Multiphase estimation procedures for forest inventories under the design-based Monte Carlo approach

Author(s):
Massey, Alexander F.

Publication Date:
2015

Permanent Link:

https://doi.org/10.3929/ethz-a-010536381 →

Diss. ETH No. 23025

Multiphase estimation procedures for forest inventories under the design-based Monte Carlo approach

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

ALEXANDER FRANCIS MASSEY
MSc. Statistics, ETH Zurich
born 18 May 1983
citizen of the United States of America and Italy

accepted on the recommendation of

PD Dr. D. Mandallaz, supervisor, examiner
Prof. Dr. H. R. Heinimann, co-examiner
Prof. Dr. J. Saborowski, co-examiner
Dr. A. Lanz, co-examiner

2015

Acknowledgements

First of all, I would like to express my gratitude to my supervisor PD Dr. Daniel Mandallaz. He has been incredibly reliable in his eagerness to provide face-to-face assistance. I greatly appreciate the opportunity to work closely with him and to benefit from his vast statistical experience. I would also like to thank Prof. Dr. Hans Rudolf Heinimann for his guidance and making it possible to continue this work with the Chair of Land Use Engineering at the ETH Zurich.

Special thanks go to the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) for financing the majority of this project. In particular, I would give many thanks to Dr. Adrian Lanz of the Forest Resources and Management group for his professional oversight and expert guidance in the field of forestry and Dr. Rita Ghosh of the Statistics Lab group for her glowing enthusiasm concerning the discussion of statistical methods. I would also like thank Christian Ginzler for his invaluable assistance with processing the stereophotography data.

In addition, I would like to thank the final doctoral examiner Prof. Dr. Joachim Saborowski for his helpful review and the evaluation of this dissertation.

Further thanks go to all the administrative staff at WSL and in particular, Maja Messerli, at ETH who helped me with many of my organizational concerns. Also, I thank my colleagues Christoph Fischer, Andreas Hill and Daniel Trüssel for their technical support using GIS and acting as a sounding board that helped clarify many of the ideas presented in this dissertation.

Contents

List of Figures

List of Tables

Summary

This cumulative dissertation explores methods for improving estimation for forest inventories by incorporating auxiliary data such as those derived from remote sensing. Results are compiled from three articles published in the *Canadian Journal of Forest Research* (see Massey et al. (2014); Massey and Mandallaz (2015a,b)). The proposed methods follow exclusively the design-based Monte Carlo setup with a focus on two- and three-phase regression estimators. The structure of this dissertation can be divided into four main parts: 1) the presentation of the general model-assisted framework in the Monte Carlo approach, 2) an application example for the estimation of state in the Swiss National Forest Inventory (NFI), 3) an extension of regression estimation to nonparametric models such as k-nearest neighbors (kNN), and 4) a reformulation of the proposed approach for the estimation of net change.

Although the methods are applicable in a variety of situations, the majority of the case studies are presented in the context of annual sampling designs. This is because the practical objective of this research was to accommodate the 2009 transition of the Swiss NFI from a periodic to an annual design where only one ninth of the overall sample of permanent plots is measured every year. The transition was primarily motivated by logistical reasons and led to a reduction in sample size that consequently increased the variance when using the standard simple random sampling estimator.

The major result was for the estimation of state where it was demonstrated that a three-phase classical regression estimator could dramatically reduce the increase in variance associated with the transition to the annual design by integrating canopy height information with growth-updated previous measurement information. While kNN regression estimators could probably have also been used effectively in place of the classical version, the most common analytical variance estimator risks potentially severe underestimation of the true variance as is demonstrated in an artificial simulation example. A bootstrap procedure is proposed that avoids this underestimation. Similar improvements are also possible for the estimation of net change, but there is a tradeoff between the estimator's gain in precision and its flexibility to estimate change over any arbitrary time frame.

Zusammenfassung

Die vorliegende kumulative Dissertation befasst sich mit der Verbesserung von Schätzmethoden in der Waldinventur durch die Einbeziehung von Hilfsinformation, u.a. Fernerkundungsdaten. Die Ergebnisse der Arbeit wurden in drei Artikeln im *Canadian Journal of Forest Research* publiziert (Massey et al. (2014); Massey and Mandallaz (2015a,b)). Die vorgeschlagenen zwei- und dreiphasigen Schätzmethoden folgen dabei ausschliesslich dem design-basierten Ansatz. Der Aufbau der Dissertation kann in vier Teile untergliedert werden: 1) eine generelle Einführung in die Methodik des design-basierten Ansatzes, 2) ein Anwendungsbeispiel für die Zustandsschätzung in der schweizerischen Landeswaldinventur, 3) eine Erweiterung der Regressionsschätzung auf nichtparametrische Modelle (u.a. k-nearest neighbors Methoden), und 4) eine Neuformulierung der vorgeschlagenen Methoden für die Schätzung von Zustandsveränderungen.

In den vorgestellten Fallstudien werden die Schätzmethoden meist auf den Fall einer jährlich wiederkehrenden Waldinventur angewendet. Ein Grund hierfür ist, dass es das praxisorientierte Ziel dieser Arbeit war, für die 2009 begonnene Umstellung der schweizerischen Landeswaldinventur von einer periodisch- zu einer jährlich-wiederkehrenden Inventur anwendbare Schätzmethoden zu entwickeln. Durch diese Umstellung, welche durch logistische Gründe motiviert war, werden zukünftig jährlich nur 1/9 aller Stichproben im Land erhoben. Diese Reduzierung der Stichprobengrösse führt folglich zu einer Vergrösserung der Varianz bei Anwendung eines herkömmlichen Schätzers unter der Annahme von zufällig verteilten Stichprobenpunkten.

Es konnte gezeigt werden, dass durch die Anwendung eines dreiphasigen klassischen Regressionsschätzers unter Einbeziehung von Kronenhöheninformationen und einer Aktualisierung früherer Messungen die Varianzerhöhung, welche durch die Umstellung der periodischen zur jährlichen Inventur bedingt ist, erheblich reduziert werden konnte. Während die Anwendung von kNN-basierten Regressionsschätzern eine mögliche Alternative zum klassischen Regressionsschätzer darstellt, konnte in einer Simulationsstudie gezeigt werden, dass in diesem Fall ein hohes Risiko besteht, die tatsächliche

Varianz deutlich zu unterschätzen. Um diese Unterschätzung zu verhindern wurde eine entsprechende Bootstrap-Methode vorgestellt. Vergleichbare Verbesserungen sind auch hinsichtlich der Schätzung von Zustandsänderungen möglich. Allerdings zeigte sich hier ein Trade-Off zwischen der realisierbaren Genauigkeitserhöhung der Schätzungen und der Flexibilität, Veränderungen über beliebige Zeiträume zu schätzen.

Chapter 1

Introduction

1.1 Context and scope

In 2009 the Swiss National Forest Inventory (NFI) turned from a periodic into an annual measurement design where only one ninth of the overall sample of permanent plots is measured every year. The main motivation for the transition was that estimates can be calculated continuously over time rather than periodically. A further incentive was that the relatively constant field measurement workload makes budgetary and logistic planning easier to optimize, potentially leading to an overall reduction in costs. However, the drawback is that the full terrestrial sample size is only calculated every 9 years in the new annual design. In order to make estimates for a *time point* represented by a grouping of 3 years, as was done in previous inventories, one must accept a reduction in sample size that leads to an unacceptably large increase in variance when using the standard simple random sampling estimator.

The initial motivation for this cumulative dissertation was to explore methods that address this problem, specifically in the design-based Monte Carlo approach. The most sensible strategy is to incorporate auxiliary information from sources such as remote sensing using multiphase regression estimation. Of course, the proposed methods are not limited to the annual designs and are general enough to increase the precision of any estimate where relevant auxiliary information is available. Whereas much of the literature on this topic tends to emphasize the use of a particular type of model, the goal of this work is to give intuition about how to use and evaluate multiple model types in the design-based framework.

1.2 Why the design-based Monte Carlo approach?

Design-based inference has long been a standard for estimation in forest inventory programs and also in governmental surveys in general. This is largely due to the fact that one does not need to believe that a model, and all of its underlying assumptions, is true, but rather that the sampling design was appropriately carried out via a realization of a controlled probability sampling scheme. Design-based inference is contrasted with model-dependent inference which is more flexible in the sense that it does not need to be associated with a probability sample and can also be applied to small areas where very little or no ground truth has been measured. Model-dependent inference, as the name suggests, is dependent on believing that the model specification is true or at least adequate. It should be emphasised that the difference between design-based and model-dependent estimators is *philosophical* in nature. Therefore, the question should not be ”which method is better?” but rather ”is the data user comfortable believing the model?”

The Monte Carlo approach refers to the practice of sampling from an infinite population of *points* where the sample is uniformly and independently distributed in the forested area. The design-based counterpart of the Monte Carlo approach is to use a finite population scheme where discrete *pixels* rather than points are the sampling units. Choosing a finite population scheme over the Monte Carlo approach most likely makes little difference in practice but the finite approach is mathematically imprecise since no shape of pixel can be tessellated across an amorphous forest area, unless one accepts a pixelated forest boundary definition. We will follow the Monte Carlo approach because it is more intuitive in the forest inventory context than finite population sampling whose theory predominantly comes from fields in social sciences where one samples from a known list of people or addresses. Furthermore, cluster sampling, which is common in many country’s NFIs, is more easily generalized in the infinite population context than it is in the finite context.

1.3 Thesis objective and structure

The objective of this thesis is not to advocate a particular model or technique but rather to explain a general family of estimators that can be used with any model type and then demonstrate their utility in the context of the Swiss NFI. As much of the current literature only focuses on two-phase estimators (a.k.a. double sampling), we will expand the existing theory to include three-phase estimators as well. The two major population

parameters we will examine are for the estimation of state (e.g. timber volume at a given time point) and the estimation of net change (e.g. change in timber volume between two time points).

The model-assisted framework is presented in very general terms in Chapter 2. All proposed point estimators are in the form of generalized difference estimators. For any chosen model there are two important questions: is the point estimator design-unbiased, and how can we estimate the design-based variance? The chapter is logically structured so that these questions can be evaluated for the chosen model. In the end the connection is made to specific well-known techniques such as double sampling for post-stratification which are special cases.

Chapter 3 provides a concrete application of this theory for the estimation of state under the new annual design of the Swiss NFI. A flexible estimation procedure using two- and three-phase regression estimators with linear models is presented with a special focus on utilizing updating techniques to account for disturbances and growth, and is applied to the second and third Swiss NFIs.

Traditionally one of the most common models used with multiphase regression estimators is based on linear regression and yields the classical regression estimator. In recent years there has been a rise in the popularity of nonparametric methods for forest mapping and estimation applications, the most common being some variant of the so-called k-nearest neighbor (kNN) estimator (for examples see Magnussen et al. (2010); McRoberts et al. (2007)). Chapter 4 makes a direct comparison of classical regression estimators to several nonparametric kernel-based estimators of which kNN is a special case.

The estimation of change is addressed in Chapter 5. The notation is restated to include the temporal component and two main variants of estimators are considered: direct and indirect. Direct estimation involves observing the change directly on the plot-level and then making that the response variable of interest. The direct approach is only possible for inventories with permanent plots and is constrained to estimating over time periods matching the duration of re-measurement cycle. Indirect estimation involves making two estimates the state at two time points and taking their difference. Two- and three-phase estimators are presented for both approaches and tested with linear and kNN models using data from the fourth Swiss NFI.

Chapter 2

Model-assisted estimation under the Monte Carlo approach

2.1 The sampling design

We use the Monte Carlo approach (infinite population model), as described in Mandallaz (2008) (Chapter 5). We have a well-defined population P of trees $i \in 1, 2, ..., N$ in forest F. For every tree we have response variable Y_i. The parameter of interest is the spatial mean $\bar{Y} = \frac{1}{\lambda(F)} \sum_{i=1}^{N} Y_i$ where $\lambda(F)$ is the surface area of F. Throughout this entire manuscript, Y_i is individual tree (stem) volume and $\bar{Y}$ is growing stock volume per unit area, however, any measurable tree-level attribute can be considered.

The first phase is a large sample s_1 of n_1 **points** $x \in F$ distributed uniformly and independently in forest F. Auxiliary information is taken in an arbitrary predefined region around each of these points that is ideally highly correlated to the response variable. It does not matter if it is qualitative, quantitative or a combination of both. The auxiliary information is contained in row vector $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)t}(x), \mathbf{Z}^{(2)t}(x))$, where $\mathbf{Z}^{(1)t}(x)$ is the first phase component known for a large sample size of points $x \in F$. $\mathbf{Z}^{(2)t}(x)$ is the second phase component known only for points $x \in s_2$ where $s_2 \subset s_1$. The second phase sample size is denoted n_2. The third phase sample $s_3 \subset s_2$ draws n_3 points and corresponds directly to the observed ground truth measured by the field team. Uniform and independent random sampling is used on an infinite population of points in a bounded domain of a plane for the selection of $s_1 \in F$. Simple random sampling without replacement is used in the selection of subsamples $s_2 \subset s_1$ and $s_3 \subset s_2$.[1] Note

[1] Although most national forest inventories typically use systematic sampling schemes with a random start, treating systematic samples as uniformly random samples is rarely a problem for point estimates

that $x \in s_2$ and $x \in s_3$ inherit i.i.d. uniformness in F. This property is important in that it will allow us to estimate theoretical (i.e. population) variance simply by taking sample copies (i.e. calculating the sample variance of the subsample, s_3).

For every $x \in s_3$ trees are selected from P with inclusion probabilities π_i equal to the area of an inclusion circle (possibly adjusted for the forest boundary) around the ith tree divided by $\lambda(F)$. Commonly known concentric circles or angle count techniques are accounted for in this formulation by defining π_i as a function of a trees diameter at breast height (DBH). For a given point x, the indicator variable for the ith tree is defined as follows:

$$I_i(x) := \begin{cases} 1 \text{ if } i \in s_3(x) \\ 0 \text{ if } i \notin s_3(x) \end{cases} \quad [2.1]$$

where $s_3(x)$ denotes the set of trees selected at point x. Y_i is evaluated for all trees i such that $I_i(x) = 1$. Using the indicator variable and the inclusion probability, we can define the local density estimate.

$$Y(x) := \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) Y_i}{\pi_i} \quad [2.2]$$

In the design-based framework the local density estimate $Y(x)$ is assumed error-free given x. This leads us to the fundamental relationship of the Monte Carlo approach

$$\mathbb{E}_x(Y(x)) = \frac{1}{\lambda(F)} \int_F Y(x) dx = \frac{1}{\lambda(F)} \sum_{i=1}^{N} Y_i = \bar{Y} \quad [2.3]$$

which still holds for both concentric circles and variable radius (angle count) sampling of trees. Historically, Monte Carlo methods were developed to calculate complicated integrals numerically though repeated random sampling. Here we have an unknown and likely complicated surface $Y(x)$ defined over an area F. [2.3] tells us that estimating the parameter of interest $\bar{Y}$ is equivalent to estimating the integral $\frac{1}{\lambda(F)} \int_F Y(x) dx$. Since the proposed sampling scheme ensures that $Y(x)$ is observed for a random sample $x \in s_3$ that is uniform in F, the simplest and most natural point estimator of $\bar{Y}$ is therefore

$$\hat{Y}_{1p} = \frac{1}{n_3} \sum_{x \in s_3} Y(x) \quad [2.4]$$

which remain asymptotically unbiased. In practice, the variance is usually overestimated particularly for small area estimation unless the range of the spatial correlation of the target variable, particularly of the residuals, is shorter than minimum distance between sample plots (see Mandallaz (2008) chapter 7).

As x is i.i.d. in s_3, it is clear that $\hat{Y}_{1p}$ is an exactly design-unbiased estimator with variance

$$\mathbb{V}(\hat{Y}_{1p}) = \frac{1}{n_3}\mathbb{V}(Y(x)) \quad [2.5]$$

which can be design-unbiasedly estimated by sample copy:

$$\hat{\mathbb{V}}(\hat{Y}_{1p}) = \frac{1}{n_3}\frac{1}{1-n_3}\sum_{x\in s_3}(Y(x) - \frac{1}{n_3}\sum_{x\in s_3}(Y(x)))^2 \quad [2.6]$$

$\hat{Y}_{1p}$ is simple and intuitive but does not utilize any information from the auxiliary vector $\mathbf{Z}^t(x)$, which is the motivation for using the model-assisted approach.

2.2 Generalized difference estimators

All model-assisted point estimators considered throughout the course of this work can be expressed in the form of a generalized difference estimator. We will start by presenting the generalized difference estimator for three-phase sampling. It is highly general, intuitively simple and leads naturally into the presentation of regression estimators. The generalized difference estimator in three-phase sampling under Monte Carlo approach is defined by Mandallaz (2014) based on ideas originating from Särndal et al. (2003) as

$$\hat{Y}_{GD,3p} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{M_1}(x) + \frac{1}{n_2}\sum_{x\in s_2}(\hat{Y}_{M_2}(x) - \hat{Y}_{M_1}(x)) + \frac{1}{n_3}\sum_{x\in s_3}(Y(x) - \hat{Y}_{M_2}(x)) \quad [2.7]$$

where $\hat{Y}_{M_1}(x)$ and $\hat{Y}_{M_2}(x)$ are model predictions available for the first and second phase sampling points, respectively, that are error-free given x. We do not specify the type of model (e.g. linear, non-linear, kNN, etc.) from which the predictions $\hat{Y}_{M_1}(x)$ and $\hat{Y}_{M_2}(x)$ are produced, nor that the models are unbiased in the model-dependent sense. For linear and nonlinear prediction models, [2.7] can be viewed in the Monte Carlo approach as a special case of the calibration estimator discussed by Wu and Sitter (2001) (see appendix of Mandallaz and Massey (2015) for an adaptation to the infinite population approach and proof).

For notational consistency we present the two-phase difference estimator as the special case of the three-phase estimator that arises when $s_1 = s_2$.

$$\hat{Y}_{GD,2p} = \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}_{M_2}(x) + \frac{1}{n_3}\sum_{x\in s_3}(Y(x) - \hat{Y}_{M_2}(x)) \quad [2.8]$$

We emphasize that the model-assisted approach presented here is purely **design-based**, not model-dependent, which may be a source of confusion for some readers. A detailed explanation of the philosophical differences between these approaches can be found in Gregoire (1998) in the finite population context and Mandallaz (2008) Chapters 4, 6 and 7 in the Monte Carlo setup. In the model-dependent approach, at a fixed point x, $Y(x)$ is viewed as the realization of a stochastic process. For example, a common model-dependent distribution assumption would be that $Y(x)$ has mean $\boldsymbol{Z}^t(x)\boldsymbol{\beta}$ and $R_{M_2}(x) := Y(x) - \hat{Y}_{M_2}(x)$ has zero mean and a spatial covariance structure such as in the geostatistical kriging procedure (see Mandallaz (2008), Chapter 7). In the design-based framework, x is random and the sample must be the result of a stochastic selection characterized by the sampling design. We are free of any distribution assumptions on $Y(x)$, $\boldsymbol{Z}(x)$ and $R_{M_2}(x)$, which can only be regarded as random because x is random (and are thus fixed for a given x). Inference is based on the theoretical distribution of estimates generated by the sampling design. According to the terminology of Särndal et al. (2003) this approach is **model-assisted**, i.e. we use models to reduce the variance but we do not assume that they are correct. Some authors use the term model-based instead of model-dependent, which is in our opinion a source of confusion because the inference is valid only if the model is true.

Given how generally [2.7] and [2.8] are currently defined, there is no way of describing their design-based variances analytically without specifying more about the type of model used and how it is fit. It should be mentioned that the design-based variance can still be calculated by relying on a resampling method such as the bootstrap.

The bootstrap routine should be selected in such a way that it mimics the sampling process that produced the original sample. Take [2.8] as an example. The simplest formulation for the given design here is to create B independent bootstrap samples each of size n_2 using SRS with replacement from the observed $x \in s_2$. From each bootstrap sample we calculate a bootstrap replicate version of [2.8] denoted $\hat{Y}^b_{GD,2p}$ where $b = 1, 2, ..., B$. The bootstrap population estimate is then

$$\hat{Y}^*_{GD,2p} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}^b_{GD,2p} \qquad [2.9]$$

and the bootstrap estimate of variance is

$$\hat{\mathbb{V}}^*(\hat{Y}_{GD,2p}) = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{Y}^b_{GD,2p} - \hat{Y}^*_{GD,2p})^2 \qquad [2.10]$$

Note that each bootstrap sample will have a random number of points $x \in s_3$, which could lead to bootstrap samples with insufficient final phase sample size to (adequately)

fit the desired model internally. This situation, albeit extremely rare for the sample sizes considered here, can be handled by either throwing out the replicate or considering a stratified bootstrap where n_3 points are resampled from s_3 and $n_2 - n_3$ are resampled independently from s_2 (see Wolter (2007) Chapter 5.6 for details). This formulation is slightly more computationally expensive but ensures a fixed sampling fraction across all bootstrap replicates. The three-phase bootstrap version would proceed in similar fashion except in the simple formulation one resamples from s_1 instead of s_2 and in the Wolter formulation one would sample n_3 points from s_3, $n_2 - n_3$ independently from s_2 and $n_1 - n_2$ from s_1. In this way, the same sizes for each phase is fixed across bootstrap replicates.

The bootstrap is useful because it can be implemented for virtually any practical model. However, it can be computationally intensive and does not produce exactly reproducible results in the sense that rerunning the same bootstrap routine on identical data will not typically produce the exact same output. Thus, it is advisable to avoid bootstrapping if a reliable analytical (i.e. closed-form) variance estimator exists.

2.3 External models

An example of a general class of models for which analytical variance estimators are guaranteed to exist are external models. External models do not depend on what type of model is considered (e.g. linear, nonlinear, kNN, etc.) but rather how they are fit. Specifically, external models are fit using a training set (sometimes called reference set) that is independent of the current inventory's sample realizations s_1, s_2 and s_3. Practically speaking, an example of a true external model is one that is fit using data from an inventory in say Austria and used to make predictions in say Switzerland. Another possibility is that it is fit using an independent sample realization of a past inventory. Yet another less evident possibility would be to fix "known" coefficients in a linear model. This is equivalent to assigning an arbitrary training set that is independent of the current sample realizations. Of course, this is almost never done in real-life as most forest inventorists would clearly choose to use s_3 to fit a predictive model, and s_3 is clearly dependent on the stochastic process of the sampling design by definition. Nevertheless working under the mathematical assumption that M_1 and M_2 are external yields some useful theoretical results.

It is easily checked using the total law of expectation that the overall expectation of the point estimate is the parameter of interest (i.e. $\mathbb{E}_{1,2,3}\hat{Y}_{GD,3p} = \bar{Y}$) and $\hat{Y}_{GD,3p}$ is exactly

design-unbiased. Furthermore, using the law of total variance, the theoretical variance takes the form (see Mandallaz (2013c))

$$\mathbb{V}(\hat{Y}_{GD,3p}) = \frac{1}{n_1}\mathbb{V}(Y(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R_{M_1}(x)) + (1-\frac{n_3}{n_2})\frac{1}{n_3}\mathbb{V}(R_{M_2}(x)) \qquad [2.11]$$

with $R_{M_1}(x)) = Y(x) - \hat{Y}_{M_1}$ and $R_{M_2}(x)) = Y(x) - \hat{Y}_{M_2}$. An exactly design-unbiased estimator of [2.11] is obtained by sample copy.

For two-phases, the point estimator is also exactly unbiased and the external variance takes the form

$$\mathbb{V}(\hat{Y}_{GD,2p}) = \frac{1}{n_2}\mathbb{V}(Y(x)) + (1-\frac{n_3}{n_2})\frac{1}{n_3}\mathbb{V}(R_{M_2}(x)) \qquad [2.12]$$

It can also be estimated via sample copy.

2.3.1 External model approach

To recap, true external models yield analytical point and variance estimators that are exactly design-unbiased for any model, regardless if that model is wrong. However, in practice, most models are not external because they are usually fit from current sample of inventory data. These are called **internal** models. It may be tempting to disregard that a model is fit internally and apply the generalized difference estimators with the appropriate external variance formulas anyway under an implicit **external model assumption**.

In practice the external variance estimator cannot be expected to be valid stricto sensu with internal models because the predictions $\hat{Y}_{M_1}(x_0)$ and $\hat{Y}_{M_2}(x_0)$ at an arbitrary point x_0 are almost always fitted internally using the current inventory data $\{\boldsymbol{Z}(x), Y(x), x \in s_3\}$. For this reason, we shall denote the empirical residual $Y(x) - \hat{Y}_{M_1}(x)$ by $\hat{R}_{M_1}(x)$ (and $\hat{R}_{M_2}(x)$) which should be used in [2.11, 2.12] in place of $R_{M_1}(x)$ and $R_{M_2}(x)$. So far, design-based asymptotic analytical results are available only for linear models, e.g. the classical regression estimator and for nonlinear models via a Monte Carlo model calibration approach (Mandallaz and Massey (2015), Appendix C). In these cases, asymptotic unbiasedness and consistency of the external variances estimates [2.11, 2.12] can be proven. This is not the case for the other nonparametric techniques used to obtain the predictions $\hat{Y}_{M_1}(x)$ and $\hat{Y}_{M_2}(x)$ that are presented in subsequent sections. Intuition suggests that using external variance estimators calculated with [2.11, 2.12] with internal models will underestimate the theoretical variance because we are essentially ignoring variation in the value of $\hat{Y}_{M_1}(x_0)$ (and $\hat{Y}_{M_2}(x_0)$) across different sample

realizations of s_3. As the design-based variances for these estimators are currently not very well understood from an analytical point of view, one must rely on finding an adequate resampling method such as the bootstrap. Chapter 4 delves in greater detail into the adequacy of the external model assumption for various model types.

2.4 Internal models

In order to analytically incorporate the effect of the sampled data on the fit of the internal model and the overall precision of the estimators, one must first specify which type of model to be considered.

In the Monte Carlo approach it is important to understand that we do not assume that $Y(x)$ has been generated by the specified model. The real forest, and hence $Y(x)$, is likely to have been generated by a far more complex mechanism, probably with variables not contained in $\boldsymbol{Z}(x)$ and with some non-linear relationships. We shall view the model as the **working** or **fitted** model. In Chapter 4 we discuss a simulation example, in which we are in the lucky situation to know the true model, i.e. the model used to generate $Y(x)$, so that we can assess the impact of using a working model which is not the true model generating $Y(x)$. To put it in simple terms, the (model-assisted) Monte Carlo objective is not to model the mechanism generating the forest but rather to provide estimators which have a much smaller variance than the ordinary sample mean $\bar{Y}_{s_3} = \frac{1}{n_3} \sum_{x \in s_3} Y(x)$.

Although many types of models are available, deriving analytical design-based results can be extremely algebraically tedious for even simple model types. For this reason we now restrict ourselves only to linear models to demonstrate the internal model approach.

2.4.1 Linear models

Two nested linear models will be considered here: one to be fitted only with the reduced component, $\mathbf{Z}^{(1)t}(x)$, of the auxiliary vector available for all first-phase points in s_1; and a second model fitted with the full auxiliary vector $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)t}(x), \mathbf{Z}^{(2)t}(x))$ available for all second-phase points in s_2. The former will be referred to as $M_{reduced}$ (*reduced* refers to the reduced number of covariates in the nested framework) and the latter as M_{full} (*full* indicates that all covariates in the auxiliary vector are used).

The reduced model

The model $M_{reduced}$

$$Y(x) = \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha} + R^{(1)}(x) \quad [2.13]$$

has a vector of theoretical regression coefficients $\boldsymbol{\alpha}$ and theoretical residual term $R^{(1)}(x)$. In complete analogy to classical least squares in a finite universe, $\boldsymbol{\alpha}$ minimizes the integral of theoretical residuals squared over all possible points x, which is infinite in the Monte Carlo approach. Thus, $\boldsymbol{\alpha}$ minimizes $\int_F R^{(1)}(x)^2 dx = \int_F (Y(x) - \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha})^2 dx$ and leads to the normal equation $(\int_F \mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)t}(x)dx)\boldsymbol{\alpha} = \int_F Y(x)\mathbf{Z}^{(1)}(x)dx$. Of course, we have incomplete information to solve for $\boldsymbol{\alpha}$ directly for the infinite population of points x because $Y(x)$ is observed only given the third phase sample while $\mathbf{Z}^{(1)t}$ is available only for $x \in s_1$. Thus, we estimate $\boldsymbol{\alpha}$ by sample copies of the normal equation leading to:

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (\frac{1}{n_3}\sum_{x\in s_3} \mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)t}(x))^{-1}\frac{1}{n_3}\sum_{x\in s_3} Y(x)\mathbf{Z}^{(1)}(x) \\ &= (\frac{1}{n_3}\mathbf{Z}^{(1)t}_{s_3}\mathbf{Z}^{(1)}_{s_3})^{-1}\frac{1}{n_3}\sum_{x\in s_3} Y(x)\mathbf{Z}^{(1)}(x) \end{aligned} \quad [2.14]$$

where $\mathbf{Z}^{(1)}_{s_3}$ is the design matrix of explanatory variables constrained to the third phase points, $x \in s_3$. In other words, $\hat{\boldsymbol{\alpha}}$ is simply the vector of regression coefficients obtained when fitting $M_{reduced}$ using ordinary least squares (**OLS**) to the third phase sample.

The full model

In the same way as $M_{reduced}$, M_{full} is defined

$$\begin{aligned} Y(x) &= \mathbf{Z}^t(x)\boldsymbol{\beta} + R(x) \\ &= \mathbf{Z}^{(1)t}(x)\boldsymbol{\beta}^{(1)} + \mathbf{Z}^{(2)t}(x)\boldsymbol{\beta}^{(2)} + R(x) \end{aligned} \quad [2.15]$$

where $\boldsymbol{\beta}^t = (\boldsymbol{\beta}^{(1)t}, \boldsymbol{\beta}^{(2)t})$ is the vector of theoretical regression coefficients. Note that $\boldsymbol{\beta}^{(1)} = \boldsymbol{\alpha}$ only if $\mathbf{Z}^{(1)}(x)$ and $\mathbf{Z}^{(2)}(x)$ are orthogonal (i.e. independent), which is almost never the case in practice. The corresponding normal equation is $(\int_F \mathbf{Z}(x)\mathbf{Z}^t(x))\boldsymbol{\beta}dx = \int_F Y(x)\mathbf{Z}(x)dx$ and $\boldsymbol{\beta}$ is estimated:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\frac{1}{n_3}\sum_{x\in s_3} \mathbf{Z}(x)\mathbf{Z}^t(x))^{-1}\frac{1}{n_3}\sum_{x\in s_3} Y(x)\mathbf{Z}(x) \\ &= (\frac{1}{n_3}\mathbf{Z}^t_{s_3}\mathbf{Z}_{s_3})^{-1}\frac{1}{n_3}\sum_{x\in s_3} Y(x)\mathbf{Z}(x) \end{aligned} \quad [2.16]$$

where $\mathbf{Z}_{s_3}$ is the design matrix containing all explanatory variables present in the auxiliary vector but restricted to $x \in s_3$. $\hat{\boldsymbol{\beta}}$ is simply the vector of regression coefficients obtained when fitting M_{full} using ordinary least squares (**OLS**) to the third-phase sample.

Zero mean property

Note that both $M_{reduced}$ and M_{full} possess the same properties as in classical OLS regression such as zero mean empirical residuals:

$$\frac{1}{n_3}\sum_{x\in s_3}\hat{R}(x)=0 \qquad \frac{1}{n_3}\sum_{x\in s_3}\hat{R}^{(1)}(x)=0 \tag{2.17}$$

where $\hat{R}^{(1)}(x) = Y(x) - \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}$ and $\hat{R}(x) = Y(x) - \mathbf{Z}^{t}(x)\hat{\boldsymbol{\beta}}$). Another property inherited by construction is the orthogonality of the true residuals to the model space (i.e. $\mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha} \perp R^{(1)}(x)$ and likewise for M_{full}). Although these models are unbiased in the model-dependent sense it should be emphasized that we will be using them in the design-based framework where we never need to make the assumption that they hold true in the model-dependent sense.

2.4.2 The classical regression estimator

The three-phases version:

The three-phase classical regression estimator occurs when one applies the nested internal models $M_{reduced}$ and M_{full} to 2.7.

$$\begin{aligned}\hat{Y}_{REG,3p} &= \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}^{(1)}(x)+\frac{1}{n_2}\sum_{x\in s_2}(\hat{Y}(x)-\hat{Y}^{(1)}(x))+\frac{1}{n_3}\sum_{x\in s_3}(Y(x)-\hat{Y}(x)) \\ &= \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}^{(1)}(x)+\frac{1}{n_2}\sum_{x\in s_2}(\hat{Y}(x)-\hat{Y}^{(1)}(x))\end{aligned} \tag{2.18}$$

where $\hat{Y}^{(1)}(x) = \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}$ and $\hat{Y}(x) = \mathbf{Z}^{t}(x)\hat{\boldsymbol{\beta}}$ are the predictions for the target variable $Y(x)$ at point x based on the reduced and full sets of explanatory variables respectively.

The estimator can be rewritten in the more useful form

$$\begin{aligned}\hat{Y}_{REG,3p} &= (\hat{\bar{\mathbf{Z}}}^{(1)}_{s_1}-\hat{\bar{\mathbf{Z}}}^{(1)}_{s_2})^t\hat{\boldsymbol{\alpha}}+(\hat{\bar{\mathbf{Z}}}_{s_2}-\hat{\bar{\mathbf{Z}}}_{s_3})^t\hat{\boldsymbol{\beta}}+\frac{1}{n_3}\sum_{x\in s_3}Y(x) \\ &= (\hat{\bar{\mathbf{Z}}}^{(1)}_{s_1}-\hat{\bar{\mathbf{Z}}}^{(1)}_{s_2})^t\hat{\boldsymbol{\alpha}}+\hat{\bar{\mathbf{Z}}}^{t}_{s_2}\hat{\boldsymbol{\beta}}\end{aligned} \tag{2.19}$$

where $\hat{\bar{\mathbf{Z}}}^{(1)}_{s_1} = \frac{1}{n_1}\sum_{x\in s_1}\mathbf{Z}^{(1)}(x)$, $\hat{\bar{\mathbf{Z}}}^{(1)}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\mathbf{Z}^{(1)}(x)$, $\hat{\bar{\mathbf{Z}}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\mathbf{Z}(x)$ and $\hat{\bar{\mathbf{Z}}}_{s_3} = \frac{1}{n_3}\sum_{x\in s_3}\mathbf{Z}(x)$.

As the true residuals are design-based uncorrelated with the predictions by construction in least squares, we get $\mathbb{V}(Y(x)) = \mathbb{V}(\hat{Y}^{(1)}(x)) + \mathbb{V}(\hat{R}^{(1)}(x))$ for $M_{reduced}$, and $\mathbb{V}(Y(x)) = \mathbb{V}(\hat{Y}(x)) + \mathbb{V}(\hat{R}(x))$ for M_{full}. Furthermore, if we accept the external model assumption (i.e. considering the internal models to be external), we can plug these decompositions into [2.11] and the theoretical variance simplifies to

$$\mathbb{V}(\hat{Y}_{REG,3p}) = \frac{1}{n_1}\mathbb{V}(\hat{Y}^{(1)}(x)) + \frac{1}{n_2}\mathbb{V}(\hat{R}^{(1)}(x)) + (1-\frac{n_3}{n_2})\frac{1}{n_3}\mathbb{V}(\hat{R}(x)) \qquad [2.20]$$

Note that sample copies could be used to estimate the terms in [2.20] and that this sample copy would likely be a little better than the sample copy of [2.11] because $\mathbb{V}(\hat{Y}^{(1)}(x))$ can be estimated using all of s_1 whereas $\mathbb{V}(Y(x))$ is only observable on s_3. In any case, here we are forgoing the external model assumption in order to develop a variance estimator which takes the effects of an internally fitted model into account (i.e. the sampling variability on $\hat{\alpha}$ and $\hat{\beta}$).

$\hat{Y}_{REG,3p}$ is still asymptotically design-unbiased and the variance estimator is written, for simplicity, in terms of g-weights (Mandallaz (2014, 2013a)):

$$g^{(1)}(x) := \hat{\bar{\mathbf{Z}}}^{(1)t}_{s_1}(\mathbf{A}^{(1)}_{s_2})^{-1}\mathbf{Z}^{(1)}(x) \qquad [2.21]$$

$$g(x) := \hat{\bar{\mathbf{Z}}}^{t}_{s_2}\mathbf{A}^{-1}_{s_3}\mathbf{Z}(x) \qquad [2.22]$$

where

$\mathbf{A}^{(1)}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)}(x)^t = \frac{1}{n_2}\mathbf{Z}^{(1)t}_{s_2}\mathbf{Z}^{(1)}_{s_2}$

$\mathbf{A}_{s_3} = \frac{1}{n_3}\sum_{x\in s_3}\mathbf{Z}(x)\mathbf{Z}(x)^t = \frac{1}{n_3}\mathbf{Z}^t_{s_3}\mathbf{Z}_{s_3}$

with $\mathbf{Z}^{(1)}_{s_2}$ and $\mathbf{Z}_{s_3}$ being the design matrices of $M_{reduced}$ and M_{full} defined over the entire second and third phases respectively. The three-phase classical regression estimator possesses the following useful calibration properties:

$$\frac{1}{n_2}\sum_{x\in s_2} g^{(1)}(x)\mathbf{Z}^{(1)}(x) = \hat{\bar{\mathbf{Z}}}^{(1)}_{s_1} \qquad \frac{1}{n_2}\sum_{x\in s_2} g^{(1)}(x) = 1 \qquad [2.23]$$

$$\frac{1}{n_3}\sum_{x\in s_3} g(x)\mathbf{Z}(x) = \hat{\bar{\mathbf{Z}}}_{s_2} \qquad \frac{1}{n_3}\sum_{x\in s_3} g(x) = 1 \qquad [2.24]$$

The $g^{(1)}(x)$-weighted mean of the auxiliary variables $\mathbf{Z}^{(1)}(x)$ over second-phase sample s_2 is equal to the mean over first-phase sample s_1 (i.e. $\hat{\bar{\mathbf{Z}}}^{(1)}_{s_1}$). Similarly, the $g(x)$-weighted mean of $\mathbf{Z}(x)$ over s_3 is equal to the mean of $\mathbf{Z}(x)$ over s_2 (i.e. $\hat{\bar{\mathbf{Z}}}_{s_2}$). The intuition follows

that it is desirable to apply these weights to $Y(x)$ provided that $Y(x)$ is well-correlated with the auxiliary vectors $\mathbf{Z}^{(1)}(x)$ and $\mathbf{Z}(x)$. The g-weights tend to 1 asymptotically and have a mean equal to 1 for $g^{(1)}(x)$ and $g(x)$. Furthermore, it is important to note that better statistical properties arise when the g-weights are used explicitly in the variance estimator than compared to a variance estimator derived by the external model assumption (Mandallaz (2013a,b)).

There is no closed analytical formula for the theoretical variance under the internal model, however, an asymptotically consistent variance estimator derived by the g-weight technique is defined (Mandallaz (2013c)) as

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{REG,3p}) = \frac{1}{n_1}\frac{\sum_{x\in s_1}(\hat{Y}^{(1)}(x)-\hat{\bar{Y}}^{(1)}_{s_1})^2}{n_1-1} &+ \frac{1}{n_2}\frac{1}{n_3}\sum_{x\in s_3}(g^{(1)}(x)\hat{R}^{(1)}(x))^2 \\ &+ \frac{1}{n_3^2}(1-\frac{n_3}{n_2})\sum_{x\in s_3}(g(x)\hat{R}(x))^2\end{aligned} \quad [2.25]$$

where $\hat{Y}^{(1)}(x) = \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}$ and $\hat{\bar{Y}}^{(1)}_{s_1} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}^{(1)}(x)$. Notice that [2.25] is asymptotically equivalent to the theoretical variance under the external model assumption given in [2.20] because of the zero mean property of the empirical residuals and the fact that the g-weights tend to 1 asymptotically.

The two-phase version:

Without undue redundancy, we also want to present the two-phase regression estimator (REG,2p). Recall that $s_1 = s_2$ for notational convenience and that there is only the model M_{full}. REG,2p simplifies to

$$\begin{aligned}\hat{Y}_{REG,2p} &= \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}(x) + \frac{1}{n_3}\sum_{x\in s_3}(Y(x)-\hat{Y}(x)) \\ &= \frac{1}{n_2}\sum_{x\in s_2}\mathbf{Z}^t(x)\hat{\boldsymbol{\beta}} + \frac{1}{n_3}\sum_{x\in s_3}(Y(x)-\mathbf{Z}^t(x)\hat{\boldsymbol{\beta}}) \\ &= (\hat{\bar{\mathbf{Z}}}_{s_2}-\hat{\bar{\mathbf{Z}}}_{s_3})^t\hat{\boldsymbol{\beta}} + \frac{1}{n_3}\sum_{x\in s_3}Y(x) \\ &= \hat{\bar{\mathbf{Z}}}^t_{s_2}\hat{\boldsymbol{\beta}}\end{aligned} \quad [2.26]$$

Analogous to the three-phase version, $\hat{Y}(x)$ and $\hat{R}(x)$ have zero empirical design-based covariance by construction and the theoretical variance under the external model assumption simplifies to

$$\mathbb{V}(\hat{Y}_{REG,2p}) = \frac{1}{n_2}\mathbb{V}(\hat{Y}(x)) + \frac{1}{n_3}\mathbb{V}(\hat{R}(x)) \quad [2.27]$$

which has a sample copy estimator that is slightly better than [2.12] because it can estimate $\hat{Y}(x)$ using the larger sample s_2.

In order to incorporate the effect of internally fitting $\hat{Y}(x)$ with a linear model, we will need the asymptotic design-based variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{s_2}$ (see Chapter 6 from Mandallaz (2008))

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \boldsymbol{A}_{s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\Big)\boldsymbol{A}_{s_2}^{-1} \qquad [2.28]$$

Using [2.28] and a first order Taylor expansion of $\bar{\boldsymbol{Z}}_1^t\hat{\boldsymbol{\beta}}_{s_2}$ we obtain the consistent g-weight variance estimate of $\hat{Y}_{reg}$ (see Mandallaz (2008), pp. 123-125 and Mandallaz (2013b), eq. [23] for details).

$$\hat{\mathbb{V}}(\hat{Y}_{REG,2p}) = \frac{1}{n_2}\frac{1}{n_3-1}\sum_{x\in s_3}(Y(x)-\bar{Y}_{s_3})^2 + \frac{1}{n_3^2}(1-\frac{n_3}{n_2})\sum_{x\in s_3}(g(x)\hat{R}(x))^2 \qquad [2.29]$$

There is only one set of g-weights that is defined the same as in [2.22]. With some algebra one can show that the sample copy estimate of [2.27] is asymptotically equivalent to [2.29]. Thus we conclude that the external variance estimate is also consistent for internal linear models.

2.4.3 Special cases of the classical regression estimator

The two-phase classical regression estimator[2] has some special cases that are worth mentioning:

- Double sampling for stratification
  - Only categorical variables are present in the model (i.e. we have an ANOVA model). The strata weights are calculated based on the first phase.
- Double sampling for regression
  - Only continuous variables are present in the model. Multiple linear regression is possible and a single intercept is possible.

[2]These special cases tend to lose their concrete interpretations when extended to three-phase regression estimation. We emphasize that one should not restrict oneself to utilizing classical regression estimators for special cases, but merely that they are easily implemented. The goal intuitively in model selection is to minimize the residuals as they typically play the most prominent part in the variance estimation.

- Double sampling for regression within strata
  - A combination of categorical and continuous variables are present in the model. When an interaction term is not present between categorical and continuous variables the regression lines within strata will have the same slope but different intercepts. When the interaction is present, the slope is allowed to vary within strata which is defined by the categorical variable.

The external variance formulas can also be used for these special cases but the g-weight variance estimator is known to have some statistical advantages. A concrete example is double sampling for post-stratification (see p. 84 in Mandallaz (2008)). By writing out the external variance estimator and the g-weight estimator, one sees that strata weights are based on the the small sample s_3 for the former whereas the latter's strata weights are calculated from the larger sample s_2. Considering that both estimators exist in closed form and are easily programmable using standard statistical software, nothing is lost by using the better g-weight version.

Chapter 3

Application: Estimation of state using linear models

The objective of this chapter is to present an application of two- and three-phase classical regression estimation of timber volume in the context of annual designs with no partial replacement. We work in the Monte Carlo setup described in Chapter 2 with internal linear models. While some authors have delved into similar model-assisted approaches using remote sensing in the two-stage (see Gregoire et al. (2011)), the three-phase setup in the specific context of regression within strata (see Lüpke et al. (2012)) and two-phase sampling with partially-exhaustive information (Mandallaz (2013b)), here three-phase regression estimation is presented using the desirable g-weight technique to demonstrate its flexibility with data taken from permanent plots at the terrestrial level and remote sensing data derived from manually interpreted aerial photographs on which tree and forest stand data are observed.

The material in this chapter has been published in the *Canadian Journal of Forest Research* (see Massey et al. (2014)).

3.1 Background

The 4th Swiss NFI is the first Swiss NFI to implement an annual measurement design where $1/9$ of its permanent remeasured terrestrial plots are assessed every year. Annual designs have the advantage of providing estimates continuously rather than periodically, but have the major drawback of a dramatically reduced sample size if the simple random sampling (SRS) estimator is used for any given year. The loss in precision due to the reduction in sample size can be mitigated by grouping years together.

For example, grouping three years of data, which seems reasonable for the Swiss NFI as this is approximately how long it took to conduct each of the previous three inventories, decreases the sample reduction from one ninth to one third, but still is expected to inflate the variance by a factor of three on both the national and regional levels. Furthermore, grouping many years together deteriorates the concept of a coherent time point associated with the point estimate.

A better approach than grouping is to enhance the estimates with auxiliary data obtained from remote sensing data (i.e. satellite images, LIDAR, stereographic photos, etc.) or with information from past inventories. A plethora of such methods have been explored in the literature. In the context of incorporating past inventory data, Eskelson et al. (2009) and Johnson et al. (2003) compared various moving averages and weighted moving averages with the goal of achieving higher estimation precision by combining annual samples from multiple years. Another popular approach involves updating plots that were not in the annual sample using the current year's realization as a reference set. Such methods in the model-dependent case include combining inventory and remote sensing data using plot-level growth models (McRoberts (2001)), spatial-temporal models (Houillier and Pierrat (1992)), imputation techniques on the tree and plot level (McRoberts (2001);Van Deusen (1997)), tree-level growth models (Lessard et al. (2001)), enhancing estimates over successive measurements using a Kalman filter (Dixon and Howitt (1979)) and mixed estimation (Van Deusen (2002)). Further model-dependent methods including regression updating, inverse weighting of new and old plots and regression combined with growth-model projections have been used to compare periodic and annual designs, for both continuous forest inventory and sampling with partial replacement (SPR) (Scott et al. (1999)). Literature concerning design-based methods for enhancing estimation under an annual design is currently scarce.

Here we will demonstrate how to integrate remote sensing data and previous inventory under an annual design in the Swiss National Forest Inventory using classical regression estimation to estimate timber volume. The main estimator of interest is a three-phase regression estimator (REG,3p) where the first phase uses information gathered from remote sensing, the second phase is based on the previous plot measurements of the entire terrestrial sample and the third phase is the current annual subsample for any given year or grouping of consecutive years. The classical two-phase regression estimator (REG,2p) will also be presented so that we can assess the efficiency of this scheme.

3.2 Study area

The Swiss national forest inventory (NFI) consists of some 160,000 systematically distributed and permanent sampling points for aerial photo interpretation, and a sub-grid of some 20,000 systematically distributed and permanently installed sampling points for terrestrial data collection (Lanz et al. (2010)). Approximately one third of the points are located in forest and have been retained for this study. The restriction to a forest sub-sample is not straight forward and deserves some explanation. Firstly, there exists no map of forest land which conforms with the NFI forest definition. Aerial photo interpretation and terrestrial inventory provide independent forest versus non-forest classifications. The classifications are highly correlated with less than 1% of the sampling points have a differing classification. In the standard estimation procedures of the NFI, the terrestrial classification is the ground truth and the classification derived from aerial photo interpretation is considered an auxiliary information in a two-phase estimation procedure (double sampling for post-stratification).

For the purpose of this study, we retained first-phase sampling points which have been classified as forest during aerial photo interpretation. From these points, the second phase is defined on a one eighth sub-grid. The third phase ideally should be based on a further sub-grid based on three consecutive years taken from the annual sub-grid implemented in current 4th Swiss NFI. However, in practice, 368 of the plots have no ground truth measured by the field crew because of a differing forest/non-forest classification. These plots are considered miss at random and were removed from the

Table 3.1: Sample size for each phase by region

Region	Forested area (HA)	n_1	n_2	n_3
Jura	201185	8064	996	331
Plateau	230014	9294	1164	379
Prealps	218596	8864	1092	347
Alps	370842	15082	1944	602
South Alps	151570	6466	821	240
Switzerland	1172207	47770	6017	1899

third phase post-factum. The sample sizes[1] are given in Table 3.1.

The 2nd and 3rd Swiss National Forest Inventories (NFI) are two-phase forest inventories where permanent terrestrial plots are defined by systematically distributed points on interpenetrating square grids. The 2nd Swiss NFI was conducted from 1993 to 1995 whereas the 3rd Swiss NFI was conducted from 2004 to 2007. In both inventories the first phase draws a sample from a regular $500m$ by $500m$ grid for auxiliary variables based on manually-interpreted digital aerial stereo-photographs that includes roughly 160,000 forest and non-forest plots. The second phase consists of roughly 20,000 plots taken on a $\sqrt{2}km$ by $\sqrt{2}km$ sub-grid for target variables. Approximately a third of these are forest plots and are evaluated based on terrestrial measurements. The survey partitions Switzerland (CH) into 5 production regions: Jura (JU), Plateau (SP), Pre-Alps (PA), Alps (AL), and South Alps (SA).

3.3 Ground Data

Trees on each plot are reidentified and remeasured every inventory using calipers according to two concentric circles measuring $200m^2$ for trees with diameter at breast height (DBH) between $12cm$ and $36cm$ and $500m^2$ for trees with DBH greater than $36cm$. There are approximately 11 trees per plot on average. An approximation of timber volume (over bark stem volume) is determined by using DBH (one-way yield table stratified by species) on all eligible trees. For details we refer the reader to Brassel and Lischke (2001). For the purposes here, it is sufficient to assume that the local density estimate is a known error-free quantity.

The 4th Swiss NFI, which began in 2009, differs from the 2nd and 3rd due to the implementation of a new annual design where 1/9 of the plots are measured every year using interpenetrating grids to allow for more up-to-date estimates to be available between inventory cycles. The drawback of this annual strategy is that these estimates are subject to a dramatically reduced sample size which provides the motivation for exploring more efficient model-assisted methods.

[1]NOTE: These sample sizes are based on a ”working forest definition” derived solely from aerial stereo-photography. In practice, some plots under this working definition are not actually measured terrestrially because they are deemed to be non-forest by the field crew. n_3 is adjusted to exclude plots for which the target variable was not measured and we assume that these responses are missing at random.

3.4 Digital aerial stereo-photography

As previously mentioned, the first phase of the Swiss NFI consists of a large sample of plots taken from a $500m$ by $500m$ grid using digital aerial stereo-photography. The main purpose of these photographs is to make a forest/non-forest decision so that field crew do not have to visit non-forest plots. In the 2nd and 3rd NFI other continuous landscape variables such as canopy height information was also assessed at these plots. The basic sampling unit consists of a $50m$ by $50m$ square interpretation area each containing 25 equally spaced raster points arranged in a 5 by 5 design. The analogue true color photos were flown at a scale of 1:30,000 and scanned at a $14\mu m$ resolution. Once digitized the photos had an aerial ground resolution of 0.42 meters and a RMS error of $< 1m$ after triangulation.

Each lattice point was assigned one thematic cover class by a photo interpreter class using a 3D stereo soft copy station (Socet Set 5.0, BAE Systems). Cover classes included tree vegetation ($> 3m$), shrub and herb vegetation ($< 3m$), soil and sand and gravel, rock, non-natural surfaces, and open water. Canopy height information was calculated for each lattice point by taking the difference between the photogrammetically measured surface elevation by the interpreter and a bilinearly interpolated $25m$ spaced terrain model provided by the Swiss Federal Office of Topography. For cases where the forest border was predefined, a forest boundary line was digitized on screen.

The forest/non-forest classification was made based on the continuous landscape described in such a way as to mirror the NFI final forest definition that will be made terrestrially by the field crew. In the Swiss NFI the main requirements are that the tree canopy height, excluding burned, cut, damaged, afforested or regeneration forests, is greater than $3m$, there is a minimum tree canopy cover of 20 percent within the forest boundary line, and there is a minimum stand width of $25m$ (for full mathematical details see Keller's section on pp.51 in Brassel and Lischke (2001)).

3.5 Sampling frame

The sampling frame is selected using the forest/non-forest decision from the aerial stereo-photographs associated with NFI 3 as a working definition for forest boundary. A small percentage of plots (about 0.9%) were designated "no decision" because of difficulty of interpretation due to external factors such as weather. Only plots designated by the aerial photos as "forest" were included in the sampling frame. A forest/non-forest

definition taken on the ground by the field crew is also possible if a non-forest stratum is allowed (note that misclassification of forested plots in the non-forest stratum is allowable), but for the purposes of evaluating the efficiency of the models in the forest area, it is more appropriate to limit the frame to the forested area.

3.6 Target Variable

The target variable of interest is the growing stock volume (VOL) in cubic meters per hectare ($\frac{m^3}{ha}$) of living trees excluding shrub forest in the accessible forest. There is a national threshold of 12 cm DBH for trees to be included in the estimates.

3.7 Auxiliary Variables

The considered first-phase auxiliary variables included the mean canopy height of the raster points identified in forest (MCH), the variance of these points (VARCH), the proportion of these raster points lying on a coniferous tree (CONPROP), the proportion over all raster points in the forest area (INFORESTPROP), the final stratification variable defined by disturbance class (defined in the next section) and region (REGION:DIST), and all possible interactions between the aforementioned variables. 5 quartiles of canopy heights were also considered.

The main variable of interest in the second phase is the plot's previous measurement for growing stock volume (PREV_VOLUME) obtained in NFI 2. PREV_VOLUME is usually one of the strongest predictors of it's current state assuming that the time to remeasurement is not unreasonably long. In theory, its predictive accuracy can be improved by accounting for two types of forest changes: (i) loss due to mortality and disturbances due to natural or unnatural causes; and (ii) growth, including both physical tree growth and in-growth (i.e. trees surpassing the selection threshold). In the design-based setup it is assumed that the predictors are known and error-free given x so we can directly input predictions derived from any external tree or plot level model into the proposed REGs as auxiliary variables. Here we will account for loss using a stratification variable (DIST) based on remote-sensing data acquired in the first phase. Although any model that provides good predictions can be used, for this case study growth will be accounted for in each plot by aggregating predictions derived from tree-level linear growth models applied to previous inventory data. This variable, denoted PREV_VOLUME_UPDATED,

does not account for volume caused by in-growth, but this is thought to be acceptable since in-growth is not expected to account for much of the overall plot volume.

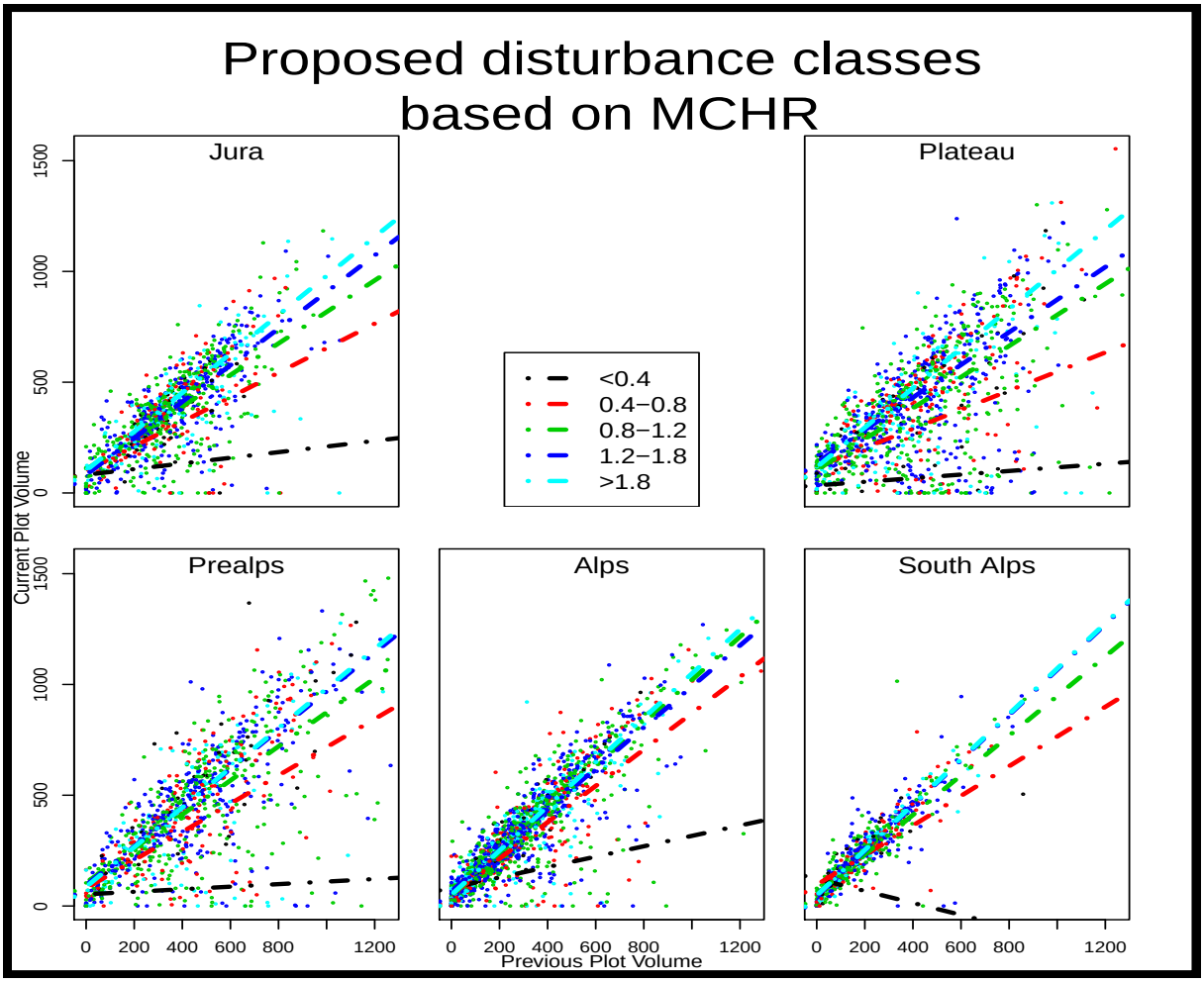
3.7.1 Accounting for disturbance

The stratification variable, DIST, is intended to account for potential loss of growing stock at each plot due to natural and human causes. DIST is derived using canopy height information from two stereo-photos taken from two different time frames. The first photo was taken just before NFI 2 (flight years occurred 1987-1993) and the second was the most recent taken before the next remeasurement in NFI 3 (flight years occurred 1998-2005). The key variable used to aid in disturbance prediction is the mean canopy height ratio (MCHR) defined as the mean canopy height of the raster points in the latter photo divided by the mean at the same raster points from the former photo. Unreasonable values for the raster heights (such as below 0 or above 55) are considered missing. The idea behind the MCHR is that if a disturbance occurred (i.e. harvesting, wind throw, etc.) in the window between the two photos then this ratio is likely to be small.

The MCHR is prone to considerable noise that may arise from the preciseness of the locations of the raster points from photo to photo and the time of year the photo was taken. Because no linear model was found where MCHR was statistically signifiant as a continuous explanatory variable to predict current plot volume, it was decided to break it into classes to be used as a stratification variable. 5 classes were selected: $(-\infty, 0.4]$, $(0.4, 0.8]$, $(0.8, 1.2]$, $(1.2, 1.8]$, and $(1.8, \infty)$.

The harvesting intensities vary by region in Switzerland so the effectiveness of these stratification classes were assessed visually. We employ the intuition behind "regression within strata" where we have a categorical variable in the first and second phases and a continuous variable (PREV_VOLUME) in the second phase. Conceptually, the effectiveness of regression within strata comes when the slope of the regression line varies between strata. The slope of the regression lines within strata are displayed in Figure 3.1. As expected the larger slopes correspond with larger values of MCHR. The correlations between past and current measurement are generally strong indicating that the stratum is suitable to fit a regression line in. In cases where the slopes between strata were too similar the strata were collapsed, as was also the case if a stratum contained too few samples corresponding to the third phase (recall that coefficients of the regression must be fit to the final phase where the response is known). Table 3.2 contains the third phase sample size within strata for plots designated as forest plots in

Figure 3.1: Regression lines for individual disturbance classes based on MCHR by region with plot volumes in $\frac{m^3}{ha}$



Legend:
The points represent all known plot volumes for Swiss NFI 2 (previous plot volume) and Swiss NFI3 (current plot volume). Each regression line depicted is fit only on a subset of these points which belong to a particular disturbance class definable for the entire first phase.

both the previous and current inventories as defined by the aerial photos.

The final disturbance class stratification (DIST) selected was $(-\infty, 0.8]$, $(0.8, 1.2]$, and $(1.2, \infty)$ within Jura, $(-\infty, 0.8]$, $(0.8, 1.2]$, and $(1.2, \infty)$ within Plateau, $(-\infty, 0.8]$, and $(0.8, \infty)$ within the Prealps, $(-\infty, 0.8]$, and $(0.8, \infty)$ within the Alps, and the South Alps was treated as its own stratum without any disturbance classes (see Table 3.2). The final stratification by region (denoted REGION:DIST) also includes a stratum for all new forest plots where the plot was not defined as forest in NFI 2 by the aerial photography

Table 3.2: Third-phase sample sizes in forest within disturbance classes

Disturbance class	Jura	Plateau	Prealps	Alps	South Alps	Switzerland
< 0.4	2^a	18^d	11^g	26^i	0^k	57
0.4-0.8	33^a	44^d	37^g	89^i	19^k	222
0.8-1.2	122^b	134^e	143^h	169^j	59^k	627
1.2-1.8	124^c	14^f	115^h	198^j	87^k	673
> 1.8	52^c	37^f	47^h	112^j	67^k	315

Legend: $a, b, c, ..., k$ represent final stratification membership after collapsing disturbance classes with similar regression line slopes and/or low third phase sample size.

decision.

3.7.2 Accounting for growth on the tree-level

Here we explain the derivation of PREV_VOLUME_UPDATED by examining two linear tree-level growth models whose variables are selected with predictive criteria such as the Akaike Information Criterion (AIC). The models expand on the idea of fully utilizing all of the previous measurements of the Swiss NFI data (note that this time we are on the tree level, not the plot level). A logically strong predictor of the future DBH would be the past DBH combined with the previous DBH increment obtained during the previous remeasurement cycle (i.e. trees must be measured twice before in order to acquire the previous increment). This of course is not possible for plots that entered the Swiss NFI sample for the first time in the previous inventory. Thus, two separate linear models are proposed to account for these two cases. DBH is used as the target response variable because it can be directly input into the one-way yield table (refer to section 3.3).

The models were trained on all available terrestrial Swiss NFI data. After assembling a comprehensive list of plausible predictive variables (e.g. past DBH, time between measurements, categorical species/region groupings, etc) accompanied with the recommended variance stabilizing transformations (including a log-transformed response variable), a backwards selection algorithm comparing AIC was implemented. AIC has good statistical properties in variable selection because it protects against overfitting. However, given the very large sample sizes of the training set (there were 55248 trees where the previous DBH growth increment exists and 120297 trees where only last known DBH existed) there were many variables that were strongly statistically sig-

nificant but not necessarily relevant. Thus, the effect on the response of the maximum value of each remaining explanatory variable was considered. In the presence of a logged response variable, y, the interpretation of the coefficients is multiplicative rather than than additive because $\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p \iff \hat{y} = \exp\hat{\beta}_0 \exp\hat{\beta}_1 x_1 ... \exp\hat{\beta}_p x_p$. So the potential effect on y of the maximal value of any given explanatory variable x_q is $\mid 1 - \exp\hat{\beta}_q \max x_q \mid$. To demonstrate the interpretation, if $\mid 1 - \exp\hat{\beta}_q \max x_q \mid = 0.0034$ then given that all the other explanatory variables stay the same, the maximal effect of x_q on y is only 0.34% implying weak predictive relevance in the model. Any variable with a maximal effect less than 7% was dropped. Note that the same idea can be applied to the effect of the median and minimum values of x_q. The following models were selected

$$\begin{aligned}
\log(DBH) &= \gamma_0 + \gamma_1 \log(PREV_{DBH}) + \gamma_2 \log(PREV_{DBH})^2 + \gamma_3 \log(PREV_{INC} + 10) \\
&+ \gamma_4 VEGUNTIL + \gamma_5 \log(BASFPH + 1) + \gamma_{6:9} SCHICHT \\
&+ \gamma_{10} log(Z25) + \gamma_{11:39} TARNR + \epsilon \qquad [3.1] \\
\log(DBH) &= \eta_0 + \eta_1 \log(PREV_{DBH}) + \eta_2 \log(PREV_{DBH})^2 + \eta_3 \log(DDOM) \\
&+ \eta_4 VEGUNTIL + \eta_5 \log(BASFPH + 1) + \eta_{6:9} SCHICHT \\
&+ \eta_{10} log(Z25) + \eta_{11:39} TARNR : VEGUNTIL + \eta_{40} \log(GWL3) \\
&+ \eta_{41:69} TARNR + \epsilon \qquad [3.2]
\end{aligned}$$

where DBH = current observed DBH, $PREV_{DBH}$ = (centered) diameter at breast height in the previous inventory, $PREV_{INC}$ = previous observed change in DBH standardized by the number of vegetative periods, $VEGUNTIL$ = number of vegetative periods, $BASFPH$ = basal area of standing living trees in m^2 per hectare, $SCHICHT$ = a categorical description of the story (i.e. upper, middle, lower, no story, story, undistinguishable), $Z25$ = altitude above sea level, $TARNR$ = a categorical variable defining region and species, $GWL3$ = a measure of site quality in terms of maximum mean annual biomass increment in kg per hectare per year (Keller (1978)), and $DDOM$ = an index describing the mean diameter of the 100 thickest trees per hectare. DDOM is calculated on the plot as the mean DBH of the largest trees. Trees over 36cm have a weight of 20 and trees with DBH between 12cm and 36cm have a weight of 50. The thickest trees are selected such that the sum of their weights is greater or equal to 100. More technical definitions of these variables can be found in Brassel and Lischke (2001).

[3.1] has an R^2 = 0.9815 and is applied in cases where the $PREV_{INC}$ is available whereas [3.2] has R^2 = 0.9676. The R^2 is high because of the strong influence of predicting the future DBH knowing the past one. Both models were fitted using data

collected in all four Swiss NFIs excluding information on the response variable collected during NFI 3 (i.e. the growth update is externally fit). PREV_VOLUME_UPDATED is derived by applying the growth models to all individual trees present on the plot in the previous measurement and summing their predicted volumes together.

3.8 Model Selection

The $M_{reduced}$ was selected based on the first-phase remote-sensing variables from Swiss NFI 3 using a forward variable selection procedure based on AIC and then an additional stepwise selection procedure with the predictors of the resulting model. The procedure was used to individually fit linear models on all regions as well as Switzerland as a whole using VOL as the response variable. The resulting models selected all contained the variables MCH, CONPROP, INFORESTPROP, REGION:DIST but differed slightly in which sets of interaction terms were included. The following model was chosen to represent Switzerland and its regions

$$\begin{aligned} VOL &= \alpha_0 + \alpha_1 MCH + \alpha_2 CONPROP + \alpha_3 INFORESTPROP \\ &+ \alpha_{4:15} REGION : DIST + \alpha_{16} MCH : INFORESTPROP \\ &+ \alpha_{17:28} MCH : REGION : DIST + \epsilon \end{aligned} \quad [3.3]$$

The model [3.3] was selected so that the interpretation of the coefficients could be intuitively inferred. For parsimony the interaction between REGION:DIST and INFORESTPROP was dropped as it was difficult to interpret and only negligibly increased the adjusted R^2 when applied to all of Switzerland.

The inclusion of CONPROP as opposed to VARCH is somewhat surprising and is likely due to a correction of underestimated MCH on plots where many conferous trees with pointier crowns are present. The raster point is less likely to be aligned with the highest part of the crown, thus explaining the positive coefficient for CONPROP in all regions. The interaction MCH:INFORESTPROP accounts for plots that are likely to be close to the forest boundary and MCH:REGION:DIST is the disturbance corrected prediction using MCH.

The second phase model, M_{full}, incorporating PREV_VOLUME and all the variables

present in [3.3] is

$$\begin{aligned} VOL &= \beta_0 + \beta_1 MCH + \beta_2 CONPROP + \beta_3 INFORESTPROP \\ &+ \beta_{4:15} REGION : DIST + \beta_{16} MCH : INFORESTPROP \\ &+ \beta_{17:28} MCH : REGION : DIST + \beta_{29} PREV\_VOLUME \\ &+ \beta_{30:41} PREV\_VOLUME : REGION : DIST + \epsilon \end{aligned} \quad [3.4]$$

The presence of the MCH:REGION:DIST interaction term obscures the interpretation of the $PREV\_VOLUME : REGION : DIST$ coefficients. When MCH:REGION:DIST is dropped from both models, the coefficients have the expected interpretation of being smaller in disturbance classes corresponding to higher expectation of harvesting. However, the goal of the model selection is prediction so both interactions were left in. It should be noted that there is a new forest stratum imbedded in the REGION:DIST variable that corresponds to plots that do not contain any information about previous measurement. If zero is imputed for $PREV\_VOLUME$ in these strata, the design matrix will be singular and the model will be unable to be calculated. A simple workaround to circumvent this computational issue is to arbitrarily impute a value of 1 if the ID number of the plot is even and 0 if it is odd. This negligibly affects the residuals of the model and allows the model to be computed.

It should also be noted that the growth update models unfortunately did not significantly improve the correlations within region between ground truth observed in Swiss NFI 3 and the update from Swiss NFI 2. This is because the time between remeasurements were all between 9 and 14 vegetative cycles. When the phase-in of the annual design is complete, the potential vegetative cycles range from 1 to 9. However, during the phase-in the maximum number of cycles is potentially as much as 15. Thus, the available Swiss NFI data was not adequate to test the efficacy of the growth models in these situations. The REG automatically adjusts for average growth (both tree growth and in-growth) automatically when the regression coefficient for PREV_VOLUME is more than 1. As a result, only one of the estimators presented here contains PREV_VOLUME_UPDATED.

3.9 Estimators considered

Nine estimators were considered in order to assess the efficiency of the three-phase REG and its update components. All estimators are presented in detail in Table 3.3. The sample sizes for each phase by region are found in Table 3.1.

Table 3.3: Estimates and models considered

	$\mathbf{Z}^{(1)t}(x)$	$\mathbf{Z}^{(2)t}(x)$
Estimate 1	REGION	...
Estimate 2	REGION	...
Estimate 3	MCH CONPROP INFORESTPROP REGION:DIST MCH:INFORESTPROP MCH:REGION:DIST	...
Estimate 4	...	PREV_VOLUME PREV_VOLUME:REGION
Estimate 5	...	PREV_VOLUME_UPDATED PREV_VOLUME_UPDATED:REGION
Estimate 6	MCH CONPROP INFORESTPROP REGION:DIST MCH:INFORESTPROP MCH:REGION:DIST	PREV_VOLUME PREV_VOLUME:REGION:DIST
Estimate 7	MCH CONPROP INFORESTPROP REGION:DIST MCH:INFORESTPROP MCH:REGION:DIST	PREV_VOLUME_UPDATED PREV_VOLUME_UPDATED:REGION:DIST
Estimate 8	MCH CONPROP INFORESTPROP STRATUM MCH:INFORESTPROP MCH:REGION	PREV_VOLUME PREV_VOLUME:REGION
Estimate 9	MCH CONPROP INFORESTPROP STRATUM MCH:INFORESTPROP MCH:REGION	PREV_VOLUME_UPDATED PREV_VOLUME_UPDATED:REGION

Legend:
: indicates a variable interaction term

Estimate 1 and 2 are based on simple random sampling without any model (i.e. mean of the plot volumes) for each region. Double sampling for poststratification using REGION as the stratification variable was used for the estimate across entire Switzerland, because it more closely corresponds to the current estimation procedure implemented in the Swiss NFI which stratifies by region. Estimate 1 is the estimator used in the Swiss NFI under the old periodic design and takes into account all forested plots in the entire terrestrial sample of NFI 3. Estimate 2 only considers forested plots corresponding to the first three years of annual subsamples in NFI 4.

Comparing Estimate 2 to Estimate 1 gives insight into the loss in precision associated with changing from a periodic to a continuous design when the standard SRS estimator is used to produce an estimate for a 3 year time period (note that the periodic Swiss NFI took three years to evaluate).

Estimates 3, 4 and 5 utilize the two-phase REG. Estimate 3 removes the second phase altogether and considers a model using only remote sensing variables in the auxiliary vector. This is the effect of ignoring the previous measurement altogether. Estimate 4 removes the first phase and only considers the previous measurement and region information at the second phase without the growth update. Estimate 5 does the same as estimate 4 but incorporates the growth update. Estimates 4 and 5 demonstrate the effect of ignoring the remote sensing data.

Estimates[2] 6, 7, 8 and 9 utilize the three-phase REG. Estimate 6 and 7 incorporate the selected model, $M_{reduced}$, and M_{full}, where the variables PREV_VOLUME and PREV_VOLUME_UPDATED are used, respectively. Estimate 7 shows the effect of the growth update in the proposed three-phase REG set-up. Finally, estimates 8 and 9 are the same as 6 and 7, respectively, except the disturbance classes are removed. However, the ”new forest” stratum is still included and is represented by dummy variable STRATUM. These estimates give insight into the efficacy of incorporating the proposed disturbance update.

3.10 Results

The estimates with their corresponding variance estimates are found in Table 3.4. The estimates are all well within 2 standard deviations of the full NFI 3 estimate that contains

[2]For Estimates 6-9 in the Prealps (PA) there was a computational singularity in the model fit due to multicollinearity associated with the interaction term between the previous measurement and the ”new forest” stratum. Thus, for estimates in the Prealps the interaction terms are dropped in $\mathbf{Z}^{(2)t}(x)$.

the full sample, which is expected given their unbiasedness. As anticipated in large samples, the differences between the two variance estimates under the external model assumption (not shown) and the variances derived using the internal model are small for each region and empirically non-existent for Switzerland as a whole (CH). In simulations the external variance was usually smaller, however here it was slightly larger for Jura (JU) and Plateau (SP).

The improvements to the standard errors should be compared to the standard estimate (i.e. Estimate 2) that only contains the first three years of annual sample to assess the gain in efficiency. The 2-phase REG that uses only remote sensing data (Estimate 3) shows a substantial improvement in variance across all regions and for CH, but not as much of an improvement as using only the previous measurement (Estimates 4 and 5) (note that for CH this also includes a simple stratification by region). The only exception was in the SP where the adjusted R^2 for Estimate 4 was the lowest (0.39) among all regions. The likely explanation is that despite the fact that the adjusted R^2 was higher than in Estimate 3 (0.34), the effect of the larger first-phase sample size in Estimate 3 was great enough to produce a lower standard error. The growth update offered only a slight improvement in Estimate 5 over Estimate 4.

Estimates 6, 7, 8 and 9 using the three-phase REG show a further decrease in standard error compared to estimates 3, 4 and 5. When the disturbance correction was removed in Estimates 8 and 9, there was a slight increase in variance (note that there is no disturbance correction in South Alps so it always remains exactly the same). Disturbance effects account for only a small proportion of the overall growing stock, so this is to be expected. As expected, there is a greater reduction in standard error in regions with heavier harvesting such as SP, and JU. The Alps (AL) and South Alps (SA) have lighter harvesting which is demonstrated by the negligible effect of their disturbance updates. Regions with fewer disturbances showed greater reductions in standard error compared to Estimate 2, when the previous measurement was included in the model.

The growth update's effect when included in the three-phase REGs was very modest as evidenced by only slight decreases in standard error in most cases. The only exception was in the Prealps where the update had a slightly adverse effect, however, in the Prealps the interaction terms were dropped from $\mathbf{Z}^{(2)t}(x)$ for Estimates 6-9 due to computational singularities in the model fit. In most cases, growth updating using only tree-level growth models provided a slight improvement during the estimation process. We note that the regression estimator internally accounts for part of the overall growth automatically, when the coefficient for the previous measurement variable is greater than one. In this case it seems that the tree-level growth update did not bring much

Table 3.4: Point Estimates and Standard Errors, in $()$, for Timber Volume in $\frac{m^3}{ha}$ and Adjusted R^2's, in [ ] for $M_{reduced}$ and M_{full} respectively.

3rd Swiss National Forest Inventory

		JU	SP	PA	AL	SA	CH
1-phase	Estimate 1	364.35 (6.51) [...]	392.30 (7.81) [...]	442.36 (9.12) [...]	308.54 (5.41) [...]	228.61 (5.99) [...]	348.95 (3.32) [0.03]
	Estimate 2	368.10 (11.30) [...]	397.39 (13.64) [...]	440.01 (15.33) [...]	309.56 (9.21) [...]	235.32 (10.88) [...]	351.41 (5.69) [0.03]
2-phase	Estimate 3	368.54 (9.64) [0.28]	397.47 (11.29) [0.34]	437.21 (13.51) [0.28]	309.32 (7.85) [0.34]	231.97 (8.89) [0.38]	350.05 (4.66) [0.36]
	Estimate 4	367.16 (9.17) [0.53]	392.81 (11.68) [0.39]	440.10 (12.80) [0.49]	310.83 (6.87) [0.72]	227.82 (7.08) [0.86]	349.74 (4.43) [0.59]
	Estimate 5	365.28 (9.08) [0.54]	392.57 (11.56) [0.41]	440.72 (12.80) [0.49]	311.08 (6.85) [0.72]	229.13 (6.97) [0.89]	349.73 (4.41) [0.60]
3-phase	Estimate 6	369.23 (8.11) [0.28; 0.60]	394.03 (10.23) [0.34; 0.52]	439.41 (12.14) [0.28; 0.52]	311.62 (6.25) [0.34; 0.75]	227.95 (6.51) [0.38; 0.87]	350.49 (4.01) [0.36; 0.65]
	Estimate 7	367.85 (7.98) [0.28; 0.62]	395.24 (10.16) [0.34; 0.53]	440.48 (12.20) [0.28; 0.52]	311.95 (6.24) [0.34; 0.75]	229.35 (6.41) [0.38; 0.89]	350.92 (4.00) [0.36; 0.65]
	Estimate 8	367.06 (8.33) [0.28; 0.58]	393.20 (10.49) [0.34; 0.47]	439.93 (12.17) [0.28; 0.51]	310.71 (6.40) [0.33; 0.74]	227.95 (6.51) [0.38; 0.87]	349.91 (4.10) [0.34; 0.62]
	Estimate 9	365.46 (8.22) [0.28; 0.59]	392.85 (10.41) [0.34; 0.49]	440.84 (12.24) [0.28; 0.51]	310.95 (6.39) [0.33; 0.74]	229.35 (6.41) [0.38; 0.89]	349.91 (4.09) [0.34; 0.63]

Legend:
Jura (JU), Plateau (SP), Prealps (PA), Alps (AL), South Alps (SA), Switzerland (CH)

improvement. Further improvement might be possible if in-growth was accounted for.

3.11 Conclusions

The 2-phase REG and 3-phase REG show clear usefulness in integrating remote sensing and past inventory data under an annual design. Furthermore, their implementation into statistical software is simplified by their close connection to classical linear regression. The estimation procedure where remote sensing is used in the first phase and the previous inventory data is included in the second phase offers substantial reduction in the variance compared to using only the current annual subsample, with the added benefit of virtually no tradeoff in bias. The updates for both growth and disturbance of the previous measurement variable provided only a modest improvement in standard error. Regardless, these estimators and this estimation scheme can clearly be recommended for use in regional and national forest estimation for growing stock volume. Their generality and flexibility allows for implementation with a wide variety of data sources including liDAR, satellite images, and high-resolution digital aerial photographs. The inclusion of a large first-phase sample as opposed to wall-to-wall data provides for great gains in computational efficiency and thus is an attractive alternative for large-scale national forest inventories.

Chapter 4

Comparing parametric and nonparametric models under the external model approach

Recall from Chapter 2.3.1 that generalized difference estimators are design-unbiased and possess analytical design-unbiased variance estimators for any choice of model provided that the model is external, i.e. fitted using a training set that is independent of the sample realized by the current inventory's design. Examples include a training set derived from a past inventory or a separate inventory conducted in another country or region. However, as it is rarely practical or economically efficient to fit using a separate inventory, a common work-around is to use the external model approach, i.e. to fit the model internally from the current inventory at hand and use the variance estimator derived for external models as if this was not the case.

In some cases it can be acceptable to treat internal models as external, as with the classical regression estimator, provided that the sample size is large enough. In the classical regression case we are mathematically reassured by asymptotic equivalency to a known consistent design-based variance estimator derived using the g-weight technique. The Monte Carlo version of the calibration estimators of Wu and Sitter (2001) shows that the external variance is asymptotically valid also for non-linear regression models (Mandallaz and Massey (2015), Appendix C). However, analogous reassurances do not exist for the family of nonparametric kernel-based methods, to which k-Nearest Neighbors (kNN) is a special case, because no practical analytical design-based variance estimator is currently known in the literature for internally fitted models. At present we are limited to evaluating the design-based properties of most kernel-based regression estimators by empirical investigation.

The scope of this chapter is to test the mathematical validity of treating internal models as external under moderate sample size for a variety of kernel-based regression estimators using empirical examples arising from a case study of a Swiss forest inventory in the Canton of Grisons as well as a rigorous simulation example. A variety of bandwidth selection algorithms and two main metrics are considered: one based on the Euclidean distance in a multidimensional space of explanatory variables and the other based on the one-dimensional Euclidean distance between predictions from a linear regression model. New nonparametric estimators are proposed, which are based on the conditional expectation of the response variable given its linear regression prediction, in an effort to address a known drawback of many kernel-regression estimators, namely the curse of dimensionality and that the predictions are bounded by the range of observed response variables. The underlying theoretical concepts of kernel-based estimators are summarized and reformulated in the design-based Monte Carlo approach to forest inventory. Finally, variance estimates derived under the external model assumption are compared with a nonparametric bootstrap and a simulation example is given where all variance estimates can be compared to the true design-based variance of the estimator for that simulation scenario. For simplicity, only two-phases are considered.

A shorter version of this chapter has been accepted for publication in the *Canadian Journal of Forest Research* (see Massey and Mandallaz (2015a)).

4.1 Recap of two-phase estimation

We are working under the same sampling design considered in Chapter 2.1 and employ the same notation for consistency. Thus s_2 represents the first phase sample selection while s_3 represents the second phase selection. We shall consider generalized difference estimators of the form

$$\hat{Y}_{GD,2p} = \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}_{M_2}(x) + \frac{1}{n_3}\sum_{x\in s_3}(Y(x) - \hat{Y}_{M_2}(x))$$

where the $R_{M_2}(x) = Y(x) - \hat{Y}_{M_2}(x)$ are the theoretical residuals, which generally do not have zero mean for external models. For linear and nonlinear prediction models, $\hat{Y}_{GD,2p}$ can be viewed in the Monte Carlo approach as a special case of the calibration estimator discussed by Wu and Sitter (2001) (see appendix of Mandallaz and Massey (2015) for an adaptation to the infinite population approach and proof). Under external model approach $\hat{Y}_{M_2}$ is an exactly design-unbiased estimator for $\bar{Y}$ and possesses the

following design-unbiased variance estimator obtained by sample copy of [2.12]

$$\hat{\mathbb{V}}(\hat{Y}_{GD,2p}) = \frac{1}{n_2}\frac{\sum_{x\in s_3}(Y(x)-\bar{Y}_{s_3})^2}{n_3-1} + \big(1-\frac{n_3}{n_2}\big)\frac{1}{n_3}\frac{\sum_{x\in s_3}(R_{M_2}(x)-\bar{R}_{s_3})^2}{n_3-1} \qquad [4.1]$$

where $\bar{Y}_{s_3} = \frac{1}{n_3}\sum_{x\in s_3} Y(x)$ and $\bar{R}_{s_3} = \frac{1}{n_3}\sum_{x\in s_3} R_{M_2}(x)$.

The two-phase classical regression $\hat{Y}_{REG,2p}$ occurs when we consider M_2 to be

$$Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R(x) \qquad [4.2]$$

and plug-in $\hat{Y}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}$ and $\hat{R}(x) = Y(x) - \hat{Y}(x)$ for $\hat{Y}_{M_2}(x)$ and $R_{M_2}(x)$ respectively. Although the g-weight variance described in [2.29] has slightly better statistical properties from a theoretical point of view, analogous versions are unfortunately not available for the nonparametric estimators discussed in the following sections. For this reason we will restrict ourselves to external variance estimates (i.e. [4.1]) for comparability.

4.2 Kernel-based regression estimators

There is a huge amount of literature on kernel-based estimators but the overwhelming majority of the papers are in the model-dependent one-dimensional framework. Here only the main results required for practical work are given. The interested reader is encouraged to consult Györfi et al. (2002) for general theory in nonparametric regression and the online technical report Mandallaz and Massey (2015) for informal proofs and further examples with alternative numerical options relevant to this chapter.

We will use primarily one-dimensional kernels of the form $K : u \in \mathbb{R} \to K(u) > 0$. The kernel $K()$ is a probability density function with zero mean and finite variance, i.e $\int_{\mathbb{R}} K(u)du = 1$, $\int_{\mathbb{R}} uK(u)du = 0$ and $\int_{\mathbb{R}} u^2K(u)du < \infty$. Popular choices are

$$\begin{aligned} K(u) &= 0.5I_{[-1,1]}(u) \quad \text{uniform kernel} \\ K(u) &= \frac{1}{\sqrt{2\pi}}\exp(-u^2/2) \quad \text{normal kernel} \\ K(u) &= \frac{3}{4\sqrt{5}}(1-\frac{u^2}{5})I_{|u|\le\sqrt{5}}(u) \quad \text{Epanechnikov kernel} \end{aligned} \qquad [4.3]$$

The first and third kernels have a finite support whereas the normal kernel does not. The concept can be easily generalized to a multidimensional kernel $K(\boldsymbol{u})$ with $\boldsymbol{u} = (u_1, u_2, \dots u_p) \in \mathbb{R}^p$, by setting $K(\boldsymbol{u}) = \prod_{l=1}^{p} K_l(u_l)$ where the $K_l(\cdot)$ are one-dimensional kernels.

A standard theoretical point of departure is to consider the conditional expectation of $Y(x)$ given the auxiliary information, i.e. $\mathbb{E}(Y(x) \mid \boldsymbol{Z}(x) \in \mathbb{R}^p)$ which requires the use of multivariate kernels to estimate multivariate densities and is subject to the curse of dimensionality problem when p is large (Lehmann (1999), p.419-420, for a dramatic example). For simplicity and in an effort to escape the curse of dimensionality, we propose using $\hat{Y}(x) \in \mathbb{R}^1$ as a concise one-dimension summary of $\boldsymbol{Z}(x) \in \mathbb{R}^p$ which leads us to consider

$$\begin{aligned} \mathbb{E}(Y(x) \mid \hat{Y}(x)) &= \mathbb{E}(\hat{Y}(x) + \hat{R}(x) \mid \hat{Y}(x)) & [4.4] \\ &= \hat{Y}(x) + \mathbb{E}(\hat{R}(x) \mid \hat{Y}(x)) & [4.5] \end{aligned}$$

To have an approximation of this conditional expectation we need estimates of the joint bivariate density of $(Y(x), \hat{Y}(x))$ or $(\hat{R}(x), \hat{Y}(x))$ and of the marginal density of $\hat{Y}(x)$. This is done with bivariate and univariate Nadaraya-Watson density estimators (essentially smoothed versions of histograms) to obtain first an estimate of the conditional density and then, by integration, the estimate of the conditional expectation (see Mandallaz and Massey (2015) for details). Estimates will be defined point-wise at any arbitrary point $x_0 \in s_1$ where for notational simplicity we denote $\hat{y}_0 := \hat{Y}(x_0)$ when needed. Using the first form (i.e. $\mathbb{E}(Y(x) \mid \hat{Y}(x))$) we are led to the estimator

$$\hat{Y}^{(1)}_{\epsilon}(x_0) := \hat{\mathbb{E}}(Y(x_0) \mid \hat{Y}(x_0)) := \frac{\frac{1}{n_3 \epsilon(n_3, \hat{y}_0)} \sum_{x \in s_3} Y(x) K\big(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon(n_3, \hat{y}_0)}\big)}{\frac{1}{n_3 \epsilon(n_3, \hat{y}_0)} \sum_{x \in s_3} K\big(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon(n_3, \hat{y}_0)}\big)} \qquad [4.6]$$

whereas the second form, $\hat{Y}(x) + \mathbb{E}(R(x) \mid \hat{Y}(x))$, leads to

$$\hat{Y}^{(2)}_{\epsilon}(x_0) = \hat{Y}(x_0) + \hat{R}^{(2)}_{\epsilon, smooth}(x_0) \qquad [4.7]$$

and

$$\hat{R}^{(2)}_{\epsilon, smooth}(x_0) := \hat{\mathbb{E}}(\hat{R}(x_0) \mid \hat{Y}(x_0)) \quad := \quad \frac{\frac{1}{n_3 \epsilon_{\hat{y}}(n_3, \hat{y}_0)} \sum_{x \in s_3} \hat{R}(x) K\big(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon_{\hat{y}}(n_3, \hat{y}_0)}\big)}{\frac{1}{n_3 \epsilon_{\hat{y}}(n_3, \hat{y}_0)} \sum_{x \in s_3} K\big(\frac{\hat{Y}(x_0) - \hat{Y}(x)}{\epsilon_{\hat{y}}(n_3, \hat{y}_0)}\big)} \qquad [4.8]$$

The Nadaraya-Watson regression estimator is simply a weighted average where the weights are calculated based on a kernel and the fact that it can be derived using the conditional expectation lends it legitimacy because the conditional expectation is the best prediction, where ”best” means minimum mean squared error. The choice of the kernel is of less importance than the bandwidth, which is denoted $\epsilon_{\hat{y}}(n_3, \hat{y}_0)$ (see e.g. Lehmann (1999)). The bandwidth is the tuning parameter that controls how smooth the predictions will be across various $\hat{Y}(x)$ (recall that $\hat{Y}(x)$ in this context is considered to be the auxiliary variable). Although the bandwidth can be arbitrarily fixed, it

makes more sense to choose an optimal bandwidth based on some prediction criterion (e.g. asymptotic integrated mean-square error). Hence, in our notation the bandwidth, $\epsilon_{\hat{y}}(n_3, \hat{y}_0)$, depends on n_3 and $\hat{y}_0$. Global bandwidths depend only on n_3 while locally varying bandwidths are allowed to expand for sparsely populated areas of the auxiliary space in the reference set and thus depend on both n_3 and $\hat{y}_0$. In the context of this presentation, we found that locally varying bandwidths lead to spurious behavior near the boundaries and very unstable bootstrap variance estimators (see Mandallaz and Massey (2015) for examples). For this reason, only global bandwidth selection strategies will be considered here.

We can now define the two-phase kernel-based regression estimators as

$$\hat{Y}_{\epsilon}^{(l)} := \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}_{\epsilon}^{(l)}(x) + \frac{1}{n_3}\sum_{x\in s_3}\hat{R}_{\epsilon}^{(l)}(x),\ l = 1, 2 \qquad [4.9]$$

where $\hat{R}_{\epsilon}^{(l)}(x) = Y(x) - \hat{Y}_{\epsilon}^{(l)}(x)$ for $l = 1, 2$. Note that the sum of the residuals is no longer zero in general. In the absence of more tractable analytical methods to derive the true design-based expectation and variance of [4.9], one conjectures that the design-unbiasedness of $\hat{Y}_{\epsilon}^{(l)}$ and [4.1], which is guaranteed for any external model, still holds at least asymptotically for internal $\hat{Y}_{\epsilon}^{(l)}$. In other words, one hopes that an analogy with the classical regression estimator exists where $\hat{Y}_{\epsilon}^{(l)}$ is asymptotically design unbiased with the asymptotically design-unbiased variance estimate

$$\hat{\mathbb{V}}(\hat{Y}_{\epsilon}^{(l)}) = \frac{1}{n_2}\hat{\mathbb{V}}(Y(x)) + (1 - \frac{n_3}{n_2})\frac{1}{n_3(n_3-1)}\sum_{x\in s_3}(\hat{R}_{\epsilon}^{(l)}(x) - \bar{\hat{R}}_{s_3}^{(l)})^2 \qquad [4.10]$$

for $l = 1, 2$ where $\hat{\mathbb{V}}(Y(x)) = \frac{1}{n_3-1}\sum_{x\in s_3}(Y(x) - \bar{Y}_{s_3})^2$, $\bar{Y}_{s_3} = \frac{1}{n_3}\sum_{x\in s_3} Y(x)$ and $\bar{\hat{R}}_{s_3}^{(l)} = \frac{1}{n_3}\sum_{x\in s_3}\hat{R}_{\epsilon}^{(l)}(x)$.

Remarks:

- There are some hidden mathematical difficulties in the previous arguments. Under a true external approach with a fixed $\boldsymbol{\beta}$ the $\hat{Y}(x)$ are i.i.d under the design, but the bivariate distribution of $(Y(x), \hat{Y}(x))$ or $(\hat{R}(x), \hat{Y}(x))$ is degenerate as all the realizations lie on a one-dimensional curve in $\mathbb{R}^2$. However, with internal models we use $\hat{\boldsymbol{\beta}}$ and this is no longer true. The $\hat{Y}(x)$'s have a correlation of order n_3^{-1}. Note that in contrast to the standard notation in the literature on non-parametric regression, $\hat{Y}(x)$ plays formally the role of the independent variable, assumed to be fixed or i.i.d. distributed on $\mathbb{R}$, and $Y(x)$ plays the role of the response variable. In a purely pragmatic approach we dismiss the correlation of the $\hat{Y}(x)$'s as being asymptotically negligible and compare empirically the estimators based on [4.6]

or [4.7]. The true theoretical mathematical properties are certainly much more difficult to derive than in the standard set-up.

- Note that to ensure consistency at x_0 one must have $\epsilon(n_3) \to 0$ and $n_3\epsilon(n_3) \to \infty$ as $n_3 \to \infty$ in the standard set-up, and intuitively also in our case since the correlation of the $\hat{Y}(x)$ is asymptotically negligible (see Mandallaz and Massey (2015)).

- If the linear model is given by a post-stratification (i.e. $\boldsymbol{Z}(x)$ is a vector of strata indicator variables), then $\hat{Y}(x)$ has a discrete distribution taking only the values of the strata means. One can check that both estimators $\hat{Y}_\epsilon^{(1)}(x_0)$ and $\hat{Y}_\epsilon^{(2)}(x_0)$ will tend to $\hat{Y}(x_0)$, the stratum mean, for $\epsilon_{\hat{y}}(n_3, \hat{y}_0)$ small enough. If the $\hat{Y}(x)$ have a continuous density function then, for $x_0 \in s_3$, both estimators will tend to $Y(x_0)$ (little smoothing and wiggled regression curve). If the bandwidth goes to infinity (strong smoothing), then $\hat{Y}_\epsilon^{(1)}(x_0)$ tends to the sample mean (the model is completely ignored) and $\hat{Y}_\epsilon^{(2)}(x_0)$ goes to $\hat{Y}(x_0)$ and nothing is gained compared to the the classical regression estimator.

- $\hat{Y}_\epsilon^{(1)}(x_0)$ is a convex weighted mean (positive weights summing up to 1) so that it cannot escape the range of observations $Y(x)$, whereas $\hat{Y}_\epsilon^{(2)}(x_0)$ can. There exist in the literature other non-parametric estimators using differently defined kernels that may lead to non-convex linear combinations of the observations and thus extrapolations. This is the case for the frequently used estimators proposed and discussed in Jennen-Steinmetz and Gasser (1988) and Gasser and Mueller (1984). Allowing the predictions to escape the range of observations can be considered an advantage because difficulty in prediction at the boundary of the auxiliary space (a.k.a. feature space) is a known drawback of many kernel-regression estimators.

4.3 Design-based kNN

As for kernel-based estimators the literature on k-Nearest Neighbors (kNN) is immense and largely model-dependent. In the forest inventory context Bafetta et al. (2009, 2011) are key references in the design-based context whereas Tomppo et al. (2008), Moeur and Stage (1996), McRoberts et al. (2007), McRoberts et al. (2011), Magnussen and Tomppo (2014), McRoberts (2012), Breidenbach and Nothdurft (2010) are more model-dependent. Also, most authors work in the finite population framework of pixels in which the first phase (denoted s_2 in the notation here) is exhaustive and unequal inclusion

probabilities on the plot level are possible. In the Monte Carlo approach, we must assume uniform i.i.d. sampling for all $x \in F$ because the notion of an inclusion probability of a given point x_0 is meaningless.

We shall use here the procedure described by Hechenbichler and Schliep (2004). For an arbitrary point $x_0 \in s_2$ the nearest neighbor in s_3, with respect to a distance $d(\cdot,\cdot)$ in $\mathbb{R}^p$ is the point $x_{(1)} \in s_3$ such that $d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(1)})) = \min_{x \in s_3} d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x))$. The second nearest neighbor $x_{(2)}$ is defined by $d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(2)})) = \min_{x \in s_3 \setminus x_{(1)}} d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x))$ and so on until we have obtained the $k+1$ nearest neighbors. We emphasize the fact when using an internal kNN model the reference set is s_3. Thus, if $x_0 \in s_3$ then we set $x_{(1)} = x_0$ (i.e. x_0 counts as its own closest neighbor). The simulations and case study presented in the following section contain only continuous variables so that distance ties among neighbors do not occur. The kNN methods with the uniform kernel simply take the mean, e.g. $\hat{Y}_{knn}(x_0) = \frac{1}{k}\sum_{i=1}^{k} Y(x_{(i)})$. In this particular case it is possible, in principle, to calculate the theoretical design-based variance of the corresponding regression estimator with an exhaustive first phase for finite populations (see Bafetta et al. (2009)). However, the uniform kernel is usually slightly less efficient than it is to give larger weights to the closer neighbors. Since explanatory variables may have different scales, each auxiliary component should be standardized. Furthermore, we should standardize the distances themselves to avoid them becoming smaller as the sample size increases. Here this is done according to

$$D\big(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})\big) = \frac{d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)}))}{d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(k+1)}))} \qquad [4.11]$$

and the weighted kNN estimator is given by

$$\hat{Y}_{knn}(x_0) = \frac{\sum_{i=1}^{k} K\big(D(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)}))\big) Y(x_{(i)})}{\sum_{i=1}^{k} K\big(D(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)}))\big)} \qquad [4.12]$$

for an arbitrary one-dimensional kernel $K(\cdot)$ (Note that the kernel is one-dimensional whereas the distance is calculated in the multidimensional feature space). This leads to the two-phase **kNN regression estimator**

$$\hat{Y}_{knn} = \frac{1}{n_2}\sum_{x \in s_2} \hat{Y}_{knn}(x) + \frac{1}{n_3}\sum_{x \in s_3} (Y(x) - \hat{Y}_{knn}(x)) \qquad [4.13]$$

Note that the predictions, $\hat{Y}_{knn}(x)$, are a weighted mean of the observations and therefore constrained to be in the observed range of the observations. Thus kNN is also a purely interpolation technique and never extrapolates, which, as already mentioned, can be a disadvantage especially when n_2 is large and n_3 is relatively small (the typical case).

As was the case for the other kernel-based regression estimators, one can use the one-dimensional metric based on the predictions alone, i.e.

$$d(\boldsymbol{Z}(x_0), \boldsymbol{Z}(x_{(i)})) = \mid \hat{\boldsymbol{\beta}}^t \big(\boldsymbol{Z}(x_0) - \boldsymbol{Z}(x_{(i)})\big) \mid$$

The resulting estimator is denoted by $\hat{Y}_{pred,knn}$ and should give us some insight into how much information is lost by using $\hat{Y}(x)$ to reduce the dimension of $\boldsymbol{Z}(x)$. Recall that for the other kernel-regression estimators the dimension reduction was necessitated by the curse of dimensionality in conjunction with a multidimensional kernel. By construction for kNN we only require a one-dimensional kernel based on distances to neighbors in the feature space. The external variance estimates are obtained as usual via [4.1] where $R_{M_2}(x) = Y(x) - \hat{Y}_{knn}(x)$ and $R_{M_2}(x) = Y(x) - \hat{Y}_{pred,knn}(x)$ respectively.

In practice one has to be careful about the tuning options available in software packages, in particular with respect to the definition of nearest neighbors when $x_0 \in s_3$. If the point itself is viewed as its nearest neighbor the optimal choices $k = 1$ can result from cross-validation and should of course not be retained as it leads to constant 0 residuals. We recommend at least $k \geq 3$ and to plot $\hat{Y}_{knn}$ and $\hat{\mathbb{V}}(\hat{Y}_{knn})$ as a function of k. In our experience, except for $\hat{Y}_{reg}$, the external variance estimate is likely too small and the bootstrap variance estimate should be preferred as will be demonstrated in Section 4.4.

4.4 Examples

To illustrate the various techniques we consider the real case study discussed in Mandallaz (2013a) and the artificial simulation example used in Mandallaz (2013b). We have compared several kernel-based estimators. The Nadaraya-Watson kernel-based estimators, [4.6] and [4.7], were calculated using the `npreg()` function from the package *np* (Hayfield and Racine (2008)) and the kNN estimators of Section 4.3 were calculated with `kknn()` from the **R**-package *kknn* (Schliep and Hechenbichler (2014)). It should be noted that `kknn()` standardizes all feature space variables by default. For the purpose of comparison, estimators discussed by Jennen-Steinmetz and Gasser (1988) and Gasser and Mueller (1984) are also considered due to their popularity and general availability for a variety of statistical software. Thus, the widely used **R**-functions `glkern()` and `lokern()` from the package *lokern* were used (for documentation see Herrmann and Maechler (2014)). We shall not discuss here the function `lokern()` because it was found that the local optimal bandwidth led to very spurious results near the

boundaries, particularly at the lower end, and highly unstable bootstrap estimates (see Mandallaz and Massey (2015) for details).

$\hat{Y}^{(1)}_{glkern}$ and $\hat{Y}^{(1)}_{npreg}$ were calculated using `glkern()` and `npreg()` respectively to calculate [4.6] while allowing the imbedded cross-validation procedures select the bandwidths. For $\hat{Y}^{(2)}_{glkern}$, `glkern()` was applied to the residual part only and the optimal bandwidth was obtained via the imbedded cross-validation procedure. For $\hat{Y}^{(2)}_{npreg}$ the optimal bandwidth obtained via the imbedded leave-one-out cross-validation procedure and applied to only the residual part led essentially to a flat horizontal line close to 0 (i.e. making $\hat{Y}^{(2)}_{npreg} \approx \hat{Y}_{reg}$). For this reason we wrote our own cross-validation procedure based on the mean squared error criteria by sequentially deleting each $x_0 \in s_3$ and refitting both $\hat{Y}(x_0)$ using the linear model as well as the smoothed residual part (using `npreg()` on the second term) to get the optimal bandwidth out of a small set of feasible values. The resulting $\hat{Y}^{(2)}_{npreg}$ then differs from $\hat{Y}_{reg}$. The same procedure was used for the simulation presented in Section 4.4.2.

The bootstrap was implemented as explained in Section 2.1 where each bootstrap sample was obtained using SRS with replacement over the s_2 (recall that $s_1 = s_2$ here because only the two-phase case is presented) from the original sample. The regression coefficients used to calculate $\hat{Y}(x)$ were recalculated for each bootstrap replicate. As for bandwidth selection, it was found that when the optimal bandwidth was recalculated in each bootstrap sample the resulting bootstrap variance estimates were very unstable due to unacceptably extreme outliers, sometimes above 1000 times the mean (examples found in Mandallaz and Massey (2015)). This was also the case, but not as extreme, for the case study. As a result, to stabilize the bootstrap variances, the optimal global bandwidth obtained from the original full sample was used in all the bootstrap samples. Likewise, the optimal choices of k in the kNN estimators were only calculated from the original samples. This was not due to instability in the bootstrap variance estimates but rather because the imbedded cross-validation routine in `kknn()` allowed for $k = 1$ for a significant proportion of the replicates, which as already discussed leads to constant zero residuals in s_3. The double sampling bootstrap routine that fixes n_2 and n_3 as discussed in Wolter (2007) was also implemented for the case study, but since it seemed to produce bootstrap variance estimates that only differed negligibly and was somewhat more involved computationally to implement, it was dropped in favor of the simpler routine that only fixed n_2. A sample of the **R**-code used for the simulation can be found in Appendix E of Mandallaz and Massey (2015).

4.4.1 Case study

The auxiliary vector $\boldsymbol{Z}(x)$ has 7 components: $Z_0(x) \equiv 1$, mean canopy height $Z_1(x)$, maximal canopy height $Z_2(x)$, 75% quantile of the canopy height, $Z_3(x)$, standard deviation of canopy height $Z_4(x)$, the LiDAR estimated volume density $Z_5(x)$ and the LiDAR estimated density of stems $Z_6(x)$. The sample sizes were $n_2 = 306, n_3 = 67$. The fitted model is

$$\begin{aligned} \hat{Y}(x) \quad &= \\ 322.57 \quad &+52.55Z_1(x) - 19.24Z_2(x) - 33.04Z_3(x) + 71.06Z_4(x) + 0.19Z_5(x) - 0.09Z_6(x) \end{aligned}$$

and the coefficient of determination is $R^2 = 0.64$.

$\hat{Y}_{knn}$ is the kNN estimator using `kknn()` with the Euclidean distance applied to a 6-dimensional feature space and a gaussian kernel. We chose $k = 7$, which is the optimal value obtained from the leave-one-out cross-validation procedure (further justifications for this choice are given in Mandallaz and Massey (2015)). $\hat{Y}_{pred,knn}$ is defined likewise except that its support is one-dimensional because it is based on the Euclidean distance between the linear predictions, $\hat{Y}(x)$. With the gaussian kernel it was decided to set $k = 9$ based on cross-validation.

Table 4.1: Global estimation of timber volume ($\frac{m^3}{ha}$) in Canton Grisons ($n_2 = 306$; $n_3 = 67$)

Estimator	point estimate	external s.e.	bootstrap s.e.
$\bar{Y}_F$	399.43	23.82	23.90
$\hat{Y}_{reg}$	384.95	16.52	18.34
$\hat{Y}^{(1)}_{glkern}$	389.89	16.08	19.36
$\hat{Y}^{(1)}_{npreg}$	388.58	16.08	17.79
$\hat{Y}^{(2)}_{glkern}$	389.79	16.08	19.41
$\hat{Y}^{(2)}_{npreg}$	387.51	16.28	18.55
$\hat{Y}_{knn}$	390.53	16.35	17.83
$\hat{Y}_{pred,knn}$	389.22	15.69	17.42

Table 4.1 displays the results. The differences between the point estimates and the sample mean are not statistically significant, which is reassuring. However, the bootstrap standard errors are roughly 15% larger than the external standard errors which is our first indication that the external variances could be underestimating their respective theoretical variances. However, for the case study this is only conjectured as the truth is not known. It should be noted that in a similar case study using the same data set

(see Mandallaz (2013a)), the g-weight variance for $\hat{Y}_{reg}$ was found to be between the external and bootstrap standard errors calculated here, as intuition would suggest.

Table 4.2: Small area estimation of timber volume ($\frac{m^3}{ha}$) in Canton Grisons

small area	G_1	G_2	G_3	G_4
$n_1 : n_2$	94:19	81:17	66:15	65:16
$\bar{Y}_G$	410.40	461.44	318.00	396.85
	(44.58)	(56.35)	(34.36)	(47.86)
	[44.17]	[55.37]	[33.93]	[47.88]
$\hat{Y}_{reg}$	397.27	426.81	327.64	366.44
	(28.80)	(35.01)	(31.64)	(36.01)
	[31.61]	[34.93]	[32.06]	[35.53]
$\hat{Y}^{(1)}_{glkern}$	405.60	427.16	340.50	364.63
	(28.82)	(34.63)	(29.65)	(34.65)
	[34.40]	[34.30]	[30.80]	[34.76]
$\hat{Y}^{(1)}_{npreg}$	397.36	428.37	341.08	369.36
	(28.65)	(35.11)	(29.17)	(34.76)
	[30.72]	[33.27]	[29.21]	[33.67]
$\hat{Y}^{(2)}_{glkern}$	405.46	426.94	340.58	364.59
	(28.81)	(34.62)	(29.67)	(34.64)
	[34.42]	[34.25]	[30.86]	[34.70]
$\hat{Y}^{(2)}_{npreg}$	400.43	430.39	334.59	363.70
	(29.61)	(34.73)	(30.09)	(35.01)
	[32.72]	[34.23]	[30.91]	[34.98]
$\hat{Y}_{knn}$	382.46	464.14	319.42	376.96
	(27.07)	(36.64)	(32.59)	(32.14)
	[28.28]	[34.77]	[28.04]	[30.02]
$\hat{Y}_{pred,knn}$	398.60	424.02	353.21	364.48
	(28.87)	(34.09)	(29.96)	(32.14)
	[29.36]	[32.42]	[28.41]	[30.62]

Legend:
The standard error based on the external variance estimator (i.e. equation [4.1]) are given in $()$ and the standard errors based on the bootstrap variance estimates (1000 replications) are given in $[]$. $\bar{Y}_G$ is the simple sample mean of plots in G_k, $k = 1, 2, 3, 4$.

For completeness we consider a small-area estimation problem. We have a partition of the entire domain F, in 4 small areas G_k, i.e. $F = \cup_{k=1}^{4} G_k$, with approximately equal

surface areas $\lambda(G_k)$. Under the simple external model approach we apply all estimators to G_k exactly as we would to F except that the regression coefficients, predictions and nearest neighbors are always based on the entire domain F. In order to apply a design-based small area estimator in this way, one implicitly is assuming that there is adequate sample in the final phase of the small area (see Mandallaz (2013b) for examples where n_2 as low as 6 was still adequate). For the bootstrap calculations, the re-sampling takes place in F.

Table 4.2 displays the results for small area. All point estimates are close to each other and not significantly different. The external variance estimates and bootstrap variances of all estimators are comparable. There is no clear trend in underestimation of the bootstrap variance by the external variance as we saw in the global case, but it should be noted that the sample sizes are relatively low. One may obtain better bootstrap results for the small areas by using a modified balanced replication as suggested in Magnussen et al. (2010). However, this was not the case here.

4.4.2 Simulations

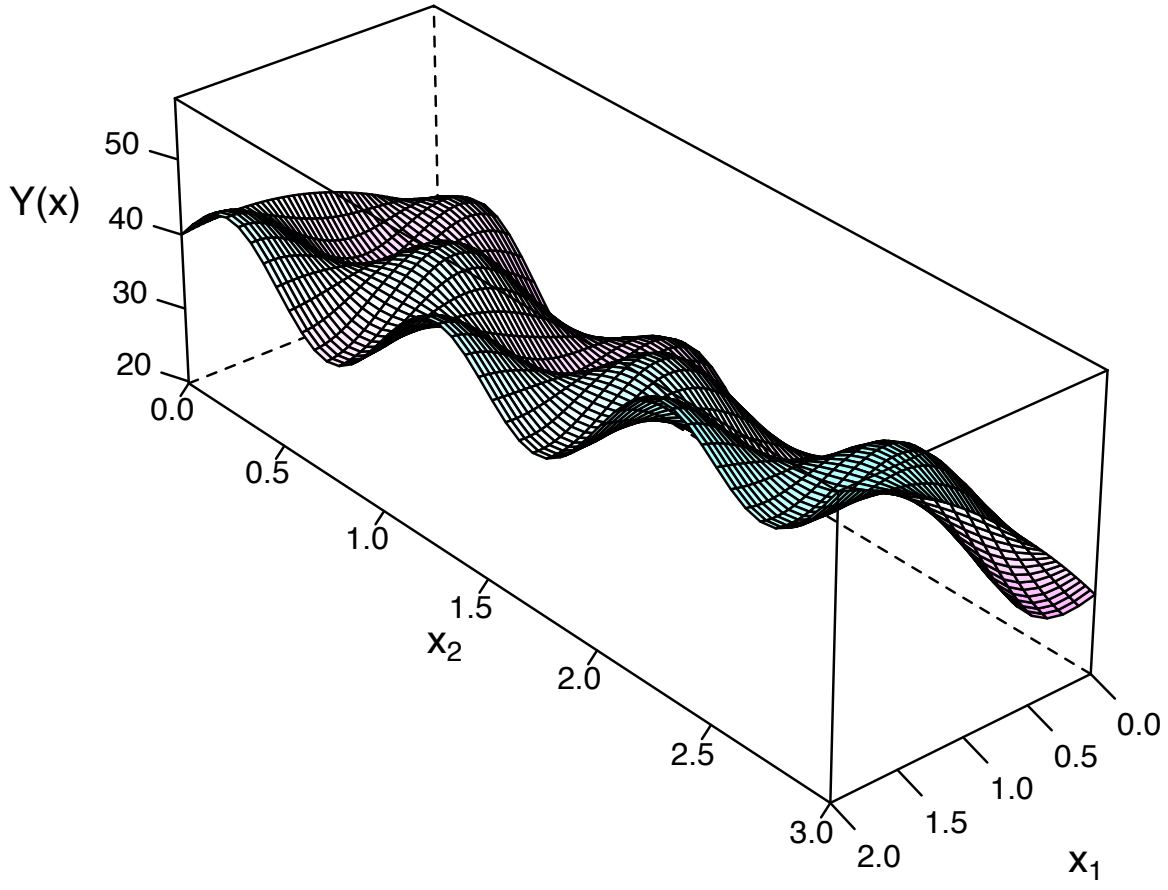
In the case study we observed some evidence that the external variance formula underestimates the theoretical design-based variance because it seemed to systematically underestimate bootstrap variance, especially in the global case. However the truth was not known so it is difficult to determine the adequacy of the implemented bootstrap. Now we present a purely artificial example where the truth is actually known and we can illustrate the theory by empirically checking the validity of the external variance formula as well as of our implementation of the bootstrap.

We present the simulation example already used in Mandallaz (2013b). The local density $Y(x)$ is defined according to the following procedure: at point $x = (x_1, x_2)^t \in \mathbb{R}^2$, the auxiliary vector is defined as $\boldsymbol{Z}(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^t \in \mathbb{R}^p$ ($p = 6$ in this example). The true parameter is $\boldsymbol{\beta}_0 = (30, 13, -6, -4, 3, 2)^t \in \mathbb{R}^6$ and the local density over the domain $F = [0, 2] \times [0, 3]$ is given by the function

$$Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0 + 6\cos(\pi x_1)\sin(2\pi x_2) := \hat{Y}_0(x) + R(x) \qquad [4.14]$$

Figure 4.1 shows a visual representation of the local density surface. Since we have an analytical representation of $Y(x)$ for the entire domain, F, we can iteratively draw samples according to our two-phase sampling design while calculating estimates, external variance estimates, and bootstrap estimates for any arbitrary estimator. The empirical variance of the point estimates across all performed iterations should offer excellent

Figure 4.1: Visual representation of simulation surface



insight to the theoretical design-based variance of an estimator. Likewise, the empirical mean allows us to observe the design-based expectation. The mean of the variance estimates, either external or bootstrap, across all iterations tells us the expectation of the respective variance estimator which we can then compare to the empirical variance.

The two sample size scenarios performed were $n_2 = 400$ and $n_3 = 100$ as well as $n_2 = 200$ and $n_3 = 50$. For each of these scenarios, two model choices were applied: the true model based on x_1, x_2, x_1^2, x_1x_2 and x_2^2; and a working model based on the subset of the true model using only x_2, x_1^2 and x_2^2. The number of neighbors for $\hat{Y}_{knn}$ was fixed at $k = 3$ (details about this decision are given in Mandallaz and Massey (2015)). For the estimator $\hat{Y}_{pred,knn}$ the optimal values of k were obtained from restricted simulations with 500 runs for each sample size and model scenario. The median of these optimal k's obtained via cross validation were used as the fixed k for each of the scenarios.

Tables 4.3 and 4.4 give the results for global estimation when the fitted model is indeed the true model and Tables 4.5 and 4.6 when the fitted model is the working model. All simulations are based on $10,000$ runs. The 95% bootstrap confidence intervals are obtained via the well-known formula $[2\hat{Y} - q^*_{0.975}, 2\hat{Y} - q^*_{0.025}]$, where q^*_α is the α-bootstrap quantile of the bootstrap estimate $\hat{Y}^*$, defined as the mean of the point estimates of the $1,000$ bootstrap samples obtained in each of the $10,000$ runs (see Davison and Hinkley (1997)).

Table 4.3: $n_2 = 400, n_3 = 100$: fitted model is the true model

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.17	0.19	0.19	0.20	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.15	0.20	0.19	0.20	0.94	0.95
$\hat{Y}^{(1)}_{npreg}$	39.15	0.20	0.18	0.20	0.94	0.94
$\hat{Y}^{(2)}_{glkern}$	39.15	0.20	0.19	0.20	0.94	0.95
$\hat{Y}^{(2)}_{npreg}$	39.17	0.20	0.18	0.20	0.94	0.95
$\hat{Y}_{pred,knn,k=11}$	39.16	0.20	0.18	0.20	0.93	0.94
$\hat{Y}_{knn,k=3}$	39.15	0.18	0.14	0.19	0.92	0.95

Table 4.4: $n_2 = 200, n_3 = 50$: fitted model is the true model

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.18	0.41	0.38	0.42	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.17	0.42	0.36	0.45	0.92	0.95
$\hat{Y}^{(1)}_{npreg}$	39.15	0.42	0.35	0.43	0.92	0.94
$\hat{Y}^{(2)}_{glkern}$	39.17	0.42	0.36	0.45	0.92	0.95
$\hat{Y}^{(2)}_{npreg}$	39.17	0.43	0.36	0.42	0.93	0.94
$\hat{Y}_{pred,knn,k=6}$	39.15	0.43	0.35	0.43	0.92	0.94
$\hat{Y}_{knn,k=3}$	39.15	0.43	0.31	0.43	0.91	0.94

Legend (Tables 4.3 and 4.4):
The true mean value is $\bar{Y} = 39.17$ and the true coefficient of determination is $R^2 = 0.83$.
$\mathbb{E}^*(\hat{P})$ is the empirical coverage probability of the 95% confidence interval based on the estimated variance and the normal approximation.
$\mathbb{E}^*(\hat{P}_{boot})$ is the empirical coverage probability based on the bootstrap confidence intervals.
$\mathbb{E}^*()$ and $\mathbb{V}^*()$ denote the empirical mean and variances over 10'000 runs.
$\mathbb{E}^*(\hat{\mathbb{V}}())$ is the empirical expectation of the external variance formula.
$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$ is the empirical expectation of the bootstrap variance obtained from 1000 bootstrap replicates.

Table 4.5: $n_1 = 400, n_2 = 100$: fitted model is not the true model

Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.18	0.28	0.26	0.27	0.95	0.95
$\hat{Y}^{(1)}_{glkern}$	39.19	0.27	0.25	0.28	0.93	0.94
$\hat{Y}^{(1)}_{npreg}$	39.18	0.28	0.24	0.27	0.92	0.94
$\hat{Y}^{(2)}_{glkern}$	39.19	0.27	0.25	0.28	0.93	0.94
$\hat{Y}^{(2)}_{npreg}$	39.18	0.27	0.25	0.27	0.93	0.94
$\hat{Y}_{knn,pred,k=15}$	39.17	0.28	0.24	0.27	0.93	0.94
$\hat{Y}_{knn,k=3}$	39.12	0.21	0.12	0.22	0.91	0.96

Table 4.6: $n_1 = 200, n_2 = 50$: fitted model is not the true model

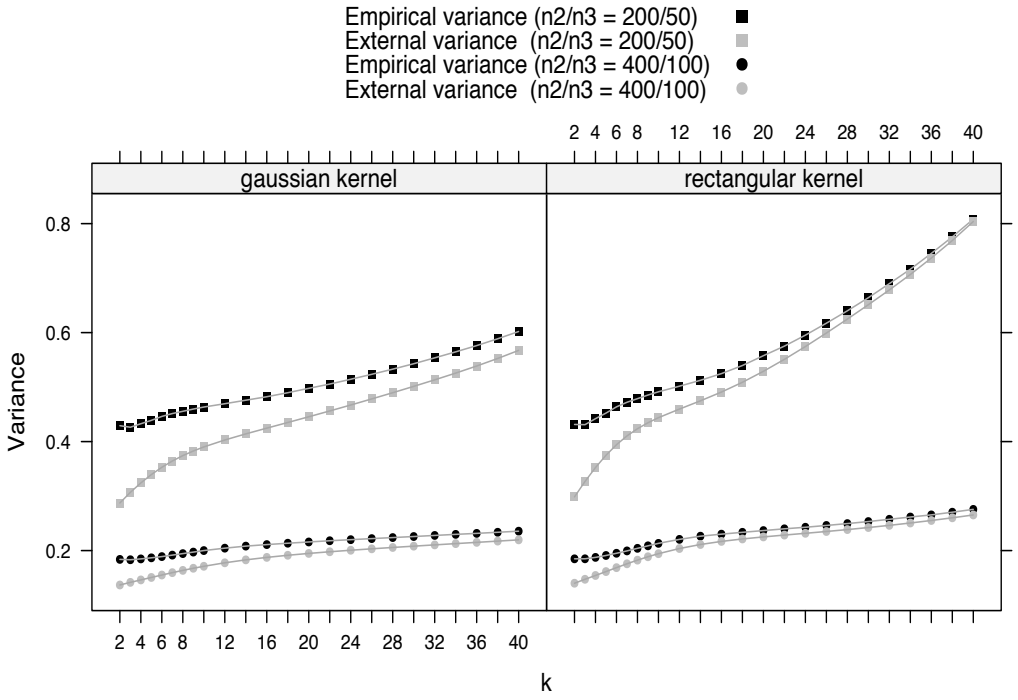
Estimator	$\mathbb{E}^*()$	$\mathbb{V}^*()$	$\mathbb{E}^*(\hat{\mathbb{V}}())$	$\mathbb{E}^*(\hat{\mathbb{V}}_{boot}())$	$\mathbb{E}^*(\hat{P})$	$\mathbb{E}^*(\hat{P}_{boot})$
$\hat{Y}_{reg}$	39.21	0.56	0.50	0.56	0.94	0.95
$\hat{Y}^{(1)}_{glkern}$	39.21	0.57	0.48	0.62	0.92	0.94
$\hat{Y}^{(1)}_{npreg}$	39.18	0.57	0.46	0.56	0.92	0.93
$\hat{Y}^{(2)}_{glkern}$	39.21	0.57	0.48	0.62	0.92	0.94
$\hat{Y}^{(2)}_{npreg}$	39.19	0.57	0.48	0.56	0.92	0.94
$\hat{Y}_{knn,pred,k=9}$	39.17	0.56	0.46	0.56	0.92	0.94
$\hat{Y}_{knn,k=3}$	39.12	0.50	0.32	0.51	0.89	0.94

Legend (Tables 4.5 and 4.6):
The columns headers are the same as in Tables 4.3 and 4.4. The true mean value is the same, i.e. $\bar{Y} = 39.17$. The fitted model did not include the explanatory variables x_1 and x_1x_2. The true coefficient of determination decreases from $R^2 = 0.83$ to $R^2 = 0.66$.

Discussion:

1. All estimators are practically design-unbiased. The empirical variances, which closely approximate the normally unobservable theoretical variance are smaller for the true model than for the working model, as expected. The external variances always underestimate the empirical variance, as mathematical intuition and the case study suggest. The underestimation appears to be much more pronounced for the nonparametric estimators than for $\hat{Y}_{reg}$. For the kNN estimator the underestimation occurs regardless of the choice of the kernel or k (see Figure 4.2).

Figure 4.2: Influence of choice of 'k' for two different kernels using kNN estimator with multidimensional support based on the complete set of explanatory variables from the true model



Empirical and external variance estimates based on 10000 simulations

2. Although these results were not presented in Tables 4.3, 4.4, 4.5 or 4.6, the estimators were also calculated using a true external approach where the models were fitted using an independent sample selection based on n_2 observations from [4.14]. As mathematically expected, the empirical variances closely matched the empirical means of the external variance estimates in all scenarios and the coverage probabilities were equal to 95%.

3. There is strong evidence that the implemented bootstrap procedure was adequate. The empirical mean of the bootstrap variance estimates are close to their associated empirical variances. Furthermore the empirical coverage probabilities of the bootstrap confidence intervals were all close to the nominal 95% probability. With respect to the nominal 95% confidence intervals the classical regression estimator $\hat{Y}_{reg}$ performs well in all investigated cases. For all other estimators one must use the bootstrap confidence intervals because their external variances clearly underestimate the empirical variances. The underestimation was less but still present in the scenario with higher sample sizes.

4. The better performance of the classical regression estimator $\hat{Y}_{reg}$ is probably due to the fact that the coefficients of determination are rather high under the true and under the working models. On the whole the kernel-based or **kNN** estimators did not perform better than than the classical regression estimator $\hat{Y}_{reg}$, with the exception being the nearest neighbor estimator $\hat{Y}_{knn}$ with the choice $k = 3$. This estimator yielded a slightly smaller empirical variance when the working model was not the true model and its variance could be correctly estimated with the bootstrap. This may suggest that $\hat{Y}_{knn}$ is better as the goodness of fit decreases. That being said, the R^2 of the working model was 0.66, which is not outstanding in the context of forest inventories, e.g. with timber volume as response variable and with LiDAR canopy measurements as explanatory variables.

5. A further theoretical advantage of $\hat{Y}_{reg}$ is that it allows for correct analytical expressions of the asymptotic variances, which currently seems intractable to derive for the other estimators. Analytical variance estimates are exactly reproducible given the same data set.

4.5 Conclusions

Intuitively, treating internally fit regression estimators as external via the application of the external variance formula can lead to systematic underestimation of the true design-based variance because one implicitly ignores the influence of the random sample selection on the realization of the model prediction. This intuition was empirically confirmed for kernel-regression estimators by simulation results that demonstrated that the observed coverage rates for kernel-based estimators fell consistently below the nominal coverage probability. The underestimation can be corrected by bootstrapping, but special attention should be given to the choice of the tuning options within the resampling

procedures because they can lead to unstable variance estimates. An acceptable way of addressing this instability is to select the optimal bandwidth on the original sample and then fix the bandwidth to that value across all bootstrap replicates. The external variance formula for the classical two-phase regression estimator, $\hat{Y}_{reg}$, on the other hand, remained acceptable even for internal models as the underestimation appears to be asymptotically negligible.

These preliminary comparisons between the various estimators show that the classical two-phase regression estimator with the external variance estimates performs on the whole at least as well as the kernel-based regression estimators with bootstrap variance. The only exception was $\hat{Y}_{knn}$, the kNN estimator with distance defined in the multidimensional feature space. $\hat{Y}_{knn}$ appears to do better when the goodness of fit of the auxiliary variables is low, but the variance estimate should be bootstrapped to avoid potentially dramatic underestimation. It should be noted that different versions of $\hat{Y}_{reg}$ are available for more complex situations such as three-phase cluster sampling with two-types of auxiliary information, and with two-stage sampling of trees at the plot level, as well as its extensions to small-area estimation (Mandallaz (2013b,a,c)), where it is not yet very clear how to proceed using the kernel-based approach. It can reasonably be expected that these conclusions will hold in general if the model is adequate, i.e. if it incorporates the most important explanatory variables (say with an $R^2 \geq 0.6$), which is certainly the case for timber volume with the standard LiDAR explanatory variables (mean canopy height being the most important).

In the context of forest inventories, the main advantage of $\hat{Y}_{reg}$ is its flexibility and the existence of an analytical design-based variance estimator. Being constructed using a standard linear model, it is relatively easy to implement and to obtain highly reproducible results that are independent of the choice of kernels, metrics and further tuning options. $\hat{Y}_{knn}$ may perform a bit better in situations where the auxiliary information leads to a decreased goodness of fit, but the variance should be obtained via bootstrap which leads to more potential arrangements of tuning parameters. Practically speaking, this means that exact reproduction of estimates, even given identical data sets, becomes more problematic and less likely to occur.

Chapter 5

Estimation of net change

The transition of the Swiss NFI in 2009 from a periodic design to an annual design was primarily motivated by a desire to produce estimates continuously over time as well as the logistic benefits associated with a regular inventory schedule, results in a loss in precision due to the reduction of available sample size for a given time point t_k. Consequently, new statistical procedures are needed that efficiently incorporate remote sensing information to estimate the state of the forest at a given time, i.e. the spatial mean of the growing stock volume at time period t_k denoted $\bar{Y}(t_k)$, and also its change over time, i.e. $\bar{Y}(t_2) - \bar{Y}(t_1)$. The estimation of the state for the Swiss NFI has been addressed in Chapter 3 via the proposal of a three-phase regression estimator. However, the estimation of net change remains a challenge under the annual design, and the estimation of the classical growth components (survivor growth, in-growth, on-growth, non-growth and depletion, see Mandallaz (2008), chapter 11) even more so, particularly if one wants to use the two-stage procedure at the plot level (which yields better estimates of the stem volume).

A common approach when estimating the change in time when permanent plots are present is to observe the change directly on the plot level and then treat that as the variable of interest. In the presence of annual designs one is thus able to estimate net change directly on a continuous basis, but only across time spans, denoted $t_2 - t_1$, that are multiples of the time it takes to complete a full cycle of the annual inventory. This approach is flexible and seamlessly allows for the implementation of model-assisted techniques such as the use of two- and three-phase regression estimators to mitigate this loss of precision due to reduced sample size while remaining design unbiased. However, if one wants to circumvent the drawback of only being able to make estimates only for predetermined time spans, one must estimate net change indirectly by taking the difference between estimates of the state of the forest at two distinct time points,

which still leads to an unbiased estimate in the design-based sense. Although it is generally thought that direct change estimation leads to lower variances when compared to indirect estimation over the same observable time span, this notion has recently been challenged by McRoberts et al. (2015) for two-phase regression estimation. This chapter explores the utility of design-based direct and indirect net change estimation in the context of the Swiss NFI and expands the notation and methodology of the two- and three-phase regression estimators presented in previous chapters. Furthermore, it provides suggestions for indirect estimators under annual designs for time spans where direct estimation is not possible.

A shorter version of this chapter has been submitted to the *Canadian Journal of Forest Research* (see Massey and Mandallaz (2015b)).

5.1 Expanded notation

First we need to expand the notation in order to include the temporal component. We want to estimate the net change between time t_1 and t_2 of the spatial mean of the timber volume densities over the *permanent* forested area F (new forested area and new non-forested area are excluded from this analysis). We have a well-defined population P of trees $i \in 1, 2, ..., N(t_k)$, for $k = 1, 2$, in forest F at time t_k and for every tree a response variable Y_i, which in this case is the individual tree's stem volume. The spatial mean at time t_k measured in cubic meters per hectare is denoted $\bar{Y}(t_k) := \frac{1}{\lambda(F)} \sum_{i=1}^{N(t_k)} Y_i(t_k)$ where $\lambda(F)$ is the surface area of F, and the parameter of interest is $\Delta_{\bar{Y}}(t_1, t_2) := \bar{Y}(t_2) - \bar{Y}(t_1)$.

We will work in a similar three-phase setup summarized as follows. The first phase is presumably derived from remote sensing data and consists of a large sample s_1 of n_1 **points** $x \in F$ assumed to be distributed uniformly and independently in F whose information is contained in row vector $\mathbf{Z}^{(1)t}(x)$. The second phase $s_2 \subset s_1$ of n_2 points is drawn by simple random sampling (SRS) without replacement from s_1 and consists of information obtained from terrestrial plots from the full cycle of the annual design (usually some growth updated form of the previous plot volume measurement at x). Thus auxiliary information for every $x \in s_2$ is contained in row vector $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)t}(x), \mathbf{Z}^{(2)t}(x))$ where $\mathbf{Z}^{(2)t}(x)$ represents the terrestrially acquired or computationally intensive remote sensing information (Mandallaz (2014)). The third-phase sample at time t_k is denoted by s_{3k} with n_{3k} points drawn by SRS without replacement from s_2. The third phase yields the ground truth under the annual design of the local densities

$Y(t_k, x)$ at time t_k where $k = 1, 2$.

Although we mathematically assume uniform independent sampling and SRS without replacement subsampling, the case study will actually use systematic sampling with a random start. However, treating systematic samples as uniformly random samples is not expected to be a problem for the same reasons discussed in Chapter 2. Since we are in the design-based framework, the local density estimate $Y(t_k, x)$ is assumed error-free given x at time t_k and the random selection x is uniform in F. We are using the Monte Carlo approach in the design-based set-up where the relation $\mathbb{E}_x(Y(t_k, x)) = \frac{1}{\lambda(F)} \int_F Y(t_k, x)dx = \frac{1}{\lambda(F)} \sum_{i=1}^{N(t_k)} Y_i(t_k) = \bar{Y}(t_k)$ holds for both fixed radius or variable radius (angle count) sampling of trees, which requires that our sample selection of x is uniformly and independently drawn in F.

We now consider two situations concerning s_{31} and s_{32}. Let L be the complete cycle length of the annual inventory measured in vegetative periods. The first situation is when $s_{31} = s_{32}$, i.e. $t_2 - t_1$ equals a multiple of L. The second occurs when $t_2 - t_1$ is not a multiple of L such that $s_{31} \cap s_{32} = \emptyset$ (for simplicity we ignore scenarios where s_{31} and s_{32} only partially overlap which can occur when t_k represent groupings of multiple years).

If $s_{31} = s_{32}$ and we neglect the auxiliary information obtained from s_1 and s_2, the direct and indirect net change estimators become mathematically equivalent. Denoting $Y_\Delta(x) := Y(t_2, x) - Y(t_1, x)$, $n_3 := n_{31} = n_{32}$ and $s_3 := s_{31} = s_{32}$,

$$\hat{\Delta}_{dir,1p}(t_1, t_2) := \frac{1}{n_3} \sum_{x \in s_3} Y_\Delta(x) \quad [5.1]$$

$$= \frac{1}{n_{32}} \sum_{x \in s_{32}} Y(t_2, x) - \frac{1}{n_{31}} \sum_{x \in s_{31}} Y(t_1, x) =: \hat{\Delta}_{indir,1p}(t_1, t_2) \quad [5.2]$$

are unbiased estimators of the parameter $\Delta_{\bar{Y}}(t_1, t_2)$ with unbiased variance estimator

$$\hat{\mathbb{V}}(\hat{\Delta}_{dir,1p}(t_1, t_2)) = \frac{1}{n_3} \frac{1}{n_3 - 1} \sum_{x \in s_3} (Y_\Delta(x) - \bar{Y}_{\Delta,s_3})^2 \quad [5.3]$$

where $\bar{Y}_{\Delta,s_3} := \frac{1}{n_3} \sum_{x \in s_3} Y_\Delta(x)$ for notational convenience.

If $s_{31} \cap s_{32} = \emptyset$ then $Y_\Delta(x)$ is unobservable and only indirect calculation is possible using [5.2], which is still design unbiased for the point estimate. However, the variance of the change will be the sum of the variances for the states at t_1 and t_2:

$$\hat{\mathbb{V}}(\hat{\Delta}_{\bar{Y}_\Delta}(t_1, t_2)) = \frac{1}{n_{32}} \frac{1}{n_{32} - 1} \sum_{x \in s_{32}} (Y(t_2, x) - \bar{Y}_{s_{32}})^2 + \frac{1}{n_{31}} \frac{1}{n_{31} - 1} \sum_{x \in s_{31}} (Y(t_2, x) - \bar{Y}_{s_{31}})^2 \quad [5.4]$$

where $\bar{Y}_{s_{3k}} = \frac{1}{n_{3k}} \sum_{x \in s_{3k}} Y(t_k, x)$ for $k = 1, 2$. The variance in [5.4] is likely to be unacceptably high because of the independence between s_{31} and s_{32}. We expect to obtain more precise design-unbiased estimators of $\Delta_{\bar{Y}}(t_1, t_2)$ by incorporating the information from the auxiliary vectors $\mathbf{Z}^{(1)t}(x)$ and $\mathbf{Z}^{t}(x)$ using model-assisted estimation.

5.2 Model-assisted estimation of net change

5.2.1 The external model approach

All estimators presented here will be based off of various generalized difference estimators, derived under the external model approach. Thus, the point estimators are of the forms [2.7] and [2.8] with external variances [2.11] and [2.12].

To recap, [2.7], [2.11], [2.8] and [2.12] remain exactly design-unbiased regardless of what models (linear, nonparametric, kNN etc.) are specified to generate the predictions, provided that we work under the external model assumption. A model is called external if the training set used to fit it does not depend on the current inventory sample realization. This ensures that $\hat{Y}_{M_1}(x)$ (likewise $\hat{Y}_{M_2}(x)$) remain i.i.d. across all $x \in s_1$, s_2 and s_3 (notice that there are no potentially complicated design-based covariance terms in [2.11] or [2.12]). In practice of course, true external models are rarely if ever used since most practitioners use s_3 as a training set, thus making the model internal. Applying formulas derived under the external model assumption to internal models may still be asymptotically acceptable (as is the case for linear models), but the adequacy of this approach is questionable for other models types such as kNN (see Chapter 4). For simplicity, all analytical estimators proposed here are derived under the external model approach and the resulting external variances are compared to the bootstrap.

5.2.2 Direct Estimation

For direct estimation of change, $Y_\Delta(x)$ must be observable. Thus, we are in a permanent plot setup and for simplicity assume that $s_{31} = s_{32}$. Only point estimators and variance estimators are given here but the interested reader is referred to Mandallaz (2008), p. 80 and Appendix B, for mathematical details concerning the derivations. The direct estimators are obtained simply by plugging $Y_\Delta(x)$, $\hat{Y}_\Delta^{(1)}(x)$ and $\hat{Y}_\Delta(x)$ into the difference estimator for $Y(x)$, $\hat{Y}_{M_1}(x)$ and $\hat{Y}_{M_2}(x)$ respectively and likewise for the

residuals. The three- and two-phase direct estimators are thus defined:

$$\hat{\Delta}_{dir,3p}(t_1,t_2) = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_\Delta^{(1)}(x) \quad + \quad \frac{1}{n_2}\sum_{x\in s_2}(\hat{Y}_\Delta(x) - \hat{Y}_\Delta^{(1)}(x)) \quad + \quad \frac{1}{n_3}\sum_{x\in s_3}(Y_\Delta(x) - \hat{Y}_\Delta(x)) \tag{5.5}$$

$$\hat{\Delta}_{dir,2p}(t_1,t_2) = \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}_\Delta(x) + \frac{1}{n_3}\sum_{x\in s_3}(Y_\Delta(x) - \hat{Y}_\Delta(x)) \tag{5.6}$$

with variance estimators

$$\hat{\mathbb{V}}(\hat{\Delta}_{dir,3p}(t_1,t_2)) = \frac{1}{n_1}\hat{\mathbb{V}}(Y_\Delta(x)) \quad + \quad (1-\frac{n_2}{n_1})\frac{1}{n_2}\hat{\mathbb{V}}(\hat{R}_\Delta^{(1)}(x)) \quad + \quad (1-\frac{n_3}{n_2})\frac{1}{n_3}\hat{\mathbb{V}}(\hat{R}_\Delta(x)) \tag{5.7}$$

$$\hat{\mathbb{V}}(\hat{\Delta}_{dir,2p}(t_1,t_2)) \quad = \quad \frac{1}{n_2}\hat{\mathbb{V}}(Y_\Delta(x)) + \big(1-\frac{n_3}{n_2}\big)\frac{1}{n_3}\hat{\mathbb{V}}(\hat{R}_\Delta(x)) \tag{5.8}$$

where on the right hand side $\hat{\mathbb{V}}$ is defined as the sample variance in s_3, $\hat{R}_\Delta(x) := Y_\Delta(x) - \hat{Y}_\Delta(x)$ and $\hat{R}_\Delta^{(1)}(x) := Y_\Delta^{(1)}(x) - \hat{Y}_\Delta^{(1)}(x)$.

$\hat{\Delta}_{dir,3p}(t_1,t_2)$ and $\hat{\Delta}_{dir,2p}(t_1,t_2)$ offer great flexibility when incorporating auxiliary information from different time points as well as sources such as remote sensing and terrestrially acquired data. The auxiliary variables chosen to be represented in $\mathbf{Z}^t(x)$ only need to be chosen so that they predict the net change, $Y_\Delta(x)$, as efficiently as possible. From a strict mathematical point of view it is irrelevant which explanatory variables are used and it can well be the case that the relevant variables for the estimation of state are not all relevant for the estimation of change and vice versa. Practically, it would seem reasonable to include pairs of remote sensing measurements, e.g. for mean canopy height (MCH), representing both t_1 and t_2. However, if a linear model is used and MCH at t_1 and MCH at t_2 are included separately in $\mathbf{Z}^t(x)$ then separate coefficients will be fit for the pair. If only a single variable representing the difference in MCH at t_1 and t_2 is included then only one coefficient is fit (see McRoberts et al. (2015) for details).

Classical regression estimators are obtained by letting $\hat{Y}_\Delta^{(1)}(x) := \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}$ and $\hat{Y}_\Delta(x) := \mathbf{Z}^t(x)\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are simply the solutions to the normal equations when regressing $Y_\Delta(x)$ on $\mathbf{Z}^{(1)t}(x)$ and $Y_\Delta(x)$ on $\mathbf{Z}^t(x)$ respectively for all $x \in s_3$. These classical regression estimators, as well as their variance estimators, are approximately design-unbiased even when they are fit internally and treated as external.

We obtain a nonparametric regression estimator by using s_3 as a reference set to produce the predictions for $\hat{Y}_{\Delta}^{(1)}(x)$ and $\hat{Y}_{\Delta}(x)$. Such estimators are discussed in Chapter 4 for kernel estimators such as the Nadaraya-Watson and kNN estimators, although only the kNN regression estimator will be considered here. There is some evidence that the kNN regression estimator can perform better than the classical regression estimator when the auxiliary variables possess a lower goodness-of-fit (Haara and Kangas (2012)). However, the external variance estimators obtained by [5.7] and [5.8] may be prone to underestimating the theoretical variance under moderate sample sizes and a bootstrap routine is recommended (see Chapter 4 for details). The simplest way to implement such a bootstrap is to resample from s_1 using SRS with replacement, allowing n_2 and n_3 to be random in each bootstrap sample. Pseudo-estimates are then computed by applying the desired estimator to each bootstrap sample. The sample variance of the pseudo-estimates is the bootstrap variance estimate. To improve stability, tuning parameters such as the choice of k should be chosen based on the original sample only and held fixed across all bootstrap replications.

Advantages and drawbacks

The main advantage of the direct change estimators presented is that their fits are optimized directly on the variable of interest. Analytical variance estimators exist greatly facilitating the reproducibility of results, but it should be noted that one often expects a slight underestimation of the variance when applying the external variance formula to internal models. Intuitively this is due to neglecting the effect of sampling variability on the fit.

The main drawback of the direct change estimation under annual designs is that there is dependency on the period of the remeasurement cycle and the remeasurement period is often not very flexible. For example, the Swiss NFI currently employs an annual design where approximately one ninth of the plots are remeasured every year. Thus the remeasurement period will always be multiples of 9 years. Net change can of course be annualized by dividing by the length of the remeasurement period, but this implies constant linear change over the interval, which is not always what is desired. If a net change estimate is desired over a period with a length different than the remeasurement period then one has to extrapolate based on this linearization.

Another potential drawback for direct change estimation is that, as it is more preferable for remote sensing information to be available at two time points rather than one, the quality of the prediction for $Y_{\Delta}(x)$ becomes more dependent on good synchronicity in

the timing of the remote sensing and terrestrial measurements so that disturbance (e.g. harvesting, wind throw, etc) can be observed on the plot level. The most unpredictable component of forest change is disturbance, either natural or man-made, and temporal gaps between the remote sensing remeasurement window and the terrestrial remeasurement window introduce blind spots for disturbance detection. Moreover, perfect temporal synchronicity is logistically problematic in data collection. In the Swiss NFI, for example, complete national coverage of leaves-on digital stereophotographs is guaranteed only every 6 years and during that time the measurements are taken regionally, whereas the annual terrestrial sample is uniformly distributed across the entire country. This leads one to expect rather low goodness of fit for $\hat{Y}_{\Delta}^{(1)}(x)$ and $\hat{Y}_{\Delta}(x)$ under the direct method, which is our motivation for also employing the kNN regression estimator in our case study.

5.2.3 Indirect Estimation

The indirect change estimation method requires static estimates of the forest state at two time points and the net change is calculated as the difference between the two estimates. We again consider two time points t_1 and t_2 and without loss of generality assume that $t_2 > t_1$. Let $s_1(t_k)$ and $s_2(t_k)$ be the first- and second-phase samples of points taken at t_k respectively. For simplicity we assume that the points included in the first and second phases are the same, i.e. $s_1 := s_1(t_1) = s_1(t_2)$ and $s_2 := s_2(t_1) = s_2(t_2)$. We consider two potential scenarios for the final phase where the local densities $Y(t_1, x)$ and $Y(t_2, x)$ are observed at the same points in s_3, i.e. $s_{31} = s_{32}$, and where they are observed at completely distinct points, i.e. $s_{31} \cap s_{32} = \emptyset$.

A lower variance can be obtained using [2.7] to estimate the forest states at t_1 and t_2. Thus we take the same sample *points* from remote sensing and the same from previous measurements at t_2 and t_1. The three- and two-phase generalized difference estimators of state at time t_k are denoted

$$\begin{aligned} \hat{Y}_{3p}(t_k) &= \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}^{(1)}(t_k,x) + \frac{1}{n_2}\sum_{x\in s_2}(\hat{Y}(t_k,x) - \hat{Y}^{(1)}(t_k,x)) \\ &\qquad + \frac{1}{n_{3k}}\sum_{x\in s_{3k}}(Y(t_k,x) - \hat{Y}(t_k,x)) \end{aligned} \qquad [5.9]$$

$$\hat{Y}_{2p}(t_k) = \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}(t_k,x) + \frac{1}{n_{3k}}\sum_{x\in s_{3k}}(Y(t_k,x) - \hat{Y}(t_k,x)) \qquad [5.10]$$

for $k = 1, 2$ and the estimators of $\Delta_{\bar{Y}}(t_1, t_2)$ are

$$\hat{\Delta}_{indir,3p}(t_1, t_2) = \hat{Y}_{3p}(t_2) - \hat{Y}_{3p}(t_1) \quad [5.11]$$
$$\hat{\Delta}_{indir,2p}(t_1, t_2) = \hat{Y}_{2p}(t_2) - \hat{Y}_{2p}(t_1) \quad [5.12]$$

We can use any model to calculate the predictions, $\hat{Y}^{(1)}(t_k, x)$. and $\hat{Y}(t_k, x)$. Under the external model assumption $\hat{Y}_{3p}(t_k)$ and $\hat{Y}_{2p}(t_k)$ are design-unbiased estimators of $\bar{Y}(t_k)$ which immediately implies that $\hat{\Delta}_{indir,3p}$ and $\hat{\Delta}_{indir,2p}$ are unbiased estimators of $\Delta_{\bar{Y}}(t_1, t_2)$. The theoretical variances will differ slightly at this point for the two scenarios where $s_{31} = s_{32}$ and $s_{31} \cap s_{32} = \emptyset$. An example derivation is given in the Appendix.

Scenario One: $s_{31} = s_{32}$

Since $s_{31} = s_{32}$ we can use the simpler notation with s_3 and n_3. The theoretical variances are

$$\begin{aligned}\mathbb{V}(\hat{\Delta}_{indir,3p}(t_1, t_2)) &= \frac{1}{n_1}\mathbb{V}(Y(x, t_2) - Y(x, t_1)) \\ &\quad + (1 - \frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R^{(1)}1(x, t_2) - R^{(1)}(x, t_1)) \\ &\quad + (1 - \frac{n_3}{n_2})\frac{1}{n_3}\mathbb{V}(R(x, t_2) - R(x, t_1))\end{aligned} \quad [5.13]$$

$$\begin{aligned}\mathbb{V}(\hat{\Delta}_{indir,2p}(t_1, t_2)) &= \frac{1}{n_2}\mathbb{V}(Y(x, t_2) - Y(x, t_1)) \\ &\quad + (1 - \frac{n_3}{n_2})\frac{1}{n_3}\mathbb{V}(R(x, t_2) - R(x, t_1))\end{aligned} \quad [5.14]$$

Note that in this scenario the $Y(x, t_2) - Y(x, t_1)$, $R^{(1)}1(x, t_2) - R^{(1)}(x, t_1)$, and $R(x, t_2) - R(x, t_1)$ are all observable because each $x \in s_3$ is observed at both t_1 and t_2. Thus, the external variance estimators are based on sample copies of each variance term in s_3.

Scenario Two: $s_{31} \cap s_{32} = \emptyset$

The theoretical variances look a little different from [5.13] and [5.14]:

$$\begin{aligned}\mathbb{V}(\hat{\Delta}_{indir,3p}(t_1,t_2)) &= \frac{1}{n_1}\mathbb{V}\big(Y(x,t_2)-Y(x,t_1)\big) \\ &\quad + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}\big(R^{(1)}1(x,t_2)-R^{(1)}(x,t_1)\big) \\ &\quad + (1-\frac{n_{31}}{n_2})\frac{1}{n_{31}}\mathbb{V}(R(x,t_2)) \\ &\quad + (1-\frac{n_{32}}{n_2})\frac{1}{n_{32}}\mathbb{V}(R(x,t_1))\end{aligned} \quad [5.15]$$

$$\begin{aligned}\mathbb{V}(\hat{\Delta}_{indir,2p}(t_1,t_2)) &= \frac{1}{n_2}\mathbb{V}(Y(x,t_2)-Y(x,t_1)) \\ &\quad + (1-\frac{n_{31}}{n_2})\frac{1}{n_{31}}\mathbb{V}(R(x,t_1)) \\ &\quad + (1-\frac{n_{32}}{n_2})\frac{1}{n_{32}}\mathbb{V}(R(x,t_2))\end{aligned} \quad [5.16]$$

In this scenario we only need to estimate $\mathbb{V}(R(x,t_1))$ and $\mathbb{V}(R(x,t_2))$ rather than $\mathbb{V}(R(x,t_2)-R(x,t_1))$ because s_{31} is conditionally independent of s_{32}. This can be done using sample copies in s_{31} and s_{32} respectively. However, the problem is that we still need to estimate $\mathbb{V}\big(Y(x,t_2)-Y(x,t_1)\big)$ and $\mathbb{V}\big(R^{(1)}1(x,t_2)-R^{(1)}(x,t_1)\big)$, but there exists no x where $Y(x,t_2)-Y(x,t_1)$ or $R^{(1)}1(x,t_2)-R^{(1)}(x,t_1)$ are observed in the realized sample. We can certainly estimate each variance using the bootstrap to remain firmly in the design-based setup, or we can attempt to construct an analytical variance estimator using an *ad hoc* solution.

In order to construct such an *ad hoc* solution for the two-phase estimator consider the decomposition

$$\begin{aligned}&\mathbb{V}(Y(t_2,x)-Y(t_1,x)) = \\ &\qquad \mathbb{V}(Y(t_2,x))+\mathbb{V}(Y(t_1,x))-2\rho_{Y_{\Delta t}}\sqrt{\mathbb{V}(Y(t_2,x))\mathbb{V}(Y(t_1,x))}\end{aligned} \quad [5.17]$$

Notice that only the correlation, $\rho_{Y_{\Delta t}}$, is inestimable from the available sample. It is reasonable to assume that $\rho_{Y_{\Delta t}} \geq 0$, so ignoring the third term completely would unsatisfactorily inflate the variance. A more efficient variance estimate is obtained if one can make a reasonably informed inference about the anticipated correlation between $Y(t_1,x)$ and $Y(t_2,x)$ from past inventories. If such information is available then one could construct a conservative model-based imputation of $\rho_{Y_{\Delta t}}$ and then apply the design-based variance formula described in [5.16] using sample copies of every variance term on the right hand side.

Theoretically, one could follow a similar procedure to impute $\rho_{R_{\Delta}^{(1)}(x)}$ and estimate $\mathbb{V}(\hat{\Delta}_{indir,3p}(t_1, t_2))$. However, this may not be reasonable. First of all, the same type of auxiliary information used in the model predictions would have to be available for historical inventories, which might be unlikely considering that technologies in remote sensing tend to change frequently over time. Second, it is not clear that one can assume a lower bound for $\rho_{R_{\Delta}^{(1)}(x)}$ or even that $\rho_{R_{\Delta}^{(1)}(x)} \geq 0$. Imputing a positive $\rho_{R_{\Delta}^{(1)}(x)}$ when it should be negative will actually lead to an underestimation of the variance, which should be avoided.

Advantages and drawbacks

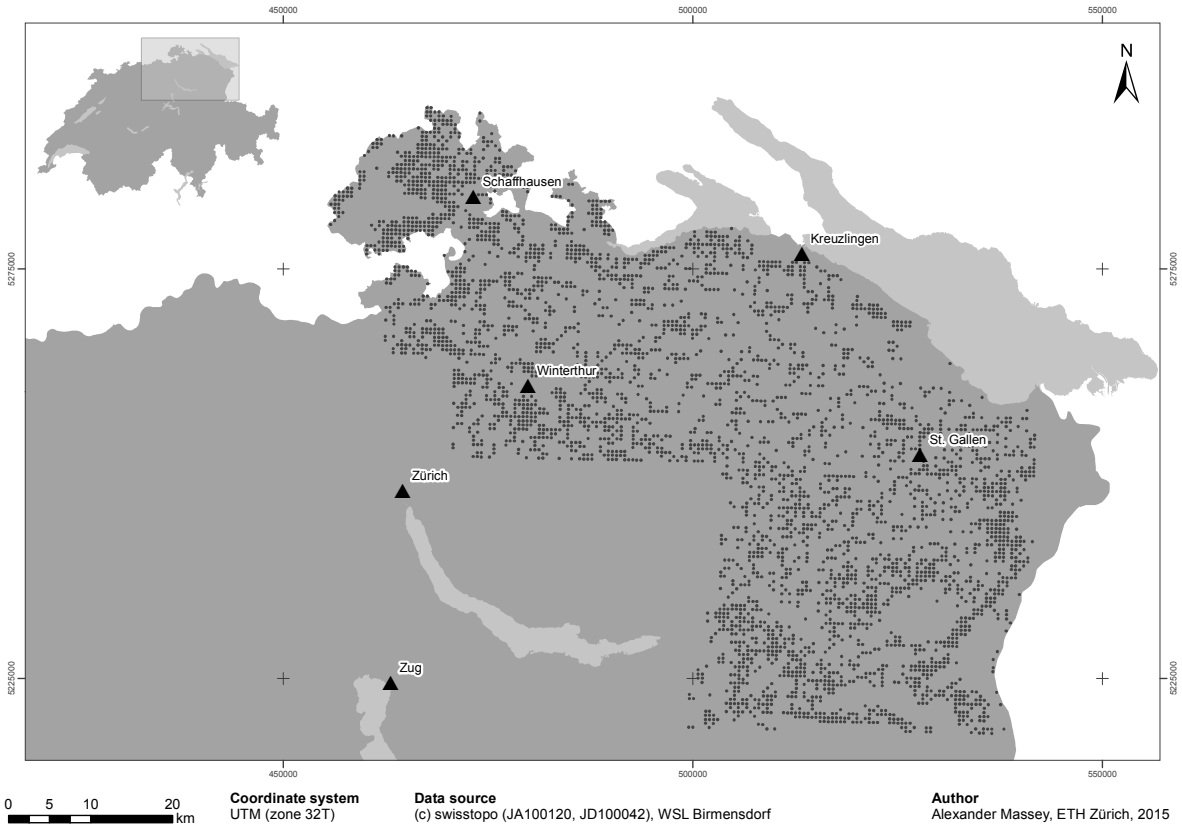
With the exception of $\hat{\Delta}_{indir,3p}(t_1, t_2)$ when $s_{31} \cap s_{32} = \emptyset$, analytical formulations of the variance estimators exist just as for the direct estimators. However, net change can be calculated over any arbitrary time span regardless of the observability of $Y_{\Delta}(x)$, which makes them more flexible. The downside to this flexibility is that one may have to rely on a bootstrap variance to stay purely design based. The proposed *ad hoc* procedure seems reasonable only for the two-phase estimator and is a bit of a mixture of the design-based and model-dependent philosophies. The other main drawback is that the models are fit to predict state, which leads us to conjecture that there exists at least some degree of suboptimality for predicting change.

5.3 Case Study

5.3.1 Study area

The study area is located in the northeast corner of Switzerland and draws information from multiple Swiss national forest inventories (NFI) as well as an updated canopy height model (CHM) derived by stereo-image matching of a digital surface model (DSM) generated by aerial stereo imagery and a previously existing digital terrain model (DTM). Figure 5.1 displays a map. Roughly 48% of the area is in the Swiss Plateau region where the incidence of man-made intervention is the heaviest, 50% is in the Jura or Prealps regions where moderate intervention occurs, and the remaining 2% lies in the Alps region where little man-made disturbance occurs. The target period over which net change is to be estimated occurs between years 2004/2005 and 2010/2011, representing a span of approximately 6 years on average. The 3771 first phase plots

Figure 5.1: Map of study area in Switzerland with locations of first-phase plots



are located on a square $500m$ by $500m$ grid and consist mainly of canopy height information obtained from the DSM. However, for the three-phase indirect estimator some canopy height information from the manually interpreted stereophotography had to be used to supplement the DSM, which only became available in 2005, because those estimators use the previous inventory plot volume as an auxiliary variable which works better in conjunction with an aerial disturbance update (Massey et al. (2014)). The second phase consists of 475 plots located on a $\sqrt{2}km$ by $\sqrt{2}km$ sub grid and contain plot volumes from the previous inventory. Note that both inventories preceding the 3rd and 4th Swiss NFIs are periodic and thus summarize a relatively fixed time point. The third phase contains 107 terrestrial plots evaluated under the annual design of the 4th Swiss NFI in 2010 and 2011. Since the 3rd NFI in the study area was conducted entirely from 2004 to 2005, we can either choose the third phase such that its plots are remeasured (i.e. $s_{31} = s_{32}$) or such that the none of final phase plots were remeasured (i.e. $s_{31} \cap s_{32} = \emptyset$) without changing the temporal period of our net change estimate. In the latter, $n_{31} = 112$ and $n_{32} = 107$.

5.3.2 Ground Data

The Swiss NFI employs a two-phase sampling design consisting of systematically distributed permanent sampling points for manual aerial photo interpretation on a rectangular $500m$ by $500m$ grid, and a $\sqrt{2}km$ by $\sqrt{2}km$ sub-grid of sampling points for terrestrial data collection (Lanz et al. (2010)). On each terrestrial plot, trees are measured using calipers according to two concentric circles measuring $200m^2$ for trees with diameter at breast height (DBH) between $12cm$ and $36cm$ and $500m^2$ for trees with DBH greater than $36cm$. An approximation of timber volume (over bark stem volume) is determined by using DBH (via one-way yield table stratified by species) on all eligible trees. For details we refer to the reader to Brassel and Lischke (2001). The 1st, 2nd and 3rd Swiss NFIs were periodic and occurred from 1983-1985, 1993-1995 and 2004-2006 respectively. The 4th Swiss NFI, which began in 2009, differs from the 1st, 2nd and 3rd due to the implementation of a new annual design where 1/9 of the plots are measured every year using interpenetrating grids ($\approx 4.2km$ by $4.2km$) to allow for more up-to-date estimates to be available between inventory cycles.

The 4th Swiss NFI began in 2009 and marks the transition from a periodic design to an annual one where $\frac{1}{9}$ of the terrestrial plots are evaluated each year. At the time of this analysis data was available from 2009-2013 for the 4th Swiss NFI. The Swiss NFI uses permanent plots, and thus pairs of plot density estimates can be calculated for any remeasured plot.

5.3.3 Digital aerial stereo-photography

Manual interpretation data

As previously mentioned, the first phase of the Swiss NFI consists of a large sample of plots taken from a $500m$ by $500m$ grid using aerial stereo-photography. The main purpose of these photographs is to make a forest/non-forest decision so that field crew do not have to visit non-forest plots. Photos used for manual interpretation are taken during the leaves on season and usually occur 2 to 3 years before terrestrial evaluation, which occurs between March and November of each year. In the 2nd and 3rd NFI other continuous landscape variables such as canopy height information was also assessed at these plots. The basic sampling unit consists of a $50m$ by $50m$ square interpretation area each containing 25 equally spaced raster points arranged in a 5 by 5 design. Note that the canopy height information from this data source was only used in the disturbance correction for the one of the indirect three-phase models because no other

data was available. All other canopy height information was derived from the digital canopy height model. Further details about this data as well as disturbance correction terms are found in Chapter 3.

Digital canopy height model

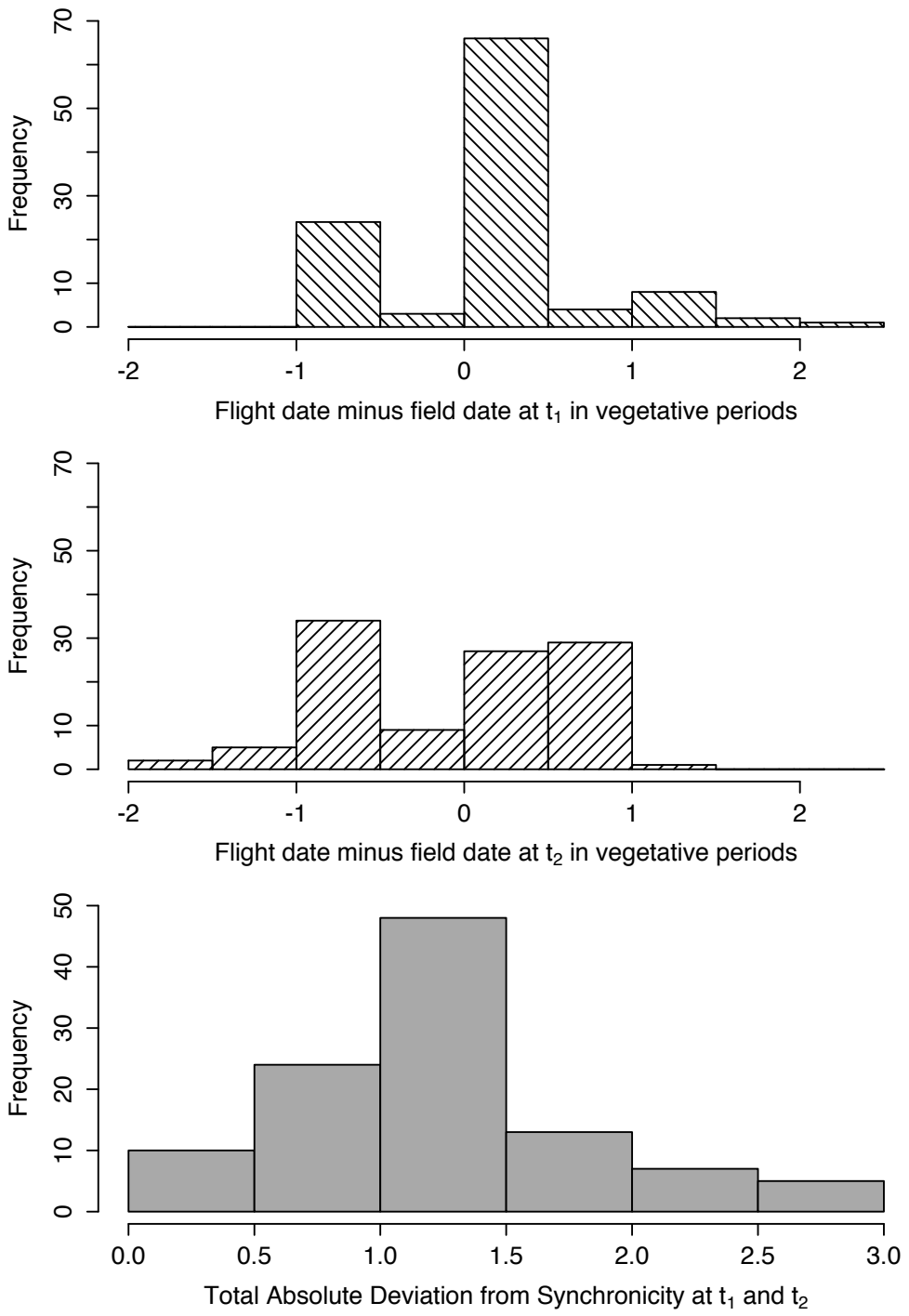
A canopy height model with a resolution of $1m$ by $1m$ was obtained by generating a digital surface model (DSM) using ADS40/ADS80 provided by the Swiss Federal Office of Topology (swisstopo) during the leaves-on season and subtracting a previously existing digital terrain model (DTM). Comprehensive details about the automated workflow and the stereo-image matching strategy is described in Ginzler and Hobi (2015). All plot-level summary statistics of the canopy characteristics are calculated using $25m$ by $25m$ interpretation areas centered over the nominal centers of the terrestrial plots. The main statistics of interest were the mean, median, and standard deviation of the canopy heights calculated over a green layer from the topographic map representing forested area. Pixels presumably describing canopy heights outside the range $0m$ to $55m$ were defined as outliers. Ideally, they should be completely removed as it is unlikely that the outlying measurement actually describes a canopy height. However, for this data set they were truncated to either $0m$ or $55m$.

The cycle of leaves-on measurements used to calculate the DSM takes a maximum of 6 years to get full nationwide coverage and is subject to the limitation that it is acquired over time by region. This makes for difficulties aligning temporally the canopy height measurements with the terrestrial NFI , which is currently acquired uniformly each year over the entire country under the new annual design. The main motivation behind restricting ourselves to the aforementioned study area is to achieve a better temporal alignment between the data sources. Histograms describing the temporal alignment in the study area at t_1 and t_2 are found in Figure 5.2. The total absolute deviation of synchronicity describes the sum of the absolute differences between the flight and field dates at t_1 and t_2 in vegetative periods and represents the amount of time where forest disturbance, either natural or manmade, could potentially lead to large discrepancies between the remote sensing prediction and the ground truth.

5.3.4 Sampling frame

From a sampling perspective it is best to have a forest versus non-forest decision available before the first phase sample selection in order to establish a well-defined sampling

Figure 5.2: Temporal asynchronicity between digital canopy height model and terrestrial measurements



Note:
The total absolute temporal deviation from synchronicity is calculated as the sum over t_1 and t_2 of the absolute value of the flight date of the aerial measurement minus the field date of the terrestrial measurement calculated in vegetative periods.

frame. However, since no map of forest land exists that conforms perfectly to the official Swiss NFI definition, we use the forest/non-forest decision obtained from manually interpreted aerial photos as opposed to the official decision that is only available for terrestrially measured plots. Although the aerial and terrestrial decisions are made independently, less than 1% of the sampling points have differing classifications. Usually aerial photo interpretation is considered as auxiliary information in a two-phase estimation procedure, but for this case study it is used primarily to create working definition of the common forested area between the third and fourth NFI. Thus all first, second and third phase plots were classified as non-bush forest by their corresponding manually interpreted aerial photos. It should also be noted that in Switzerland, as many other European countries, land that is classified as forest can not become non-forest by legal decision in most cases.

5.3.5 Target Variable

The target variable of interest is the change between times t_1 (i.e. 2004/2005) and t_2 (i.e. 2010/2011) of the growing stock volume ($VOL_\Delta := VOL_{t_2} - VOL_{t_1}$) in cubic meters per hectare ($\frac{m^3}{ha}$) of living trees excluding shrub forest in the common accessible forest. There is a threshold of 12 cm DBH for trees to be included in the estimates.

5.3.6 Auxiliary Variables

The main auxiliary variables of interest from remote sensing are derived from the mean canopy height at time t_k (MCH(t_k)), the median canopy height at time t_k (Q50CH(t_k)) and the standard deviation of the canopy height at time t_k (STD(t_k)) calculated over the forested part of the interpretation area. Transformations of these variables were considered such as taking the square and taking the absolute value. For the direct estimators various parameterizations are available depending whether one includes a single variable for the change between t_1 and t_2 or two separate variables. For example one may allow MCH(t_1) and MCH(t_2) to be included in the model independently resulting in two separate coefficients or one can take a single variable with value MCH(t_2)-MCH(t_1) resulting in a single coefficient. The terrestrial auxiliary variables of interest are derived from previous plot-level measurements of volume and stem count in the 2nd and 3rd NFIs.

5.3.7 Model Selection

The model selection was based on a combination of intuition and an exhaustive variable selection algorithm using the Mallows's Cp criterion. Initially, the selection algorithm was used to select models for the two-phase estimators using only remote sensing in the first phase. The three-phase models were chosen such that they include the entire selected two-phase model as well as a terrestrially derived variable based on the previous measurement and a disturbance correction term. The selected models are displayed in Table 5.1. For the indirect estimators the same structure of model variables was kept for each estimate of state at times t_1 and t_2 for simplicity. The same auxiliary variables were used in the kNN regression estimators.

5.3.8 Imputating correlation for the 4th Swiss National Forest Inventory

The *ad hoc* variance estimator for $\hat{\Delta}_{indir,2p}$ requires correlations to be estimated based on previous inventories. Past data for timber density was taken from all four Swiss NFIs. For every combination of remeasured plots, the number of vegetative cycles between remeasurements is calculated and rounded to the nearest integer. We call this the time lag (denoted LAG) and for each time lag the sample Pearson correlation coefficient can be calculated. For each of these correlation estimates, a lower bound of a 99% confidence interval can be calculated using Fisher's Z' transformation.

In order to make a conservative imputation for $\rho_{Y_{\Delta t}}$, one can model on the lower bound of the 99% confidence interval. The desired model should have an intercept at 1 and converge near to 0 as the years between remeasurements gets larger. This can be obtained by fitting the following linear model and backtransforming the logged response variable with the appropriate bias correction

$$log(\rho_{Y_{\Delta t}}) = \beta_0 + \beta_1 LAG + \epsilon \qquad [5.18]$$

Of course one would not expect the correlation structure over time to be the same across regions with different harvesting intensities. Regions with similar harvesting intensities should be grouped so that sufficient sample size is present for each LAG (for this study the sample size remained above 80 for each LAG). For Switzerland the regional groupings are Jura/Prealps, Plateau and Alps/South Alps. Using the same procedure as described before in each region one obtains the anticipated correlations presented in Figure 5.3. It should be noted that the variance of the point estimates for

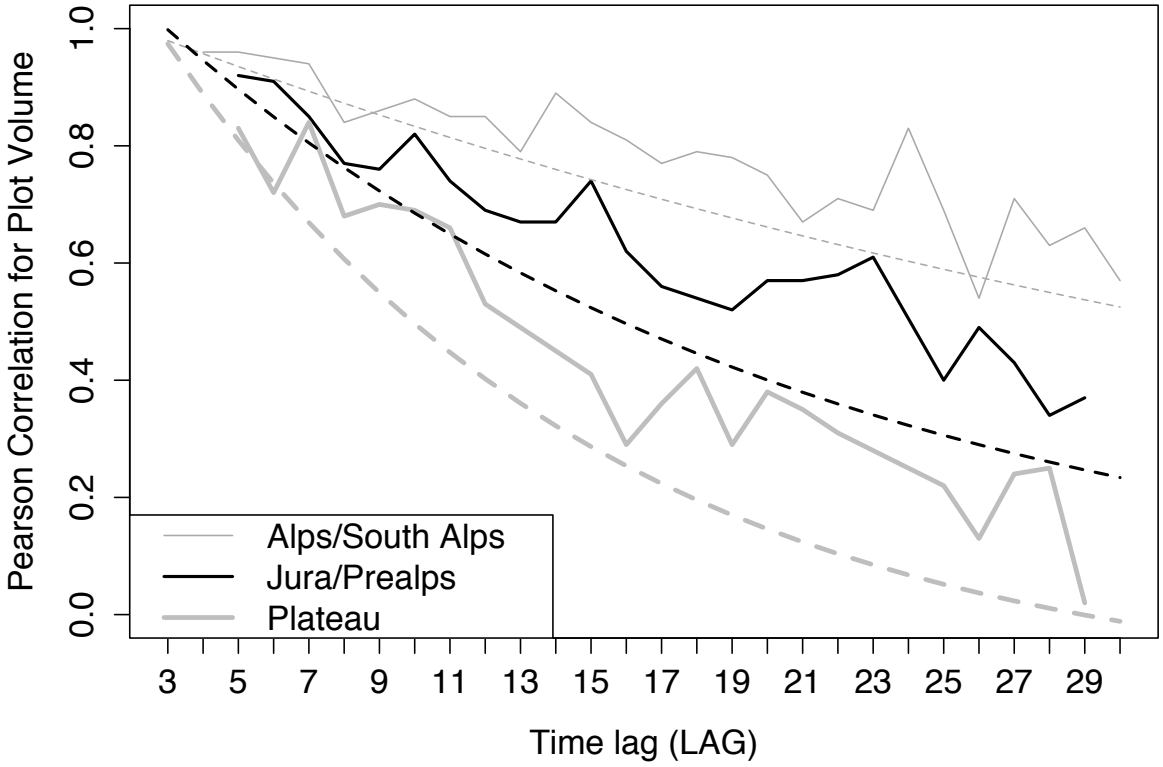
Table 5.1: Auxiliary variables selected for considered models

Estimator	$\mathbf{Z}^{(1)t}(x)$	$\mathbf{Z}^{(2)t}(x)$
$\hat{\Delta}_{dir,1p}(t_1,t_2)$	...	...
$\hat{\Delta}_{dir,2p}(t_1,t_2)$	$MCH_\Delta(t_1,t_2)$ $\lvert STD_\Delta(t_1,t_2)\rvert$	...
$\hat{\Delta}_{dir,3p}(t_1,t_2)$	$MCH_\Delta(t_1,t_2)$ $\lvert STD_\Delta(t_1,t_2)\rvert$	$PREV\_SC_\Delta(t_1,t_2)$
$\hat{\Delta}_{indir,1p}(t_1,t_2)$	...	...
$\hat{\Delta}_{indir,2p}(t_1,t_2)$	$MCH(t_k)$ $(MCH(t_k))^2$ $Q50CH(t_k)$ $(Q50CH(t_k))^2$ $STD(t_k)$	...
$\hat{\Delta}_{indir,3p}(t_1,t_2)$	$MCH(t_k)$ $(MCH(t_k))^2$ $Q50CH(t_k)$ $(Q50CH(t_k))^2$ $STD(t_k)$	$PREV\_VOL(t_k)$ $DIST(t_k)$

Legend:
$MCH_\Delta(t_1,t_2) := MCH(t_2) - MCH(t_1)$
$\lvert STD_\Delta(t_1,t_2)\rvert :=$ absolute value of $STD(t_2) - STD(t_1)$
$PREV\_SC_\Delta(t_1,t_2) :=$ change in stem count from the 2nd to 3rd Swiss NFI
$MCH(t_k) :=$ mean canopy height at time t_k for $k = 1, 2$
$Q50CH(t_k) :=$ median canopy height at time t_k for $k = 1, 2$
$STD(t_k) :=$ standard deviation of canopy height at time t_k for $k = 1, 2$
$PREV\_VOL(t_k) :=$ plot volume from NFI preceding time t_k for $k = 1, 2$
$DIST(t_k) :=$ the disturbance update associated with $PREV\_VOL(t_k)$ corresponding to the change in MCH since the previous inventory
Note:
The indirect estimator, $\hat{\Delta}_{indir,2p}(t_1,t_2)$, uses the same model variables for $\hat{Y}_{2p}(t_1)$ and $\hat{Y}_{2p}(t_2)$ (likewise for $\hat{\Delta}_{indir,3p}(t_1,t_2)$). Also remark the digital canopy height model is used to derive all variables except $DIST(t_1)$ which is based on the change in MCH from the manual interpretation data since the digital canopy height model only began in 2005.

given years seem to increase with time. This can be compensated for by using weighted least squares weights proportional to either LAG or LAG^2.

Figure 5.3: Correlations By Regional Grouping



Legend:
Solid lines represent the linear interpolation of estimated correlations for each time lag.
Dotted lines represent the conservative prediction generated by backtransforming a linear logged response model on the 99% lower bound of the confidence interval.

5.3.9 Estimators considered

The two- and three-phase direct and indirect estimators were compared for both scenarios one (i.e. $s_{31} = s_{32}$) and two (i.e. $s_{31} \cap s_{32} = \emptyset$). For each of these estimators and scenarios, both kNN and classical linear regression models were tested using the same auxiliary variables. The kNN models were fit using the function `kknn()` from the **R**-package *kknn* (Schliep and Hechenbichler (2014)). For comparability only Gaussian kernels were considered and the choice of k was made using leave-one-out cross validation. It should be noted that `kknn()` automatically scales all auxiliary variables to have equal standard deviations. The classical regression estimators were calculated using `lm()` and the bootstrap was calculated using `boot()` from the **R**-package *boot* (Canty and Ripley (2015)), which is easily adapted for parallel processing and very

computationally efficient. The bootstrap resampling was applied to the first phase only as described in Chapter 4. It should be noted that the bootstrap algorithm for double sampling proposed by Wolter (2007) was also tested but it did not differ substantially and was more involved to implement using the parallel features built into `boot()`.

5.3.10 Results

The estimates with their corresponding standard error estimates are found in Table 5.2. None of the point estimates are statistically different from each other. As the mean and the classical regression estimators are known to be (asymptotically) unbiased, this indicates some assurance that all estimators are in fact unbiased , even those using the kNN model which possesses less clear mathematical results. However, the direct estimator appears to be far more stable than the indirect estimator. The direct estimators where $s_{31} = s_{32}$ performed far more predictably compared to their indirect counterparts. Two phases are better than one phase and three phases is slightly better than two phases. The modest improvement over two phases is validated by the fact that the full model had only a slightly better R^2. The analogous indirect estimators however got worse when more auxiliary information was included in the models, despite exhibiting values of R^2 that seemed superior at first glance. Although these R^2's refer to how well the models predict state and not the target variable net change, some of them were as high as 0.92, which may seem counterintuitive that they actually led to less precise estimates than if no auxiliary information had been observed at all. Table 5.2 only displays the R^2's obtained by the linear models because the classical interpretation of R^2 as the percentage of the variance explained by the model is no longer valid for kNN. The indirect estimators in scenario two where net change on the plot level was never observed returned to the more intuitive pattern of two phase being better than one-phase but worse than three phase, but at the cost of significantly higher standard errors. This implies that if net change is observed then it should be calculated directly using three phases if possible.

The analytical standard errors derived under the external model assumption appear to be adequate when used with linear models. The external standard errors of the kNN estimators were consistently smaller than their bootstrap counterparts indicating an overall underestimation of their true design-based variance. The underestimation of the classical regression estimators was more negligible in magnitude which is to be expected considering that they have fewer parameters to estimate. Overall kNN did not seem to offer much of an advantage over classical regression, which might be expected

given that the goodness of fit for the direct models is neither exceptionally low nor high (Massey and Mandallaz (2015a)). The external variance using the *ad hoc* method also appears to be adequate as verified by the bootstrap.

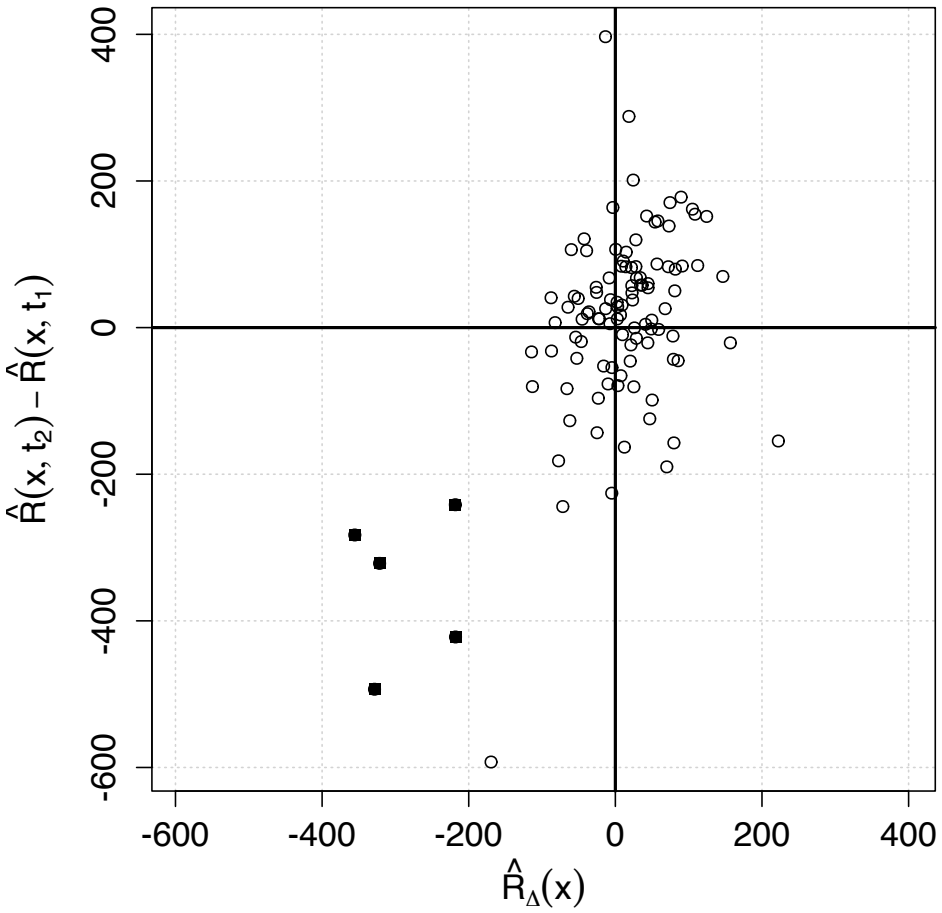
5.3.11 Discussion

The most disturbing feature of the results is the complete breakdown of the indirect estimator when $s_{31} = s_{32}$ for two and three phases, which did not occur for the direct versions. Consider the two-phase cases. Notice that the first terms in [5.8] and [5.14] and all sample sizes are equivalent so it suffices to examine only the residual variance terms to compare this estimator to the two-phase direct estimator. $\hat{R}(x, t_2) - \hat{R}(x, t_1)$ and $\hat{R}_{\Delta}(x)$ from the classical regression estimators are plotted against each other in Figure 5.4. We see that both the direct and indirect estimators performed poorly on many of the same plots and that the effect of these outliers is more pronounced for the indirect method. Overall, the indirect residual terms are more dispersed which is reasonable considering that the model fit is optimized to predict state rather than change directly. Furthermore, its lack of efficiency seems systematic for outlier and non-outlier plots. This indicates that if direct estimation is possible, it should be preferred to indirect estimation. In Figure 5.5 the direct model predictions for plot level-change are compared to the ground truth. Plots closer to the dotted line indicate better model predictions. Immediately it should be obvious that most plots grow modestly and a small minority of plots experience some kind of dramatic disturbance. The direct model predicted forest disturbance well for all except 5 plots, marked with solid dots. It should be noted that the direct two-phase model's failure to detect forest disturbances on these 5 plots resulted in a 33% increase in standard error. Even on a handful of final phase plots, failed disturbance detection can be particularly detrimental to an estimator's precision. As a result, factors contributing to failed disturbance detection must not be ignored, in particular imprecise locations for plot coordinates and temporal asynchronicity between the remote sensing and terrestrial measurements.

5.4 Conclusions

Two- and three-phase regression estimation shows promise for improving the precision of net change estimation in the design-based context. If the change is observed on the plot level the direct method should be employed over the indirect method. The indirect

Figure 5.4: Plot-level comparison of direct versus indirect residual terms for the two-phase classical regression estimators

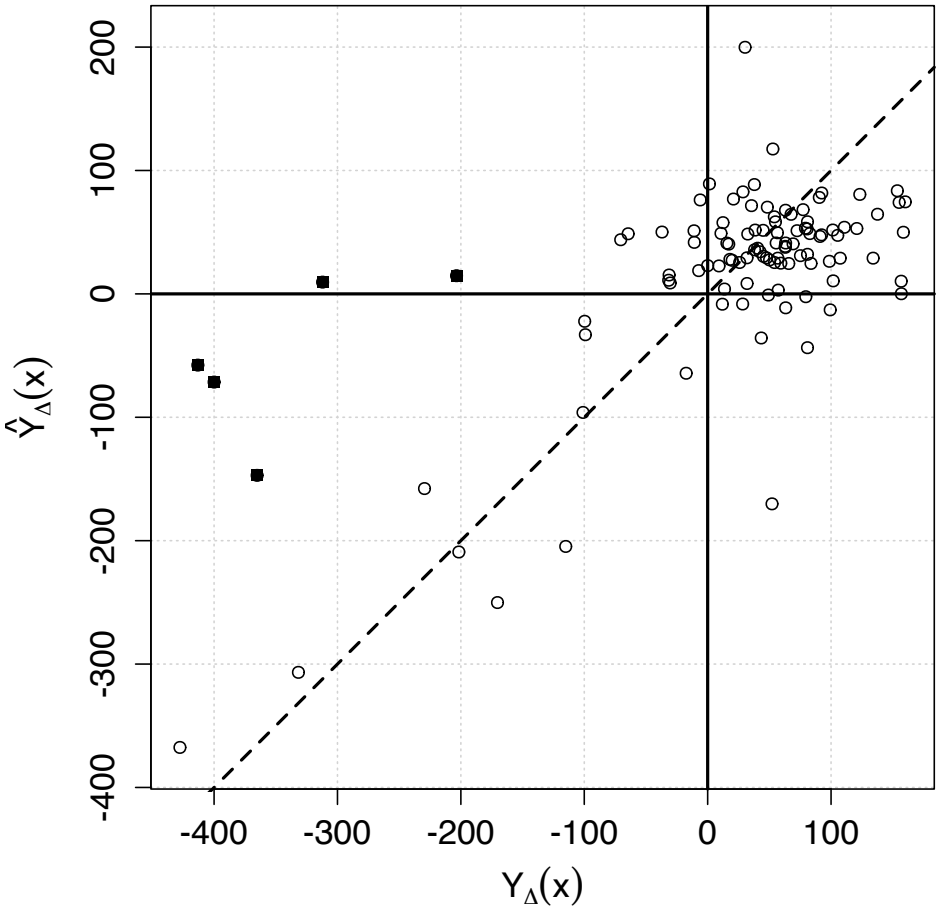


Legend:
Solid dots represent outlying plots where major disturbance went undetected by the direct model.
$\hat{R}_\Delta(x)$ corresponds to the direct estimator whereas $\hat{R}(x, t_2) - \hat{R}(x, t_1)$ corresponds to the indirect estimator.

estimators can still be helpful when direct estimation is impossible, which would be the case for inventories with temporary plots or annual designs where the target time period is not a multiple of the remeasurement cycle. The proposed *ad hoc* procedure for imputing the correlation from previous inventories appears to be a reasonable option for circumventing the issue of not observing plot-level remeasurements, but to remain purely design-based, the bootstrap must be used.

For choosing a model, both kNN and linear models appear to offer the same magni-

Figure 5.5: Plot-level predictions versus ground truth for the two-phase direct classical regression estimator



Legend:
Solid dots represent outlying plots where major disturbance went undetected by the model.

tudes of precision improvement. However, kNN should be used in conjunction with a bootstrap variance estimator where the choice of k and kernel are fixed across all bootstrap samples. The advantage of using linear models via the two- and three-phase classical regression estimators is that the analytical variance estimates derived under the external variance assumption appear to remain valid. The existence of reliable analytical variance estimators allows for simplified implementation and exactly reproducible results.

The issue of temporal alignment between remote sensing and terrestrial measurements

appears to be of greater importance in change estimation than it is in state estimation as the failed detection of even a small number of plots with large disturbances have the potential to dramatically inflate the variance. This creates a dilemma for large-scale forest inventories that have operational incentives to collect remote sensing data regionally rather than uniformly across the entire country because establishing good temporal synchronicity requires harmonizing aerial and terrestrial campaigns.

Table 5.2: Point estimates and standard errors for net change ($\frac{m^3}{ha}$)

Estimator	Model	Scenario	Estimate	Ext s.e.	Boot s.e.		k	R^2
$\hat{\Delta}_{dir,1p}(t_1,t_2)$	–	One	15.22	11.65	11.54		–	–
$\hat{\Delta}_{dir,2p}(t_1,t_2)$	reg	One	19.27	8.69	9.54		–	0.46
$\hat{\Delta}_{dir,3p}(t_1,t_2)$	reg	One	13.09	8.67	8.76		–	0.46, 0.48
$\hat{\Delta}_{dir,2p}(t_1,t_2)$	kNN	One	14.40	7.43	9.14		9	–
$\hat{\Delta}_{dir,3p}(t_1,t_2)$	kNN	One	13.14	6.82	8.60		9, 5	–
$\hat{\Delta}_{indir,1p}(t_1,t_2)$	–	One	15.22	11.65	11.54		–	–
$\hat{\Delta}_{indir,2p}(t_1,t_2)$	reg	One	2.73	13.63	12.74	t_1:	–	0.37
						t_2:	–	0.39
$\hat{\Delta}_{indir,3p}(t_1,t_2)$	reg	One	-4.74	14.19	16.91	t_1:	–	0.37, 0.78
						t_2:	–	0.39, 0.92
$\hat{\Delta}_{indir,2p}(t_1,t_2)$	kNN	One	4.41	12.88	14.68	t_1:	10	–
						t_2:	7	–
$\hat{\Delta}_{indir,3p}(t_1,t_2)$	kNN	One	-1.88	11.33	16.01	t_1:	10, 6	–
						t_2:	7, 5	–
$\hat{\Delta}_{indir,1p}(t_1,t_2)$	–	Two	20.75	38.61	38.51		–	–
$\hat{\Delta}_{indir,2p}(t_1,t_2)$	reg	Two	15.64	31.09*	32.30	t_1:	–	0.28
						t_2:	–	0.15
$\hat{\Delta}_{indir,3p}(t_1,t_2)$	reg	Two	-11.59	–	18.40	t_1:	–	0.28, 0.71
						t_2:	–	0.15, 0.70
$\hat{\Delta}_{indir,2p}(t_1,t_2)$	kNN	Two	11.62	28.03*	31.30	t_1:	10	–
						t_2:	7	–
$\hat{\Delta}_{indir,3p}(t_1,t_2)$	kNN	Two	4.29	–	21.30	t_1:	10, 6	–
						t_2:	7, 5	–

Legend:
Model "reg" indicates that Classical Regression Estimators were calculated.
Model "kNN" indicates that kNN Regression Estimators were calculated using with Gaussian kernels.
Scenario One indicates $s_{31} = s_{32}$.
Scenario Two indicates $s_{31} \cap s_{32} = \emptyset$.
Ext s.e. corresponds to the analytical standard errors derived under the external model assumption.
Boot s.e. corresponds to the bootstrap standard errors with 25000 replications.
* *ad hoc* variances using an imputed $\hat{\rho} = 0.67$.
Remark: No suitable analogy for R^2 exists for kNN which is why it is not reported.

Chapter 6

Synthesis

6.1 Main findings

The proposed three-phase classical regression estimator where previous measurement information is updated for growth and disturbance and integrated with canopy height information derived from airborne measurements shows great promise for reducing the increase in variance associated with the transition of the Swiss NFI to the annual design. In controlled tests using previously completed periodic inventories, the increase of variance for the estimation of timber volume was reduced from 294% to 145% compared to the estimator based on the full periodic sample while remaining unbiased. Comparing the three-phase estimator to the two-phase estimator revealed that the previous measurement is a powerful predictor that can be used to enhance auxiliary information derived from remote sensing.

Practically speaking, the classical regression estimators are preferable to nonparametric kernel-based regression estimators for the following reasons:

1. The findings suggest that if the model captures the main features of the underlying process, then it is advisable to use the classical regression estimator because it performs at least as well as other nonparametric kernel-based estimators including kNN.

2. The g-weight variance estimator helps to account for the effect of fitting a linear model internally and exists in a closed-form for both two- and three-phases which makes it easy to implement.

3. The analytical variance estimators for the considered nonparametric models were derived by falsely assuming that the models were externally fit. This false assump-

tion under moderate sample size can lead to underestimation of the theoretical variance thus necessitating a bootstrap procedure to be considered.

4. The aforementioned underestimation associated with external model approach empirically seems to be less pronounced for linear models. Furthermore, it can actually be proven analytically that the external variance estimator and the g-weight variance estimator for linear models are asymptotically equivalent.

5. Many well-known variance reduction techniques such as double sampling for post-stratification are simply special cases of classical regression estimators.

6. All results can be easily generalized to cluster sampling in the Monte Carlo approach (see Mandallaz (2013c)).

As for the estimation of net change, direct estimation is preferable whenever it is possible. When direct estimation is not possible, such as for temporary plot designs, indirect estimation can still improve precision. In any case, the temporal alignment of the remote sensing measurements with the terrestrial measurements appears to pose more problems for the estimation of net change than it for the estimation of state.

6.2 Limitations and criticisms

A potential limitation for the proposed three-phase technique for estimating timber volume is that the second phase based on full terrestrial sample is technically not drawn using simple random sampling (SRS) as is mathematically assumed. In fact, once the transition to the annual design is complete then the third phase will always consist of previous measurements that are older than the second phase plots not in the third phase. It should be noted that it is common practice to apply estimators that assume SRS when it is not in practice (e.g. with systematic grids). The implications of this assumption violation for this particular estimator is not as well understood. It could be conjectured that practically speaking this would only be a problem in situations where there were strong growth or loss trends over the entire remeasurement cycle. Adequate growth and disturbance updates would likely include a temporal variable that could control for any bias, but as of yet this has not been experimentally tested. Diagnostic procedures for identifying pathological growth or loss trends is a potential area for further work.

Another potential criticism is that the proposed Monte Carlo approach depends on a well-defined forested area from which the first phase is sampled. This means defining

the forested area *prior* to the first phase selection, probably derived by a green layer generated by remote sensing. However other forest definitions, e.g. ones used for legal purposes, cannot be expected to always match the remote sensing classification.

The final criticism is more cosmetic in nature and concerns the non-additivity of the estimates. Additivity can make sense for totals, but does not have an immediate connection to densities, which was the target variable here. While it is sometimes possible to convert densities into additive totals if the forest areas for the different stratification classes are known (e.g. to make regionally additive estimates), this is frequently not the case. For example, some plots may contain both coniferous and deciduous trees which means that the estimates of total timber volume of conifers and deciduous trees cannot be made additive through the use of stratification. It needs to be noted that imposing additivity post-estimation is counterproductive if the goal is to optimize precision.

6.3 Implications

Despite these limitations, the Monte Carlo framework for the multiphase regression estimators is well-established, corresponds closely with the existing finite population sampling theory and is simpler to implement. Given the availability of high quality remote sensing data, multiphase regression estimation in the design-based Monte Carlo approach offers generality in terms of the types of models that can be applied and robustness in terms of model misspecification. Furthermore, the classical regression estimator is analytically the best understood from a mathematical perspective and can be easily implemented with common statistical software such as **R** and SAS. Since the inference is design-based, automated model selection procedures can be applied with minimal concern about variable selection, which allows for automated workflow to produce estimates.

Finally, if forest inventorists are inclined to use another type of model for whatever reason, they can always bootstrap or examine the adequacy of violating the external model assumption using an artificial simulation.

Appendix

We now give an example of the derivation of one of the indirect net change estimators, $\hat{\Delta}_{indir,2p}$, as an outline of the technique by which the theoretical expectation and variances of all the indirect estimators can be derived. For this example only scenario two is considered, which means that we assume s_2 consists of the same sample of points at t_1 and t_2, $s_{31} \cap s_{32} = \emptyset$, and $t_1 < t_2$. Recall that s_2 is the result of uniform independent sampling from F and s_{3k} is SRS without replacement from s_2. Working under external model assumption, we want to demonstrate the two following results:

$$\begin{aligned}
\mathbb{E}_{2,3}(\hat{\Delta}_{indir,2p}) &= \mathbb{E}(\hat{Y}_{2p}(t_2)) - \mathbb{E}(\hat{Y}_{2p}(t_1)) = \bar{Y}(t_2) - \bar{Y}(t_1) \\
\mathbb{V}_{2,3}(\hat{\Delta}_{indir,2p}) &= \frac{1}{n_2}\mathbb{V}(Y(t_2,x) - Y(t_1,x)) \\
&\quad + (1 - \frac{n_{31}}{n_2})\frac{1}{n_{31}}\mathbb{V}(R(t_1,x)) + (1 - \frac{n_{32}}{n_2})\frac{1}{n_{32}}\mathbb{V}(R(t_2,x))
\end{aligned}$$

For the first result, it suffices to show that $\mathbb{E}(\hat{Y}_{2p}(t_k)) = \bar{Y}(t_k)$ for $k \in 1, 2$.

$$\begin{aligned}
\mathbb{E}_{2,3}(\hat{Y}_{2p}(t_k)) &= \mathbb{E}_2\mathbb{E}_{3|2}(\hat{Y}_{2p}(t_k)) \\
&= \mathbb{E}_2\Big(\frac{1}{n_2}\sum_{x\in s_2}\hat{Y}(x) + \mathbb{E}_{3|2}\Big(\frac{1}{n_{3k}}\sum_{x\in s_{3k}}\big(Y(x) - \hat{Y}(x)\big)\Big)\Big) \\
&= \mathbb{E}_2\Big(\frac{1}{n_2}\sum_{x\in s_2}\hat{Y}(x)\Big) + \mathbb{E}_2\Big(\frac{1}{n_2}\sum_{x\in s_2}Y(x)\Big) - \mathbb{E}_2\Big(\frac{1}{n_2}\sum_{x\in s_2}\hat{Y}(x)\Big) \\
&= \bar{Y}(t_k)
\end{aligned}$$

So $\hat{\Delta}_{indir,2p}$ is exactly unbiased under the external model assumption. For the theoretical variance we will use the decomposition
$\mathbb{V}_{2,3}(\hat{\Delta}_{indir,2p}) = \mathbb{V}_2(\mathbb{E}_{3|2}(\hat{\Delta}_{indir,2p})) + \mathbb{E}_2(\mathbb{V}_{3|2}(\hat{\Delta}_{indir,2p}))$. For the first term we get

$$\mathbb{V}_2(\mathbb{E}_{3|2}(\hat{\Delta}_{indir,2p})) = \mathbb{V}_2\Big(\frac{1}{n_2}\sum_{x\in s_2}Y(t_2,x) - \frac{1}{n_2}\sum_{x\in s_2}Y(t_1,x)\Big) = \frac{1}{n_2}\mathbb{V}(Y(t_2,x) - Y(t_1,x))$$

The subscripted 2 in the variance and expectation is shorthand emphasizing the sampling procedure generating the first phase in F and is dropped for notational

simplicity in the final result. For the second term we get

$$\begin{aligned}
\mathbb{E}_2(\mathbb{V}_{3|2}(\hat{\Delta}_{indir,2p})) &= \mathbb{E}_2\Big(\mathbb{V}_{3|2}\Big(\frac{1}{n_2}\sum_{x\in s_2}\big(\hat{Y}(t_2,x)-\hat{Y}(t_1,x)\big) \\
&\quad + \frac{1}{n_{32}}\sum_{x\in s_{32}} R(t_2,x) - \frac{1}{n_{31}}\sum_{x\in s_{31}} R(t_1,x)\Big)\Big) \\
&\quad + \mathbb{E}_2\Big(\mathbb{V}_{3|2}\Big(\frac{1}{n_{32}}\sum_{x\in s_{32}} R(t_2,x)\Big)\Big) + \mathbb{E}_2\Big(\mathbb{V}_{3|2}\Big(\frac{1}{n_{31}}\sum_{x\in s_{31}} R(t_1,x)\Big)\Big) \\
&= (1-\frac{n_{31}}{n_2})\frac{1}{n_{31}}\mathbb{V}(R(t_1,x)) + (1-\frac{n_{32}}{n_2})\frac{1}{n_{32}}\mathbb{V}(R(t_2,x))
\end{aligned}$$

The analogous derivation for three-phases follows similar arguments.

Bibliography

Bafetta, F., Corona, P., and Fatottorini, L. (2011). Design-based diagnostics for k-NN estimators of forest resources. *Can. J. For. Res.*, **41**:pp. 59–72.

Bafetta, F., et al. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remote sensed data in forest inventory. *Remote Sensing of Environment*, **113**:pp. 463–479.

Brassel, P. and Lischke, H. (2001). Swiss National Forest Inventory: Methods and Models of the Second Assessment. Technical report, WSL Swiss Federal Research Institute, Birmensdorf.

Breidenbach, J. and Nothdurft, A. (2010). Comparison of nearest neighbours approaches for small area estimation of tree species specific forest inventory attributes in central Europe using airborne laser scanner data. *Eur. J. Forest Res.*, **129**:pp. 833–846.

Canty, A. and Ripley, B. D. (2015). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-15.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics.

Dixon, B. and Howitt, R. (1979). Continuous Forest Inventory Using a Linear Filter. *Forest Science*, **20**(4):pp. 675–689.

Eskelson, B., Temesgen, H., and Barrett, T. (2009). Estimating Current Forest Attributes from Paneled Inventory Data Using Plot-Level Imputation: A Study from the Pacific Northwest. *Forest Science*, **55**.

Gasser, T. and Mueller, H. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, **11**:pp. 171–185.

Ginzler, C. and Hobi, M. (2015). Countrywide Stereo-Image Matching for Updating Digital Surface Models in the Framework of the Swiss National Forest Inventory. *Remote Sensing*, **7**:pp. 4343–4370.

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

Gregoire, T., et al. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. *Canadian Journal of Forest Research*, **41**:pp. 83–95.

Györfi, L., et al. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Haara, A. and Kangas, A. (2012). Comparing K Nearest Neighbours Methods And Linear Regression – Is There Reason To Select One Over The Other? *Mathematical and Computational Forestry And Natural-Resource Sciences*, **4**:pp. 50–65.

Hayfield, T. and Racine, J. S. (2008). Nonparametric Econometrics: The np Package. *Journal of Statistical Software*, **27**(5).

Hechenbichler, K. and Schliep, K. (2004). Weighted k-Nearest Neighbor Techniques and Ordinal Classification. Technical report, University of Munich and Massey University.

Herrmann, E. and Maechler, M. (2014). *lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth*. R package version 1.1-6.

Houillier, F. and Pierrat, J. (1992). Application des modèles statistiques spatio-temporels aux èchantillonnages forestiers successifs. *Canadian Journal of Forest Research*, **20**:pp. 1988–1995.

Jennen-Steinmetz, C. and Gasser, T. (1988). A Unifying Approach to Nonparametric Regression Estimation. *Journal of The American Statistical Association*, **83**:pp. 1084–1089.

Johnson, D., Williams, M., and Czaplewski, R. (2003). Comparison of estimators for rolling samples using forest Inventory and analysis data. *Forest Science*, **49**:pp. 50–63.

Keller, W. (1978). Einfacher ertragskundlicher Bonitätsschlüssel für Waldbestände in der Schweiz. Technical Report 54, Eidgenössische Anstalt für das forstliche Versuchswesen.

Lanz, A., et al. (2010). Switzerland. In E. Tomppo, T. Gschwantner, M. Lawrence, R. McRoberts, and Winzeler, editors, *National Forest Inventories Pathways for Common Reporting*, chapter 36, pp. 555–565. Springer.

Lehmann, E. (1999). *Elements of Large-Sample Theory*. Springer Texts in Statistics, New York.

Lessard, V., McRoberts, R., and Holdaway, M. (2001). Imputation on model-based updating techniques for annual forest inventories. *Forest Science*, **47**:pp. 301–310.

Lüpke, N., Hansen, J., and Saborowski, J. (2012). A three-phase sampling procedure for continuous forest inventory with partial re-measurement and updating of terrestrial sample plots. *European Journal of Forest Research*, **131**:pp. 1979–1990.

Magnussen, S., McRoberts, R. E., and Tompo, E. (2010). A resampling variance estimator for the k nearest neighbours technique. *Can. J. For. Res.*, **40**:pp. 648–658.

Magnussen, S. and Tomppo, E. (2014). The k-nearest neighbor technique with local linear regression. *Scandinavian Journal of Forest Research*, **29**:pp. 120–131.

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Chapman and Hall, Boca Raton FL.

Mandallaz, D. (2013a). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research*, **43**:pp. 441–449.

Mandallaz, D. (2013b). New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. *Canadian Journal of Forest Research*, **43**:pp. 1023–1031.

Mandallaz, D. (2013c). Regression estimators in forest inventories with three-phase sampling and two multivariate components of auxiliary information. Technical report, ETH Zurich, Department of Environmental Systems Science. Available from http://e-collection.library.ethz.ch.

Mandallaz, D. (2014). A three-phase sampling extension of the generalized regression estimator with partially exhaustive information. *Canadian Journal of Forest Research*, **44**(4):pp. 383–388.

Mandallaz, D. and Massey, A. (2015). Regression and non-parametric estimators for two-phase forest inventories in the design-based Monte Carlo approach. Technical report, ETH Zurich, Department of Environmental Systems Science, http://e-collection.library.ethz.ch.

Massey, A. and Mandallaz, D. (2015a). Comparison of classical, kernel-based and nearest neighbors regression estimators using the design-based Monte Carlo approach for two-phase forest inventories. *Canadian Journal of Forest Research*, **42**:pp. 1480–1488.

Massey, A. and Mandallaz, D. (2015b). Design-based regression estimation of net change for forest inventories. *Canadian Journal of Forest Research*.

Massey, A., Mandallaz, D., and Lanz, A. (2014). Integrating remote sensing and past inventory data under the new annual design of the Swiss National Forest Inventory using three-phase design-based regression estimation. *Canadian Journal of Forest Research*, **44**(10):pp. 1177–1186.

McRoberts, R. (2001). Imputation on model-based updating techniques for annual forest inventories. *Forest Science*, **47**:pp. 322–330.

McRoberts, R., et al. (2007). Estimating areal means and variance of forest attributes using the k-Nearest Neighbors technique and satellite imagery. *Remote Sensing of Environment*, **111**:pp. 466–480.

McRoberts, R., et al. (2011). Parametric bootstrap and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. *Remote Sensing of Environment*, **115**:pp. 3165–3174.

McRoberts, R., et al. (2015). Indirect and direct estimation of forest biomass change using forest inventory and airborne laser scanning data. *Remote Sensing of Environment*, **164**:pp. 36–42.

McRoberts, R. E. (2012). Estimating forest attributes parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management*, **272**:pp. 3–12.

Moeur, M. and Stage, A. (1996). Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science*, **41**:pp. 337–359.

Särndal, C., Swenson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling.*. Springer Series in Statistics, New York.

Schliep, K. and Hechenbichler, K. (2014). *kknn: Weighted k-Nearest Neighbors*. R package version 1.2-5.

Scott, C., Köhl, M., and Schnellbächer, H. J. (1999). A Comparison of Periodic and Annual Forest Surveys. *Forest Science*, **45**(3):pp. 433–451.

Tomppo, E., et al. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, **112**:pp. 1982–1999.

Van Deusen, P. (1997). Annual first inventory statistical concepts with emphasis on multiple imputation. *Canadian Journal of Forest Research*, **27**:pp. 379–384.

Van Deusen, P. (2002). Comparison of some annual forest inventory estimators. *Canadian Journal of Forest Research*, **32**:pp. 1992–1995.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer Science+Business Media, New York, second edition.

Wu, C. and Sitter, R. (2001). A Model Calibration Approach to Using Complete Auxiliary Information From Survey Data. *Journal of the American Statistical Association*, **96**:pp. 185–193.