Integrating remotely sensed data into forest resource inventories: the impact of model and variable selection on estimates of precision

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Forstwissenschaften und Waldökologie
der Georg–August–Universität Göttingen

vorgelegt von
Philip Henrich Mundhenk
geboren in Hamburg

Göttingen, 2014

Erster Gutachter: Prof. Dr. Christoph Kleinn
Zweiter Gutachter: Prof. Dr. Thomas Kneib
Dritter Gutachter: Dr. Steen Magnussen

Tag der mündlichen Prüfung: 26. Mai 2014

Abstract

The past two decades have demonstrated a great potential for airborne Light Detection and Ranging (LiDAR) data to improve the efficiency of forest resource inventories (FRIs). In order to make efficient use of LiDAR data in FRIs, the data need to be related to observations taken in the field. Various modeling techniques are available that enable a data analyst to establish a link between the two data sources. While the choice for a modeling technique may have negligible effects on point estimates, different model techniques may deliver different estimates of precision.

This study investigated the impact of various model and variable selection procedures on estimates of precision. The focus was on LiDAR applications in FRIs. The procedures considered included stepwise variable selection procedures such as the Akaike Information Criterion (AIC), the corrected Akaike Information Criterion (AICc), and the Bayesian (or Schwarz) Information Criterion. Variables have also been selected based on the condition number of the matrix of covariates (i.e., LiDAR metrics) and the variance inflation factor. Other modeling techniques considered in this study were ridge regression, the least absolute shrinkage and selection operator (Lasso), partial least squares regression, and the random forest algorithm. Stepwise variable selection procedures have been considered in both, the (design-based) model-assisted, as well as in the model-based (or model-dependent) inference framework. All other techniques were investigated only for the model-assisted approach.

In a comprehensive simulation study, the effects of the different modeling techniques on the precision of population parameter estimates (mean aboveground biomass per hectare) were investigated. Five different datasets were used. Three artificial datasets were simulated; two further datasets were based on FRI data from Canada and Norway. Canonical vine copulas were employed to create synthetic populations from the FRI data. From all populations simple random samples of different size were repeatedly drawn and the mean and variance of the mean were estimated for each sample. While for the model-based approach only a single variance estimator was investigated, for the model-assisted approach three alternative estimators were examined.

The results of the simulation studies suggest that blind application of stepwise variable selection procedures lead to overly optimistic estimates of precision in LiDAR-assisted FRIs. The effects were severe for small sample sizes ($n = 40$ and $n = 50$). For large samples ($n = 400$) overestimation of precision was negligible. Good performance in terms of empirical standard errors and coverage rates were obtained for ridge regression, Lasso, and the random forest algorithm. This study concludes that the use of the latter three modeling techniques may prove useful in future LiDAR-assisted FRIs.

Zusammenfassung

Die letzten zwanzig Jahre haben gezeigt, dass die Integration luftgestützter Lasertechnologien (Light Detection and Ranging; LiDAR) in die Erfassung von Waldressourcen dazu beitragen kann, die Genauigkeit von Schätzungen zu erhöhen. Um diese zu ermöglichen, müssen Feldaten mit LiDAR-Daten kombiniert werden. Diverse Techniken der Modellierung bieten die Möglichkeit, diese Verbindung statistisch zu beschreiben. Während die Wahl der Methode in der Regel nur geringen Einfluss auf Punktschätzer hat, liefert sie unterschiedliche Schätzungen der Genauigkeit.

In der vorliegenden Studie wurde der Einfluss verschiedener Modellierungstechniken und Variablenauswahl auf die Genauigkeit von Schätzungen untersucht. Der Schwerpunkt der Arbeit liegt hierbei auf LiDAR Anwendungen im Rahmen von Waldinventuren. Die Methoden der Variablenauswahl, welche in dieser Studie berücksichtigt wurden, waren das Akaike Informationskriterium (AIC), das korrigierte Akaike Informationskriterium (AICc), und das bayesianische (oder Schwarz) Informationskriterium. Zudem wurden Variablen anhand der Konditionsnummer und des Varianzinflationsfaktors ausgewählt. Weitere Methoden, die in dieser Studie Berücksichtigung fanden, umfassen Ridge Regression, der *least absolute shrinkage and selection operator* (Lasso), und der Random Forest Algorithmus. Die Methoden der schrittweisen Variablenauswahl wurden sowohl im Rahmen der Modell-assistierten als auch der Modell-basierten Inferenz untersucht. Die übrigen Methoden wurden nur im Rahmen der Modell-assistierten Inferenz untersucht.

In einer umfangreichen Simulationsstudie wurden die Einflüsse der Art der Modellierungsmethode und Art der Variablenauswahl auf die Genauigkeit der Schätzung von Populationsparametern (oberirdische Biomasse in Megagramm pro Hektar) ermittelt. Hierzu wurden fünf unterschiedliche Populationen genutzt. Drei künstliche Populationen wurden simuliert, zwei weitere basierten auf in Kanada und Norwegen erhobenen Waldinveturdaten. *Canonical vine copulas* wurden genutzt um synthetische Populationen aus diesen Waldinventurdaten zu generieren. Aus den Populationen wurden wiederholt einfache Zufallsstichproben gezogen und für jede Stichprobe wurden der Mittelwert und

die Genauigkeit der Mittelwertschätzung geschäzt. Während für das Modell-basierte Verfahren nur ein Varianzschätzer untersucht wurde, wurden für den Modell-assistierten Ansatz drei unterschiedliche Schätzer untersucht.

Die Ergebnisse der Simulationsstudie zeigten, dass das einfache Anwenden von schrittweisen Methoden zur Variablenauswahl generell zur Überschätzung der Genauigkeiten in LiDAR unterstützten Waldinventuren führt. Die verzerrte Schätzung der Genauigkeiten war vor allem für kleine Stichproben ($n = 40$ und $n = 50$) von Bedeutung. Für Stichproben von größerem Umfang ($n = 400$), war die Überschätzung der Genauigkeit vernachlässigbar. Gute Ergebnisse, im Hinblick auf Deckungsraten und empirischem Standardfehler, zeigten Ridge Regression, Lasso und der Random Forest Algorithmus. Aus den Ergebnissen dieser Studie kann abgeleitet werden, dass die zuletzt genannten Methoden in zukünftige LiDAR unterstützten Waldinventuren Berücksichtigung finden sollten.

Acknowledgements

First of all I would like to thank my supervisor Prof. Dr. Christoph Kleinn for his support and guidance during my time at the Chair of Forest Inventory and Remote Sensing. His trust in me made this doctoral thesis possible. I would also like to thank Dr. Steen Magnussen who contributed significantly in developing research ideas and was of great help during my time in Göttingen and Victoria. One day we will watch a match of the Hamburger SV together.

I would like to thank Prof. Dr. Thomas Kneib for his support. His doctoral students have been of great help, too. I really appreciate their patience with a non-statistician. Thank you Jule!

It would not have been possible to write this doctoral thesis without the support of my colleagues and friends at the Chair of Forest Inventory and Remote Sensing and within the research training group "Scaling Problems in Statistics".

I would also like to thank all colleges who gave me the opportunity to visit them abroad to exchange ideas, to enjoy the rain, or have barbecues: Dr. Jose Javier Corral-Rivas in Durango, Dr. Marco Aurelio González Tagle in Linares, Prof. Dr. Hans-Erik Andersen in Seattle, and Prof. Dr. Timothy Gregoire in New Haven.

The data for this study were provided by Joanne White and Mike Wulder (Pacific Forestry Center, Canadian Forest Service, Canada), and Liviu Ene, Erik Næsset, and Terje Gobakken (Norwegian University of Life Sciences). I really appreciate their willingness to share the data. I am much obliged to Joanne who patiently answered all my questions regarding the LiDAR data.

This research was conducted within the Research Training Group 1644 "Scaling problems in statistics" and would not have been possible without the financial support of the Deutsche Forschungsgemeinschaft (DFG).

I would like to thank my family for their constant support and trust. Finally, I would like to express sincere thanks to my wife Marion.

Contents

List of Figures

List of Tables

Nomenclature

$\bar{y}$	Estimate of the sample mean for y
$\boldsymbol{x}_k$	Vector of ancilary data for element k, $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp}, \ldots, x_{kP})'$
$\mathscr{S}$	Set of all possible samples that can be drawn from a finite population U under a given sample design $p(S)$
μ_y	Parametric mean of the target variable y
π_k	Inclusion probability of element k
J	Number of ancilary variables $j = 1, 2, \ldots, J$
k	Label of elements in the finite population U
N	Finite population size
n	Sample size
$p(S)$	Sample design
S	Sample of size n
s	Standard deviation of $y_{k \in S}$
s^2	Variance of $y_{k \in S}$
SE	Estimated standard error
U	Finite population consisting of $1, 2, \ldots, k, \ldots, N$ elements
$U - S$	Non-sampled set of the finite population U, i.e., $k \in U - S$
AGB	Aboveground biomass
AIC	Akaike Information Criterion
AICc	Corrected Akaike Information Criterion

AVSE	Avregare estimated standard error
BIC	Bayesian Information Criterion
CON	Variable selection based on the condition number
COV	Coverage rate
EFF	Efficiency
ESE	Empirical standard error
FRI	Forest resource inventory
FULL	Saturated regression model
LiDAR	Light Detection and Ranging
NMBU	Norwegian University of Life Sciences
PLSR	Partial least squares regression
RF	Random forest
RSS	Residual sum of squares
SI	Estimator for SRSwoR (no use of auxiliary information)
SRSwoR	Simple random sample without replacement
VIF	Variance inflation factor
VIFB	Variable selection based on the variance inflation factor and best-subset selection

Part I.

Introduction

1. Rationale

1.1. Models in forest resource assessments

In order to manage a forest resource in a sustainable manner, accurate and timely information about the resource is needed. The aim of forest resource assessments (FRIs) is to provide this information in a meaningful, methodological sound and transparent manner. A first step in any survey — including FRIs — is to clearly define its objectives. Once the survey's objectives have been set they need to be translated into measurable goals, that is, a subject matter needs to be translated into a survey problem (Valliant *et al.*, 2013).

The goal of many FRIs is to provide information about one or more population characteristics, such as the total aboveground biomass of trees in a forest, and/or the number of stems per hectare. Before such information can be produced, data need to be collected. These data are usually obtained by direct measurements of one or more attributes of trees, as, for example, the diameter at breast height (DBH), or tree height. However, it generally proves too expensive or otherwise impractical to record attributes of *all* trees in a forest. The information provided by most FRI is, therefore, based on sampling. In a sample survey only a part of the population, i.e., forest, is observed, these observations are then used to produce summary statistics for the whole population. Typically in FRIs, a defined number of sample plots is established within the forest and on all trees that fall within the plot area, attributes of trees are measured.

When a sample survey is conducted, the data in the sample is used to estimate one or more summary statistics for the population. To improve the precision of estimates many FRIs incorporate auxiliary data. Frequently remotely sensed data, such as satellite imagery or aerial photographs, are integrated into the design and/or estimation stage of a FRI. Since about two decades Light Detection and Ranging (LiDAR) technologies are increasingly used in FRIs (García *et al.*, 2010).

To make efficient use of LiDAR or other remotely sensed data at the estimation stage, a relationship between field data and the remotely sensed data needs to exist. Moreover, an analyst must be able to capture the relationship in form of a statistical model. In FRIs the exact form of a (potential) association is often not known. Model formulation is therefore not only based on prior knowledge or "careful thinking" (Burnham & Anderson, 2002), but instead the data at hand is screened for potential relationships in an exploratory manner.

Various statistical modeling tools, either parametric, semi-parametric, or non-parametric, are available that allow to capture the association between a target or response variable Y, and one or more auxiliary variables or covariates X. A FRI data analyst needs to choose among these many tools. As vividly described by Selvin & Stuart (1966, page 20), a survey data analyst may be liken "to a hunter stalking an unknown quarry through an unfamiliar landscape with an arsenal of complex weapons". This "arsenal" has grown remarkably large over the past decades. Choosing the "weapons" wisely is one of the major challenges a data analyst faces today.

1.2. Model choice

Model choice and definition comprises at least two aspects, the mathematical form of the model or algorithm and the variables that may enter the model. Often — especially when working with survey data — an important third aspect needs to be considered that is intimately linked to model choice: which estimator should be used to obtain an estimate of a population quantity? And how should uncertainty attached to this estimate be quantified? The three aspects should not be treated separately. However, more often than not the process of model formulation and variable selection is separated from the inference (Chatfield, 1995). The model that is finally used to compute an estimate is simply taken as given. Such a strategy ignores the uncertainty that may evolve during the process of model formulation and application.

If the sample data at hand is used to formulate and fit a specific model, an analyst has to consider two situations, (a) relationships that exist might not be detected, and (b) relationships that seem to be supported by the data are detected but are actually spurious (Anderson *et al.*, 2001). The latter often leads to what is known as "overfitting". In the "classical" linear regression setting, the likelihood of fitting spurious effects usually increases the more covariates are available and/or the smaller the ratio between the

number of observations relative to the number of covariates. In LiDAR-assisted FRIs the number of covariates is often large, and many of them might have low or no predictive power. Moreover, in many FRIs the ratio between observations and covariates is small.

Generally, when a model is used to estimate a population parameter, the precision of this parameter depends on how well the model describes the relationships in the sample data. The better the model captures this relationship, the higher the precision. In LiDAR-assisted FRIs an analyst usually tries to identify a model that has good predictive power; the precision of coefficient estimates is of minor interest. Hence, when the sample data is fitted too hard estimates of precision may be overly optimistic.

Today, models are used in almost all scientific disciplines (Fahrmeir *et al.*, 2013). Model and variable selection — and the problems associated with it — have been covered in several statistical text books (see Burnham & Anderson (2002); Miller (2002); Claeskens & Hjort (2008)). However, survey sampling statistics differ in many aspects from other branches of statistics. Probably surprising to scientists from other disciplines, in survey theory and practice, the role of the model, has been controversially discussed for more than 40 years (Särndal, 2010). It is not so much a question of whether a model should be used or not, but more of *how* it should be integrated into the inference. Särndal (2010) provides an interesting account of this (ongoing) discussion.

In the "classical" design-based model-assisted framework, valid inference does not depend on the correctness of the model (Särndal *et al.*, 1992). The model is used to *assist* in estimation. No assumptions are made about a stochastic process that generated the data. However, this does not mean that model choice is without any consequences in the model-assisted approach. As Lumley (2011, page 83) noted, "Any model can estimate a summary of the population [...], but only some models estimate useful summaries".

So far, only few studies have investigated the effects of model and variable selection on estimates of precision in model-assisted approaches. Two notable exceptions are provided by Silva & Skinner (1997) and Knobelspies & Münnich (2008). To the author's knowledge no publication exists that has systematically assessed the effects of model and variable selection in model-assisted approaches in FRIs. This somehow surprises for at least two reasons: (a) design-based approaches dominate in FRIs (Gregoire, 1998), and (b) the types of remotely sensed data frequently used in FRIs provide the analyst often with a vast set of potentially useful covariates. This is particularly true for LiDAR technologies. For LiDAR, often more than 50 covariates are available for usually a small number of ground observations.

There exist alternatives to the model-assisted approach. In the 1960's, the design-based (model-assisted) approach became contrasted with model-based or model-dependent inference (Särndal, 2010). For the latter, the model-based approach, valid inference depends on a correctly specified model. An analyst seeks to find a model that describes the process that generated the population data. As in the model-assisted approach, model formulation is frequently data-driven. Since in the model-based approach inference depends on the model, problems of model uncertainty are likely to be more apparent.

1.3. General aim of the study

It is important to note that either approach, the model-based and model-assisted, is based on a solid theoretical basis — they simply differ (Gregoire, 1998). In both inference frameworks a working model needs to be defined that is, at least in most FRIs, obtained by screening the available data. The general aim of this study is to investigate the impact of this data-driven screening on estimates of precision within the model-assisted, as well as model-based inference framework. The focus will be on FRIs in which LiDAR data is integrated at the estimation stage.

Before a more detailed definition of this study's objectives is provided (Chapter 5), Chapter 2 will provide a brief overview of LiDAR technologies in FRIs. Special emphasis will be put on what type of modeling techniques have been used in LiDAR-assisted FRIs.

Chapter 3 provides a brief review of the (design-based) model-assisted and model-based approaches to inference. The main purpose is to (a) show how models are integrated into the estimation stage and (b) to highlight the differences between the two inference frameworks.

In Chapter 4 the model and variable selection procedures considered in this study are briefly described.

2. The use of LiDAR in forest resource assessments

2.1. LiDAR technologies in forest resource assessments

Light Detection and Ranging (LiDAR) technologies refer to active remote sensing sensors that emit laser energy. When the laser pulse emitted by the LiDAR device hits an object, the energy is reflected back to the emitter. The time elapsed is used to determine distances.

LiDAR technologies are classified in either discrete return or full waveform recording (Wulder *et al.*, 2012). In forestry application the former dominates (Wulder *et al.*, 2012, 2013). For discrete return LiDAR one or more (often up to four) returns are recorded for each emitted pulse. Full waveform LiDARs, in contrast, provide sub-meter canopy profiles (Wulder *et al.*, 2012). While waveform LiDARs usually have a large footprint, that is, a laser beam of several meter radius, for discrete return LiDAR the laser beam diameter is typically in the range of centimeters or decimeters, i.e., small footprint (McGaughey, 2013).

Depending on to which platform the LiDAR device is attached, one may further distinguish between, terrestrial, airborne, or spaceborn scanners. The use of spaceborne laser data in a forestry context has been limited so far (examples are provided by Lefsky *et al.* (2011) and Popescu *et al.* (2011)). In most LiDAR-assisted FRIs, airborne laser scanners (ALS) are used.

If an airplane or helicopter, to which the LiDAR sensor is attached, moves over an area, the flying altitude and geographical position of the sensor is constantly recorded. For discrete return LiDARs, thousands of pulses are emitted every second, and from these returns a so-called LiDAR point cloud is obtained (see Figure 2.1). For each point in the cloud the x, y and z coordinate is recorded.

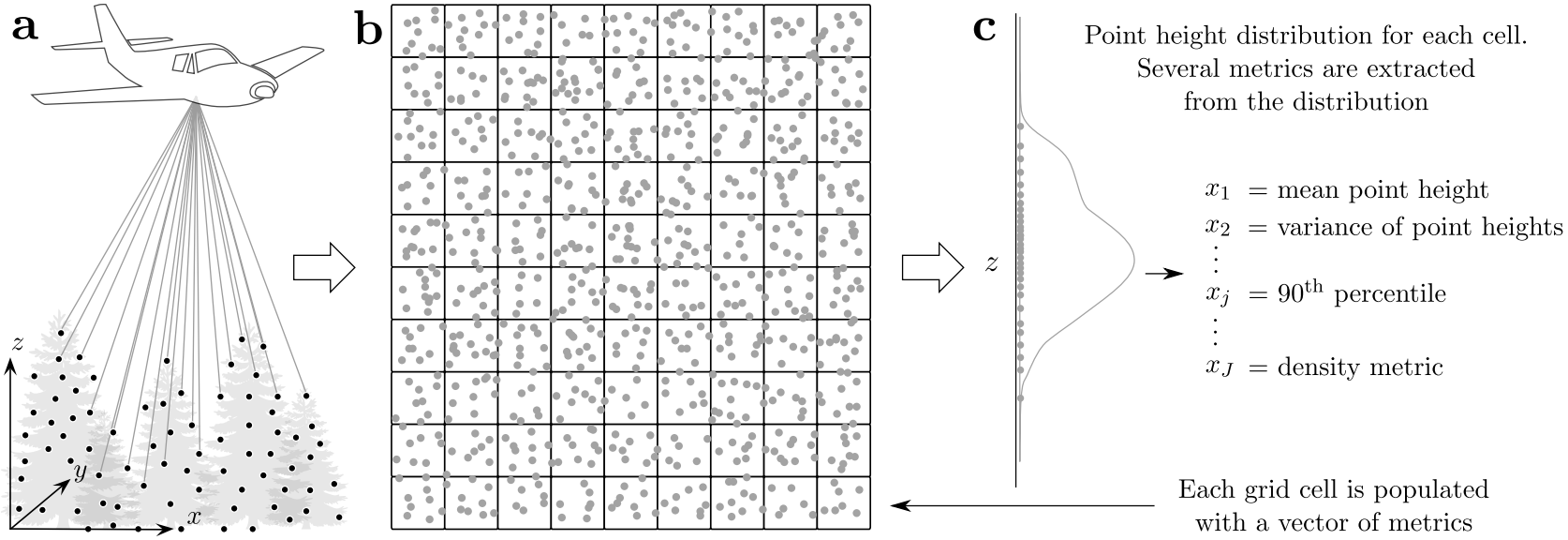

Figure 2.1.: Simplified description of a (small-footprint, discrete return) LiDAR application in a FRI. (**a**) an airplane moves of a forest environment. A laser, attached to the airplane, emits laser pulses. The time the signal needs to travel to the object and back to the sensor is recorded and used to determine distances. The geographic position of the airplane is known, such that a 3-D point cloud is obtained, where the x, y and z coordinate of each point is determined. (**b**) the point cloud is gridded into cells. The cell size usually corresponds to the size of a field plot. (**c**) from the points within a cell, a point height distribution is obtained. Several metrics (e.g., mean point height, variation of point heights, etc.) are extracted from each point height distribution. Finally, each grid cells is populated with a vector of metrics.

In forestry, the point cloud data obtained from the laser scanner is used for many different purposes. In FRI applications the cloud is usually rasterized into equally sized, non-overlapping, and often square-shaped grid cells. When LiDAR data is used in combination with data obtained in the field, a grid cell size that matches in size with field plots is usually chosen.

Once the point cloud has been gridded into cells, a point height distribution is obtained for each cell (see Figure 2.1). From this distribution several so-called LiDAR metrics are computed. These metrics form summary statistics of the point height distribution. Typical examples of metrics are the mean height of points in a cell, the variance of point heights, or the fraction of points above a defined threshold. Software packages like FUSION© (McGaughey, 2013) often compute more than 100 different metrics from the raw point cloud.

The list of LiDAR applications in FRIs is long and growing (Magnussen *et al.*, 2010). Several studies have shown that LiDAR data have good predictive power of desired biophysical variables that are of interest in FRIs. LiDAR applications have successfully targeted the estimation and prediction of tree and canopy height (Naesset, 1997; Magnussen & Boudewyn, 1998; Clark *et al.*, 2004; Jensen *et al.*, 2006; Maltamo *et al.*, 2006b), diameter and basal area distributions (Gobakken & Næsset, 2005; Hudak *et al.*, 2006; Breidenbach *et al.*, 2008; Salas *et al.*, 2010), timber volume (Nilsson, 1996; Næsset, 1997; Lefsky *et al.*, 1999; Maltamo *et al.*, 2004; Jensen *et al.*, 2006; Maltamo *et al.*, 2006a; Dalponte *et al.*, 2011), forest productivity (Lefsky *et al.*, 2005), forest structure (Jaskierniak *et al.*, 2011; Kane *et al.*, 2010; Miura & Jones, 2010; Latifi *et al.*, 2012), stem density (Næsset & Bjerknes, 2001; Maltamo *et al.*, 2004; Hudak *et al.*, 2006), understory light conditions (Alexander *et al.*, 2013), forest fuel parameters (Andersen *et al.*, 2005; Erdody & Moskal, 2010; García *et al.*, 2011), aboveground biomass (Drake *et al.*, 2003; Andersen *et al.*, 2011; Frazer *et al.*, 2011b; Gleason & Im, 2012; Ahmed *et al.*, 2013; Næsset *et al.*, 2013a), forest carbon (Patenaude *et al.*, 2004; Gonzalez *et al.*, 2010; Asner *et al.*, 2012; Hudak *et al.*, 2012; Stephens *et al.*, 2012), or change in aboveground biomass over time (Næsset *et al.*, 2013b; Skowronski *et al.*, in press).

Many more examples of successful application of LiDAR data in FRIs exist. Nelson (2013) provides a review of early applications of LiDAR, and Hyyppä *et al.* (2008), Koch (2010), Wulder *et al.* (2012) and Wulder *et al.* (2013) provide reviews of how laser technologies have been integrated into forestry applications.

Hesitation to use LiDAR technologies in the context of FRIs is generally argued on non-technical grounds (Magnussen *et al.*, 2010). While satellite imagery is often readily available for large areas and free of charge (e.g., Landsat 8 imagery), collecting LiDAR data for large areas is comparatively expensive. Moreover, while LiDAR data is collected often only once for a given application, satellite imagery is often available over short time intervals. For example, Landsat products are available on a monthly basis.

2.2. Relating field and LiDAR data in FRIs

2.2.1. Choosing a modeling technique

In order to make efficient use of the LiDAR data in FRIs, the field and remotely sensed data need to be linked using a statistical model. If the LiDAR point cloud has been gridded into cells, and a set of metrics is available for each cell, these metrics may be related to the information obtained for the field plots, e.g., aboveground biomass per plot. Such an approach is often referred to as an area-based approach (ABA) (Næsset, 2002; Wulder *et al.*, 2013).

Regardless of the target variable, one of the first issues a data analyst faces is to decide on how the relationship should be modeled. In most LiDAR-assisted FRIs, parametric approaches such as simple and multiple linear regression techniques have been used (Garcia-Gutierrez *et al.*, 2014). Early examples of regression modeling are provided by Næsset (1997) and Means *et al.* (1999). In several studies multiplicative models have been employed (Næsset, 2002). Here, the target and explanatory variables are ordinarily transformed using the log or square-root transform. Predictions made by these models need to be back-transformed.

To avoid the need of transforming variables back to the original scale, Ene *et al.* (2012) used generalized linear models (GLMs) with a square-root link function to relate aboveground biomass to LiDAR data for Norwegian forests. Using GLMs has also been advocated by Gregoire *et al.* (2008). Whether a transformation of the target and/or explanatory variables is deemed necessary or not depends, among other things, on the target variable investigated. Nord-Larsen & Riis-Nielsen (2010), for example, noted that, after visual inspection of the data, they saw no reason for transformation because the relationship between the target variable and the LiDAR data showed a linear pattern. In their study they used LiDAR data to predict dominant height for different forest types

in Denmark. However, (Næsset, 2002), for example, looked at the same target variable and transformed variables using the log-transform.

2.2.2. Variable selection

Since many LiDAR metrics can be extracted from the point cloud a data analyst needs to filter out those that are useful for a given application. Using all of them is generally not recommended and rarely done in practice. However, even if a small number of metrics is purposefully selected, a large number of potentially useful models may still be established; in particular when interactions between variables are considered.

In some applications, subject matter dictates which metrics are to be included into the model. However, in many FRI-LiDAR applications statistical subset selection procedures are employed (Garcia-Gutierrez *et al.*, 2014). Stepwise regression procedures, either forward or backward selection, or mixed, are common. Examples of their use are provided by Gobakken & Næsset (2005); Hudak *et al.* (2006); Vincent *et al.* (2012); Ene *et al.* (2012). Dalponte *et al.* (2011), for example, used F-tests and a significance level of 0.05 to drop or retain variables. In many studies criteria such as the Akaike Information Criterion (AIC), the corrected AIC (AICc), or the Bayesian Information Criterion (BIC) or variants of these criteria have been used.

Due to the increasing computational power of modern computers and faster search algorithms, best-subset selection has become prominent in variable selection. In best-subset selection separate models are fitted to all possible combinations of covariates (Hastie *et al.*, 2009). Using different criteria such as maximum R^2, or Mallow's C, the "best" model is selected. This approach appears to be popular in LiDAR applications, see e.g., Hudak *et al.* (2006), Tonolli *et al.* (2011), Zhao *et al.* (2012), Rana *et al.* (2014). However, even with modern computers, best subset selection is currently still prohibitive if too many covariates are available.

Another common approach is to formulate a set of candidate models based on "careful thinking". Strunk *et al.* (2011) and Nyström *et al.* (2012), for example, first defined a set of candidate models and then used statistical software tools that guided the final selection. This approach of combining expert knowledge with automated variable selection usually leads again to a final — supposedly — "best" model.

In their book, Burnham & Anderson (2002) suggest to not select a single "winner" but to consider all candidate models as potentially useful. Using model averaging techniques

based on the Akaike or Bayesian Information Criterion has become increasingly popular in the past decade. However, to the author's knowledge these techniques have not yet been used in LiDAR applications in FRIs.

In many studies several procedures are combined. Jensen *et al.* (2006), for example, used best-subset selection. Before the final model was selected the number of possible models was substantially reduced upon review of selection procedures including the AIC, AICc, and Mallows's C_p.

2.2.3. Multicollinearity

Since plot-level LiDAR metrics are usually computed from the same point cloud, many of the metrics correlate (strongly) with each other. In a modeling context, issues of multicollinearity may be a concern. To reduce collinearity between covariates different approach have been used in LiDAR applications. Variance inflation factors (VIFs) have frequently been employed to identify highly correlated covariates. Variables showing high VIFs have then been removed from the model. Here, a choice has to be made when to retain or remove a variable. Some researchers choose a maximum VIF of 10 (Penner *et al.*, 2013) before dropping a variable, others used a threshold VIF of 5 (d'Oliveira *et al.*, 2012). No universally accepted rules appears to exist.

In some studies the number of LiDAR covariates was reduced by using principal component analysis (PCA) or canonical correlation analysis (CCA) techniques, see e.g., Lefsky *et al.* (2005). Sherrill *et al.* (2008), and Stephens *et al.* (2012). Nord-Larsen & Riis-Nielsen (2010) and Nord-Larsen & Schumacher (2012) used cluster analysis to identify correlated groups of LiDAR metrics; from each group the variable that correlated most with the target variable was selected. Tinkham *et al.* (2012) used Person's correlation coefficient to identify and select correlated variables. Stephens *et al.* (2012) suggested to use partial least squares (PLS) regression. They argued that PLS may prove particularly useful when a large number of highly correlated LiDAR metrics is available.

2.2.4. Model validation

When a model is calibrated to a single dataset, it is often of interest how the model performs on formerly unseen data. Different methods for model validation have been used. Jensen *et al.* (2006) and Frazer *et al.* (2011b), for example, divided their datasets into a training and a validation dataset. In this approach, the model is formulated

and fitted to the training set and subsequently its performance is evaluated by making predictions for the test set. However, there are no predefined rules of how large the different splits should be relative to each other. A popular choice is to use 3/4 of the data to train the model and use the remaining 1/3 for model validation. The decision depends, among other things, on the total number of available sample observations.

If only few observations are available splitting the sample data into two parts becomes infeasible. In that case k-fold cross-validation provides an alternative. In k-fold cross-validation the sample data is randomly divided into k groups, or folds. One of the k folds is treated as a validation set, and the remaining $k-1$ folds are used for training the model. This procedure is repeated k times; each time a different fold serves as the validation set (Hastie *et al.*, 2009). Here, an analyst needs to decide how many folds to use. Popular choices are to split the data into 3, 5, or 10 folds. Jakubowski *et al.* (2013), for example, used 10 fold cross-validation to assess trade-offs between LiDAR pulse density and measurement accuracies.

If the number of observations is small, another options for model validation is to use leave-one-out cross-validation (LOOCV). LOOCV is closely related to k-fold cross-validation. Here, the sample data is divided into as many folds as there are observations. To predict the value of one observation, that same observation will not be used for model fitting. LOOCV was, for examples, used by Magnussen *et al.* (2010), Bright *et al.* (2012), Nyström *et al.* (2012), Li *et al.* (in press).

2.2.5. Non-parametric approaches

For large area FRIs, such as National Forest Inventories (NFIs) non-parametric techniques such as k-nearest-neighbour (kNN) are frequently used to relate optical satellite imagery to field observations (McRoberts & Tomppo, 2007). However, for LiDAR applications only few studies have used non-parametric modelling techniques. Recent examples of using the random forest (RF) algorithm (Breiman, 2001), are given in Latifi *et al.* (2010), Gleason & Im (2012), and Penner *et al.* (2013). These applications have targeted at the prediction of standing timber and biomass in forests in Germany and the US. Support vector machines (SVM) were tested by García *et al.* (2011) for forest fuel type mapping, and Gleason & Im (2012) used SVM to estimate forest biomass. Breidenbach *et al.* (2010, 2012) used k-nearest-neighbour (kNN) techniques to predict standing timber and number of stems for individual forest stands in Germany and Nor-

way. Penner *et al.* (2013) used kNN to predict top height, merchantable basal area, and gross merchantable volume in boreal forests in Ontario, Canada.

The application of RF is relatively recent in FRIs but receives increasing attention (Brosofske *et al.*, 2014). RF does not require that a model is specified and it can cope with situations where there are more variables than observations, collinearities, or both (Penner *et al.*, 2013). Furthermore, for RF variable transformation is not necessary, as non-linear relationships between the target and explanatory variables are captured in a tree-based structure.

2.3. Inference

In most forestry related LiDAR applications, a model is used to make predictions for those parts of the populations that have not been sampled on the ground. In the area-based approach, mentioned above, this means that predictions of the target variable are made for all LiDAR grid cells for which an observation of the target variable is not available. In this setting, "finding" a model that has good predictive power is essential. The quality of the predictions depends on how well the postulated model captures the structure in the population. However, even if the model describes the data generating process well, it is very likely that the final prediction will not equal the "truth". For valid inference, this difference between the (unknown) "truth" and the prediction needs to be expressed in probabilistic terms (McRoberts, 2011). The following chapter provides an overview of two different modes of inference in survey sampling.

3. Theoretical background

3.1. Design-based inference

3.1.1. General framework

In this chapter the notation given in Särndal *et al.* (1992) was largely adopted. In the design-based approach we consider a finite population U consisting of N elements,

$$U = \{u_1, u_2, \ldots, u_k, \ldots, u_N\}.$$

For simplicity, the kth element in U will be represented by its label k. The finite population can thus be written as

$$U = \{1, 2, \ldots, k, \ldots, N\}.$$

Attached to each element $k \in U$ is the value of a *study* or *target* variable y. The population vector of y is given by

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_k, \ldots, y_N)'.$$

In the design-based approach these values are treated as fixed numbers. The population is, therefore, called *fixed* and finite. No assumptions are made about the stochastic process that has generated the population data.

The population mean of the target variable y is given by

$$\mu_y = N^{-1} \sum_{k \in U} y_k. \tag{3.1}$$

Since the population vector $\boldsymbol{y}$ consists of fixed constant, μ_y is a fixed number, too. If all y_k in the population are observed, the parametric mean can be calculated. Note, in (3.1) $\sum_{k \in U}$ indicates that the sum is taken over all elements in the population U.

We assume that the values of the target variable are unknown to us. To obtain an estimate of the population mean (3.1) a *probability* sample S, $S \subseteq U$, of size n is drawn from U. If $n = N$, the entire population is sampled, i.e., a census is conducted, and the population mean can be calculated. Here, we will assume that the sample size n is generally small. Furthermore, it is — as usually — assumed that there are no measurement errors.

For the time being, we assume that a simple random sample without replacement (SRSwoR) is drawn. Without replacement means, that once an element has been selected it cannot be selected again. Under SRSwoR, with fixed N and n, each element has the same probability of ending up in the sample. The inclusion probability of element k is given by

$$\pi_k = \frac{n}{N}.$$

Which elements are selected into the sample is determined by a random process. If n and N are fixed, there is a finite set of distinct samples that can be drawn from U. This set will be denoted by $\mathscr{S} = \{S_1, S_2, \ldots, S_i\}$. Under SRSwoR, the cardinality, or size (indicated by $|\cdot|$), of the set is given by

$$|\mathscr{S}| = \binom{N}{n} = \frac{N!}{n!(N-n)!}. \tag{3.2}$$

The probability of selecting one specific sample from this set is given,

$$p(S) = 1/\binom{N}{n}. \tag{3.3}$$

The function $p(S)$ is frequently called the sample design in survey sampling literature. For designs other than SRSwoR the probability of selecting a sample might be different from (3.3).

We assume that a probability sample is drawn from U. A probability sample needs to satisfy certain conditions (Särndal *et al.*, 1992, page 8):

1. The set of possible samples, $\mathscr{S}$, that can be drawn from U under a given $p(\cdot)$ can be defined.
2. The probability $p(S)$ of selecting a specific sample is known.
3. Each element in the population has a positive probability of selection.
4. One sample is selected by a random mechanism, and each sample S receives exactly the probability $p(S)$.

To be able to draw a sample from a finite population a list is needed that contains all elements in the population, the so-called sample frame. Following Särndal *et al.* (1992, page 9), we define the sampling frame as "any material or device used to obtain observational access to the finite population of interest". We further assume that the sampling frame is complete, that is, each element in the population can be accessed from the sampling frame.

Remark: In most FRIs the population and sampling frame does not consist of a finite set of elements. Usually an *aerial* sampling frame is assumed. Within a forest covering an area A, n sample points are randomly placed and around each point a field plot of either fixed or variable size is established. One or more attributes are then recorded on each tree is included within the plot. One plot, and not a tree, represents one independent observation. Since points have, by definition, no dimension, infinitely many points can be selected within A. The concept of a fixed and finite population U consisting of a set of $k = 1, 2, \ldots, N$ elements, is therefore not directly transferable to most FRIs. However, if remotely sensed data, as for example satellite imagery, are integrated at the estimation phase, the concept of a finite population is often useful and the infinite population approach of FRIs needs to be converted to a finite population.

Once a probability sample has been drawn from the population, the population mean of the target variable can be estimated by (Cochran, 1977)

$$\bar{y} = n^{-1} \sum_{k \in S} y_k. \tag{3.4}$$

Here, the sum, $\sum_{k \in S} y_k$, is taken over all elements k in the sample S. Note, the above estimator (3.4) can alternatively be written as

$$\bar{y} = N^{-1} \sum_{k \in U} y_k I_k, \tag{3.5}$$

where

$$I_k = \begin{cases} 1 & \text{if element } k \text{ is selected into the sample} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

Since the values y_k are treated as fixed constants, the random mechanism in (3.5) is induced by the random variable I_k.

The expectation of $\bar{y}$ is given by,

$$E(\bar{y}) = \sum_{S \in \mathscr{S}} p(S) \bar{y}(S). \tag{3.7}$$

That is, the expectation of $\bar{y}$ is the weighted sum of all possible estimates of $\bar{y}$ under the design. Since $E(\bar{y}) = \mu_y$, the estimator (3.4) is unbiased.

Furthermore,

$$\lim_{n \to \infty} E(\bar{y}) = \mu_y. \tag{3.8}$$

For a definition of asymptotic consistency see Särndal *et al.* (1992, page 166).

Since n and N (as defined above) are finite and fixed, the definition of asymptotic unbiasedness in (3.8) can not be directly transferred into the sample survey context. A "workaround" for the finite population and sample size setting is to imagine a sequence of increasing populations (and sample) sizes, where n and N both tend to infinity (see Särndal *et al.* (1992, page 167) for details). The practical importance of asymptotic unbiasedness is that, when n grows sufficiently large, the estimator (3.4) is considered nearly unbiased (and nearly consistent).

3.1.2. Variance estimation

The variance of the mean estimator is defined as

$$V(\bar{y}) = \sum_{S \in \mathscr{S}} p(S)[\bar{y}(S) - E(\bar{y})]^2, \tag{3.9}$$

and can, for a sample of size n, be calculated by (Cochran, 1977)

$$V(\mu_y) = \frac{\sigma^2}{n}, \tag{3.10}$$

where

$$\sigma^2 = N^{-1} \sum_{k \in U} (y_k - \mu_y)^2 \tag{3.11}$$

is the population variance. The square-root of (3.10),

$$SE_\mu = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \tag{3.12}$$

gives the standard error. Under SRSwoR, the variance of $\bar{y}$ is estimated by

$$\hat{V}(\bar{y}) = (1 - f)\frac{s^2}{n}, \tag{3.13}$$

where $f = \frac{n}{N}$ is the sample fraction, and

$$s^2 = (n-1)^{-1} \sum_{k \in S} (y - \bar{y})^2. \tag{3.14}$$

The standard error is estimated in analogy to (3.12) as

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}. \tag{3.15}$$

Confidence intervals for a given α are obtained as follows,

$$\mathbb{P}(\bar{y} - SE_{\bar{y}} \times t_{\frac{\alpha}{2},\nu} < \mu_y < \bar{y} + SE_{\bar{y}} \times t_{\frac{\alpha}{2},\nu}) = .95 \tag{3.16}$$

for $\alpha = 0.05$, assuming a t-distribution with $\nu = n - 1$ degrees of freedom.

A key feature of design-based inference is that the estimates that can be computed from all samples that are permissible under the given design and their distribution are the only basis of inference. The distribution of possible estimates is frequently called the randomization distribution, and design-based inference is, therefore, sometimes referred to as randomization inference.

3.1.3. Using auxiliary information to improve the estimation

Whereas in the preceding section (3.1) no additional information was integrated into the estimation, in this section auxiliary information will be integrated into the estimation. We assume that for each element k in the finite population U the values of one or more auxiliary variables

$$\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kJ})',$$

are available. These values are assumed known at the outset. That is, the vector $\boldsymbol{x}_k$ is accessible for all $k \in U$ after the sample S has been selected from U. Again, we assume that a SRSwoR has been drawn.

We can make use of the auxiliary information in various ways. Here, we will consider model-assisted design-based estimation using the regression estimator (REG). If the auxiliary variables strongly correlate with the target variable, large gains in efficiency can be expected when the REG is used instead of the variance estimators given in the preceding section.

For the time being we will assume that only one single auxiliary variable, x_k, is available $\forall k \in U$. If a sample S is drawn from U, the data tuple (y_k, x_k) are observed on all $k \in S$. Using the sample data, $\bar{y}$ and $\bar{x}$ can be unbiasedly estimated using the estimator given in (3.4). However, since x_k is known $\forall k \in U$, the "true" population mean of x_k, i.e., $\mu_x = N^{-1} \sum_{k \in U} x_k$, is also known (assuming that measurement errors are absent or

at least negligible). Now, if x correlates positively with y and the relationship is strong, then we would expect that, if $\bar{x} < \mu_x$, we also have that $\bar{y} < \mu_y$.

The regression estimator of the mean is defined as

$$\bar{y}_{\text{REG}} = \bar{y} - b(\mu_x - \bar{x}), \tag{3.17}$$

where $\bar{y}$ is estimated using the estimator given in (3.4),

$$\mu_x = N^{-1} \sum_{k \in U} x_k, \tag{3.18}$$

is the known population mean of the auxiliary variable,

$$\bar{x} = N^{-1} \sum_{k \in U} x_k I_k = n^{-1} \sum_{k \in S} y_k, \tag{3.19}$$

is the estimated mean of the (known) population mean μ_x, and

$$b = \frac{\sum_{k \in S}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k \in S}(x_k - \bar{x})^2}, \tag{3.20}$$

is an estimate of the population regression coefficient

$$B = \frac{\sum_{k \in U}(x_k - \mu_x)(y_k - \mu_y)}{\sum_{k \in U}(x_k - \mu_x)^2}. \tag{3.21}$$

Equation (3.17) shows that the design-unbiased estimate, $\bar{y}$, is "adjusted" by the quantity $b(\mu_x - \bar{x})$. Note that the estimator (3.17) is *not* unbiased. The bias is given by

$$E[\bar{y}_{\text{REG}} - \mu_y] = E[\bar{y} - \mu_y] + E[b(\mu_x - \bar{x})]. \tag{3.22}$$

However, if (3.17) portrays the population point scatter reasonably well, the bias is negligible (Gregoire & Valentine, 2008) in particular for large samples.

The REG estimator (3.17) can alternatively be written using so-called calibrated, or g-weights

$$\bar{y}_{\text{REG}} = (w_k \times g_k \times y_{k\in S})/N, \tag{3.23}$$

where w_k is the reciprocal of π_k, i.e., $w_k = 1/\pi_k$, and (Lehtonen & Pahkinen, 2004)

$$g_k = \left[1 + \frac{\mu_x - \bar{x}}{\frac{n-1}{n} s_x^2} \times (x_k - \bar{x})\right], \tag{3.24}$$

and

$$s_x^2 = (n-1)^{-1} \sum_{k\in S} (x_k - \bar{x})^2. \tag{3.25}$$

The REG can easily be extended to situations where more than one auxiliary variable is available. For $J > 1$ the estimator (3.17) becomes

$$\bar{y}_{\text{REG}} = \bar{y} + (\boldsymbol{\mu}_x - \bar{\boldsymbol{x}})'\boldsymbol{b}, \tag{3.26}$$

where the mean vector of population means of the J auxiliary variables is given by

$$\boldsymbol{\mu}_x = N^{-1} \sum_{k\in U} \boldsymbol{x}_k, \tag{3.27}$$

and

$$\bar{\boldsymbol{x}} = n^{-1} \sum_{k\in S} \boldsymbol{x}_k \tag{3.28}$$

is the sample mean of the $\boldsymbol{x}_k \in S$,

$$\boldsymbol{b} = (b_1, b_2, \ldots, b_j, \ldots, b_J)' = (\boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{X}'_{k\in S}\boldsymbol{y}_{k\in S}, \tag{3.29}$$

with

$$\boldsymbol{X}_{k\in S} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,J} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,J} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & \cdots & x_{n,J} \end{pmatrix}, \tag{3.30}$$

and

$$\boldsymbol{y}_{k\in S} = (y_1, y_2, \ldots, y_k, \ldots, y_n)' \tag{3.31}$$

is the vector of sampled observations of the population vector $\boldsymbol{y}$.

The estimator can be rewritten as

$$\bar{y}_{\text{REG}} = \left(\sum_{k\in S} \left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'(\boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{x}_k/\pi_k \right] w_k y_k \right)/N, \tag{3.32}$$

where

$$\boldsymbol{t}_x = \sum_{k\in U} \boldsymbol{x}_k, \tag{3.33}$$

are the population (known) totals of the J auxiliary variables, and

$$\hat{\boldsymbol{t}}_x = \sum_{k\in S} \frac{\boldsymbol{x}_k}{\pi_k} \tag{3.34}$$

are the estimated J population totals using the sample data.

As aforementioned, the weights π_k do not need to be the same for all $k \in U$. If weights are allowed to vary among the population elements the REG is frequently called the generalized regression (GREG) estimator (see Särndal *et al.* (1992)).

Further note that the g-weights

$$g_k = 1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x)'(\boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{x}_k/\pi_k \tag{3.35}$$

do not depend on y, and can therefore be used for estimating different target variables.

A third expression of the REG estimator is given by

$$\bar{y}_{\text{REG}} = N^{-1} \sum_{k \in U} \hat{y}_k + n^{-1} \sum_{k \in S} e_k, \tag{3.36}$$

where predictions, $\hat{y}_k = \boldsymbol{x}_k' \boldsymbol{b}$, are made for the entire population, and $e_k = y_k - \hat{y}_k$. Note, the use of this estimator requires that the values $\boldsymbol{x}_k$ are known for all population elements. For the estimator given in (3.26) only the population means of the J auxiliary variables and the $\boldsymbol{x}_k \in S$ are needed.

3.1.4. Variance estimation for the regression estimator

Different variance estimators for the REG exist. An asymptotically approximately design-unbiased estimator is given by (Lehtonen & Pahkinen, 2004)

$$V(\bar{y}) \doteq \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sigma^2_{E_k} \tag{3.37}$$

where

$$\sigma^2_{E_k} = (N-1)^{-1} \sum_{k \in U} (E_k - \bar{E})^2, \tag{3.38}$$

is the variance of the population regression residuals, with

$$\bar{E} = N^{-1} \sum_{k \in U} E_k \tag{3.39}$$

and

$$E_k = y_k - \hat{y}_k \qquad \forall k \in U, \tag{3.40}$$

are the population regression residuals obtained by

$$\hat{y}_k = B x_k \qquad \forall k \in U \tag{3.41}$$

An estimator for (3.37) is given by (Lehtonen & Pahkinen, 2004)

$$\hat{V}(\bar{y}) \doteq \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)s_{e_k}^2 \tag{3.42}$$

where

$$s_{e_k}^2 = (n-1)^{-1}\sum_{k\in S}(e_k - \bar{e})^2, \tag{3.43}$$

is the variance of the regression residuals, with

$$\bar{e} = n^{-1}\sum_{k\in S} e_k \tag{3.44}$$

and

$$e_k = y_k - \hat{y}_k, \tag{3.45}$$

are the regression residuals obtained by

$$\hat{y}_k = bx_k \tag{3.46}$$

for a single auxiliary variable, x. The estimator (3.42) is frequently used when more than one auxiliary variable is used. That is, when

$$\hat{y}_k = \boldsymbol{x}_k'\boldsymbol{b} \tag{3.47}$$

Generally, from (3.42) it follows that the better the regression model fits the *sample* data, the smaller (3.42), and, since the estimate of the variance is solely based on the residuals, the more precise the estimate.

The variance estimator (called the simple variance estimator hereafter) can be rewritten as

$$\hat{V}(\bar{y}_{\mathrm{REG}}) \doteq \hat{V}(\bar{y})(1 - R^2) \tag{3.48}$$

where $\hat{V}(\bar{y})$ is given in (3.13), and

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}. \tag{3.49}$$

Here,

$$SS_{\text{res}} = \sum_{k \in S} (y_k - \hat{y}_k)^2 = \sum_{k \in S} e_k^2, \tag{3.50}$$

and

$$SS_{\text{tot}} = \sum_{k \in S} (y_k - \bar{y})^2. \tag{3.51}$$

Equation (3.48) underlines that precision depends on how well the chosen working regression model fits the sample data.

Silva & Skinner (1997) proposed an alternative version for (3.42), which is a generalization of the estimator given in Cochran (1977, equation 7.2.9, page 195) (see also Fuller (2009, equation 2.2.4, page 101)),

$$\hat{V}(\bar{y}_{\text{REG,J}}) = \frac{1-f}{n} s_{e_k}^2 \tag{3.52}$$

where

$$s_{e_k}^2 = (n - J - 1)^{-1} \sum_{k \in S} e_k^2 \tag{3.53}$$

and

$$e_k = (y_k - \bar{y}) - (\boldsymbol{x}_k - \bar{\boldsymbol{x}})' \boldsymbol{b} \tag{3.54}$$

In contrast to (3.48), this estimator takes account of multiple auxiliary variables in the denominator. A third, more conservative estimator using g-weights is given by Lehtonen & Pahkinen (2004)

$$\hat{V}(\bar{y}_{\mathrm{REG,g}}) = \left(1 - \frac{n}{N}\right)\left(\frac{n-1}{n-J}\right)s^2_{e^*_k}, \tag{3.55}$$

where

$$s^2_{e^*_k} = (n-1)^{-1}\sum_{k\in S}(e^*_k - \bar{e}^*)^2, \tag{3.56}$$

is the variance of the g-weighted regression residuals, with

$$\bar{e}^* = n^{-1}\sum_{k\in S} e^*_k, \tag{3.57}$$

and

$$e^*_k = e_k g_k. \tag{3.58}$$

The g-weights are given in (3.35).

In summary, in the design-based inference framework, the precision of estimates is assessed by considering all possible estimates of the target parameter permissible under the design. A random mechanism determines which elements end up in the sample. The distribution of the target variable in the population is not of interest; it is the distribution of estimates that forms the basis for inference. When auxiliary information is incorporated into the estimation, the estimators maintain their design properties, at least asymptotically.

3.2. Model-based inference

3.2.1. General framework

In this section a brief introduction to model-based approaches to inference is provided. The underlying assumptions in the model-based approach differ from those in the model-assisted approach. The differences between the frameworks may be best explained by considering what is treated as fixed and what as random. In the model-based framework, the population values y_k are regarded as outcomes of random variables

$$Y_1, Y_2, \ldots, Y_k, \ldots, Y_N.$$

In the design-based approach to inference, in contrast, the population values

$$y_1, \ldots, y_k, \ldots, y_N$$

are treated as fixed but unknown constants. The only random mechanism is introduced by the sampling procedure, that is, whether element k ends up in the sample or not. No assumption is made about the structure of the population (Gregoire, 1998). This is different in the model-based approach.

In model-based inference a so-called superpopulation model is assumed to have generated the population data. A simple model to adopt is one that satisfy the following conditions

$$\begin{aligned} \mathrm{E}[y_k] &= \mu_y \\ V[y_k] &= \sigma_y^2 \end{aligned} \tag{3.59}$$

with y_k and y_l independent when $k \neq l$. This simple model is often called the *common mean* or *homogeneous population model*.

In contrast to the design-based approach, the sample S is treated as fixed in the model-based approach. All expectations and variances are conditional on the sample that has been selected (Chambers & Clark, 2012). Since the values y_k are outcomes of a stochastic process, the population mean is considered to be a random variable, too. If observations of the target variable are made on all elements in the population, i.e., a census is conducted, the mean of all elements represents one possible outcome of this random variable.

If a sample S is drawn from the finite population U, the values $y_1, y_2, \ldots, y_k, \ldots, y_n$ are observed. The sample based estimate of μ_y is calculated in the same way as in the design-based approach, namely

$$\bar{y} = n^{-1} \sum_{k \in S} y_k \tag{3.60}$$

An estimate of the population total is given by

$$\hat{\tau}_{y,\mathrm{MB}} = \sum_{k \in S} y_k + \sum_{k \in U-S} \hat{y}_k = \frac{N}{n} \sum_{k \in S} y_k, \tag{3.61}$$

where $k \in U - S$ is the non-sampled set. Equation (3.61) makes clear that in the model-based approach, estimating a population quantity is viewed as a prediction problem. In model-based inference the joint probability distribution of $Y_1, \ldots, Y_N$ supplies the link between sampled and non-sampled elements (Lohr, 1999). As Royall (1992, page 225) noted, "Estimating a finite population mean from a sample is equivalent to predicting the mean of the non-sample values".

The estimator (3.61) is *model-unbiased* (Lohr, 1999). If the model correctly specifies the superpopulation model, then

$$\mathrm{E}_\zeta[\hat{\tau}_{y,\mathrm{MB}} - \tau_y] = \frac{N}{n} \sum_{k \in S} \mathrm{E}[Y_k] - \sum_{k \in U} \mathrm{E}[Y_k] = 0, \tag{3.62}$$

and, hence, the estimator is unbiased over repeated realizations of the population. However, if the model does not hold, the adverse effects on inference may be substantial (Hansen *et al.*, 1983; McRoberts, 2011). Note that, here, the subscript ζ (the Greek letter zeta) indicates that expectation is with respect to the model.

3.2.2. Variance estimation

The estimator for the variance of the mean (and total) for the common mean model is similar to the estimator in the design-based approach (see Lohr (1999, page 56))

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \tag{3.63}$$

where s^2 is estimated as in (3.14). When a model different from (3.59) is adopted, the estimators might differ.

3.2.3. Using auxiliary information to improve the estimation

As in the model-assisted approach introduced in Section 3.1, auxiliary information can be incorporated into the estimation to improve efficiency. We, again, assume that the data tuples (y_k, x_k) are observed on all $k \in S$. The linear population model (REGB) assuming a single auxiliary variable, $J = 1$, is given by

$$\begin{aligned} \mathrm{E}[y_k|x_k] &= \beta_0 + \beta_1 x_k \\ V[y_k|x_k] &= \sigma^2 \end{aligned} \tag{3.64}$$

where y_k and y_l are independent when $i \neq l$. An ordinary least squares (OLS) estimate of β_1 is obtained in a similar way as in (3.20), and β_0 is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{3.65}$$

The prediction variance for the model-based regression estimator (under the linear population model), is defined as (Chambers & Clark, 2012),

$$V(\bar{y}_{\mathrm{REGB}}) = \sigma^2 \left[\left(1 - \frac{n}{N} \right) + \frac{(\mu_x - \bar{x})^2}{(1 - n^{-1}) s_x^2} \right], \tag{3.66}$$

where $\mu_x = N^{-1} \sum_{k \in U} x_k$, and

$$s_x^2 = (n-1)^{-1} \sum_{k \in S} (x_k - \bar{x})^2. \tag{3.67}$$

Note, the only unknown quantity in (3.66) is σ^2. From standard regression theory it is known that σ^2 can be unbiasedly estimated by

$$\hat{\sigma}^2 = (n-2)^{-1} \sum_{k \in S} (y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2. \tag{3.68}$$

Under the linear population model the variance is then (Chambers & Clark, 2012)

$$\hat{V}(\bar{y}_{\text{REGB}}) = \hat{\sigma}^2\left[\left(1 - \frac{n}{N}\right) + \frac{(\mu_x - \bar{x})^2}{(1 - n^{-1})s_x^2}\right]. \tag{3.69}$$

The linear population model can easily be extended to more than one auxiliary variable, $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kJ})'$. The model (3.59) then becomes

$$\begin{aligned} \mathrm{E}[y_k|\boldsymbol{x}_k] &= \boldsymbol{x}_k'\boldsymbol{\beta} \\ V[y_k|\boldsymbol{x}_k] &= \sigma^2 \end{aligned} \tag{3.70}$$

where y_k and y_l are independent conditional on $\boldsymbol{X}$ when $k \neq l$. The best linear unbiased estimator for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{k\in S}'\boldsymbol{X}_{k\in S})\boldsymbol{X}_{k\in S}'\boldsymbol{y}_{k\in S}. \tag{3.71}$$

An estimator for the population mean is (Chambers & Clark, 2012)

$$\bar{y}_{\text{REGB}} = \left(\sum_{k\in S} y_k + \sum_{k\in U-S} \boldsymbol{x}'\hat{\boldsymbol{\beta}}\right)/N. \tag{3.72}$$

The prediction variance is given by (Chambers & Clark, 2012)

$$V(\bar{y}_{\text{REGB}}) = \left[\sigma^2\left((N-n) + \boldsymbol{\tau}_x'(\boldsymbol{X}_{k\in S}'\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{\tau}_x\right)\right]/N^2, \tag{3.73}$$

where

$$\boldsymbol{\tau}_x = \sum_{k\in U-S} \boldsymbol{x}_k. \tag{3.74}$$

Finally, the prediction variance is estimated by (Chambers & Clark, 2012)

$$\hat{V}(\bar{y}_{\text{REGB}}) = \left[\hat{\sigma}^2\left((N-n) + \boldsymbol{\tau}_x'(\boldsymbol{X}_{k\in S}'\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{\tau}_x\right)\right]/N^2 \tag{3.75}$$

where

$$\hat{\sigma}^2 = (n-J)^{-1} \sum_{k \in S} (y_k - \boldsymbol{x}_k' \hat{\boldsymbol{\beta}})^2 = (n-J)^{-1} \sum_{k \in S} \epsilon_k^2. \tag{3.76}$$

3.3. The role of the model

Särndal *et al.* (1992, page 239) note that "We do not require that the model be 'true' in the sense of depicting the process by which the population data may have been generated. We only believe that the population data can be fairly well described by the model". This holds true for the (design-based) model-assisted approach. Having sample data at hand, the challenge an analyst faces is to express "fairly well" in quantitative terms.

As shown by Särndal *et al.* (1992), the regression estimator is consistent for the population mean (or total) even if the model does not mimic the data generating process. For variance estimation, however, the situation is less clear. Given a sample of fixed size n and $J > 2$ the estimators given in the preceding Section 3.1.4 will provide different estimates of precision.

For example, the estimator given in (3.42), called the simple estimator hereafter, takes no account of the number of auxiliary variables that enter the model. The estimated variance is solely based on the regression residuals. Asymptotically, the estimator (3.42) is, for a given model, approximately unbiased, that is, if $n \to \infty$ and $N \to \infty$. However, in many surveys, including most FRIs, the sample size is small relative to the population size. Asymptotic properties may therefore be of little comfort.

If the simple variance estimator is used, the estimated variance will be smaller the better the model fits the sample data. If many auxiliary variables are available for only a few observations an analyst should have no problem to identify a model that provides "good" predictions for observed values. From standard regression theory it is known that the more variables are added, the higher the R^2. For that reason, adding more variables will always increase the precision of the regression estimator. However, when the same model is fitted to a different sample the residual variance will likely be larger. Unless an analyst is so fortunate that an estimated regression model is also nearly optimal for the entire population one should — in theory at least — expect that the average residual variance for the same model when computed over all possible samples (minus the one

that was observed) would be greater. Blind application of the simple variance estimator may, thus, lead to overly optimistic estimates of variance.

In most model-assisted applications of LiDAR in FRIs, the simple estimator (3.42) was employed (see e.g., d'Oliveira *et al.* (2012); Strunk *et al.* (2012)). As noted in Chapter 2, the full set of available LiDAR metrics has rarely been used. Instead, different (automated) variable selection procedures, such as the AIC or BIC, are employed which aim at removing those variables that are not related to the target variable. Nonetheless, as Strunk *et al.* (2012) noted "Unfortunately, model selection can result in overly optimistic inferences [...]. In our case we attempted to protect ourselves from this optimism by combining automated model selection with expert knowledge. This reduces the chance that a best model is selected due to an artifact in the data which may only effectively represent a trend present in the sample." That variable selection can lead to an overestimation of precision in a model-assisted context has also been mentioned by Silva & Skinner (1997). An interesting example of the consequences of variable selection on the R^2, on which the simple variance estimator depends, is provided by Chatfield (1995, page 423, Example 3).

It is often not clear which approach one should use to identify a model that delivers a reliable estimate of precision. Strunk *et al.* (2012) used best-subset selection based on the BIC to select the — supposedly — "best" model. A different approach or different selection criteria may have lead to a different model. Given the many techniques for variable selection that an analyst might consider for LiDAR applications, the choice he makes may directly impact estimates of precision.

In contrast to the simple variance estimator, the estimator (3.52) takes into account the number of auxiliary variables that enter the model. If $J > 2$ the estimated variance will, thus, be larger if applied to the same sample (and fitted model).

The estimator in (3.55) explicitly takes into account the uncertainty in the auxiliary variables. This can be seen when looking at the definition of the g-weights given in (3.35). Here, the term $\boldsymbol{X}'_{k\in S}\boldsymbol{X}$ usually becomes more variable the more explanatory variables are added to the model (Silva & Skinner, 1997). However, in model-assisted FRI applications the latter two estimators have rarely been applied.

In contrast to the model-assisted approach, where the distribution of possible estimates permissible under the design is the basis for inference, in the model-based approach the inference is based on the model. Thus, in the latter approach the specification of the model directly influences point estimates, such as the predicted mean or total. Choosing

the model carefully becomes even more important in the model-based than in the model-assisted approach. Moreover, in the model-based approach assumptions are made about the distribution of the regression residuals, ϵ_k, namely that they are independently and identically distributed with zero mean and constant variance (Lohr, 1999). As a consequence, model diagnostics, e.g., plotting standardized residuals, becomes an integral part of model formulation in the model-based framework.

James *et al.* (2013, page 92) provide a list of potential challenges that may occur when a linear regression model is fitted to a particular dataset:

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

Not all challenges are equally relevant in model-based inference. While challenges 1–5 certainly merit consideration, number 6 is probably of minor importance when the aim of using a model is prediction. However, for variance estimation, multicollinearity may, in the model-based approach, have an impact. As it is the case for the g-weighted variance estimator, the model-based estimator for the prediction variance includes the therm $\boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S}$. Reducing the set of variables to only those that are strongly related to the target variable (and do not correlate strongly with each other) may therefore lead to a decrease of the prediction variance. Hence, variable selection and collinearity most likely affect estimates of precision in the model-based approach.

Note, in the model-assisted framework none of the challenges lead to invalid inference. However, point 1, 4 and 5 may decrease the efficiency of the estimation. For the simple model-assisted variance estimator (that depends on the regression residuals only) and the estimator given in (3.52) collinearity does not directly affect precision. However, for the g-weighted variance estimator (3.55) collinearity may be of concern.

Over the past decades numerous modeling techniques have been developed that try to tackle specific challenges in modeling. To approach problems where the number of covariates is large relative to the number of observations, ridge regression (Hoerl &

Kennard, 1970) or partial least squares (PLS) regression might, for example, be considered. These two techniques may also prove useful for problems of multicollinearity. In a model-assisted context, a variable selection procedure that takes into account the condition number of the auxiliary variables, i.e., the largest eigenvalue of the cross-product matrix $CP = \boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S}$ divided by the smallest eigenvalue of CP, was proposed by Silva & Skinner (1997). The least absolute shrinkage and selection operator, or LASSO (Tibshirani, 1996), provides another option.

Non-parametric modeling approaches have rarely been applied in model-assisted estimation. An example of the application of generalized additive models is provided by (Opsomer *et al.*, 2007). Semi-parametric modeling techniques have been used by Breidt *et al.* (2005, 2007). Although random forests (RF) may not fit well into the "classical" model-assisted inference framework, they do have some features that may prove useful when working with datasets that exhibit complex structures — such as LiDAR data.

In this study, several model and variable selection procedures were considered. The following chapter provides a brief overview. The choice of procedures was largely based on practical considerations, rather than on theoretical justifications. Procedures that are commonly employed in LiDAR-assisted FRIs were selected (see Chapter 2). The different techniques will be described briefly; more detailed information can be found in, for example, Hastie *et al.* (2009), Fahrmeir *et al.* (2013), and James *et al.* (2013).

4. Modelling

4.1. Linear regression

4.1.1. Full model

When a set of covariates is available a data analyst may decide to include all of them into the model. This saturated or full model was the first model that was considered in this study. Model coefficients are estimated in analogy to the methods described in Chapter 3. As mentioned in Chapter 2, the full model is rarely applied in LiDAR-assisted FRIs. However, it was included as a "benchmark" model in this study.

4.1.2. Stepwise selection

Different stepwise selection procedures were considered as they are widely used in LiDAR-applications (e.g., Næsset, 2002; Gobakken *et al.*, 2012).

In stepwise selection a subset of the J available variables is selected from the full set of covariates following statistical selection criteria. One can distinguish between three different approaches: (a) forward stepwise selection, (b) backward stepwise selection, and (c) a hybrid approach.

Forward stepwise regression starts with a model containing no covariates. Then variables are added sequentially. At each step the covariate that provides the greatest additional improvement to the fit is added (James *et al.*, 2013). Forward stepwise selection is a so-called *greedy algorithm* producing a nested sequence of models (Hastie *et al.*, 2009). Different criteria such as the AIC, or BIC (see below) are used to evaluate improvements (or decline) in model performance. In forward selection the procedure terminates if no further improvement is possible (Fahrmeir *et al.*, 2013).

An alternative to forward stepwise selection is backward stepwise selection, or backward elimination. Here, the full model containing all potential covariates is considered

first. Variables are iteratively removed that lead to the greatest improvement of model performance. The procedure terminates if no further improvements are possible.

A hybrid stepwise selection approach is a combination of both, forward and backward selection. Here, variables are added to the model at each iteration. However, after adding a covariate the algorithm may also remove a covariate which no longer provide an improvement of model performance. In this study the hybrid approach has been applied, as this approach best mimics best-subset selection (James *et al.*, 2013). In best-subset selection all 2^J possible models are separately fitted to the data and the best model, according to a predefined criteria, is selected. However, because of the large number of iterations in the simulation studies, best-subset selection was computationally prohibitive when the full set of covariates was considered.

Akaike Information Criterion (AIC)

For stepwise selection algorithms a criterion needs to be defined that determines whether adding (or removing) a covariate leads to an improvement (or decline) in model performance. The fist criterion that was used in this study was the Akaike Information Criterion (AIC) defined as

$$\text{AIC} = -2\ \log(\mathcal{L}) + 2J \tag{4.1}$$

where J is the number of coefficients in the model, and $\log(\mathcal{L})$ is the logarithm of the maximized likelihood function of the estimated model. For OLS the AIC can alternatively expressed as

$$\text{AIC} = n\ \log(s_e^2) + 2J$$

where (as defined in Chapter 3)

$$s_e^2 = (n-1)^{-1} \sum_{k \in S} e_k^2.$$

Corrected Akaike Information Criterion (AICc)

The term $2J$ in equation (4.1) penalizes model complexity. However, the AIC does not necessarily lead to the most parsimonious model, and there is a risk of overfitting (Claeskens & Hjort, 2008). A corrected version of the AIC is the AICc that has been proposed by Hurvich & Tsai (1989). The AICc is defined as

$$\text{AICc} = \text{AIC}\frac{2J(J+1)}{n-J-1}. \tag{4.2}$$

The AICc puts a stronger penalty on the number of parameters in the model and has been recommend for small n (Burnham & Anderson, 2002). The same authors suggested to always employ the AIC instead of the AICc, as the latter converges to the AIC, if n gets sufficiently large.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) developed by Schwarz (1978) is closely related to the AIC. The only difference is that for the BIC the number of covariates that enter the model is multiplied by $\log(n)$ instead of 2. The BIC is defined as

$$\text{BIC} = -2\,\log(\mathcal{L}) + \log(n)\;J. \tag{4.3}$$

The BIC, like the AICc, puts, thus, a stronger penalty on the number of covariates compared to the AIC.

AIC and variance inflation factor (VIF)

When the aim of using a model is prediction only, multi-collinearity is often of minor concern (Burnham & Anderson, 2002). However, as has been noted in Section 3.3 in Chapter 3, multicollinearity may affect estimates of precision for some variance estimators. Therefore, variance inflation factors (VIFs) have been computed for the sampled

data in order to identify highly correlated covariates. The VIF is defined as (Fahrmeir *et al.*, 2013)

$$\text{VIF} = \frac{1}{(1 - R_j^2)} \tag{4.4}$$

where R_j^2 is obtained from a regression of $\boldsymbol{x}_j$ onto all other covariates $\boldsymbol{x}_{J-j}$.

In this study, the VIF was first calculated for each covariate using the full model. If any VIF_j was above or equal 10, the variable with the highest VIF was removed. Next, the model was refitted and the VIF was computed again for all covariates that remained in the model. The procedure was repeated until no covariate with a VIF larger 10 remained in the model.

Since the above procedure does not remove variables that are not related to the target variable, the final model was selected using a stepwise procedure based on the AIC (as described above).

Best-subset selection and variance inflation factor

The VIF procedure was also combined with best-subset selection. First, the VIF was used to remove highly correlated variables (as described above), and afterwards best-subset selection was used to choose the best model from the set of candidate models. Mallows' C_p statistic was used to identify the final model. The C_p statistic is computed as

$$C_p = n^{-1}(\sum_{k \in S}[y_k - \hat{y}_k]^2 + 2Js_e^2).$$

Like the AIC, Mallow's C_p puts a penalty on the number of variables that enter the model.

Condition number

Silva & Skinner (1997, page 26) propose the following variable selection procedure (which is a a modification of the procedure originally developed by Bankier *et al.* (1992)):

1. Compute the cross-products matrix CP = $\boldsymbol{X}^{*\prime}_{k\in S}\boldsymbol{X}^{*}_{k\in S}$ considering all the columns initially available (saturated subset).

2. Compute the Hermite canonical form of CP, say H (see Rao (1973, page 18)), and check for singularity by looking at the diagonal elements of H. Any zero diagonal elements in H indicate that the corresponding columns of $\boldsymbol{X}^{*\prime}_{k\in S}\boldsymbol{X}^{*}_{k\in S}$ (and $\boldsymbol{X}^{*}_{k\in S}$) are linearly dependent on other columns (see Rao (1973, page 27)). Each of these columns is eliminated by deleting the corresponding rows and columns from $\boldsymbol{X}^{*\prime}_{k\in S}\boldsymbol{X}^{*}_{k\in S}$.

3. After removing any linearly dependent columns, the condition number $c = \lambda_{\max}/\lambda_{\min}$ of the reduced CP matrix is computed, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest of the eigenvalues of CP, respectively. If $c < L$, a specified value, stop and use all the auxiliary variables remaining.

4. Otherwise perform backward elimination as follows. For every k, drop the kth row and column from CP, and recompute the eigenvalues and the condition number of the reduced matrix. Compute the condition number reductions $r_k = c - c_k$ where c_k is the condition number after dropping the kth row and column from CP. Determine $r_{\max} = \max_k(r_k)$ and $k_{\max} = \{k : r_{\max} = r_k\}$ and eliminate the column $k_{\max}$ by deleting the $k_{\max}$ row and column from CP. Make $c = c_{k_{\max}}$ and iterate while $c \geq L$ and $q \geq 2$, starting each new iteration with the reduced CP matrix resulting from the previous one.

Note, $\boldsymbol{X}^{*}_{k\in S}$ above is similar to $\boldsymbol{X}_{k\in S}$ (defined in 3.30) except that a vector of 1's of length n was added as a first column. For L a value of 30 was chosen.

4.1.3. Regularization

Ridge regression

OLS coefficient estimates may become numerically unstable when the ratio between the the number of observations and the number of covariates is small. This problem is often referred to as the "small n, large P" (where P is the number of covariates) problem. Similar problems surface in case of multicollinearity. Regularization techniques may prove useful in these situations. In this study two regularization techniques were considered: ridge regression and the least absolute shrinkage and selection operator (LASSO).

Ridge regression is not very different from OLS, except that a slightly different quantity is minimized (James *et al.*, 2013). Ridge regression coefficients are obtained by minimizing

$$\text{RSS} + \lambda \sum_{j=1}^{J} \beta_j^2 \tag{4.5}$$

where $\lambda \geq 0$ is a tuning parameter which needs to be determined separately, and

$$\text{RSS} = \sum_{k \in S} \left(y_k - \hat{y}_k \right)^2.$$

where RSS is the residual sum of squares. The ridge estimator may be rewritten as

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'_{k \in S} \boldsymbol{X}_{k \in S} + \lambda \boldsymbol{I}_J)^{-1} \boldsymbol{X}'_{k \in S} \boldsymbol{y}_{k \in S} \tag{4.6}$$

where $\boldsymbol{I}_J$ is the squared $J \times J$ identity matrix with ones on the main diagonal and zeros elsewhere.

Ridge regression corresponds to OLS when $\lambda = 0$. However, whenever $\lambda > 0$ the ridge estimator will shrink the coefficients towards zero. If $\lambda \to \infty$, the coefficients will become zero.

The initial motivation of developing the ridge estimator was to solve non-singular problems, i.e., when $\boldsymbol{X}'_{k \in S} \boldsymbol{X}_{k \in S}$ is not of full rank (Hoerl & Kennard, 1970). However, the estimator (4.5) may also be useful when covariates are highly correlated. Ridge regression tries to alleviate the problem of multicollinearity by adding a size constraint on the coefficients (Hastie *et al.*, 2009).

Note that for different λ different estimates of coefficients will be obtained. Also note, that ridge solutions are not equivariant when covariates are scaled (Hastie *et al.*, 2009). In this study, covariates have been normalized, i.e.,

$$\boldsymbol{x}_j^* = \frac{\boldsymbol{x}_{kj} - \bar{\boldsymbol{x}}_j}{\boldsymbol{s}_{x_j}}.$$

To determine λ cross-validation may be used. For a grid of λ values the cross-validation errors is computed, and usually the largest value within one standard error of the min-

imum λ was selected. The minimum refers to the value of λ where the cross-validated error is smallest.

Least absolute shrinkage and selection operator

Closely related to ridge regression is the least absolute shrinkage and selection operator (Lasso). As for the ridge estimator a penalty term is added to the OLS solution. The form of the penalty is, however, different for the Lasso. The Lasso coefficients minimize the quantity (Hastie *et al.*, 2009)

$$\mathrm{RSS} + \lambda \sum_{j=1}^{J} |\beta_j| \tag{4.7}$$

The Lasso uses a so-called L_1-norm penalty, whereas for ridge regression the L_2-norm is used. A notable difference between the two is that the Lasso shrinks coefficients to exactly zero when λ is sufficiently large. In ridge regression coefficients will not become exactly zero unless $\lambda = \infty$. The Lasso, therefore, *selects* variables (like the procedures above), whereas ridge regression *only shrinks* the coefficient towards zero.

Note, by adding the penalty term to the Lasso and ridge, the least squares estimate is not unbiased anymore. Both approaches trade-off between variance and bias, that is, a small bias is introduced to obtain a large reduction in variance.

4.2. Partial least squares regression (PLSR)

As an alternative to selecting variables or shrinking coefficients, one may consider to transform the covariates that enter the regression model. These techniques are referred to as *dimension reduction* techniques. Here, the number of explanatory variables is reduced and this reduced set of (transformed) variables is then used in the classical (multiple) regression setting.

Common techniques of dimension reduction include principal component analysis regression (PCR) and partial least squares regression (PLS). For PCR linear combinations, or directions, of covariates $X_1, X_2, \ldots, X_j, \ldots, X_J$ are computed. These combinations are called principal components $Z_1, Z_2, \ldots, Z_M$, where $J > M$. In that way the dimension of the $n \times J$ covariate matrix, $\boldsymbol{X}_{k \in S}$, is reduced. The first principal component lies

in the direction of the largest variation in the data, the second in the second largest direction, and so forth (Faraway, 2004). The key idea of PCR is to obtain a small set of transformed variables that are able to capture the variability in the response sufficiently well. A drawback of PCR is that when directions are identified the response variable is ignored. PLS, on the other hand, extracts linear combinations of the covariates, that explain both response and predictor variation (James *et al.*, 2013).

4.3. Random forests (RF)

In this study the performance of a single non-parametric method was investigated, namely the random forest (RF) algorithm. Although applications of RF are relatively recent in FRIs, they seem to become increasingly prominent (Brosofske *et al.*, 2014).

RF is based on regression trees, and is, therefore, conceptually very different from "classical" linear regression setting.

Building a regression tree involves at least two steps (James *et al.*, 2013).

1. The covariate space[1], i.e., the set of possible values for the J explanatory variables, is divided into D distinct and non-overlapping regions, $R_1, R_2, \ldots, R_d, \ldots, R_D$.
2. For every observation that falls within the same region R_d, the same prediction is made. This prediction is simply the mean of the response variable of all observations within the region R_d.

The goal is to identify a configuration of regions, $R_1, R_2, \ldots, R_d, \ldots, R_D$, that minimizes

$$RSS = \sum_{d=1}^{D} \sum_{k \in R_d} (y_k - \hat{y}_{R_d})^2,$$

where $\hat{y}_{R_d}$ is the mean response for the observations in the dth region.

Splitting the feature space into all possible partitions is usually computationally infeasible. Therefore, a greedy approach known as *recursive binary splitting* is commonly employed. In this top-down approach a first covariate is selected that splits the entire set of responses into two regions R_1 and R_2 given a particular cutpoint, a, of the covariate. The covariate and the cutpoint are chosen in such a way that the reduction of

[1] In machine-learning terminology this space is usually referred to as the *feature space*.

the RSS is maximized. This partitioning is repeated within the two regions R_1 and R_2. Once subregions are obtained the process continues until a predefined stopping rule is reached, e.g., until no region contains more than 5 observations (James *et al.*, 2013). The procedure of partitioning results in a tree-like structure and that is why this technique is called regression tree. The tree can be used to make predictions for combinations of the covariates that are not present in the sample.

The problem of regression trees is, that the structure in the *sample* is very well described but the tree may perform poorly when predictions are to be made for formerly unseen data. An alternative approach is to not construct a single tree but to consider many.

In bagging, a random sample with replacement is drawn from the original sample S and a regression tree is constructed for this so-called bootstrap sample. A large number of bootstrap samples, e.g., 500, are drawn and for each sample a tree is constructed. To predict the response for the kth observation the average of predictions made from all trees may be used. However, it can be shown that on average, each bagged tree uses roughly two-thirds of the observations in the original sample S when bootstrapping is used (James *et al.*, 2013). The observations not used are referred to as out-of-bag (OOB) observations. Instead of using the average of all trees to make a prediction, the prediction for the kth observations can be obtained by only considering those trees where the kth observation was not included. Conceptually, this approach corresponds to cross-validation (Hastie *et al.*, 2009).

It has to be considered that the trees may all exhibit a very similar structure. This frequently happens when there is one very strong predictor and several other predictors that are only moderately associated with the response. It is likely that the strong predictor will be used at the top split in most — or probably all — trees. To alleviate the problem of correlated trees, Breiman (2001) proposed the random forest (RF) algorithm. For RFs not all features are used for partitioning the feature space in a tree. At each split a random sample of usually $m \approx \sqrt{J}$ predictors is considered. For that reason, in on average $(J - m)/J$ of the splits the strong predictor will not be considered. This leads to decorrelating the trees (James *et al.*, 2013) and usually results in more stable predictions.

5. Objectives

5.1. Objectives, hypothesis & research questions

As outlined in Chapter 2 and 4, a wide range of modeling techniques are used in LiDAR-assisted FRIs. In Chapter 3 it has been shown that an analyst does not only have to choose between different modeling techniques, but also among the two different inference frameworks. Moreover, within the two frameworks a choice has to be made which variance estimator one should employ. Most, if not all of the decision an analyst makes may directly affect estimates of precision.

The objective of this study is to assess the effect of different statistical model and variable selection procedures on estimates of precision in a model-assisted and model-based inference framework. The focus is on if and how these procedures affect precision estimates in LiDAR-assisted FRIs.

In this study the following hypotheses are made:

1. Different statistical model and variable selection procedures lead to different estimates of precision, regardless of which estimator is used. This holds true in the model-assisted, as well as in the model-based inference framework.
2. Model-assisted inference:
    a) When variance estimators are used that are based solely on model residuals, blind application of stepwise variable selection procedures, such as the AIC or BIC, lead to overly optimistic estimates of precision. Currently applied methods in LiDAR-assisted FRIs underestimate variances of target parameter estimates.
    b) "Modern" modeling techniques, such as the Lasso, lead to unbiased estimates of precision, and therefore outperform stepwise variable selection procedures.

c) The random forest algorithm can capture associations in complex datasets and, therefore, provides unbiased estimates of precision when applied to the simple model-assisted variance estimator.

3. When variance estimators are used that account for the variability in the auxiliary variables, variable selection procedures based on the condition number of the auxiliary variable matrix will improve efficiencies and lead to correct estimates of precision in LiDAR-assisted FRIs.

4. Using stepwise variable selection procedures will generally lead to overly optimistic estimates of precision in LiDAR-assisted FRIs, regardless of what modeling technique, variance estimator or inference framework is selected.

5.2. Structure of this document

The remainder of this document is organized as follows. In Chapter 6 the five datasets used in this study are described. Three datasets are "artificial" datasets; they have been simulated for this study. The remaining two datasets are based on FRI and LiDAR data from Canada and Norway. Synthetic populations have been created from the latter two. The methods used to obtain the synthetic populations are described in detail in Section 6.4.

The impact of model and variable selection on estimates of precision was assessed in simulation studies. Chapter 7 provides an overview of how these simulation studies have been carried out. Detailed information is provided on which estimators were considered and how estimates where computed.

In Chapter 8 the results of the simulation studies are presented for the model-assisted framework. First, results for the artificial datasets are provided. In Section 8.2 and 8.3 results for the two synthetic populations are given. Chapter 9 gives the results of the simulation study for model-based inference.

In Chapter 10 the results are discussed in detail and in Chapter 11 general conclusion and an outlook that does also embrace suggestions for further research are provided.

Part II.

Materials & Methods

6. Data

6.1. Artificial datasets

Three artificial dataset were generated for this study. All datasets consist of $N = 1000$ observations and $J = 21$ variables. The first variable in each dataset is denoted Y, and the remaining 20 variables are denoted $X_1, X_2, \ldots, X_{20}$. The values of each variable were randomly sampled from a normal distribution $Z \sim \mathcal{N}(\mu, \sigma^2)$. The first variable, Y, will serve as the target variable and the remaining 20 variables, $X_1, X_2, \ldots, X_{20}$ as auxiliary variables.

For the first dataset (NOISE) values have been sampled independently for each variable, with $\mu = 100$ and $\sigma^2 = 100$. The Pearson product-moment correlation coefficient, ρ, is approximately zero among all variables, i.e., $\rho_{Y|X_j} \approx 0$. For the simulated data the absolute average $\bar{\rho}_{Y|X_j}$ was 0.026. This dataset was created to evaluate the performance of model and variable selection procedures when the target variable is not related to any of the covariates (and the covariates not with each other).

The second dataset (COR) consist of 21 variables that strongly correlate with each other. The average correlation coefficient is $\bar{\rho}_{Y|X_j} = 0.95$, and is roughly equal across all X_js. The average correlation between the X_js is $\bar{\rho}_{X_j|X_{j+1}} = 0.90$. This dataset was created to assess how variable selection procedures perform when the covariates are strongly correlated with each other (and with the target variable).

In the third dataset (DCOR) the correlations between Y and $X_1, X_2, \ldots, X_{20}$ constantly decrease. For the first covariate X_1 the correlation coefficient was $\rho_{Y|X_1} = 0.93$, for X_2 it was $\rho_{Y|X_2} = 0.74$, $\rho_{Y|X_3} = 0.51$, $\rho_{Y|X_4} = 0.38$, $\rho_{Y|X_5} = 0.29$, $\rho_{Y|X_6} = 0.21$, $\rho_{Y|X_7} = 0.13$, $\rho_{Y|X_8} = 0.10$, $\rho_{Y|X_9} = 0.07$, $\rho_{Y|X_{10}} = 0.08$, $\rho_{Y|X_{11}} = 0.0.08$, $\rho_{Y|X_{12}} = 0.08$, $\rho_{Y|X_{13}} = 0.02$, $\rho_{Y|X_{14}} = 0.06$, $\rho_{Y|X_{15}} = 0.05$, $\rho_{Y|X_{16}} = -0.02$, $\rho_{Y|X_{17}} = 0.07$, $\rho_{Y|X_{18}} = 0.05$, $\rho_{Y|X_{19}} = 0.03$, $\rho_{Y|X_{20}} = 0.01$. The correlation among covariates was low except for X_1 and X_2 ($\rho_{X_1|X_2} = 0.68$), and for X_1 and X_3 ($\rho_{X_1|X_3} = 0.36$). The motivation to create

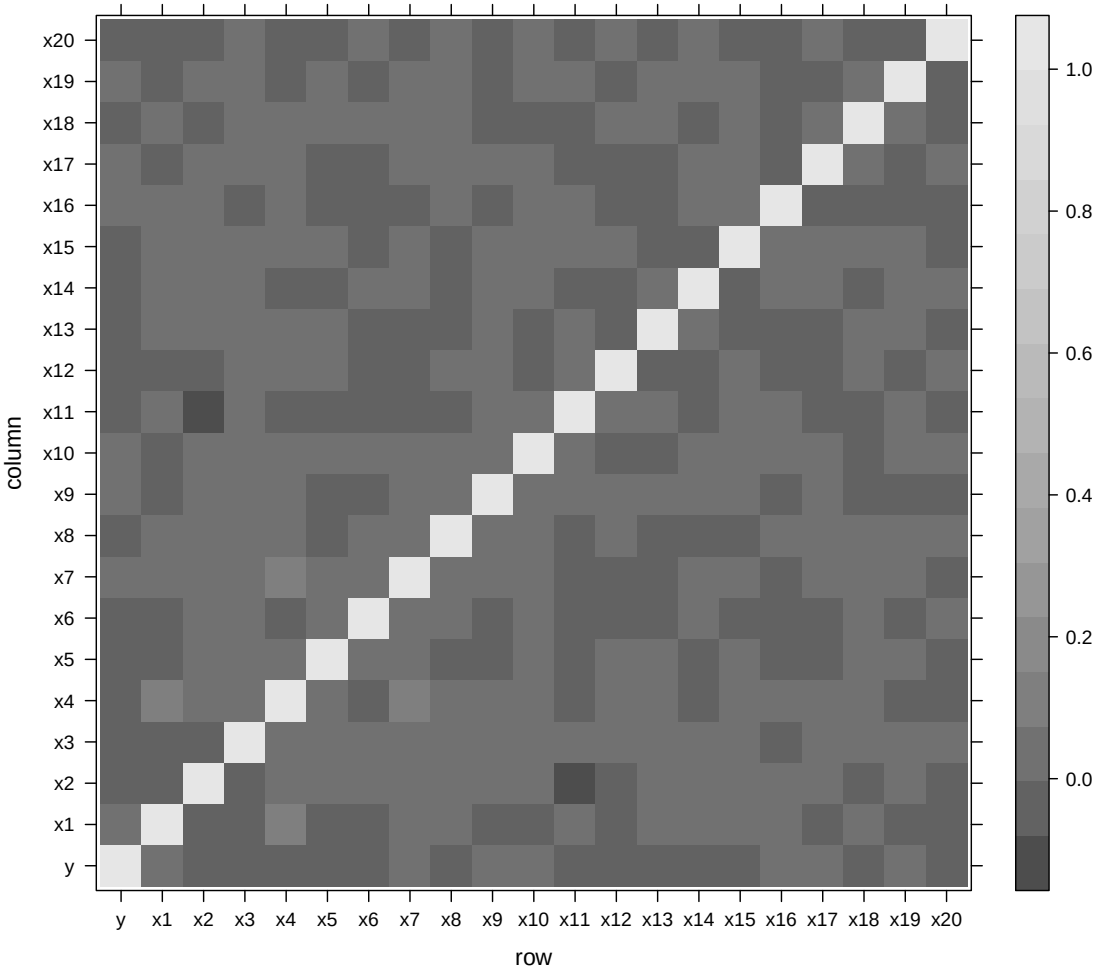

Figure 6.1.: Correlation structure in the artificial datasets NOISE (the scale bar refers to the Pearson correlation coefficient).

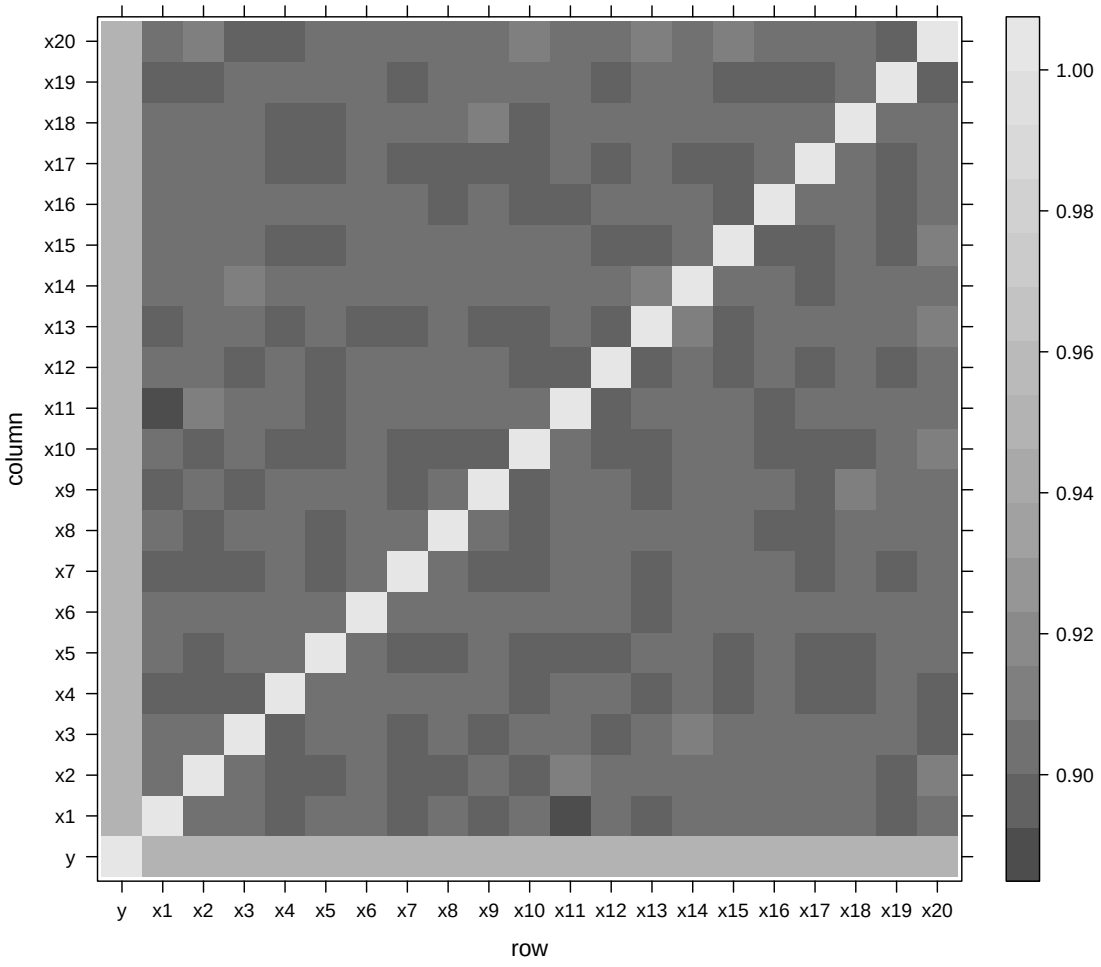

Figure 6.2.: Correlation structure in the artificial datasets COR (the scale bar refers to the Pearson correlation coefficient).

this dataset was to evaluate how variable selection procedures perform when the dataset consists of a few strong predictors and many unrelated covariates.

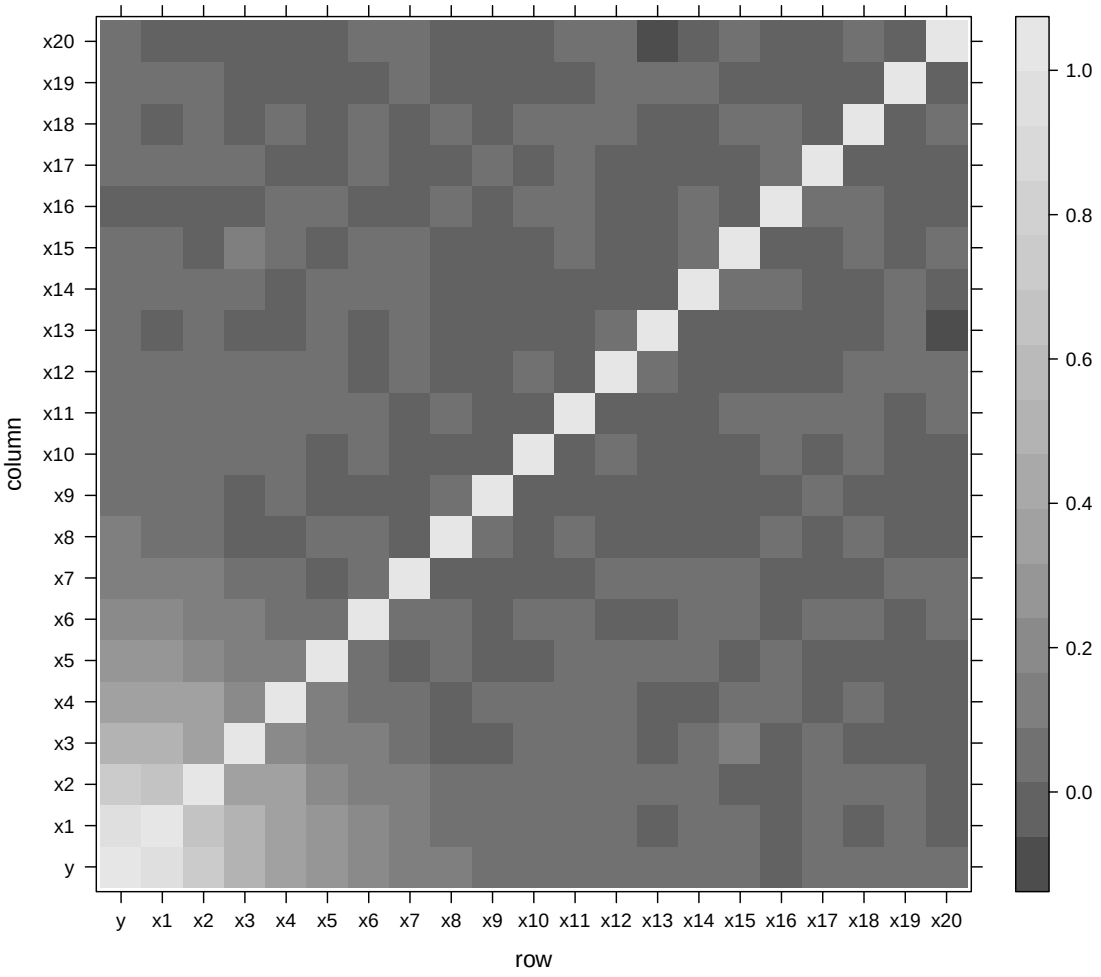

Figure 6.3.: Correlation structure in the artificial datasets DCOR (the scale bar refers to the Pearson correlation coefficient).

All datasets were generated using the statistical language and environment R (R Core Team, 2015). The code can be found in Appendix A. Figures 6.1, 6.2, and 6.3 depict the general structure of the simulated datasets.

LiDAR data often exhibit complex data structures, i.e., multi-collinearity, "white noise", non-linear effects, etc. The three artificial datasets provide a useful basis to assess the effect of these structures on model and variable selection separately.

6.2. Hinton (HIN)

6.2.1. Study area

Hinton Wood Products (HWP), a division of West Fraser Mills Ltd., manages a large Forest Management Agreement (FMA) area in west-central Alberta, Canada. The FMA covers almost one million hectares and is comprised of five natural sub-regions: Upper Foothills, Lower Foothills, Montane, Sub-Alpine and Alpine (see Natural Regions Committee (2006)). Large chapters of the area consist of pure coniferous stands (80%), while the remaining 20% consist of pure deciduous (8%) and mixed stands (12%) (Hinton Wood Products, 2010). The dominant tree species in the coniferous areas is lodgepole pine (*Pinus contorta* Douglas).

The dataset from Hinton (HIN hereafter) was provided by the Pacific Forestry Center (PFC), Canadian Forest Service, Natural Resources Canada (Joanne White and Mike Wulder). The following description of the field and LiDAR datasets is largely based on Frazer *et al.* (2011a) and White *et al.* (2013).

6.2.2. Field data

HWP maintains a Permanent Growth Sample (PGS) program consisting of more than 3,200 fixed area sample plots (Frazer *et al.*, 2011a). In this study only plots that have been remeasured since 2002 and for which the expected planimetric error in GPS (Global Positioning System) plot positioning was available were used. Data from in total $n = 788$ sample plots were available[1]. The plots are systematically spread over the study area (see Figure 6.4). Plot centers were established at the intersections of the Alberta legal survey grid section lines (Hinton Wood Products, 2008), approximately 2 kilometers apart from each other.

[1]The original dataset consisted of $n = 957$ field plots. Several plots have been identified as outliers during the development of the calibration dataset (see Frazer *et al.* (2011a) and Section 6.2.4).

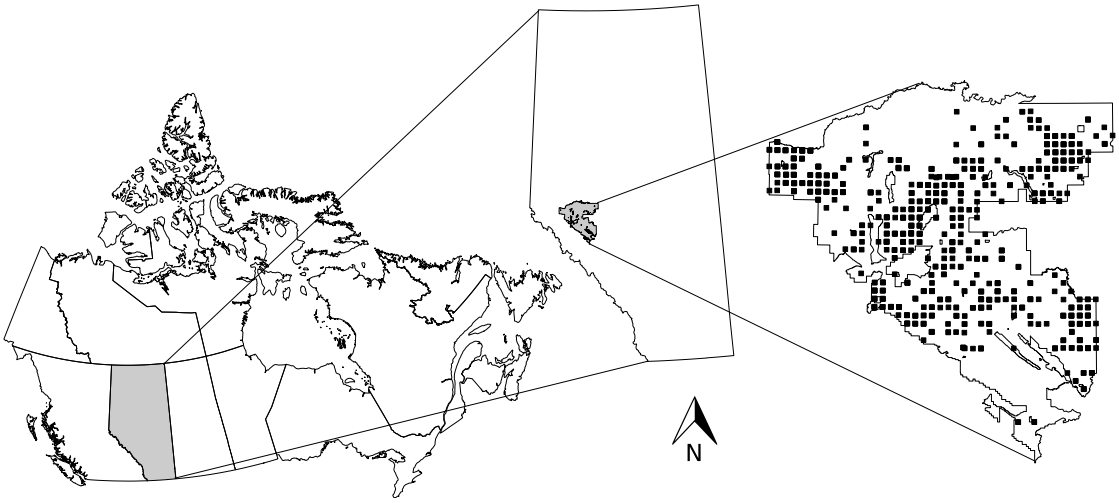

Figure 6.4.: Location of the HIN study area. Left: Canada; middle: Province of Alberta; right: Hinton Wood Products Forest Management Agreement (FMA) area. Black squares show field plot locations.

Each of the 788 PGS squared plots is either of size 20.1×20.1 m (i.e., 0.04 ha) or 28.5×28.5 m (i.e., 0.08 ha). Small plots were arranged in clusters of four subplots (see Figure 6.5), whereas large plots consist of only one plot. Locations of plot centers were determined with unknown planimetric accuracies using GPS.

On each plot tree attributes were recorded on trees (in total 55,652) that were $\geq$ 2 m in height. For each plot an estimate of the total live aboveground biomass (AGB) in megagrams per hectare (Mg ha^{-1}) was available. AGB estimates were obtained from measurements of diameter at breast height (DBH; in cm), tree height (m), species code, crown-class mode, and other data obtained for each tree (see Hinton Wood Products (2008) for details). Estimates were provided by Hinton Wood Products. Biomass components, such as branches, bark, and foliage, were estimated using height and diameter-based national allometric regression equations from Lambert *et al.* (2005) and Ung *et al.* (2008).

Three different stand heights (i.e., average height [m] of the tallest 100 trees ha^{-1} [using four and eight stems per .04 ha and .08 ha plot, respectively]), average height (m), and the 75^{th} percentile of tree heights (m) above 7.1 cm DBH) were computed for each plot. In addition the quadratic mean diameter (cm) and basal area (m^2 ha^{-1}), were calculated and estimates of merchantable stem volume (m^3 ha^{-1}), and total stem volume (m^3 ha^{-1}) were obtained. Summary statistics for the HIN field data is provided in Table 6.1.

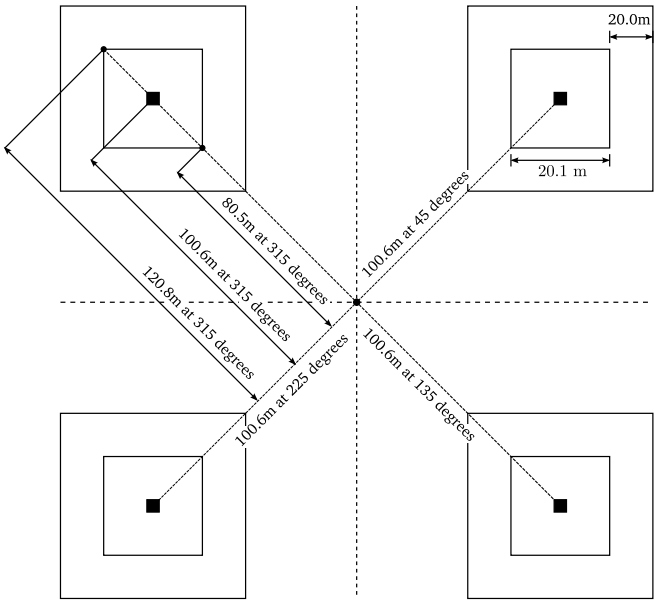

Figure 6.5.: HIN cluster plot (large plots consist of a single square of 28.5×28.5 m; see Hinton Wood Products (2008) for details).

Table 6.1.: Descriptive statistics for ground-reference measurements obtained from the PGS dataset. The majority of PGS plots belonged to the 'conifer' forest cover type (n = 572), followed by 'mixed' (n = 129) and 'deciduous' (n = 87) cover classes (taken from Frazer *et al.* (2011a)).

Attributes	Units	Min.	Q1[a]	Median	Q3[b]	Max.	Mean	SD[c]	N[d]
Basal area	m^2 ha^{-1}	0.01	6.42	17.14	30.81	64.07	19.40	14.82	786
Height (Top)	m	2.10	8.60	13.12	19.84	34.89	14.24	7.31	788
Height (Mean)	m	2.10	5.84	9.30	13.89	26.73	10.39	5.56	788
Height (75th Pct.)	m	3.90	8.63	11.65	17.15	32.20	13.20	6.07	725
QMD[e]	cm	1.20	9.15	11.91	18.29	38.83	13.91	7.21	788
Volume (merch.)	m^3 ha^{-1}	0.47	13.29	53.96	196.40	639.20	115.60	131.09	596
Volume (total)	m^3 ha^{-1}	0.01	20.70	81.18	203.60	659.20	128.30	132.38	779
Biomass (total)	Mg ha^{-1}	0.01	19.22	60.70	135.80	385.00	84.61	77.95	787

[a]Q1 = 25^{th} percentile
[b]Q3 = 75^{th} percentile
[c]SD = standard deviation
[d]N = total number of PGS plots in the sample
[e]QMD = quadratic mean diameter

6.2.3. LiDAR data

Between 2004 and 2007, multiple-discrete-return (maximum 4 laser returns), small-footprint (< 30 cm) LiDAR data were collected by fixed-wing aircraft for the entire HIN study area. The bulk of the data were acquired in 2005 and 2006. The nominal post spacing was ≈ .75 point per square meter. All LiDAR points (x, y, and z coordinates) were georeferenced using a UTM Zone 11 North projection, and NAD83 (horizontal) and CGVD28 (vertical) datums. Points were classified as either 'ground' or 'non-ground' points using TerraScan software[2]. Points classified as 'ground' were used to construct a one meter 'bare-earth' digital elevation model (DEM). The DEM was used to derive above ground LiDAR point heights. The freeware FUSION/LDV[3] (McGaughey, 2013) was used to compute LiDAR canopy height and density metrics, for grid cells of size 25×25 m. A grid cell size of 25 m meters was chosen as a compromise between the two different plot sizes of 20.1 and 28.5 m.

The original dataset consisted of 13,885,234 grid cells. In this study only a subset of 157,053 grid cells was used. These grid cells covered all forest stands in which field plots were established (in total 552 stands). All cells that covered non-sampled stands were removed.

Table 6.2 provides an overview of the available LiDAR metrics. Some metrics were not available for the plot data (m15, m17, m19, m20), these metrics have been removed from the LiDAR dataset. As plot sizes differ, the minimum and maximum point heights (metric m6 and m7) were removed. The variance of point heights (m11) was removed because it has a 1:1 relationship with the standard deviation of point heights (m10). All metrics that were used for data analysis are indicated by an asterisk.

6.2.4. Development of the calibration dataset

Frazer *et al.* (2011a) used the coordinates of plot centers to clip the original LiDAR point cloud into 788 (25×25 m) grid cells. From the plot level point clouds the metrics listed in Table 6.2 were computed using FUSIN/LDV. Hence, for each ground observation a set of 36 LiDAR metrics was available.

[2]http://www.terrasolid.fi/

[3]http://forsys.cfr.washington.edu/fusion/fusionlatest.html. The software was developed at the United States Department of Agriculture (USDA), Forest Service, Pacific Northwest Research Station.

Table 6.2.: List of the 36 LiDAR metrics computed using FUSION/LDV software. The second column (Sel.var. = selected variables) indicates whether the variable has been selected (✳) for the simulation study (see Chapter 7).

Variable	Sel.var.	Description
m6		Minimum point height >2m (LHMIN)
m7		Maximum point height >2m (LHMAX)
m8	✳	Average of point heights >2m (LHMEAN)
m9	✳	Mode of point height >2m (LHMODE)
m10	✳	Standard deviation of point heights >2m (LHSD)
m11		Variance of point heights >2m (LHVAR)
m12	✳	Coefficient of variation of point heights >2m (LHCOV)
m13	✳	Interquartile range of point heights >2m (LHIQR)
m14	✳	Coefficient of skewness (product moment) of point heights >2m (LHSKEW)
m15		Coefficient of kurtosis (product moment) of point heights >2m (LHKURT)
m16	✳	Average absolute deviation of point heights >2m (LHAAD)
m17		First L-moment (mean) of point heights >2m (LHL1)
m18	✳	Second L-moment (variance) of point heights >2m (LHL2)
m19		Third L-moment (skewness) of point heights >2m (LHL3)
m20		Fourth L-moment (kurtosis) of point heights >2m (LHL4)
m21	✳	Second L-moment ratio (coefficient of variation) of point heights >2m (LHLCOV)
m22	✳	Third L-moment ratio (coefficient of skewness) of point heights >2m (LHLSKEW)
m23	✳	Fourth L-moment ratio (coefficient of kurtosis) of point heights >2m (LHLKURT)
m24	✳	1^{st} percentile of point heights >2m (LH01)
m25	✳	5^{th} percentile of point heights >2m (LH05)
m26	✳	10^{th} percentile of point heights >2m (LH10)
m27	✳	20^{th} percentile of point heights >2m (LH20)
m28	✳	25^{th} percentile of point heights >2m (LH25)
m29	✳	30^{th} percentile of point heights >2m (LH30)
m30	✳	40^{th} percentile of point heights >2m (LH40)
m31	✳	50^{th} percentile of point heights >2m (LH50)
m32	✳	60^{th} percentile of point heights >2m (LH60)
m33	✳	70^{th} percentile of point heights >2m (LH70)
m34	✳	75^{th} percentile of point heights >2m (LH75)
m35	✳	80^{th} percentile of point heights >2m (LH80)
m36	✳	90^{th} percentile of point heights >2m (LH90)
m37	✳	95^{th} percentile of point heights >2m (LH95)
m38	✳	99^{th} percentile of point heights >2m (LH99)
m50	✳	Percent canopy density (cover) at 2m (CC2M)
m56	✳	Percent canopy density (cover) at mean canopy height (CCMEAN)
m57	✳	Percent canopy density (cover) at modal canopy height (CCMODE)

6.3. Hedmark

Study area

The second study area is located in south-eastern Norway (see Figure 6.6) and incorporates all of Hedmark County (HED). The size of the area is ≈2.73 million hectares of which ≈53.7% percent is covered by forests. Whereas the northern part of HED consist

of mountainous areas with highest altitudes of 2178 m above sea level, altitudes decrease almost linearly towards the south (Ene *et al.*, 2012). Norway spruce (*Picea abies* [L.] Karst.) and Scots pine (*Pinus sylvestris* L.) represent the dominant trees species in HED.

The information provided in the next two sections is largely based on Ene *et al.* (2012) and Gobakken *et al.* (2012). The dataset HED was provided by the Ecology and Natural Resource Management Department of the Norwegian University of Life Sciences (NMBU; Liviu Ene, Erik Næsset, and Terje Gobakken). For HED no datasets for the field plots and LiDAR data were available. A synthetic population (see Section 6.4 and Ene *et al.* (2012, 2013b) for how the population was simulated) was provided by the NMBU. The field and LiDAR data on which the synthetic populations are based, are, therefore, described only briefly.

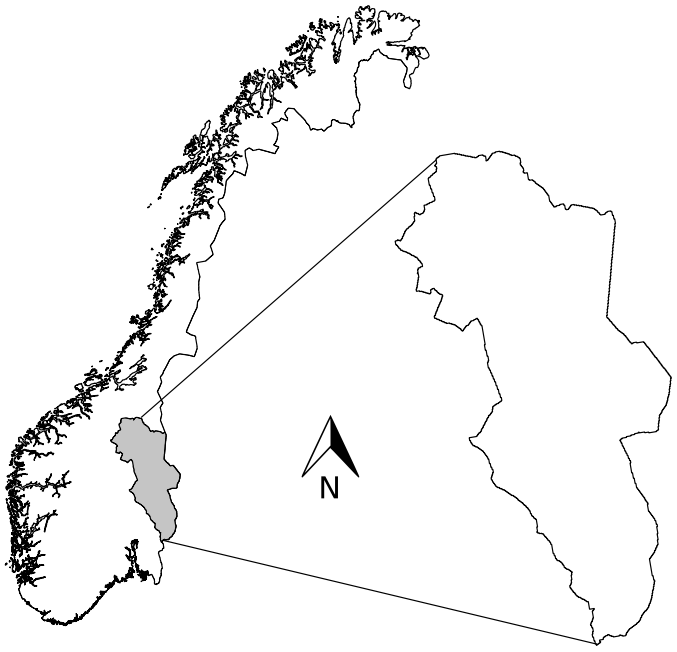

Figure 6.6.: Location of the Hedmark County, Norway. Left: Norway; right: Hedmark County.

6.3.1. Field data

For HED field data were obtained from the permanent Norwegian National Forest Inventory (NFI) grid. For the Norwegian NFI, fixed-area, circular field plots are established on a grid of 3×3 km width. HED contains 2,309 plots that are revisited every five years.

On 1,483 sample plots measurements were taken between 2005 and 2007. Ene *et al.* (2012) used 662 of the available plots. The plot center coordinates were determined using differential post-processing of dual-frequency GPS and Global Navigation Satellite System (GLONASS) measurements (Ene *et al.*, 2012). The accuracy of the center positioning reported was on average .25 m.

On all plots, trees having a DBH $\geq$ 5 cm were callipered. In addition the height of trees was obtained for ten sample trees. These 10 trees were selected proportionally to stem basal area using an adjustable basal area factor (Tomter *et al.*, 2010; Ene *et al.*, 2012).

For each tree of $\geq$ 1.3 m height the AGB was predicted using tree species specific allometric equations (Marklund, 1988) using the height and DBH of each tree as predictor variables. For each plot tree-level biomass predictions were aggregated.

Table 6.3.: List of the 11 LiDAR metrics available for the synthetic HED population.

Variable	Description
D0	Canopy density metric 0[a]
D10	Canopy density metric 10
D20	Canopy density metric 20
D40	Canopy density metric 40
D80	Canopy density metric 80
D90	Canopy density metric 90
h20	20^{th} percentile of height distribution
h30	30^{th} percentile of height distribution
h40	40^{th} percentile of height distribution
h90	90^{th} percentile of height distribution
hmax	Maximum echo height

[a]Proportions of laser echoes above a specified threshold, e.g., one tenth of point heights, to total number of echoes (see Gobakken *et al.* (2012) for details).

6.3.2. LiDAR data

For HED the LiDAR data were obtained during leaf-on conditions between July and September 2006. TerraScan software was used to distinguish between 'ground' and 'non'ground' returns. Only first returns were used by Ene *et al.* (2012). The average point density was approximately 2.8 m^{-2}. The LiDAR metrics available for HED are listed in Table 6.3. More details can be found in Gobakken *et al.* (2012). In addition to

the LiDAR metrics, variables from optical remotely sensed data (Landsat 5 TM[4]) were available. These variables include the at sensor reflectance values for bands one to six.

6.4. Synthetic populations

6.4.1. Rationale

The LiDAR and field data available for HIN were used to simulate a synthetic population. The motivation for creating the population was to obtain a dataset for which the target variable as well as several auxiliary variables were known for each element (i.e., LiDAR pixel). The synthetic population provided a useful database for the simulation study described later in this text (see Chapter 7).

The creation of the synthetic population can be divided into two steps. First, functions that describe the multivariate dependency structure among continuous random variables, called copulas, were fitted to the field/LiDAR data. From the fitted copula model a large number of observations was sampled. In a second step the sampled copula data were used to impute observations of AGB for all LiDAR pixels for which no AGB estimate was available. The methods described below follow closely the approach of Ene *et al.* (2013a).

6.4.2. Copula

A copula is a multivariate distribution function that describes the dependency structure among continuous random variables. For a random vector

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_J),$$

the continuous marginal cumulative distribution functions (c.d.f.s) are given by

$$F_1, F_2, \ldots, F_J.$$

As Sklar (1959) showed, the joint c.d.f. H of $\boldsymbol{X}$ can be written as

[4] http://landsat.gsfc.nasa.gov/?p=3217

$$H(\boldsymbol{x}_k) = C\{F_1(x_{k1}), F_2(x_{k2}), \ldots, F_J(x_{kJ})\},$$

where C is a unique function that captures the multivariate dependencies among the individual probability distributions (Genest & Favre, 2007). The probability functions, F_j, can be estimated independently from the dependency structure represented by the copula. These distribution functions may either be modeled parametrically or non-parametrically (e.g., kernel density estimators, or empirical cumulative distributions). While the dependency structure is modeled by the copula, the rank correlations among the marginals are preserved by their corresponding uniform margins. As rank correlations are used, even complex relationships, e.g., non-linear associations, can be captured by the copula.

From a fitted copula model, multivariate observations can be randomly drawn from the uniform multivariate distribution. What makes copulas particularly useful is that in these samples the dependency structure found in the original data is preserved. In the following the data sampled from the copula will be called *copula data*.

For copula modeling the original variables need to be transformed into uniform margins, and, after a random sample is drawn from the copula, the variables need to be back-transformed to their original scale. Therefore, the distribution function for the marginals must be known. Ene *et al.* (2013a) used empirical cumulative distribution functions (e.c.d.f.). In this study kernel density estimators where employed to determine the distribution of the marginals. The R (R Core Team, 2015) package `ks` (Duong, 2014) was used for both, fitting kernels to the original random variables and to back-transform the observations obtained from the copula.

For bivariate copulas, i.e., two random variables, several families of copula functions are readily available, for high-dimensional data, however, the choice is limited (Brechmann & Schepsmeier, 2013). Recently Aas *et al.* (2009) proposed to decompose high-dimensional datasets into hierarchical pair-copula constructs. These constructs exhibit tree-like structures, where a first variable is defined as a root node of a tree and the pair-wise dependencies between the selected variable and all other variables are modeled using bivariate copulas. Subsequently, a second variable is selected and placed in a second root node. Pair-wise dependencies are modeled for the remaining variables conditional on the variable placed in the first root. Graphical representations of these tree structures are known as vines (Bedford & Cooke, 2002). Figure 6.7 provides an

example of a vine structure. Details on the construction of vines for copula modeling are provided in Brechmann & Schepsmeier (2013).

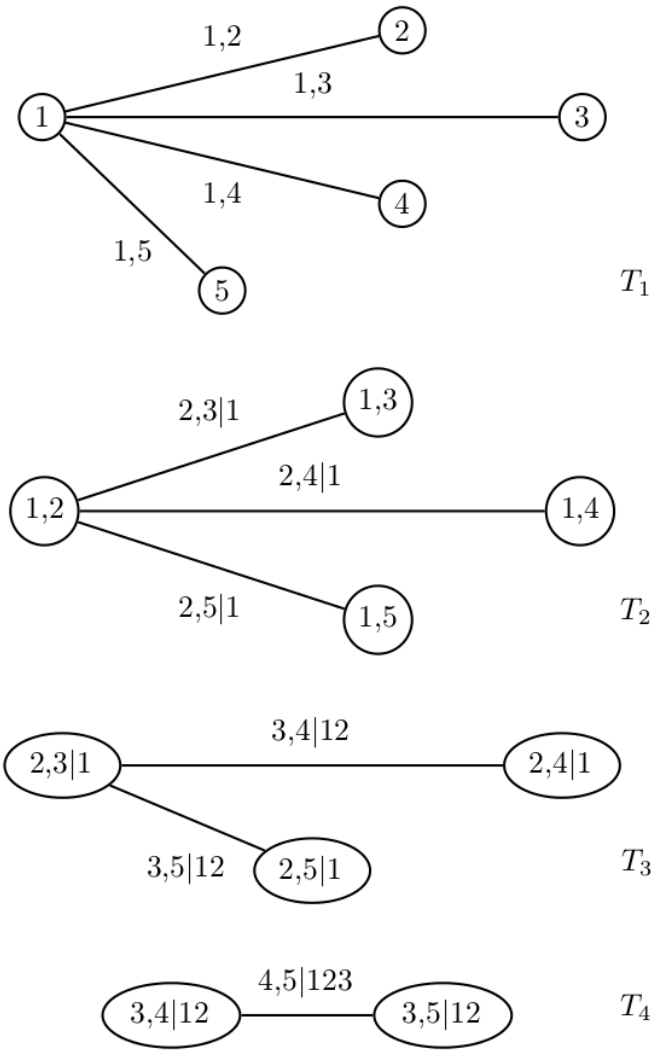

Figure 6.7.: Example of a five dimensional Canonical vine (C-vine) tree (taken from Brechmann & Schepsmeier (2013)).

To construct the tree-like structure, root nodes have to be selected from the available variables. In this study the procedures proposed by Czado *et al.* (2012) were used. In the approach described in Czado *et al.* (2012) all pair-wise Kendall's τ values between

all J variables (and observations) are estimated. The variable placed in the first root node maximizes

$$\hat{P}_j := \sum_{j=1}^{J} |\hat{\tau}_{jp}|,$$

where $j = 1, 2, \ldots, J$, and $p = 1, 2, \ldots, J-1$ (see Czado *et al.* (2012, equation 4.1)). After the first root node has been identified, the second root node is selected among the remaining $J-1$ variables. The procedure continuous until all $J-1$ root nodes have been identified.

6.4.3. Computation

To obtain copula models for HIN the R packages `VineCopula` and `CDVine` (Brechmann & Schepsmeier, 2013) were used. First, the original variables (29 LiDAR metrics and the target variable AGB) were transformed into uniform margins using the function `pkde()` from the `ks` package. The approach proposed by Czado *et al.* (2012) was then used to define the tree structure, i.e., identify the root nodes. Next, pair-copula families were selected and a C-vine copula model was fitted to the data using the function `RVineStructureSelect()`.

From the fitted copula model a large sample of 1,000,000 observations was selected. Subsequently, the uniform margins sampled from the copula model were back-transformed to their original scale. The final copula data consisted of 998,635 observations. For each observation the AGB as well as the 29 LiDAR metrics were available. In Figure 6.8 original observations of AGB (obtained from the field data) are plotted against copula data for a subset of four variables. Table 6.4 shows the Pearson product-moment correlation coefficients between the original field data and the copula data. The correlation coefficients between AGB and the LiDAR metrics are slightly larger for the original field data compared to the coefficients for the copula data.

6.4.4. Imputation

The copula data were used to impute AGB values for each of the 157,053 LiDAR cells that were available for HIN. For imputation kNN was used with $k = 1$. Most similar neighbour (MSN) (Moeur & Stage, 1995) was used as a measure of nearness (R package `yaImpute`

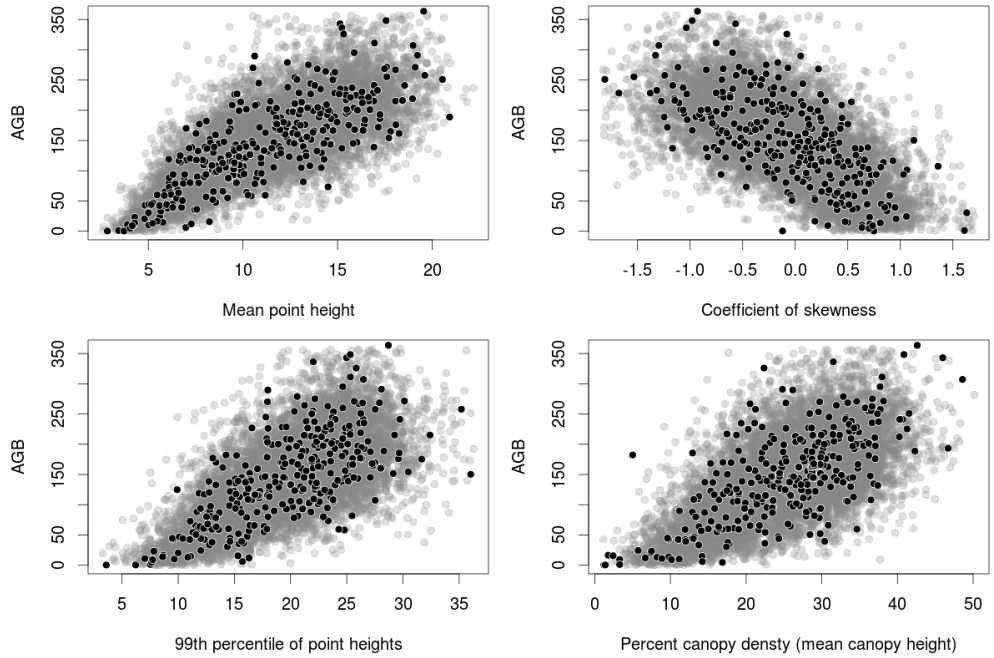

Figure 6.8.: Observations of AGB and LiDAR metrics from the original dataset (black dots), plotted against values obtained from the copula (gray dots); HIN data.

(Crookston & Finley, 2008) with method `msn`). After imputation the dataset was merged with the original dataset ($n = 788$), resulting in a dataset of in total 157,839 complete observations, i.e., AGB and LiDAR. The mean AGB for the synthetic population was 82.34 Mg ha^{-1} (field data 84.61 Mg ha^{-1}), and the the standard deviation 81.98 Mg ha^{-1} (77.95 Mg ha^{-1}).

A random sample of 50,000 observations was drawn from the synthetic HIN population. The size of the population was reduced in order to reduce computational costs. All field observations where retained. A comparison of descriptive statistics for AGB and the LiDAR metrics revealed no significant differences (using non-parametric test, i.e., Wilcoxon signed-rank test) between the reduced and original synthetic population.

The size of the synthetic HED population was reduced in a similar manner. The original population (provided by NMBU) consisted of 200,000 observations; 50,000 observations were randomly sampled. No significant differences between the reduced and complete synthetic populations were found.

Table 6.4.: Pearson product-moment correlation coefficients between AGB and LiDAR metrics for the original field observations and the copula data (HIN).

LiDAR metric	Field data	Copula data
m8	0.78	0.76
m9	0.67	0.64
m10	0.60	0.63
m12	−0.22	−0.20
m13	0.46	0.52
m14	−0.65	−0.65
m16	0.58	0.60
m21	−0.23	−0.22
m22	−0.65	−0.64
m23	0.04	−0.01
m24	0.42	0.43
m25	0.59	0.60
m26	0.65	0.65
m27	0.71	0.71
m28	0.74	0.72
m29	0.75	0.73
m30	0.76	0.75
m31	0.77	0.76
m32	0.77	0.76
m33	0.77	0.76
m34	0.76	0.75
m35	0.75	0.74
m36	0.73	0.72
m37	0.71	0.70
m38	0.67	0.67
m50	0.54	0.52
m56	0.67	0.64
m57	0.04	−0.04

For HED variables were individually transformed to linearize the relationship between the target variable AGB and the covariates. For D0, D10, D20, D40, and D90, powers of 2, 2, 1.75, 1.5, and 1.25 were taken, and for h20, h30, h40, h90, and hmax powers of 1.25, 1.25, 1.5, 2.5, and 2.25, respectively. To all Landsat bands, b1 – b6, the log-transform was applied. D80 and HOH were not transformed. For HIN only the LiDAR metrics m25, m26, m27, m28 and m29 were transformed (see Table 6.2) using the log-transform. Figures 6.9 and 6.10 show the correlation structure in the two datasets.

It is important to note that transformations alter the structure in the datasets. Transformation were applied to facilitate a fair comparison between the different model techniques, especially between the parametric approaches and random forests. The main purpose of this study was not to estimate the "true" AGB for HED and HIN, but to assess the impact of model and variable selection on estimates of precision. Estimates

of AGB means or totals of this study may not be directly compared to other findings for the same study areas, e.g., Ene *et al.* (2012), Frazer *et al.* (2011a), or White *et al.* (2013).

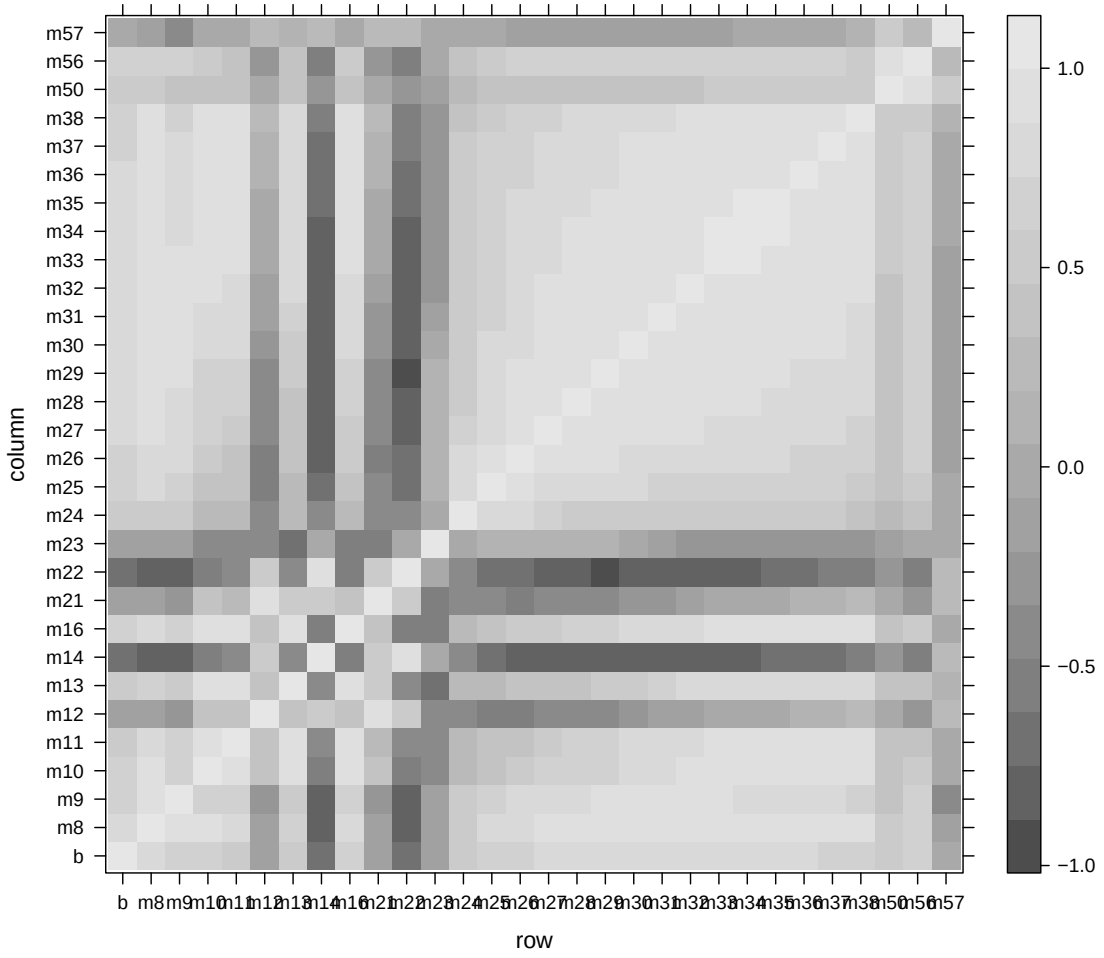

Figure 6.9.: Correlation structure in the artificial datasets HIN (the scale bar refers to the Pearson correlation coefficient).

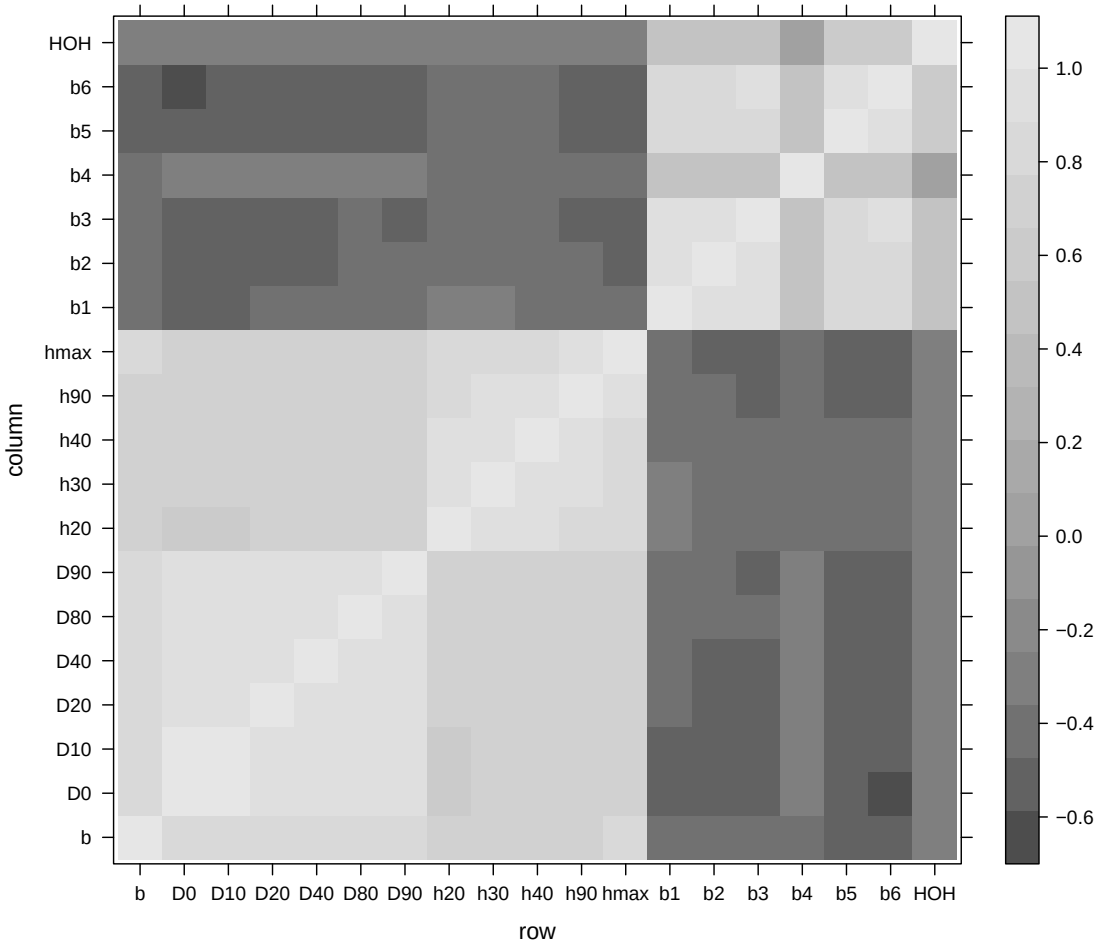

Figure 6.10.: Correlation structure in the artificial datasets HED (the scale bar refers to the Pearson correlation coefficient).

7. Simulation study

7.1. Outline of the simulation studies

The artificial and synthetic populations described in the preceding Chapter 6 served as the data basis for simulation studies as follows:

1. A simple random sample without replacement (SRSwoR) of size n was taken from a population U. For all elements in the sample the values of the target variable y as well as the values of J auxiliary variables were observed. The values of y remained unknown for all elements of the population that were not in the sample, i.e., for the set $U - S$. The values of the J auxiliary variables were known for all elements in U.

   For each sample,

   a) the mean of the target variable as well as the standard error of the mean was estimated (no use of the auxiliary information).

   b) a (model-assisted) regression estimator was used to estimate the population mean and the standard error of the mean. The working model for the regression estimator was obtained by using one of the model and variable selection procedures described below (Section 4).

   c) a model-based estimator was used to estimate the population mean and the standard error of the mean. To obtain the working model the model and variable selection procedures described below were used (Section 4).

2. The procedures described in step 1 were repeated P times. Each sample is identified by an index $p = 1, 2, \ldots, P$.

After a working model was formulated and fitted to each sample, the following information was extracted:

1. Estimated model coefficients
2. Model residuals
3. Sum of the predicted values of the target variable for all elements in U obtained from the working model.
4. Sum of the predicted values of the target variable for all elements $U - S$ obtained from the working model.

The populations used for the simulation studies were the three artificial populations (i.e., NOISE, COR, and DCRO described in Section 6.1 in Chapter 6), as well as the two synthetic populations HIN and HED (see Section 6.2 and 6.3 in Chapter 6).

From the three artificial populations $P = 2000$ SRSwoR samples of size $n = 40$ were drawn. For HIN and HED SRSwoR samples of size $n_1 = 50$, $n_2 = 100$, $n_3 = 200$, and $n_4 = 400$ were selected. From each of the synthetic populations, $P = 50000$ samples were drawn for each of the four sample sizes. All samples were saved in a `list` object in R and all subsequent computations were done using the same software package.

7.2. Computation — implementation in R

Saturated model: The full regression model including all covariates was fitted to each sample using the R function `lm()`. The function `predict()` and `residuals()` where used to predict values for the whole population and to extract the model residuals, respectively. The same functions were used for all variable selection techniques for which the `lm()` function was applied.

AIC: The R function `stepAIC()` from package `MASS` (Venables & Ripley, 2002) was used for stepwise variable selection using the AIC as a criterion for model performance. The function `lm()` was used for model fitting.

AICc: For computation the function `stepAICc()` was used for variable selection. The R script for the function can be downloaded from the following website (last accessed April 19, 2014): http://wwwuser.gwdg.de/ cscherb1/stepAICc.txt. Function `lm()` was applied to fit the model.

BIC: The function `stepAIC()` with argument `k = log(n)` was used for selecting variables and the function `lm()` was used for model fitting. Function `lm()` was applied to fit the model.

VIF: The function `vif()` from package `faraway` (Faraway, 2011) was used to compute the VIF for each `lm()` object. The variable with the highest VIF was removed from the dataset. The model was refitted using all remaining variables. This procedure was repeated until all VIFs were below 10. The `stepAIC()` function (on `lm()`) was applied afterwards.

VIFB: The same procedure as for **VIF** was applied, except that instead of the AIC, an exhaustive search for the best model, using C_p as a criterion, was used. The model with the lowest C_p was selected. The function `regsubsets()` from package `leaps` (Lumley, 2009) was used. Note that **VIFB** was only computed for the three artificial datasets (NOISE, COR, and DCOR).

Condition number: An R script was written that implements the procedure proposed by Silva & Skinner (1997). The function `lm()` was applied to fit the model.

Ridge regression: Function `cv.glmnet()` from package `glmnet` (Friedman *et al.*, 2010) was used for ridge regression. By default the function scales the input variables and performs 10-fold cross-validation to obtain the value for λ.

Lasso: The Lasso was fitted using the same functions as for **Ridge**, except that the value for `alpha` was changed from 0 (ridge penalty) to 1 (lasso penalty).

Partial least squares regression: For the simulation studies the function `plsr()` from package `pls` (Mevik *et al.*, 2013) was used. A single (the first) component was extracted. Variables have been scaled using `scale = TRUE`.

Random forests: For the simulation studies random forests where fitted to each of the p samples. The function `randomForest()` from the R package `randomForest` (Liaw & Wiener, 2002) was used with default settings, i.e., 500 trees. For each of the p samples the importance of predictor variables was extracted from the `randomForest` object. The importance of a variable is assessed by computing the mean decrease of the mean squared error (MSE) of the variable over all splits where the variable was used. The MSE is defined as

$$\text{MSE} = n^{-1} \sum_{k \in S} (\hat{y}_k - y_k)^2.$$

The abbreviations that will be used for the different model and variable selection procedures in the remainder of this text are given in Table 7.1.

Table 7.1.: List of modeling techniques used in the simulation studies.

FULL	Saturated model (no variable selection)
AIC	Variable selection based on Akaike's Information Ciriterion
AICc	Corrected version of the AIC
BIC	Bayesian Information Criterion (or Schwarz Criterion)
VIF	Selection based on the variance inflation factor combined with the AIC
VIFB	Selection based on the variance inflation factor combined with best-subset selection
CON	Variable selection based on the condition number (Silva & Skinner, 1997)
Ridge	Ridge regression
Lasso	Least absolute shrinkage and selection operator (lasso)
PLSR	Partial least squares regression
RF	Random forest algorithm

7.3. Analysis

7.3.1. Estimators

Once all modeling techniques had been applied to all P samples, several estimators were used to estimate the population mean and its precision. The following estimators were considered:

1. The SRSwoR (SI) estimator for population mean (ignoring the auxiliary information; this estimator will be abbreviated SI hereafter)

$$\bar{y}_{\mathrm{SI}} = n^{-1} \sum_{k \in S} y_k, \tag{7.1}$$

2. the model-assisted (MA) mean estimator

$$\bar{y}_{\mathrm{MA}} = N^{-1} \sum_{k \in U} \hat{y}_k + n^{-1} \sum_{k \in S} e_k, \tag{7.2}$$

3. and the model-based (or model-dependent; MD) mean estimator

$$\bar{y}_{\mathrm{MD}} = \left(\sum_{k \in S} y_k + \sum_{k \in U-S} \boldsymbol{x}' \hat{\boldsymbol{\beta}} \right) / N. \tag{7.3}$$

The standard error for the estimator (7.1) was estimated by

1.
$$\mathrm{SE}_{\mathrm{SI}} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}} \quad \text{with} \quad s^2 = (n-1)^{-1}\sum_{k\in S}(y - \bar{y})^2. \tag{7.4}$$

2. To estimate the precision for (7.2) three different estimators where used; namely,

$$\mathrm{SE}_{\mathrm{Simple}} \doteq \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)s^2_{e_k}} \quad \text{with} \quad s^2_{e_k} = (n-1)^{-1}\sum_{k\in S}(e_k - \bar{e})^2, \tag{7.5}$$

$$\mathrm{SE}_{\mathrm{Fuller}} = \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)s^2_{e_k}} \quad \text{with} \quad s^2_{e_k} = (n-J-1)^{-1}\sum_{k\in S} e_k^2 \tag{7.6}$$

and

$$\mathrm{SE}_{\mathrm{Sarndal}} = \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{n-1}{n-J}\right)s^2_{e^*_k}}, \tag{7.7}$$

with

$$s^2_{e^*_k} = (n-1)^{-1}\sum_{k\in S}(e^*_k - \bar{e}^*)^2 \quad \text{where} \quad e^*_k = e_k g_k, \tag{7.8}$$

where g_k is given in (3.35).

3. For the model-based estimator (7.3) the standard error was estimated by

$$\mathrm{SE}_{\mathrm{MD}} = \sqrt{\left[\hat{\sigma}^2\left((N-n) + \boldsymbol{\tau}_x'(\boldsymbol{X}'_{k\in S}\boldsymbol{X}_{k\in S})^{-1}\boldsymbol{\tau}_x\right)\right]/N^2}. \tag{7.9}$$

All estimators are explained in more detail in Chapter 3.

Not all estimators can readily be used for all modeling techniques. Table 7.2 provides an overview of which estimators where used for which modeling technique. For ridge regression, the Lasso, PLSR, and RF, only the simple variance estimator (7.5) was employed. For the first three, the R^2 can be computed in a similar way as for single or multiple regression models. For RF, a "pseudo-R^2" was computed by substituting the

Table 7.2.: List of estimators used for the different modeling techniques described in Section 4. A star (*) indicates that the estimator was used.

Modeling technique	SE_{SI}	SE_{Simple}	SE_{Fuller}	$SE_{Sarndal}$	SE_{MD}
FULL	*	*	*	*	*
AIC	*	*	*	*	*
AICc	*	*	*	*	*
BIC	*	*	*	*	*
VIFAIC	*	*	*	*	*
CON	*	*	*	*	*
RR	*	*	-	-	-
LASSO	*	*	-	-	-
PLS	*	*	-	-	-
RF	*	*	-	-	-

sum of squared residuals by the sum of squared differences between the observations and out-of-bag predictions provided by RF.

As mentioned in in Section 3.2 in Chapter 3, for model-based inference, model residuals are assumed to independent and identically distributed with zero mean and constant variance. For the simulation studies these assumptions were not tested at each iteration. As SRSwoR was applied and variables have been transformed beforehand, it was expected that model assumptions are met.

7.3.2. Evaluating the performance of estimators

The following summary statistics were estimated for each model and variable selection procedure:

1. BIAS

$$\text{BIAS} = \sum_{S_p \in \mathscr{S}_P} [\bar{y}(S_p) - \mu_y]/P$$

where $\mathscr{S}_P$ is the set of the P samples that were drawn from the populations in the simulation studies (i.e., $|\mathscr{S}_P| = 2000$ for the artificial datasets, and $|\mathscr{S}_P| = 50000$ for each sample size for HIN and HED), S_p represents one sample and $\bar{y}(S_p)$ is the estimated mean for that sample; μ_y is the true population mean. The relative bias (rBIAS) is the bias divided by μ_y.

2. Empirical standard error (ESE)

$$\text{ESE} = \sqrt{\frac{\sum_{S_p \in \mathscr{S}_P} [\bar{y}(S_p) - \overline{\bar{y}(S_p)}]^2}{(P-1)}},$$

where $\overline{\bar{y}(S_p)}$ is the mean of estimated means. The ESE gives the standard deviation of estimated means. The relative empirical standard error in percent is given by dividing the ESE by $\overline{\bar{y}(S_p)}$ times 100.

3. Average estimated standard error (AVSE)

$$\text{AVSE} = \frac{\sum_{S_p \in \mathscr{S}_P} \text{SE}_{\bar{y}}(S_p)}{P}$$

The relative AVSE in percent is given by dividing the AVSE by $\overline{\bar{y}(S_p)}$ times 100.

4. Coverage (COV%)

$$\text{COV\%} = \frac{\sum_{S_p \in \mathscr{S}_P} T_p}{P}$$

where

$$T_p = \begin{cases} 1 & \text{if the estimated mean } \bar{y}(S_p) \text{ lies within the 95\% confidence interval} \\ 0 & \text{otherwise.} \end{cases}$$

The construction of confidence intervals is covered in Section 3.16.

5. Efficiency (EFF)

$$\text{EFF} = \frac{\text{AVSE}}{\overline{\text{SE}_{\text{SI}}}} \times 100$$

where $\overline{\text{SE}_{\text{SI}}}$ is the mean standard error of the SI sample mean without using auxiliary data. For SI the efficiency is 100. Note that the smaller the value of EFF, the more efficient is the alternative estimator, i.e., efficiencies increase when EFF decrease.

A good model and variable selection procedure was considered as one that

1. that results in negligible bias (BIAS), e.g., rBIAS $< 1\%$,
2. that has a small empirical standard error (ESE),
3. that has a small mean estimated standard error (AVSE),
4. that has a coverage rate (COV%) of approximately 0.95,

5. that is efficient (i.e., EFF < 100),
6. and where the AVSE and ESE roughly match. An undesirable situation would be one in which the AVSE is significantly smaller than the ESE. This would mean that precision is overestimated.

Part III.

Results

8. Model-assisted inference

8.1. Artificial datasets

8.1.1. Dataset NOISE

For the three artificial datasets, results for the three variance estimators (considered in the model-assisted approach) are presented for each dataset separately. The first dataset is NOISE, followed by COR and DCOR.

Summary statistics for the NOISE simulation study ($n = 40$; 2,000 iterations) are given in Table 8.1. In the NOISE dataset the covariates were not related to the target variable (nor with each other), and, therefore, efficiencies should be approximately 100 for all procedures. The parametric standard error (see equation 3.12 in Chapter 3) for a sample size of $n = 40$ was 1.55.

For the simple variance estimator, $\hat{V}_{\text{Simple}}$, precision was overestimated for all modeling techniques, except for RF. The relative ESE was larger than the parametric standard error for all procedures, although only slightly for Ridge and Lasso. The difference between the relative ESE and AVSE was large, except for Lasso, Ridge and RF. Therefore, coverage rates were below 0.95 for all other modeling techniques. All stepwise variable selection procedures, i.e., AIC, AICc, and BIC, grossly overestimated precision when the simple variance estimator was used for the NOISE dataset.

For the variance estimator after Fuller, $\hat{V}_{\text{Fuller}}$, the picture slightly changed. The differences between the relative ESE and AVSE decreased, however, precision was still overestimated by all modeling techniques. For FULL efficiency was above 100. Coverage rates are all below 0.90.

When the variance estimator after Särndal, $\hat{V}_{\text{Sarndal}}$, was applied to the NOISE dataset, the difference between the relative ESE and AVSE further decreased and were almost

Table 8.1.: Results for the three variance estimators: $\hat{V}_{\text{Simple}}$, $\hat{V}_{\text{Fuller}}$, and $\hat{V}_{\text{Sarndal}}$. Relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %; DIFF = (ESE - AVSE) / ESE×100), efficiency (EFF), and coverage rates (COV) for dataset NOISE.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF	EFF	COV
SI	0.10	1.434	1.531	−6.75	100.0	0.95
$\hat{V}_{\text{Simple}}$						
AIC	0.14	1.959	1.205	38.48	78.7	0.79
AICc	0.14	1.874	1.279	31.77	83.5	0.83
BIC	0.13	1.876	1.315	29.90	85.9	0.84
CON	0.16	1.896	1.102	41.87	72.0	0.77
FULL	0.17	2.081	1.101	47.09	71.9	0.72
Lasso	0.10	1.437	1.495	−4.02	97.6	0.95
PLSR	0.14	1.614	1.198	25.78	78.2	0.87
RF	0.11	1.489	1.587	−6.60	103.6	0.96
Ridge	0.10	1.437	1.496	−4.09	97.7	0.95
VIF	0.14	1.959	1.205	38.48	78.7	0.79
VIFB	0.12	1.850	1.290	30.25	84.3	0.84
$\hat{V}_{\text{Fuller}}$						
AIC	0.14	1.959	1.313	32.96	85.8	0.83
AICc	0.14	1.875	1.348	28.08	88.0	0.85
BIC	0.13	1.876	1.372	26.90	89.6	0.86
CON	0.16	1.896	1.452	23.45	94.8	0.86
FULL	0.17	2.081	1.538	26.12	100.4	0.87
VIF	0.14	1.959	1.313	32.96	85.8	0.83
VIFB	0.12	1.850	1.352	26.93	88.3	0.85
$\hat{V}_{\text{Sarndal}}$						
AIC	0.14	1.959	1.419	27.59	92.6	0.86
AICc	0.14	1.875	1.409	24.84	92.0	0.87
BIC	0.13	1.876	1.418	24.42	92.6	0.88
CON	0.16	1.896	1.846	2.68	120.5	0.94
FULL	0.17	2.081	2.083	−0.09	136.0	0.94
VIF	0.14	1.959	1.419	27.59	92.6	0.86
VIFB	0.12	1.850	1.406	24.01	91.8	0.87

[a]Method = model or variable selection procedure

zero for FULL and CON. However, for these two modeling techniques efficiency was well above 100.

Graphical comparisons for the estimators are provided in Annex B on page 126.

8.1.2. Dataset COR

For the dataset COR, the parametric standard error for a sample size of $n = 40$ was 1.57. Summary statistics for COR for all three variance estimators are given in Table 8.2 on page 79. In the COR dataset all variables were strongly correlated with the target variable, as well as with each other.

Table 8.2.: Results for the three variance estimators: $\hat{V}_{\text{Simple}}$, $\hat{V}_{\text{Fuller}}$, and $\hat{V}_{\text{Sarndal}}$. Relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %; DIFF = (ESE - AVSE) / ESE×100), efficiency (EFF), and coverage rates (COV) for dataset COR.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
SI	0.06	1.582	1.571	0.71	100.0	0.95
$\hat{V}_{\text{Simple}}$						
AIC	0.00	0.171	0.086	49.59	5.5	0.68
AICc	−0.00	0.180	0.093	48.16	5.9	0.69
BIC	−0.00	0.179	0.092	48.45	5.9	0.68
CON	−0.02	0.369	0.356	3.54	22.7	0.94
FULL	0.00	0.163	0.081	50.18	5.2	0.69
Lasso	0.00	0.163	0.101	37.84	6.4	0.79
PLSR	0.00	0.120	0.112	6.40	7.1	0.93
RF	0.01	0.346	0.318	8.08	20.2	0.93
Ridge	0.00	0.133	0.121	8.82	7.7	0.93
VIF	0.02	0.348	0.326	6.46	20.7	0.94
VIFB	0.02	0.348	0.326	6.47	20.8	0.94
$\hat{V}_{\text{Fuller}}$						
AIC	0.00	0.171	0.101	40.75	6.4	0.76
AICc	−0.00	0.180	0.105	41.55	6.7	0.75
BIC	−0.00	0.179	0.105	41.61	6.7	0.75
CON	−0.02	0.369	0.361	2.27	23.0	0.95
FULL	0.00	0.163	0.113	30.43	7.2	0.82
VIF	0.02	0.348	0.332	4.74	21.1	0.95
VIFB	0.02	0.348	0.332	4.75	21.1	0.95
$\hat{V}_{\text{Sarndal}}$						
AIC	0.00	0.171	0.117	31.68	7.4	0.82
AICc	−0.00	0.180	0.117	34.84	7.5	0.79
BIC	−0.00	0.179	0.117	34.66	7.5	0.80
CON	−0.02	0.369	0.364	1.28	23.2	0.95
FULL	0.00	0.163	0.154	5.21	9.8	0.92
VIF	0.02	0.348	0.336	3.57	21.4	0.95
VIFB	0.02	0.348	0.336	3.58	21.4	0.95

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

High efficiencies were found for all modeling techniques and variance estimators. For the procedures that removed highly correlated variables, i.e., CON, VIF, and VIFB, the EFF was above 20 and much larger than for the stepwise variable selection procedures. CON, VIF, and VIFB, however, showed smaller differences between the relative ESE and AVSE and, therefore, the bias in variance estimation was small compared to the stepwise selection procedures. RF showed results similar to the procedures VIF and VIFB.

For PLSR and Ridge efficiencies were high (i.e., low EFF) and coverage rates were close to 0.95, revealing good performance on the dataset COR for these two modeling strategies.

For the procedures that selected variables, the average number of covariates included in the models (over 2,000 iterations) were: AIC (11.87), AICc (9.41), BIC (9.72), CON (2.00), VIF (2.56), VIFB (2.56), and Lasso (16.54).

For the full model (FULL) large differences were found among the three different variance estimators. While for the simple variance estimator, the difference between the relative ESE and AVSE was 50%, the differences was much smaller when the variance estimator after Särndal was used. For $\hat{V}_{\text{Sarndal}}$, DIFF% was roughly 5% for FULL, and revealed good performance when compared to all other procedures. For CON the difference was small for all variance estimators, but also much larger than for FULL. Table 8.3 provides a comparison between the three different variance estimators for dataset COR.

Although FULL yielded reasonable results for the variance estimator after Särndal, PLSR and Ridge still outperform FULL. The minimum relative AVSEs for these techniques were lower (using the simple variance estimator) than for FULL (using the variance estimator after Särndal).

8.1.3. Dataset DCOR

For DCOR, the parametric standard error for a sample size of $n = 40$ was 1.59. In DCOR the correlation between the target variable and the 20 auxiliary variables constantly decreased. The first variable was most strongly correlated with the target variable, i.e., $\rho_{Y|X_1} = 0.93$.

Using the simple variance estimator, the precision was overestimated by the stepwise variable selection procedures, i.e., AIC, AICc, and BIC. Larger differences between the relative ESE and AVSE were found for FULL VIF, and CON. Here, precision was over-

Table 8.3.: Comparison of the relative AVSE $\hat{V}_{\text{Sarndal}}$, $\hat{V}_{\text{Fuller}}$, $\hat{V}_{\text{Simple}}$ (dataset COR).

Method[a]	ESE (%)	$\hat{V}_{\text{Sarndal}}$ AVSE (%)	$\hat{V}_{\text{Fuller}}$ AVSE (%)	$\hat{V}_{\text{Simple}}$ AVSE (%)
SI	1.58	1.57	1.57	1.57
AIC	0.17	0.12	0.10	0.09
AICc	0.18	0.12	0.11	0.09
BIC	0.18	0.12	0.10	0.09
CON	0.37	0.36	0.36	0.36
FULL	0.16	0.15	0.11	0.08
Lasso	0.16	–	–	0.10
PLSR	0.12	–	–	0.11
RF	0.34	–	–	0.31
Ridge	0.13	–	–	0.12
VIF	0.35	0.34	0.33	0.33
VIFB	0.35	0.34	0.33	0.33

[a]Method = model or variable selection procedure

estimated by over 50%, over 48%, and over 43%, respectively. Small differences and good coverage rates were obtained for Lasso and RF.

For the variance estimator after Fuller, the picture slightly changes. The largest differences between the relative ESE and AVSE were found for VIF. For the variance estimator after Särndal, the difference between the relative ESE and AVSE dropped to 7.29% for FULL.

For CON the EFF was larger than 100 when the variance estimator after Särndal was employed. Table 8.5 provides an overview of how often the five most powerful predictor variables were selected by the different variable selection procedures.

For CON the variables X_{17}, X_{16}, X_{15}, X_{13}, X_{12}, and X_{11} were selected in 99% of the 2,000 iterations, although the correlation coefficient with the target variable, $\rho_{Y|X_j}$, is below 0.1. The most powerful predictor X_1 ($\rho_{Y|X_1} = 0.93$) was selected only two times.

In summary, stepwise variable selection procedures, i.e., AIC, AICc, and BIC, overestimated precision in all artificial datasets, regardless of which variance estimator was used. For CON and FULL, as well as for VIF and VIFB, the structure of the dataset (either noisy, strongly correlated covariates, or both) largely determined how the different variance estimators performed. The Lasso and RF performed well on the dataset DCOR. Ridge and PLSR, in contrast, performed well on the COR dataset where covariates are strongly correlated. VIF and VIFB overestimated precision grossly on the dataset DCOR. Although they showed good coverage rates on the dataset COR, relative ESE and AVSE were much larger than for most other variable selection procedures.

Table 8.4.: Results for the three variance estimators: $\hat{V}_{\text{Simple}}$, $\hat{V}_{\text{Fuller}}$, and $\hat{V}_{\text{Sarndal}}$. Relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %; DIFF = (ESE - AVSE) / ESE×100), efficiency (EFF), and coverage rates (COV) for dataset DCOR.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
SI	−0.03	1.571	1.568	0.15	100.0	0.96
$\hat{V}_{\text{Simple}}$						
AIC	0.06	0.743	0.423	43.07	26.9	0.76
AICc	0.05	0.699	0.455	34.87	29.0	0.80
BIC	0.03	0.684	0.470	31.21	30.0	0.82
CON	−0.06	1.803	1.023	43.30	65.2	0.75
FULL	0.07	0.804	0.388	51.75	24.7	0.68
Lasso	−0.01	0.628	0.567	9.78	36.1	0.93
PLSR	−0.02	0.830	0.629	24.20	40.1	0.87
RF	−0.01	0.927	0.915	1.28	58.3	0.95
Ridge	−0.01	0.808	0.541	33.05	34.5	0.83
VIF	0.12	1.078	0.556	48.39	35.4	0.75
VIFB	0.09	1.031	0.596	42.22	38.0	0.79
$\hat{V}_{\text{Fuller}}$						
AIC	0.06	0.743	0.468	37.04	29.8	0.80
AICc	0.05	0.699	0.484	30.82	30.8	0.82
BIC	0.03	0.684	0.494	27.84	31.4	0.84
CON	−0.06	1.804	1.304	27.69	83.1	0.84
FULL	0.07	0.805	0.542	32.62	34.5	0.82
VIF	0.12	1.078	0.617	42.73	39.3	0.78
VIFB	0.09	1.031	0.636	38.32	40.5	0.82
$\hat{V}_{\text{Sarndal}}$						
AIC	0.06	0.743	0.513	30.93	32.7	0.83
AICc	0.05	0.699	0.511	26.90	32.6	0.85
BIC	0.03	0.684	0.515	24.65	32.8	0.86
CON	−0.06	1.804	1.596	11.49	101.7	0.91
FULL	0.07	0.805	0.746	7.29	47.5	0.92
VIF	0.12	1.078	0.679	37.02	43.3	0.82
VIFB	0.09	1.031	0.675	34.53	43.0	0.84

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

8.2. Hedmark

8.2.1. Simple variance estimator

Results of the simulation studies for the two synthetic populations HED and HIN are presented separately for each estimator.

Table 8.5.: Percentage of how often a variable was selected by the different variable selection procedures after 2,000 iterations (dataset DCOR).

Method[a]	X_1	X_1	X_1	X_1	X_5
AIC	100.00	83.70	50.60	37.80	35.90
AICc	100.00	77.80	36.90	23.80	20.70
BIC	100.00	73.70	31.10	19.20	17.00
CON	0.00	0.50	13.20	45.60	85.30
VIF	66.90	82.40	65.50	48.50	39.90
VIFB	66.90	80.20	55.70	36.20	26.00
Lasso	100.00	85.70	32.00	14.50	9.40

[a]Method = model or variable selection procedure

Summary statistics for the simple variance estimator for HED are presented in Table 8.6 on page 87. Using regression estimation increased precision of estimates for all modeling techniques and sample sizes as compared to SI.

For a sample size of $n = 50$, precision was overestimated for all modeling techniques. For RF, PLSR, and Ridge the difference between the relative ESE and AVSE was small, however. These three techniques revealed coverage rates close to 0.95. Coverage rates for the stepwise variable selection procedures AIC, AICc, and BIC, were around 0.80, and precision was overestimated by about one third.

Coverage rates for a sample size of $n = 100$ were close to 0.95. The relative performance between $n = 50$ and $n = 100$ was almost equal for the different techniques. When the sample size further increased to $n = 200$ and $n = 400$, precision was not overestimated. Coverage rates for all model and variable selection procedures were close to or equal to 0.95. Hence, the difference between the relative ESE and AVSE was below 5% for all modeling techniques[1]. Efficiencies among the different sample sizes changed only little. EFF was below 50 for all procedures and sample sizes.

Table 8.7 on page 88 lists the average number of variables that were included in the working models by the different variable selection procedures.

[1]Note, because of an error in the R code that was written for this study (i.e., predicted values based on all trees instead of out-of-bag predictions were extracted in the code), results for RF are only presented for sample sizes of $n = 50$ and $n = 100$. The error was detected close to the end of this study and due to the time needed for the simulations, it was not possible to rerun simulations.

8.2.2. Variance estimator after Fuller

As expected, for the variance estimator $\hat{V}_{\text{Fuller}}$ estimated precision was lower than for $\hat{V}_{\text{Simple}}$ for all modeling techniques considered. Results for $\hat{V}_{\text{Fuller}}$ can be found in Table 8.8 on page 88.

For a sample size of $n = 50$, precision was again overestimated for all model and variable selection procedures. The difference between the relative ESE and AVSE was largest for AIC, followed by AICc, and BIC. For these procedures coverage rates were all below 0.85. Coverage rates close to 0.95 were found for CON. However, for CON the relative ESE, as well as the relative AVSE was larger.

For a sample size of $n = 100$ results only change in absolute terms, i.e., the relative ESE and AVSE decreased, the differences between the different modeling techniques were almost equal. However, when sample sizes increased further to $n = 200$ and $n = 400$, the differences between CON and all other procedures changed. The relative ESE was much larger for CON than for AIC, AICc, BIC, FULL, and VIF. For a sample size of $n = 400$ variances were estimated unbiasedly by CON. The coverage rates for the other techniques, however, were almost equal to 0.95, indicating no overestimation of precision.

8.2.3. Variance estimatior after Särndal

When the variance estimator $\hat{V}_{\text{Sarndal}}$ was used for the HED dataset, variances further increased, at least for small sample sizes. With a sample size reaching $n = 400$, all three variance estimators provided almost identical estimates of precision.

For $\hat{V}_{\text{Sarndal}}$ the differences between the different modeling techniques were almost identical compared to the results obtained for the estimator after Fuller. A notable difference is the reduced overestimation of precision for FULL, which dropped from 27.26% to 8.52%.

8.3. Hinton

8.3.1. Simple variance estimator

Summary statistics for the simple variance estimator for the Hinton dataset (HIN) can be found in Table 8.10 on page 90. For Hinton, in total 29 variables (see Table 6.2 on page 55) were available.

For $n = 50$ precision was overestimated for all stepwise variable selection procedures (AIC, AICc, and BIC), as well as for FULL, CON, and VIF. For FULL precision was overestimated by over 66% and for AIC by almost 60%. The relative ESE for FULL (7.76%) was larger than for SI (7.10%). However, while SI estimated the variance correctly (7.16%), the relative AVSE for FULL was 2.61%. The AIC behaved in almost the same manner. In contrast, for PLSR, Ridge and RF, precision was estimated correctly.

When sample sizes increased to $n = 100$, variances were still underestimated if stepwise variable selection procedures were employed. However, the differences became smaller and for none of the modeling techniques relative ESEs were larger than the ESE for SI.

For $n = 200$, coverage rates for all procedures were above or equal to 0.90. However, FULL and stepwise procedures still overestimated precision. The lowest ESE was found for VIF (for both $n = 200$ and $n = 400$). When sample sizes increased to $n = 400$, standard errors were overestimated by several modeling techniques, including CON, Lasso, PLSR, RF, Ridge, and VIF.

8.3.2. Variance estimator after Fuller

For the estimator $\hat{V}_{\text{Fuller}}$, considerable differences among the modeling techniques are observed when the sample size changes. For $n = 50$, CON outperforms stepwise selection procedures, regarding both relative ESE and coverage rates. Differences between CON and VIF are small, however. When the sample size increases, ESEs drop much faster for AIC, AICc, BIC, FULL, and VIF than for CON. For a sample size of $n = 400$ relative ESEs were all well below the relative ESE for CON.

8.3.3. Variance estimator after Särndal

For a sample size of $n = 50$, the variance estimator $\hat{V}_{\text{Sarndal}}$ showed efficiencies of 98.9 for FULL. This is in sharp contrast to the simple estimator where the efficiency was 36.6. Hence, while the simple variance estimator grossly overestimated precision, $\hat{V}_{\text{Sarndal}}$ estimated variances correctly.

However, this did not hold for stepwise variable selection procedures. Here, variances were still underestimated by 40–36%. Coverage rates for the stepwise procedures were, therefore, below or equal to 0.80.

When the sample size increased overestimation of precision decreased for AIC, AICc, and BIC. For a sample size of $n = 400$ precision was correctly estimated for all variable selection procedures.

Table 8.11 on page 91 lists the average number of variables that were included in the working models by the different variable selection procedures.

Table 8.6.: $\hat{V}_{\text{Simple}}$; Relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %; DIFF = (ESE - AVSE) / ESE×100), efficiency (EFF), and coverage rates (COV) for dataset Hedmark.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	−0.26	14.005	13.682	2.31	100.0	0.93
AIC	0.30	6.952	4.398	36.74	32.0	0.78
AICc	0.28	6.906	4.534	34.35	33.0	0.80
BIC	0.32	6.875	4.661	32.20	33.9	0.81
CON	0.17	7.757	6.687	13.79	48.6	0.91
FULL	0.20	7.151	4.204	41.22	30.6	0.75
Lasso	0.29	6.887	5.913	14.14	43.0	0.90
PLSR	0.27	6.991	6.428	8.05	46.8	0.93
RF	−0.22	6.860	6.469	5.70	47.1	0.93
Ridge	0.30	7.083	6.435	9.15	46.8	0.91
VIF	0.13	6.837	5.340	21.90	38.8	0.87
$n = 100$						
SI	0.25	9.625	9.784	−1.66	100.0	0.95
AIC	0.29	4.612	3.606	21.82	36.8	0.87
AICc	0.30	4.635	3.630	21.70	37.1	0.88
BIC	0.33	4.637	3.738	19.38	38.2	0.89
CON	0.23	5.210	4.807	7.72	49.1	0.93
FULL	0.32	4.657	3.524	24.34	36.0	0.86
Lasso	0.30	4.766	4.271	10.38	43.6	0.91
PLSR	0.29	4.951	4.667	5.74	47.7	0.92
RF	−0.03	4.358	4.167	4.37	42.6	0.93
Ridge	0.30	4.835	4.491	7.11	45.9	0.93
VIF	0.27	4.571	3.972	13.11	40.6	0.91
$n = 200$						
SI	0.16	7.094	6.944	2.13	100.0	0.94
AIC	0.11	2.922	2.696	7.74	38.8	0.93
AICc	0.11	2.916	2.700	7.42	38.9	0.94
BIC	0.12	2.939	2.757	6.20	39.7	0.94
CON	0.14	3.655	3.434	6.04	49.5	0.94
FULL	0.09	2.916	2.666	8.58	38.4	0.93
Lasso	0.15	3.000	2.982	0.60	43.0	0.94
PLSR	0.07	3.311	3.329	−0.53	48.0	0.94
Ridge	0.11	3.088	3.083	0.16	44.4	0.94
VIF	0.08	2.943	2.840	3.49	40.9	0.94
$n = 400$						
SI	−0.02	4.924	4.901	0.46	100.0	0.96
AIC	0.07	2.040	1.948	4.48	39.7	0.94
AICc	0.07	2.042	1.949	4.54	39.7	0.94
BIC	0.06	2.053	1.974	3.85	40.2	0.94
CON	0.09	2.408	2.396	0.52	48.8	0.95
FULL	0.08	2.036	1.938	4.79	39.5	0.93
Lasso	0.07	2.101	2.072	1.40	42.2	0.94
PLSR	0.03	2.394	2.357	1.53	48.1	0.94
Ridge	0.06	2.135	2.125	0.46	43.3	0.94
VIF	0.04	2.045	2.018	1.31	41.1	0.94

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Table 8.7.: Average number of variables that were included in the working model (out of 18; after 50,000 iterations; Hedmark).

Sample size	AIC	AICc	BIC	CON	VIF
50	8.41	6.90	5.90	4.30	4.29
100	8.26	6.55	5.46	4.61	4.92
200	8.88	8.55	5.80	4.76	5.83
400	9.92	9.77	6.71	4.93	6.88

Table 8.8.: $\hat{V}_{\text{Fuller}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hedmark.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	−0.26	14.005	13.682	2.31	100.0	0.93
AIC	0.30	6.954	4.771	31.39	34.7	0.82
AICc	0.28	6.908	4.832	30.06	35.1	0.82
BIC	0.32	6.877	4.910	28.61	35.7	0.83
CON	0.17	7.759	6.918	10.85	50.3	0.91
FULL	0.20	7.153	5.203	27.26	37.8	0.83
VIF	0.13	6.839	5.529	19.16	40.2	0.88
$n = 100$						
SI	0.25	9.625	9.784	−1.66	100.0	0.95
AIC	0.29	4.614	3.747	18.79	38.3	0.89
AICc	0.30	4.637	3.757	18.98	38.4	0.89
BIC	0.33	4.638	3.826	17.50	39.1	0.90
CON	0.23	5.211	4.896	6.04	50.0	0.94
FULL	0.32	4.659	3.873	16.87	39.6	0.90
VIF	0.27	4.573	4.054	11.34	41.4	0.92
$n = 200$						
SI	0.16	7.094	6.944	2.13	100.0	0.94
AIC	0.11	2.922	2.751	5.86	39.6	0.94
AICc	0.11	2.916	2.753	5.61	39.7	0.94
BIC	0.12	2.939	2.791	5.05	40.2	0.94
CON	0.14	3.655	3.466	5.16	50.0	0.94
FULL	0.09	2.916	2.788	4.41	40.2	0.94
VIF	0.08	2.943	2.875	2.30	41.4	0.94
$n = 400$						
SI	−0.02	4.924	4.901	0.46	100.0	0.96
AIC	0.07	2.040	1.971	3.40	40.2	0.94
AICc	0.07	2.042	1.971	3.47	40.2	0.94
BIC	0.06	2.053	1.988	3.15	40.5	0.94
CON	0.09	2.408	2.408	0.03	49.1	0.95
FULL	0.08	2.036	1.981	2.70	40.4	0.94
VIF	0.04	2.045	2.033	0.57	41.4	0.94

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Table 8.9.: $\hat{V}_{\text{Sarndal}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hedmark.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	−0.26	14.005	13.682	2.31	100.0	0.93
AIC	0.30	6.954	5.247	24.55	38.2	0.86
AICc	0.28	6.908	5.229	24.31	38.0	0.86
BIC	0.32	6.877	5.232	23.93	38.0	0.86
CON	0.17	7.759	7.120	8.23	51.8	0.93
FULL	0.20	7.153	6.544	8.52	47.6	0.91
VIF	0.13	6.839	5.786	15.40	42.1	0.90
$n = 100$						
SI	0.25	9.625	9.784	−1.66	100.0	0.95
AIC	0.29	4.614	3.937	14.67	40.2	0.91
AICc	0.30	4.637	3.924	15.37	40.1	0.91
BIC	0.33	4.638	3.947	14.89	40.3	0.91
CON	0.23	5.211	5.022	3.63	51.3	0.94
FULL	0.32	4.659	4.295	7.81	43.9	0.92
VIF	0.27	4.573	4.177	8.66	42.7	0.94
$n = 200$						
SI	0.16	7.094	6.944	2.13	100.0	0.94
AIC	0.11	2.922	2.834	3.02	40.8	0.95
AICc	0.11	2.916	2.833	2.87	40.8	0.95
BIC	0.12	2.939	2.846	3.18	41.0	0.95
CON	0.14	3.655	3.512	3.91	50.6	0.94
FULL	0.09	2.916	2.945	−0.98	42.4	0.95
VIF	0.08	2.943	2.932	0.37	42.3	0.95
$n = 400$						
SI	−0.02	4.924	4.901	0.46	100.0	0.96
AIC	0.07	2.040	2.002	1.84	40.8	0.94
AICc	0.07	2.042	2.002	1.95	40.8	0.94
BIC	0.06	2.053	2.011	2.06	41.0	0.94
CON	0.09	2.408	2.422	−0.56	49.4	0.95
FULL	0.08	2.036	2.033	0.12	41.5	0.95
VIF	0.04	2.045	2.058	−0.62	41.9	0.95

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Table 8.10.: $\hat{V}_{\text{Simple}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hinton.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	0.35	7.053	7.162	−1.55	100.0	0.95
AIC	0.01	6.941	2.805	59.59	39.3	0.60
AICc	−0.08	6.058	3.119	48.51	43.7	0.71
BIC	0.12	6.154	3.142	48.94	44.0	0.71
CON	0.03	4.594	4.417	3.86	61.9	0.94
FULL	0.00	7.757	2.613	66.32	36.6	0.52
Lasso	0.01	4.741	4.392	7.36	61.5	0.92
PLSR	−0.14	4.283	4.262	0.49	59.7	0.94
RF	−0.24	4.611	4.692	−1.76	65.7	0.95
Ridge	0.07	4.621	4.563	1.26	63.9	0.94
VIF	−0.02	4.618	3.941	14.66	55.2	0.91
$n = 100$						
SI	0.25	4.962	5.087	−2.52	100.0	0.95
AIC	−0.08	3.461	2.540	26.63	50.1	0.85
AICc	0.09	3.413	2.583	24.31	50.9	0.85
BIC	−0.15	3.332	2.705	18.83	53.3	0.88
CON	−0.06	3.338	3.176	4.87	62.6	0.94
FULL	0.09	3.632	2.440	32.82	48.1	0.82
Lasso	0.02	3.295	3.084	6.41	60.8	0.93
PLSR	0.04	3.250	3.050	6.13	60.2	0.94
RF	−0.03	3.333	3.236	2.92	63.8	0.94
Ridge	−0.00	3.327	3.166	4.84	62.4	0.94
VIF	−0.12	3.195	2.879	9.88	56.8	0.92
$n = 200$						
SI	−0.17	3.675	3.583	2.50	100.0	0.95
AIC	−0.02	2.178	1.935	11.15	53.9	0.91
AICc	0.01	2.182	1.942	10.98	54.1	0.91
BIC	0.01	2.135	2.005	6.06	55.9	0.94
CON	−0.03	2.329	2.279	2.14	63.5	0.94
FULL	0.00	2.203	1.897	13.91	52.9	0.90
Lasso	−0.05	2.203	2.161	1.92	60.2	0.94
PLSR	−0.00	2.197	2.176	0.93	60.6	0.94
Ridge	−0.05	2.223	2.196	1.23	61.2	0.94
VIF	0.01	2.106	2.071	1.65	57.7	0.94
$n = 400$						
SI	−0.05	2.512	2.541	−1.18	100.0	0.96
AIC	−0.00	1.469	1.408	4.17	55.4	0.94
AICc	−0.00	1.470	1.410	4.09	55.4	0.94
BIC	−0.01	1.457	1.439	1.23	56.6	0.95
CON	−0.01	1.596	1.616	−1.29	63.6	0.95
FULL	0.01	1.494	1.395	6.67	54.9	0.94
Lasso	−0.02	1.481	1.510	−1.97	59.4	0.95
PLSR	−0.01	1.493	1.541	−3.21	60.6	0.96
Ridge	−0.02	1.485	1.526	−2.71	60.0	0.96
VIF	−0.02	1.438	1.473	−2.45	57.9	0.96

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Table 8.11.: Average number of variables that were included in the working model (out of 29; after 50,000 iterations; Hinton).

Sample size	AIC	AICc	BIC	CON	VIF
50	15.31	10.05	9.87	2.94	4.27
100	11.95	10.00	6.35	2.86	4.72
200	11.68	10.83	5.89	2.73	5.07
400	12.54	12.04	6.13	2.69	5.46

Table 8.12.: $\hat{V}_{\text{Fuller}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hinton.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	0.35	7.053	7.162	−1.55	100.0	0.95
AIC	0.01	6.937	3.330	52.00	46.7	0.68
AICc	−0.08	6.054	3.446	43.08	48.3	0.76
BIC	0.12	6.150	3.460	43.75	48.5	0.76
CON	0.03	4.591	4.504	1.90	63.1	0.94
FULL	−0.00	7.753	3.988	48.55	55.9	0.70
VIF	−0.02	4.615	4.076	11.69	57.1	0.92
$n = 100$						
SI	0.25	4.962	5.087	−2.52	100.0	0.95
AIC	−0.08	3.460	2.692	22.21	53.1	0.87
AICc	−0.09	3.412	2.708	20.63	53.4	0.88
BIC	0.15	3.331	2.779	16.57	54.8	0.90
CON	0.06	3.337	3.205	3.96	63.2	0.94
FULL	0.09	3.631	2.880	20.67	56.8	0.89
VIF	−0.12	3.194	2.934	8.14	57.9	0.92
$n = 200$						
SI	−0.17	3.675	3.583	2.50	100.0	0.95
AIC	0.02	2.178	1.990	8.66	55.4	0.92
AICc	−0.01	2.182	1.992	8.70	55.5	0.92
BIC	0.01	2.135	2.030	4.89	56.6	0.94
CON	0.03	2.329	2.289	1.71	63.8	0.94
FULL	−0.00	2.203	2.046	7.13	57.0	0.92
VIF	0.01	2.106	2.093	0.63	58.3	0.94
$n = 400$						
SI	−0.05	2.512	2.541	−1.18	100.0	0.96
AIC	−0.00	1.469	1.429	2.75	56.2	0.95
AICc	0.00	1.470	1.429	2.73	56.2	0.95
BIC	−0.01	1.457	1.448	0.59	57.0	0.95
CON	0.01	1.596	1.620	−1.51	63.7	0.95
FULL	0.01	1.494	1.446	3.21	56.9	0.95
VIF	−0.02	1.438	1.482	−3.02	58.3	0.96

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Table 8.13.: $\hat{V}_{\text{Sarndal}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hinton.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	0.35	7.053	7.162	−1.55	100.0	0.95
AIC	0.01	6.937	4.096	40.95	57.4	0.77
AICc	−0.08	6.054	3.851	36.39	54.0	0.80
BIC	0.12	6.150	3.877	36.97	54.3	0.80
CON	0.03	4.591	4.610	−0.40	64.6	0.95
FULL	0.00	7.753	7.058	8.96	98.9	0.92
VIF	−0.02	4.615	4.203	8.94	58.9	0.93
$n = 100$						
SI	0.25	4.962	5.087	−2.52	100.0	0.95
AIC	−0.08	3.460	2.869	17.07	56.6	0.89
AICc	−0.09	3.412	2.851	16.43	56.2	0.89
BIC	0.15	3.331	2.870	13.83	56.6	0.91
CON	0.06	3.337	3.253	2.52	64.2	0.94
FULL	−0.09	3.631	3.503	3.53	69.1	0.93
VIF	−0.12	3.194	3.003	5.99	59.2	0.93
$n = 200$						
SI	−0.17	3.675	3.583	2.50	100.0	0.95
AIC	−0.02	2.178	2.044	6.19	56.9	0.93
AICc	−0.01	2.182	2.041	6.47	56.9	0.92
BIC	0.01	2.135	2.053	3.82	57.2	0.94
CON	0.03	2.329	2.297	1.37	64.0	0.94
FULL	−0.00	2.203	2.218	−0.65	61.8	0.95
VIF	0.01	2.106	2.112	−0.27	58.8	0.94
$n = 400$						
SI	−0.05	2.512	2.541	−1.18	100.0	0.96
AIC	−0.00	1.469	1.450	1.30	57.0	0.95
AICc	−0.00	1.470	1.450	1.34	57.0	0.95
BIC	−0.01	1.457	1.458	−0.06	57.3	0.95
CON	−0.01	1.596	1.624	−1.75	63.9	0.95
FULL	0.01	1.494	1.503	−0.55	59.1	0.95
VIF	0.02	1.438	1.490	−3.62	58.6	0.96

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

9. Model-based inference

9.1. Artificial datasets

For model-based inference only a single variance estimator was considered (see equation 7.9 in Section 7.3). The results of the simulation study for the datasets NOISE, COR, and DCOR are given in Table 9.1.

For the NOISE dataset, the relative ESE for SI was smaller than the relative ESE for FULL and CON, leading to efficiencies above 100. However, coverage rates for these two modeling strategies were close to 0.95. In contrast, coverage rates for AIC, AICc, VIF, and VIFB were all below 0.9, and differences between the relative ESE and AVSE were positive. Hence, precision was overestimated.

For the dataset COR precision was overestimated for all stepwise selection procedures. Although for COR, VIF, and VIFB, precision was estimated unbiasedly, the relative AVSE was relatively larger compared to the relative ESE for the stepwise procedures. Similar to the estimator after Särndal, FULL showed coverage rates close to 0.95, and a small ESE.

The same was observed for the dataset DCOR. The largest differences between the relative ESE and AVSE for the DCOR were found for VIF and VIFB.

Generally, the results for the three artificial datasets were almost identical to the results obtained for the estimator after Särndal above. The same is true for the two synthetic populations HED and HIN.

Table 9.1.: $\hat{V}_{\mathrm{MD}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset NOISE (top), COR (middle), and DCOR (bottom).

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
NOISE						
SI	0.10	1.443	1.541	−6.75	100.0	0.95
AIC	0.14	1.970	1.444	26.70	93.7	0.86
AICc	0.14	1.884	1.428	24.20	92.7	0.87
BIC	0.13	1.889	1.438	23.84	93.3	0.88
CON	0.06	1.908	1.908	−0.01	123.8	0.94
FULL	0.17	2.094	2.169	−3.63	140.8	0.96
VIF	0.14	1.970	1.444	26.70	93.7	0.86
VIFB	0.12	1.860	1.423	23.47	92.4	0.88
COR						
SI	0.09	1.575	1.563	0.71	100.0	0.95
AIC	0.09	0.170	0.117	30.92	7.5	0.82
AICc	0.13	0.180	0.117	34.62	7.5	0.80
BIC	0.11	0.178	0.117	34.39	7.5	0.80
CON	0.16	0.367	0.361	1.53	23.1	0.95
FULL	0.16	0.162	0.158	2.78	10.1	0.93
VIF	0.13	0.346	0.333	3.81	21.3	0.95
VIFB	0.12	0.346	0.333	3.81	21.3	0.95
DCOR						
SI	−0.10	1.576	1.574	0.15	100.0	0.96
AIC	0.12	0.745	0.518	30.40	32.9	0.83
AICc	0.09	0.700	0.514	26.56	32.7	0.84
BIC	0.11	0.685	0.518	24.41	32.9	0.85
CON	−0.12	1.809	1.633	9.71	103.8	0.92
FULL	0.15	0.807	0.760	5.84	48.3	0.93
VIF	0.09	1.092	0.686	37.15	43.6	0.82
VIFB	0.12	1.046	0.678	35.19	43.1	0.84

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

9.2. Hedmark

For HED, precision was overestimated for all modeling techniques. For a sample size of $n = 50$, the difference between the relative ESE and AVSE was largest for the three stepwise procedures and VIF and VIFB. The largest relative ESE was found for CON.

When the sample size increased, the difference between CON and all other procedures increased. While the ESE for CON decreased slowly, the ESE of all other techniques decreased relatively fast.

Table 9.2.: $\hat{V}_{\mathrm{MD}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hedmark.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	−0.26	14.005	13.682	2.31	100.0	0.93
AIC	−0.12	6.958	5.143	26.08	37.6	0.85
AICc	−0.21	6.929	5.093	26.50	37.3	0.85
BIC	−0.22	6.944	5.072	26.95	37.1	0.85
CON	0.18	7.759	7.120	8.23	50.5	0.93
FULL	−0.15	7.151	6.532	8.66	47.8	0.92
VIF	−0.13	6.941	5.618	19.06	41.1	0.89
$n = 100$						
SI	0.25	9.625	9.784	−1.66	100.0	0.95
AIC	−0.12	4.640	3.835	17.34	39.4	0.90
AICc	−0.24	4.659	3.826	17.88	39.3	0.90
BIC	−0.22	4.669	3.844	17.67	39.5	0.90
CON	0.17	5.211	5.022	3.36	51.3	0.94
FULL	−0.18	4.689	4.224	9.91	43.4	0.93
VIF	−0.12	4.629	4.070	12.08	41.8	0.92
$n = 200$						
SI	0.16	7.094	6.944	2.13	100.0	0.94
AIC	−0.21	2.944	2.761	6.22	40.0	0.93
AICc	−0.23	2.938	2.760	6.06	40.0	0.93
BIC	−0.24	2.967	2.773	6.56	40.2	0.93
CON	0.	3.655	3.512	3.91	50.6	0.94
FULL	−0.22	2.937	2.881	1.90	41.8	0.94
VIF	−0.26	2.988	2.864	4.17	41.5	0.93
$n = 400$						
SI	−0.10	4.924	4.901	0.46	100.0	0.96
AIC	−0.15	2.044	1.960	4.13	40.2	0.92
AICc	−0.19	2.049	1.960	4.36	40.2	0.92
BIC	−0.11	2.075	1.965	5.28	40.3	0.91
CON	0.22	2.408	2.422	−0.56	49.4	0.95
FULL	−0.24	2.040	1.998	2.05	41.0	0.93
VIF	−0.14	2.058	2.017	2.02	41.4	0.93

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

9.3. Hinton

Similar to the results obtained for HED and the three artificial datasets, the results for HIN resembled the findings for the variance estimator after Särndal.

For a sample size of $n = 50$, CON performed best in terms of coverage rate and magnitude of the relative ESE. This changed when the sample size increased. For a sample size of $n = 400$, the coverage rates were all close or equal to 0.95 and efficiencies were all < 60, except for CON.

Table 9.3.: $\hat{V}_{\text{MD}}$; relative bias (rBias in %), relative empirical standard error (ESE %), relative average standard error (AVSE %), difference between ESE and AVSE (DIFF %), efficiency (EFF), and coverage rates (COV) for dataset Hinton.

Method[a]	rBias (%)	ESE (%)	AVSE (%)	DIFF[b](%)	EFF	COV
$n = 50$						
SI	0.35	7.053	7.162	−1.55	100.0	0.95
AIC	−0.01	6.931	4.155	40.04	58.3	0.78
AICc	−0.08	6.061	3.864	36.24	54.2	0.80
BIC	−0.12	6.169	3.870	37.26	54.3	0.80
CON	−0.09	4.591	4.610	−0.40	64.6	0.95
FULL	−0.04	7.739	7.182	7.20	100.8	0.93
VIF	−0.12	4.628	4.150	10.33	58.3	0.92
$n = 100$						
SI	0.26	4.962	5.087	−2.52	100.0	0.95
AIC	−0.08	3.455	2.861	17.21	56.5	0.90
AICc	−0.09	3.412	2.836	16.89	56.0	0.90
BIC	−0.12	3.349	2.825	15.64	55.8	0.90
CON	−0.11	3.337	3.253	3.52	64.2	0.94
FULL	−0.09	3.623	3.535	2.43	69.9	0.94
VIF	−0.13	3.184	2.952	7.30	58.3	0.93
$n = 200$						
SI	−0.17	3.676	3.584	2.50	100.0	0.95
AIC	−0.00	2.174	2.039	6.23	56.9	0.93
AICc	−0.00	2.176	2.035	6.45	56.8	0.93
BIC	−0.00	2.141	2.035	4.95	56.8	0.94
CON	−0.03	2.329	2.297	1.37	61.8	0.94
FULL	−0.01	2.197	2.229	−1.45	62.2	0.95
VIF	−0.00	2.123	2.087	1.71	58.2	0.95
$n = 400$						
SI	−0.05	2.510	2.539	−1.18	100.0	0.96
AIC	−0.00	1.472	1.446	1.76	56.9	0.95
AICc	−0.00	1.472	1.445	1.80	56.9	0.95
BIC	−0.01	1.471	1.446	1.69	56.9	0.94
CON	−0.00	1.569	1.624	−1.75	63.9	0.95
FULL	0.02	1.494	1.503	−0.60	59.2	0.96
VIF	−0.00	1.449	1.474	−1.70	58.0	0.95

[a]Method = model or variable selection procedure
[b]DIFF = (ESE - AVSE) / ESE×100

Part IV.

Discussion & Conclusions

10. Discussion

This study investigated the impact of several model and variable selection procedures on estimates of precision in a model-assisted, as well as model-based inference framework. Results of simulation studies suggest that different modeling strategies lead to different estimates of precision in both inference frameworks. In this chapter, the different model and variable selection procedures employed in this study will be discussed.

10.1. Stepwise selection procedures

In this study, stepwise variable selection procedures that used the AIC, AICc, and BIC as a criterion, generally lead to overly optimistic estimates of precision in both, the model-assisted and model-based inference framework. This was in particular observed when the sample size was small relative to the number of covariates. It has been noted by, for example, Claeskens & Hjort (2008), that the AIC does not necessarily identify the most parsimonious model. Hence, there is no guarantee that the AIC does not overfit the sample data.

In the simulation study that was conducted on the dataset NOISE (see Section 8.1.1 in Chapter 8), the "overfitting property" of the AIC, AICc, and BIC in finite samples was probably most apparent. Although there was no relationship between the target variable and the covariates, estimates of precision suggested high efficiencies. However, the bias of estimates of precision largely depended on the sample size. For sufficiently large sample sizes, $n = 400$, coverage rates were equal to 0.95.

When n is small relative to the number of potential covariates, Burnham & Anderson (2004) suggested to always use the AICc instead of the AIC. They argued that the AICc is less prone to detecting spurious effects and asymptotically converge to the AIC. However, in this study it was shown that the AICc did not generally perform better than the AIC when the sample size was ≤ 100. In all simulation studies differences between precision estimates for AIC and AICc were small, except for small sample sizes

for Hinton were coverage rates of AIC were lower than for AICc. The same holds for the BIC.

The differences among precision estimates observed for the saturated model and for models obtained after applying stepwise variable selection procedures showed that the choice of the variance estimator matters. While for the simple variance estimator overestimation was reduced by using stepwise procedures, the reverse was observed for the variance estimator proposed by Särndal *et al.* (1992). For the g-weighted variance estimator the difference between average estimated standard errors and empirical standard errors were smaller for the saturated model than for the model obtained after stepwise selection.

As the g-weighted variance estimator performed similarly to the model-based variance estimator, the effects of stepwise selection were almost identical among model-assisted (using the g-weighted variance estimator) and model-based estimates of precision.

10.2. Variance inflation factors

The use of the variance inflation factor to remove correlated variables (combined with the AIC) had no effect when applied to the dataset NOISE. This is not surprising, since the potential covariates were not related in NOISE. Therefore, results were similar to the findings for the AIC. However, for the COR dataset, the two procedures VIF and AIC performed very differently. Removing highly correlated variables from the model lead to unbiased estimates of precision and, compared to stepwise procedures, to relatively high empirical standard errors. In contrast, for the AIC, precision was overestimated by almost 50%, but empirical standard errors were small compared to the VIF procedure. For the dataset COR, this pattern was observed for all variance estimators considered in this study. For the dataset DCOR, as well as for the two synthetic populations, empirical standard errors were still lower for stepwise selection procedures; overestimation of precision was, however, larger for VIF.

Results obtained for the two synthetic populations indicate that for LiDAR data, where covariates are often highly correlated, estimates of precision are less biased when the variance inflation factor is used to remove correlated variables. This was observed for all variance estimators employed. For the Hedmark data the bias reduction of variance estimates was, however, less pronounced than for the Hinton dataset. For Hedmark the average number of variables included in the model for BIC was roughly equal to

the number of variables selected by the VIF procedure. For Hinton, in contrast, the BIC selected on average more variables than VIF. For large sample sizes, employing the variance inflation factor in combination with the AIC performed best among all modeling techniques considered in this study.

For the three artificial datasets the differences between VIF combined with the AIC and the VIF combined with best-subset selection (based on Mallow's C_p) was small. Modest differences were only found for the NOISE dataset.

In this study the threshold for the variance inflation factor was set to 10. This decision was entirely subjective. A different threshold value would, most likely, lead to different results.

10.3. Condition number

For the procedure CON, results were mixed. For the datasets NOISE and COR, CON lead to results comparable to VIF. For DCOR the procedure performed poorly. For the latter, the CON approach almost always removed the most powerful predictor variables to reduce the condition number of the cross-product matrix.

For the two synthetic populations the performance of CON was remarkably different. While for Hedmark, the empirical standard errors were high for all sample sizes and variance estimators, compared to other modeling techniques, CON performed well for small sample sizes for Hinton. However, when the sample size increased, other procedures outperformed CON substantially. This indicates that the performance of CON relative to the other modeling techniques depends on the sample size and correlation structure in the data. The large differences between the VIF and CON approaches (for the synthetic populations) might be attributed to (a) that the the AIC has been applied to VIF after correlated variables have been removed, and (b) the way CON and VIF remove variables. For the variance inflation factor covariates are considered individually. For the condition number, in contrast, all covariates are considered at once. The generally poor performance of CON may be attributed to the fact that it does not consider the target variable. CON may be useful approach for calibration, as shown by Bankier *et al.* (1992), however, when a target variable is specified, alternative approaches seem to outperform CON.

10.4. Regularization

In contrast to the aforementioned modeling techniques, ridge regression and Lasso were applied only when the simple (model-assisted) variance estimator was employed. Both modeling techniques showed promising results in terms of coverage rates, as well as empirical standard errors. In particular for small sample sizes overestimation of precision was small. For Hinton and Hedmark, ridge regression performed slightly better than the Lasso in small samples. For larger sample sizes coverage rates were almost identical. Although regularization techniques have, to the author's knowledge, not been applied to LiDAR data in FRI applications, results from the simulation studies suggest that they might outperform commonly applied variable selection procedures such as the AIC and BIC. However, for large sample sizes the combination of the variance inflation factor combined with AIC lead to smaller empirical standard errors. But even if sample sizes are large, regularization techniques performed only slightly worse than the VIF procedure.

As an alternative regularization technique, elastic nets might be considered. However, in a pilot study (that was conducted before the large-scale simulation studies) elastic nets showed very similar behavior compared to the Lasso. Therefore, elastic nets were not included in this study.

10.5. Partial least squares regression

Partial least squares regression proved to be useful for highly correlated covariates (dataset COR) and small sample sizes. With regard to the former, partial least squares might, therefore, be particularly suited for LiDAR data. However, when sample sizes increased for Hinton and Hedmark partial least squares were outperformed by other modeling strategies. This may be caused by the fact, that only one component was selected for modeling. Alternatively, cross-validation might be used to obtain a suitable number of components. This approach was, however, not investigated in this study.

10.6. Random forests

LiDAR applications in FRIs are dominated by parametric approaches. In this study, the random forest (RF) algorithm was included. It has to be mentioned, that RF has, to the author's knowledge, not been applied in a model-assisted context. In this study, the

naïve assumption was made, that RF provides a measure of the "variance explained" (a "pseudo-R^2"). This measure was used as a substitute for R^2 in the simple model-assisted variance estimator given in Chapter 3. Whether this is justified from a statistical point of view might be questionable. Therefore, results obtained for RF should be interpreted with caution. The decision to include RF was largely due to the increasing popularity in FRIs.

In this study RF performed relatively well. For the synthetic populations coverage rates were close to 0.95 were RF was employed, and, thus, precision was hardly overestimated. Only for the dataset NOISE, RF showed much larger empirical standard errors than alternative procedures.

The good performance of RF for Hedmark may be attributed to how the target variable AGB was related to the covariates. Although variables have been transformed for Hedmark, it was impossible to remove non-linear effects entirely. In that case parametric models may fail to capture the associations as good as RF. The differences between RF and the other modeling techniques might change when a different target variable is considered. For example, plot level mean tree height is often linearly related to various LiDAR metrics.

10.7. Further comments

10.7.1. Cross-validation

As outlined in Chapter 2, cross-validation is frequently used to evaluate the performance of a postulated model. Cross-validation was also applied in the pilot study conducted for this project. However, results from cross-validation seem to be highly dependent on the number of folds. When the dataset was split into two, a training and test set, estimates of variance were generally large (in particular for small sample sizes), and where the variance was considerably overestimated. The picture changes when the number of folds was changed to three, five, and ten. However, for different datasets, sample sizes, and fold sizes, no general trend could be observed of how cross-validation should be performed.

10.7.2. Expert knowledge

In this study only purely automatic approaches were tested — without any human interaction during the simulation runs. This is, of course, different in practical applications. Here an analyst may interact at various stages during data analysis. For example, while in practice a regression model that reveals an $R^2 < 0.3$ may not be considered as useful, this model has always been estimated in the simulation studies.

Moreover, when a large set of potential covariates is available, an analyst may first evaluate which variables are potentially useful before any of the techniques considered in this study are applied. For example, the number of potential variables may be reduced to four or five variables before a stepwise procedure, such as AIC or BIC, is used. In several applications of LiDAR in FRIs, only few or only a single variable entered the final model (e.g., Asner *et al.*, 2012).

Furthermore, adding interaction terms between LiDAR metrics may improve model performance. Ene *et al.* (2012), for example, used an interaction term for top height and a density metric. In this study interaction terms were not considered, largely due to computational costs. If interactions among variables are allowed, the set of candidate models becomes extremely large. Moreover, when even more models are considered as candidate models the risk of identifying a model that fits well to the sample data but poorly to unseen data may further increase.

10.7.3. Alternative modeling techniques

Numerous other modeling techniques could have been considered for this study, including boosted regression (or boosting), generalized additive models, smoothing splines, and many more. Multi-model inference and Bayesian approaches could also be considered. The selection of methods for this study was obviously somewhat subjective and limited.

Breidt *et al.* (2005), Opsomer *et al.* (2007), and Wang & Bellhouse (2009), for example, all used semi- and non-parametric approaches in complex survey designs. The techniques these authors considered, e.g., penalized splines, are generally more flexible than the classical linear model. However, as the results from this study suggest, these techniques would require large sample sizes. Applied to moderate and small sample sizes, the risk of overestimating precision may even be larger for these techniques than for those considered in this study.

However, in this study the focus was on techniques that are frequently applied in LiDAR-assisted FRIs. Ridge regression and the Lasso were included because they are known to alleviate problems in modeling that might be encountered when working with LiDAR data and are relatively easy to apply. Results from the simulation studies revealed that these regularization techniques performed comparatively well.

However, the goal of this study was not to identify a single modeling technique that performs best for any given dataset. The aim was to evaluate if, and how different techniques affect estimates of precision. The setting for the simulation studies was purposefully kept simple, i.e., simple random sampling without replacement was used. For other sample designs, such as systematic and/or stratified sampling, the general findings of this study may not become obsolete or invalid. However, one should not expect that the results from this study are readily transferable to entirely different sample designs. For systematic sampling, which is commonly applied in FRIs, one should expect, that the risk of overly optimistic estimates of precision is lower than for simple random sampling.

Moreover, in this study only two datasets were used that are based on FRI data. Results may differ for other study areas. For that reason, datasets were simulated to evaluate specific data structures. As the findings among the different datasets varied, further research is needed to evaluate if the results obtained from this study can be transfered to other LiDAR datasets.

11. Conlusions

The following conclusions can be drawn from this study:

1. Model and variable selection in LiDAR-assisted FRIs affect estimates of precision in the model-assisted, as well as model-based inference framework.

2. Blind application of stepwise variable selection procedures leads to overly optimistic estimates of precision in LiDAR-assisted FRIs. The underestimation of variances can be severe when sample sizes are small, that is, when the ratio between the number of observations and LiDAR metrics is relatively small. For large sample sizes, the negative bias of precision estimates becomes negligible.

3. Pre-selection of variables, by means of variance inflation factors, coupled with stepwise selection procedures, lead to unbiased estimates of variances for small samples.

4. In this study, the use of the condition number to remove covariates provided mixed results. Based on these findings, its use cannot be recommended for LiDAR-assisted FRIs.

5. Results of the simulation studies suggest that ridge regression and the Lasso lead to high efficiencies and unbiased estimates of precision for small, as well as large sample sizes.

6. Partial least squares regression performs well when the sample size is small and covariates are highly correlated. This conclusion holds when a single component is used.

7. Random forests provide a useful alternative to parametric approaches.

This study concludes that the use of ridge regression, Lasso, and random forests may prove useful in future LiDAR-assisted FRIs.

Bibliography

Aas, K., Czado, C., Frigessi, A. & Bakken, H., 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182 – 198.

Ahmed, R., Siqueira, P. & Hensley, S., 2013. A study of forest biomass estimates from lidar in the northern temperate forests of new england. *Remote Sensing of Environment*, 130(0):121 – 135.

Alexander, C., Moeslund, J.E., Bøcher, P.K., Arge, L. & Svenning, J.C., 2013. Airborne laser scanner (lidar) proxies for understory light conditions. *Remote Sensing of Environment*, 134(0):152 – 161.

Andersen, H.E., McGaughey, R.J. & Reutebuch, S.E., 2005. Estimating forest canopy fuel parameters using LIDAR data. *Remote Sensing of Environment*, 94(4):441 – 449.

Andersen, H.E., Strunk, J., Temesgen, H., Atwood, D. & Winterberger, K., 2011. Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior alaska. *Canadian Journal of Remote Sensing*, 37(06):596–611.

Anderson, D.R., Burnham, K.P., Gould, W.R. & Cherry, S., 2001. Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, 29(1):311–316.

Asner, G.P., Mascaro, J., Muller-Landau, H., Vieilledent, G., Vaudry, R., Rasamoelina, M., Hall, J. & Breugel, M., 2012. A universal airborne lidar approach for tropical forest carbon mapping. *Oecologia*, 168(4):1147–1160.

Bankier, M.D., Rathwell, S. & Majkowski, M., 1992. Two step generalized least squares estimation in the 1991 canadian census. In *Proceedings of the Survey Research Methods Section, SSC Annual Meeting*, pages 764–769.

Bedford, T. & Cooke, R.M., 2002. Vines – a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068.

Brechmann, E.C. & Schepsmeier, U., 2013. Modeling dependence with C- and D-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3):1–27.

Breidenbach, J., Gläser, C. & Schmidt, M., 2008. Estimation of diameter distributions by means of airborne laser scanner data. *Canadian Journal of Forest Research*, 38(6):1611–1620.

Breidenbach, J., Nothdurft, A. & Kändler, G., 2010. Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes in central europe using airborne laser scanner data. *European Journal of Forest Research*, 129(5):833–846.

Breidenbach, J., Næsset, E. & Gobakken, T., 2012. Improving k-nearest neighbor predictions in forest inventories by combining high and low density airborne laser scanning data. *Remote Sensing of Environment*, 117(0):358 – 365.

Breidt, F.J., Claeskens, G. & Opsomer, J.D., 2005. Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Breidt, F.J., Opsomer, J.D., Johnson, A.A. & Ranalli, M.G., 2007. Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33(1):35–44.

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

Bright, B.C., Hicke, J.A. & Hudak, A.T., 2012. Estimating aboveground carbon stocks of a forest affected by mountain pine beetle in idaho using lidar and multispectral imagery. *Remote Sensing of Environment*, 124(0):270 – 281.

Brosofske, K.D., Froese, R.E., Falkowski, M.J. & Banskota, A., 2014. A review of methods for mapping and prediction of inventory attributes for operational forest management. Preprint; available at `http://www.ingentaconnect.com/content/saf/fs/pre-prints/content-forsci12134`.

Burnham, K.P. & Anderson, D.R., 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

Burnham, K. & Anderson, D., 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Chambers, R. & Clark, R., 2012. *An introduction to model-based survey sampling with applications*. Oxford University Press.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A*, 158(3):419–466.

Claeskens, G. & Hjort, N.L., 2008. *Model selection and model averaging*. Cambridge University Press Cambridge.

Clark, M.L., Clark, D.B. & Roberts, D.A., 2004. Small-footprint lidar estimation of sub-canopy elevation and tree height in a tropical rain forest landscape. *Remote Sensing of Environment*, 91(1):68–89.

Cochran, W.G., 1977. *Sampling Techniques*. John Wiley & Sons, New York, 3 edition.

Crookston, N.L. & Finley, A., 2008. yaimpute: An r package for knn imputation. *Journal of Statistical Software*, 23(10):1–16.

Czado, C., Schepsmeier, U. & Min, A., 2012. Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, 12(3):229–255.

Dalponte, M., Martinez, C., Rodeghiero, M. & Gianelle, D., 2011. The role of ground reference data collection in the prediction of stem volume with lidar data in mountain areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(6):787 – 797.

d'Oliveira, M.V., Reutebuch, S.E., McGaughey, R.J. & Andersen, H.E., 2012. Estimating forest biomass and identifying low-intensity logging areas using airborne scanning lidar in antimary state forest, acre state, western brazilian amazon. *Remote Sensing of Environment*, 124(0):479 – 491.

Drake, J.B., Knox, R.G., Dubayah, R.O., Clark, D.B., Condit, R., Blair, J.B. & Hofton, M., 2003. Above-ground biomass estimation in closed canopy neotropical forests using lidar remote sensing: Factors affecting the generality of relationships. *Global Ecology and Biogeography*, 12(2):147–159.

Duong, T., 2014. *ks: Kernel smoothing*. R package version 1.9.1.

Ene, L.T., Næsset, E. & Gobakken, T., 2013a. Model-based inference for k-nearest neighbours predictions using a canonical vine copula. *Scandinavian Journal of Forest Research*, 28(3):266–281.

Ene, L.T., Næsset, E., Gobakken, T., Gregoire, T.G., Ståhl, G. & Holm, S., 2013b. A simulation approach for accuracy assessment of two-phase post-stratified estimation in large-area lidar biomass surveys. *Remote Sensing of Environment*, 133(0):210 – 224.

Ene, L.T., Næsset, E., Gobakken, T., Gregoire, T.G., Ståhl, G. & Nelson, R., 2012. Assessing the accuracy of regional lidar-based biomass estimation using a simulation approach. *Remote Sensing of Environment*, 123:579–592.

Erdody, T.L. & Moskal, L.M., 2010. Fusion of lidar and imagery for estimating forest canopy fuels. *Remote Sensing of Environment*, 114(4):725 – 737.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B., 2013. *Regression: Models, Methods and Applications*. Springer, Heidelberg.

Faraway, J., 2011. *faraway: Functions and datasets for books by Julian Faraway.* R package version 1.0.5.

Faraway, J.J., 2004. *Linear models with R.* Chapman & Hall/CRC, London.

Frazer, G., Hobart, G., White, J. & Wulder, M., 2011a. Predictive modelling of forest inventory attributes using airborne LiDAR and ground-reference measurements derived from Hinton Wood Products' Permanent Growth Sample (PGS) program. Technical report, Canadian Forest Service, Canadian Wood Fibre Centre, Pacific Forestry Centre, Victoria, BC, and Hinton Wood Products (A division of West Fraser Mills, Ltd.), Edmonton, AB.

Frazer, G., Magnussen, S., Wulder, M. & Niemann, K., 2011b. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of lidar-derived estimates of forest stand biomass. *Remote Sensing of Environment*, 115(2):636–649.

Friedman, J., Hastie, T. & Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fuller, W.A., 2009. *Sampling statistics*. John Wiley & Sons.

García, M., Riaño, D., Chuvieco, E. & Danson, F.M., 2010. Estimating biomass carbon stocks for a mediterranean forest in central Spain using LiDAR height and intensity data. *Remote Sensing of Environment*, 114(4):816–830.

García, M., Riaño, D., Chuvieco, E., Salas, J. & Danson, F.M., 2011. Multispectral and LiDAR data fusion for fuel type mapping using support vector machine and decision rules. *Remote Sensing of Environment*, 115(6):1369 – 1379.

Garcia-Gutierrez, J., Gonzalez-Ferreiro, E., Riquelme-Santos, J.C., Miranda, D., Dieguez-Aranda, U. & Navarro-Cerrillo, R.M., 2014. Evolutionary feature selection to estimate forest stand variables using lidar. *International Journal of Applied Earth Observation and Geoinformation*, 26(0):119 – 131.

Genest, C. & Favre, A., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347.

Gleason, C.J. & Im, J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sensing of Environment*, 125(0):80 – 91.

Gobakken, T. & Næsset, E., 2005. Weibull and percentile models for LiDAR-based estimation of basal area distribution. *Scandinavian Journal of Forest Research*, 20(6):490–502.

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Holm, S., Ørka, H.O. & Astrup, R., 2012. Estimating biomass in hedmark county, norway using national forest inventory field plots and airborne laser scanning. *Remote Sensing of Environment*, 123(0):443 – 456.

Gonzalez, P., Asner, G.P., Battles, J.J., Lefsky, M.A., Waring, K.M. & Palace, M., 2010. Forest carbon densities and uncertainties from Lidar, quickbird, and field measurements in california. *Remote Sensing of Environment*, 114(7):1561 – 1575.

Gregoire, T. & Valentine, H., 2008. *Sampling Strategies for Natural Resources And The Environment*. Applied Environmental Statistics. Chapman & Hall/CRC, Boca Raton.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, 28(10):1429–1447.

Gregoire, T.G., Lin, Q.F., Boudreau, J. & Nelson, R., 2008. Regression estimation following the square-root transformation of the response. *Forest Science*, 54(6):597–606.

Hansen, M.H., Madow, W.G. & Tepping, B.J., 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning*. Springer, New York, 2 edition.

Hinton Wood Products, 2008. *Permanent Growth Sample Program Manual*. A division of West Fraser Mills Ltd.; Woodlands Department, 18 edition.

Hinton Wood Products, 2010. Mountain pine beetle forest management plan. Technical report no. 3.

Hoerl, A.E. & Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hudak, A.T., Crookston, N.L., Evans, J.S., Falkowski, M.J., Smith, A.M., Gessler, P.E. & Morgan, P., 2006. Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data. *Canadian Journal of Remote Sensing*, 32(2):126–138.

Hudak, A.T., Strand, E.K., Vierling, L.A., Byrne, J.C., Eitel, J.U., Martinuzzi, S. & Falkowski, M.J., 2012. Quantifying aboveground forest carbon pools and fluxes from repeat LiDAR surveys. *Remote Sensing of Environment*, 123(0):25 – 40.

Hurvich, C.M. & Tsai, C.L., 1989. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X. & Maltamo, M., 2008. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing*, 29(5):1339–1366.

Jakubowski, M.K., Guo, Q. & Kelly, M., 2013. Tradeoffs between lidar pulse density and forest measurement accuracy. *Remote Sensing of Environment*, 130(0):245 – 253.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An introduction to statistical learning*. Springer, New York.

Jaskierniak, D., Lane, P.N., Robinson, A. & Lucieer, A., 2011. Extracting LiDAR indices to characterise multilayered forest structure using mixture distribution functions. *Remote Sensing of Environment*, 115(2):573 – 585.

Jensen, J.L.R., Humes, K.S., Conner, T., Williams, C.J. & DeGroot, J., 2006. Estimation of biophysical characteristics for highly variable mixed-conifer stands using small-footprint lidar. *Canadian Journal of Forest Research*, 36(5):1129–1138.

Kane, V.R., McGaughey, R.J., Bakker, J.D., Gersonde, R.F., Lutz, J.A. & Franklin, J.F., 2010. Comparisons between field-and lidar-based measures of stand structural complexity. *Canadian journal of forest research*, 40(4):761–773.

Knobelspies, M. & Münnich, R., 2008. Variablenselektion bei gebundener hochrechnung. *Austrian Journal of Statistics*, 37:335–34.

Koch, B., 2010. Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):581–590.

Lambert, M.C., Ung, C.H. & Raulier, F., 2005. Canadian national tree aboveground biomass equations. *Canadian Journal of Forest Research*, 35(8):1996–2018.

Latifi, H., Fassnacht, F. & Koch, B., 2012. Forest structure modeling with combined airborne hyperspectral and lidar data. *Remote Sensing of Environment*, 121(0):10 – 25.

Latifi, H., Nothdurft, A. & Koch, B., 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/lidar-derived predictors. *Forestry*, 83(4):395–407.

Lefsky, M., Ramond, T. & Weimer, C., 2011. Alternate spatial sampling approaches for ecosystem structure inventory using spaceborne lidar. *Remote Sensing of Environment*, 115(6):1361 – 1368.

Lefsky, M., Turner, D., Guzy, M. & Cohen, W., 2005. Combining lidar estimates of aboveground biomass and landsat estimates of stand age for spatially extensive validation of modeled forest productivity. *Remote Sensing of Environment*, 95(4):549 – 558.

Lefsky, M.A., Harding, D., Cohen, W., Parker, G. & Shugart, H., 1999. Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern maryland, usa. *Remote Sensing of Environment*, 67(1):83–98.

Lehtonen, R. & Pahkinen, E., 2004. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley & Sons Ltd, West Sussex, 2nd edition.

Li, M., Im, J., Quackenbush, L.J. & Liu, T., in press. Forest biomass and carbon stock quantification using airborne lidar data: A case study over huntington wildlife forest in the adirondack park. *IEEE*.

Liaw, A. & Wiener, M., 2002. Classification and regression by randomforest. *R News*, 2(3):18–22.

Lohr, S.L., 1999. *Sampling: design and analysis*. Cengage Learning.

Lumley, T., 2009. *leaps: regression subset selection (using Fortran code by Alan Miller).* R package version 2.9.

Lumley, T., 2011. *Complex surveys: A guide to analysis using R.* John Wiley & Sons, New Jersey.

Magnussen, S. & Boudewyn, P., 1998. Derivations of stand heights from airborne laser scanner data with canopy-based quantile estimators. *Canadian Journal of Forest Research*, 28(7):1016–1031.

Magnussen, S., Næsset, E. & Gobakken, T., 2010. Reliability of lidar derived predictors of forest inventory attributes: A case study with norway spruce. *Remote Sensing of Environment*, 114(4):700 – 712.

Maltamo, M., Eerikäinen, K., Pitkänen, J., Hyyppä, J. & Vehmas, M., 2004. Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sensing of Environment*, 90(3):319 – 330.

Maltamo, M., Eerikäinen, K., Packalén, P. & Hyyppä, J., 2006a. Estimation of stem volume using laser scanning-based canopy height metrics. *Forestry*, 79(2):217–229.

Maltamo, M., Hyyppä, J. & Malinen, J., 2006b. A comparative study of the use of laser scanner data and field measurements in the prediction of crown height in boreal forests. *Scandinavian Journal of Forest Research*, 21(3):231–238.

Marklund, L.G., 1988. Biomass functions for pine, spruce and birch in sweden. *Rapport-Sveriges Lantbruksuniversitet, Institutionen foer Skogstaxering (Sweden).*

McGaughey, R.J., 2013. *FUSION/LDV: Software for LiDAR Data Analysis and Visualization.* US Forest Service. Version 3.30.

McRoberts, R.E., 2011. Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, 115(2):715 – 724.

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

Means, J.E., Acker, S.A., Harding, D.J., Blair, J.B., Lefsky, M.A., Cohen, W.B., Harmon, M.E. & McKee, W.A., 1999. Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the western cascades of oregon. *Remote Sensing of Environment*, 67(3):298–308.

Mevik, B.H., Wehrens, R. & Liland, K.H., 2013. *pls: Partial Least Squares and Principal Component regression*. R package version 2.4-3.

Miller, A., 2002. *Subset selection in regression*. CRC Press.

Miura, N. & Jones, S.D., 2010. Characterizing forest ecological structure using pulse types and heights of airborne laser scanning. *Remote Sensing of Environment*, 114(5):1069 – 1076.

Moeur, M. & Stage, A.R., 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science*, 41(2):337–359.

Naesset, E., 1997. Determination of mean tree height of forest stands using airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 52(2):49–56.

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

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. & Bjerknes, K.O., 2001. Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. *Remote Sensing of Environment*, 78(3):328–340.

Næsset, E., Bollandsås, O.M., Gobakken, T., Gregoire, T.G. & Ståhl, G., 2013a. Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne lidar: A case study with post-stratification to provide "activity data". *Remote Sensing of Environment*, 128(0):299 – 314.

Næsset, E., Gobakken, T., Bollandsås, O.M., Gregoire, T.G., Nelson, R. & Ståhl, G., 2013b. Comparison of precision of biomass estimates in regional field sample surveys and airborne lidar-assisted surveys in hedmark county, norway. *Remote Sensing of Environment*, 130(0):108 – 120.

Natural Regions Committee, 2006. Natural regions and subregions of Alberta. Compiled by D.J. Downing and W.W. Pettapiece.

Nelson, R., 2013. How did we get here? An early history of forestry LiDAR. *Canadian Journal of Remote Sensing*, 39(S1):S1–S12.

Nilsson, M., 1996. Estimation of tree heights and stand volume using an airborne lidar system. *Remote Sensing of Environment*, 56(1):1 – 7.

Nord-Larsen, T. & Riis-Nielsen, T., 2010. Developing an airborne laser scanning dominant height model from a countrywide scanning survey and national forest inventory data. *Scandinavian journal of forest research*, 25(3):262–272.

Nord-Larsen, T. & Schumacher, J., 2012. Estimation of forest resources from a country wide laser scanning survey and national forest inventory data. *Remote Sensing of Environment*, 119(0):148 – 157.

Nyström, M., Holmgren, J. & Olsson, H., 2012. Prediction of tree biomass in the forest–tundra ecotone using airborne laser scanning. *Remote Sensing of Environment*, 123(0):271 – 279.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. & Kauermann, G., 2007. Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102(478):400–409.

Patenaude, G., Hill, R., Milne, R., Gaveau, D., Briggs, B. & Dawson, T., 2004. Quantifying forest above ground carbon content using lidar remote sensing. *Remote Sensing of Environment*, 93(3):368 – 380.

Penner, M., Pitt, D. & Woods, M., 2013. Parametric vs. nonparametric lidar models for operational forest inventory in boreal ontario. *Canadian Journal of Remote Sensing*, 39(05):426–443.

Popescu, S.C., Zhao, K., Neuenschwander, A. & Lin, C., 2011. Satellite lidar vs. small footprint airborne lidar: Comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level. *Remote Sensing of Environment*, 115(11):2786 – 2797.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rana, P., Tokola, T., Korhonen, L., Xu, Q., Kumpula, T., Vihervaara, P. & Mononen, L., 2014. Training area concept in a two-phase biomass inventory using airborne laser scanning and rapideye satellite data. *Remote Sensing*, 6(1):285–309.

Rao, C.R., 1973. *Linear statistical inference and its applications*. John Wiley & Sons.

Royall, R.M., 1992. The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series*, pages 225–240.

Salas, C., Ene, L., Gregoire, T.G., Næsset, E. & Gobakken, T., 2010. Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models. *Remote Sensing of Environment*, 114(6):1277 – 1285.

Särndal, C.E., 2010. Models in survey sampling. In Carlson, M., Nyquist, H. and Villani, M. (eds), Official Statistics–Methodology and Applications in Honour of Daniel Thorburn.

Särndal, C.E., Swensson, B. & Wretman, J., 1992. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer, New York, second edition.

Schwarz, G., 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Selvin, H.C. & Stuart, A., 1966. Data-dredging procedures in survey analysis. *The American Statistician*, 20(3):20–23.

Sherrill, K.R., Lefsky, M.A., Bradford, J.B. & Ryan, M.G., 2008. Forest structure estimation and pattern exploration from discrete-return lidar in subalpine forests of the central rockies. *Canadian Journal of Forest Research*, 38(8):2081–2096.

Silva, P.N. & Skinner, C.J., 1997. Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1):23–32.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Skowronski, N.S., Clark, K.L., Gallagher, M., Birdsey, R.A. & Hom, J.L., in press. Airborne laser scanner-assisted estimation of aboveground biomass change in a temperate oak–pine forest. *Remote Sensing of Environment*.

Stephens, P.R., Kimberley, M.O., Beets, P.N., Paul, T.S., Searles, N., Bell, A., Brack, C. & Broadley, J., 2012. Airborne scanning lidar in a double sampling forest carbon inventory. *Remote Sensing of Environment*, 117(0):348 – 357.

Strunk, J., Temesgen, H., Andersen, H.E., Flewelling, J.P. & Madsen, L., 2012. Effects of lidar pulse density and sample size on a model-assisted approach to estimate forest inventory variables. *Canadian Journal of Remote Sensing*, 38(05):644–654.

Strunk, J.L., Reutebuch, S.E., Andersen, H.E., Gould, P.J. & McGaughey, R.J., 2011. Model-assisted forest yield estimation with light detection and ranging. *Western Journal of Applied Forestry*, 27(2):53–59.

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

Tinkham, W.T., Smith, A.M., Hoffman, C., Hudak, A.T., Falkowski, M.J., Swanson, M.E. & Gessler, P.E., 2012. Investigating the influence of lidar ground surface errors on the utility of derived forest inventories. *Canadian Journal of Forest Research*, 42(3):413–422.

Tomter, S.M., Hylen, G. & Nilsen, J.E., 2010. National forest inventory reports: Norway. In *National forest inventories—Pathways for common reporting*, pages 411–424. Springer, Dordrecht.

Tonolli, S., Dalponte, M., Neteler, M., Rodeghiero, M., Vescovo, L. & Gianelle, D., 2011. Fusion of airborne lidar and satellite multispectral data for the estimation of timber volume in the southern alps. *Remote Sensing of Environment*, 115(10):2486 – 2498.

Ung, C.H., Bernier, P. & Guo, X.J., 2008. Canadian national biomass equations: new parameter estimates that include British Columbia data. *Canadian Journal of Forest Research*, 38(5):1123–1132.

Valliant, R., Dever, J.A. & Kreuter, F., 2013. *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

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

Vincent, G., Sabatier, D., Blanc, L., Chave, J., Weissenbacher, E., Pélissier, R., Fonty, E., Molino, J.F. & Couteron, P., 2012. Accuracy of small footprint airborne lidar in its predictions of tropical moist forest stand structure. *Remote Sensing of Environment*, 125(0):23 – 33.

Wang, Z. & Bellhouse, D.R., 2009. Semiparametric regression model for complex survey data. *Survey Methodology*, 35(2):247.

White, J., Wulder, M. & Frazer, G., 2013. A lidar-based structurally guided sampling design to augment hinton wood products' permanent growth sample (pgs) program and support predictive modelling of forest inventory attributes from lidar. Technical

report, Canadian Forest Service, Canadian Wood Fibre Centre, Pacific Forestry Centre, Victoria, BC, and Hinton Wood Products (A division of West Fraser Mills, Ltd.), Edmonton, AB.

Wulder, M., Coops, N., Hudak, A., Morsdorf, F., Nelson, R., Newnham, G. & Vastaranta, M., 2013. Status and prospects for lidar remote sensing of forested ecosystems. *Canadian Journal of Remote Sensing*, 39(s1):S1–S5.

Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W. & Gobakken, T., 2012. Lidar sampling for large-area forest characterization: A review. *Remote Sensing of Environment*, 121(0):196–209.

Zhao, F., Guo, Q. & Kelly, M., 2012. Allometric equation choice impacts lidar-based forest biomass estimates: A case study from the sierra national forest, CA. *Agricultural and Forest Meteorology*, 165(0):64 – 72.

Part V.

Annexes

A. Annex

A.1. Annex A

R code for the simulated populations described in Chapter 6.

```
## Population size
N <- 2000

## NOISE
NOISE <- data.frame(y = rnorm(N, 100, 10))
for(i in 2:21){
    NOISE[,i] <- rnorm(N, 100, 10)
    names(NOISE)[i] <- paste("x", i - 1, sep = "")
}

## COR
COR <- data.frame(y = rnorm(N, 100, 10))
for(i in 2:21){
    COR[,i] <- rnorm(N, COR$y, 3.25)
    names(COR)[i] <- paste("x", i - 1, sep = "")
}

## DCOR
DCOR <- data.frame(y = rnorm(N, 100, 10))
for(i in 2:21){
    DCOR[,i] <- rnorm(N, DCOR$y, i^2)
    names(DCOR)[i] <- paste("x", i - 1, sep = "")
}
```

A.2. Annex B

Figures for the simulation studies from Chapter 8 and 9.

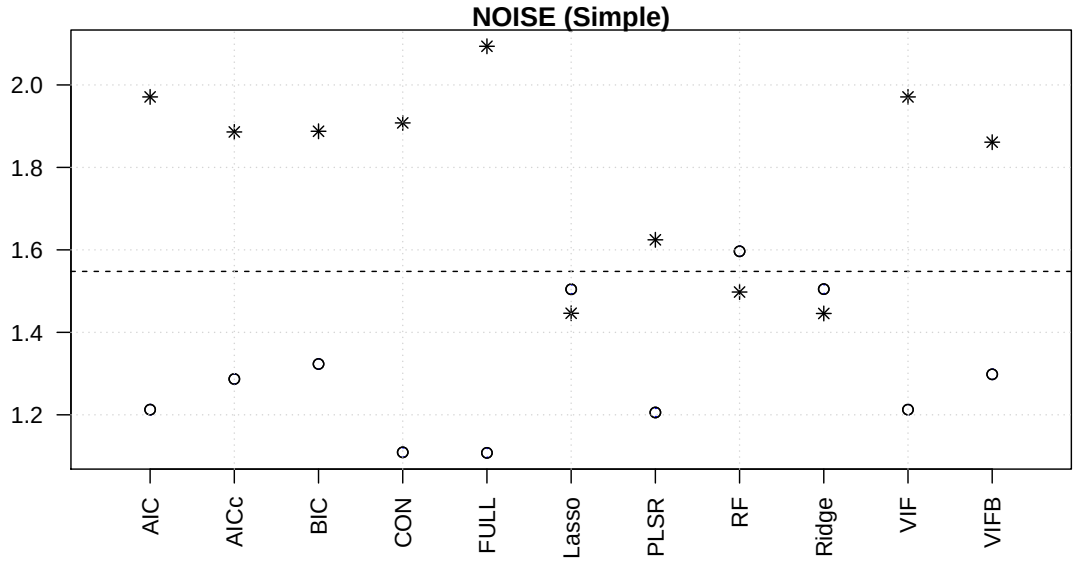

Figure A.1.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; over 2,000 iterations; circles) and empirical standard error (stars) for the dataset NOISE. The dashed vertical line depicts the parametric standard error of SI.

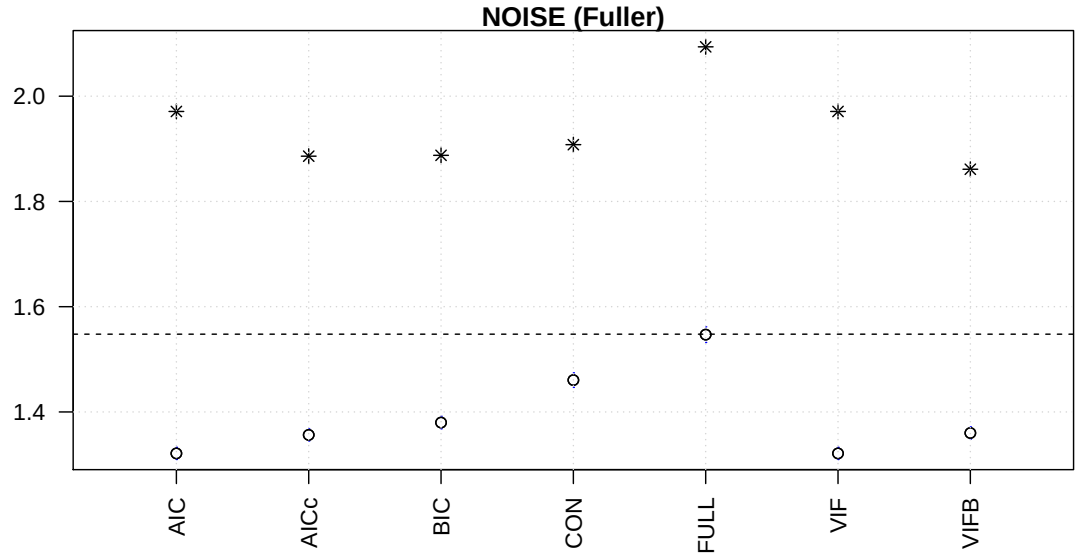

Figure A.2.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset NOISE. The dashed vertical line depicts the parametric standard error of SI.

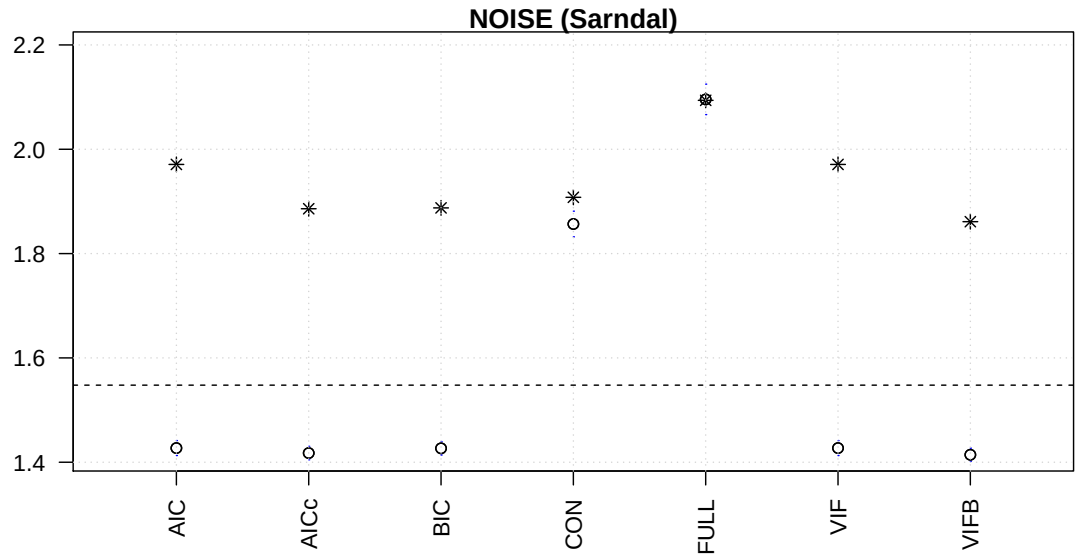

Figure A.3.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset NOISE. The dashed vertical line depicts the parametric standard error of SI.

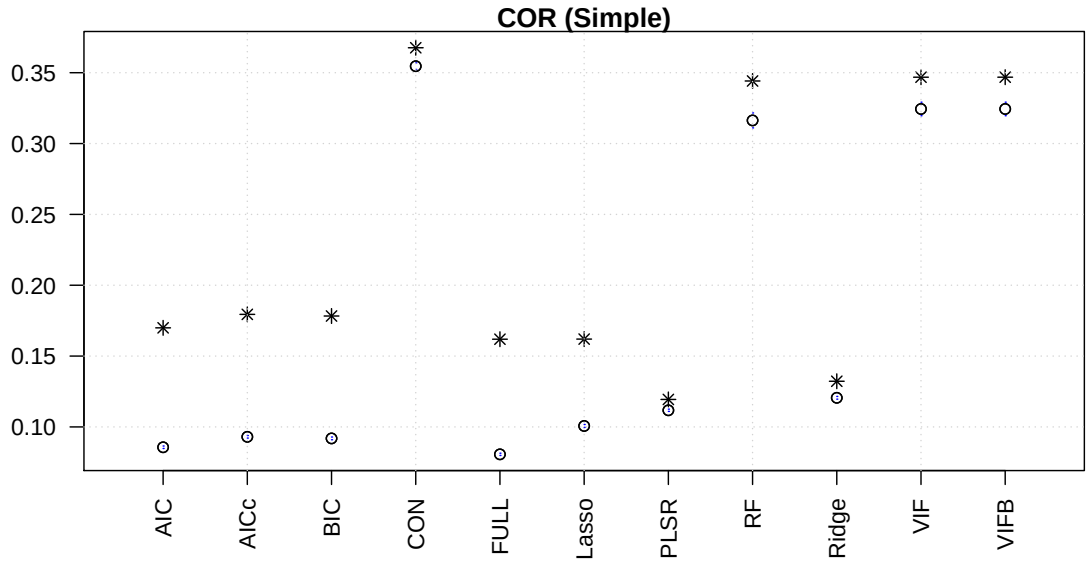

Figure A.4.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset COR.

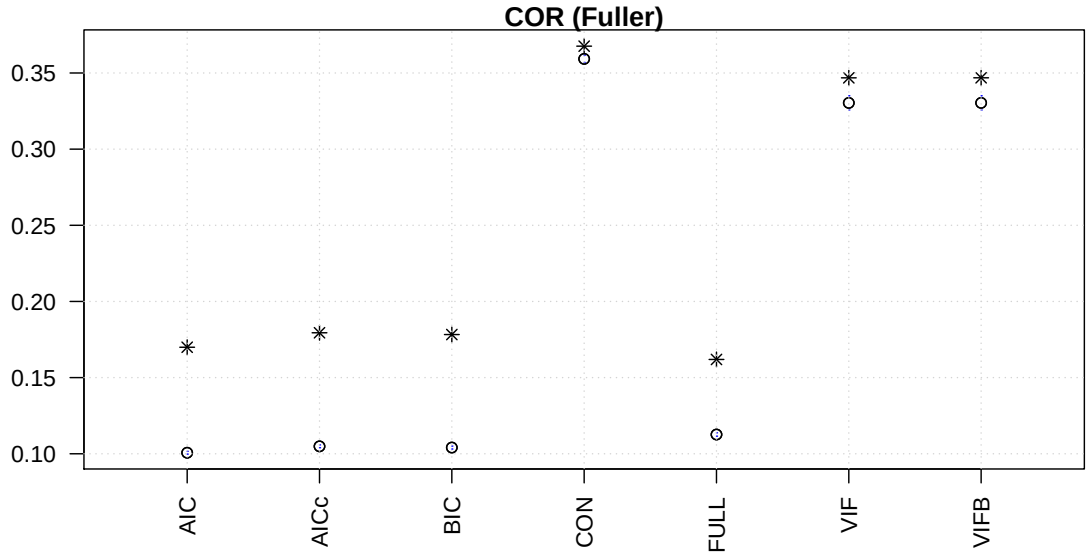

Figure A.5.: Simple variance estimator ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset COR.

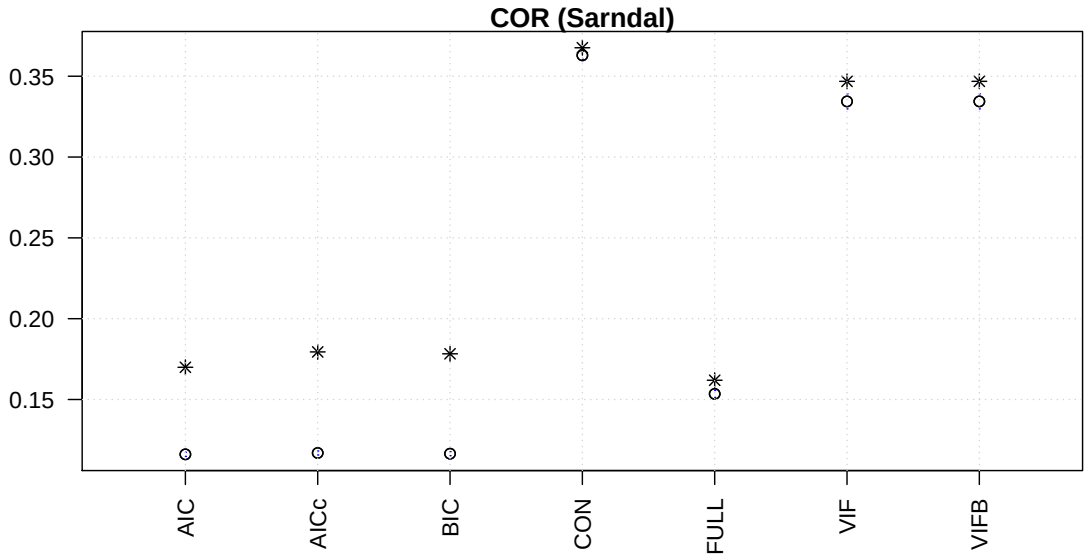

Figure A.6.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset COR.

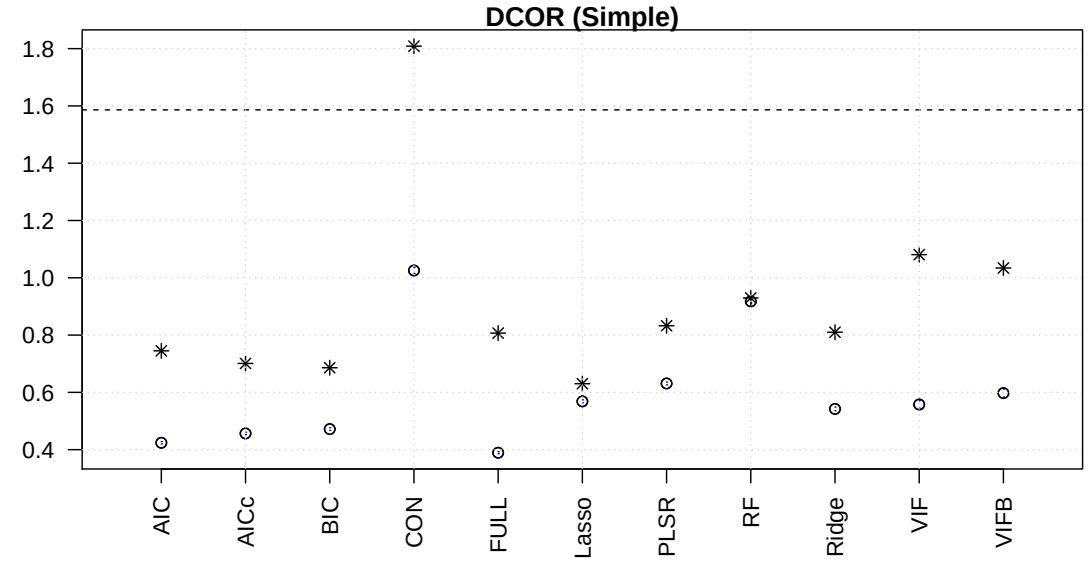

Figure A.7.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset DCOR.

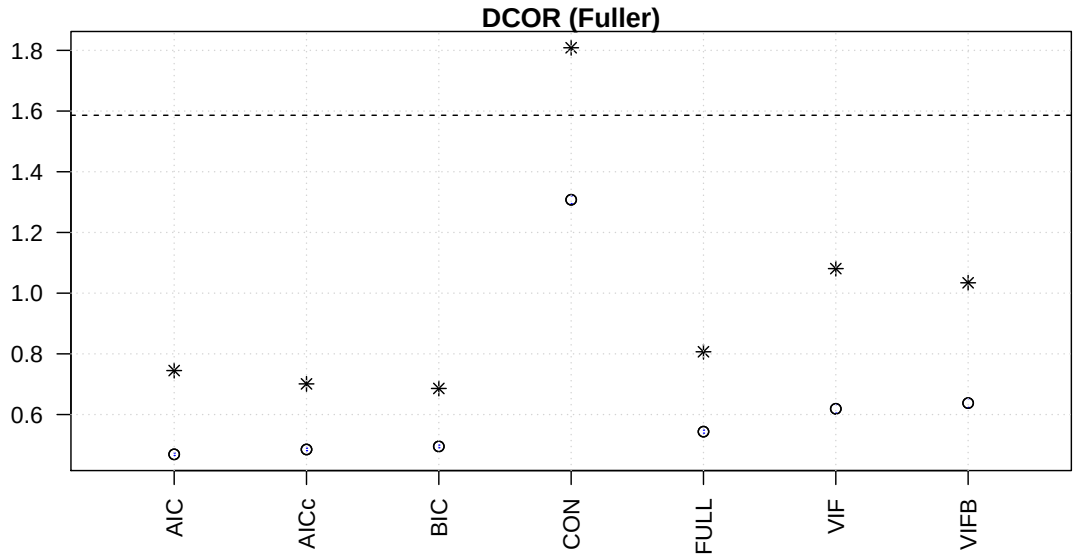

Figure A.8.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset DCOR.

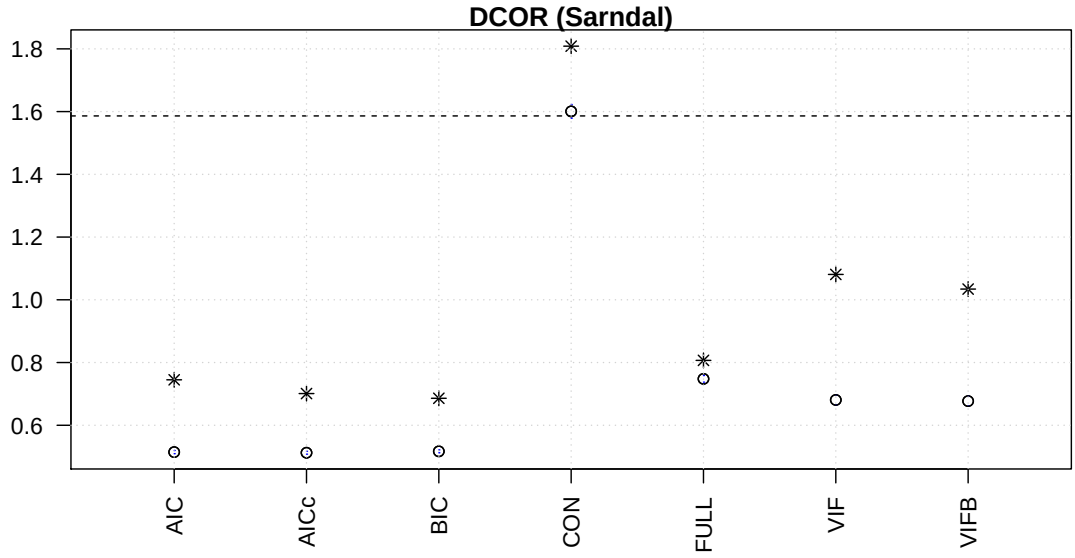

Figure A.9.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset DCOR.

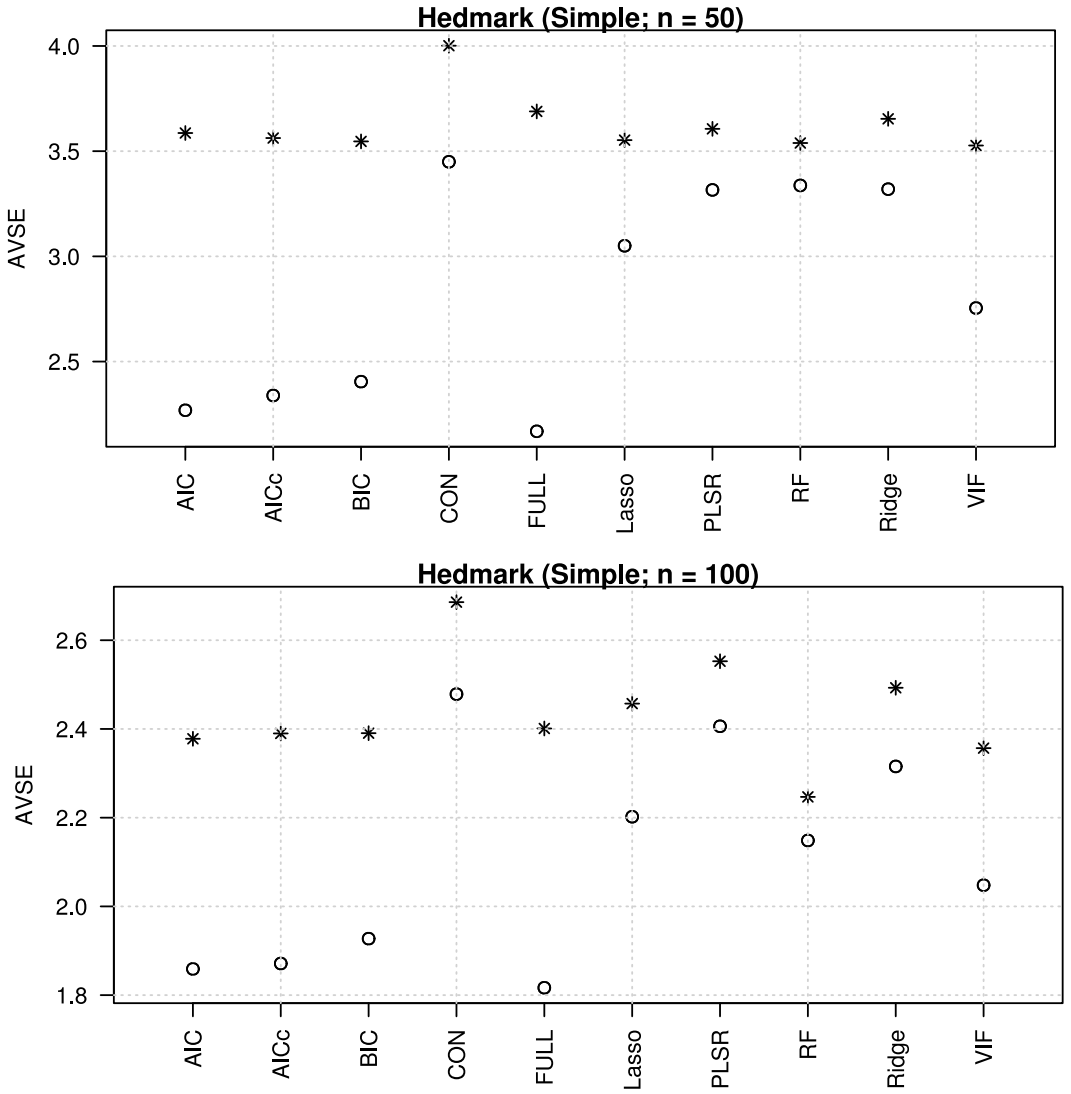

Figure A.10.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 50$, bottom: $n = 100$).

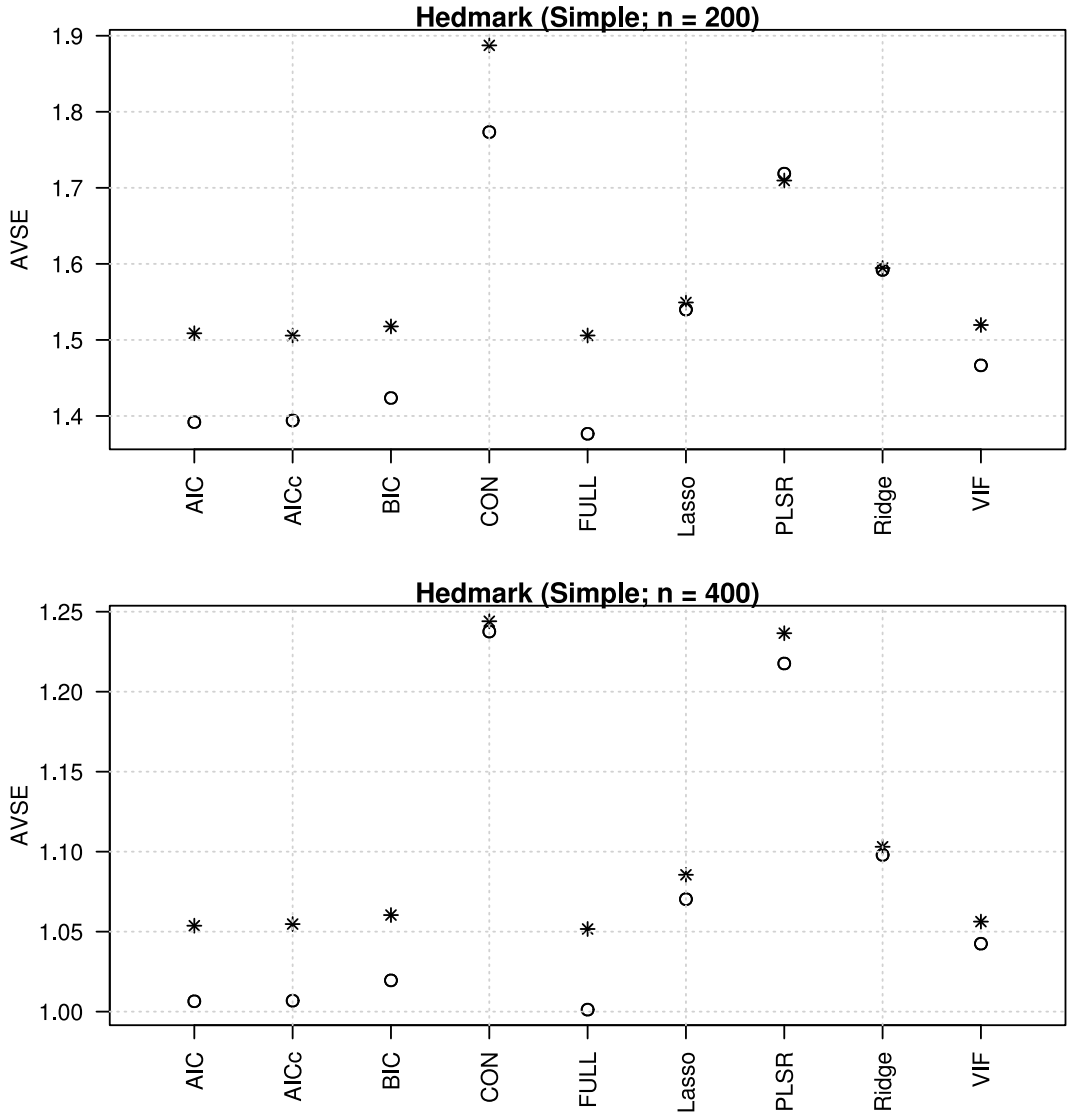

Figure A.11.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 200$, bottom: $n = 400$).

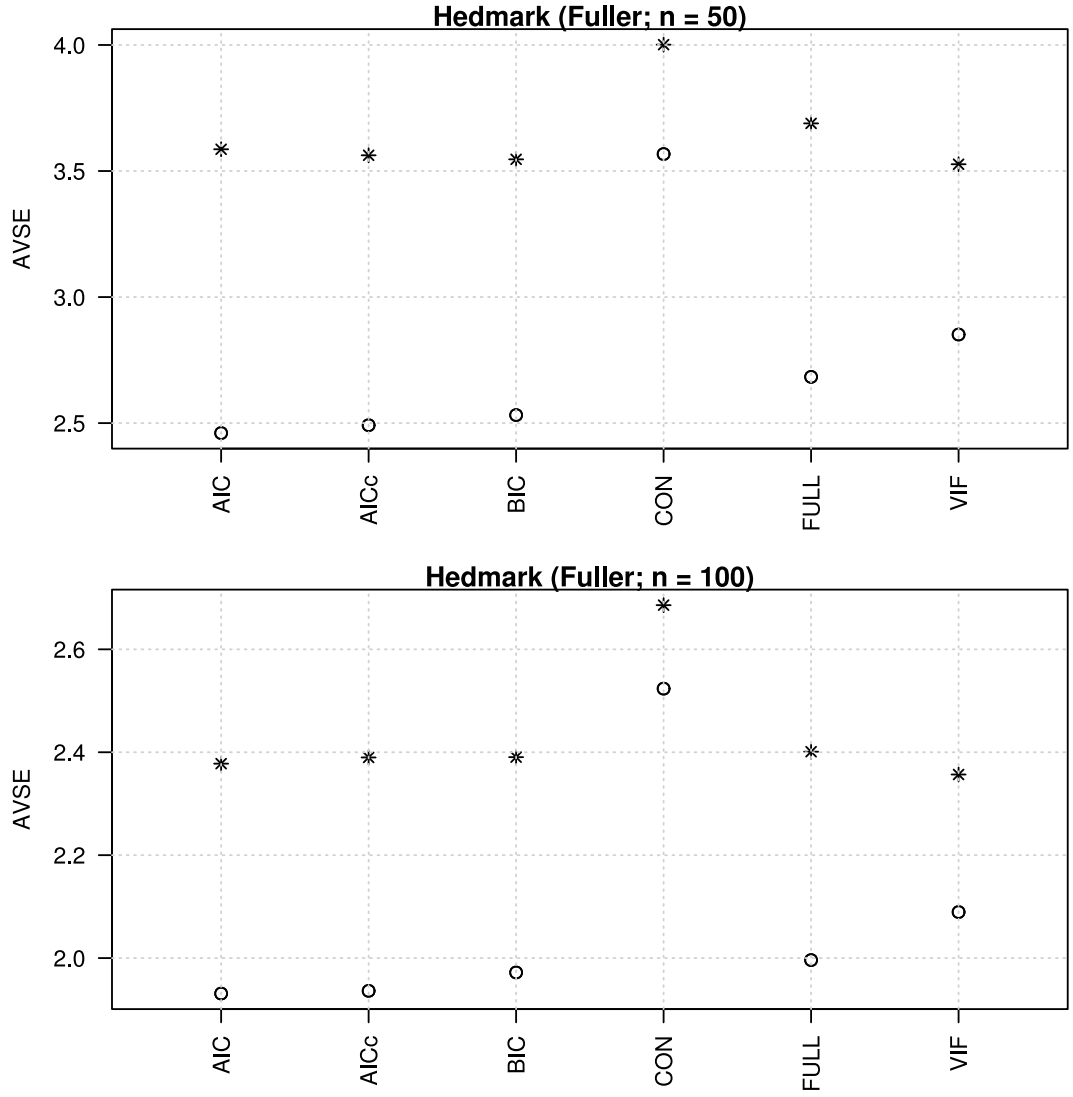

Figure A.12.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 50$, bottom: $n = 100$).

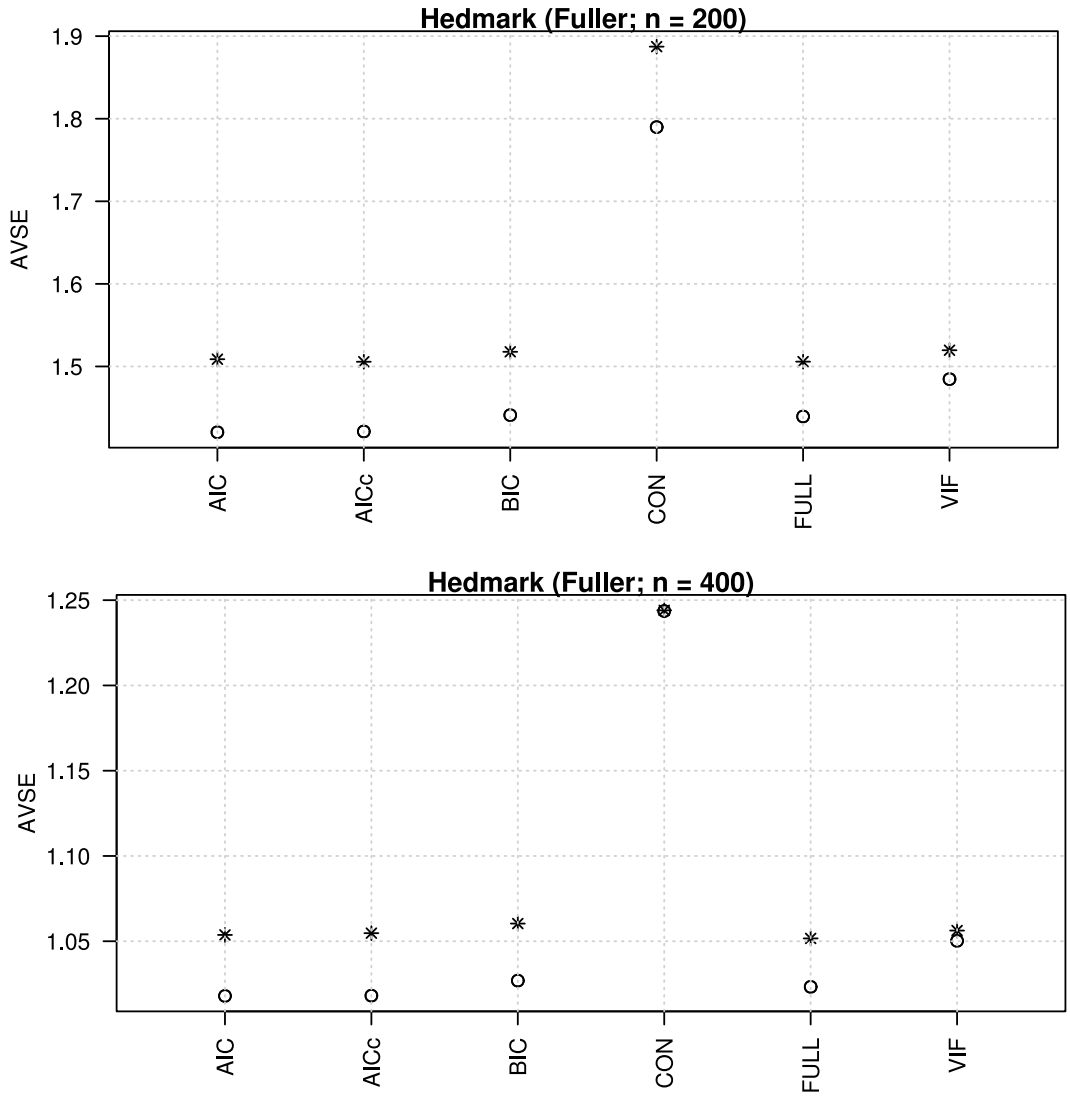

Figure A.13.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 200$, bottom: $n = 400$).

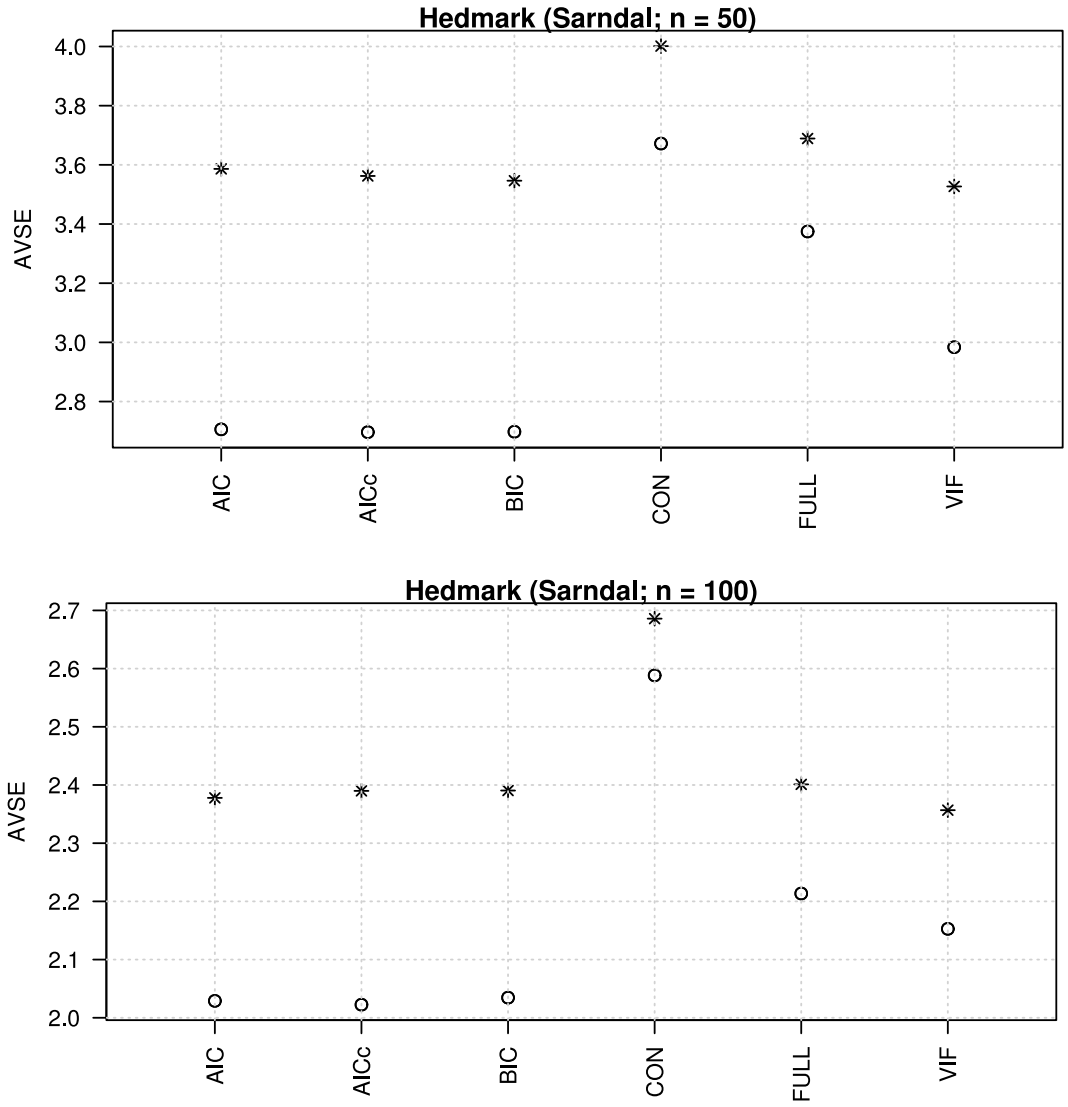

Figure A.14.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 50$, bottom: $n = 100$).

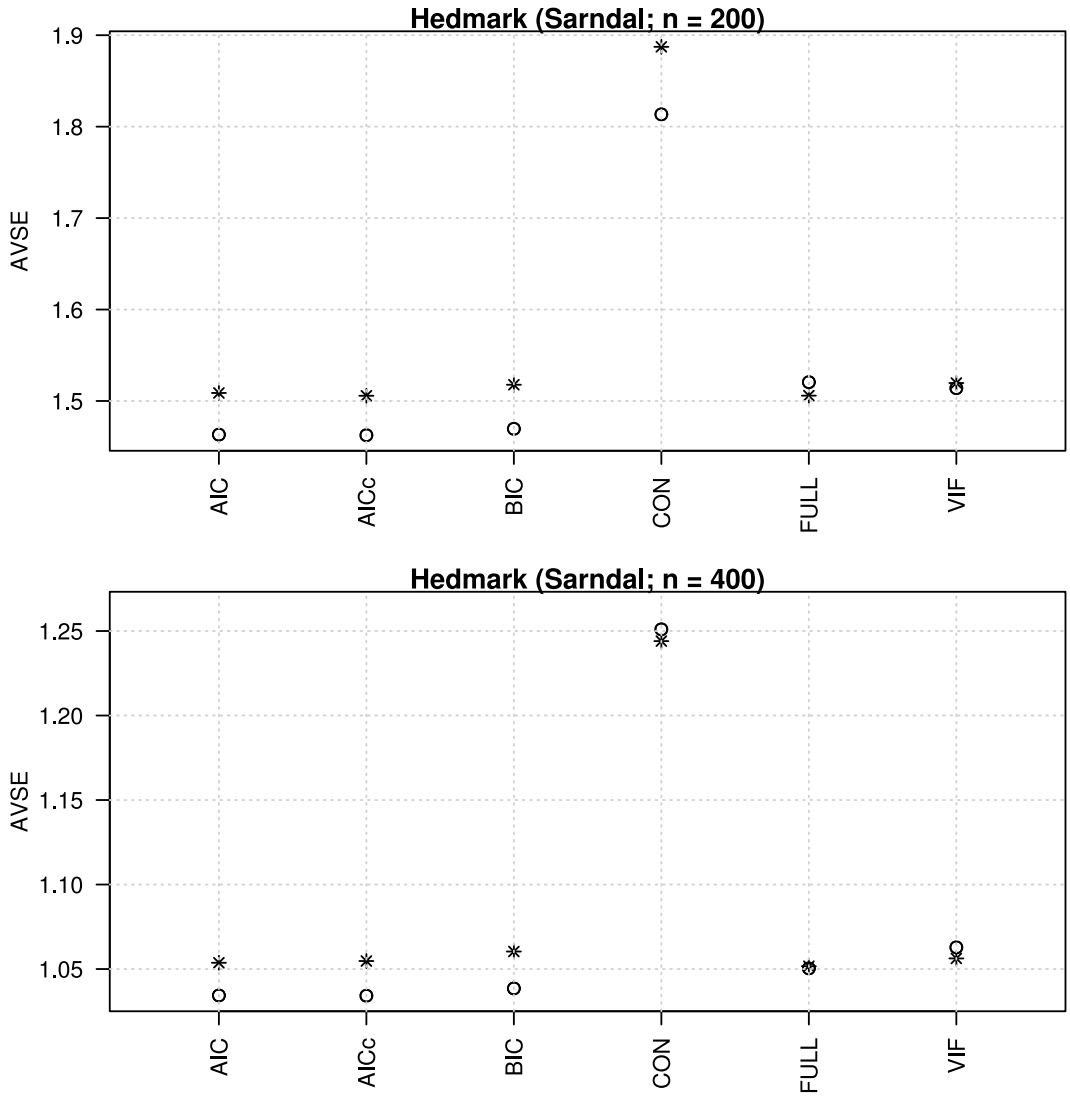

Figure A.15.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark (top: $n = 200$, bottom: $n = 400$).

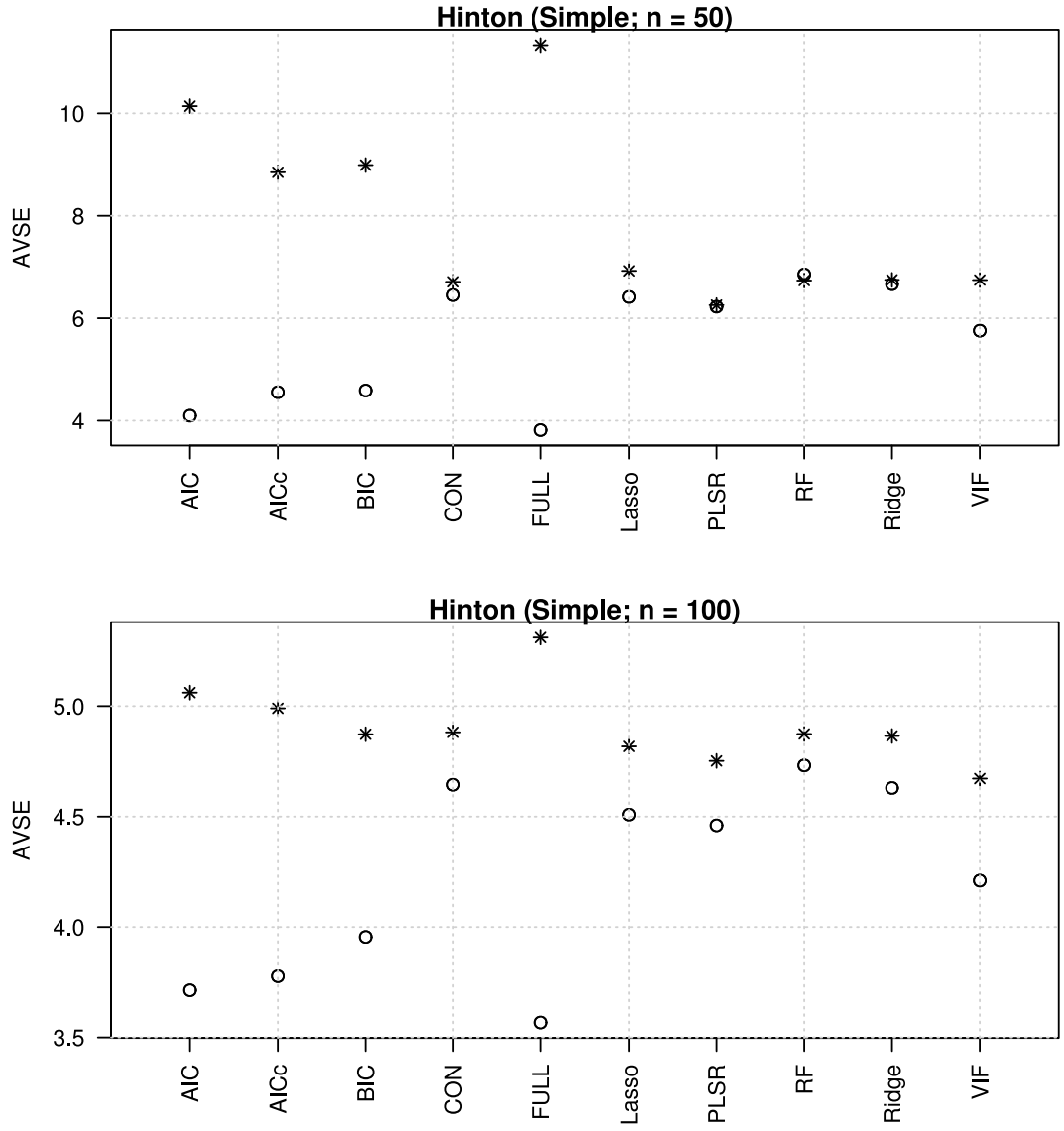

Figure A.16.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 50$, bottom: $n = 100$).

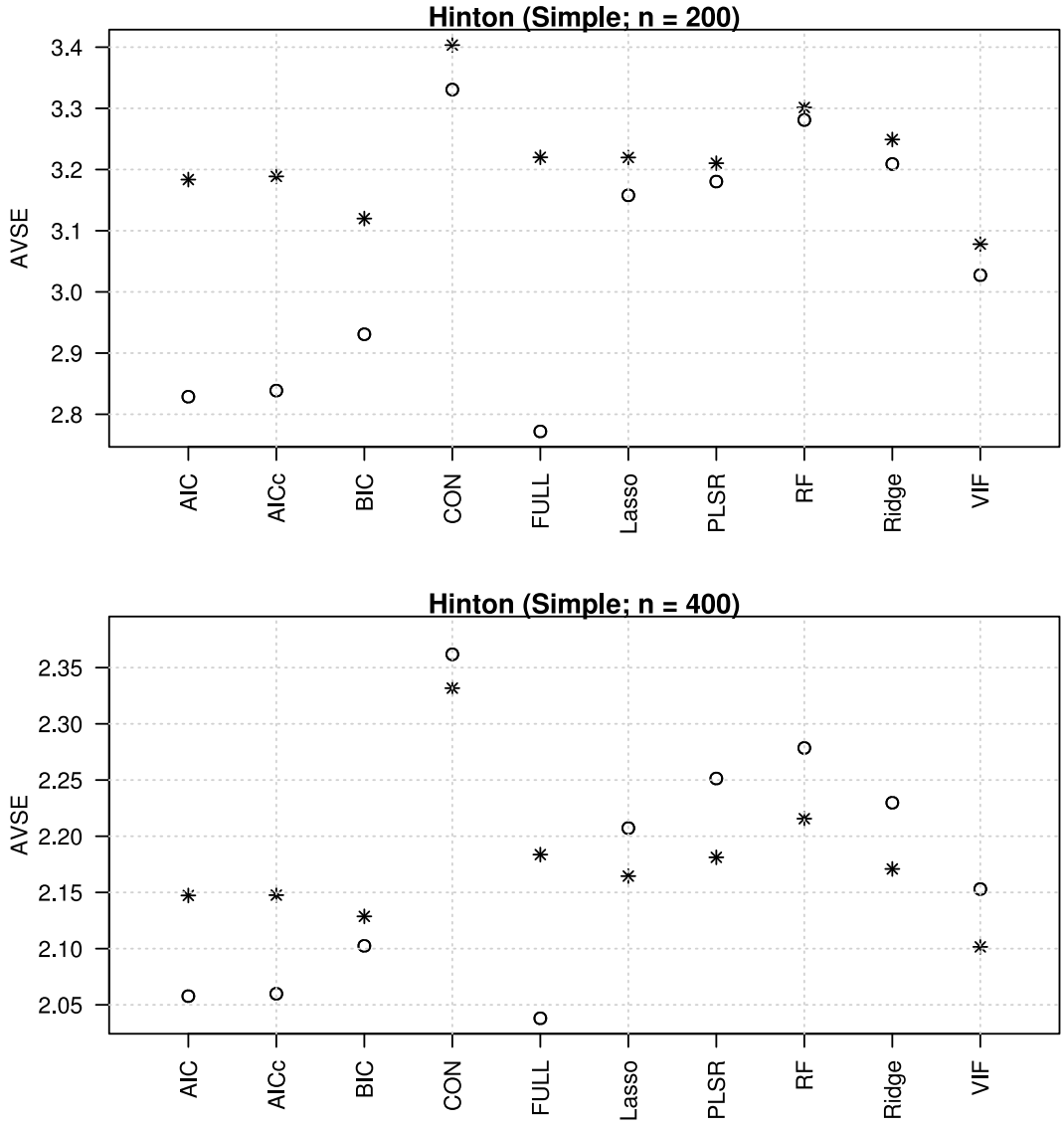

Figure A.17.: Simple variance estimator ($\hat{V}_{\text{Simple}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 200$, bottom: $n = 400$).

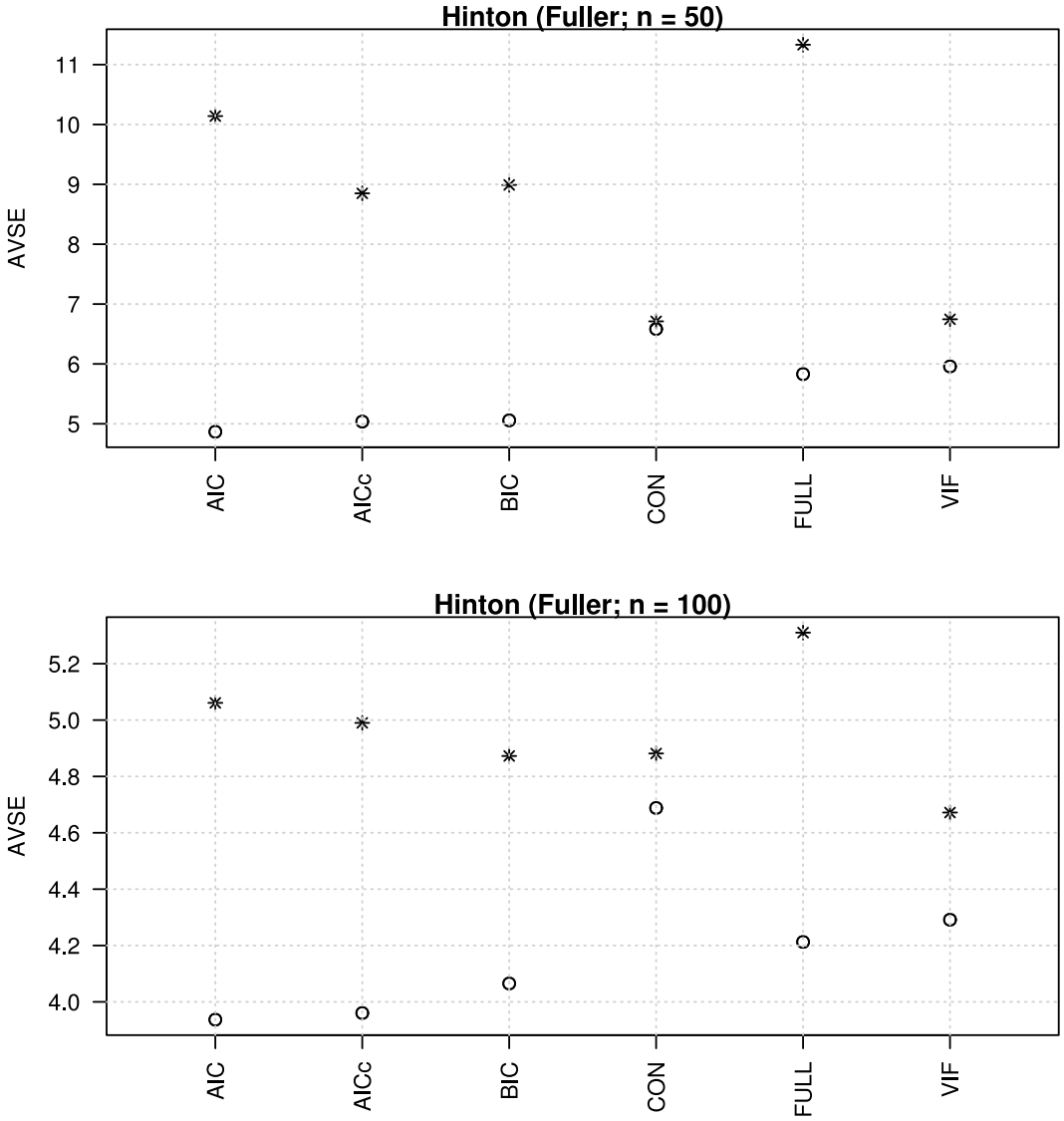

Figure A.18.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 50$, bottom: $n = 100$).

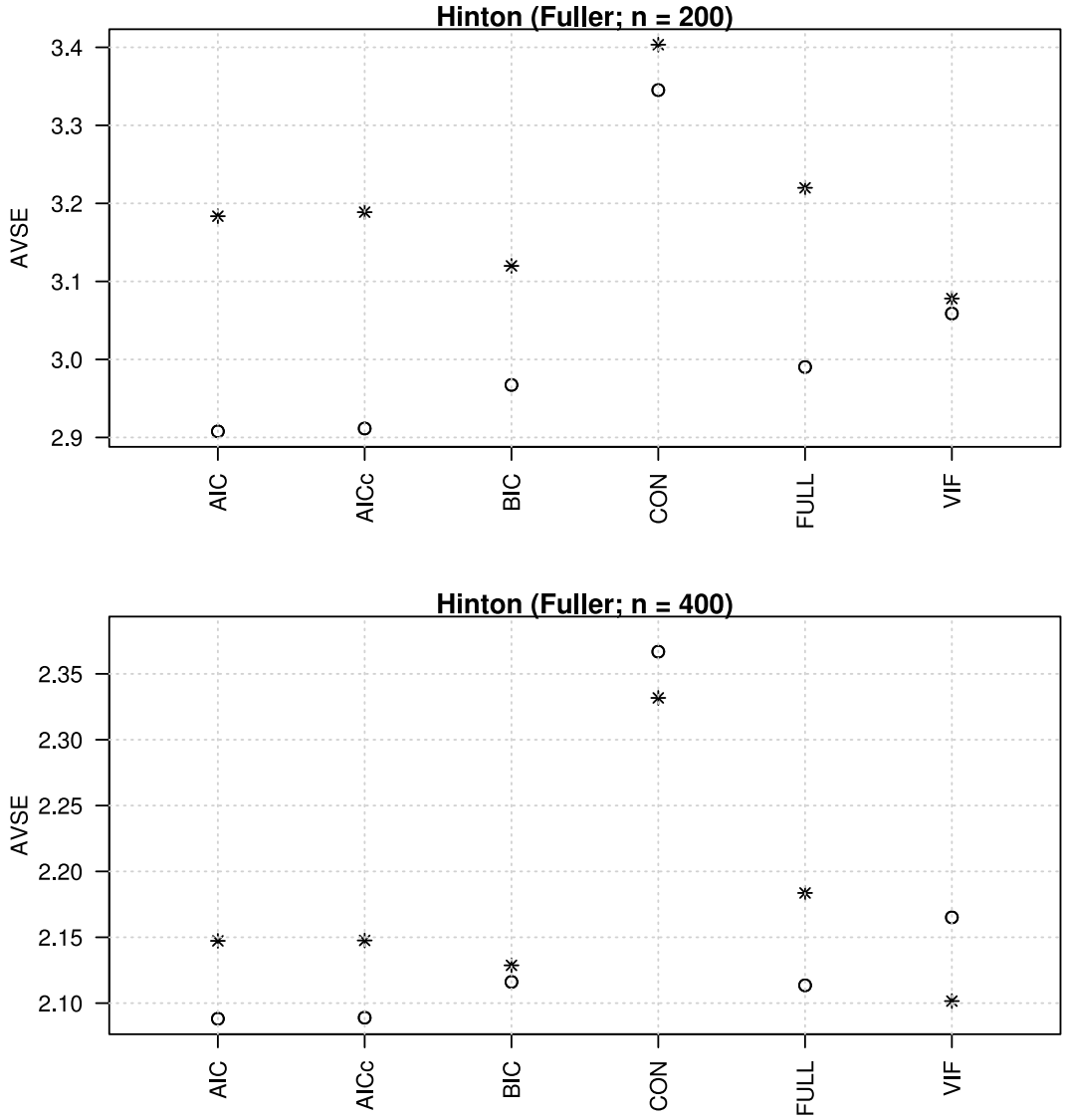

Figure A.19.: Variance estimator after Fuller ($\hat{V}_{\text{Fuller}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 200$, bottom: $n = 400$).

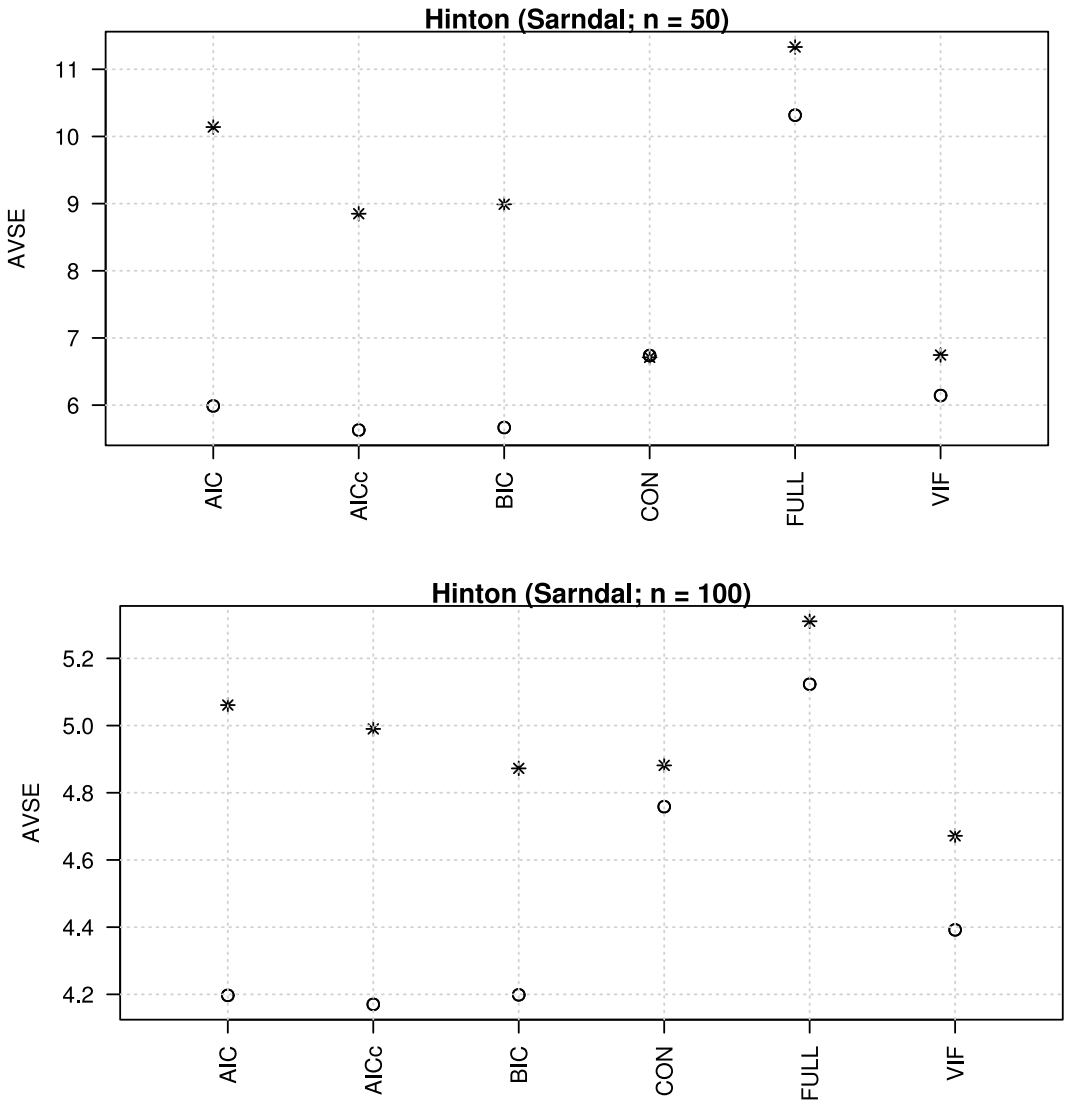

Figure A.20.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 50$, bottom: $n = 100$).

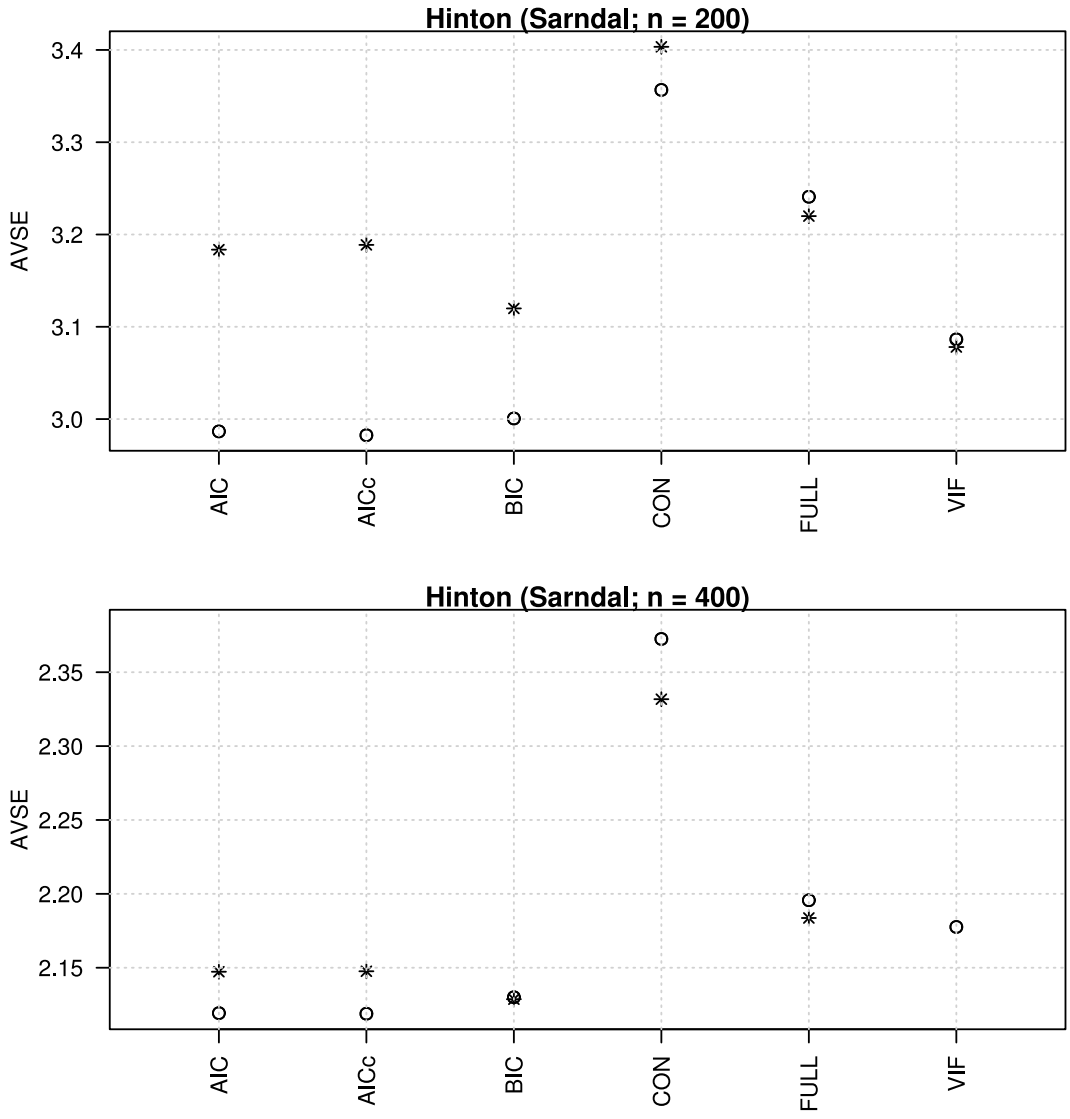

Figure A.21.: Variance estimator after Särndal ($\hat{V}_{\text{Sarndal}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton (top: $n = 200$, bottom: $n = 400$).

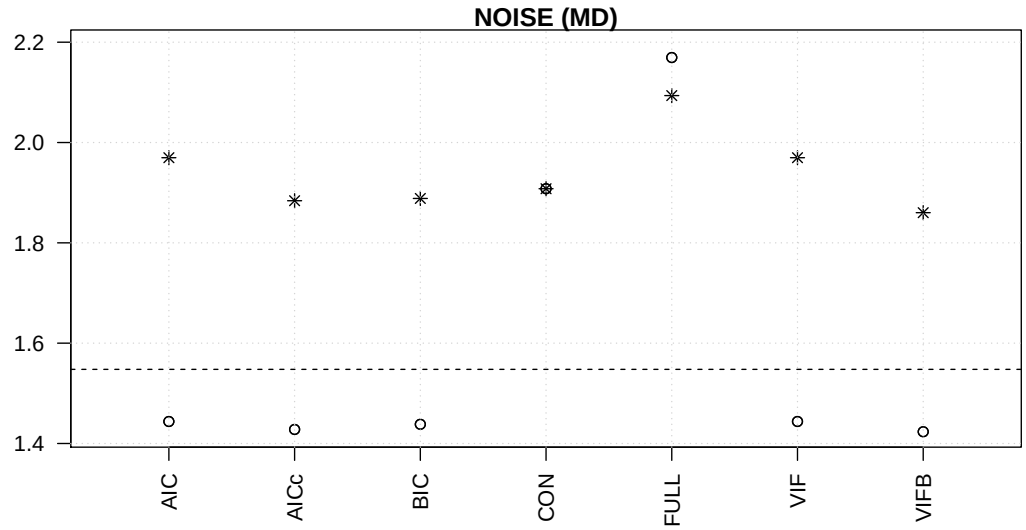

Figure A.22.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset NOISE.

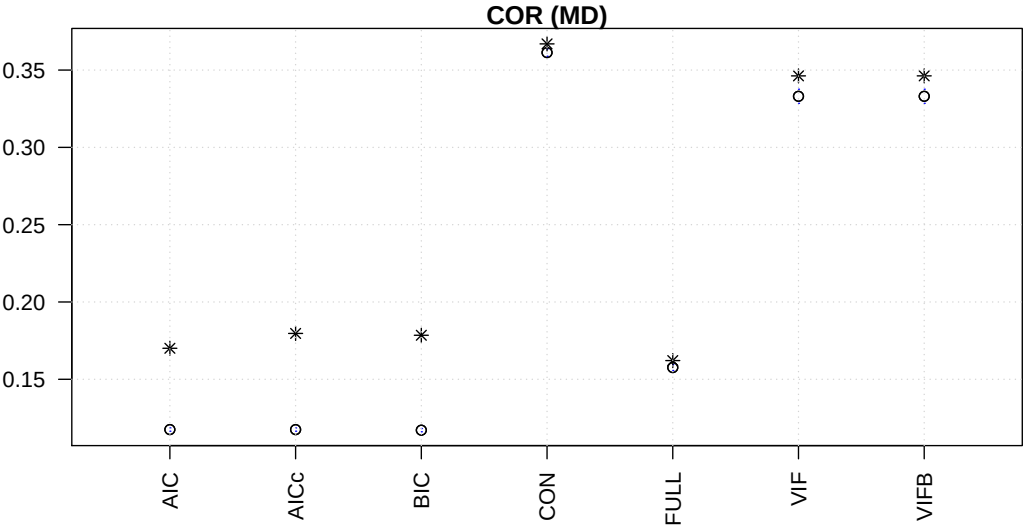

Figure A.23.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset COR.

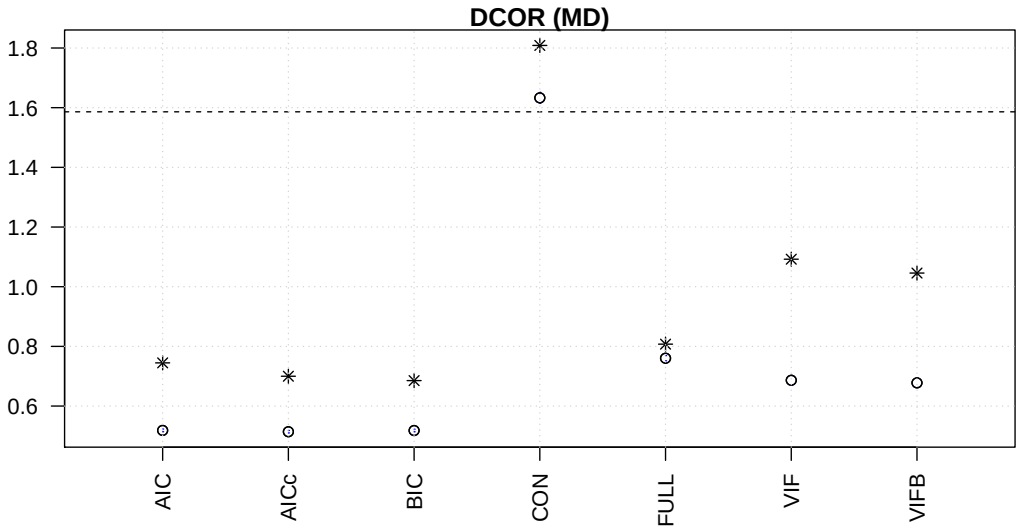

Figure A.24.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset DCOR.

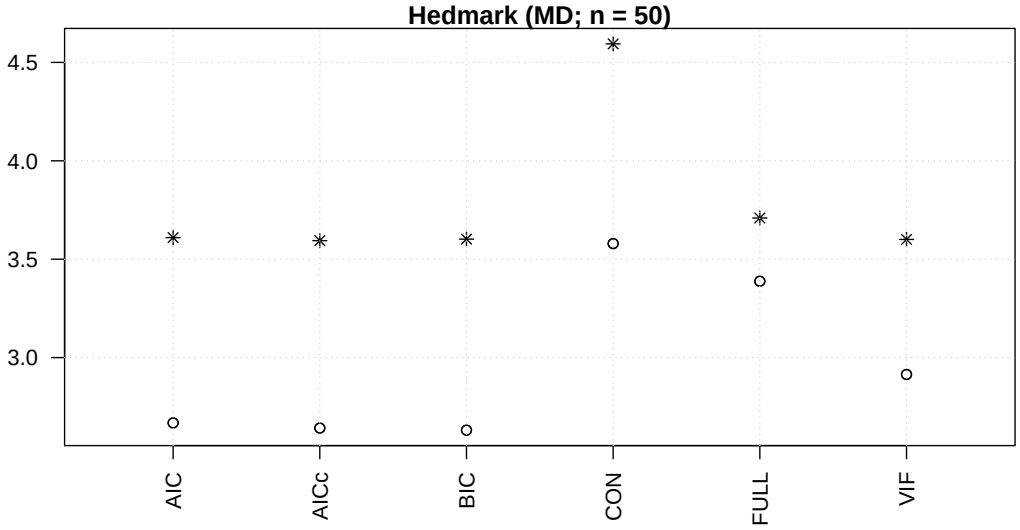

Figure A.25.: Variance estimator $\hat{V}_{\mathrm{MD}}$); mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark ($n = 50$).

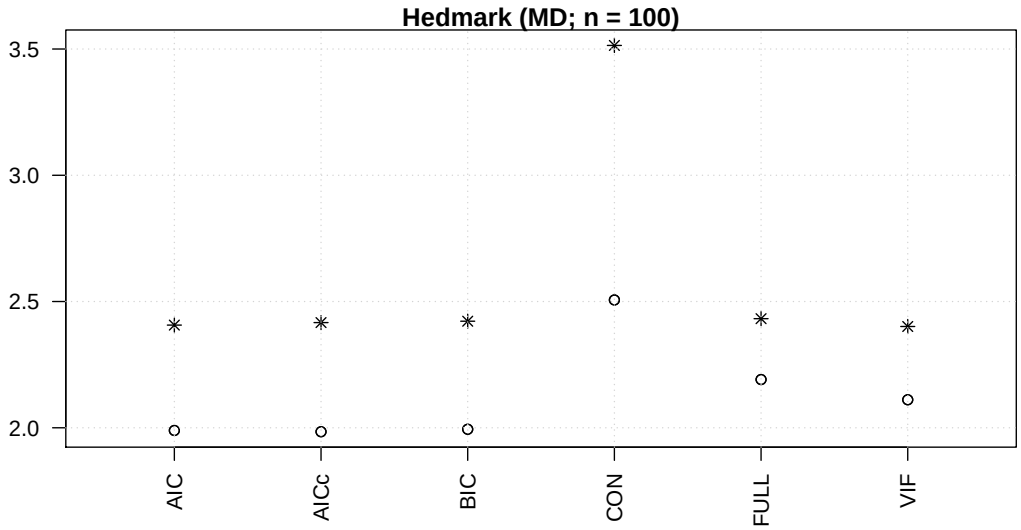

Figure A.26.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark ($n = 100$).

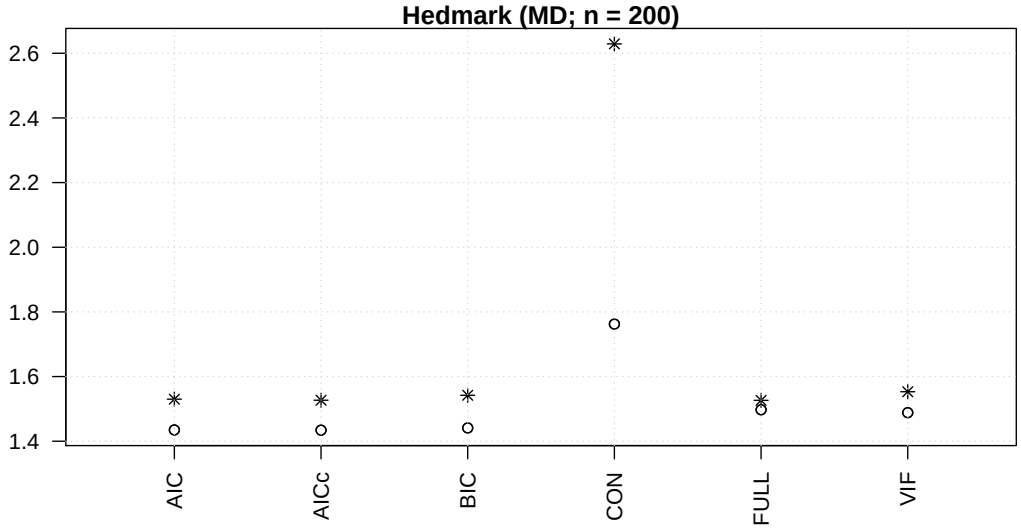

Figure A.27.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark ($n = 200$).

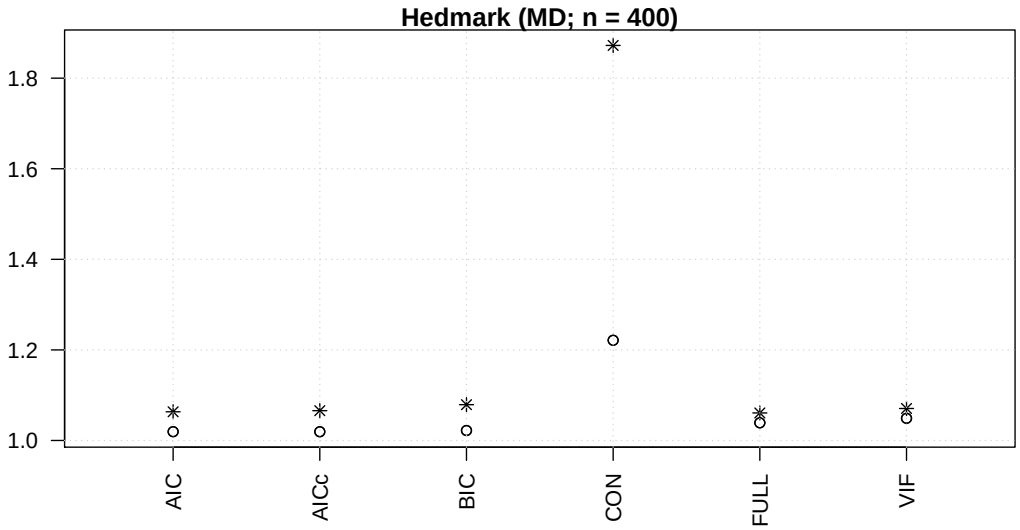

Figure A.28.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hedmark ($n = 400$).

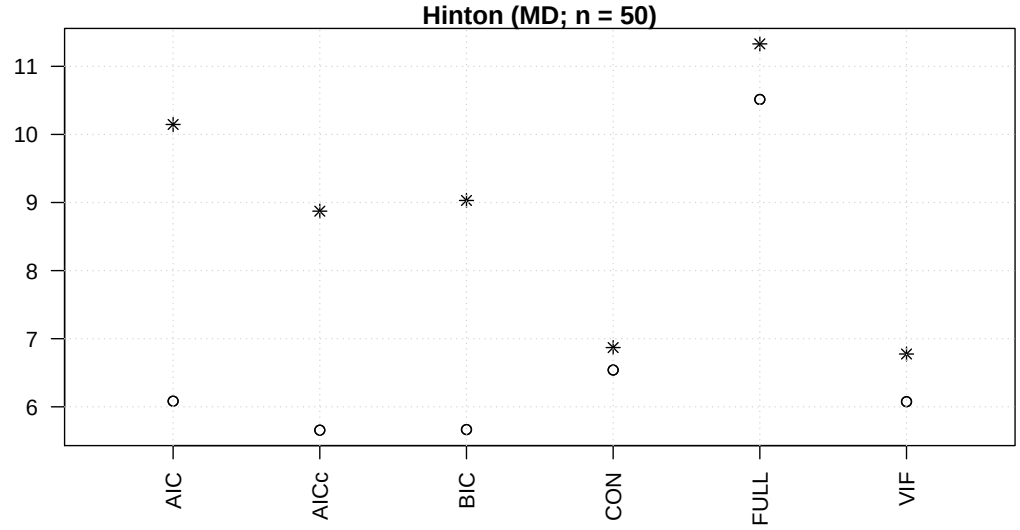

Figure A.29.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton ($n = 50$).

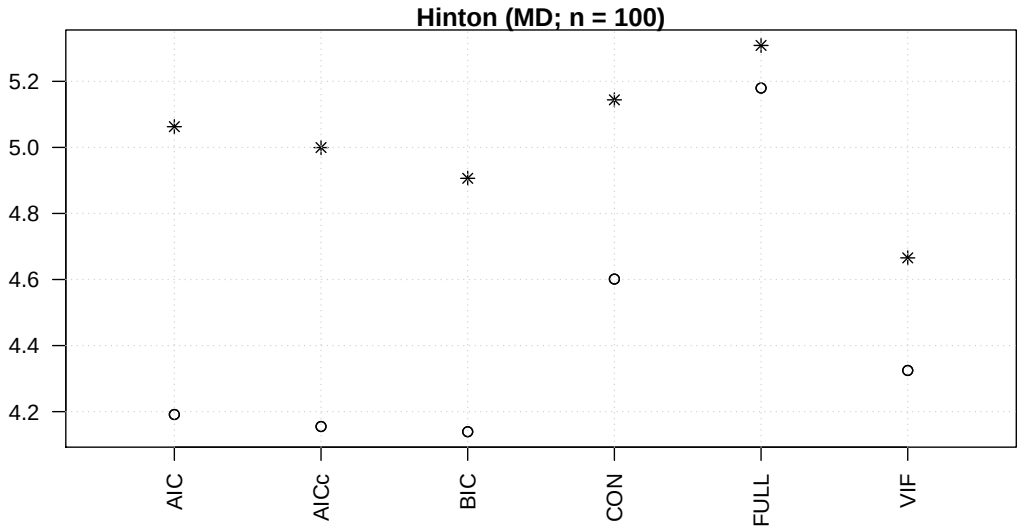

Figure A.30.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton ($n = 100$).

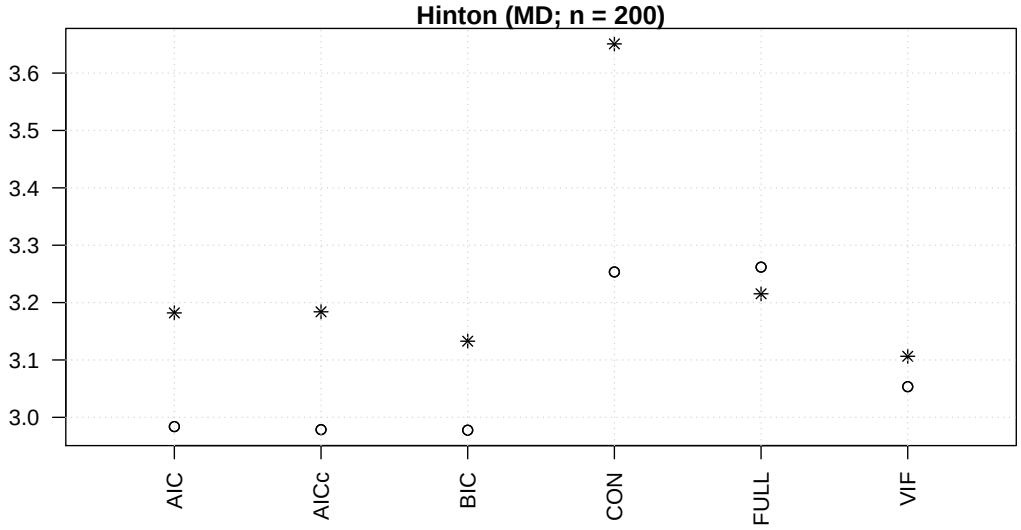

Figure A.31.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton ($n = 200$).

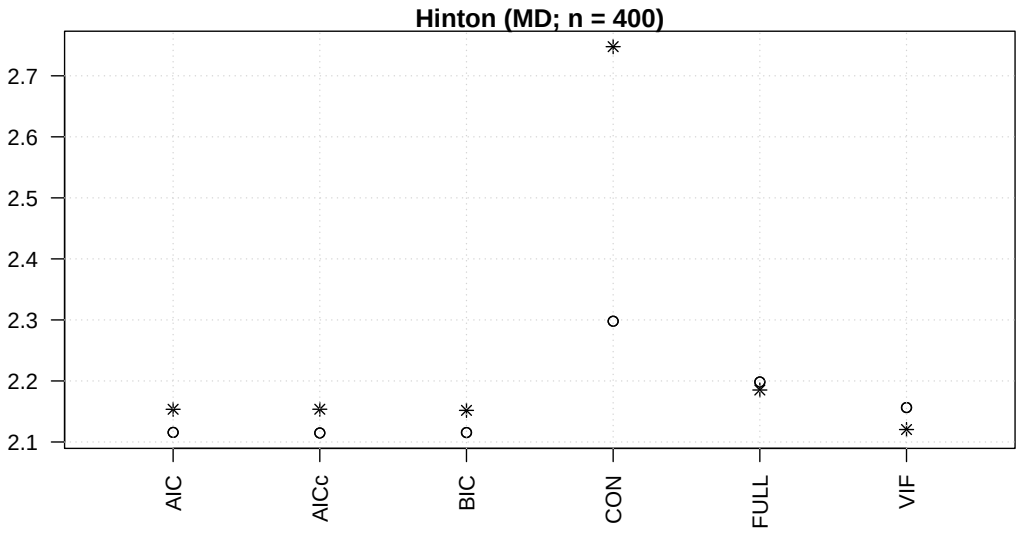

Figure A.32.: Variance estimator $\hat{V}_{\mathrm{MD}}$; mean estimated standard error (AVSE; circles) and empirical standard error (stars) for the dataset Hinton ($n = 400$).

Curriculum Vitae

Personal data

Name:	Philip Henrich Mundhenk
Date of birth:	August 20, 1979
Place of birth:	Hamburg

Education

Since October 2010	PhD student at the Chair of Forest Inventory and Remote Sensing and within the Research Training Group "Scaling problems in statistics"
March 2006 - June 2008:	Master of Science in "Tropical and International Forestry" Georg–August–University of Göttingen
October 2001 - October 2005	Bachelor of Science in "Forest Sciences and Forest Ecology" Georg–August–University of Göttingen