasted against ground-based field surveys, which have been the primary alternative for assessing thinning eligibility for the BLM in western Oregon and many other forest management organizations. Although we did not conduct a formal cost assessment in this analysis, some particular considerations for operational applications of the commercial thinning classifier can be identified.

Assessing thinning eligibility using a ground-based field survey typically requires the use of stand exams that employ direct estimators of thinning eligibility attributes (section 2.4). These direct estimators are attained at the stand level using a number of fixed- or variable-radius field plots that measure attributes necessary to derive VOL, DEN and BA. Records of past management and aerial photo interpretation can be used to narrow down the list of stands potentially eligible for thinning. Each of the candidate stands would then need to be sampled with field plots to determine thinning eligibility. The list of candidate stands is potentially large, particularly in complex landscapes and heterogeneous forests, or when management records are limited, making thorough field sampling a potentially costly endeavor.

The proposed CT classification method differs from ground-based field surveys with respect to cost in two important ways. First, a LiDAR acquisition for the entire study area is required to produce the analysis. This represents a significant upfront cost for land

managers, reported anywhere between 0.75 to 3.80 USD (1 to 5 CAD) per hectare, which is primarily a function of study area size and pulse density (Jakubowski et al. 2013; Yebra et al. 2015). Suitability of the CT method using lower cost remote sensing datasets, such as digital aerial photogrammetry (e.g. Strunk et al. 2019), or lower point density LiDAR acquisitions is left for further research. Second, the analysis also requires a set of fixed-radius field plots. Notably, the sample size at which classifier performances stabilized (n = 200 to 300) is orders of magnitude smaller than what might be required in the ground-based field survey case.

Whether or not the CT classification method costs less and is more operationally feasible than ground-based field surveys involves a complex interplay between current cost of LiDAR acquisitions, cost of field staff, the size of the study area and the number of stands under consideration (see Wulder et al. 2008). Other costs induced by the CT classification method are also possible, as it requires more technical infrastructure to process LiDAR data as well as expertise in diagnosing and fitting the proposed classification methods. Likewise, some of the benefits of the CT classification method, such as its increased spatial resolution of predictions, are difficult to quantify in this regard. In addition, land managers are unlikely to use the LiDAR acquisition only for CT classification. LiDAR is valuable for creating many different analyses or processes such as improving digital elevation models (Reutebuch et al 2003), mapping streams (Gaspa et al. 2016), assessing slope stability (Waugh and Shakoor 2015), planning road networks (Grigolato et al. 2017), evaluating habitat quality (Garabedian et al. 2017), and timber harvest operations planning (Pascual et al. 2016).. Ultimately, the CT classification method seems appropriate for forest management inventories where the cost of field

sampling and number of stands is high relative to the cost of a LiDAR acquisition, (e.g., for large and remote forested areas) and where there are multiple benefits for the LiDAR acquisition or if the LiDAR data has already been acquired.

2.4.4 Implications for Forest Management Planning

The evaluation metrics used in this study are limited in that they only produce global measures of performance for the entire training data set. Forest land managers may be interested in levels of uncertainty for arbitrary aggregations of pixels, i.e. extant forest stands, or potential project areas for CT operations. Derivation of uncertainty for subpopulations in the parametric case is well-defined in the small area estimation (SAE) literature in the context of forestry (e.g. Mauro et al. 2016). SAE also includes extensions for logistic regression (Rao and Molina 2015, p. 92). However, measures of uncertainty for subpopulations in the non-parametric case is not as thoroughly explored. Some approaches have been presented in the literature, including attempts to frame non-parametric methods in the context of a population model (Kim and Tomppo 2006; McRoberts et al. 2007; Nothdurft et al. 2009) and resampling methods, such as bootstrapping (McRoberts et al. 2011). Resampling methods for large scale estimation and prediction problems, such as forest inventory, can present significant computational considerations, such as population-level simulation methods (González-Manteiga et al. 2008a), and algorithm-specific methods tend to overwhelmingly focus on the k-Nearest Neighbors method in the regression setting, such as those cited. Further inquiry into management-oriented classifiers could examine the trade-offs of parametric and non-parametric classifiers with regard to development of area-level uncertainty measures. Although we demonstrated that the parametric logistic regression lacks in predictive

performance compared to RF and XGB, this may be outweighed by its deeper connection to small area estimation theory, especially if analytic small-area specific measures of uncertainty are important in the planning process.

2.5 Conclusion

In general, we conclude that commercial thinning eligibility can be discriminated using LiDAR derived height and density metrics using a LiDAR acquisition of moderate pulse density at 8.1 points / m^2. Using $\widehat{AUC}_{KF}$ as a primary performance measure, XGB and RF exhibited similar predictive performance, followed by LOG. We have also assessed the performance of each classifier across a sample size gradient, using $\widehat{AUC}_{merge}$, and have shown that the performance of XGB, RF and LOG stabilize at sample sizes that are feasible for many large scale FMIs. We ultimately recommend the use of XGB for use in predicting CT eligibility, but stress the need for automated tuning methods to mitigate its large computational cost in the hyperparameter search.

Ultimately, management-oriented classifiers should be considered in the larger context of forest management planning. The possible criteria for selecting an area for CT operations are numerous, but some examples include: the proportion of pixels in a specified area predicted as thinning eligible; the level of confidence in those predictions and aggregations therein, the current market conditions for the extracted product; the available resources of the land management agency to execute the planned project; and so on. Future research should focus on the level of importance each of these factors has on optimally allocating areas to CT operations, wherein a commercial thinning classifier and its performance may play a major role.

2.6 Acknowledgments

We would like to give thanks to Kalvin Bailey, Drs. Jacob Strunk, and Peter Gould for their motivation, ideas and insights regarding early ideas and assistance for early drafts. We thank the Bureau of Land Management for providing the data for this research and other support.

3 Manuscript II

Model-based Estimation of Forest Inventory Attributes Using the Semi-Individual Tree Crown Approach

Bryce Frank, Francisco Mauro, Hailemariam Temesgen

Submitted to:

Canadian Journal of Forest Research

Canadian Science Publishing

65 Auriga Drive, Suite 203

Ottowa, ON K2E 7W6, Canada

Abstract

The use of individual tree detection methods to support forest management inventories has been a research topic for over two decades, but a formal assessment of these methods to produce stand-level predictions of forest attributes and their corresponding measures of error is lacking. We leverage model-based estimation methods, including small area estimation methods, in conjunction with the semi-individual tree crown approach (s-ITC), to produce predictions and measures of error for tree volume (VOL), basal area (BA), stem density (DEN) and quadratic mean diameter (QMD) at the level of individual forest stands and the entire study area.

The study area is the Panther Creek watershed in Oregon, USA for which a set of 78 field plots and aerial lidar auxiliary information is available. We compared the point predictions and measures of error of stand-level attributes using the s-ITC method to the more operationally common area-based approach (ABA). For large area estimation, the s-ITC and ABA demonstrated similar performance. For small areas, s-ITC models demonstrated advantages in stands where the sample size increased greatly as a result of the segmentation and was less prone to extrapolation than ABA models, while the ABA demonstrated better performance in stands with small sample sizes and stands without sampled population units for all variables. Our findings motivate further research into these niche applications where s-ITC models may consistently outperform ABA models.

3.1 Introduction

Interest in leveraging tree segmentation for supporting forest management planning is an active area of research in the remote sensing and forest management literature. Forest managers are increasingly attracted to benefits of tree segmentation methods, including opportunities for species identification (Nevalainen et al. 2017; Shi et al. 2018), direct measurement of tree height for use in allometric models (Xu et al. 2018, 2019), and a widening scope of other potential attributes to predict variables such as presence of wildlife habitat (Jeronimo et al. 2018) and tree mortality (Kim et al. 2009; Wing et al. 2015). A common use-case for tree segmentation methods is the prediction of forest inventory attributes at the scale of detected tree crowns that can be used to construct maps of forest inventory attributes at very fine resolutions (e.g. Silva et al. 2016; Xu et al. 2018). These crown level predictions can provide a basis for forest inventory predictions at scales larger than detected tree crowns by aggregating individual predictions to the stand- level.

Approaches to generating predictions of forest inventory attributes that leverage tree segmentation methods are myriad, and are typically classified into two distinct categories: individual tree crown (ITC) and semi-individual tree crown (s-ITC). The s-ITC and ITC methods differ importantly in their conceptualization of the population of interest. In most manifestations, the ITC method makes the explicit assumption that a detected tree segment in the auxiliary data represents a physical tree, and a tree matching method is employed to attach auxiliary information to a measured tree in the sample (Breidenbach and Astrup 2014 p. 114). The population, therefore, is a set of trees, from which a sample is collected. The s-ITC method, in contrast, makes no such assumption,

and a detected tree segment in the auxiliary data represents none, one or several trees. All trees within a delineated segment are assigned to that segment, and a response variable for each segment is computed therein. The population for the s-ITC method is a set of segments, which are population units composed of a number of trees.

The s-ITC method may be attractive for forest managers who desire predictions at the scale of individual tree crowns. S-ITC is therefore an alternative in areas with complex and multi-layered forest structures where ITC methods tend to have larger omission errors. When using the ITC method, predictions for subpopulation parameters, such as mean attributes in a particular forest stand, will tend to be negatively biased due to omission error in the tree detection and matching process (Hyyppa 1999; Peuhkurinen et al. 2011). Therefore, correction factors are needed to adjust aggregations of ITC predictions, as in the case of stand-level predictions. For the s-ITC method, these omission problems do not exist, because it is possible to segment the population in such a manner that it creates a compact tessellation (i.e., the segmentation method does not leave gaps) of the study area that include all trees. S-ITC is still able to provide predictions at the resolution of detected tree crowns, but the interpretation that one segment is equivalent to one tree is lost. Therefore, s-ITC is a compromise between the more operationally common area-based approach (ABA) and the ITC method.

The ABA combines fixed radius plots with coincident remote sensing data in a modeling procedure to produce forest inventory predictions for a set of regular grid cells cast across the study area (e.g. Næsset 2002; 2014). The ABA is advantageous in that it does not require the measurement of tree positions because only the position of the plot is required for georeferencing remote sensing auxiliary information. However, the output

resolution of predictions is constrained to the resolution of the regular grid, and can preclude prediction of attributes at finer scales.

Breidenbach et al. (2010) were the first study to use the term “semi-individual tree crown”, and the formalization of detected tree crowns as a population type has its origin in Flewelling (2006). To date, efforts in the investigation of s-ITC has largely focused on k-NN predictions of forest inventory attributes. In the former study, the authors used a multivariate k-NN method to generate segment level predictions of species-specific and all-species timber volume. They used an upscaling procedure on detected segments to assess the semi-ITC method at approximately the same resolution as an area-based method and found comparable accuracies at this scale. Rahlf et al. (2015) applied the semi-ITC using digital aerial photogrammetric point clouds (DAP) as an auxiliary dataset using a similar upscaling procedure and found similar accuracies.

Importantly, the ABA has been deployed with explicit consideration of predictions for parameters of subpopulations of interest, such as the total or mean volume of timber in a forest stand, compartment, or county using small area estimation (SAE) methods (Breidenbach and Astrup 2012; Goerndt et al. 2013, Mauro et al. 2016; Magnussen and Breidenbach 2017). In many forest inventories it is typical for subpopulations, such as forest management stands, to have very small sample sizes, i.e. “small areas” (Rao and Molina 2015 pp. 1-3), which can preclude reliable estimation of stand-level parameters using direct design-based estimators. In other cases, some stands may be entirely unsampled, which requires the use of synthetic prediction (Rao and Molina 2015). Small area estimation (SAE) methods allow for the consideration of these two cases and obtaining predictions and their corresponding uncertainty measures, such

as stand specific root mean squared errors. These uncertainty measures are analogous to the standard errors of direct estimators produced from field-based exams given the assumptions of employed models are met. Thus, they can be used as quality controls to determine the need for extra field data collection and establish the reliability of a model-based prediction of a stand-level parameter.

SAE provides model-based techniques that leverage linear mixed models (see Rao and Molina 2015 Section 1.5) for predicting stand-level parameters and their uncertainties for both sampled and unsampled stands. The model coefficients are estimated using pairs of observed quantities of interest and auxiliary information. For sampled stands, empirical best linear unbiased predictors (EBLUPs) that are weighted averages of a direct field-based estimate and a synthetic component are typically used. For unsampled stands, synthetic predictions of stand-level parameters are made assuming the model holds for unsampled stands using only the fixed effect part of the mixed model (e.g. Goerndt et al. 2011). The relative amount of synthetic predictions is perhaps the defining characteristic of SAE in forest inventory contexts, as typically few stands have sampled population units within them.

An important result of many model-based SAE analyses is the estimate of the random effect variance, i.e. the share of the variance that can be explained by subpopulation membership. If the random effect variance is large, it implies that SAE methods that adjust predictions for different subpopulations, typically attained via use of the EBLUP, are necessary (Datta et al. 2011). Such adjustments are only possible in subpopulations with sampled population units. In the case of the ABA using lidar auxiliary information, the random effect variance has been found to be large for some

forest inventory attributes. Mauro et al. (2017) found significant random effect standard deviations for volume (m^3 ha^{-1}), stems per hectare, basal area (m^2 ha^{-1}), and quadratic mean diameter (cm) for a study area in southwestern Oregon. Breidenbach et al. (2018) found random effect standard deviations for timber volume (m^3 ha^{-1}) between 30 and 37 when using digital aerial photogrammetry as auxiliary data for a study area in southeastern Norway using the ABA. However, such an assessment has not been conducted in the case of the s-ITC method. If a large random effect variance is present in the case of s-ITC, then the risk of biased predictions of stand-level parameters may be non-negligible if predictions are not adjusted for sampled stands, or if synthetic prediction for unsampled stands is desired.

As tree segmentation methods continue to be employed for forest inventory predictions, a rigorous assessment of the error properties of the s-ITC method is necessary, and can have major implications for forest inventory operations planning. We employ SAE methods, specifically the unit-level model (Rao and Molina 2015, pp. 78-81) to produce predictions of four forest inventory variables including stem volume (VOL), basal area (BA), stem density (DEN), and quadratic mean diameter (QMD) for two different population types: 1) a population of detected crown segments produced in the manner of s-ITC, and 2) a population of grid cells produced in the manner of the ABA. Particularly, we focus on the relative performances of these inventory models at the scale of individual unit predictions, forest management stands and at the scale of the entire study region using uncertainty measures derived for small areas. In so doing, we develop recommendations for the application of the s-ITC method for producing stand-level predictions of forest-inventory attributes.

3.2 Materials

3.2.1 Study Area

The study was conducted in the Panther Creek watershed located in northwestern Oregon, USA. The area is composed of approximately 2,580 hectares of forest ranging from 100 m to 700 m in elevation, with an annual precipitation of 1500 mm. The forest types range from planted stands of Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco), to natural stands of mixed species including western hemlock (*Tsuga heterophylla* (Raf.) Sarg), western red cedar (*Thuja plicata* Donn ex D. Don), red alder (*Alnus rubra* Bong.), grand fir (*Abies grandis* (Douglas ex D. Don) Lindley), and other minor species. Various forest management actions have been conducted in the area, and a patchwork of management is apparent, including variable retention harvest, thinning and recent reforestation.

Forest stands and other areas were delineated using visual photographic interpretation methods. A total of 144 delineations were produced, of which 15 were removed in a later interpretation phase that revealed they were either non-forested (e.g. lakes, residential areas, etc.) or recently harvested near the time of the field data collection date. The remaining 129 forest stands compose what will be referred to as the study region.

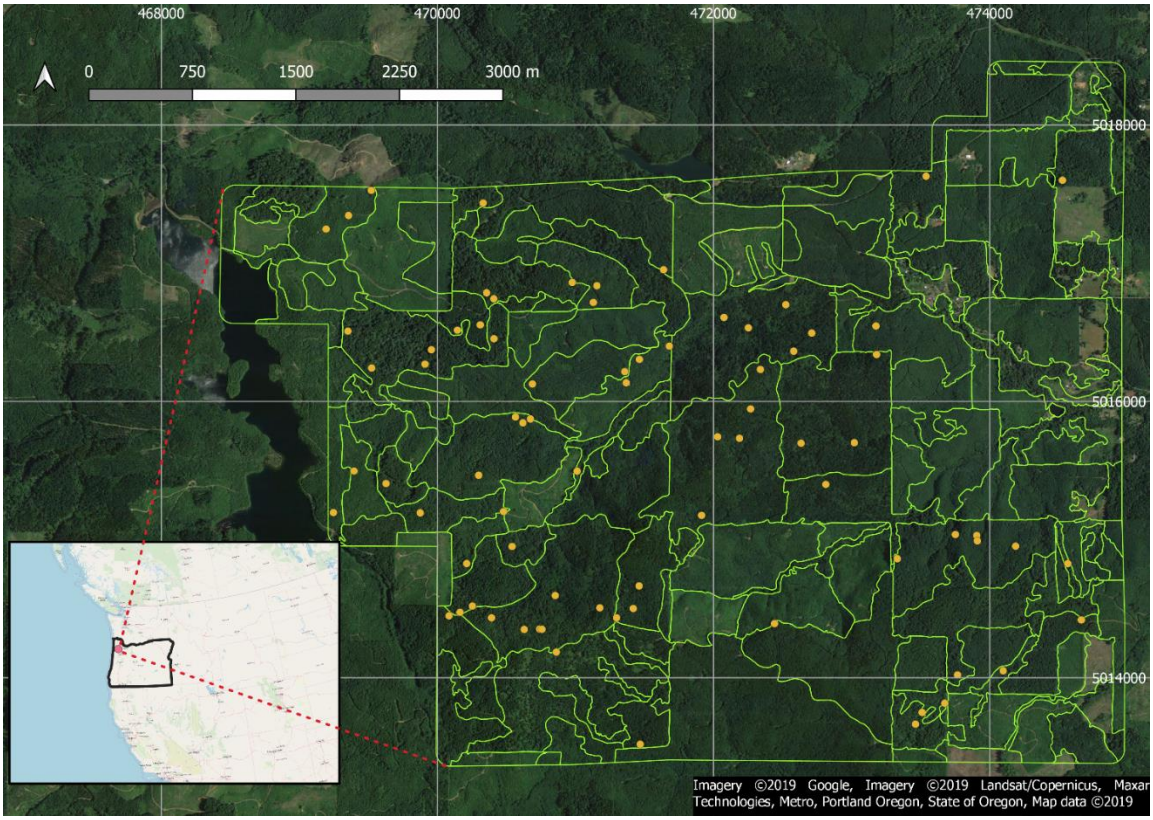

Figure 3.1: Vicinity map of the Panther Creek watershed situated in the northwest of Oregon, USA. Forest stand delineations are indicated in green and field plot positions are indicated in orange. UTM Zone 10 N coordinates are given at the margins of the figure, with UTM grid lines demarcated in gray.

3.2.2 Field Data

A field campaign conducted between July 2009 and May 2010 produced a set of 78 field plots with a fixed radius of 16 meters (≈ 0.08 ha in area) as part of a larger program not explicitly implemented for the objectives of this study. The sampling conducted in the Panther Creek watershed contains a mixture field plots from a probability-based sample selection mechanism, and a non-random sampling design. Stand species compositions were visually estimated using color infrared photos, identifying three major groups including conifer, mixed conifer in association with hardwoods, and riparian zones. 54 conifer stands were identified and divided into nine strata based on 90th percentile lidar

height. Five stands were assigned to the mixed group, which formed the tenth stratum, and two stands were assigned to the riparian group, which formed eleventh stratum.

For the conifer group, two sampling procedures were performed, including a design-based sample (C-D) and a non-random sampling design intended for model-based estimation (C-M). For the C-D procedure, one stand in each of the 9 strata was selected using probability proportional to size. Within each selected stand, three plot centers were randomly positioned. For the C-M procedure, one plots were selected within each of the nine stands by selecting a grid cell that represented median conditions in terms of number of first returns and 90^{th} percentile lidar height. For the mixed group and riparian group, all stands were selected for sampling, and two plots randomly positioned in each stand. The remaining 36 plots were part of a separate sampling program intended to sample various soil structures in the study area, were not selected with a probability-based design and are not dependent on forest conditions or structures (Flewelling and McFadden 2011).

Field crews located the pre-allocated centers using sub-meter grade GPS units and installed plot centers at the GPS measured positions. These GPS measured positions were assessed against an existing cadastral survey and were found to be within 0.25 m (Flewelling and McFadden 2011). All trees within the plot radius that had a diameter greater than 0.5 cm were measured for diameter and height and had their species recorded. Each tree position was recorded relative to the established plot center by measuring its horizontal distance to the nearest 0.1 m and azimuth. For all trees, predictions of cubic stem volume were computed using the National Volume Estimator Library (Wang 2019). For the purposes of this study, the predicted cubic volume,

including top and stump, for each tree in the ground data tree list are treated as known quantities.

3.2.3 Lidar Data Acquisition and Processing

The aerial lidar acquisition for this study was collected July 15, 2010 during leaf on conditions using a Leica ALS60 sensor mounted on a Cessna Caravan 208B. The aircraft was flown at approximately 900 meters above ground level with a scan angle of $\pm 14^{\circ}$ from nadir. The average pulse density for the acquisition is 20.01 pulses per square meter. The raw lidar acquisition was normalized for terrain elevation by filtering ground and non-ground points by applying the ground-filtering algorithm developed in Zhang et al. (2003). These ground points were used to form an intermediate bare earth model. Empty cells of this bare earth model were interpolated using a nearest neighbor interpolation algorithm to produce the final bare earth model, by which the elevations of each lidar point were subtracted to produce a normalized point cloud of the study area. Implementations of the ground filtering, interpolation and normalization algorithms were provided by the *pyfor* package (Frank 2019).

3.3 Methods

3.3.1 Constructing Population Units

3.3.1.1 Grid Cells

For the ABA, the population units are a set of grid cells that covers the entire study area, such that the grid cell size equals the size of the field plots (0.08 hectares). For each grid cell in the population, a vector of lidar predictors was produced (Table A 8). All grid cells were assigned to the stand in which their geometric center fell, and cells that are assigned to the same stand are considered a subpopulation. Grid cells and field plots are

considered to represent the same population type for the purposes of this study. VOL ($m^3 ha^{-1}$), BA ($m^2 ha^{-1}$), DEN ($stems\ ha^{-1}$), and QMD (cm) were computed at this scale using the entire plot level tree-lists. Mean values and standard deviations of VOL, BA, DEN, and QMD for the field plots are given in Table 3.1. The means and standard deviations tend to agree across field plots and segments, with some disagreement apparent in the VOL attribute. This can be a result of fewer segments included in the sample when segments are large, i.e. in stands with larger crown widths (see section 3.5.3).

Table 3.1: Sample means and standard deviations for forest inventory attributes for the sample of field plots and the sample of segments. For the segment attributes, segments for each plot were first scaled to the scale of the field plots using segment areas as weights.

	VOL ($m^3 ha^{-1}$)		**BA ($m^2 ha^{-1}$)**		**QMD (cm)**		**DEN ($stems\ ha^{-1}$)**	
Source	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Field Plots	601.3	389.4	48.7	23.7	31.9	12.8	659.3	277.2
Segments	542.6	340.8	46.0	22.3	30.1	12.1	687.8	302.2

3.3.1.2 Grid Cells

For the s-ITC, the population units are a set of segments derived from a canopy height model. To produce a population of segments, we utilized a combination of a variable window local maxima filter (Popescu et al. 2002) and a Voronoi tessellation. For the entire study area, an intermediate canopy height model was produced at a resolution of 0.33 meters, where each pixel of the canopy height model represented the lidar return of maximum height. A median filter with a 3 x 3 kernel was passed over the intermediate canopy height model to produce the final canopy height model. Local maxima of this canopy height model were found using a two-stage filter. A fixed window local maxima

filter was passed over the canopy height model such that a distance of at least 2 meters must exist between detected maxima. This provided a set of candidate maxima for a variable window local maxima filter that was passed over the canopy height model in a second phase. For each candidate maximum, an allometric equation defined a search window such that only the highest maximum in the search window was determined to be a final local maximum. We used the allometric equation defined in (Popescu et al. 2002) that relates the height of a given pixel to the search window width:

$$w(\boldsymbol{s}) = a + b \cdot z(\boldsymbol{s})^2 \tag{3.1}$$

where a and b are fixed coefficients and $z(\mathbf{s})$ is the value of the canopy height model at position $\boldsymbol{s}$. We modified the coefficients provided by Popescu et al. (2002) to provide window widths consistent with the observed tree crown radii in the canopy height model using visual inspection such that $a = 2$ and $b = 0.004$.

The variable window local maxima output is a set of points in the study area. We used a Voronoi tessellation over these points to provide the final set of segments for analysis (Figure 3.2). This procedure is similar to that of the segmentation procedure proposed by Silva et al. (2016), but does not constrain the individual segment extents based on low values in the canopy height model. This was considered appropriate as we desired a compact tessellation of the study area (i.e., a division leaving no gaps) that would eliminate the possibility of omitting trees from the sample.

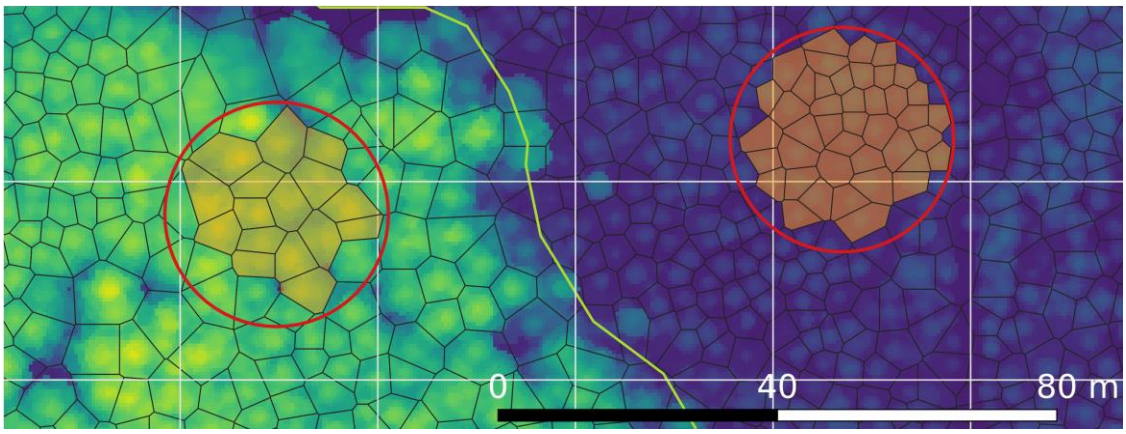

Figure 3.2: Canopy height model (blue to green background) displayed underneath delineated segments (black lines), and grid cells (white lines). Two example field plots are shown (red circles) with included segments shaded in orange.

For each segment, a vector of lidar predictors was computed by clipping the extent of the study-area level point cloud by the extent of each segment. In total, a set of 30 predictors were computed for each segment (Table A 8). In the same manner as the grid cells, segments were assigned to the stand in which their geometric center fell, and a subpopulation is defined as a set of segments that all belong to the same stand. Each segment that was entirely within the plot radius plus the addition of a small constant of 0.5 meters was considered a sampled segment. For each sampled segment, VOL, BA, DEN and QMD were calculated using only those trees whose positions were measured inside the segment. Means and standard deviations of forest attributes are given in Table 3.1 after upscaling the observations to the plot level.

3.3.2 Unit-Level Model

We employ the unit-level model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e} \tag{3.2}$$

to cover both, the cell and the segment cases. In equation (3.2), $y = \left(y_{11}, \ldots, y_{M\,N_M}\right)^T$ is a vector of per-unit area values where y_{ij} denotes the jth observation in the ith stand, M represents the total number of areas and N_i represents the number of population units in the ith stand. $\boldsymbol{X}$ is a $N\ x\ p$ design matrix of lidar covariates and a column of ones to accommodate an intercept. $\boldsymbol{\beta}$ is a $p\ x\ 1$ vector of regression coefficients. $\boldsymbol{Z}$ is a $N\ x\ M$ matrix that assigns population units to forest stands. For example, the elements of the jth column of $\boldsymbol{Z}$ take values of 1 if the ith element of that column is in stand j, and 0 otherwise. $\boldsymbol{v} = (v_1, \ldots, v_M)^T$ is a $M\ x\ 1$ vector of independently and identically distributed random effects such that $\boldsymbol{v} \sim MVN(\boldsymbol{0}, \boldsymbol{G})$. $\boldsymbol{e} = \left(e_{11}, \ldots, e_{M\,N_M}\right)^T$ is a vector of model errors that explain deviations from the stand-level random effect such that its elements $\boldsymbol{e} \sim MVN(\boldsymbol{0}, \boldsymbol{R})$. Additionally, $\boldsymbol{v}$ is independent of $\boldsymbol{e}$. This defines the basic unit level (e.g. Rao and Molina 2015 Ch. 7). Because there is a mismatch in population units between a model developed using circular field plots and the population units upon which predictions are made, i.e. grid cells, we will use the general term "ABA model" when referring to (3.2) if a set of field plots is used to construct the model and "s-ITC model" when referring to (3.2) if a set of segments was used to construct the model. Additionally, the unit-level model, (3.2), is defined for the entire population. In some instances it will be necessary to refer only to the sampled portions of the model. To do so we will use the sub-index S to refer to a matrix or vector that has rows corresponding to unobserved population units removed.

Specific structures are assumed for the variance-covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$. The random effects are assumed independent and identically distributed such that $\boldsymbol{G} =$

$\sigma_v^2 \boldsymbol{I}_M$. Additionally, we incorporate the possibility for heteroskedasticity on the residuals with a general variance function model (e.g. Pinheiro and Bates 2006, p. 206):

$$\boldsymbol{R} = \underset{(i,j)}{diag}(\sigma_e^2 mcp_{ij}^{2\eta}) \tag{3.3}$$

where mcp_{ij} is a variance function covariate that represents the most correlated predictor, σ_e^2 is the residual variance and η is a constant controlling heteroskedasticity. It will be convenient to use the relationship $mcp_{ij}^{\eta} = k_{ij}$ for later formulae, where k_{ij} is a unit-specific variance weight.

3.3.3 Target Parameters

We consider target parameters that are linear combinations of the model coefficients, $\boldsymbol{\beta}$ and the error vectors $\boldsymbol{v}$ and $\boldsymbol{e}$ (e.g. Rao and Molina 2015 p. 98). Let μ_α represent a generic target parameter for a set of population units indexed by α. For example, μ_α may represent a mean of a forest attribute for a particular stand, or a study-region level mean (i.e. a mean of all population units in the study region). In all cases the parameter can be expressed as a linear combination of the model components:

$$\mu_\alpha = \boldsymbol{l}_\alpha^T \boldsymbol{\beta} + \boldsymbol{m}_\alpha^T \boldsymbol{v} + \boldsymbol{q}_\alpha^T \boldsymbol{e}_\alpha. \tag{3.4}$$

Such a generic formulation affords the construction of different target parameters, provided they are linear combinations as in (3.4), by constructing different vectors $\boldsymbol{l}$, $\boldsymbol{m}$, and $\boldsymbol{q}$. We will focus on two specific forms for this study. First, we consider stand-level means. This implies

$$\boldsymbol{l}_i = \left(\frac{\sum_{j=1}^{N_i} h_{ij} \boldsymbol{x}_{ij}^T}{h_{i\cdot}} \right)^T \tag{3.5}$$

which is a $px1$ vector of weighted means of the unit-level predictors in the ith area, where N_i is the number of population units in the ith stand and $h_{i\cdot} = \sum_{j=1}^{N_i} h_{ij}$. Note that in the case of stand-level mean prediction $\boldsymbol{m}_i$ is a vector of zeroes except for the ith position that contains a one. When referring to a generic stand-level parameter, we will use the sub-index i. Second, a study-region mean, μ_τ, is constructed by letting

$$\boldsymbol{l}_\tau = \left(\frac{1}{h_{\cdot\cdot}} \sum_{i=1}^{M} \sum_{j=1}^{N_i} h_{ij} \boldsymbol{x}_{ij}^T \right)^T \tag{3.6}$$

and

$$\boldsymbol{m}_\tau = \left(\frac{h_{i\cdot}}{h_{\cdot\cdot}}, \ldots, \frac{h_{M\cdot}}{h_{\cdot\cdot}} \right)^T \tag{3.7}$$

which are the weighted means of the predictor vectors and a vector of area proportions for each stand in the study region respectively, where $h_{\cdot\cdot} = \sum_{i=1}^{M} h_{i\cdot}$. We will use the sub-index τ when referring to the study-region level parameter.

For both parameter types the term $\boldsymbol{q}_\alpha^T \boldsymbol{e}_\alpha$ is the weighted mean of the errors of a set of population units. For large N_α this term becomes negligible and is typically disregarded for models with independent errors (e.g. Rao and Molina 2015 p. 98). Thus, in all cases we consider target parameters of the form

$$\mu_\alpha = \boldsymbol{l}_\alpha^T \boldsymbol{\beta} + \boldsymbol{m}_\alpha^T \boldsymbol{v}. \tag{3.8}$$

3.3.4 Predictions for Target Parameters

Predictions for the target parameters, $\hat{\mu}_\alpha$, were obtained using the empirical best linear unbiased predictor (EBLUP) by first obtaining an estimate of the variance parameters $\boldsymbol{\delta} = (\sigma_v^2, \sigma_e^2)^T$ via restricted maximum likelihood (REML) using the R package *nlme* (Pinheiro et al. 2019). Estimation of the variance parameters provides a basis for the estimated variance-covariance matrix $\widehat{\boldsymbol{V}}_S = \boldsymbol{Z}_S \widehat{\boldsymbol{G}}_S \boldsymbol{Z}_S^T + \widehat{\boldsymbol{R}}_S$ because $\widehat{\boldsymbol{V}}_S$ depends only on $\widehat{\boldsymbol{\delta}} = (\sigma_v^2, \sigma_e^2)^T$. An estimate of the slope coefficient vector $\boldsymbol{\beta}$ is obtained:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_S^T \widehat{\boldsymbol{V}}_S{}^{-1} \boldsymbol{X}_S \right)^{-1} \boldsymbol{X}_S^T \widehat{\boldsymbol{V}}_S{}^{-1} \boldsymbol{y}_S \tag{3.9}$$

A prediction for the target area parameter can be made using the EBLUP:

$$\hat{\mu}_\alpha = \boldsymbol{l}_\alpha^T \widehat{\boldsymbol{\beta}} + \boldsymbol{m}_\alpha^T \widehat{\boldsymbol{v}} \tag{3.10}$$

where the predicted area-level random effect vector is:

$$\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{G}} \boldsymbol{Z}_S^T \widehat{\boldsymbol{V}}_S^{-1} (\boldsymbol{y}_S - \boldsymbol{X}_S \widehat{\boldsymbol{\beta}}) \tag{3.11}$$

Note that for unsampled areas, $\hat{v}_i = 0$ and we obtain the synthetic predictor $\hat{\mu}_i = \boldsymbol{l}_i \widehat{\boldsymbol{\beta}}$.

3.3.5 Model Selection

For some general attribute $\zeta \in \{VOL, BA, DEN, QMD\}$ and population type $\rho \in \{grid\ cell, segment\}$, we sought the selection of a model that satisfied linearity between predictors and the response, whose Pearson's standardized residuals expressed homoscedasticity. Models were selected using a three-phase procedure. In the first phase, the objective was to select variables for the fixed part of the model, thereby attaining a

preliminary estimate of the slope coefficients $\boldsymbol{\beta}$. From an initial pool of thirty predictor variables, the five best candidate models for a given level of $p \in \{1, 2, 3, 4, 5\}$ predictors plus an intercept were found via adjusted R^2 by an exhaustive search implemented by the *leaps* package in R (Miller 2017). This resulted in a set of 25 candidate models, five for each level of p. Denote the models from this first phase as $mod_{\zeta\rho 1}$. The models in $mod_{\zeta\rho 1}$ were plotted by their adjusted R^2 values and number of predictors, and parsimony was determined by finding the number of predictors at which adjusted R^2 began to level off. Models around this region were explored further via graphical inspection of residual plots. Those models that appeared to satisfy linearity assumptions moved to the second phase and compose the set of models denoted as $mod_{\zeta\rho 2}$. In the second phase, heteroscedasticity was introduced via the variance function described in (3.4) using 0, 0.5 and 1 as values for η. The models in $mod_{\zeta\rho 2}$ were refit with this variance function. The standardized residuals were inspected, and those models that appeared homoscedastic moved to the third phase and compose the set $mod_{\zeta\rho 3}$. If more than one model remained, the model with the highest adjusted R^2 was selected as the final model. In total, eight models were selected as final models, two (one s-ITC model and one ABA model) for each of four attributes.

3.3.6 Mean Squared Error Estimators

The mean squared error estimator for the predicted target parameter $\hat{\mu}_\alpha$ is a function of the estimated variance component vector $\widehat{\boldsymbol{\delta}}$ and can be expressed as the sum of two components, assuming the sampling fraction, $\frac{n_i}{N_i}$, for all areas is negligible:

$$\widehat{MSE}_{\alpha} = g_{1\alpha}(\hat{\boldsymbol{\delta}}) + g_{2\alpha}(\hat{\boldsymbol{\delta}}) \tag{3.12}$$

The first term quantifies uncertainty caused by the random-effect variance:

$$g_{1\alpha}(\hat{\boldsymbol{\delta}}) = \boldsymbol{m}_{\alpha}^{T}\left(\hat{\boldsymbol{G}} - \hat{\boldsymbol{G}}\boldsymbol{Z}_{S}^{T}\hat{\boldsymbol{V}}_{S}^{-1}\boldsymbol{Z}_{S}\hat{\boldsymbol{G}}\right)\boldsymbol{m}_{\alpha} \tag{3.13}$$

The second term quantifies uncertainty in the estimate of the coefficient vector $\boldsymbol{\beta}$:

$$g_{2\alpha} = \boldsymbol{d}_{\alpha}^{T}\left(\boldsymbol{X}_{S}^{T}\hat{\boldsymbol{V}}_{S}^{-1}\boldsymbol{X}_{S}\right)^{-1}\boldsymbol{d}_{\boldsymbol{\alpha}} \tag{3.14}$$

where $\boldsymbol{d}_{\alpha}^{T} = \boldsymbol{l}_{\alpha}^{T} - \boldsymbol{m}_{\alpha}^{T}\hat{\boldsymbol{G}}\boldsymbol{Z}_{S}^{T}\hat{\boldsymbol{V}}_{S}^{-1}\boldsymbol{X}_{S}$. For synthetic predictions, $\boldsymbol{d}_{\alpha}^{T} = \boldsymbol{l}_{\alpha}^{T}$. Equations (3.13) and (3.14), as well as other relevant details are described in Rao and Molina (2015 pp. 108-109).

The sum of the components $g_{1\alpha}$ and $g_{2\alpha}$ quantify the uncertainty of the BLUP for the prediction of the target parameter. That is, they do not account for the uncertainty involved in the estimation of the variance component vector $\boldsymbol{\delta}$, and predictions of area-level MSEs are therefore negatively biased for the EBLUP. A bias correction term, $g_{3\alpha}$ accounts for this additional uncertainty, but this correction term is $O(m^{-1})$ (Rao and Molina 2015 p. 110) and expected to be negligible when m is large (e.g. Mauro et al. 2016, m = 54) and is omitted from this survey. We therefore employ root mean squared error estimators for the BLUP.

$$\widehat{RMSE}_{\alpha} = \sqrt{\widehat{MSE}_{\alpha}} \tag{3.15}$$

Estimated coefficients of variation were also produced for each population type and attribute:

$$\widehat{CV}_{\alpha} = \frac{\widehat{RMSE}_{\alpha}}{\hat{\mu}_{\alpha}} \cdot 100 \tag{3.16}$$

3.3.7 Unit-Level Prediction Accuracy Assessment

To quantify the accuracy of unit-level predictions, a leave-one-out cross-validation procedure was used. For each attribute α and population type ρ four cross-validation measures were produced including absolute and relative bias ($\widehat{mBias}$ and $\widehat{rBias}$) and absolute and relative root mean square error ($\widehat{mRMSE}$ and $\widehat{rRMSE}$).

$$\widehat{mBias} = \frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij}^{*})}{n} \tag{3.17}$$

$$\widehat{rBias} = \frac{\widehat{mBias}}{\bar{y}} \cdot 100 \tag{3.18}$$

$$\widehat{mRMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{ij} - \hat{y}_{ij}^{*}\right)^{2}}{n}} \tag{3.19}$$

$$\widehat{rRMSE} = \frac{\widehat{mRMSE}}{\bar{y}} \cdot 100 \tag{3.20}$$

where $\hat{y}_{ij}^{*}$ is the prediction of the jth population unit in the ith area made from constructing the selected model with that observation removed and $\overline{y}$ is the mean of the observations in the sample.

3.4 Results

3.4.1 Selected Models and Unit-Level Prediction Accuracy

The standardized residuals of the eight models are displayed in Figure 3.3 and their parameter estimates and cross-validation measures are presented in Tables 3.2 and 3.3. Figure 3.3 indicates some minor heteroskedasticity remaining in the DEN and BA models, even after the inclusion of the variance function, but otherwise the ABA and s-ITC models appear to have symmetrically distributed residuals. ABA models appear to have lower residual variance than s-ITC. Across population types, a number of differences in model parameter estimates are apparent, as depicted in Tables 3.2 and 3.3. For all inventory attributes, s-ITC models had consistently higher values of $\hat{\sigma}_v$ than the respective ABA models. Most notably, the ABA models for VOL and BA had near-zero estimates for $\hat{\sigma}_v$. With the exception of QMD, $\hat{\sigma}_e$ was consistently higher for s-ITC models as well. The variance function model term η was consistent for VOL and BA, but was not required (i.e. $\eta = 0$) for the DEN and QMD models in the ABA case to correct heteroskedasticity.

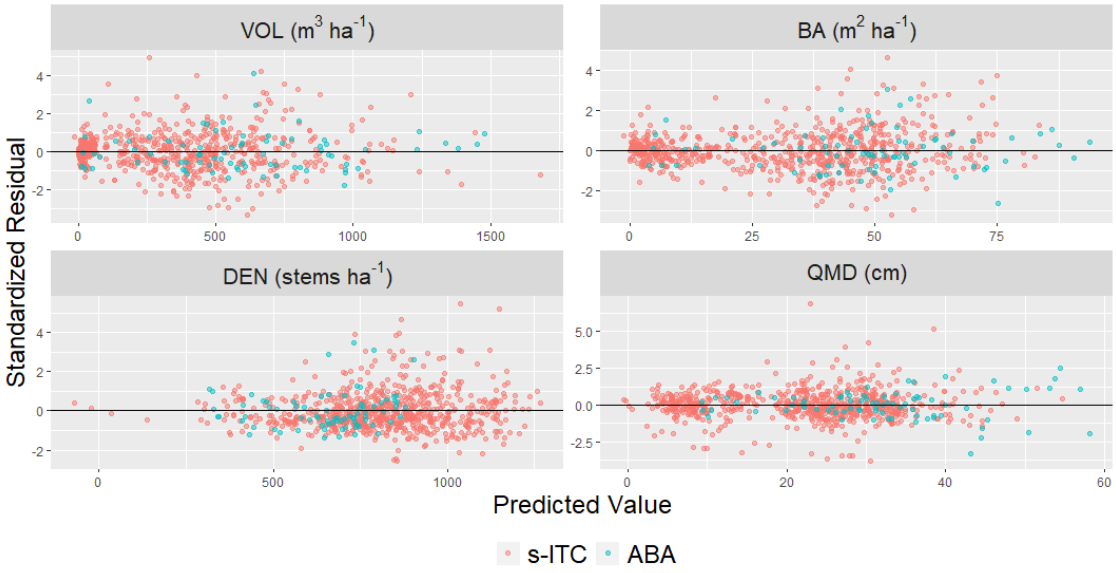

Figure 3.3: Pearson's standardized residuals of s-ITC (pink) and ABA models (blue) for the four forest attributes.

Within population types, there are notable differences in the relationship between $\hat{\sigma}_v$ and $\hat{\sigma}_e$. For SAE methods, high values of $\hat{\sigma}_v$ compared to low values of $\hat{\sigma}_e$ is an important indicator of the possibility of bias in predictions of stand-level means if the random effect variance is ignored. In addition, for models with constant error variance the leading term of MSE_i, i.e., $g_{1i}(\boldsymbol{\delta})$, is proportional to the term $\gamma_i = \frac{\sigma_v^2}{(\sigma_v^2 + \frac{\sigma_e^2}{n_i})}$. Therefore, for models explaining similar amounts of variance, those with smaller values of γ_i should be preferred (Rao and Molina 2015, p. 178) Particularly for VOL and BA in the ABA model, $\hat{\sigma}_v$ is near zero.

The general accuracy measures, $mRMSE$ and $rRMSE$ are observably different across population types. Generally, the s-ITC models express a much larger $\widehat{mRMSE}$ values than the respective ABA models, with a roughly two-fold increase between ABA and s-ITC models for VOL and BA, while the relative increase between DEN and QMD

models is not as large. None of the models expressed bias issues as estimated by the cross-validation procedure.

Table 3.2: Selected predictors, parameter estimates and general performance criteria for models produced using s-ITC approach.

Variable	Model	Predictor	Coefficient	Std. Error	η	$\widehat{\sigma}_v$	$\widehat{\sigma}_e$	mRMSE	rRMSE	mBias	rBias
VOL ($m^3 ha^{-1}$)	ABA	Intercept mean_z mean_z_sq	-8.09 10.50 0.83	12.90 4.02 0.15	0.5	0*	7.37	157.91	26.26%	-0.39	-0.06%
	s-ITC	Intercept mean_z_sq	-20.27 1.25	16.66 0.04	0.5	61.65	8.23	169.61	51.62%	-0.16	-0.05%
BA ($m^2 ha^{-1}$)	ABA	Intercept P_60	0.74 1.98	1.50 0.08	0.5	0*	2.33	11.90	24.45%	0.00*	-0.01%
	s-ITC	Intercept vol_cov	-2.24 2.50	1.69 0.08	0.5	5.7	3.53	15.37	48.22%	-0.02	-0.05%
DEN (cm)	ABA	Intercept P_80 vol_cov	935.42 -33.84 34.32	77.92 7.06 9.21	0	73.17	228.14	241.45	36.62%	-1.37	-0.21%
	s-ITC	Intercept canopy_relief_ratio P_95	210.17 1301.40 -10.26	96.32 152.93 3.41	0.5	175.32	614.55	461.71	56.83%	-0.56	-0.07%
QMD ($stems\ ha^{-1}$)	ABA	Intercept canopy_relief_ratio P_60	16.79 -28.59 1.26	3.57 7.88 0.08	0	1.43	5.83	6.29	20.00%	-0.02	-0.05%
	s-ITC	Intercept P_80 Pct_r_1_above_2m	1.68 0.99 3.09	0.98 0.05 0.87	0.5	2.72	1.54	7.03	32.22%	0.00*	0.00%

* Indicates that the quantity was < 0.01.

3.4.2 Large-area Estimation

For the entire study region we produced a prediction of the target parameter using the respective ABA and s-ITC models by aggregating area-level EBLUP and synthetic predictions (section 3.3). Table 3.3 displays the predictions along with their $\widehat{RMSE}_\tau$ and $\widehat{CV}_\tau$. Some moderate disagreement exists among predictions generated with ABA and s-ITC models, most notably for VOL. With regard to the error, $g_{1\tau}$ is negligible for all models leaving $g_{2\tau}$ as the remaining source of error. $\widehat{RMSE}_\tau$ and $\widehat{CV}_\tau$ is comparable across population types, with ABA models out-performing s-ITC models for BA and DEN, while s-ITC models out-performed for QMD. For VOL, both models performed comparably.

Table 3.3: Large-area predictions of forest attributes and error components. Note that for all models $g_{1\tau}$ was < 0.01 and is not included in this table.

Variable	$\widehat{\mu}_\tau$		$g_{2\tau}$		$\widehat{RMSE}_\tau$		$\widehat{CV}_\tau$	
	ABA	**s-ITC**	**ABA**	**s-ITC**	**ABA**	**s-ITC**	**ABA**	**s-ITC**
VOL ($m^3 ha^{-1}$)	395.73	417.22	159.94	166.05	12.65	12.89	3.12%	3.09%
BA ($m^2 ha^{-1}$)	35.32	36.03	1.06	1.50	1.22	1.03	2.91%	3.40%
DEN ($stems\ ha^{-1}$)	696.27	701.13	1225.02	1690.60	35.00	41.12	5.03%	5.86%
QMD (cm)	25.37	24.97	0.77	0.36	0.87	0.60	3.47%	2.39%

3.4.3 Small-area Estimation

For each area, we produced a prediction of the stand-level target parameter using the respective ABA and s-ITC models. Figure 3.4 displays these area-level predictions and whether or not that particular stand was sampled. Between ABA and s-ITC models, a high level of agreement, i.e. the points in the figure are clustered around the diagonal line, is apparent for VOL, BA, and QMD. For VOL, slight disagreement exists at point predictions exceeding 500 m^3/ha, with the s-ITC model producing slightly higher

predictions than the ABA model. This disagreement is more obvious for unsampled stands than sampled stands. For BA, disagreement exists mostly at the extreme small end of the predicted values, i.e. $< 5m^2/ha$, with the ABA model producing slightly higher predictions than the s-ITC model. This same pattern exists for QMD, but occurring at predicted values $< 15cm$. The disagreement between ABA and s-ITC models is most apparent for DEN. The grid cell model appears to saturate at a level of 900 stems per hectare, with marked differences in the upper range of segment predictions. This holds for both sampled and unsampled stands.

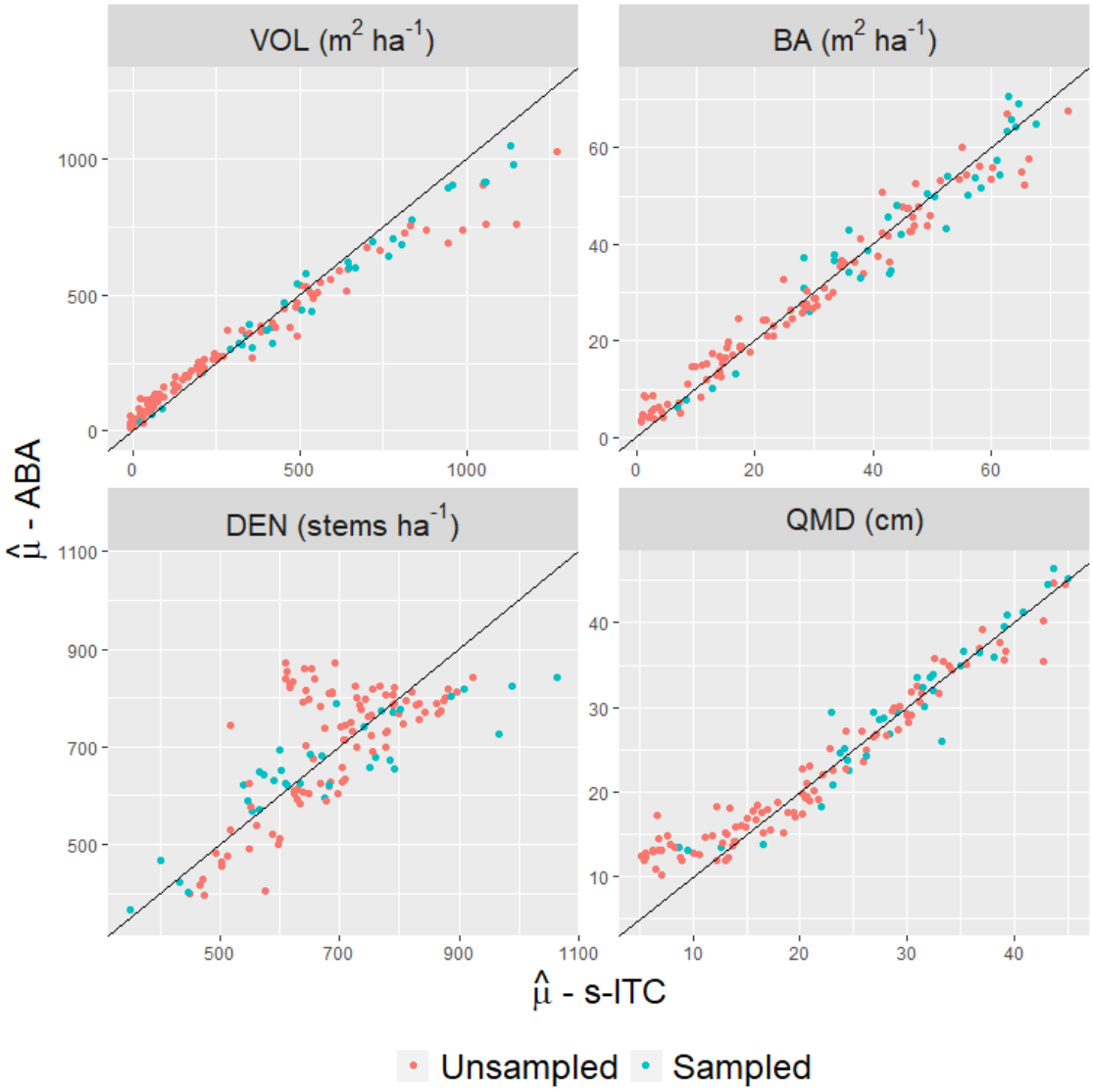

Figure 3.4: Predictions of area-level parameters for VOL, BA, DEN and QMD for s-ITC models and ABA models. Whether or not the stand had at least one field plot is indicated by the color of the point.

For all stand-level target parameter predictions, we produced the model-based mean squared error estimates of sampled and unsampled stands, and coefficients of variation in Table 3.4. With the exception of DEN, the ABA models had lower median measures of error than that of the s-ITC models for both sampled and unsampled stands. For models where random effect variance is near zero, this implies that the realized

random effect is close to zero and that no substantial error is introduced from the random effect component. This was the case for ABA VOL and BA models, and the difference between the measures of error between sampled and unsampled stands is accordingly negligible.

Table 3.4: Medians of estimated mean squared errors for small-areas partitioned by sampled and unsampled stands. In the "Sampled" column "S" indicates a stand with sampled population units and "U" indicates a stand without sampled population units.

Variable	**Sampled**	**Median $\widehat{RMSE}_\iota$**		**Median $\widehat{CV}_\iota$**	
		ABA	**s-ITC**	**ABA**	**s-ITC**
VOL	S	9.7	29.9	2.3%	4.9%
$(m^3 ha^{-1})$	U	9.4	48.4	4.5%	18.7%
BA	S	1.1	2.5	2.3%	6.1%
$(m^2 ha^{-1})$	U	1.2	5.4	4.6%	21.0%
DEN	S	32.3	22.8	4.8%	3.7%
$(stems\ ha^{-1})$	U	89.4	183.0	12.1%	26.4%
QMD	S	1.2	1.7	3.5%	5.8%
(cm)	U	1.9	2.6	10.0%	12.7%

Figure 3.5 indicates the percentage decrease in $\widehat{RMSE}_\iota$ as it relates to a percentage increase in the number of sampling units when going from a grid cell population to a segment population for sampled stands. We refer to this increase in n_i as "segment induced replication." When segment induced replication is large, the decrease in $\widehat{RMSE}_\iota$relative to the ABA model tends to be large. This effect is most clear for DEN and QMD models, with more erratic behavior for the VOL and BA models. When segment induced replication is low, $\widehat{RMSE}_\iota$ tended to increase from ABA to s-ITC models for all variables. This effect is most extreme for the VOL and BA models.

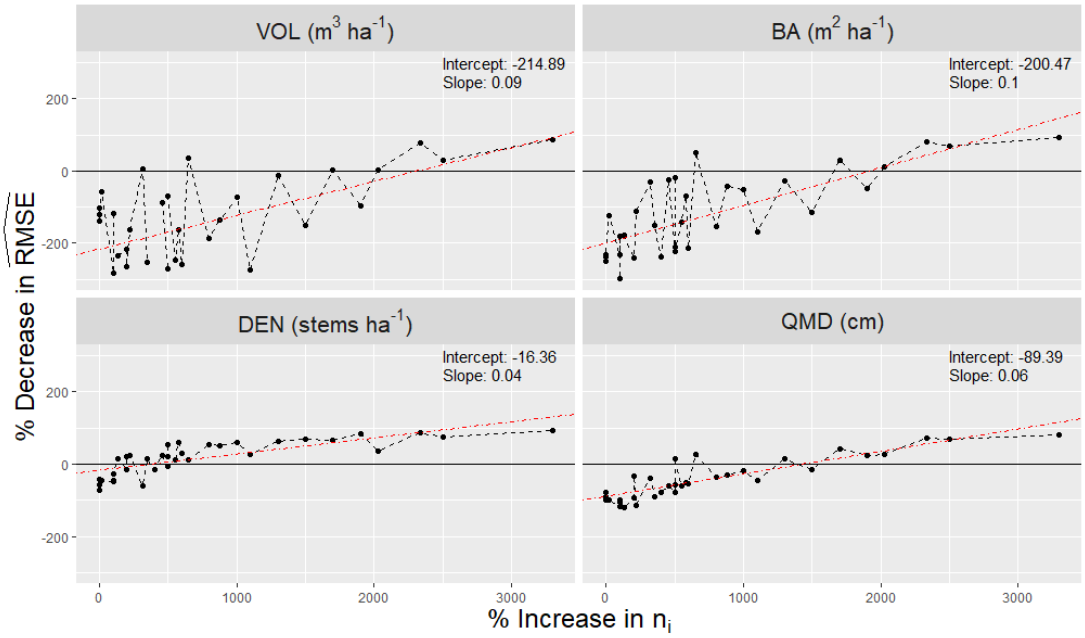

Figure 3.5: The percent increase in n_i, calculated via $\frac{n_{si}-n_{ci}}{n_{ci}} \cdot 100$, when moving from an ABA model to an s-ITC model plotted against the percent decrease in $\widehat{RMSE}_i$, calculated via $\frac{\widehat{RMSE}_{si}-\widehat{RMSE}_{ci}}{\widehat{RMSE}_{ci}} \cdot 100$, for VOL, BA, DEN and QMD in sampled stands. A simple linear regression (red) was fit to each attribute to demonstrate the trend, with slopes and intercepts reported for each attribute.

Figure 3.6 displays the error components, g_{1i} and g_{2i} for sampled stands, ranked in by percentage decrease of $\widehat{RMSE}_i$ in descending order. For models where the random effect variance is large, such as the VOL, BA, and QMD s-ITC models, the relative share of g_{1i} is a very large portion of the total mean squared error for many of the sampled stands. For models where this is not the case, such as the DEN s-ITC model, and DEN, BA and VOL ABA models, the relative share of g_{1i} is low. For most stands, all s-ITC models with the exception of DEN appear to have a lower value of g_{2i} than the respective ABA model, with a notably lower share of the total mean squared error estimate. The ordering of the stands in the figure demonstrates the impact of segment induced replication on the individual error components as well as the overall mean

squared error estimate. While segment induced replication appears to reduce both components in most cases, the impact is most clear for g_{1i} for all s-ITC models with large random effect variance estimates.

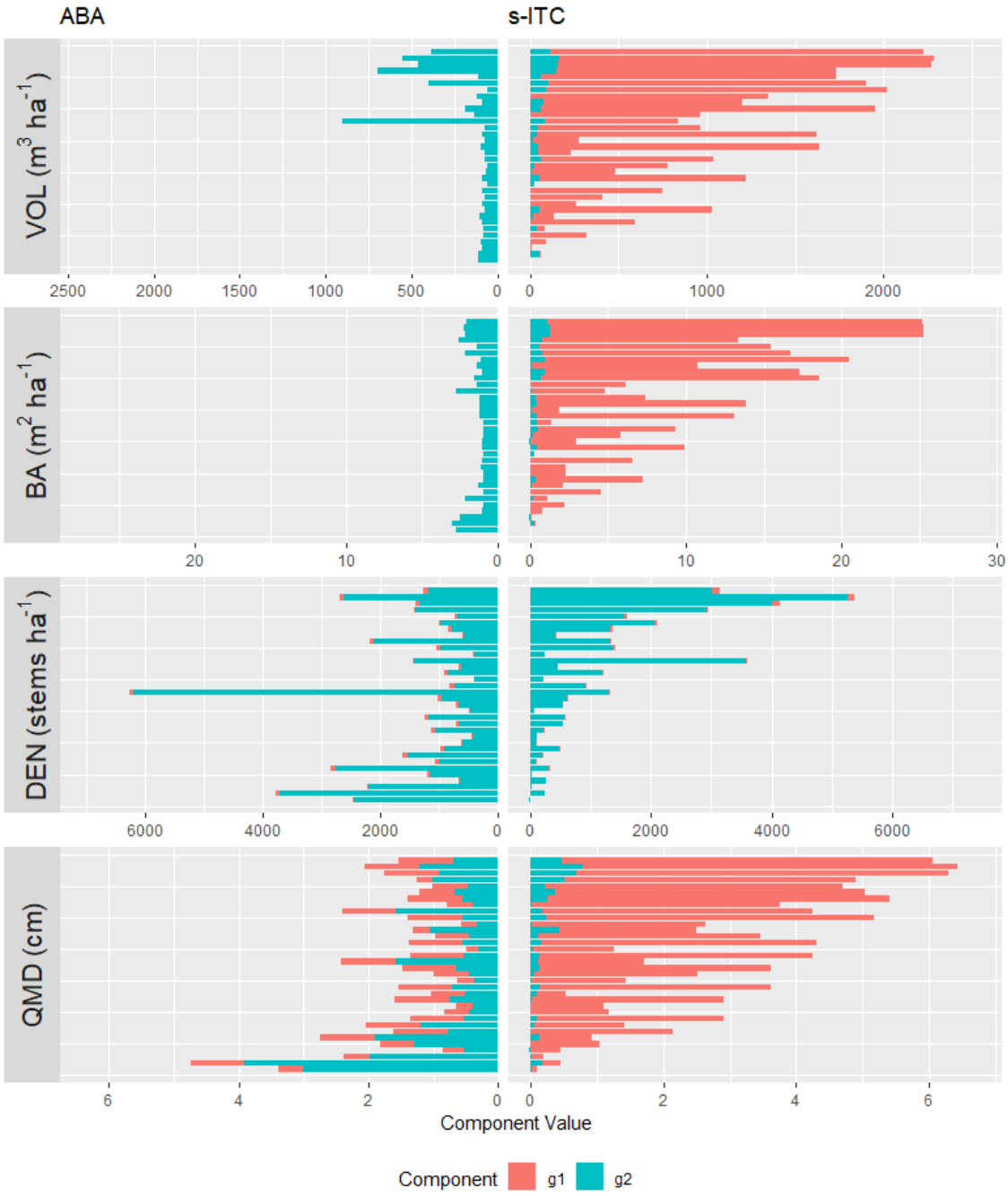

Figure 3.6: Error components for segment and cell models for VOL, BA, DEN and QMD in sampled stands. Component rows are ranked by the percentage increase of n_i (as described in Figure 3.5) induced by segmentation such that the top most row is the stand with the least segment induced replication.

3.5 Discussion

3.5.1 Unit-Level Prediction Accuracy Assessment

The accuracy measures described in section 3.3.7 are effective at communicating the average uncertainty for predictions of individual population units, as opposed to $\widehat{MSE_\alpha}$ which concern aggregations of population units and uncertainty of model parameter estimates. We observed larger values of $m\widehat{RMSE}$ and $r\widehat{RMSE}$ for the cross-validation assessment of the s-ITC models compared to the ABA models (2). This difference is to be expected when dealing with a hierarchy of primary population units (segments, plots) and secondary population units (trees). A decline in the spatial extent of primary population unit, while the secondary population unit remains a fixed entity, will likely result in a higher variance for forests with heterogenous vertical and horizontal structures such as those observed in this study. This relative difference is consistent with previous semi-ITC assessments (e.g. Breidenbach et al. 2010). For both grid cell and segment populations, QMD had consistently the best performance measured by $r\widehat{RMSE}$, followed by BA, VOL and DEN.

3.5.2 Contribution of Error Components

Model-based estimation methods, particularly mean squared error estimators for small area target parameters, provide meaningful partitions of the mean squared error for a prediction that allocate different parts of the error to different sources. While we observed generally poorer performance from s-ITC models for small-area predictions (Table 3.4), this was not uniformly the case (Figures 3.5, 3.6). A thorough discussion of the error components for both s-ITC and ABA models provides insight into the behavior of the error, opportunities for mitigation, and intuition for why certain models may outperform

others in specific situations. We will describe and assess the stand-level error components g_{1i} and g_{2i} in turn.

For a given estimate of model coefficients, an increase in the number of sampled population units in an area is the primary way to reduce the impact of the $g_{1\alpha}$ error component. A clear example of this is the negligible share of the mean squared error estimate expressed in the large-area estimation for this study that has access to all sampled units (see Table 3.3 and equation 3.12). Consider a simplified expression for the stand-specific error component g_{1i} that holds for sampled stands (Rao and Molina 2015 p. 176):

$$g_{1i} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{a_i}}\left(\frac{\hat{\sigma}_e^2}{a_i}\right) = \hat{\gamma}_i\left(\frac{\hat{\sigma}_e^2}{a_i}\right), \tag{3.21}$$

where $a_i = \sum_{j=1}^{n_i} k_{ij}^{-2}$. Note $a_i = n_i$ for models where $\eta = 0$, which is a simpler case useful for illustration. The first term, $\hat{\gamma}_i$, is a number bounded between 0 and 1, and can be interpreted as the proportion of total variance of the target parameter prediction that is due to the random effect variance. For a given model and its estimated parameters, increasing n_i will reduce g_{1i} conditioned on similar variance function covariates.

Figure 3.7 displays the g_{1i} for all sampled stands. For the VOL and BA models, it is clear that the share of g_{1i} for ABA models is near zero and approximately equal for any given level of n_i. For these models, $\hat{\gamma}_i$ is negligible. For all s-ITC models this is not the case, and reduction of this term is primarily due to an increase in sample size. While it is possible for g_{1i} to reduce to a level similar to that of the ABA models, it requires much larger sample sizes, in the realm of 50 for the DEN model and 75 for the VOL, BA, and QMD models. The source of this increased sample size comes primarily from the

particular segmentation method used. For the VWLM + Voronoi method, shorter values of the canopy height model implied larger numbers of detected tree tops in a given vicinity, which in turn increase the number of polygons produced from the tessellation. The degree to which this segment induced replication will reduce g_{1i} is also a function of fixed plot radius, as this will limit the number of segments included in the sample as plot radius declines. In our study the plot radius is 16 meters, which is typically regarded as a large plot size for forest management inventories.

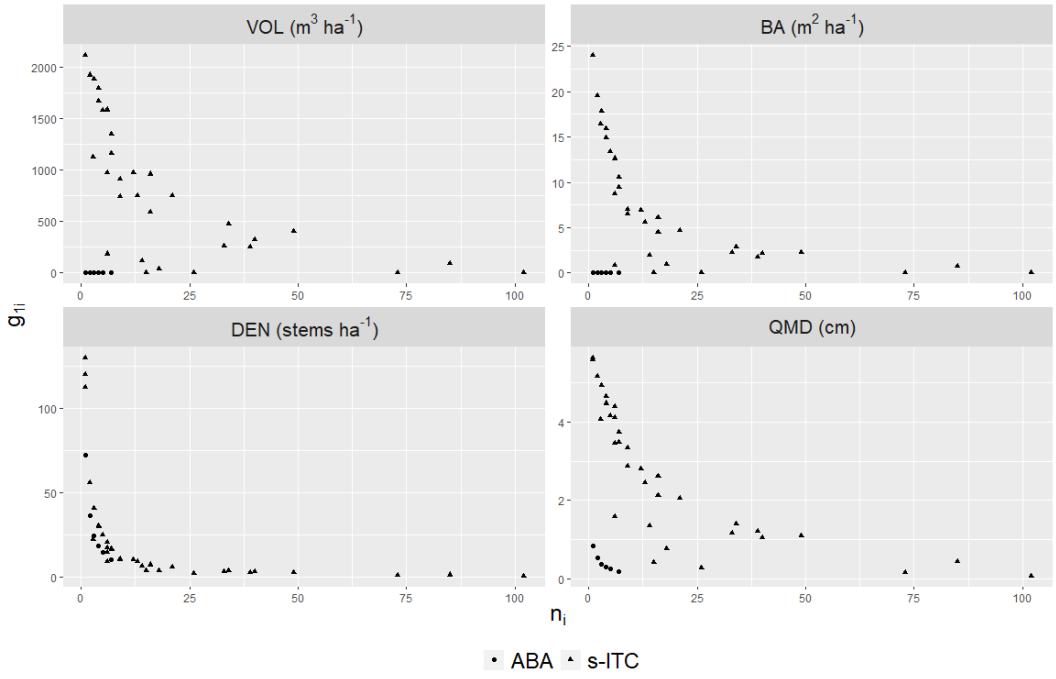

Figure 3.7: The error component g_{1i} for all sampled stands in the study area plotted against the stand-specific sample size n_i.

The error component g_{2i} can be considered a measure of uncertainty that incorporates the difference between an aggregation of all covariates that exist in the stand, $\boldsymbol{l}_i$, and an aggregation of the covariates of the sampled units in the stand used to fit

the model weighted by $\hat{\gamma}_i$, and the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$. Similarly to g_{1i}, g_{2i} can be re-written for the special case of stand-level means:

$$g_{2i} = (\boldsymbol{l}_i - \hat{\gamma}_i \boldsymbol{x}_{ia})^T \widehat{Var}(\widehat{\boldsymbol{\beta}})(\boldsymbol{l}_i - \hat{\gamma}_i \boldsymbol{x}_{ia}) \,, \tag{3.22}$$

where $\boldsymbol{x}_{ia} = \sum_{j=1}^{n_i} \frac{\boldsymbol{x}_{ij}}{k_{ij}^2 a_i}$ is a mean vector of sampled predictors weighted by the variance predictors k_{ij}^2. For a fixed $\hat{\gamma}_i$, the outer terms, $(\boldsymbol{l}_i - \hat{\gamma}_i \boldsymbol{x}_{ia})$, will be large when there is a high degree of dissimilarity between the covariates in the sample and the covariates in the population. High levels of g_{2i} can therefore indicate a higher level of extrapolation required to generate a prediction for a given stand. Importantly, the difference terms can be mitigated by controlling $\hat{\gamma}_i$ via an increase in n_i. Additionally, a more precise estimate of $\boldsymbol{\beta}$ will reduce g_{2i} globally for all predictions, synthetic or otherwise.

Figure 3.8 displays the error component g_{2i} for all stands in the study area. The color of each point represents the sample size, n_i, and it is clear that the overall error component can be drastically reduced in some cases with larger sample sizes. This effect is more evident for the s-ITC models. For the VOL, BA and QMD, a stark difference is apparent between s-ITC and ABA models at the tails of the stand-level predicted values. ABA models tend to have higher values of g_{2i} in these stands which suggest a higher degree of extrapolation, this effect is most extreme for large VOL predictions, while the effect is most extreme for small QMD predictions. Notably, we observed the highest levels of disagreement for the prediction of stand-level target parameters for these variables in these stands, specifically for VOL (Figure 3.4), which suggest that ABA and s-ITC models can disagree when large degrees of extrapolation are required by either or

both models. For s-ITC models a wider range of predictor values and a larger sample size are available in the model fitting stage, which can lead to a reduction in the distance terms in equation (3.22) and a more precise estimate of $\boldsymbol{\beta}$. Such an effect is evident in the QMD predictions, where the s-ITC predictions exhibit a much lower and more stable g_{2i} across the prediction range.

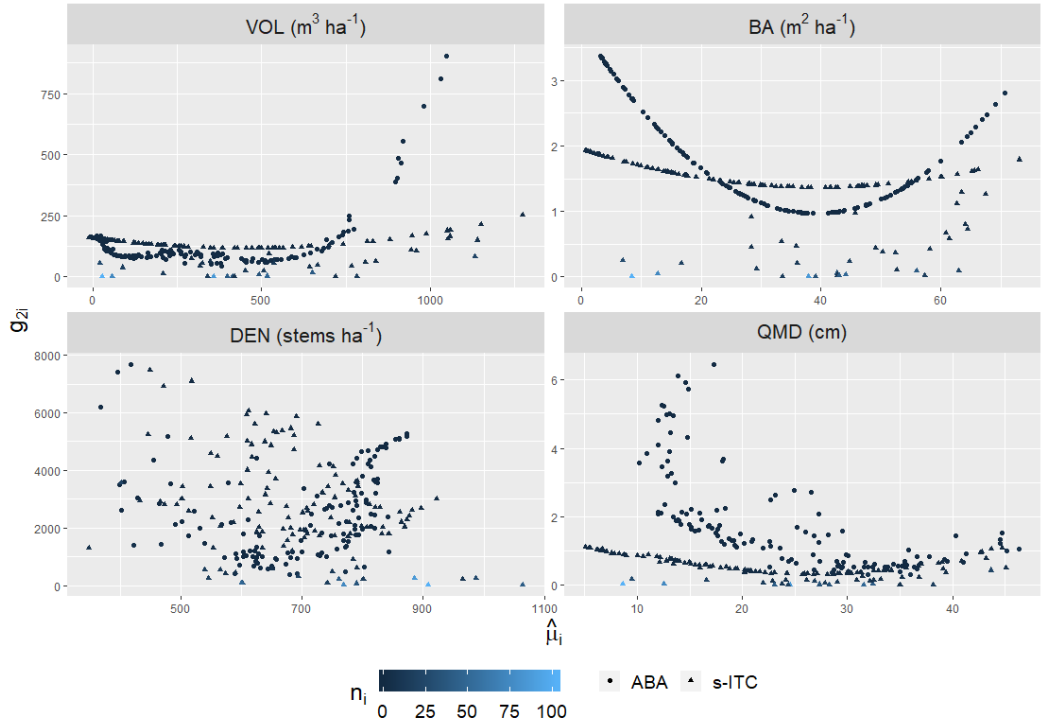

Figure 3.8: The error component g_{2i} for all stands in the study area plotted against the stand-level prediction $\hat{\mu}_i$.

3.5.3 Peculiarities of a Segment Population

A number of features of the segment population distinguish it from the population of regular grid cells used in the area-based approach. Most apparent is that the average size of segments are much smaller than that of grid cells (Figure 3.2). This change in spatial extent may make the segment population more prone to measurement error of observed variables within each population unit. To calculate forest attributes of a segment requires

knowledge of the position of each tree within the larger fixed-area plot. As segment size declines, the probability of mis-assigning a tree to one segment increases if measurement error of tree positions is present. In our study, we treated the tree positions as known, but further research could propagate this additional source of error into estimates of $\widehat{RMSE}_i$. For operational applications, it should be stressed that the required measurement of tree positions presents an increase in the cost of field data collection campaigns, which is not strictly required for the ABA.

Sampling a segment population via fixed radius plots implies a spatial structure of sample observations that are tightly clustered, which presents an opportunity to estimate potential spatial correlation parameters. In an early phase of the analysis, we did not observe strong patterns in empirical semi-variograms for the s-ITC or the ABA that indicated any spatial correlation present in the sample. While many pairwise distances are available at close ranges for s-ITC, e.g. those pairs that exist in the same fixed radius plot, relatively few exist outside of this range. A lack of pairwise observations at close- and mid-range distances often precludes reliable estimation of spatial correlation parameters (Mauro et al. 2017b). A failure to observe spatial correlation patterns in the available data should not be conflated with a lack of its presence in the actual state, and the decision to ignore a possible spatial correlation can lead to a risk of under-estimating the mean squared error of stand-level predictions, as discussed for the ABA in Magnussen et al. (2017). Assessing the impact of ignoring a potential spatial correlation requires specialized datasets not typically found in forest inventories (e.g. Mauro et al. 2017b) or simulation studies (e.g. Magnussen et al. 2017), and we leave such a question for the case of s-ITC for further research.

When constructing target parameters for both ABA and s-ITC models, a geometric mismatch occurs (e.g. Figure 3.2) between stands as they are delineated, as they are constructed from a set of cells assigned to that stand and as they are constructed from a set of segments assigned to that stand. We observed mean absolute deviations of 0.12 (0.7%) hectares and 0.3 (1.7%) hectares between segment and cell populations respectively as they differ from the stand boundaries and expect the impact on stand-level and study region predictions to be minor. Such a disagreement is not accounted for in the presented mean squared error estimators, but may lead to discrepancies in the estimate of stand-level means. For the s-ITC the impact of this geometric mismatch will be smaller than that of the ABA given the observed mean absolute deviations, which is one potential advantage of the method. If stand-level predictions are desired for the area-as-delineated, setting $h_{i.}$ to the delineated area during the construction of $\boldsymbol{l}_i$ is one potential alternative, which is equivalent to rescaling the mean prediction to the desired area. Further corrections, such as splitting population units by stand boundaries, tightly related to resolution dependence (e.g. Packalen et al. 2019), may introduce bias for unit-level predictions and beget further research before serious consideration can be given.

3.5.4 Implications for Forest Management Inventories

As lidar assisted forest inventories continue to support forest management decisions, explicit consideration of the uncertainty at the resolution where decisions are made is needed. For many forest management organizations, this decision making process is done at the stand level (see Kangas et al. 2015 pp. 13-14). At the same time, the interest and technology around tree segmentation has seen marked increase in recent years (Wang et

al. 2016; Zhen et al. 2016), yet studies that examine the appropriateness of segmentation methods to support stand-level decision making are rare.

We have shown that, for this particular data set, models that use detected tree segments as a population exerted higher random effect variances than their ABA counterparts. For synthetic predictions of stand-level means of forest inventory attributes, the importance of this difference in random effect variance is paramount, and implies that the ABA will provide synthetic predictions of stand level means with lower $\widehat{RMSE}_i$ than the s-ITC counterparts. In sampled stands, however, s-ITC models were able to outperform ABA models in some cases. We identified two situations when this may occur: when segment induced replication was strong and when ABA were required to extrapolate to a higher degree than that of s-ITC models. For the first situation, segment induced replication is expected to be large in younger stands as a feature of the segmentation method. For the second situation, ABA models demonstrated extrapolation issues for stand with large VOL, small QMD and the large and small extremes of BA predictions. These two situations are niches where s-ITC models may be able to support reductions in $\widehat{RMSE}_\iota$ relative to their ABA counterparts.

S-ITC also comes with a number of less tangible benefits, such as an increase in the spatial resolution of unit-level predictions, increased interpretability of population units and the ability to predict unit-level attributes such as maximum diameter or dominant species, for elements whose size is closer to actual trees. Importantly, the s-ITC method requires a large computational effort to generate canopy height models and delineate segments for further analysis. Such an undertaking may be infeasible for large areas and, combined with the required measurement of tree positions, s-ITC is inherently

a more complex approach to forest management inventory than the ABA. However, given that we noted reduction in $\widehat{RMSE}_i$ for only those stands that are sampled, and population units are treated as independent if they are located in different stands, segmentation can be done only for those stands that are sampled, greatly reducing the computational effort for large areas if augmentation of area-level parameter estimation via s-ITC in sampled areas is desired. In this way, we see the s-ITC method as an additional tool for forest management inventories, rather than a pure replacement of an already existing ABA system if the primary objective is to produce minimum $\widehat{RMSE}_i$ for all stands in a study area.

3.6 Conclusion

We assessed small area unit-level models produced using the semi-ITC method for the prediction of area-level means of four forest inventory attributes and compared them to models produced using the area-based approach. For the case of synthetic predictions, i.e. predictions of totals per unit area for unsampled stands, the random effect variance limits the s-ITC relative to models produced using the area-based approach. For the case of sampled stands, we observed a reduction in $\widehat{RMSE}_i$ in stands where increased replication occurred for all four forest inventory variables. This increased replication results from a segmentation method that explicitly considers a predicted crown area in the segmentation process, and therefore begets a large increase in the number of sampled population units.

3.7 Acknowledgements

We would like to thank the Bureau of Land Management for data acquisition and Dr. Jim Flewelling for assistance in its use. This study was supported by the grant: Predicting Thinning Eligibility with Airborne LiDAR Data (Bureau of Land Management).

4 Manuscript III

Predicting Forest Inventory Variables with Bivariate Fay-Herriot Models in Blacks Mountain Experimental Forest, California, USA

Bryce Frank

Francisco Mauro

Temesgen Hailemariam

Martin Ritchie

Abstract

Recently, Fay-Herriot (FH) models have been investigated in forest inventory contexts for providing predictions of forest attributes at the resolution of forest management stands, especially in cases where accurate geopositioning of field plots is not available. FH models can be extended to consider multiple variables simultaneously, and have been shown to reduce model-based mean squared errors in some contexts. The objective of this study was to assess the performance of bivariate FH models for predicting standing timber volume (VOL), basal area (BA), stem density (DEN), mean height (HT) and mean diameter (DIA) against their univariate counterparts for the Blacks Mountain Experimental Forest in northern California, USA. Performance was assessed using model-based mean squared errors for the entire study area and for individual forest management stands. For the entire study area assessment, at least one bivariate model exists that reduced mean-squared errors by 15%, 5%, 19%, 9% and 39% for VOL, BA, DEN, HT and DIA respectively for large area estimates. For stand-level assessments, mean-squared errors were also reduced by at least one bivariate model for each attribute. Bivariate models composed of pairs of direct estimators that tended to be strongly correlated were those that resulted in the greatest reductions of mean-squared error at both scales.

4.1 Introduction

In many cases, forest management planning revolves around point estimates and measures of uncertainty at the scale of forest stands. Small area estimation (SAE) is a methodology that explicitly considers the uncertainty assessment of predictions of subpopulations of interest, which can be adapted for use in the context of forest stands. Many existing forest inventory systems report measures of error at this scale, e.g. via stand exams that rely on probability-based sampling designs (Finley et al. 2014). In recent years SAE methods that rely on parametric linking models have seen increased use as a model-based alternative that reports measures of error similar to those attained in design-based stand exams, while leveraging the information provided with remote sensing data such as aerial lidar data (Goerndt et al. 2011; Breidenbach and Astrup 2012; Mauro et al. 2016).

In many of these studies, a particular SAE method referred to as the unit-level model is used as the vehicle for generating forest inventory predictions. Unit-level models are somewhat synonymous with models developed using the forest inventory-specific term " the area-based approach" (ABA, e.g. Næsset 2014) as they rely on auxiliary information that can be assigned to particular population units (e.g. field plots or grid cells). In forest inventory contexts, this implies that geopositioning information for field plots is available and of adequate precision to match spatially explicit auxiliary data to field plots. In cases where a probability-based design is available, but field plot positions are not available or are of inadequate precision, Fay-Herriot (FH) models offer opportunities to improve forest inventory predictions and rely only on the knowledge of forest stand membership rather than precise geopositioning to relate field data to auxiliary

information. In FH models, the direct estimate from a field-based survey of a stand is used as the response variable, and predictors are derived at the area level, enabling inference at the scale of forest stands.

FH models have been investigated for use in forest inventories in past studies. The primary distinction between FH models and unit-level models is that the smallest scale at which predictions can be made is at the scale of areas, or domains, such as forest management stands or strata, whereas unit-level models can make predictions for individual population units, such as grid cells or tree segments The auxiliary data used to make predictions with FH models are at the scale of domains, and thus it is generally accepted that some loss of precision is inevitable when adopting the FH model over the unit-level model (Mauro et al. 2017; Breidenbach et al. 2018). However, better precision is attained than that of design-based estimators that do not incorporate the use of auxiliary information, especially when area-specific sample sizes are small (Mauro et al. 2017a; Ver Planck et al. 2018). Furthermore, FH models provide the ability to generate synthetic predictions, i.e. predictions for unsampled stands, that are based on a linking model developed from the sample data, which provide the derivation of optimal estimators and area-specific measures of error under the assumed model (Rao and Molina 2015 p. 5).

To date, only univariate Fay-Herriot (UVFH) models have been considered in the context of forest inventory. Recently, however, FH models and estimation of their components have been extended to consider multiple variables, known as the multivariate Fay-Herriot model (MVFH) (González-Manteiga et al. 2008b; Benavent and Morales 2016). The MVFH is able to take advantage of the correlations that exist among attributes of interest to increase the precision of model-based predictions. In this study we

considered the special case of the MVFH model that contain two response variables only, i.e. the bivariate FH (BVFH) model. BVFH models have been shown in other contexts to reduce model-based mean squared error estimates of subpopulations (Benavent and Morales 2016; Ubaidillah et al. 2019). However, these studies typically have access to large domain-specific sample sizes and do not consider the case of synthetic prediction. In many forest inventories, domain-specific sample sizes may be small and the number of stands where synthetic prediction is required may be large (e.g. Mauro et al. 2017).

While BVFH models in particular have not been investigated explicitly in the context of forest inventories, some studies that consider multivariate models in the context of model-based estimation for forest inventories have been conducted. Recently, Hou et al. (2019) presented the use of the best linear unbiased predictor (BLUP) as a means for simultaneously predicting large area target parameters, with a specific focus on young forest stands, and noted a 55-76% increase in precision for broadleaf and conifer density and broadleaf and conifer mean height. Mauro et al. (2019b), a study primarily focused on estimating change in forest attributes with repeat lidar acquisitions and field measurements, investigated the use of a multivariate unit-level model for use in change estimation compared to a univariate model that directly considered change and found that the multivariate model was less prone to extrapolation risk. Babcock et al. (2012) assessed the use of multivariate models in a Bayesian setting for predicting various tree-level attributes in conjunction with lidar covariates and found a reduction in RMSPE for tree height, crown ratio and diameter at breast height compared to analogous univariate models.

The objective of this study is to assess the performance of UVFH and BVFH models for all pairs among the attributes VOL, BA, DEN, HT and DIA. Prediction performance will be assessed at the scale of the entire study area, i.e. a large area assessment, and at the scale of individual forest stands, i.e. a small area assessment using model-based mean squared error estimators. For small areas, we consider both the case where field samples are available and where field samples are unavailable (i.e. synthetic prediction). Models will be developed using an aerial lidar acquisition and a set of 164 fixed radius field plots across 24 sampled stands. We also investigate the impact of the correlation of direct estimators on model performance.

4.2 Materials

4.2.1 Study Area

This study was conducted in the Blacks Mountain Experimental Forest (BMEF), a 3,715 ha forested area managed by the United States Forest Service located in northeastern California, USA. The Köppen climate type of the area is classified as a Mediterranean climate. The elevation of the BMEF ranges from 1700 m to 2100 m with gentle to moderate slopes. The forest is dominated by Ponderosa pine (*Pinus ponderosa*), while Incense cedar (*Calocedrus decurrens*), white fir (*Abies concolor*) and Jeffrey pine (*Pinus jeffreyi*) maintain a smaller but consistent presence. The forest structure for this area consists of many large gaps between tree canopies, resulting in a heterogenous spatial pattern and highly variable vertical and horizontal structures.

4.2.2 Sampling Design

The BMEF is delineated into 106 forest stands, of which 24 were selected and sampled in the field as part of a larger study. The field sampling campaign for this larger study began

in 2009 with the establishment of permanent sample plots. A regular grid was cast over the study area, with nodes placed 100 meters apart. Grid nodes were established using survey grade GPS observations and manual traverse methods and are reported to have an accuracy of 15 cm or better (Oliver 2000). This grid served as the basis for establishing sample plot locations within each of the 24 sampled stands. Within a given stand selected for sampling, a grid node was selected at random. Then, sample plot locations were established by traversing in diagonal trajectories from this location. The resulting sampling grid is therefore 282 m x 282 m (Mauro et al. 2019b) within each stand. Within each stand the sampling design is therefore systematic random sampling. Figure 4.1 displays the BMEF, with selected field plots indicated in orange. At each sampled position, a permanent 16 m fixed radius plot was established, measured in the summer of 2016. Within each plot, all live trees larger than 9 cm and all dead trees larger than 12 cm were measured for DBH and height and had their species recorded. Volumes for all trees were predicted using the National Volume Estimation Library (NVEL) and are considered the true volumes for the purposes of this study. Observations for VOL and BA were calculated from the field data by summing the individual tree-level observations and dividing by the field plot size in hectares. Observations for DIA and HT were the means of the individual tree-level observations on each plot. Observations for DEN were the number of trees observed divided by the field plot size in hectares.

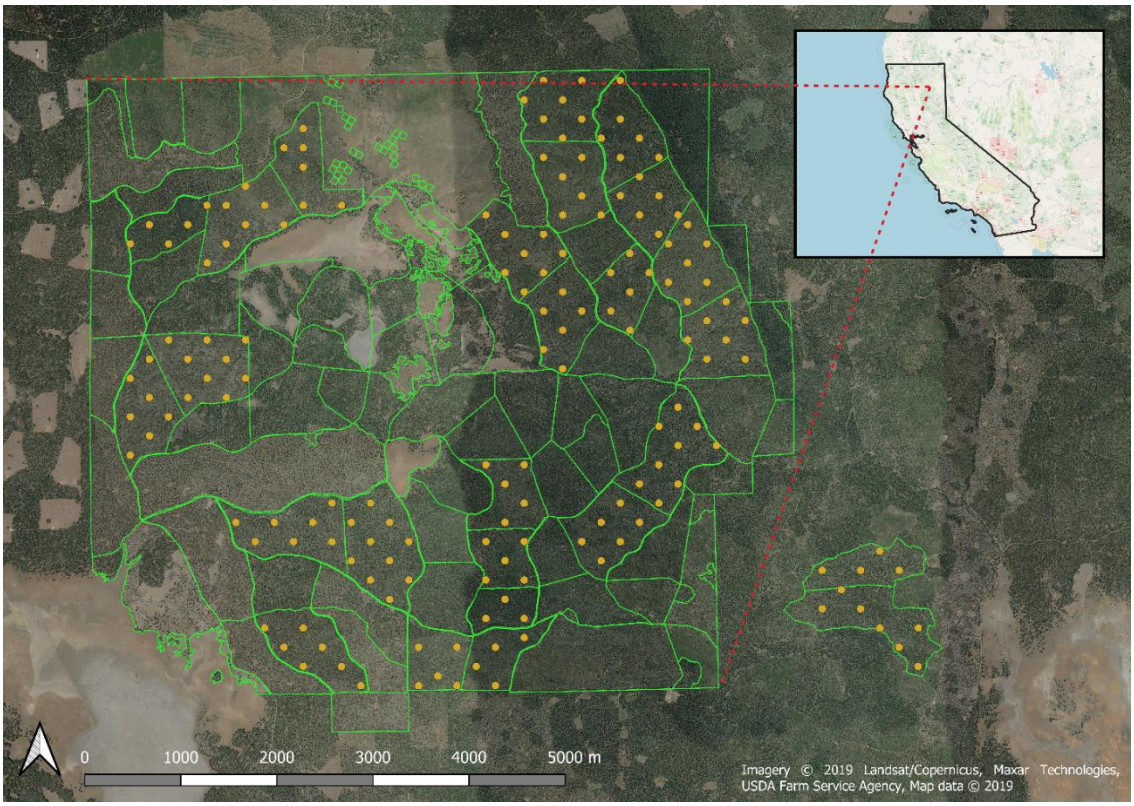

Figure 4.1: Blacks Mountain Experimental Forest (BMEF) displayed along with the 106 forest stand delineations indicated in green, of which 26 are sampled with fixed radius plots indicated in orange. Note that the small square delineations in the northern section are areas excluded from the study area and do not represent forest stand delineations.

4.2.3 Lidar Acquisition and Processing

An aerial lidar dataset was acquired during the summer of 2015 using a Leica ALS 50 sensor at a flying altitude of 900 m and a scanning angle of +/- 14°. The resulting point density was 14.5 ppm^2. The raw lidar acquisition was normalized for terrain elevation by first filtering the ground and non-ground points via the algorithm detailed in Zhang et al. (2003) and interpolating the ground points with a nearest neighbor interpolation to produce a bare earth model. The elevation of the bare earth model beneath each point was subtracted to produce a normalized point cloud of the study area.

A set of 15 lidar height metrics were computed by casting a grid with resolution of 28.35 meters over the entire study area and computing metrics for each grid cell from

the normalized point cloud. A list of metrics and their descriptions are included in Table A 9. For each stand, the grid cell values within the stand boundary, determined by the grid cell centroid, were summarized by calculating their means and variances for a total of 30 means and variances for each stand. This set of means and variances are used as a set of stand-level predictors.

In addition to the summarized height metrics, we implemented a variable window local maxima (VWLM) tree detection algorithm to provide an additional auxiliary data source (Popescu et al. 2002). The variable window local maxima tree detection algorithm operates on a canopy height model derived from the normalized point cloud described above. We produced an intermediate canopy height model with a resolution of 0.33 m. Following this, we used a median filter with a 3x3 kernel to smooth the intermediate canopy height model to produce a final canopy height model, upon which the tree detection algorithm was used. The VLWM first operates by detecting maxima using a fixed window local maxima algorithm then, for each detected maxima from this set of candidate maxima, it determines the maxima that has the highest elevation among all candidates in a variable sized window using the equation:

$$w(\boldsymbol{s}) = a + b \cdot z(\boldsymbol{s})^2 \tag{4.2}$$

where $w(\boldsymbol{s})$ is the window width at position $\boldsymbol{s}$, $z(\boldsymbol{s})$ is the height of the canopy height model at position $\boldsymbol{s}$ and a and b are coefficients. We set the values for a and b using visual inspection of the detected tree positions and the underlying canopy height model to values of 2.1 and 0.007 respectively. For each stand the number of detected tree positions was divided by the stand area to create the predicted number of trees per hectare, *tpha_pred*.

4.3 Methods

4.3.1 Multivariate Fay-Herriot Model

The multivariate area-level parameter is a vector of realized values for a set of R forest attributes. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iR})^T$ be the multivariate parameter for a forest stand indexed by i. The multivariate Fay-Herriot model (MVFH) involves a vector of direct estimators, $\widehat{\boldsymbol{\theta}}_i = \left(\hat{\theta}_{i1}, \dots, \hat{\theta}_{iR}\right)^T$, for $\boldsymbol{\theta}_i$, typically attained via a probability-based sampling design. The vector of direct estimators can be decomposed into the realized area-level parameter $\boldsymbol{\theta}_i$ and a vector of errors:

$$\widehat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{e}_i \tag{4.2}$$

where $\boldsymbol{e}_i$ represents the deviance that is a result of the uncertainty of the direct estimator. When referring to the jth attribute of the area-level parameter vector, we reserve the notation u_j.

Direct estimators for five forest inventory variables, VOL ($m^3 ha^{-1}$), BA ($m^2 ha^{-1}$), DEN ($stems\ ha^{-1}$), DIA (cm) and HT (m) were produced in each forest stand in the BMEF. For a given stand, let n_i represent the number of field plots. Given that the sampling design within each area is systematic (section 2.2) direct estimates for $\boldsymbol{\theta}_i$ can be produced using the direct estimator

$$\hat{\theta}_{ir} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ijr}, \tag{4.3}$$

i.e. the sample mean (Thompson 2012, p. 159), where y_{ijr} represents the jth observation of the rth attribute in area i. The elements of the variance-covariance matrix $Var\left(\widehat{\boldsymbol{\theta}}_i\right) = \boldsymbol{V}_{ei}$ are estimated via

$$\widehat{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{ik}) = \frac{1}{N_i^2} \sum_{l=1}^{n_i} w_{il}(w_{il} - 1)(y_{ilj} - \hat{\theta}_{ij})(y_{ilk} - \hat{\theta}_{ik}), \tag{4.4}$$

where $w_{il} = \frac{1}{\pi_{il}}$ is the inverse of the inclusion probability of the lth population unit in the ith stand (Särndal et al. 2003 p. 170; Benavent and Morales 2016). (3.4) is derived assuming simple random sampling with replacement was conducted. Although systematic random sampling was conducted within each area, it is not possible to derive an unbiased estimator for the sample covariance or sample variance in the case of a single random start (Särndal et al. 2003 pp. 74-75). However, the simple random sampling estimator presented in (3) is a naïve estimator frequently used in this scenario (Thompson 2012 pp. 162-163), and the resultant bias can be expected to be small if spatial autocorrelation is weak (Babcock et al. 2018) and, additionally, the bias is positive, resulting in conservative estimates of estimator variance (Wolter 1984). Applying (3.4) to all pairs of direct estimators in area i yields the estimated variance covariance matrix $\widehat{\boldsymbol{V}}_{ei}$. In addition to the estimated variance-covariance matrix, we also produced the estimated the pairwise correlations, with elements defined by:

$$\widehat{Corr}(\hat{\theta}_{ij}, \hat{\theta}_{ik}) = \frac{\widehat{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{ik})}{\sqrt{\widehat{Var}(\hat{\theta}_{ij})\widehat{Var}(\hat{\theta}_{ik})}}, \tag{4.5}$$

Once the direct estimators and their variance-covariance matrices have been established, a linking model is imposed on $\boldsymbol{\theta}_i$ via the multivariate linear regression model:

$$\boldsymbol{\theta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{v_i} \tag{4.6}$$

where $\boldsymbol{X}_i$ is a block-diagonal matrix of predictors, and each block, $\boldsymbol{x}_{ir}^T$, is a vector of dimension p_r and

$$\boldsymbol{X}_i = \underset{1 \le r \le R}{diag}(\boldsymbol{x}_{ir}^T). \tag{4.7}$$

This allows for different numbers of predictors for each attribute indexed by the subscript r. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_R^T)^T$ is a column vector of regression coefficients, composed of sub-vectors each of dimension p_r. Finally, $\boldsymbol{v}_i = (v_{i1}, \dots, v_{iR})^T$ is a vector of random effects $\sim_{ind} MVN(\boldsymbol{0}, \boldsymbol{V}_{vi})$ for $i = 1, \dots, D$. The variance-covariance matrices $\boldsymbol{V}_{vi}$ are assumed diagonal, composed of variable-specific random-effect variance parameters σ_{vr}^2 such that $\boldsymbol{V}_{vi} = \boldsymbol{V}_{vj}\ \forall\ (i, j)$ and $\boldsymbol{V}_{vi} = \underset{1 \le r \le R}{diag}(\sigma_{vr}^2)$.

The MVFH model can be constructed for the entire set of areas:

$$\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e} \tag{4.8}$$

where

$$\boldsymbol{\theta} = \underset{1 \le i \le D}{\text{col}}(\boldsymbol{\theta}_i),\ \boldsymbol{X} = \underset{1 \le i \le D}{\text{col}}(\boldsymbol{X}_i),\quad \boldsymbol{v} = \underset{1 \le i \le D}{\text{col}}(\boldsymbol{v}_i),\quad \boldsymbol{e} = \underset{1 \le i \le D}{\text{col}}(\boldsymbol{e}_i)$$

and $\boldsymbol{Z}$ is a $D\ x\ D$ matrix that assigns the observations to their respective areas. In the case of the FH model, $\boldsymbol{Z} = \boldsymbol{I}$. Such a formulation specifies the variance-covariance matrices for $\boldsymbol{v}$ and $\boldsymbol{e}$:

$$\boldsymbol{V}_v = \underset{1 \le i \le D}{diag}(\boldsymbol{V}_{vi}),\quad \boldsymbol{V}_e = \underset{1 \le i \le D}{diag}(\boldsymbol{V}_{ei})$$

where the operator col stacks matrices vertically and the operator $diag$ constructs a block-diagonal matrix using all sub-matrices indicated in the expression.

4.3.2 Target Parameters

The MVFH, defined in (4.6), affords derivation of the empirical best linear unbiased predictor (EBLUP) for target parameters that can be expressed as linear transformations of the model coefficients $\boldsymbol{\beta}$ and the random effects $\boldsymbol{v}$. Consider a generic target parameter indexed by α:

$$\boldsymbol{\theta}_\alpha = \boldsymbol{L}_\alpha \boldsymbol{\beta} + \boldsymbol{M}_\alpha \boldsymbol{v} \tag{4.9}$$

where $\boldsymbol{L}_\alpha$ and $\boldsymbol{M}_\alpha$ are matrices of dimension $(R\ x\ p)$ and $(R\ x\ RD)$ respectively. The exact formulations of $\boldsymbol{L}_\alpha$ and $\boldsymbol{M}_\alpha$ will depend on the structure of the target parameter. We consider two cases that are of common interest in forest inventories, that of area-specific means, $\boldsymbol{\theta}_i$, and aggregations of all area-specific means to produce a large-area parameter $\boldsymbol{\theta}_\tau$ for the entire BMEF study region. For the first case we obtain:

$$\boldsymbol{L}_i = \boldsymbol{X}_i \tag{4.10}$$

$$\boldsymbol{M}_i = \underset{1 \le r \le R}{\text{col}}(\boldsymbol{k}_r^T)^T \tag{4.11}$$

where $\boldsymbol{k}_r$ is a vector of dimension RD where the element at position $R(i-1)+r$ is 1 and 0 otherwise such that $\boldsymbol{M}_i^T \boldsymbol{v} = \boldsymbol{v}_i$. For the second case we obtain:

$$\boldsymbol{L}_\tau = \frac{1}{h_.} \sum_{i=1}^{D} h_i \boldsymbol{X}_i \tag{4.12}$$

$$\boldsymbol{M}_\tau = \underset{1 \le i \le D}{\text{col}} \left(\frac{h_i}{h_.} \boldsymbol{I} \right)^T \tag{4.13}$$

where h_i is the area of stand i and $h_\cdot = \sum_{i=1}^{D} h_i$. The terms $\boldsymbol{L}_\tau$ and $\boldsymbol{M}_\tau^T \boldsymbol{v}$ are interpretable as the means of the stand-specific design matrices and stand-specific realized random effects weighted by stand area.

Prediction for target parameters in both cases is attained via the EBLUP:

$$\widehat{\boldsymbol{\theta}}_\alpha^E = \boldsymbol{L}_{\boldsymbol{\alpha}} \widehat{\boldsymbol{\beta}} + \boldsymbol{M}_{\boldsymbol{\alpha}} \widehat{\boldsymbol{v}} \tag{4.14}$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^T \boldsymbol{y}_S \tag{4.15}$$

and

$$\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{V}}_{vS} \boldsymbol{Z}_S^T \widehat{\boldsymbol{V}}_S^{-1} (\boldsymbol{y}_S - \boldsymbol{X}_S \widehat{\boldsymbol{\beta}}) \tag{4.16}$$

are the predicted random effects. We reserve the subscript S to refer to a matrix that has rows corresponding to non-sampled stands removed and the notation D_S to refer to the number of sampled stands. For stands that do not have samples we obtain the synthetic predictor

$$\widehat{\boldsymbol{\theta}}_\alpha^{Syn} = \boldsymbol{L}_{\boldsymbol{\alpha}} \widehat{\boldsymbol{\beta}}. \tag{4.17}$$

Note that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{v}}$ depend only on the unknown variance components $\boldsymbol{\delta} = (\sigma_{v1}^2, \ldots, \sigma_{vR}^2)^T$. Estimation of these components can be attained via restricted maximum likelihood (REML). We employ the methodology developed in Benavent and Morales (2016 pp. 374-5) to attain $\widehat{\boldsymbol{\delta}}$ via REML using the R programming language.

4.3.3 Mean Squared Error Estimators

The mean squared error estimator of the EBLUP quantifies the uncertainty of target parameter predictions and is analogous to standard errors employed in design-based

forest exams provided model assumptions hold. Note that (4.6) is a general linear mixed model with block-diagonal covariance structure. MSEs for EBLUPs constructed from this model (i.e. (4.7)) using REML estimation of variance components are derived in Rao and Molina (2015 pp. 108-111) and Benavent and Morales (2016 pp. 375-7).

The mean squared error estimator of the EBLUP for the rth attribute in the ith area can be expressed as the rth diagonal element of the sum of three matrices:

$$\widehat{MSE}(\hat{\theta}_{\alpha r}^{E}) = [\boldsymbol{G}_{1\alpha}(\widehat{\boldsymbol{\delta}}) + \boldsymbol{G}_{2\alpha}(\widehat{\boldsymbol{\delta}}) + 2\boldsymbol{G}_{3\alpha}(\widehat{\boldsymbol{\delta}})]_{r,r} \tag{4.18}$$

where

$$\boldsymbol{G}_{1\alpha}(\widehat{\boldsymbol{\delta}}) = \boldsymbol{M}_{\alpha}(\widehat{\boldsymbol{V}}_{v} - \widehat{\boldsymbol{V}}_{v}\boldsymbol{Z}_{S}^{T}\widehat{\boldsymbol{V}}_{S}^{-1}\boldsymbol{Z}_{S}\widehat{\boldsymbol{V}}_{v})\boldsymbol{M}_{\alpha}^{T}, \tag{4.19}$$

$$\boldsymbol{G}_{2\alpha}(\widehat{\boldsymbol{\delta}}) = (\boldsymbol{L}_{\alpha} - \boldsymbol{M}_{\alpha}\boldsymbol{V}_{v}\boldsymbol{Z}_{S}^{T}\boldsymbol{V}_{S}\boldsymbol{X}_{S})^{T}(\boldsymbol{X}_{S}^{T}\boldsymbol{V}_{S}^{-1}\boldsymbol{X}_{S})(\boldsymbol{L}_{\alpha} - \boldsymbol{M}_{\alpha}\boldsymbol{V}_{v}\boldsymbol{Z}_{S}^{T}\boldsymbol{V}_{S}\boldsymbol{X}_{S}), \tag{4.20}$$

$$\boldsymbol{G}_{3\alpha}(\widehat{\boldsymbol{\delta}}) = \boldsymbol{M}_{\alpha}\left[\sum_{k=1}^{m}\sum_{l=1}^{m} Cov(\hat{\delta}_{k}, \hat{\delta}_{l})\boldsymbol{C}_{\alpha}^{k}\widehat{\boldsymbol{V}}_{\alpha}\boldsymbol{C}_{\alpha}^{l}\right]\boldsymbol{M}_{\alpha}^{T}. \tag{4.21}$$

where $Cov(\hat{\delta}_{k}, \hat{\delta}_{l})$ is the klth element of the Fisher information matrix and $\boldsymbol{C}_{i}^{k}$ is the partial derivative of $\boldsymbol{V}_{i}$ with respect to the kth element of $\boldsymbol{\delta}$ (proof provided in section A.3.3). We reserve the notation $g_{k\alpha r} = [\boldsymbol{G}_{k\alpha}]_{r,r}$ to refer to the kth error component for the rth attribute. For simplicity and conciseness, we exclusively refer to only the first attribute such that $g_{k\alpha 1} = g_{k\alpha}$. These error components can be interpreted as various sources of error: $g_{1\alpha}$ captures the uncertainty caused by random effect variance, and typically declines with an increase in sample size for a given area, $g_{2\alpha}$ captures the uncertainty caused by the estimate of the fixed-effects $\boldsymbol{\beta}$ and $g_{3\alpha}$ captures the uncertainty of the estimate of the variance components $\boldsymbol{\delta}$. For the synthetic predictor, the mean squared cross product error is defined as

$$MSE\left(\hat{\theta}_{ir}^{Syn}\right) = E\left[\left(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_i\right)\left(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_i\right)^T\right] \tag{4.22}$$

which only considers the fixed-effects part of the model. This leads to a reduced version of (17):

$$\begin{aligned}\widehat{MSE}\left(\hat{\theta}_{ir}^{Syn}\right) &= \left[\boldsymbol{G}_{1i}^{Syn}\left(\widehat{\boldsymbol{\delta}}\right) + \boldsymbol{G}_{2i}^{Syn}\left(\widehat{\boldsymbol{\delta}}\right)\right]_{r,r} \\ &= \left[\widehat{\boldsymbol{V}}_{vi} + \boldsymbol{X}_i^T\left(\sum_{j=1}^{D}\boldsymbol{X}_j^T\widehat{\boldsymbol{V}}_j\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_i\right]_{r,r}.\end{aligned} \tag{4.23}$$

4.3.4 Model Selection

The selection of predictors for the MVFH affords a vast number of potential covariates to be included in the model. For example, it is possible for the same predictor to be included if that predictor corresponds to different attributes in the response vector. To simplify the model selection process, we chose to conduct model selection for the univariate Fay-Herriot (UVFH) for each of the five forest attributes. Such an approach is attractive for forest inventories where univariate model forms have already been established or reliance on model selection from past literature is defensible.

To select UVFH models for the selected forest attributes we employed a two-phase model selection procedure. Let the index $\xi = \{VOL, BA, DEN, HT, DIA\}$ represent the forest attribute for which the UVFH model selection is being conducted. In the first phase a set of 25 multiple linear regression models were obtained for each attribute by an exhaustive search implemented by the leaps package in R (Miller 2017). For each of one through four predictor variables plus an intercept, the 5 best models were found via adjusted R^2 and compose the set $mod_{\xi 1}$. Models in this set were plotted by their number of predictors and adjusted R^2 values. A number of candidate models were explored around the region where adjusted R^2 began to level off relative to the number of predictors. Models in this region that appeared to satisfy linearity assumptions moved to

the second phase and compose the set $mod_{\xi 2}$. The remaining models were inspected graphically to assess the assumption of constant variance and normality in the random effects. Of those that satisfied this condition, the one with the highest adjusted R^2 value was chosen as the final model.

4.4 Results

4.4.1 Selected Models

Five UVFH models were selected after following the model selection procedure described in section 4.3.4. The selected models, their predictor variables, coefficient estimates, standard errors and estimates of random effect standard deviation are reported in Table 4.1. For the VOL, HT and BA UVFH models, the selected predictors are the means of predictors typically selected for unit-level models, where high-height predictors for VOL and HT (e.g. *p_90_mean*), and a middle-height predictor for BA (e.g. *mean_z_mean*) are typically selected in linear regression models (Sheridan et al. 2015; Shin et al. 2016). For DEN, HT and DIA models, the estimate of the coefficient for the tree-detection based predictor was significant at $\alpha = 0.05$. The sign of this coefficient is positive with respect to DEN, negative with respect to HT, and negative with respect to DIA. Estimates of random effect standard deviation are 11.71, 1.94, 35.63, 0.61 and 1.28 for VOL, BA, DEN, HT and DIA, models respectively.

Table 4.1: Selected UVFH models for VOL, BA, DEN, HT and DIA, including estimates for regression coefficients, standard errors and $\hat{\sigma}_v$.

Attribute	**Predictor**	**Coefficient**	**Std. Error**	$\hat{\sigma}_v$
VOL $(m^3 ha^{-1})$	Intercept P_90_mean	-77.91 17.98	16.06 1.29	11.71
BA $(m^2 ha^{-1})$	Intercept Mean_z_mean	3.89 5.64	1.93 0.52	1.94
DEN $(stems\ ha^{-1})$	Intercept Tpha_pred P_95_var	-243.43 2.19 3.54	59.08 0.27 1.17	35.63
HT (m)	Intercept Tpha_pred P_90_mean	13.88 -0.02 0.32	0.73 0.002 0.06	0.61
DIA (cm)	Intercept Tpha_pred Pct_r_1_above_mean_mean	33.89 -0.07 40.23	1.85 0.01 10.67	1.28

The five UVFH models are paired with each of the remaining four UVFH models to generate a total of twenty BVFH models using the procedure described in section 3.5. By considering only the parameter estimates of the original UVFH model it is possible to observe changes in the parameter estimates and their standard errors before and after pairing with another UVFH model. Table 4.2 displays the relative change in standard error of the parameter estimates, or the random effect variance for each of the twenty BVFH models. A mixture of results is apparent, and it is not always beneficial to adopt a bivariate pairing. For example, the standard errors and random effect variances markedly increase for VOL-DEN, VOL-DIA, BA-DEN, BA-HT, and BA-DIA. However, for each u_1 at least one pairing results in a decrease in at least some of the parameter variances, such as the VOL-BA, BA-VOL, DEN-BA, HT-DIA and DIA-DEN models.

Table 4.2: Percentage change between UVFH and BVFH models. The text above each table indicates the variable u_1, and the corresponding labels adjacent to each sub-table indicate the variable u_2. Each sub-table is composed of the coefficient for each predictor ascribed to u_1 and its random-effect standard deviation σ_{v1}. The percentage values indicate a change in the standard error of the parameters across models for the regression coefficients and a change in σ_{v1}, $\frac{BV-UV}{BV} \cdot 100$ such that BV and UV represent a generic bivariate or univariate parameter respectively. Cooler colors indicate a decline in the coefficient standard error or σ_{v1} by adopting the BV model over the UV model, warmer colors indicate an increase.

VOL

BA		DEN	
Intercept	-13.2 %	Intercept	2.6 %
p_90_mean	-12.2 %	p_90_mean	1.1 %
σ_{v1}	-26 %	σ_{v1}	11.2 %
Intercept	6.9 %	Intercept	14.6 %
p_90_mean	4.9 %	p_90_mean	11.1 %
σ_{v1}	22.7 %	σ_{v1}	40.3 %
HT		DIA	

BA

VOL		DEN	
Intercept	-3.3 %	Intercept	10.2 %
mean_z_mean	-9 %	mean_z_mean	2.9 %
σ_{v1}	9.8 %	σ_{v1}	36.1 %
Intercept	6.2 %	Intercept	13.6 %
mean_z_mean	1.6 %	mean_z_mean	6.8 %
σ_{v1}	21.4 %	σ_{v1}	37.5 %
HT		DIA	

DEN

VOL		BA	
Intercept	-19.3 %	Intercept	-28.2 %
tphapred	-21.8 %	tphapred	-30.8 %
p_95_var	4.3 %	p_95_var	-5.5 %
σ_{v1}	16.2 %	σ_{v1}	-2.8 %
Intercept	-6.5 %	Intercept	-18 %
tphapred	-7.5 %	tphapred	-18.6 %
p_95_var	-5 %	p_95_var	-18.3 %
σ_{v1}	-0.4 %	σ_{v1}	-16.7 %
HT		DIA	

HT

VOL		BA	
Intercept	-0.9 %	Intercept	1 %
tphapred	-2.9 %	tphapred	-2.5 %
p_90_mean	-5 %	p_90_mean	-4.1 %
σ_{v1}	2.7 %	σ_{v1}	5.2 %
Intercept	4.4 %	Intercept	-10.8 %
tphapred	-0.9 %	tphapred	-7.4 %
p_90_mean	-7 %	p_90_mean	-13.8 %
σ_{v1}	15.1 %	σ_{v1}	-7.2 %
DEN		DIA	

DIA

VOL		BA	
Intercept	5.3 %	Intercept	2.9 %
tphapred	0.1 %	tphapred	-4.6 %
pct_r_1_above_mean_mean	-1 %	pct_r_1_above_mean_mean	-5.2 %
σ_{v1}	16.3 %	σ_{v1}	13.3 %
Intercept	-29.1 %	Intercept	-17.3 %
tphapred	-33.4 %	tphapred	-18.8 %
pct_r_1_above_mean_mean	-33 %	pct_r_1_above_mean_mean	-20.1 %
σ_{v1}	-32.3 %	σ_{v1}	-15.7 %
DEN		HT	

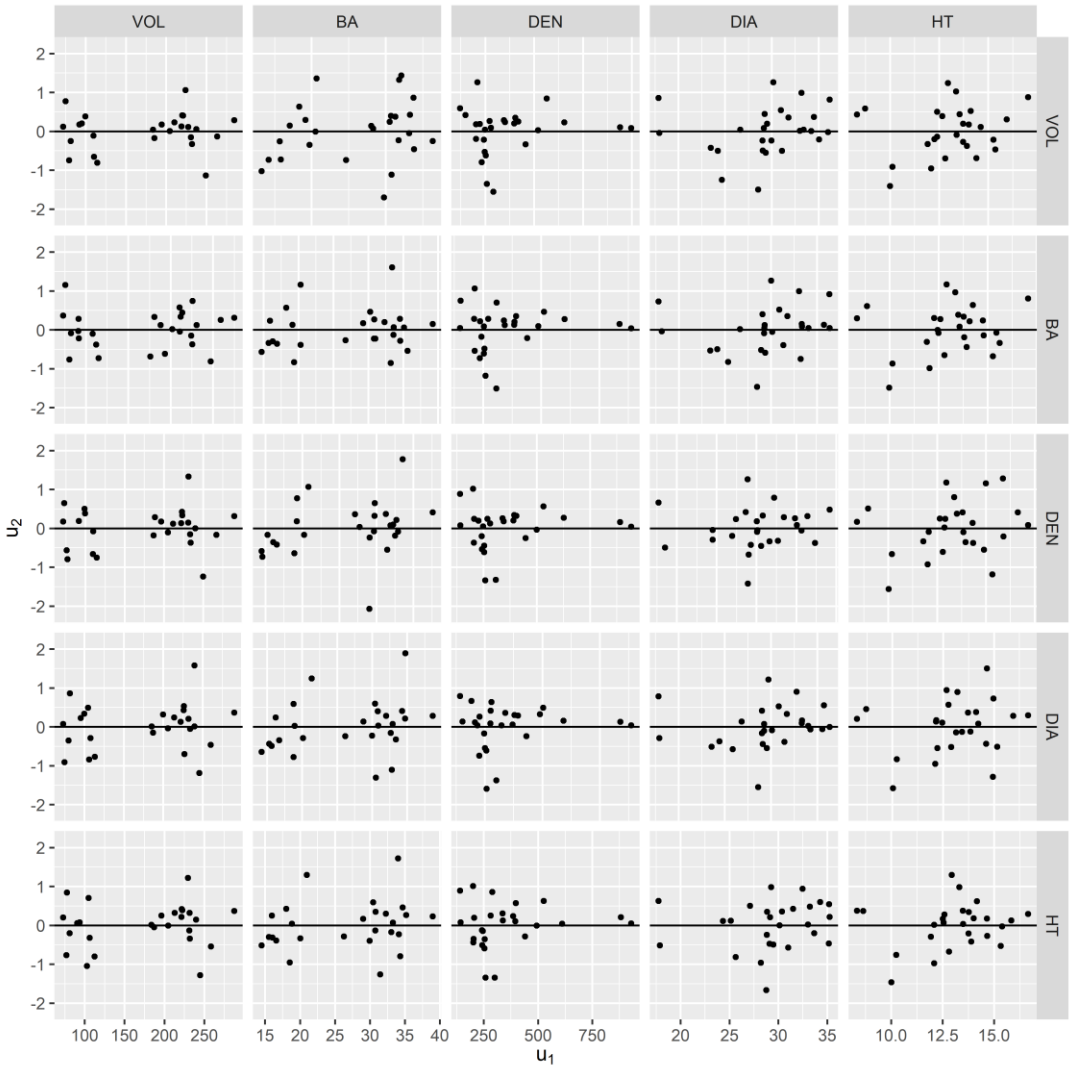

Figure 4.2: Predicted random effects standardized by the estimated random effect standard deviation for all UVFH and BVFH models. UVFH models are given along the diagonal.

Figure 4.2 displays the standardized predicted random effects for all UVFH and BVFH models for the 24 sampled stands in the BMEF. Some minor heteroskedasticity is apparent in the DEN UVFH and BVFH models. All other models appear homoscedastic. This suggests that the assumption of constant variance (section 4.3.1) is appropriate. Only minor changes in random effect predictions are apparent between UVFH and BVFH models.

4.4.2 Large Area Estimation

We produced predictions for large area parameters via the methodology presented in section 4.3.2 and 4.3.3. The point predictions, error components $g_{1\tau}$, $g_{2\tau}$, $g_{3\tau}$ and $\widehat{RMSE}$ are presented in Table 4.3 along with the error components' relative share and $\widehat{CV}$ %. Point estimates are highly consistent between UVFH and BVFH models for the large area parameter prediction. The relative share of the error components for all models indicates that $g_{2\tau}$ composes the largest amount of the total error, followed by $g_{1\tau}$ and $g_{3\tau}$. Models that resulted in the minimum $\widehat{RMSE}$ across all-pairings are consistent with those models that had reduced parameter standard errors and random effect standard deviation as described in Table 4.2. Similarly, not all pairings resulted in smaller $\widehat{RMSE}$ for the large-area estimate, this seems to be driven by larger values of $g_{1\tau}$ which, given that the predictor values and sample sizes are constant across models, implies a larger random effect variance for these models.

Table 4.3: Large area estimates and measures of uncertainty for each UVFH and BVFH model. UVFH models are indicated where u_2 is marked with two hyphens. The error components $g_{k\tau}$ and their relative share, $\frac{g_{k\tau}}{\widehat{MSE}}$, of the total error is displayed. Values < 0.01 are indicated as 0.

u_1	u_2	$\hat{\theta}_{\tau 1}$	$g_{1\tau}$	$g_{1\tau}$ (%)	$g_{2\tau}$	$g_{2\tau}$ (%)	$g_{3\tau}$	$g_{3\tau}$(%)	$\widehat{RMSE}$	$\widehat{CV}$ %	$\widehat{RMSE}$ % Diff.
VOL	--	136.70	2.02	10.17%	17.79	89.37%	0.05	0.46%	4.46	3.26%	--
	BA	**133.57**	**1.11**	**7.39%**	**13.83**	**92.11%**	**0.04**	**0.49%**	**3.87**	**2.90%**	**-15.14%**
	DEN	137.15	2.47	12.23%	17.62	87.34%	0.04	0.44%	4.49	3.27%	0.67%
	HT	136.69	2.96	13.34%	19.12	86.25%	0.04	0.41%	4.71	3.44%	5.23%
	DIA	138.49	3.83	14.78%	21.99	84.82%	0.05	0.41%	5.09	3.68%	12.39%
BA	--	23.93	0.05	13.06%	0.36	86.38%	0.00	0.55%	0.65	2.70%	--
	VOL	**23.15**	**0.06**	**16.05%**	**0.31**	**83.56%**	**0.00**	**0.38%**	**0.61**	**2.65%**	**-5.47%**
	DEN	23.21	0.09	19.14%	0.40	80.48%	0.00	0.38%	0.70	3.03%	8.23%
	HT	23.70	0.08	17.15%	0.37	82.37%	0.00	0.48%	0.67	2.84%	4.05%
	DIA	23.64	0.10	19.10%	0.41	80.45%	0.00	0.44%	0.72	3.03%	9.83%
DEN	--	324.23	18.28	8.51%	196.12	91.31%	0.19	0.18%	14.66	4.52%	--
	VOL	328.60	24.21	12.31%	172.06	87.48%	0.20	0.21%	14.02	4.27%	-4.50%
	BA	323.34	17.12	9.76%	158.05	90.03%	0.18	0.21%	13.25	4.10%	-10.61%
	HT	321.00	17.92	9.71%	166.18	90.05%	0.22	0.24%	13.58	4.23%	-7.88%
	DIA	**316.78**	**12.55**	**8.23%**	**139.62**	**91.52%**	**0.19**	**0.25%**	**12.35**	**3.90%**	**-18.66%**
HT	--	12.79	0.01	14.20%	0.03	85.33%	0.00	0.47%	0.19	1.51%	--
	VOL	12.80	0.01	15.11%	0.03	84.45%	0.00	0.44%	0.19	1.49%	-1.32%
	BA	12.80	0.01	15.34%	0.03	84.21%	0.00	0.45%	0.19	1.51%	0.57%
	DEN	12.92	0.01	16.81%	0.03	82.86%	0.00	0.33%	0.20	1.54%	3.10%
	DIA	**12.78**	**0.00**	**13.74%**	**0.03**	**85.80%**	**0.00**	**0.46%**	**0.18**	**1.39%**	**-8.87%**
DIA	--	31.07	0.02	7.14%	0.31	92.62%	0.00	0.24%	0.57	1.85%	--
	VOL	31.00	0.03	9.82%	0.28	89.92%	0.00	0.26%	0.56	1.81%	-2.30%
	BA	31.06	0.03	10.40%	0.25	89.30%	0.00	0.30%	0.53	1.72%	-7.67%
	DEN	**31.89**	**0.01**	**6.31%**	**0.16**	**93.38%**	**0.00**	**0.31%**	**0.41**	**1.30%**	**-38.62%**
	HT	30.81	0.02	6.83%	0.22	92.85%	0.00	0.32%	0.49	1.58%	-17.68%

4.4.3 Small Area Estimation

For each forest stand and for each possible model pair, point predictions of the parameter of interest were produced along with estimates of their mean squared error. Figure 4.3 displays the distribution of $\widehat{RMSE}$ disaggregated by sampled (non-synthetic) and unsampled (synthetic) predictions. As is consistent with previous results, it is clear that not all pairings provide a gain in performance, measured by

$\widehat{RMSE}$, but those pairings that demonstrated improvements in parameter precision (Table 4.2) or large area estimation mean squared error (Table 4.3) also tended to have lower median $\widehat{RMSE}$ for synthetic and non-synthetic predictions. Generally, synthetic predictions have a smaller range of $\widehat{RMSE}$ values, but higher median $\widehat{RMSE}$ values than non-synthetic predictions. We provide further summaries, including error components and synthetic and non-synthetic summaries as a table in Appendix II.

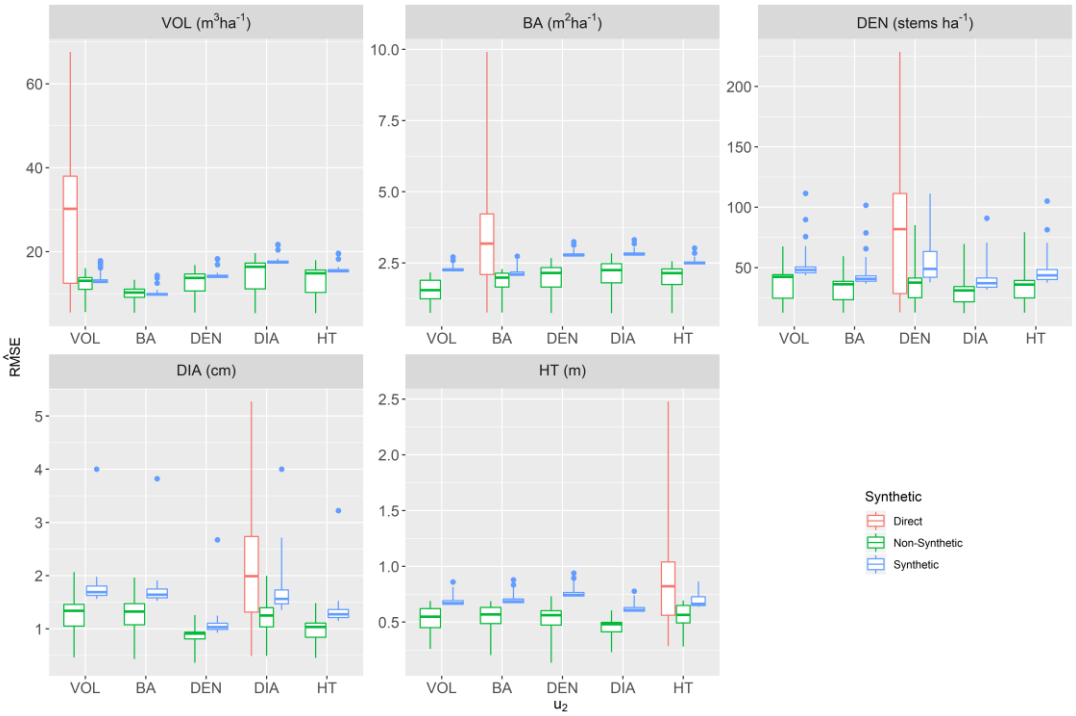

Figure 4.3: Box plots for all RMSEs for all stands in the BMEF. Sub-plot titles (gray) indicate u_1 and the x-axis of each sub-plot indicates u_2. The RMSEs for the univariate models are those where $u_1 = u_2$. RMSEs for synthetic predictions are indicated in blue, and RMSEs for EBLUP predictions are indicated in red. Medians are indicated by the line within the box, box hinges represent the interquartile range (IQR), and the upper and lower whiskers indicate the largest or smallest value no further than 4 * IQR respectively. All other stands are considered outlying and are indicated with points.

Figure 4.4 displays the change in g_{1i} and g_{2i} when adopting a bivariate pairing. Generally, models that provided smaller $\widehat{RMSE}$ showed large reductions in g_{1i} with comparatively modest reductions in g_{2i}. Conversely, models that performed

worse showed large increases in g_{1i}, this is especially evident for VOL-DIA and BA-DIA models. This implies that BVFH models that improved over their UVFH counterparts did so primarily by a reduction in estimated random effect variance, followed by an increase in precision of the coefficient parameters.

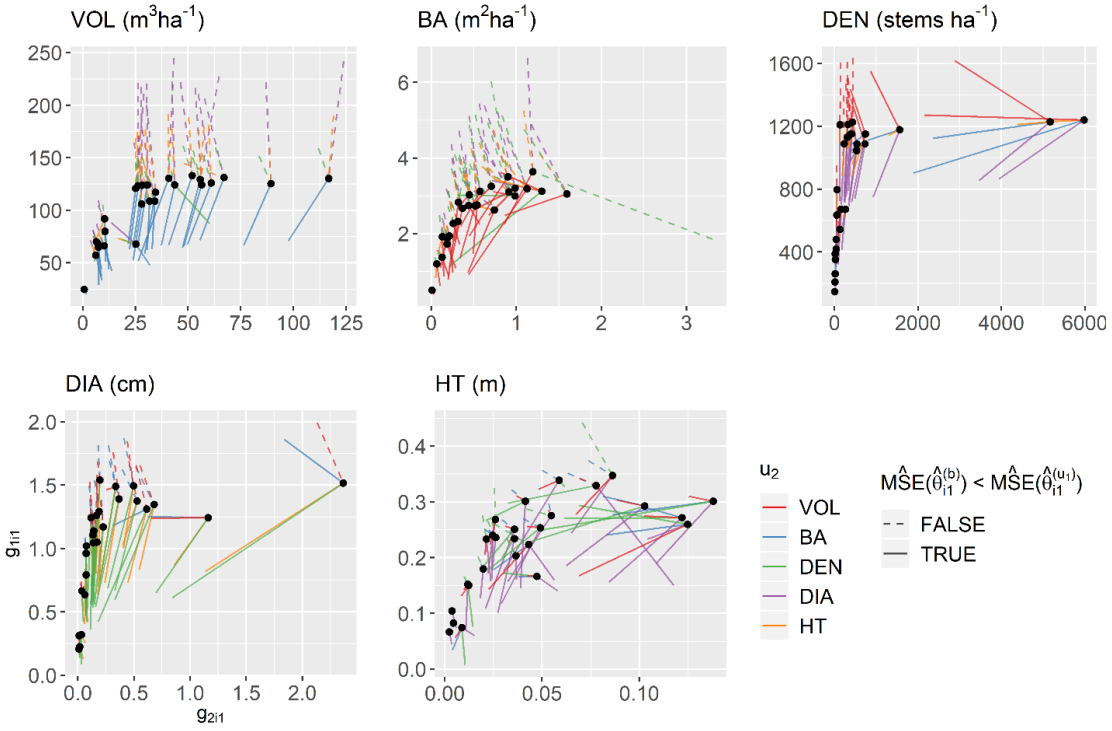

Figure 4.4: For each u_1, the univariate model g_{1i} and g_{2i} values for each sampled stand are indicated by the black points. Colored vectors emanating from each point indicate the change in g_{1i} and g_{2i} when adopting the associated bivariate model. If the mean squared error for the stand declined when adopting the bivariate model, the line is solid, otherwise it is dashed.

Figure 4.5 displays the correlation of the direct estimators, $\widehat{Corr}(\hat{\theta}_{ij}, \hat{\theta}_{ik})$, for all sampled stands in the BMEF plotted against the relative percentage decrease in $\widehat{RMSE}$. A distinctive pattern is evident at the extreme ranges of the correlation coefficient for all models. With the exception of BA and VOL, strong negative correlation coefficients resulted in BVFH models with lower $\widehat{RMSE}$ than their univariate counterparts. For all attributes, strong positive correlations reduced

$\widehat{RMSE}$. Pairs that displayed consistent behavior with respect to the correlation coefficient, such as the VOL-BA pair and the DIA-HT pair appear to benefit the most from a BVFH pairing in that their reduction is large and consistent for all sampled stands. Pairs that did not demonstrate consistent behavior or demonstrated weak correlation, e.g. VOL-HT, VOL-DEN, and BA-DIA, in many cases exhibited an increase in $\widehat{RMSE}$ relative to the UVFH model.

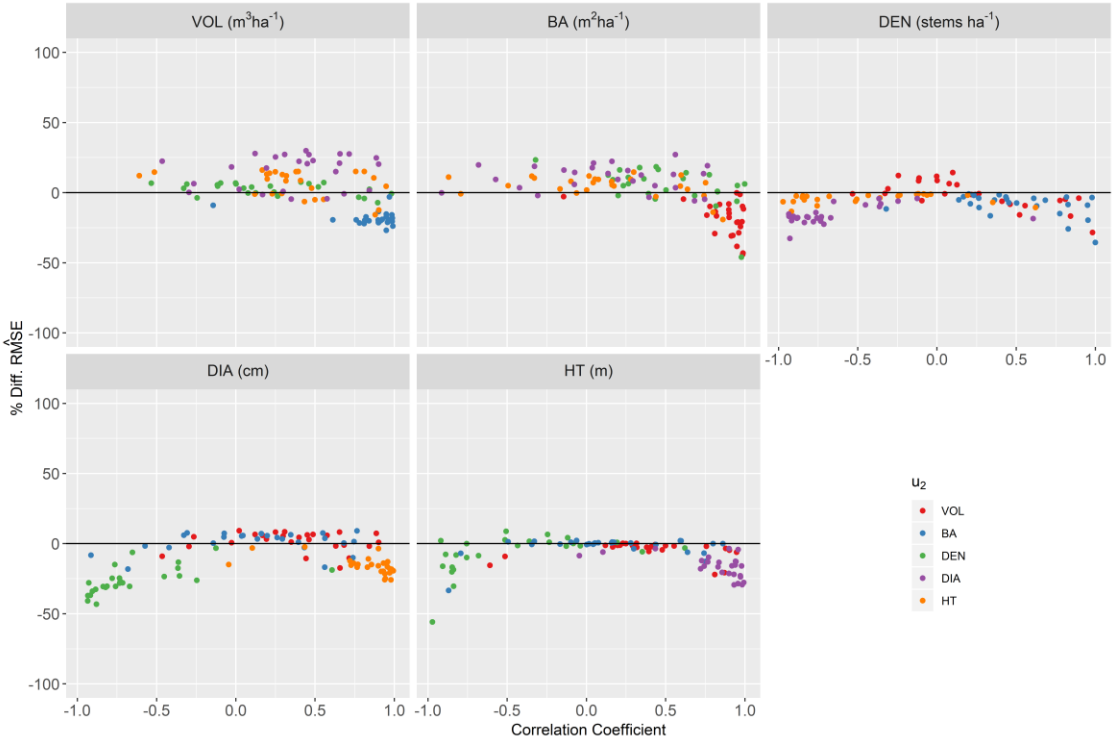

Figure 4.5: Correlation coefficient, $\widehat{Corr}(\hat{\theta}_{ij}, \hat{\theta}_{ik})$, plotted against the percentage difference RMSE for sampled stands only. Sub-plot titles indicate u_1 and colors indicate u_2.

4.5 Discussion

FH models can be used to link remote sensing data to direct estimates produced at the stand level. We presented an extension of the FH model by considering two responses simultaneously in the BVFH model and produced predictions of small and large area forest attributes. For large area predictions, at least one BVFH model can lead to

decreases in $\widehat{RMSE}$ for all of the attributes we considered. For small area predictions, this result also held for the same models. Upon inspection of the error components (Figure 4.4, Appendix II), it seems evident that the primary source of change in error for small area predictions is a change in the g_1 error component. Because the predictors used in the model and the sample data are fixed, this implies a reduction in the estimate of random effect variance components. For large area predictions (Table 4.3), the g_1 error component plays a smaller role and thus the reduction in standard error (Table 4.2) of the regression coefficients serves as the primary source of benefit for BVFH models.

We demonstrated that an important indicator of the potential success of a BVFH model in reducing $\widehat{RMSE}$ for small and large area predictions is the behavior of the correlations of direct estimators at the small area level (Figure 4.6). This is consistent with previous studies that leverage BVFH models. Ubaidillah et al. (2019) explored BVFH models for jointly prediction household consumption per capita expenditure on food and non-food goods, which are known to have strong correlation, and found that the BVFH models produced lower random effect variance estimates (i.e. $\widehat{\boldsymbol{\delta}}$) than their UVFH counterparts. Additionally, the authors provide a simulation study exploring the effect of the correlation coefficient on mean squared error estimates and found that stronger correlation plays are more dominant role in reduction than domain-specific sample size in reducing mean squared error. Importantly, the authors note that BVFH models reduce mean squared error estimates on average across all simulations. However, we have access to only one realization of the superpopulation model, therefore it can be expected to see an increase with

respect to $\widehat{RMSE}$ in some cases, especially when correlation coefficients are low (e.g. Table 4.3, Figures 4.4 and 4.5).

Adopting BVFH methodology adds an additional step to the model selection process in that it requires an assessment of any number of potentially paired attributes. In an initial phase of this study, only a smaller number of pairings were considered (VOL, BA and DEN only), however we did not observe a reduction in $\widehat{RMSE}$ when adopting the BVFH model when considering the DEN attribute. Seeking additional pairings readily available from the sample data, e.g. DIA and HT, even when their prediction may not be directly relevant to the objective of a forest inventory can still provide advantages to reduction in $\widehat{RMSE}$ for the target attribute (Table 4.3, Figure 4.3) if those attributes are strongly correlated (Figure 4.5). The strength of correlation will be a function of forest structure and type as well as domain-specific sample sizes. Additionally, a number of other direct estimators were available that we did not consider, such as the mean heights, diameters, volumes and densities disaggregated by species or by coniferous and broadleaf classes (Hou et al. 2019). Further research and application of BVFH models could explore these potential variables and their appropriateness for use via linear linking models. BVFH models also need not be constrained to attribute-attribute modeling such as that presented in this study. Another common application of MVFH in general is that of time series analysis, i.e. generating small area predictions that leverage the correlations of time variant direct estimates (Rao and Yu 1994), e.g. from a series of field surveys throughout time. Such a formulation may require auto-regressive covariance structures, provided in Benavent and Morales (2016).

The application of BVFH models in forest inventory contexts requires some special considerations. For many forest inventories, synthetic prediction is required for large proportions of unsampled domains, given the typically limited total sample size and large number of total domains. Employment of BVFH models in previous small area estimation literature uniformly rely on large domain-specific sample sizes and do not directly consider the task of aggregating the errors of multiple synthetic and non-synthetic estimates (Benavent and Morales 2016; Ubaidillah et al. 2019). Furthermore, domain-specific sample sizes in forest inventory can be small. One trait of Fay-Herriot models in general is that the sampling variances are assumed to be known quantities. For small domain-specific sample sizes, such as those used in this study, such an assumption may be untenable. One solution proposed in the literature is to implement variance smoothing techniques (Goerndt et al. 2012; Rao and Molina 2015 p. 148) with multivariate extensions provided in Isaki et al. (2000). However, in an early phase of this study we implemented variance smoothing techniques and did not observe any large gains in efficiency between UVFH models and BVFH models. Accommodation for sampling variances that are treated as unknown quantities have been provided for UVFH in the hierarchical Bayesian setting (You and Chapman 2006) and in the empirical Bayes setting (Rivest and Vandal 2003).

4.6 Conclusion

We conclude that, for all forest attributes among VOL, BA, DEN, DIA, and HT at least one BVFH model pairing exists that will reduce the model-based mean squared error estimate relative to that of the UVFH model. This held for large area estimates and synthetic and non-synthetic small area estimates. For large area estimates, this

reduction was primarily due to a decrease in an error component that describes the uncertainty of regression coefficient estimates. For small area estimates, the reduction was primarily due to a decrease in estimated random effect variance. Importantly, the use of a BVFH model does not guarantee a reduction in $\widehat{RMSE}$, and the stability (i.e. spread) and strength (i.e. mean level of correlation) of stand-level correlations was found to be an important indicator of the potential success of a BVFH model to reduce $\widehat{RMSE}$ relative to an analogous UVFH model.

5 Conclusion

RS-FI is a diverse body of methodologies that exhibit different strengths and weaknesses for solving an array of problems in forest management inventories. Both non-parametric and model-based approaches to forest inventory are continuing to proliferate in the applied and theoretical literature along with new sources of remote sensing data. Having presented three manuscripts that explore and address specific issues in forest management inventories that incorporate remote sensing, we will evaluate and discuss the contributions, limitations and potential for future research of each manuscript in turn.

Manuscript I addressed a specific forest planning task, that of identifying stands for commercial thinning. While consideration of this task had been conducted in RS-FI for the Nordic context (Pippuri et al. 2012; Korhonen et al. 2013), such an assessment had not been done for Douglas-fir stands in the Pacific Northwest. Furthermore, past studies relied on professional expertise to assign stands or plots to "thinnable" categories based on ocular assessments of field plots in physical space, which may not be practical for already existing inventories. Our methodology leveraged variables that are typically measured (stem density, basal area, etc.), such that it can be applied more readily to already existing forest inventory data. Additionally, we assessed the performance of three classifiers at different sample sizes to motivate the use of the thinning classifier for inventories where only a limited sample size is available (i.e. < 200). While sample size determination for design-based inference is well established (Thompson 2012 Ch. 4), we are not aware of any theoretical literature that provides sample-size determination for classification tasks

that leverage non-parametric methods. This study serves as a rare example of considering sample size for classification tasks in forest inventory and, while specific to the problem of commercial thinning, may provide impetus for the use of the classifier as an operational tool in existing forest inventories.

A number of limitations exist in Manuscript I that provide the basis for further research for management-oriented classifiers, such as a commercial thinning classifier. We conducted the analysis on a subset of available forest inventory data by eliminating stands that were not considered Douglas-fir dominated based on an available set of auxiliary information referred to as the Forest Operations Inventory (FOI) database. This approach has two specific drawbacks. First, it assumes the FOI database species classifications correctly quantify which stands are Douglas-fir dominated and which are not. However, the FOI database was created using a combination of aerial photograph interpretation, field visits, and other auxiliary data (Bureau of Land Management, 2015), and the species classification is likely prone to some level of error. Further research could examine the potential for discriminating different types of species or species groups (e.g. Bont et al. 2020) in tandem with the commercial thinning (or other management-specific) classifiers. Second, the use of the FOI database to determine the species compositions of forest stands precludes application of the method where such information is unavailable. This further motivates the classification of species or species groups from available auxiliary data for wider adoption of the method.

The thinning decision boundaries we employed, based primarily on hard decision boundaries, defines “thinning eligibility” as a crisp set (Pickens and Hof

1991), and models the response as a binary response. However, the decision to thin or not to thin may not be so rigid, and could depend on a number of factors such as species composition and tree vigor (Tappeiner et al. 2015 p. 187). While we did produce continuous predictions of commercial thinning eligibility (Figure 2.8), which is somewhat similar to a fuzzy set, the modeling procedure itself relied on the assumption of thinning eligibility as a crisp set. In real situations, membership to this set may be more or less certain, e.g. for plots that exhibit a large amount of defect, or presence and absence of various other kinds of species not typically targeted in commercial thinning operations. Modeling of such a phenomena may be more appropriate using methods directed towards compositional modeling (e.g. Pawlowsky-Glahn et al. 2015) and may more accurately reflect the uncertainty involved in determining the thinning eligibility of a field plot.

Manuscript II investigated the appropriateness of tree segmentation methods for producing stand-level estimates of timber volume, basal area, quadratic mean diameter and stem density and compared the results to that of the more operationally common area-based approach. While studies and software that consider the task of tree segmentation continue to proliferate (e.g. Silva et al. 2016; Roussel 2018), assessment of these methods to produce stand-level estimates are quite rare. We consider manuscript II to be an important bridge between model-based small area estimation, that provides reliable and well-established measures of error at the scale of forest management stands, and tree segmentation methods that further fragment the population. While our results indicated that, generally, it is preferable to use the ABA for producing stand-level estimates for the variables we considered, there are specific

situations where s-ITC provides important advantages, especially for the case of sampled, younger, forest management stands. S-ITC models also have access to a larger sample size, which in some cases reduced the risk of extrapolation relative to ABA models.

While manuscript II is an important step forward in formalizing the assessment of error in stand-level forest inventories that leverage tree segmentation methods, many issues remain. Foremost, the VWLM segmentation method may not be an appropriate choice for forests that exhibit a high degree of horizontal complexity and contain many gaps in the canopy. In the northwestern United States, such stands may exist in Ponderosa pine (*Pinus ponderosa*) dominated stands in eastern Oregon and Washington, or may exist in areas with partial canopy closure. In these situations, the resultant segments may be highly irregular or too large to fit into a plot as an artifact of the Voronoi tessellation, which necessarily occupies these empty areas. These situations suggest the use of a forest mask (e.g. Ostafin et al. 2017) as a preliminary step to constrain the segment sizes. However, implementation of a forest mask introduces the risk of omitting some forested areas (i.e. omission), or including areas that are not forested (i.e. commission) due to errors in the masking process itself (e.g. Finley et al. 2008). Future research could examine the propagation of this forest-mask-induced error to estimates of stand-level forest inventory variables and could be applied more generally to the ABA as well as the s-ITC method.

Manuscript II also relied on the semi-ITC population framework, which is distinct from the stricter ITC population framework (Breidenbach and Astrup 2014 pp. 115-116). However, this is not strictly necessary, and it is feasible to conduct a

small area estimation assessment using the ITC framework. However, the assumption that all population units are well defined (i.e. they are observable and have attached auxiliary data) no longer holds due to the omission and commission error of tree detection algorithms, and requires a deeper consideration toward the types of models employed. Propagation of omission and commission error to stand-level assessments could be incorporated along the lines of, e.g., Xu et al. (2019), and motivates further research that explicitly considers the prediction of small area target parameters using the ITC approach.

Manuscript III investigated the use of bivariate Fay-Herriot (BVFH) models for prediction of stand-level forest attributes. While BVFH have been discussed in the SAE literature as early as 2008 (González-Manteiga et al. 2008b), the extension of restricted maximum likelihood (REML) for estimation of multiple random-effect variance parameters, necessary for practical use, has only recently been established (Benavent and Morales 2016). Research into BVFH at the time of writing was restricted to public health and demographics studies that had access to large domain-specific sample sizes and did not consider the task of synthetic prediction as part of the analyses. However, forest inventories tend to have limited domain-specific sample sizes and the number of forest management stands that require synthetic prediction is often large. Manuscript III sought to directly address these specific issues in a study that compared the performance of BVFH against UVFH models. This manuscript was also exploratory, because BVFH models tend to perform better than UVFH when paired with strongly correlated variables, it is therefore advantageous to discover forest attributes that are strongly correlated for successful use of BVFH models.

Manuscript III established that, for the forest attributes we considered, at least one bivariate pairing existed that reduced both synthetic and non-synthetic measures of error relative to the UVFH models. Additionally, we derived a mean squared error component for arbitrary aggregations of synthetic and non-synthetic stand-level predictions for the multivariate Fay-Herriot (MVFH) model (Section 4.10).

Manuscript III is primarily limited through the assumptions imposed on the UVFH and BVFH models. Relaxing these assumptions may provide the basis for further research. For example, the variance-covariance matrix of the random effects was assumed diagonal. Benavent and Morales (2016) provide extensions for the MVFH that allow for auto-regressive and heteroskedastic auto-regressive variance-covariance structures. Models with these variance-covariance structures are typically more appropriate for MVFH models used in a time-series setting (see Rao and Yu 1994). By assuming a block-diagonal variance-covariance structure, we also assume that no spatial correlation structure exists among the random effects. Again, a larger dataset with pairs of stands at a range of distances would be required for assessing this assumption, and estimation of a spatial correlation parameter, well established in the UVFH literature (e.g. Marhuenda et al. 2013), would need to be extended to the MVFH setting. Notably, Ver Planck et al. (2018) incorporated the possibility of a conditional autoregressive process on the random effects in a Bayesian setting for UVFH models. Finally, unique to nearly all Fay-Herriot analyses, the sampling variance is assumed known (Rao and Molina 2015 p. 77), which fails to propagate to the final mean squared error estimates (see Rao and Molina 2015 pp. 148-151). For forest inventory, where domain-specific sample sizes are frequently low, this may be

an important source of error. While some theoretical literature exists that develops methodology assuming unknown sampling variances, the impact of ignored uncertainty of estimated sampling variances has yet to be investigated in RS-FI contexts. We believe adopting a Bayesian framework, as outlined in You and Chapman (2006) for UVFH, would provide in-roads toward this objective.

This dissertation has explored parametric and non-parametric predictive modeling methods for use in RS-FI contexts, and has provided important advancements in their use for specific forest inventory problems. We have identified a clear division between the uncertainty assessment of non-parametric and parametric methods throughout the text. While non-parametric methods offer clear advantages in terms of predictive performance than parametric methods in some situations, parametric methods offer a clearer foundation of error assessment for small areas. We consider the expansion of parametric methods for forest inventory problems that are currently occupied by non-parametric methods as an important objective for RS-FI to develop measures of error for small areas, as was done with respect to individual tree segmentation methods in the Manuscript II. Such an objective will require developments in small area estimation theory, remote sensing data manipulation, and other tasks. One example is the use of k-NN methods for producing predictions of species-specific forest inventories (Eskelson et al. 2009). We see compositional modeling, specifically Dirichlet models, as an opportunity for parametric methods to gain in-roads in species specific inventories.

Bibliography

Babcock, C., Finley, A.O., Bradford, J.B., Kolka, R., Birdsey, R., and Ryan, M.G. 2015. LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients. Remote Sensing of Environment **169**: 113–127.

Babcock, C., Finley, A.O., Gregoire, T.G., and Andersen, H.-E. 2018. Remote sensing to reduce the effects of spatial autocorrelation on design-based inference for forest inventory using systematic samples. Available from https://arxiv.org/abs/1810.08588 [accessed 13 December 2018].

Babcock, C., Matney, J., Finley, A.O., Weiskittel, A., and Cook, B.D. 2012. Multivariate spatial regression models for predicting individual tree structure variables using LiDAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **6**(1): 6–14.

Bailey, J.D., and Tappeiner, J.C. 1998. Effects of thinning on structural development in 40- to 100-year-old Douglas-fir stands in western Oregon. Forest Ecology and Management **108**(1–2): 99–113. doi:10.1016/S0378-1127(98)00216-3.

Bell, J.F., and Dilworth, J.R. 1990. Log scaling and timber cruising. OSU Bookstores, Corvallis, OR, US.

Benavent, R., and Morales, D. 2016. Multivariate Fay–Herriot models for small area estimation. Computational Statistics & Data Analysis **94**: 372–390.

Bont, L.G., Hill, A., Waser, L.T., Bürgi, A., Ginzler, C., and Blattert, C. 2020. Airborne-laser-scanning-derived auxiliary information discriminating between broadleaf and conifer trees improves the accuracy of models for predicting timber volume in mixed and heterogeneously structured forests. Forest Ecology and Management **459**: 117856.

Brainerd, S.M., and Rolstad, J. 2002. Habitat selection by Eurasian pine martens Martes martes in managed forests of southern boreal Scandinavia. Wildlife biology **8**(1): 289–297.

Breidenbach, J., and Astrup, R. 2012. Small area estimation of forest attributes in the Norwegian National Forest Inventory. Eur J Forest Res **131**(4): 1255–1267. doi:10.1007/s10342-012-0596-7.

Breidenbach, J., and Astrup, R. 2014. The semi-individual tree crown approach. *In* Forestry applications of airborne laser scanning. Springer. pp. 113–133.

Breidenbach, J., Magnussen, S., Rahlf, J., and Astrup, R. 2018. Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment **212**: 199–211. doi:10.1016/j.rse.2018.04.028.

Breidenbach, J., Næsset, E., Lien, V., Gobakken, T., and Solberg, S. 2010. Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. Remote Sensing of Environment **114**(4): 911–924. doi:10.1016/j.rse.2009.12.004.

Breiman, L. 2001a. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science **16**(3): 199–231.

Breiman, L. 2001b. Random forests. Machine learning **45**(1): 5–32.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. *In* Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, San Francisco, CA. pp. 785–794.

Corveil, R., and Searle, S. 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics. **18**(1): 31-38.

Cottam, G., and Curtis, J.T. 1949. A method for making rapid surveys of woodlands by means of pairs of randomly selected trees. Ecology **30**(1): 101–104.

Curtis, R.O. 1982. A simple index of stand density for Douglas-fir. Forest Science **28**(1): 92–94.

Dalponte, M., Orka, H.O., Gobakken, T., Gianelle, D., and Naesset, E. 2013. Tree species classification in boreal forests with hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing **51**(5): 2632–2645.

Datta, G.S., Hall, P., and Mandal, A. 2011. Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data. Journal of the American Statistical Association **106**(493): 362–374. doi:10.1198/jasa.2011.tm10036.

Diamantidis, N.A., Karlis, D., and Giakoumakis, E.A. 2000. Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence **116**(1–2): 1–16.

Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., and Hudak, A.T. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research **24**(3): 235–246.

Farrell, S.L., Collier, B.A., Skow, K.L., Long, A.M., Campomizzi, A.J., Morrison, M.L., Hays, K.B., and Wilkins, R.N. 2013. Using LiDAR-derived vegetation metrics for high-resolution, species distribution models for conservation planning. Ecosphere **4**(3): 1–18.

Fedrigo, M., Newnham, G.J., Coops, N.C., Culvenor, D.S., Bolton, D.K., and Nitschke, C.R. 2018. Predicting temperate forest stand types using only structural profiles from discrete return airborne lidar. ISPRS Journal of Photogrammetry and Remote Sensing **136**: 106–119. doi:10.1016/j.isprsjprs.2017.11.018.

Finley, A.O., Banerjee, S., and McRoberts, R.E. 2008. A Bayesian approach to multi-source forest area estimation. Environ Ecol Stat **15**(2): 241–258. doi:10.1007/s10651-007-0049-5.

Finley, A.O., Banerjee, S., Weiskittel, A.R., Babcock, C., and Cook, B.D. 2014. Dynamic spatial regression models for space-varying forest stand tables. Environmetrics **25**(8): 596–609.

Flewelling, J.W. 2006. Forest inventory predictions from individual tree crowns: regression modeling within a sample framework. *In* Proceedings of the eighth annual forest inventory and analysis symposium; 2006 October 16-19; Monterey, CA.

Flewelling, J.W., and McFadden, G. 2011. LiDAR data and cooperative research at Panther Creek, Oregon. Proceedings of SilviLaser.

Flewelling, J.W., and Raynes, L.M. 1993. Variable-shape stem-profile predictions for western hemlock. Part I. Predictions from DBH and total height. Canadian journal of forest research **23**(3): 520–536.

Ford, K. 2020, January 7. Traditional Methods for Bureau of Land Management Thinning Selection. Personal communication.

Forest Operations Inventory vegetation: spatial data standard. 2015. Bureau of Land Management. Available from https://www.blm.gov/sites/blm.gov/files/FOI_Vegetation_Spatial_Data_Standard_1.pdf [accessed 17 January 2020].

Frank, B. 2019. brycefrank/pyfor. Python. Available from https://github.com/brycefrank/pyfor [accessed 25 September 2019].

Freeman, E.A., Moisen, G.G., Coulston, J.W., and Wilson, B.T. 2015. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research **46**(3): 323–339.

Friedman, J.H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery **1**(1): 55–77.

Garabedian, J.E., Moorman, C.E., Nils Peterson, M., and Kilgo, J.C. 2017. Use of LiDAR to define habitat thresholds for forest bird conservation. Forest Ecology and Management **399**: 24–36. doi:10.1016/j.foreco.2017.05.024.

Gaspa, M.C., De La Cruz, R.M., Olfindo Jr, N.T., Borlongan, N.J.B., and Perez, A.M.C. 2016. Integration of manual channel initiation and flow path tracing in extracting stream features from lidar-derived DTM. *In* Remote Sensing for Agriculture, Ecosystems, and Hydrology XVIII. International Society for Optics and Photonics, Edinburg, UK. p. 999804.

Goerndt, M.E., Monleon, V.J., and Temesgen, H. 2011. A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. Can. J. For. Res. **41**(6): 1189–1201. doi:10.1139/x11-033.

Goerndt, M.E., Monleon, V.J., and Temesgen, H. 2012. Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. Forest Science **59**(5): 536–548.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., and Santamaría, L. 2008a. Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical Computation and Simulation **78**(5): 443–462.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., and Santamaría, L. 2008b. Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. Computational Statistics & Data Analysis **52**(12): 5242–5252. doi:10.1016/j.csda.2008.04.031.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S.P., and Condit, R. 2017. The continuous population approach to forest inventories and use of information in the design. Environmetrics **28**(8): e2480.

Graham, R.T., Harvey, A.E., Jain, T.B., and Tonn, J.R. 1999. Effects of thinning and similar stand treatments on fire behavior in western forests. United States Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, Oregon. Available from http://digitalcommons.usu.edu/barkbeetles/55/ [accessed 23 April 2017].

Gregoire, T.G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. **28**(10): 1429–1447. doi:10.1139/x98-166.

Grigolato, S., Mologni, O., and Cavalli, R. 2017. GIS applications in forest operations and road network planning: An overview over the last two decades. Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering **38**(2): 175–186.

Grosenbaugh, L.R. 1958. Point-sampling and line-sampling probability theory, geometric implications, synthesis. USDA Forest Service, Southern Forest Experiment Station, Occasional Paper 160.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media.

Hawbaker, T.J., Keuler, N.S., Lesak, A.A., Gobakken, T., Contrucci, K., and Radeloff, V.C. 2009. Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design. Journal of Geophysical Research: Biogeosciences **114**(G2).

Hou, Z., Mehtätalo, L., McRoberts, R.E., Ståhl, G., Tokola, T., Rana, P., Siipilehto, J., and Xu, Q. 2019. Remote sensing-assisted data assimilation and simultaneous inference for forest inventory. Remote Sensing of Environment **234**: 111431. doi:10.1016/j.rse.2019.111431.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment **112**(5): 2232–2245. doi:10.1016/j.rse.2007.10.009.

Hutter, F., Hoos, H.H., and Leyton-Brown, K. 2011. Sequential model-based optimization for general algorithm configuration. *In* International Conference on Learning and Intelligent Optimization. Springer, Rome, Italy. pp. 507–523.

Hyyppa, J. 1999. Detecting and estimating attributes for single trees using laser scanner. Photogramm J Finland **16**: 27–42.

Hyyppa, J., Kelle, O., Lehikoinen, M., and Inkinen, M. 2001. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. IEEE Transactions on geoscience and remote sensing **39**(5): 969–975.

Iles, K. 2003. A sampler of inventory topics. Kim Isles & Associates Ltd., Nanimo. British Columbia, Canada.

Instrucciones generales para la redacción de proyectos de ordenación de montesgestionados por la Comunidad de Madrid. 2012. Comunidad de Madrid.

Isaki, C.T., Tsay, J.H., and Fuller, W.A. 2000. Estimation of census adjustment factors. Survey Methodology **26**(1): 31–42.

Jakubowski, M.K., Li, W., Guo, Q., and Kelly, M. 2013. Delineating Individual Trees from Lidar Data: A Comparison of Vector- and Raster-based Segmentation Approaches. Remote Sensing **5**(9): 4163–4186. doi:10.3390/rs5094163.

Jeronimo, S.M., Kane, V.R., Churchill, D.J., McGaughey, R.J., and Franklin, J.F. 2018. Applying LiDAR individual tree detection to management of structurally diverse forest landscapes. Journal of Forestry **116**(4): 336–346.

Kangas, A., Kurttila, M., Hujala, T., Eyvindson, K., and Kangas, J. 2015. Decision Support for Forest Management. *In* 2nd edition. Springer International Publishing. Available from https://www.springer.com/gp/book/9783319235219 [accessed 5 September 2019].

Kankare, V., Vastaranta, M., Holopainen, M., Räty, M., Yu, X., Hyyppä, J., Hyyppä, H., Alho, P., Viitala, R., Kankare, V., Vastaranta, M., Holopainen, M., Räty, M., Yu, X., Hyyppä, J., Hyyppä, H., Alho, P., and Viitala, R. 2013. Retrieval of Forest Aboveground Biomass and Stem Volume with Airborne Scanning LiDAR. Remote Sensing **5**(5): 2257–2274. doi:10.3390/rs5052257.

Kim, H.-J., and Tomppo, E. 2006. Model-based prediction error uncertainty estimation for k-nn method. Remote Sensing of Environment **104**(3): 257–263. doi:10.1016/j.rse.2006.04.009.

Kim, Y., Yang, Z., Cohen, W.B., Pflugmacher, D., Lauver, C.L., and Vankat, J.L. 2009. Distinguishing between live and dead standing tree biomass on the North Rim of Grand Canyon National Park, USA using small-footprint lidar data. Remote Sensing of Environment **113**(11): 2499–2510. doi:10.1016/j.rse.2009.07.010.

Korhonen, L., Pippuri, I., Packalén, P., Heikkinen, V., Maltamo, M., and Heikkilä, J. 2013. Detection of the need for seedling stand tending using high-resolution remote sensing data. Silva Fennica **47**(2): 1–20.

Landis, J.R., and Koch, G.G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics **33**: 363–374.

Lindberg, E., Holmgren, J., Olofsson, K., Wallerman, J., and Olsson, H. akan. 2010. Estimation of tree lists from airborne laser scanning by combining single-tree and area-based methods. International Journal of Remote Sensing **31**(5): 1175–1192.

Magnussen, S., and Breidenbach, J. 2017. Model-dependent forest stand-level inference with and without estimates of stand-effects. Forestry (Lond) **90**(5): 675–685. doi:10.1093/forestry/cpx023.

Magnussen, S., Breidenbach, J., and Mauro, F. 2017. The challenge of estimating a residual spatial autocorrelation from forest inventory data. Canadian Journal of Forest Research **47**(11): 1557–1566.

Maltamo, M., and Packalen, P. 2014. Species-specific management inventory in Finland. *In* Forestry applications of airborne laser scanning. *Edited by* M. Maltamo, E. N\a esset, and J. Vauhkonen. Springer. pp. 241–252.

Mandallaz, D. 1991. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. PhD Thesis, ETH Zurich.

Marhuenda, Y., Molina, I., and Morales, D. 2013. Small area estimation with spatio-temporal Fay–Herriot models. Computational Statistics & Data Analysis **58**(0): 308–325. doi:10.1016/j.csda.2012.09.002.

Matern, B. 1956. On the geometry of the cross-section of a stem.

Mauro, F., Frank, B., Monleon, V.J., Temesgen, H., and Ford, K. 2019a. Prediction of diameter distributions and tree-lists in southwestern Oregon using LiDAR and stand-level auxiliary information. Canadian Journal of Forest Research **49**(7): 775–787.

Mauro, F., Molina, I., García-Abril, A., Valbuena, R., and Ayuga-Téllez, E. 2016. Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. Environmetrics **27**(4): 225–238. doi:10.1002/env.2387.

Mauro, F., Monleon, V.J., Temesgen, H., and Ford, K.R. 2017a. Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. PLOS ONE **12**(12): e0189401. doi:10.1371/journal.pone.0189401.

Mauro, F., Monleon, V.J., Temesgen, H., and Ruiz, L.A. 2017b. Analysis of spatial correlation in predictive models of forest variables that use LiDAR auxiliary information. Can. J. For. Res. **47**(6): 788–799. doi:10.1139/cjfr-2016-0296.

Mauro, F., Ritchie, M., Wing, B., Frank, B., Monleon, V., Temesgen, H., and Hudak, A. 2019b. Estimation of Changes of Forest Structural Attributes at Three Different Spatial Aggregation Levels in Northern California using Multitemporal LiDAR. Remote Sensing **11**(8): 923.

McGaughey, R. 2016. FUSION/LDV: Software for LIDAR Data Analysis and Visualization. *In* FUSION Version 3.60+. United Stated Department of Agriculture, Pacific Northwest Research Station.

McRoberts, R.E., Magnussen, S., Tomppo, E.O., and Chirici, G. 2011. Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sensing of Environment **115**(12): 3165–3174.

McRoberts, R.E., and Tomppo, E.O. 2007. Remote sensing support for national forest inventories. Remote sensing of environment **110**(4): 412–419.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., and Heikkinen, J. 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sensing of Environment **111**(4): 466–480.

Melin, M., Matala, J., Mehtätalo, L., Suvanto, A., and Packalen, P. 2015. Detecting moose (Alces alces) browsing damage in young boreal forests from airborne laser scanning data. Canadian Journal of Forest Research **46**(1): 10–19.

Mentch, L., and Hooker, G. 2017. Formal Hypothesis Tests for Additive Structure in Random Forests. Journal of Computational and Graphical Statistics **26**(3): 589–597. doi:10.1080/10618600.2016.1256817.

Miller, T.L. based on F. code by A. 2017. leaps: Regression Subset Selection. Available from https://CRAN.R-project.org/package=leaps [accessed 3 September 2019].

Mizuniwa, Y., Nakahata, C., and Aruga, K. 2016. Comparative analyses on the cycle time, productivity, and cost between commercial thinning and clear-cutting operations in Nasu-machi Forest Owners' Co-operative, Tochigi Prefecture, Japan. Journal of forest research **21**(2): 99–104.

Moore, J., and Lin, Y. 2019. Determining the extent and drivers of attrition losses from wind using long-term datasets and machine learning techniques. Forestry **92**(4): 425–435. doi:10.1093/forestry/cpy047.

Moreno-Torres, J.G., Sáez, J.A., and Herrera, F. 2012. Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems **23**(8): 1304–1312.

Mund, J.-P., Wilke, R., Körner, M., and Schultz, A. 2015. Detecting multi-layered forest stands using high density airborne LiDAR data. Journal for Geographic Information Science **1–2015**: 178–188. doi:10.1553/giscience2015s178.

Næsset, E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment **80**(1): 88–99.

Næsset, E. 2014. Area-based inventory in Norway–from innovation to an operational reality. *In* Forestry applications of airborne laser scanning. Springer. pp. 215–240.

Nelson, R. 2013. How did we get here? An early history of forestry lidar. Canadian Journal of Remote Sensing **39**(s1): S6–S17. doi:10.5589/m13-011.

Nevalainen, O., Honkavaara, E., Tuominen, S., Viljanen, N., Hakala, T., Yu, X., Hyyppä, J., Saari, H., Pölönen, I., and Imai, N. 2017. Individual tree detection and classification with UAV-based photogrammetric point clouds and hyperspectral imaging. Remote Sensing **9**(3): 185.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society **97**(4): 558–625.

Ng, A.Y., and Jordan, M.I. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *In* Advances in neural information processing systems. Vancouver, BC, CA. pp. 841–848.

Nielsen, D. 2016. Tree Boosting With XGBoost-Why Does XGBoost Win "Every" Machine Learning Competition? Master's Thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Nothdurft, A., Saborowski, J., and Breidenbach, J. 2009. Spatial prediction of forest stand variables. Eur J Forest Res **128**(3): 241–251. doi:10.1007/s10342-009-0260-z.

Oliver, W.W. 2000. Ecological research at the Blacks Mountain Experimental Forest in northeastern California.

Ostafin, K., Iwanowski, M., Kozak, J., Cacko, A., Gimmi, U., Kaim, D., Psomas, A., Ginzler, C., and Ostapowicz, K. 2017. Forest cover mask from historical topographic maps based on image processing. Geoscience Data Journal **4**(1): 29–39.

Packalen, P., Strunk, J., Packalen, T., Maltamo, M., and Mehtätalo, L. 2019. Resolution dependence in an area-based approach to forest inventory with airborne laser scanning. Remote sensing of environment **224**: 192–201.

Pascual, A., Pukkala, T., Rodríguez, F., and de-Miguel, S. 2016. Using spatial optimization to create dynamic harvest blocks from LiDAR-based small interpretation units. Forests **7**(10): 220. doi:10.3390/f7100220.

Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, R. 2015. Modeling and analysis of compositional data. John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research **12**(Oct): 2825–2830.

Peuhkurinen, J., Mehtätalo, L., and Maltamo, M. 2011. Comparing individual tree detection and the area-based statistical approach for the retrieval of forest stand characteristics using airborne laser scanning in Scots pine stands. Canadian Journal of Forest Research **41**(3): 583–598.

Pickens, J.B., and Hof, J.G. 1991. Fuzzy goal programming in forestry: An application with special solution problems. Fuzzy Sets and Systems **39**(3): 239–246. doi:10.1016/0165-0114(91)90095-8.

Pinheiro, J., and Bates, D. 2006. Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. 2019. nlme: Linear and Nonlinear Mixed Effects Models. Available from https://CRAN.R-project.org/package=nlme [accessed 4 September 2019].

Pippuri, I., Kallio, E., Maltamo, M., Peltola, H., and Packalén, P. 2012. Exploring horizontal area-based metrics to discriminate the spatial pattern of trees and need for first thinning using airborne laser scanning. Forestry (Lond) **85**(2): 305–314. doi:10.1093/forestry/cps005.

Pippuri, I., Suvanto, A., Maltamo, M., Korhonen, K.T., Pitkänen, J., and Packalen, P. 2016. Classification of forest land attributes using multi-source remotely sensed data. International Journal of Applied Earth Observation and Geoinformation **44**: 11–22. doi:10.1016/j.jag.2015.07.002.

Popescu, S.C., Wynne, R.H., and Nelson, R.F. 2002. Estimating plot-level tree heights with lidar: local filtering with a canopy-height based variable window size. Computers and Electronics in Agriculture **37**(1): 71–95. doi:10.1016/S0168-1699(02)00121-7.

Prasad, N., and Rao, J.N.K., The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association **85**(409): 163-171.

Rahlf, J., Breidenbach, J., Solberg, S., and Astrup, R. 2015. Forest parameter prediction using an image-based point cloud: a comparison of semi-ITC with ABA. Forests **6**(11): 4059–4071.

Rao, J.N., and Yu, M. 1994. Small-area estimation by combining time-series and cross-sectional data. Canadian Journal of Statistics **22**(4): 511–528.

Rao, J.N.K., and Molina, I. 2015. Small Area Estimation. *In* 2nd edition. John Wiley & Sons, Inc.

Räty, J., Packalen, P., Kotivuori, E., and Maltamo, M. 2019. Fusing diameter distributions predicted by area-based approach and individual tree detection in coniferous-dominated forests. Canadian Journal of Forest Research (ja).

Reutebuch, S.E., McGaughey, R.J., Andersen, H.-E., and Carson, W.W. 2003. Accuracy of a high-resolution lidar terrain model under a conifer forest canopy. Canadian journal of remote sensing **29**(5): 527–535.

Rivest, L.-P., and Vandal, N. 2003. Mean squared error estimation for small areas when the small area variances are estimated. *In* Proceedings of the International Conference on Recent Advances in Survey Sampling. Laboratory for Research in Statistics and Probability, Carleton University ….

Roussel, J.-R. 2018. lidR. Available from https://github.com/Jean-Romain/lidR.

Sanchez-Lopez, N., Boschetti, L., and Hudak, A.T. 2018. Semi-Automated Delineation of Stands in an Even-Age Dominated Forest: A LiDAR-GEOBIA Two-Stage Evaluation Strategy. Remote Sensing **10**(10): 1622. doi:10.3390/rs10101622.

Sandino, J., Pegg, G., Gonzalez, F., and Smith, G. 2018. Aerial mapping of forests affected by pathogens using UAVs, hyperspectral sensors, and artificial intelligence. Sensors **18**(4): 944.

Särndal, C.-E., Swensson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer Science & Business Media.

Senécal, J.-F., Doyon, F., and St-Onge, B. 2018. Discrimination of canopy gaps and non-regenerating openings in old-growth temperate deciduous forests using airborne LiDAR data. Canadian Journal of Forest Research **48**(7): 774–782.

Sheridan, R., Popescu, S., Gatziolis, D., Morgan, C., and Ku, N.-W. 2015. Modeling forest aboveground biomass and volume using airborne LiDAR metrics and forest inventory and analysis data in the Pacific Northwest. Remote Sensing **7**(1): 229–255.

Shi, Y., Wang, T., Skidmore, A.K., and Heurich, M. 2018. Important LiDAR metrics for discriminating forest tree species in Central Europe. ISPRS Journal of Photogrammetry and Remote Sensing **137**: 163–174. doi:10.1016/j.isprsjprs.2018.02.002.

Shin, J., and Temesgen, H. 2018. Generating Tree-Lists by Fusing Individual Tree Detection and Nearest Neighbor Imputation Using Airborne LiDAR Data. Open Journal of Forestry **8**(04): 500.

Shin, J., Temesgen, H., Strunk, J.L., and Hilker, T. 2016. Comparing modeling methods for predicting forest attributes using LiDAR metrics and ground measurements. Canadian Journal of Remote Sensing **42**(6): 739–765.

Silva, C.A., Hudak, A.T., Vierling, L.A., Loudermilk, E.L., O'Brien, J.J., Hiers, J.K., Jack, S.B., Gonzalez-Benecke, C., Lee, H., and Falkowski, M.J. 2016. Imputation of individual Longleaf Pine (Pinus palustris Mill.) Tree attributes from field and LiDAR data. Canadian journal of remote sensing **42**(5): 554–573.

Singh, K.K., Davis, A.J., and Meentemeyer, R.K. 2015. Detecting understory plant invasion in urban forests using LiDAR. International Journal of Applied Earth Observation and Geoinformation **38**: 267–279.

Strîmbu, V.F., and Strîmbu, B.M. 2015. A graph-based segmentation algorithm for tree crown extraction using airborne LiDAR data. ISPRS Journal of Photogrammetry and Remote Sensing **104**: 30–43.

Strunk, J., Packalen, P., Gould, P., Gatziolis, D., Maki, C., Andersen, H.-E., and McGaughey, R.J. 2019. Large Area Forest Yield Estimation with Pushbroom Digital Aerial Photogrammetry. Forests **10**(5): 397.

Suzuki, N., and Hayes, J.P. 2003. Effects of Thinning on Small Mammals in Oregon Coastal Forests. The Journal of Wildlife Management **67**(2): 352–371. doi:10.2307/3802777.

Takayama, T., and Iwasaki, A. 2016. Optimal Wavelength Selection on Hyperspectral Data with Fused LASSO for Biomass Estimation of Tropical Rain Forest. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences **3**(8): 101–108.

Tappeiner, J.C., Maguire, D.A., Harrington, T.B., and Bailey, J. 2015. Silviculture and ecology of western US forests. Oregon State University Press. Available from http://agris.fao.org/agris-search/search.do?recordID=US201300114139 [accessed 4 May 2017].

Temesgen, H., Goerndt, M.E., Johnson, G.P., Adams, D.M., and Monserud, R.A. 2007. Forest measurement and biometrics in forest management: status and future needs of the Pacific Northwest USA. Journal of Forestry **105**(5): 233–238.

Thompson, S.K. 2012. Sampling. John Wiley & Sons, Hoboken, USA.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1): 267–288.

Tomppo, E., Korhonen, K.T., Heikkinen, J., and Yli-Kojola, H. 2001. Multisource inventory of the forests of the Hebei Forestry Bureau, Heilongjiang, China. Siliva Fennica **44**(8): 931–948.

Tomppo, E., Malimbwi, R., Katila, M., Mäkisara, K., Henttonen, H.M., Chamuya, N., Zahabu, E., and Otieno, J. 2014. A sampling design for a large area forest inventory: case Tanzania. Canadian Journal of Forest Research **44**(8): 931–948.

Ubaidillah, A., Notodiputro, K.A., Kurnia, A., and Mangku, I.W. 2019. Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. Journal of Applied Statistics: 1–17.

Valbuena, R., Maltamo, M., and Packalen, P. 2016a. Classification of forest development stages from national low-density lidar datasets: a comparison of machine learning methods. Revista de Teledetección (45): 15–25.

Valbuena, R., Maltamo, M., and Packalen, P. 2016b. Classification of multilayered forest development classes from low-density national airborne lidar datasets. Forestry. doi:10.1093/forestry/cpw010.

Ver Planck, N.R., Finley, A.O., Kershaw Jr, J.A., Weiskittel, A.R., and Kress, M.C. 2018. Hierarchical Bayesian models for small area estimation of forest variables using LiDAR. Remote Sensing of Environment **204**: 287–295.

Wang, Y. 2019. Volume Estimator Library Equations. US Forest Service.

Wang, Y., Hyyppä, J., Liang, X., Kaartinen, H., Yu, X., Lindberg, E., Holmgren, J., Qin, Y., Mallet, C., Ferraz, A., Torabzadeh, H., Morsdorf, F., Zhu, L., Liu, J., and Alho, P. 2016. International Benchmarking of the Individual Tree Detection Methods for Modeling 3-D Canopy Structure for Silviculture and Forest Ecology Using Airborne Laser Scanning. IEEE Transactions on Geoscience and Remote Sensing **54**(9): 5011–5027. doi:10.1109/TGRS.2016.2543225.

Waugh, M., and Shakoor, A. 2015. Characterizing Slope Stability of Colluvial Soils in Ohio Using LiDAR Data. *In* Engineering Geology for Society and Territory-Volume 2. Springer. pp. 249–253.

White, J., Tompalski, P., Vastaranta, M., Wulder, M.A., Saarinen, N., Stepper, C., and Coops, N.C. 2017. A model development and application guide for generating an enhanced forest inventory using airborne laser scanning data and an area-based approach. Natural Resources Canada= Ressources naturelles Canada.

White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., and Tompalski, P. 2016. Remote sensing technologies for enhancing forest inventories: A review. Canadian Journal of Remote Sensing **42**(5): 619–641.

Wilson, D.S., and Puettmann, K.J. 2007. Density management and biodiversity in young Douglas-fir forests: Challenges of managing across scales. Forest Ecology and Management **246**(1): 123–134. doi:10.1016/j.foreco.2007.03.052.

Wing, B.M., Ritchie, M.W., Boston, K., Cohen, W.B., and Olsen, M.J. 2015. Individual snag detection using neighborhood attribute filtered airborne lidar data. Remote Sensing of Environment **163**: 165–179.

Wolter, K.M. 1984. An investigation of some estimators of variance for systematic sampling. Journal of the American Statistical Association **79**(388): 781–790.

Wulder, M.A., Bater, C.W., Coops, N.C., Hilker, T., and White, J.C. 2008. The role of LiDAR in sustainable forest management. The Forestry Chronicle **84**(6): 807–826.

Xia, Y., Liu, C., Li, Y., and Liu, N. 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications **78**(15): 225–241. doi:10.1016/j.eswa.2017.02.017.

Xu, Q., Li, B., Maltamo, M., Tokola, T., and Hou, Z. 2019. Predicting tree diameter using allometry described by non-parametric locally-estimated copulas from tree dimensions derived from airborne laser scanning. Forest ecology and management **434**: 205–212.

Xu, Q., Man, A., Fredrickson, M., Hou, Z., Pitkänen, J., Wing, B., Ramirez, C., Li, B., and Greenberg, J.A. 2018. Quantification of uncertainty in aboveground biomass estimates derived from small-footprint airborne LiDAR. Remote sensing of environment **216**: 514–528.

Yang, L., Zhang, X., Liang, S., Yao, Y., Jia, K., and Jia, A. 2018. Estimating Surface Downward Shortwave Radiation over China Based on the Gradient Boosting Decision Tree Method. Remote Sensing **10**(2): 185.

Yebra, M., Marselis, S., Van Dijk, A., Cary, G., and Chen, Y. 2015. Using LiDAR for forest and fuel structure mapping: options, benefits, requirements and costs.

You, Y., and Chapman, B. 2006. Small area estimation using area level models and estimated sampling variances. Survey Methodology **32**(1): 97.

Yu, X., Hyyppä, J., Karjalainen, M., Nurminen, K., Karila, K., Vastaranta, M., Kankare, V., Kaartinen, H., Holopainen, M., and Honkavaara, E. 2015. Comparison of laser and stereo optical, SAR and InSAR point clouds from air-and space-borne sources in the retrieval of forest inventory attributes. Remote Sensing **7**(12): 15933–15954.

Zhang, K., Chen, S.-C., Whitman, D., Shyu, M.-L., Yan, J., and Zhang, C. 2003. A progressive morphological filter for removing nonground measurements from airborne LIDAR data. IEEE transactions on geoscience and remote sensing **41**(4): 872–882.

Zhen, Z., Quackenbush, L.J., and Zhang, L. 2016. Trends in Automatic Individual Tree Crown Detection and Delineation—Evolution of LiDAR Data. Remote Sensing **8**(4): 333. doi:10.3390/rs8040333.

Appendices

A.1 Chapter 2 Appendix

Table A 1: Summary of LiDAR predictors computed using points above 2 meters at ground level.

Metric	**Description**
elev_min (m)	Minimum height
elev_max (m)	Maximum height
elev_ave (m)	Mean height
elev_stddev (m)	Standard deviation of height
elev_CV	Coefficient of variation of height
elev_IQ	Interquartile range of height
elev_AAD	Average absolute deviation of height
elev_p01	1st percentile height
elev_p05	5th percentile height
elev_p10	10th percentile height
elev_p20	20th percentile height
elev_p25	25th percentile height
elev_p30	30th percentile height
elev_p40	40th percentile height
elev_p50	50th percentile height
elev_p60	60th percentile height
elev_p75	75th percentile height
elev_p80	80th percentile height
elev_p85	85th percentile height
elev_p90	90th percentile height
elev_p95	95th percentile height
elev_p99	99th percentile height
first_cover_above2m	Percentage of first returns above 2 m
all_cover_above2m	Percentage of returns above 2 m
all_1st_cover_above2m	Number of first returns above 2 m / total number of first returns
first_cover_above_mean	Percentage of first returns above mean height
all_cover_above_mean	Percentage of all returns above mean height
all_1st_cover_above_mean	Number of first returns above mean / total number of first returns

Table A 2: Random Forests full sample hyperparameters.

n_estimators	**max_features**	**min_samples_leaf**
500	1	6
1000	2	11
	3	16
	4	21
	5	26
	6	31
	7	36
	8	41
	9	46
	10	51
		56
		61
		66
		71
		76
		81
		86
		91
		96
		101

Table A 3: XGB full sample hyperparameters.

n_estimators	learning_rate	max_depth	subsample	colsample_bytree
1	0.01	1	0.1	0.1
6	0.03	6	0.3	0.3
11	0.05	11	0.5	0.5
16	0.07	16	0.7	0.7
21	0.1	21	0.9	0.9
26	0.15	26	1	1
31	0.2	31		
36	0.25			
41	0.3			
46				
51				
56				
61				
66				
71				
76				
81				
86				
91				
96				
101				

Table A 4: Random Forests downsample hyperparameters.

n_estimators	**max_features**	**min_samples_leaf**
500	1	1
1000	2	4
	3	7
	4	10
	5	13
	6	16

Table A 5: XGB downsample hyperparameters.

n_estimators	learning_rate	max_depth	subsample	colsample_bytree
50	0.01	1	0.1	0.1
55	0.03	6	0.3	0.3
60	0.05	11	0.5	0.5
65	0.07	16		
70	0.1	21		
75		26		
80		31		
85				
90				
95				
100				

Table A 6: Confusion matrices for LOG, RF and XGB methods.

LOG		**Predicted**		
		NO_THIN	THIN	Omission Error
Observed	NO_THIN	**515**	50	8.8%
	THIN	56	**32**	63.6%
Commission Error		9.8%	60.9%	**OA:** 83.7%

RF		**Predicted**		
		NO_THIN	THIN	Omission Error
Observed	NO_THIN	**527**	38	6.7%
	THIN	41	**47**	46.6%
Commission Error		7.2%	44.7%	**OA:** 87.9%

XGB		**Predicted**		
		NO_THIN	THIN	Omission Error
Observed	NO_THIN	**543**	22	3.9%
	THIN	35	**53**	39.7%
Commission Error		6.1%	29.3%	**OA:** 91.1%

A.2 Chapter 3 Appendix

Table A 7: Definition of major notation.

Target Parameters and Their Components			
Notation	**Description**	**Notation**	**Description**
μ_α	The target parameter for an area indexed by α.	h_{ij}	The area occupied by the jth population unit in the ith area in hectares.
$h_{i\cdot}$	The sum of the areas of the population in the ith area in hectares.	$h_{\cdot\cdot}$	The sum of the areas of all population units in the entire study region in hectares.
Models and Their Components			
Notation	**Description**	**Notation**	**Description**
$\boldsymbol{y}$	A vector of observable quantities of the response variable for all population units.	$\boldsymbol{e}$	A vector of residuals.
$\boldsymbol{X}$	A design matrix of lidar covariates and an intercept for all population units.	$\boldsymbol{\beta}$	A vector of regression coefficients.
$\boldsymbol{Z}$	A binary matrix that assigns population units to areas.	$\boldsymbol{G}$	The variance-covariance matrix of $\boldsymbol{v}$.
$\boldsymbol{v}$	A vector of realized random effects.	$\boldsymbol{R}$	The variance-covariance matrix of $\boldsymbol{e}$.
σ_e^2	The residual variance.	σ_v^2	The random-effect variance.
N_i	The number of population units in stand i.	n_i	The sample size in stand i.
Results Assessment Measures			
Notation	**Description**	**Notation**	**Description**
$\widehat{MSE}_\alpha$, $R\widehat{MSE}_\alpha$	The estimated mean squared error and root mean squared error of the predicted target parameter α.	$\widehat{CV}_\alpha$	The estimated coefficient of variation of the predicted target parameter.
$g_{1\alpha}$	The mean squared error component for the target parameter for an area indexed by α concerning the random effect variance.	$mRMSE$, $rRMSE$	Cross-validation measures of RMSE and relative RMSE for unit-level predictions.
$g_{2\alpha}$	The mean squared error component for the target parameter for an area indexed by α concerning the uncertainty from the estimate of $\widehat{\boldsymbol{\beta}}$.	$mBias$, $rBias$	Cross-validation measures of bias and relative bias for unit-level predictions.

Table A 8: Lidar predictors and descriptions.

Predictor Name	**Description**
p_1, p_10, p_20, p_25, p_30, p_40, p_50, p_60, p_70, p_75, p_80, p_90, p_95, p_99	The percentile of the z-dimension indicated by the trailing number. For example _95 describes the elevation at which 95% of the lidar points fall below.
max_z	The maximum z value.
min_z	The minimum z value.
mean_z	The mean z value.
stddev_z	The standard deviation of the z values.
var_z	The variance of the z values.
mean_z_sq	The square of the mean z value.
vol_cov	The product of the mean z value and the pct_r_1_above_2 metric.
pct_all_above_2	The proportion of all returns above 2 meters.
pct_all_above_mean	The proportion of all returns above the mean z value.
pct_r_1_above_2	The proportion of first returns above 2 meters.
pct_r_1_above_mean	The proportion of all returns above 2 meters.
r_1, r_2, r_3, r_4	The number of returns indicated by the trailing number. For example, r_1 indicates the number of first returns.
area	The area of the population unit (only included for s-ITC models)

A.3 Chapter 4 Appendix

A.3.1 Lidar Predictors and Descriptions

For each grid cell in the study region the following metrics were computed. Following this, each metric was aggregated to a stand level mean and variance for each stand. The stand-level means and variances of each metric serve as the predictors used in the model selection procedure (section 3.5).

Table A 9: Lidar predictors and descriptions.

Predictor Name	**Description**
p_20, p_40, p_50, p_60, p_70, p_80, p_90, p_95	The percentile of the z-dimension indicated by the trailing number. For example _95 describes the elevation at which 95% of the lidar points fall below.
mean_z	The mean z value.
stddev_z	The standard deviation of the z values.
var_z	The variance of the z values.
mean_z_sq	The square of the mean z value.
vol_cov	The product of the mean z value and the pct_r_1_above_2 metric.
canopy_relief_ratio	(mean_z – min_z) / (max_z – min_z)
pct_r_1_above_mean	The proportion of all returns above 2 meters.
tpha_pred	The number of predicted tree tops divided by the area of the stand in hectares.

A.3.2 Summary of Small Area Mean Squared Error Estimates

Table A 10: Across all non-synthetic and synthetic predictions, the median error components, the median relative shares, as well as median $\widehat{RMSE}$ values are reported.

		Non-Synthetic							***Synthetic***					***Direct***
u_1	u_2	g_1	$g_1(\%)$	g_2	$g_2(\%)$	g_3	$g_3(\%)$	$\widehat{RMSE}$	$g_1(\%)$	$g_1(\%)$	g_2	$g_2(\%)$	$\widehat{RMSE}$	$\widehat{RMSE}$
VOL	--	119.05	69.82%	29.40	17.24%	20.04	11.75%	13.06	137.08	83.19%	27.69	16.81%	12.84	30.28
	BA	62.31	58.88%	26.21	24.76%	17.96	16.97%	10.29	75.12	78.24%	20.89	21.76%	9.80	
	DEN	130.02	68.64%	26.18	13.82%	22.11	11.67%	13.76	169.44	86.03%	27.50	13.97%	14.03	
	HT	151.99	68.77%	27.26	12.33%	21.53	9.74%	14.87	206.27	87.27%	30.10	12.73%	15.37	
	DIA	203.43	75.67%	27.95	10.40%	28.37	10.55%	16.40	269.97	88.74%	34.27	11.26%	17.44	
BA	--	2.75	69.81%	0.48	12.19%	0.58	14.61%	1.99	3.78	85.20%	0.66	14.80%	2.11	3.18
	VOL	1.59	66.14%	0.34	14.17%	0.31	12.87%	1.55	4.55	89.82%	0.52	10.18%	2.25	
	DEN	3.32	71.80%	0.31	6.81%	0.49	10.55%	2.15	7.00	91.09%	0.68	8.91%	2.77	
	HT	3.57	77.51%	0.40	8.71%	0.60	13.07%	2.14	5.57	89.45%	0.66	10.55%	2.49	
	DIA	4.05	80.03%	0.36	7.02%	0.64	12.65%	2.25	7.14	90.88%	0.72	9.12%	2.80	
DEN	--	1065.76	75.14%	244.66	17.25%	91.96	6.48%	37.66	1269.40	53.06%	1123.04	46.94%	48.91	82.00
	VOL	1299.64	72.49%	184.80	10.31%	110.24	6.15%	42.34	1715.19	74.21%	596.15	25.79%	48.08	
	BA	938.52	71.47%	195.63	14.90%	84.01	6.40%	36.24	1198.27	72.08%	464.20	27.92%	40.77	
	HT	883.69	68.08%	195.01	15.02%	110.11	8.48%	36.03	1258.95	66.08%	646.16	33.92%	43.65	
	DIA	696.12	71.95%	165.37	17.09%	99.71	10.31%	31.10	880.81	63.79%	500.06	36.21%	37.16	
HT	--	0.24	74.57%	0.04	11.32%	0.04	13.83%	0.57	0.37	83.46%	0.07	16.54%	0.66	0.82
	VOL	0.23	75.30%	0.03	10.69%	0.04	12.88%	0.55	0.39	86.16%	0.06	13.84%	0.67	
	BA	0.25	76.73%	0.03	9.66%	0.04	13.23%	0.57	0.41	86.39%	0.06	13.61%	0.69	
	DEN	0.24	76.33%	0.03	7.89%	0.03	10.84%	0.56	0.49	88.00%	0.07	12.00%	0.74	
	DIA	0.15	65.96%	0.03	12.14%	0.04	15.20%	0.48	0.32	85.71%	0.05	14.29%	0.61	
DIA	--	1.16	73.60%	0.15	9.83%	0.21	13.31%	1.25	1.63	66.83%	0.81	33.17%	1.56	2.00
	VOL	1.31	72.97%	0.14	7.71%	0.23	12.86%	1.34	2.21	77.24%	0.65	22.76%	1.69	
	BA	1.28	72.63%	0.14	7.84%	0.23	12.85%	1.33	2.10	77.77%	0.60	22.23%	1.64	
	DEN	0.55	65.69%	0.12	14.23%	0.14	16.51%	0.91	0.75	70.00%	0.32	30.00%	1.03	
	HT	0.73	67.90%	0.14	13.16%	0.18	16.64%	1.04	1.16	71.48%	0.46	28.52%	1.27	

A.3.3 Extension of $G_{3\alpha}$ for Aggregations of Multivariate Predictions

For the MVFH, Benavent and Morales (2016) provide derivations of the $\boldsymbol{G}_3$ component for the special case of individual area-level predictions (i.e. $\boldsymbol{G}_{3i}$). However, it may be of interest to produce this error component for arbitrary aggregations of multivariate predictions, as in the case of aggregating several synthetic and non-synthetic area-level predictions (i.e. $\boldsymbol{G}_{3\tau}$). We provide a proof that considers this more general case.

Before proceeding it is necessary to define notation that describes the asymptotic behavior of scalar and matrix-valued functions. Let $f(D_S) = O(g(D_S))$ represent a scalar-valued function such that $\lim_{D_S \to \infty} |\frac{f(D_S)}{g(D_S)}| < \infty$. Similarly, $f(D_S) = o(g(D_S))$ represents a scalar-valued function such that $\lim_{D_S \to \infty} |\frac{f(D_S)}{g(D_S)}| = 0$. When $f(D_S)$ is a matrix-valued function such that the elements of that matrix are $O(g(D_s))$ we reserve the notation $\left[O\left(g(D_s)\right)\right]_{m\ x\ l}$ where m and l are the number of rows and columns of the matrix, respectively.

The $\boldsymbol{G}_{3\alpha}$ error component quantifies the mean squared cross product error between the EBLUP and the BLUP:

$$\boldsymbol{G}_{3\alpha} \approx E[(\widehat{\boldsymbol{\theta}}_{\alpha}^{E} - \widehat{\boldsymbol{\theta}}_{\alpha}^{B})\,(\widehat{\boldsymbol{\theta}}_{\alpha}^{E} - \widehat{\boldsymbol{\theta}}_{\alpha}^{B})^{T}]$$

where

$$\widehat{\boldsymbol{\theta}}_{\alpha}^{B} = \boldsymbol{L}_{\alpha}\widehat{\boldsymbol{\beta}}^{B} + \boldsymbol{M}_{\alpha}\widehat{\boldsymbol{v}}^{B}$$

and $\widehat{\boldsymbol{\beta}}^{B} = (\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{y}_S$ and $\widehat{\boldsymbol{v}}^{B} = \boldsymbol{V}_u\boldsymbol{Z}_S^T\boldsymbol{V}_S^{-1}(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}^{B})$ and $\boldsymbol{Z}_S$ is a $RD_S\ x\ DR$ matrix, i.e. $\boldsymbol{Z}$ with rows removed that correspond to observations in unsampled stands. Note that $\boldsymbol{G}_{3\alpha}$ only approximates this expectation. This is due to

the omission of higher-order terms in a Taylor series expansion that is used for the approximation. Letting $f(\boldsymbol{\delta}) = \widehat{\boldsymbol{\theta}}_{\alpha}^{B}$ and $f(\widehat{\boldsymbol{\delta}}) = \widehat{\boldsymbol{\theta}}_{\alpha}^{E}$ the Taylor series expansion of $f(\widehat{\boldsymbol{\delta}})$ around $\boldsymbol{\delta}$ allows for the following approximation:

$$(\widehat{\boldsymbol{\theta}}_{\alpha}^{E} - \widehat{\boldsymbol{\theta}}_{\alpha}^{B})\,(\widehat{\boldsymbol{\theta}}_{\alpha}^{E} - \widehat{\boldsymbol{\theta}}_{\alpha}^{B})^{T} \approx \boldsymbol{S}_{\alpha}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{T}\boldsymbol{S}_{\alpha}^{T}$$

where $\boldsymbol{S}_{\alpha} = \frac{\partial \widehat{\boldsymbol{\theta}}_{\alpha}^{B}}{\partial \widehat{\boldsymbol{\delta}}}$ is a $R\ x\ R$ matrix of partial derivatives. Letting $\boldsymbol{s}_{\alpha}^{(k)} = \frac{\partial \widehat{\boldsymbol{\theta}}_{\alpha}^{B}}{\widehat{\delta}_{k}}$ we obtain the summation

$$\boldsymbol{S}_{\alpha}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{T}\boldsymbol{S}_{\alpha}^{T} = \sum_{i=1}^{R}\sum_{j=1}^{R}(\hat{\delta}_{i} - \delta_{i})(\hat{\delta}_{j} - \delta_{j})\boldsymbol{s}_{\alpha}^{(i)}\boldsymbol{s}_{\alpha}^{(j)}$$

and the expectation:

$$E[\boldsymbol{S}_{\alpha}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{T}\boldsymbol{S}_{\alpha}^{T}] = \sum_{i=1}^{R}\sum_{j=1}^{R}E\left[(\hat{\delta}_{i} - \delta_{i})(\hat{\delta}_{j} - \delta_{j})\boldsymbol{s}_{\alpha}^{(i)}\boldsymbol{s}_{\alpha}^{(j)}\right].$$

The remaining tasks are to attain the summation term on right-hand side of the above expression, which constitutes $\boldsymbol{G}_{3\alpha}$, and to establish the asymptotic behavior of any neglected terms.

Theorem 1

$$E\left[(\hat{\delta}_{i} - \delta_{i})(\hat{\delta}_{j} - \delta_{j})\boldsymbol{s}_{\alpha}^{(i)}\boldsymbol{s}_{\alpha}^{(j)}\right] = cov(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)})\boldsymbol{C}^{(i)}\boldsymbol{V}_{\alpha}\boldsymbol{C}^{(j)T} + [O(D_{S}^{-1})]_{R\ x\ R}$$

Proof.

Via Lemma 1 we attain that the derivatives $\boldsymbol{s}_{\alpha}^{(i)} = \left[s_{\alpha 1}^{(i)}, \ldots, s_{\alpha R}^{(i)}\right]^{T}$ and $\boldsymbol{s}_{\alpha}^{(j)}$, defined analogously, are linear functions of the sampled model errors $\boldsymbol{w} = \boldsymbol{v}_{S} + \boldsymbol{e}_{S}$ such that

$$s_{\alpha r}^{(i)} = \left(\boldsymbol{F}_{\alpha r}^{(i)} + \boldsymbol{C}_{\alpha r}^{(i)}\right)^{T}\boldsymbol{w}$$

where $\boldsymbol{F}^{(i)} = \left[\boldsymbol{F}_{\alpha 1}^{(i)T}, \dots, \boldsymbol{F}_{\alpha R}^{(i)T}\right]^T$ and $\boldsymbol{C}^{(i)} = \left[\boldsymbol{C}_{\alpha 1}^{(i)T}, \dots, \boldsymbol{C}_{\alpha R}^{(i)T}\right]^T$. Furthermore, we assume that the estimator of the variance components $\widehat{\boldsymbol{\delta}}$ is an unbiased, consistent and translation invariant estimator of the form $\hat{\delta}_i = k + \boldsymbol{w}^T \boldsymbol{A}_i \boldsymbol{w}$ where $\boldsymbol{A}_i = diag\{[O(D_S{}^{-1})]_{RxR \dots}, [O(D_S{}^{-1})]\} + O(D_S^{-2})$ which is a standard result of, e.g., the restricted maximum likelihood estimator of $\hat{\delta}_i$ used in section 3.2 (Corbeil and Searle 1976). Plugging this into the difference between the estimated and actual variance component we obtain $\hat{\delta}_i - \delta_i = \boldsymbol{w}^T \boldsymbol{A}_i \boldsymbol{w} - E[\boldsymbol{w}^T \boldsymbol{A}_i \boldsymbol{w}] = q_i - E[q_i]$. Estimators of this form and under the assumption that $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{V_s})$ allows for direct application of Lemma 2. By letting $\boldsymbol{\lambda}_1 = \boldsymbol{F}_{\alpha r}^{(1)} + \boldsymbol{C}_{\alpha r}^{(1)}$, $\boldsymbol{\lambda}_2 = \boldsymbol{F}_{\alpha r'}^{(2)} + \boldsymbol{C}_{\alpha r'}^{(2)}$, $s_1 = s_{\alpha r}^{(i)}$, $s_2 = s_{\alpha r'}^{(i)}$, $q_1 = \boldsymbol{w}^T \boldsymbol{A}_i \boldsymbol{w}$ and $q_2 = \boldsymbol{w}^T \boldsymbol{A}_j \boldsymbol{w}$ we apply the result of Lemma 2 and obtain:

$$E\left[s_{\alpha r}^{(i)} s_{\alpha r'}^{(j)} \left(\hat{\delta}_i - \delta_i\right)\left(\hat{\delta}_i - \delta_i\right)\right]$$

$$= cov\left(s_{\alpha r}^{(i)}, s_{\alpha r'}^{(j)}\right) cov\left(\hat{\delta}_i, \hat{\delta}_j\right)$$

$$+ \ 8\left(\boldsymbol{F}_{\alpha r}^{(i)} + \boldsymbol{C}_{\alpha r}^{(i)}\right)\boldsymbol{V}_S \boldsymbol{A}_i \boldsymbol{V}_S \boldsymbol{A}_j \boldsymbol{V}_S \left(\boldsymbol{F}_{\alpha r'}^{(j)} + \boldsymbol{C}_{\alpha r'}^{(j)}\right)^T$$

In matrix and vector notation we obtain:

$$E\left[\left(\hat{\delta}_i - \delta_i\right)\left(\hat{\delta}_j - \delta_j\right)\boldsymbol{s}_\alpha^{(i)} \boldsymbol{s}_\alpha^{(j)}\right]$$

$$= cov\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) cov\left(\hat{\delta}_i, \hat{\delta}_j\right)$$

$$+ 8\left(\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)}\right)\boldsymbol{V}_S \boldsymbol{A}_i \boldsymbol{V}_S \boldsymbol{A}_j \boldsymbol{V}_S \left(\boldsymbol{F}^{(j)} + \boldsymbol{C}^{(j)}\right)^T.$$

Where $\boldsymbol{F}^{(j)}$ and $\boldsymbol{C}^{(j)}$ are given in Lemma 1. From Lemma 3 we obtain:

$$cov\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) = \boldsymbol{C}^{(i)} \boldsymbol{V}_S \boldsymbol{C}^{(j)T} + [O(D_S^{-1})]_{RxR}$$

And from Lemma 5 we obtain:

$$8\left(\boldsymbol{F}^{(i)}+\boldsymbol{C}^{(i)}\right)\boldsymbol{V}_S\boldsymbol{A}_i\boldsymbol{V}_S\boldsymbol{A}_j\boldsymbol{V}_S\left(\boldsymbol{F}^{(j)}+\boldsymbol{C}^{(j)}\right)^T=[O(D_S^{-2})]_{RxR}$$

Therefore:

$$E\left[\left(\delta_i-\hat{\delta}_i\right)\left(\delta_j-\hat{\delta}_j\right)\boldsymbol{s}_\alpha^{(i)}\boldsymbol{s}_\alpha^{(j)}\right]=cov\left(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}\right)\boldsymbol{C}^{(i)}\boldsymbol{V}_\alpha\boldsymbol{C}^{(j)T}+[O(D_S^{-1})]_{R\,x\,R}$$

With the expectation defined, and the asymptotic properties of the neglected terms established, we obtain:

$$\boldsymbol{G}_{3\alpha}(\boldsymbol{\delta})=\sum_{k=1}^{m}\sum_{l=1}^{m}Cov\left(\hat{\delta}_k,\hat{\delta}_l\right)\boldsymbol{C}_\alpha^k\boldsymbol{V}_\alpha\boldsymbol{C}_\alpha^l$$

Lemma 1

$$\boldsymbol{s}_\alpha^{(i)}=[\boldsymbol{F}^{(i)}+\boldsymbol{C}^{(i)}]\boldsymbol{w}$$

where $\boldsymbol{w}=\boldsymbol{Z}_S^T\boldsymbol{v}_S+\boldsymbol{e}_S$ is the error of the model, including sampling error and $\boldsymbol{F}^{(i)}$ and $\boldsymbol{C}^{(i)}$ are matrices attained below.

Proof.

Let $\boldsymbol{R}=\boldsymbol{V}_u\boldsymbol{Z}_S^T\boldsymbol{V}_S^{-1}$. By letting $\boldsymbol{Q}=\left(\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{X}_S\right)^{-1}$, $\boldsymbol{w}=\boldsymbol{Z}_S^T\boldsymbol{v}_S+\boldsymbol{e}_S$ and rearranging $\widehat{\boldsymbol{\theta}}_\alpha^B$ we obtain:

$$\widehat{\boldsymbol{\theta}}_\alpha^B=\boldsymbol{L}_\alpha\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta}+\boldsymbol{L}_\alpha\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{w}+\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{X}_S\boldsymbol{\beta}+\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{w}-$$

$$\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta}-\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S\boldsymbol{V}_S^{-1}\boldsymbol{w}.\text{e}$$

Noting that $\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{X}_S=\boldsymbol{I}$ we obtain the simplification:

$$\widehat{\boldsymbol{\theta}}_\alpha^B=\boldsymbol{L}_\alpha\boldsymbol{\beta}+\boldsymbol{L}_\alpha\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{w}+\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{w}-\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S\boldsymbol{V}_S^{-1}\boldsymbol{w}.$$

Letting $\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}$ we obtain:

$$\widehat{\boldsymbol{\theta}}_\alpha^B=\boldsymbol{L}_\alpha\boldsymbol{\beta}+\boldsymbol{L}_\alpha\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}\boldsymbol{w}+\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{A}\boldsymbol{w}$$

and

$$\boldsymbol{s}_\alpha^{(i)}=\boldsymbol{L}_\alpha\boldsymbol{Q}\boldsymbol{X}_S^T\frac{\partial\boldsymbol{V}_S^{-1}}{\partial\delta_i}\boldsymbol{A}\boldsymbol{w}-\boldsymbol{M}_\alpha\boldsymbol{R}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\frac{\partial\boldsymbol{V}^{-1}}{\partial\delta_i}\boldsymbol{A}\boldsymbol{w}+\boldsymbol{M}_\alpha\frac{\partial\boldsymbol{R}}{\partial\delta_i}\boldsymbol{A}\boldsymbol{w}.$$

Reorganizing the expression we obtain

$$\boldsymbol{s}_{\alpha}^{(i)} = \left[(\boldsymbol{L}_{\alpha} - \boldsymbol{M}_{\alpha}\boldsymbol{R}\boldsymbol{X}_{S})\boldsymbol{Q}\boldsymbol{X}_{S}^{T}\frac{\partial \boldsymbol{V}_{S}^{-1}}{\partial \delta_i}\,\boldsymbol{A} - \boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_{S}\boldsymbol{Q}\boldsymbol{X}_{S}^{T}\boldsymbol{V}_{S}^{-1} + \boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}\right] w$$

$$= [\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)}]\boldsymbol{w}$$

where $\boldsymbol{F}^{(i)} = (\boldsymbol{L}_{\alpha} - \boldsymbol{M}_{\alpha}\boldsymbol{R}\boldsymbol{X}_{S})\boldsymbol{Q}\boldsymbol{X}_{S}^{T}\frac{\partial \boldsymbol{V}_{S}^{-1}}{\partial \delta_i}\,\boldsymbol{A} - \boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_{S}\boldsymbol{Q}\boldsymbol{X}_{S}^{T}\boldsymbol{V}_{S}^{-1}$ and $\boldsymbol{C}^{(i)} = \boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}$.

Lemma 2

This Lemma is identical to Lemma 3 provided in Benavent and Morales (2016), however, without loss of generality, the input vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ change when applying the result of this Lemma in Theorem 1. Under the assumption that $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{V}_s)$ and Let $s_1 = \boldsymbol{\lambda}_1^T\boldsymbol{w}$, $s_2 = \boldsymbol{\lambda}_2^T\boldsymbol{w}$, $q_1 = \boldsymbol{w}^T\boldsymbol{A}_1\boldsymbol{w}$ and $q_2 = \boldsymbol{w}^T\boldsymbol{A}_2\boldsymbol{w}$ where $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are non-stochastic vectors and matrices we obtain:

$$E[s_1 s_2 (q_1 - E[q_1])(q_2 - E[q_2])]$$

$$= cov(q_1, q_2)cov(s_1, s_2) + 8\boldsymbol{\lambda}_1\boldsymbol{V}_S\boldsymbol{A}_1\boldsymbol{V}_S\boldsymbol{A}_2\boldsymbol{\lambda}_2$$

Proof.

Applying Lemma A.2 of (Prasad and Rao 1990) we obtain:

$$E[s_1 s_2 (q_1 - E[q_1])(q_2 - E[q_2])] = \boldsymbol{\lambda}_1^T E[\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{A}_1\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{A}_2\boldsymbol{w})\boldsymbol{w}^T]\boldsymbol{\lambda}_2 -$$

$$E[q_1]\boldsymbol{\lambda}_1^T E[\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{A}_2\boldsymbol{w})\boldsymbol{w}^T]\boldsymbol{\lambda}_2 - E[q_2]\boldsymbol{\lambda}_1^T E[\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{A}_1\boldsymbol{w})\boldsymbol{w}^T]\boldsymbol{\lambda}_2 +$$

$$E[q_1]E[q_2]\boldsymbol{\lambda}_1^T\boldsymbol{V}_S\boldsymbol{\lambda}_2.$$

Lemma A.1 of (Prasad and Rao 1990) provides expansions of the terms $E[\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{A}_1\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{A}_2\boldsymbol{w})\boldsymbol{w}^T]$ and $E[\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{A}_i\boldsymbol{w})\boldsymbol{w}^T]$ that are omitted here for brevity. Applying these expansions we obtain:

$$E[s_1 s_2 (q_1 - E[q_1])(q_2 - E[q_2])]$$

$$= E[q_1]E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2 + 2E[q_1]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V} \boldsymbol{\lambda}_2 + 2E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_1 \boldsymbol{V} \boldsymbol{\lambda}_2$$

$$+ 8\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S \boldsymbol{\lambda}_2 + 2tr(\boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S)\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2$$

$$- E[q_1]E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2 - 2E[q_1]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S \boldsymbol{\lambda}_2 - E[q_1]E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2$$

$$- 2E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{\lambda}_2 + E[q_1]E[q_2]\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2$$

$$= 2tr(\boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S)\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2 + 8\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S \boldsymbol{\lambda}_2$$

Noting that $cov(q_1, q_2) = 2tr(\boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S)$ and $cov(s_1, s_2) = \boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{\lambda}_2$ we obtain:

$$E[s_1 s_2 (q_1 - E[q_1])(q_2 - E[q_2])] = cov(q_1, q_2)cov(s_1, s_2) + 8\boldsymbol{\lambda}_1^T \boldsymbol{V}_S \boldsymbol{A}_1 \boldsymbol{V}_S \boldsymbol{A}_2 \boldsymbol{V}_S \boldsymbol{\lambda}_2$$

Lemma 3

$$cov\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) = \boldsymbol{C}^{(i)} \boldsymbol{V}_S \boldsymbol{C}^{(j)T} + [O(D_S^{-1})]_{R\,x\,R}$$

Proof.

Via Lemma 1 we obtain $\boldsymbol{s}^{(i)} = \left[\boldsymbol{C}^{(i)} + \boldsymbol{F}^{(i)}\right]\boldsymbol{w}$ we obtain:

$$cov\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) = cov\left(\left[\boldsymbol{C}^{(i)} + \boldsymbol{F}^{(i)}\right]\boldsymbol{w}, \left[\boldsymbol{C}^{(i)} + \boldsymbol{F}^{(i)}\right]\boldsymbol{w}\right)$$

$$= \left[\boldsymbol{C}^{(i)} + \boldsymbol{F}^{(i)}\right] cov(\boldsymbol{w}) \left[\boldsymbol{C}^{(i)} + \boldsymbol{F}^{(i)}\right]^T$$

$$= \boldsymbol{C}^{(i)} \boldsymbol{V}_S \boldsymbol{C}^{(j)T} + \boldsymbol{C}^{(i)} \boldsymbol{V}_S \boldsymbol{F}^{(j)T} + \boldsymbol{F}^{(i)} \boldsymbol{V}_S \boldsymbol{C}^{(j)T} + \boldsymbol{F}^{(i)} \boldsymbol{V}_S \boldsymbol{F}^{(j)T}$$

From Lemma 4 we note that $\boldsymbol{F}^{(i)} = [O(D_S^{-1})]_{RxRD_S}$, $\boldsymbol{C}^{(i)} = [O(1)]_{RxRD_S}$ and similarly for $\boldsymbol{F}^{(j)}$ and $\boldsymbol{C}^{(j)}$. Additionally, assuming the elements of $\boldsymbol{V}_S$ are bounded from above, $\boldsymbol{V}_S = [O(1)]_{RD_S\,x\,RD_S}$. This suggests the latter three terms in the above expression have elements of $O(D_S^{-1})$. Applying the asymptotic behavior of these matrices provides:

$$cov\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) = \boldsymbol{C}^{(i)} \boldsymbol{V}_S \boldsymbol{C}^{(j)T} + [O(D_S^{-1})]_{RxR}$$

Lemma 4

$$\boldsymbol{C}^{(i)} = [O(1)]_{RxRD_S}, \boldsymbol{F}^{(i)} = [O(D_S^{-1})]_{RxRD_S}$$

Proof.

Via Lemma 1 we obtain:

$$\boldsymbol{C}^{(i)} = \boldsymbol{M}_\alpha \frac{\partial \boldsymbol{R}}{\partial \delta_i} = \boldsymbol{M}_\alpha [\boldsymbol{W}_i \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1} - \boldsymbol{V}_u \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{W}_i] \boldsymbol{V}_S^{-1}$$

where $\boldsymbol{W}_i = \frac{\partial \boldsymbol{V}_S}{\partial \delta_i} = \underset{1 \le i \le D_S}{diag}([O(1)_{RxR}])_{RD_S\, x\, RD_S}$. $\boldsymbol{W}_i \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1}$ is a block-diagonal matrix whose elements are, in the worst case, $O(1)$. Similar results are obtained for $\boldsymbol{V}_u \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{W}_i$, noting that $\boldsymbol{V}_S^{-1} = \underset{1 \le i \le D_S}{diag}([O(1)_{RxR}])_{RD_S\, x\, RD_S}$. This establishes $[\boldsymbol{W}_i \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1} - \boldsymbol{V}_u \boldsymbol{Z}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{W}_i] \boldsymbol{V}_S^{-1}$ as a block-diagonal matrix whose block-specific elements are $O(1)$. We assume that $\boldsymbol{M}_\alpha = [O(1)]_{RxRD}$. The elements of $\boldsymbol{C}^{(i)}$ are therefore $R \cdot O(1) = O(1)$.

Via Lemma 1 we obtain:

$$\boldsymbol{F}^{(i)} = (\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R} \boldsymbol{X}_S) \boldsymbol{Q} \boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i} \boldsymbol{A} - \boldsymbol{M}_\alpha \frac{\partial \boldsymbol{R}}{\partial \delta_i} \boldsymbol{X}_S \boldsymbol{Q} \boldsymbol{X}_S^T \boldsymbol{V}_S^{-1}$$

Considering the first summand, $(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R} \boldsymbol{X}_S) \boldsymbol{Q} \boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i} \boldsymbol{A}$**,** we establish that $\boldsymbol{R}$ is a block-diagonal matrix, with diagonal blocks of size R whose elements are $O(1)$ and 0 otherwise. Noting this we obtain

$$(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R} \boldsymbol{X}_S) = [O(D_S)]_{Rxp}$$

Turning to $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{X}_S \boldsymbol{Q} \boldsymbol{X}_S^T \boldsymbol{V}_S^{-1})$**,** from Benavent and Morales (2016) we obtain $\boldsymbol{X}_S \boldsymbol{Q} \boldsymbol{X}_S^T = [O(D_S^{-1})]_{RD_S x RD_S}$. Therefore

$$\boldsymbol{A} = ([O(1)]_{RD_S x\, RD_S} - [O(D_S^{-1})]_{RD_S x RD_S} [\underset{1 \le i \le D_S}{diag}([O(1)_{RxR}])]_{RD_S\, x\, RD_S})$$

$$\boldsymbol{A} = [_{1\le i\le D_S}^{diag}([O(D_S^{-1})_{RxR}])]_{RD_S\, x\, RD_S}$$

For the first summand we obtain

$$(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R}\boldsymbol{X}_S)\boldsymbol{Q}\boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i}\, \boldsymbol{A}$$

$$= [O(D_S)]_{Rxp}[O(D_S^{-1})]_{pxp}[O(1)]_{pxRD_S}[_{1\le i\le D_S}^{diag}([O(1)_{RxR}])]_{RD_S\, x\, RD_S}$$

$$[_{1\le i\le D_S}^{diag}([O(D_S^{-1})_{RxR}])]_{RD_S\, x\, RD_S}$$

The block-diagonality, where the dimensions of each block are not dependent on D_S, assists in reducing the expression to

$$(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R}\boldsymbol{X}_S)\boldsymbol{Q}\boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i}\, \boldsymbol{A} =$$

$$[O(D_S)]_{Rxp}[O(D_S^{-1})]_{pxp}[O(1)]_{pxRD_S}[_{1\le i\le D_S}^{diag}([O(D_S^{-1})_{RxR}])]_{RD_S\, x\, RD_S}.$$

The first three terms can also be simplified, noting that $O(D_S)\cdot\ O(D_S^{-1}) = O(1)$, obtaining:

$$(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R}\boldsymbol{X}_S)\boldsymbol{Q}\boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i}\, \boldsymbol{A} = [O(1)]_{RxRD_S}[_{1\le i\le D_S}^{diag}([O(D_S^{-1})_{RxR}])]_{RD_S\, x\, RD_S}.$$

Therefore

$$(\boldsymbol{L}_\alpha - \boldsymbol{M}_\alpha \boldsymbol{R}\boldsymbol{X}_S)\boldsymbol{Q}\boldsymbol{X}_S^T \frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i}\, \boldsymbol{A} = [O(D_S^{-1})]_{RxRD_S}.$$

For the second summand, $\boldsymbol{M}_\alpha \frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}$, we obtain:

$$\boldsymbol{M}_\alpha \frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1}$$

$$= [O(1)]_{RxRD_S}[_{1\le i\le D_S}^{diag}([O(D_S^{-1})_{RxR}])]_{RD_S\, x\, RD_S}[O(1)]_{RD_S xRD_S}$$

Again, leveraging the block-diagonality of $\boldsymbol{V}_S^{-1}$ we obtain

$$\boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1} = [O(D_S^{-1})]_{RxRD_S}$$

Noting that $O(D_S^{-1}) - O(D_S^{-1}) = O(D_S^{-1})$, we finally obtain:

$$\boldsymbol{F}^{(i)} = (\boldsymbol{L}_{\alpha} - \boldsymbol{M}_{\alpha}\boldsymbol{R}\boldsymbol{X}_S)\boldsymbol{Q}\boldsymbol{X}_S^T\frac{\partial \boldsymbol{V}_S^{-1}}{\partial \delta_i}\,\boldsymbol{A} - \boldsymbol{M}_{\alpha}\frac{\partial \boldsymbol{R}}{\partial \delta_i}\boldsymbol{X}_S\boldsymbol{Q}\boldsymbol{X}_S^T\boldsymbol{V}_S^{-1} = [O(D_S^{-1})]_{RxRD_S}$$

Lemma 5

$$8\big(\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)}\big)\boldsymbol{V}_S\boldsymbol{A}_i\boldsymbol{V}_S\boldsymbol{A}_j\boldsymbol{V}_S\big(\boldsymbol{F}^{(j)} + \boldsymbol{C}^{(j)}\big)^T = [O(D_S^{-2})]_{RxR}$$

Proof.

Via Lemma A.3 of (Prasad and Rao 1990) we obtain $\boldsymbol{V}_S\boldsymbol{A}_i\boldsymbol{V}_S\boldsymbol{A}_j\boldsymbol{V}_S = [O(D^{-2})]_{RD_S\,x\,RD_S}$. Via Lemma 4 we obtain $\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)} = [O(D_S^{-1})]_{R\,x\,RD_S}$. In asymptotic notation we obtain:

$$8\big(\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)}\big)\boldsymbol{V}_S\boldsymbol{A}_i\boldsymbol{V}_S\boldsymbol{A}_j\boldsymbol{V}_S\big(\boldsymbol{F}^{(j)} + \boldsymbol{C}^{(j)}\big)^T$$

$$= 8[O(D_S^{-1})]_{R\,x\,RD_S}[O(D_S^{-2})]_{RD_S\,x\,RD_S}[O(D_S^{-1})]_{RD_S\,x\,R}$$

$$8\big(\boldsymbol{F}^{(i)} + \boldsymbol{C}^{(i)}\big)\boldsymbol{V}_S\boldsymbol{A}_i\boldsymbol{V}_S\boldsymbol{A}_j\boldsymbol{V}_S\big(\boldsymbol{F}^{(j)} + \boldsymbol{C}^{(j)}\big)^T = [O(D_S^{-2})]_{RxR}$$