Comparison of estimators in one-phase two-stage Poisson sampling in forest inventories

Master Thesis

Author(s):
Massey, Alexander

Publication date:
2011

Permanent link:
https://doi.org/10.3929/ethz-a-006380897

Master Thesis:

Comparison of Estimators in One-Phase Two-Stage Poisson Sampling in Forest Inventories

Alexander Massey

February 21, 2011

Referees:
PD Dr. Daniel Mandallaz, Department of Environmental Sciences, and
Prof. Dr. Werner Stahel, Department of Mathematics, ETH Zurich

Abstract

This thesis investigates the performances of various estimators in one-phase (purely terrestrial) two-stage forest inventories, where trees in the first-stage are selected by concentric circles (approximate PPS) and a subset thereof are selected by Poisson sampling for further measurements to get an accurate estimation of the timber volume. Poisson sampling is used because it is easy to implement in field work. However, this comes with the drawback of a random second-stage sample size that can drive up the variance. The widely accepted remedy in analogous situations in survey sampling is to add a stabilizing factor to the estimator that compensates for this randomness and presumably lowers the variance. In this paper the effectiveness of three formulations of such stabilizing factors are examined in the context of timber density estimation. These factors are applied to the residual component of a generalized two-stage density estimator and tested using data from the 3rd Swiss National Forest Inventory taken in 2003. These factors introduce a negligible bias. Contrary to empirical findings in general survey sampling and asymptotic results, the adjusted estimators did not perform really better than the original unadjusted two-stage estimator.

Acknowledgements

I would like to thank Professor Werner Stahel, Department of Mathematics, ETH Zurich and PD Dr. Daniel Mandallaz, Department of Environmental Sciences, ETH Zurich for their role as referees, as well as Edgard Kaufmann from the Swiss Research Institute for Snow, Forest and Landscape in Birmensdorf for his support with data management aspects pertaining to the Swiss National Forest Inventory. I also would like to thank Dr. Richard Valliant from the Joint Program in Survey Methodology at the University of Maryland for his helpful advice concerning replication variance estimation and members of the Current Population Survey Branch at the US Census Bureau for their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many applications costs to measure the response variable Y_i are high. For instance, in forest inventory a good determination of the volume may require that one records the diameter at breast height, DBH ($1.3m$ above ground), as well as the diameter at $7m$ above ground (D_7) and total height (H) in order to utilize a three-way yield table. However, one could rely on a coarser, but cheaper, approximation of the volume based only on DBH. Nonetheless, it may be most sensible to assess those three parameters only on a sub-sample of trees. We now formalize this simple idea.

For each point $x \in s_2$ trees are drawn with probabilities π_i. The set of selected trees is denoted by $s_2(x)$. From each of the selected trees $i \in s_2(x)$ one gets an approximation $\tilde{Y}_i$ of the exact value Y_i. From the finite set $s_2(x)$ one draws a subsample $s_3(x) \subset s_2(x)$ of trees. For each tree $i \in s_3(x)$ one then measures the exact variable Y_i. Let us now define the second stage indicator variable

$$J_i(x) = \begin{cases} 1 \text{ if } i \in s_3(x) \\ 0 \text{ if } i \notin s_3(x) \end{cases} \tag{1.1}$$

In our context, the sub-index 1 refers to the first phase in which one collects the auxiliary information in the large sample s_1 (this will not be discussed here). The sub-index 2 refers to the second phase, when one gathers the terrestrial information from first-stage trees. Finally, the sub-index 3 refers to the sampling procedure of the second-stage trees out of the first-stage trees. We shall use the notation $\mathbb{E}_{2,3}(.), \mathbb{V}_{2,3}(.), \mathbb{E}_{3|2}(.), \mathbb{V}_{3|2}(.)$ for the overall expectation and variance under the random selections $(2,3)$, as well as for the conditional expectation and variance of the second stage procedure, given the second-phase and first-stage selections. We recall that we can calculate the expected value and variance of an arbitrary random variable Z depending on a random selection $(2,3)$ with

$$\begin{aligned} \mathbb{E}_{2,3}(Z) &= \mathbb{E}_2\left(\mathbb{E}_{3|2}(Z)\right) \\ \mathbb{V}_{2,3}(Z) &= \mathbb{E}_2\left(\mathbb{V}_{3|2}(Z)\right) + \mathbb{V}_2\left(\mathbb{E}_{3|2}(Z)\right) \end{aligned} \tag{1.2}$$

Hence, because $I_i(x)J_i(x) = J_i(x)$, we arrive at

$$\begin{aligned} \mathbb{E}_{2,3}(J_i(x)) &= \mathbb{E}_2\mathbb{E}_{3|2}(J_i(x)I_i(x)) \\ &= \mathbb{E}_2 I_i(x)\mathbb{E}_{3|2}(J_i(x)|I_i(x)) \\ &= \mathbb{P}(J_i(x) = 1|I_i(x) = 1)\mathbb{P}(I_i(x) = 1) := p_i\pi_i \end{aligned} \tag{1.3}$$

$\pi_i =: \frac{\lambda(K_i \cap F)}{\lambda(F)}$ are the first-stage inclusion probabilities such that $\lambda(K_i \cap F)$ is the inclusion area of the i-th tree, possibly adjusted for boundary effects ($\lambda(G)$ denotes

the surface area of any subset G in the plane). The **second-stage** conditional inclusion probabilities are $p_i = \mathbb{P}(J_i(x) = 1 | I_i(x) = 1)$.

We assume that trees in $s_2(x)$ are sampled **independently of each other**, so that $p_{ij} = \mathbb{P}(J_i(x)J_j(x) = 1 | I_i(x)I_j(x) = 1) = p_i p_j$. Thus, we have **Poisson sampling** at the second stage. The advantage of this proposed scheme is that a field crew can collect the required information on first-stage trees one by one, then enter these data in a portable computer. Using a software generating appropriate random numbers, the crew can then determine immediately whether further measurements will be taken. Of course, other schemes are possible, but these are not necessarily better nor as easily implemented because one needs a list of all first-stage trees at point x and, possibly, at others too.

Chapter 2

Point estimates

To construct better point estimates, we must use **the residuals** $R_i = Y_i - \tilde{Y}_i$ which are known only for trees $i \in s_3(x)$. The **true local density** corresponds formally to exhaustive second stage sampling, i.e. $p_i \equiv 1$, and is given by

$$Y(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) Y_i}{\pi_i} \tag{2.1}$$

Note that $\frac{1}{\lambda(F)\pi_i}$ is the tree extrapolation factor f_i with dimension $\frac{1}{ha}$.
The local density based solely on the approximation $\tilde{Y}_i$ is defined as

$$Y_0(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) \tilde{Y}_i}{\pi_i} \tag{2.2}$$

The local density of the residuals is defined as

$$R(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) R_i}{\pi_i} \tag{2.3}$$

which plays a purely theoretical role because it is not directly observable (though it can be estimated).

We shall consider three procedures to estimate $R(x)$

$$\begin{aligned} R_1(x) &= \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) J_i(x) R_i}{\pi_i p_i} \\ R_2(x) &= R_1(x) \frac{\sum_{i=1}^{N} I_i(x) p_i}{\sum_{I=1}^{N} I_i(x) J_i(x)} \\ R_3(x) &= R_1(x) \frac{\sum_{i=1}^{N} I_i(x)}{\sum_{i=1}^{N} \frac{I_i(x) J_i(x)}{p_i}} \end{aligned} \tag{2.4}$$

and the corresponding adjusted estimates at point x for $k = 1, 2, 3$

$$Y_k(x) = Y_0(x) + R_k(x) \tag{2.5}$$

$Y_1(x)$ is the generalized local density introduced by D. Mandallaz (2008). The other two estimates are adapted from suggestions of Grosenbaugh (see e.g. H.T Schreuder et al. 1968) and Särndal et al. (2003) in the context of one-stage sampling. It is understood that the residual terms $R_k(x)$ are set to zero whenever

$n_2(x) = \sum I_i(x)J_i(x) = 0$, i.e. when no second-stage trees are drawn, which usually occurs if all trees at x have very small DBH because the optimal schemes rest upon $\lambda(F)\pi_i p_i \propto \mid R_i \mid$ (Probability Proportional to Error, D. Mandallaz (2008)). One drawback of Poisson sampling is that the random sample size inflates the variance: for exact PPE the variance of $R_1(x)$ depends only on the sample size. Several theoretical and empirical investigations in one-phase one-stage sampling suggest that the variance of the modified estimate $R_2(x)$ and $R_3(x)$ can be expected to be smaller. The correction term for $Y_2(x)$ is intuitively appealing, as it adjusts for the difference between expected and observed sample sizes at the second stage. It is less intuitive for $Y_3(x)$ because one estimates the first-stage population's size, which is $\sum_{i=1}^N I_i(x)$ and therefore known!

Since $Y_i = \tilde{Y}_i + R_i$, one clearly has $\mathbb{E}_2\mathbb{E}_{3|2}Y_1(x) = \frac{1}{\lambda(F)}\sum_{i=1}^N Y_i = \bar{Y}$, so that $Y_1(x)$ is an unbiased estimate, even if the prediction model is not correct, i.e. if $\frac{1}{\lambda(F)}\int_F R(x)dx \neq 0$. This is generally not the case for $Y_0(x)$, which can be severely biased, particularly for small area estimation because the prediction model for the Y_i^* can lead locally to substantial overestimation as well as underestimation.

Given a sample s_2 of n_2 points uniformly and independently distributed in F, we define the following point estimates

$$\hat{Y}_k = \frac{1}{n_2}\sum_{x\in s_2} Y_k(x) \;\; k = 0, 1, 2, 3 \tag{2.6}$$

2.1 Bias

We shall show that $Y_2(x)$ and $Y_3(x)$ have a bias, which can be expected to be negligible in practice. In the following, we shall primarily use heuristic arguments as exact calculations are not available. In particular, we will frequently use the first order Taylor approximation $\mathbb{E}(\frac{X}{Y}) \approx \frac{\mathbb{E}(X)}{\mathbb{E}(Y)}$ for the expectation of the ratio of random variables and likewise for a product $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, particularly with respect to $\mathbb{E}_2$, and also second order Taylor expansions for $\mathbb{E}_{3|2}$. For $R_2(x)$ and $R_3(x)$ we have to work conditionally on the events $\{n_2(x) > 0\} = \{I_x = 1\}$ and $\{n_2(x) = 0\} = \{I_x = 0\}$. $I_x = 0$ is thus notation for the situation where no second stage trees are selected at point x.

One has $\mathbb{P}(I_x = 1) = 1 - \Pi_{i=1}^N (1-p_i)^{I_i(x)} =: 1 - p_0(x)$. We shall need the approximation $1 - p_i \approx \exp(-p_i)$ which can be expected to be good if the majority of the p_i is not too large. This leads to $1 - p_0(x) \approx 1 - \exp(-\sum_{i=1}^N I_i(x)p_i)$ and $\mathbb{E}_2\mathbb{P}(I_x = 1) \approx 1 - e^{-m_2}$, where $m_2 = \sum_{i=1}^N \pi_i p_i$ is the expected number of second stage trees per plot. We note that we have the conditional inclusion probabilities $\mathbb{P}(J_i(x) = 1 \mid I_x = 1) = \frac{p_i}{1-p_0(x)} = p_i'$ and $\mathbb{P}(J_k(x)J_l(x) = 1 \mid I_x = 1) = p_{kl}' = \frac{p_k p_l}{1-p_0(x)}$. This leads for $k = 2, 3$ to

$$\mathbb{E}_{3|2,I_x=1}R_k(x) \approx \frac{1}{\lambda(F)}\sum_{i=1}^N \frac{I_i(x)R_i}{\pi_i} = R(x)$$

because the $1 - p_0(x)$ in numerators and denominators cancel. Furthermore,

$$\mathbb{E}_{3|2}R_k(x) = 0 \cdot \mathbb{P}(I_x = 0) + \mathbb{E}_{3|2,I_x=1}R_k(x)\mathbb{P}(I_x = 1) \approx R(x)(1 - p_0(x))$$

Using $\mathbb{E}_2\Big((1-p_0(x))R_k(x)\Big) \approx \mathbb{E}_2(1-p_0(x))\mathbb{E}_2 R_k(x)$, we obtain

$$\mathbb{E}_2(R_k(x)) \approx (1-e^{-m_2})\frac{1}{\lambda(F)}\sum_{i=1}^{N} R_i = (1-e^{-m_2})\bar{R} \tag{2.7}$$

$$\mathbb{E}_2(Y_k(x)) \approx \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i = \bar{Y} - e^{-m_2}\bar{R} \tag{2.8}$$

Therefore the bias can be expected to be negligible in practice.

For completeness let us consider the second order approximation for the bias. With $f(u,v,) = \frac{u}{v}$ and the second-order Taylor expansion for $f(u,v)$ at the point $f(\mathbb{E}U, \mathbb{E}V)$ we obtain for $\mathbb{E}() = \mathbb{E}_{3|2,I_x=1}()$

$$\mathbb{E}(\frac{U}{V}) \approx \frac{\mathbb{E}U}{\mathbb{E}V}\Big(1 - \frac{\mathbb{E}UV}{\mathbb{E}U\mathbb{E}V} + \frac{\mathbb{E}V^2}{\mathbb{E}^2 V}\Big)$$

Using exactly the same techniques as those leading to (3.11) below we get for $E_2\mathbb{P}(I_x=1)\mathbb{E}_{3|2,I_x=1}R_k(x)$ the result

$$\begin{aligned}\mathbb{E}_2\mathbb{E}_{3|2}R_k(x) &\approx (1-e^{-m_2})\bar{R}\Big(1-(1-e^{-m_2})(\frac{1}{m_1}-\frac{1}{m_1})\Big)\\ &= (1-e^{-m_2})\bar{R}\end{aligned}$$

Hence, the 1st order approximation for the bias is equivalent to the 2nd order approximation. Using $\hat{\bar{R}} = \frac{1}{n_2}\sum_{x\in s_2}\sum_{i=1}^{N}\frac{I_i(x)J_i(x)R_i}{\pi_i p_i}$ and $\hat{m}_2 = \frac{1}{n_2}\sum_{x\in s_2}\sum_{i=1}^{N} I_i(x)J_i(x)$ the estimate of $e^{-m_2}\bar{R}$ for the Swiss National Inventory is ≈ -0.13, which corresponds to an absolute relative bias of less than 0.04%. More accurate analytical results seem to be out of reach. One can of course question the validity of these asymptotic considerations as the number of second-stage trees per plot is small (2 on average for the Swiss NFI). This is the motivation behind the pooled estimate to be introduced in chapter 4.

Chapter 3

Variances

3.1 Asymptotic variances

The variance of $Y_1(x)$ variance can be obtained readily from (1.2) and is given by

$$\mathbb{V}_{2,3}Y_1(x) = \mathbb{V}_2 Y(x) + \frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{(1-p_i)R_i^2}{\pi_i p_i} \tag{3.1}$$

This is simply the variance of the true local density augmented by the second-stage variance.

The calculation of the variances of $Y_2(x)$ and $Y_3(x)$ is a bit more cumbersome. First note that

$$\begin{aligned} \mathbb{V}_{2,3}Y_k(x) &= \mathbb{E}_2\mathbb{V}_{3|2}Y_k(x) + \mathbb{V}_2\mathbb{E}_{3|2}Y_k(x) \\ &= \mathbb{E}_2\mathbb{V}_{3|2}R_k(x) + \mathbb{V}_2\mathbb{E}_{3|2}Y_k(x) \;\; k=2,3 \end{aligned} \tag{3.2}$$

By conditioning on $I_x = 1$ we have

$$\begin{aligned} \mathbb{E}_{3|2}Y_k(x) &= Y_0(x) + (1-p_0(x))\mathbb{E}_{3|2,I_x=1}R_k(x) \\ &\approx Y_0(x) + (1-p_0(x))R(x) = Y(x) - p_0(x)R(x) \end{aligned}$$

With (2.7) we get

$$\mathbb{V}_2\mathbb{E}_{3|2}Y_k(x) = \mathbb{E}_2(Y(x) - p_0(x)R(x))^2 - (\bar{Y} - e^{-m_2}\bar{R})^2$$

For the second term in (3.2) we get after some algebra

$$\mathbb{V}_2\mathbb{E}_{3|2}Y_k(x) \approx \mathbb{V}_2 Y(x) + (e^{-m_2})^2\mathbb{V}_2 R(x) \;\; k=2,3 \tag{3.3}$$

and for the first term in (3.2) we obtain as two terms $\mathbb{P}(I_x = 1)(\mathbb{E}_{3|2,I_x=1}R_k(x))$ cancel

$$\begin{aligned} \mathbb{V}_{3|2}R_k(x) &= \mathbb{E}_{I_x}\mathbb{V}_{3|2,I_x}R_k(x) + \mathbb{V}_{I_x}\mathbb{E}_{3|,2,I_x}R_k(x) \\ &= \mathbb{P}(I_x=1)\mathbb{V}_{3|2,I_x=1}R_k(x) + \mathbb{E}_{I_x}((\mathbb{E}_{3|2,I_x}R_k(x))^2) - \left(\mathbb{E}_{I_x}\mathbb{E}_{3|2,I_x}R_k(x)\right)^2 \\ &\approx \mathbb{P}(I_x=1)\mathbb{E}_{3|2,I_x=1}R_k(x)^2 - (\mathbb{P}(I_x=1))^2R^2(x) \end{aligned} \tag{3.4}$$

which leads to

$$\mathbb{E}_2\mathbb{V}_{3|2}R_k(x) \approx (1-e^{-m_2})\mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_k^2(x) - (1-e^{-m_2})^2\mathbb{E}_2R^2(x) \tag{3.5}$$

Therefore, we obtain by using (3.2,3.3,3.4)

$$\begin{aligned} \mathbb{V}_{2,3}Y_k(x) &= \mathbb{V}_2 Y(x) + (e^{-m_2})^2\mathbb{V}_2 R(x) \\ &+ (1-e^{-m_2})\mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_k^2(x) - \left(1-e^{-m_2}\right)^2\mathbb{E}_2R^2(x) \end{aligned} \tag{3.6}$$

The last step is therefore to calculate $\mathbb{E}_{3|2,I_x=1}R_k^2(x)$ for $k = 2, 3$.

Using 1st Order Approximation to Estimate $\mathbb{E}_{3|2,I_x=1}R_2^2(x)$

We now limit ourselves to the case $k = 2$. Using the denominator of the stabilization term found in (2.4), we get

$$\begin{aligned}\mathbb{E}_{3|2,I_x=1}\big(\sum_{i=1}^N I_i(x)J_i(x)\big)^2 &= \frac{1}{1-p_0(x)}\sum_{i=1}^N I_i(x)p_i + \frac{1}{1-p_0(x)}\sum_{i\neq j} I_i(x)I_j(x)p_ip_j \\ &= \frac{1}{1-p_0(x)}\big(\sum_{i=1}^N I_i(x)p_i(1-p_i) + (\sum_{i=1}^N I_i(x)p_i\big)^2\big)\end{aligned}$$

If we set

$$\alpha(x) = \frac{\big(\sum_{i=1}^N I_i(x)p_i\big)^2}{\mathbb{E}_{3|2,I_x=1}\big(\sum_{i=1}^N I_i(x)J_i(x)\big)^2}$$

then

$$\begin{aligned}\mathbb{E}_{3|2,I_x=1}\frac{(\sum_{i=1}^N I_i(x)p_i)^2}{(\sum_{i=1}^N I_i(x)J_i(x))^2} &\approx \frac{(\sum_{i=1}^N I_i(x)p_i)^2}{\mathbb{E}_{3|2,I_x=1}((\sum_{i=1}^N I_i(x)J_i(x))^2} \\ &= (1-p_0(x))\alpha(x) \leq (1-p_0(x)\end{aligned}$$

We have

$$\mathbb{E}_{3|2,I_x=1}R_2^2(x) = \mathbb{E}_{3|2,I_x=1}\alpha(x)\beta(x) \leq \beta(x)$$

because the $1-p_0(x)$ cancel out. We have

$$\begin{aligned}\mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_2^2(x) &= \mathbb{E}_2\alpha(x)\beta(x) \leq \mathbb{E}_2\beta(x) \\ &= \frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{R_i^2}{\pi_i p_i} + \frac{1}{\lambda^2(F)}\sum_{i\neq j}\frac{\pi_{ij}}{\pi_i\pi_j}R_iR_j \qquad (3.7)\end{aligned}$$

Recall that by the properties of the Horwitz-Thomson estimator one has

$$\mathbb{V}_2R(x) = \frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{(1-\pi_i)R_i^2}{\pi_i} + \frac{1}{\lambda^2(F)}\sum_{i\neq j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}R_iR_j$$

and consequently

$$\begin{aligned}\frac{1}{\lambda^2(F)}\sum_{i\neq j}\frac{\pi_{ij}R_iR_j}{\pi_i\pi_j} &= \mathbb{V}_2R(x) - \frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{R_i^2}{\pi_i} + \bar{R}^2 \\ &= \mathbb{E}_2R^2(x) - \frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{R_i^2}{\pi_i}\end{aligned}$$

Using (3.7) we obtain

$$\mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_2^2(x) \leq \mathbb{E}_2R^2(x) + \frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{R_i^2(1-p_i)}{\pi_ip_i} \qquad (3.8)$$

With (3.6,3.8) we obtain after some elementary algebra the result

$$\begin{aligned}\mathbb{V}_{2,3}Y_2(x) &\leq \mathbb{V}_2Y(x) + (1-e^{-m_2})\frac{1}{\lambda^2(F)}\sum_{i=1}^N\frac{R_i^2(1-p_i)}{\pi_ip_i} \\ &+ e^{-m_2}\mathbb{E}_2R^2(x) - \bar{R}^2(e^{-m_2})^2 \qquad (3.9)\end{aligned}$$

The last two terms in (3.9) can be expected to be much smaller than the leading term of the second stage variance, i.e. $(1-e^{-m_2})\sum_{i=1}^N \frac{R_i^2(1-p_i)}{\pi_i p_i}$. In any case, for large m_2, $Y_1(x)$ and $Y_2(x)$ are, as intuitively expected, asymptotically equivalent.

Using 2nd Order Approximation to Estimate $\mathbb{E}_{3|2,I_x=1}R_2^2(x)$

The above result is primarily based on the approximation $\mathbb{E}\frac{U^2}{V^2} \approx \frac{\mathbb{E}U^2}{\mathbb{E}V^2}$ where $U = \frac{1}{\lambda(F)}\sum_{I=1}^N \frac{I_i(x)J_i(x)R_i}{\pi_i p_i}$, $V = \sum_{i=1}^N I_i(x)J_i(x)$ and **where $\mathbb{E}$ stands for $\mathbb{E}_{3|2,I_x=1}$.**

We now look at the 2nd order Taylor expression for $f(u,v) = \frac{u^2}{v^2}$. We will need the following partial derivatives

$$f_u = \frac{2u}{v^2}, f_{uu} = \frac{2}{v^2}, f_{uv} = \frac{-4u}{v^3}, f_v = \frac{-2u^2}{v^3}, f_{vv} = \frac{6u^2}{v^4}$$

The second-order Taylor expansion for $f(u,v)$ at the point $f(\mathbb{E}U, \mathbb{E}V)$ for the random variables U and V is

$$\begin{aligned} f(U,V) &= f(\mathbb{E}U,\mathbb{E}V) + (U-\mathbb{E}U)f_u(\mathbb{E}U,\mathbb{E}V) + (V-\mathbb{E}V)f_v(\mathbb{E}U,\mathbb{E}V) \\ &+ \frac{1}{2}\Big((U-\mathbb{E}U)^2 f_{uu}(\mathbb{E}U,\mathbb{E}V) + 2(U-\mathbb{E}U)(V-\mathbb{E}V)f_{uv}(\mathbb{E}U,\mathbb{E}V) + (V-\mathbb{E}V)^2 f_{vv}(\mathbb{E}U,\mathbb{E}V)\Big) \end{aligned}$$

After some simple algebra, this leads to the 2nd order approximation

$$\mathbb{E}\Big(\frac{U^2}{V^2}\Big) = \Big(\frac{\mathbb{E}U}{\mathbb{E}V}\Big)^2\Big(\frac{\mathbb{E}U^2}{\mathbb{E}^2U} - 4\frac{\mathbb{E}UV}{(\mathbb{E}U)\mathbb{E}(V)} + 3\frac{\mathbb{E}V^2}{\mathbb{E}^2V} + 1\Big) \tag{3.10}$$

One has

$$\begin{aligned} \mathbb{E}U &= \frac{1}{1-p_0(x)}\frac{1}{\lambda(F)}\sum_{i=1}^N \frac{I_i(x)}{\pi_i}R_i = \frac{R(x)}{1-p_0(x)} \\ \mathbb{E}V &= \frac{1}{1-p_0(x)}\sum_{i=1}^N I_i(x)p_i \\ \Big(\frac{\mathbb{E}U}{\mathbb{E}V}\Big)^2 &= \frac{R^2(x)}{(\sum_{i=1}^N I_i(x)p_i)^2} \end{aligned}$$

After tedious but simple algebra one obtains the terms

$$\begin{aligned} \frac{\mathbb{E}U^2}{\mathbb{E}^2U} &= (1-p_0(x))\Big(1 + \frac{\frac{1}{\lambda^2(F)}\sum_{i=1}^N \frac{I_i(x)(1-p_i)R_i^2}{\pi_i^2 p_i}}{R^2(x)}\Big) \\ \frac{\mathbb{E}V^2}{\mathbb{E}^2V} &= (1-p_0(x))\Big(1 + \frac{\sum I_i(x)p_i(1-p_i)}{(\sum_{i=1}^N I_i(x)p_i)^2}\Big) \\ \frac{\mathbb{E}UV}{\mathbb{E}U\mathbb{E}V} &= (1-p_0(x))\Big(1 + \frac{1}{\sum_{i=1}^N I_i(x)p_i} - \frac{\frac{1}{\lambda(F)}\sum_{I=1}^N \frac{I_i(x)p_iR_i}{\pi_i}}{R(x)\sum_{i=1}^N I_i(x)p_i}\Big) \end{aligned}$$

To calculate the expectation $\mathbb{E}_2$ we shall use approximations of the form $\mathbb{E}_2 A(x)B(x) \approx \mathbb{E}_2A(x)\mathbb{E}_2B(x)$ and $\mathbb{E}_2\frac{A(x)}{B(x)} \approx \frac{\mathbb{E}_2A(x)}{\mathbb{E}_2B(x)}$. Furthermore note that $\sum_{i=1}^N \pi_i p_i^2 = \sum_{i=1}^N \pi_i p_i p_i \approx m_2\frac{\sum_{i=1}^N p_i}{N} \approx m_2\frac{m_2}{m_1} = \frac{m_2^2}{m_1}$ and that $\frac{1}{\lambda(F)}\sum_{i=1}^N \frac{I_i(x)R_ip_i}{\pi_i} \approx \frac{m_2}{m_1}R(x)$. Collecting the pieces together we end up with

$$\begin{aligned} \mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_2^2(x) &= (1-e^{-m_2})\frac{1}{\lambda^2(F)}\sum_{i=1}^N \frac{(1-p_i)R_i^2}{\pi_i p_i} \\ &+ (1-e^{-m_2})(1 + \frac{1}{m_1} - \frac{1}{m_2})\mathbb{E}_2R^2(x) \end{aligned} \tag{3.11}$$

Using (3.6) we finally obtain the result

$$\begin{aligned}\mathbb{V}_{2,3}(Y_2(x)) &= \mathbb{V}_2(Y(x)) + \big(1-e^{-m_2}\big)^2 \frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{R_i^2(1-p_i)}{\pi_i p_i} \\ &- \big(1-e^{-m_2}\big)^2 \big(\frac{1}{m_2} - \frac{1}{m_1}\big)\mathbb{E}_2 R^2(x) \\ &+ (e^{-m_2})^2 \mathbb{E}_2 R^2(x) - (e^{-m_2})^2 \bar{R}^2 \end{aligned} \tag{3.12}$$

Using 2nd Order Approximation to Estimate $\mathbb{E}_{3|2,I_x=1}R_3^2(x)$

The techniques for deriving the asymptotic variance formula for $Y_3(x)$ is exactly the same. Picking up at (3.6), we can easily estimate all the terms except $\mathbb{E}_{3|2,I_x=1}R_3^2(x)$. For this we will focus only on applying a second order approximation. Again referring to (3.10), we need the following pieces

$$U = \frac{1}{\lambda(F)} \sum_{I=1}^{N} \frac{I_i(x)J_i(x)R_i}{\pi_i p_i}, V = \sum_{i=1}^{N} \frac{I_i(x)J_i(x)}{p_i}$$

$$\begin{aligned}\mathbb{E}U &= \frac{R(x)}{1-p_0(x)} \\ \mathbb{E}V &= \frac{\sum_{i=1}^{N} I_i(x)}{1-p_0(x)} \\ \Big(\frac{\mathbb{E}U}{\mathbb{E}V}\Big)^2 &= \frac{R^2(x)}{(\sum_{i=1}^{N} I_i(x))^2} \\ \frac{\mathbb{E}U^2}{\mathbb{E}^2U} &= (1-p_0(x))\Big(1 + \frac{\frac{1}{\lambda^2(F)}\sum_{i=1}^{N} \frac{I_i(x)(1-p_i)R_i^2}{\pi_i^2 p_i}}{R^2(x)}\Big) \\ \frac{\mathbb{E}V^2}{\mathbb{E}^2V} &= (1-p_0(x))\big(1 + \frac{\sum \frac{I_i(x)(1-p_i)}{p_i}}{(\sum_{i=1}^{N} I_i(x))^2}\big) \\ \frac{\mathbb{E}UV}{\mathbb{E}U\mathbb{E}V} &= (1-p_0(x))\Big(1 - \frac{1}{\sum_{i=1}^{N} I_i(x)} + \frac{\frac{1}{\lambda(F)}\sum_{i=1}^{N} \frac{I_i(x)R_i}{\pi_i p_i}}{R(x)\sum_{i=1}^{N} I_i(x)}\Big)\end{aligned}$$

We can use similar approximations as those that were used to derive (3.11). In particular, $\sum_{i=1}^{N} \frac{R_i}{p_i} \approx \frac{m_1}{m_2}\sum_{i=1}^{N} R_i$ and that $\sum_{i=1}^{N} \frac{\pi_i(1-p_i)}{p_i} \approx \frac{m_1}{m_2}\sum_{i=1}^{N} \pi_i(1-p_i)$. We get

$$\begin{aligned}\mathbb{E}_2\mathbb{E}_{3|2,I_x=1}R_3^2(x) &= \mathbb{E}_2(\sum_{i=1}^{N} I_i(x))^2 \mathbb{E}_{3|2,I_x=1}(\frac{U}{V})^2 \approx (1-e^{-m_2})\frac{1}{\lambda^2(F)}\sum_{i=1}^{N}\frac{(1-p_i)R_i^2}{\pi_i p_i} \\ &+ \mathbb{E}_2 R^2(x)(1-e^{-m_2})(1+\frac{1}{m_1}-\frac{1}{m_2})\end{aligned} \tag{3.13}$$

Thus, the 2nd order taylor approximation in (3.13) leads to the same result as presented in (3.11). Consequently the asymptotic variance formula of $Y_3(x)$ is expected to be the same as in (3.12). It is worthwhile to point out that for large m_2 the estimators $Y_1(x)$, $Y_2(x)$ and $Y_3(x)$ are equivalent.

Anticipated gain over $\mathbb{V}Y_1$

Another noteworthy point is that within (3.12) we can estimate all the terms to give us an idea of the anticipated gain in $\mathbb{V}_{2,3}Y_k(x)$, $k = 2, 3$ over $\mathbb{V}_{2,3}Y_1(x)$. Let us denote the second-stage variance at x by

$$\mathbb{V}_{3|2}R_1(x) = V(x) = \frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{I_i(x)(1-p_i)R_i^2}{\pi_i^2 p_i}$$

which can be unbiasedly estimated by

$$\hat{V}(x) = \frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{I_i(x)J_i(x)(1-p_i)R_i^2}{\pi_i^2 p_i^2}$$

According to (3.1) we have also

$$\mathbb{V}_{2,3}(\hat{Y}_1) = \frac{1}{n_2}\mathbb{V}_2 Y(x) + \frac{1}{n_2}\mathbb{E}_2 V(x) \tag{3.14}$$

which as shown in Mandallaz (2008) can be unbiasedly estimated by

$$\hat{\mathbb{V}}(\hat{Y}_1) = \frac{1}{n_2(n_2-1)} \sum_{x \in s_2} (Y_1(x) - \hat{Y}_1)^2$$

This yields at once the following unbiased estimate for the variance of the unobservable true density $Y_(x)$

$$\hat{\mathbb{V}}_2(Y(x)) = n_2 \hat{\mathbb{V}}(\hat{Y}_1) - \frac{1}{n_2} \sum_{x \in s_2} \hat{V}(x) \tag{3.15}$$

$\bar{R}$ can be estimated by

$$\hat{\bar{R}} = \frac{1}{n_2} \sum_{x \in s_2} \hat{R}(x)$$

where $\hat{R}(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x)J_i(x)R_i}{\pi_i p_i}$
We also have

$$\hat{m}_2 = \frac{1}{n_2} \sum_{x \in s_2} \sum_{i=1}^{N} I_i(x)p_i \text{ or } \frac{1}{n_2} \sum_{x \in s_2} \sum_{i=1}^{N} I_i(x)J_i(x)$$

These alternative methods of estimating $\hat{m}_2$ correspond to the average expected second stage sample size per plot and the average observed second stage sample size per plot respectively. Analytically there does not seem to be any advantage to using one version over the other so we have chose the latter for simplicity. However, it should be noted that empirically this was not the case. We forego discussion about the effect of this until Chapter 5.
Let us set

$$\hat{\mathbb{E}}_2 R^2(x) = \frac{1}{n_2} \sum_{x \in s_2} \hat{R}^2(x)$$

We then have

$$\begin{aligned}
\mathbb{E}_{2,3}\hat{R}^2(x) &= \mathbb{V}_{2,3}\hat{R}(x) + \left(\mathbb{E}_{2,3}\hat{R}(x)\right)^2 \\
&= \mathbb{E}_2\mathbb{V}_{3|2}\hat{R}(x) + \mathbb{V}_2\mathbb{E}_{3|2}\hat{R}(x) + \bar{R}^2 \\
&= \mathbb{E}_2\mathbb{V}_{3|2}\hat{R}(x) + \mathbb{V}_2 R(x) + \bar{R}^2 \\
&= \mathbb{E}_2\Big(\frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{I_i(x)p_i(1-p_i)R_i^2}{\pi_i^2 p_i^2}\Big) + \mathbb{E}_2 R^2(x)
\end{aligned}$$

Consequently the intuitive estimate $\hat{\mathbb{E}}_2 R^2(x)$ of $\mathbb{E}_2 R^2(x)$ has a bias equal to

$$\frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{(1-p_i)R_i^2}{\pi_i p_i}$$

This bias can be estimated by

$$\frac{1}{n_2} \sum_{x \in s_2} \frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{I_i(x) J_i(x)(1-p_i)R_i^2}{\pi_i^2 p_i^2}$$

and then removed. Hence, we can estimate all the components of the asymptotic variance formula given in (3.12).

From this approximation we can get a rough idea of what relative decrease in variance to expect from using $Y_2(x)$ over $Y_1(x)$. For the 3rd Swiss National Inventory this relative improvement is expected to be $\approx 1.75\%$. Note that this estimate is not a substitute for a suitable variance estimator and it is difficult to see if the second order approximation (3.12) is better than the first order one (3.9).

For $\hat{Y}_3$, the anticipated gain is expected to be the same $\hat{Y}_2$ at least asymptotically. However, we will see that this is not supported by empirical evidence where $\hat{Y}_2$ produced consistently lower variances. Let us also emphasize that the most important term is

$$\frac{1}{\lambda^2(F)} \sum_{i=1}^{N} \frac{R_i^2(1-p_i)}{\pi_i p_i} = \mathbb{V}_{2,3} R_1(x)$$

which is obtained without further approximations in the second order expansion (3.10).

For those interested in assessing the validity of the asymptotic variance formula, refer to tables A.7 and A.8. Table A.8 displays relative differences of the variance estimates of $\hat{Y}_k$, $k = 2, 3$ to the variance estimate $\hat{Y}_1$ as well as their anticipated gains based on (3.12) while Table A.7 presents a simple comparison of the asymptotic variance formulae to the empirical ones (to be defined in the next section). On the bright side, we see that the asymptotic variance formulae roughly match the empirical results implying that the 2nd order approximation was somewhat adequate. However, the estimates for the anticipated gain predicted improvement in the variance. This was frequently not observed with the empirical variance estimates, especially as the first stage sample size was reduced. There also seems to be evidence that the asymptotic variance formula is somewhat more appropriate for $\hat{Y}_2$ than for $\hat{Y}_3$ and for Jura than for other, more mountainous, regions.

3.2 Empirical variances

The estimated variances of the $\hat{Y}_k$ can be obtained by using the standard formulae

$$\hat{\mathbb{V}}(\hat{Y}_k) = \frac{1}{n_2(n_2-1)} \sum_{x \in s_2} (Y_k(x) - \hat{Y}_k)^2 \ , \ k = 0, 1, 2, 3 \tag{3.16}$$

This result is obvious for $\hat{Y}_0$ and somewhat surprising for $\hat{Y}_k$, $k = 1, 2, 3$ (Mandallaz (2008)).

Chapter 4

The Pooled Estimator

The small number of second-stage trees at point x suggests to pool observations across the n_2 points $x \in s_2$. First, let us note that we can rewrite the point estimate $\hat{Y}_1$ as

$$\begin{aligned} \hat{Y}_1 &= \frac{1}{n_2}\sum_{x\in s_2} Y_0(x) + \frac{1}{n_2}\sum_{x\in s_2} R_1(x) = \frac{1}{n_2}\sum_{x\in s_2} Y_1(x) \\ &= \frac{1}{n_2\lambda(F)}\sum_{i=1}^{N}\frac{T_i(x)\tilde{Y}_i}{\pi_i} + \frac{1}{n_2\lambda(F)}\sum_{i=1}^{N}\frac{S_i R_i}{\pi_i p_i} \end{aligned} \tag{4.1}$$

where we have set

$$T_i = \sum_{x\in s_2} I_i(x)\ ,\ S_i = \sum_{x\in s_2} I_i(x)J_i(x)$$

Note that $\hat{Y}_1$ is the mean of n_2 generalized Horwitz-Thomspon (HT) estimators and not a HT estimator itself because it is not based on inclusion probabilities such as $1-(1-\pi_i)^{n_2}$. It can be viewed as an Hansen-Hurwitz estimator under sampling with replacement (see problem 4.5 in Mandallaz (2008)). Although, in practice, inventories are usually performed with systematic grids of points, we shall proceed as if the points are independently uniformly distributed in F, which is justified for extensive inventories. In this theoretical framework a tree can be sampled many times and we have after simple algebra the following important relations

$$\begin{aligned} \mathbb{E}_2 T_i &= n_2\pi_i \\ \mathbb{E}_{3|2} S_i &= T_i p_i \\ \mathbb{E}_{2,3} S_i &= n_2\pi_i p_i \\ \mathbb{E}_{3|2} S_i^2 &= T_i p_i(1-p_i) + T_i^2 p_i^2 \\ \mathbb{E}_{3|2} S_i S_j &= T_i T_j p_i p_j \\ \mathbb{E}_2 T_i^2 &= n_2\pi_i(1-\pi_i) + n_2^2\pi_i^2 \\ \mathbb{E}_2 T_i T_j &= n_2^2\pi_i\pi_j + n_2(\pi_{ij}-\pi_i\pi_j) \end{aligned} \tag{4.2}$$

It is a tedious, but simple, exercise to check that the above equations lead of course to the same overall variance for $\mathbb{V}_{2,3}(\hat{Y}_1)$ based on (3.1), namely $\frac{1}{n_2}\mathbb{V}_{2,3}(Y_1(x))$. We consider the pooled version of $R_2(x)$, that is

$$R_{2,p} = \Big(\frac{\frac{1}{n_2\lambda(F)}\sum_{i=1}^{N}\frac{S_i R_i}{\pi_i p_i}}{\sum_{i=1}^{N} S_i}\Big)\sum_{i=1}^{N} T_i p_i \tag{4.3}$$

and the corresponding point estimate

$$\hat{Y}_{2,p} = \frac{1}{n_2 \lambda(F)} \sum_{i=1}^{N} \frac{T_i \tilde{Y}_i}{\pi_i} + R_{2,p} \tag{4.4}$$

The justification is that the overall number of second stage trees $\sum_{i=1}^{N} S_i$ is large and that therefore the validity of the asymptotic calculations is better. The probability that the overall sample has no second stage trees is

$$p_0(s_2) = \mathbb{P}(S_i = 0, \forall i \mid s_2) = \prod_{x \in s_2} p_0(x) = \prod_{x \in s_2} \prod_{I=1}^{N} (1 - p_i)^{I_i(x)}$$

which is approximately the same as

$$\prod_{x \in s_2} \prod_{I=1}^{N} e^{-p_i I_i(x)} = e^{-\sum_{i=1}^{N} T_i p_i}$$

Therefore, we have on average for the probability of an empty second stage sample

$$\mathbb{E} p_0(s_2) \approx e^{-n_2 \sum_{i=1}^{N} \pi_i p_i} = e^{-n_2 m_2}$$

This probability is so small that it can be considered to be zero for all practical purposes. A nice consequence of that is that we can bypass the conditional expectation and variance (on the event $s_2 \neq \emptyset$).

The first order approximation yields at once

$$\mathbb{E}_{3|2} R_{2,p} \approx \frac{1}{n_2 \lambda(F)} \sum_{i=1}^{N} \frac{T_i R_i}{\pi_i p_i} \tag{4.5}$$

and hence also $\hat{Y}_{2,p} \approx \hat{Y}_1$ and likewise for the variances. Tedious second order calculations similar to those presented previously show that the bias of $\hat{Y}_{2,p}$ is of order $\bar{R} \cdot O(n_2^{-1})$ instead of $\bar{R} \cdot e^{-m_2}$, so that it is negligible in large samples. The first order approximation for the variance yields, by noting that $\mathbb{E}(\frac{X}{Y})^2 \approx \frac{\mathbb{E}X^2}{\mathbb{E}Y^2} \leq \frac{\mathbb{E}X^2}{\mathbb{E}^2 Y}$,

$$\mathbb{V}_{3|2} R_{2,p} \approx\leq \mathbb{E}_{3|2} \Big(\frac{1}{n_2 \lambda(F)} \sum_{i=1}^{N} \frac{S_i R_i}{\pi_i p_i} \Big)^2 - \Big(\frac{1}{n_2 \lambda(F)} \sum_{i=1}^{N} \frac{T_i R_i}{\pi_i} \Big)^2$$

Taking the expectation with respect to $\mathbb{E}_2$ and using (4.2) we see after tedious algebra that

$$\mathbb{V}_{2,3} \hat{Y}_{2,p} \approx\leq \frac{1}{n_2} \mathbb{V}_2 Y(x) + \frac{1}{n_2 \lambda^2(F)} \sum_{i=1}^{N} \frac{(1 - p_i) R_i^2}{\pi_i p_i} = \mathbb{V}_{2,3} \hat{Y}_1 \tag{4.6}$$

Again, the second order Taylor approximation confirms that the above equality is correct up to terms of order $O(n_2^{-1})$. This is, of course, a somewhat disappointing result due primarily to the fact that the correction factor $\frac{\sum_{i=1}^{N} T_i p_i}{\sum_{i=1}^{N} S_i}$ is expected to be very close to 1. Nonetheless it is a worthwhile exercise in that even negative results are worth knowing.

To estimate the variance of $\hat{Y}_{2,p}$ we have used the jackknife.

Chapter 5

Results

Table A.1 displays the results from the 3rd Swiss National Forest Inventory (2003). Notice that all point estimates are well within 2 standard deviations of each other. Since the sample size is large and $\hat{Y}_1$ is design unbiased, we confirm that the biases of $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_{2,p}$ are in fact negligible. Therefore, it should be sufficient to compare these estimators based on their variances alone. Figure B.1 allows us to visualize standard errors presented in Table A.1 together.

There appeared to be no dramatic improvement of the any of the estimators when compared to $\hat{Y}_1$. $\hat{Y}_2$ consistently performed the same or slightly better than $\hat{Y}_1$ across most regions and for the total. $\hat{Y}_3$, on the other hand, consistently produced slightly higher standard errors when compared to $\hat{Y}_1$ and $\hat{Y}_2$. The pooled estimator, $\hat{Y}_{2,p}$, was virtually identical to $\hat{Y}_1$ in all regions except for the Alps and the South Alps. It was the only estimator that did not consistently rank the same across all regions. It should be noted that the delete-one jackknife estimator can be expected to have a positive bias when the 2nd stage sample size is small. Fuller (2009) discusses the possible adjustments to correct this; however, in the present circumstance it seems unnecessary since the motivation behind the pooled estimator was to compensate for the bias of $\hat{Y}_2$ and $\hat{Y}_3$, which is most certainly negligible. In any case, this is a possible explanation for the apparently worse performance of $\hat{Y}_{2,p}$ in regions with smaller m_2.

It is worth pointing out that $\hat{Y}_0$ is included in the Table A.1 as a point of reference to gauge the size of the $\bar{R}$ compared to the point estimates. While its variance is lower than the other estimates, its MSE is higher than that of $\hat{Y}_1$, especially in local estimation because it can be biased. If this was not the case the second stage in the Swiss National Forest Inventory would be unnecessary and wasteful.

Adjusting the Sample Size

In addition to comparing the performance of these estimators for the Swiss National Forest Inventory, we are also interested in assessing the influence of sample size. To do this we need to create subsamples from the original full sample that mimic the same geographical structure. Figure B.2 is an overview of all first stage plots. Notice that the outline of Switzerland is visible and upon close inspection so is the outline of out-of-forest areas such as lakes and the highest parts of the Alps. Our goal was to thin the plot density in such a way that we can still make out the largest features visually. For simplicity, the target sub-grids correspond to a 1/2 sample, a 1/4 sample, a 1/8 sample and a 1/16 sample. With every successive reduction of half the remaining sample, a set of plots was generated to check the geographic structure of the overview and of an arbitrary zoomed view. Figure B.3 illustrates

this by example for the 1/2 sample.

The final thinning procedure that was selected was as follows:

1. Divide all X and Y coordinates by 1000.
2. Remove every odd X coordinate. This is incidentally the same as removing every even Y. The result is approximately a half sample size selection of the full sample. From now on this is will be referred to as the Half Sample.
3. Using the Half Sample from the previous step, sort the Y coordinates in ascending order and remove every 2nd item. This is the Quarter Sample.
4. Using the Quarter Sample, sort the X coordinates in ascending order and remove every 2nd item. This is the Eighth Sample.
5. Using the Eighth Sample, sort the Y coordinates in ascending order and remove every 2nd item. This is the Sixteenth Sample.

Figure B.4 shows the same zoomed snapshot with selected sample for all subsamples. Figure B.5 shows the geographic overview for all subsamples selected. The first stage selection from any of these samples is treated as if it were selected under simple random sampling. It is widely considered that this is acceptable as long as the forest does not show any periodicity that interacts with the systemic sampling structure to produce sampling bias. Given the plot overviews of the subsamples selected, this seems unlikely to be the case.

Tables A.2 - A.5 give the results of the estimators in question while Figure B.6 gives a more visual comparison. It is more clear now that none of the proposed estimators provide any reliable improvement to $\hat{Y}_1$ for this data. $\hat{Y}_3$ consistently produces larger standard errors than $\hat{Y}_1$ and $\hat{Y}_2$ and the pooled estimator $\hat{Y}_{2,p}$ seems to have problems specifically in the South Alps region. In any case, it is clear that the stabilizing factors that were applied when using $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_{2,p}$ have very limited practical advantage over the standard two stage estimator $\hat{Y}_1$ in this context.

Interpretation

To gain further insight concerning the usefulness of the second stage, one should look at the trade-off between variance and costs. Set

$$\theta = \frac{\mathbb{V}(Y_1(x))}{\mathbb{V}(Y(x))} = \frac{\mathbb{V}(Y(x)) + \mathbb{E}_2(V(x))}{\mathbb{V}(Y(x))} = 1 + \frac{\mathbb{E}_2(V(x))}{\mathbb{V}(Y(x))}$$

θ is the expected increase in variance using $\hat{Y}_1(x)$ compared to the variance we would have gotten had we selected ALL trees to be in the second stage (represented by $\mathbb{V}(Y(x))$). Equivalently, one can consider the proportion of the second-stage variance with respect to the total variance, i.e.

$$\eta = \frac{\mathbb{E}_2(V(x))}{\mathbb{V}(Y_1(x))} = \frac{\mathbb{V}(Y_1(x)) - \mathbb{V}(Y(x))}{\mathbb{V}(Y_1(x))} = 1 - \frac{1}{\theta}$$

Therefore

$$\theta = \frac{1}{1-\eta}$$

According to (3.15) η can be estimated by

$$\hat{\eta} = \frac{\frac{1}{n_2}\sum_{x \in s_2} \hat{V}(x)}{n_2 \hat{\mathbb{V}}(\hat{Y}_1)} * 100$$

Remark that the n_2 is part of the denominator because we are looking at the total variance over all plots which decreases at a rate of $\frac{1}{n_2}$. Table A.6 contains the estimates of η. With $\eta \approx 0.03$ we have $\theta \approx 1.03$. The ratio of standard errors is $\sqrt{\theta} \approx 1.015$, i.e. an increase of $\approx 1.5\%$.

Let's have a look at the costs within plots (travel and installation costs are independent of the π_i and p_i). We have for the first-stage costs $c_{21} \approx 2$ min per tree and for the second-stage $c_{22} \approx 5$ min per tree (total time for a crew of 2 persons, D. Mandallaz (2008)). With $m_1 = 11.5$ and $m_2 = 2.2$ we have the cost ratio for the SNI

$$\frac{m_1 c_{21} + m_2 c_{22}}{m_1(c_{21} + c_{22})} \approx 43\%$$

That is: we can reduce the costs within plot by 67% by taking an increase in standard error of 1.5% into account. However, the total costs for estimating the timber volume is roughly only 10% of the total costs in the second SNI. Note that travel and installation costs contribute to roughly 80% of the total costs and that the field crews record numerous other data than volume. However, using $Y_1(x)$ instead of $Y(x)$, results in saving ≈ 6000 hours or roughly $600'000$ SFr. Thus, the second-stage variance is indeed small but not without practical relevance. Using Y_0 saves further 2000 hours, but at the costs of local biases.

It is now intuitively clear from Table A.6 that the proposed estimators, which attempt to improve $\hat{Y}_1$ by reducing variance arising from the randomness of the second stage sample size, will at best have only a small impact on the total variance.

It may also be helpful to examine the stabilization factors for each plot to see if they give any insights. $\hat{Y}_0$ is equivalent to a one phase two stage estimator whose residual term has a stabilization factor always equal to 0. $\hat{Y}_1$ has the constant 1. $\hat{Y}_2$ has stabilization factors equal to $\frac{\sum_{i=1}^{N} I_i(x) p_i}{\sum_{I=1}^{N} I_i(x) J_i(x)}$ and $\hat{Y}_3$ equal to $\frac{\sum_{i=1}^{N} I_i(x)}{\sum_{i=1}^{N} \frac{I_i(x) J_i(x)}{p_i}}$ for each plot x. The pooled estimator's factor, $\frac{\sum_{i=1}^{N} T_i p_i}{\sum_{i=1}^{N} S_i}$, will also be constant across all plots but it is random since it is generated from the data. These factors are presented graphically in Figure B.7.

It is immediately clear that the magnitude of the stabilization factors of $\hat{Y}_3$ is much larger than those of $\hat{Y}_2$ as seen by the different scales of the Y-axes of the plots. This may suggest that $\hat{Y}_3$ overcompensates when it attempts to correct for variability of the second stage sample size, which leads to slightly higher variances overall when compared to $\hat{Y}_2$. $\hat{Y}_3$ also appears to be somewhat more prone to extreme stabilization factors in the Alps region which was a region that it did not test well under reduced sample size.

We also see that $\hat{Y}_{2,p}$ seems to have an larger than average stabilization factor in the South Alps. This implies that more trees on average were expected to be selected than were actually measured for the second stage. Recall that two alternative methods for estimating m_2 were presented in section 3.1: $\frac{1}{n_2} \sum_{x \in s_2} \sum_{i=1}^{N} I_i(x) p_i$ and $\frac{1}{n_2} \sum_{x \in s_2} \sum_{i=1}^{N} I_i(x) J_i(x)$. The stabilizing factor for $\hat{Y}_{2,p}$ is in fact equal to a ratio of these two methods. Analytically we expect this ratio to be around 1 on average but the empirical results presented in Figure B.7 seem to be paint a different picture. Figure B.8 shows the stabilizing factors for just $\hat{Y}_{2,p}$ by region under various sample size alongside a graph depicting the percentage of percent of plots where no second stage trees were selected. This indicates that in the South Alps and Alps there is a persistent tendency to expect more second stage trees than were actually observed. This can be attributed to a practical problem in the field where a tree can be selected for the second stage but not actually measured for it. This can happen, for instance, if a tree is not 7 meters tall or unmeasurable for some other reason beyond the control of the field technician. In any case, this difference

between the theory and practical application of the sampling scheme seems to cause a slight inflation of the average stabilizing factors in regions with smaller trees. This may also help to explain the breakdown of the $\hat{Y}_{2,p}$ in the South Alps.

Chapter 6

Conclusions

Several theoretical and empirical investigations in the context of one-stage Poisson sampling suggest that adjusting estimators with a stabilization factor will lead to a reduction in variance, particularly if PPS sampling techniques are used, in which case a large part of the variance is due to the random sample size. It was hoped that similar adjustments could also be useful in the slightly different context of forest inventory, where Poisson sampling is used for second-stage sampling to estimate timber volume with three-way yield tables (based on height and two diameter measurements), as is the case in the Swiss National Forest Inventory. All three of the proposed estimators, $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_{2,p}$, had negligible biases. However, they failed to consistently improve the variance as compared to the unadjusted two-stage Hansen-Hurwitz estimator $\hat{Y}_1$, although the asymptotic variances suggested that this should be the case. In large samples $\hat{Y}_2$ was slightly better than the original unadjusted estimator $\hat{Y}_1$. The stabilization factors are very sensitive to the variability of the number of second-stage trees (on average around 2 per plot, which is much smaller than sample sizes generally considered in survey sampling). This was particularly apparent for $\hat{Y}_3$. In any case, the proportion of the total variance accounted for by the second-stage variance is relatively small so that only a modest improvement could be expected. Considering that $\hat{Y}_1$ is simpler to implement, unbiased and tends to lead to the lowest variance when the first stage sample size is reduced, it remains an excellent choice, followed by $\hat{Y}_2$. It also has an advantage over the one-stage estimator $\hat{Y}_0$ (based on DBH alone) which can be severely biased for small area estimation, while keeping costs much lower than would be incurred by using the three-way yield table for all trees selected in the first-stage. It is intuitively clear that the same findings will also hold under cluster sampling and two-phase sampling.

Bibliography

[1] Brandli, U.B., Brassel, P., Duc, P., Keller, M., Kohl, M., Herold, A., Kaufmann, E., Lischke, H., Paschedag, I., Schnellbacher, H. J., Schwyzer, A., Stierlin, H. R., Strobel, T., Traub, B., Ulmer, U. and Zinggeler, J. (2001) *Swiss National Forest Inventory: Methods and Models of the Second Assessment.* WSL Swiss Federal Research Institute, Birmensdorf.

[2] Fuller, W. A. (2009) *Sampling Statistics.* Wiley Series in Survey Methodology, New Jersey.

[3] Furnival, G. M., Gregoire T.G. and Grosenbaugh L.R. (1987) Adjusted Inclusion Probabilities with 3P Sampling. *Forest Science.*, **33** 617-631.

[4] Gregoire, T.G. and Valentine, H. (2007) *Sampling Strategies for Natural Resources and the Environment.* Chapman and Hall, New York.

[5] Mandallaz, D. (2008) *Sampling Techniques for Forest Inventories.* Chapman and Hall, New York.

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

[7] Williams, M.S. and Schreuder, H.T. (1998) Outlier-Resistant Estimators for Poisson Sampling: a note. *Canadian Journal of Forest Research*, **28**, 794-797

[8] Williams, M.S., Schreuder, H.T. and Terrazas, G.H. (1998) Poisson Sampling - The Adjusted and Unadjusted Estimator Revisited. *United States Department of Agriculture Forest Service.* Res Note RMRS-RN-4 Sept. 1998.

[9] Wolter, K. (2007) *Introduction to Variance Estimation.* Springer Science + Business Media, LLC, New York.

Appendix A

Tables

Table A.1: **Point Estimates and Errors, in (), for Timber Volume in $\frac{m^3}{ha}$.**

Full Sample

Region	n_2	m_1	m_2	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{2,p}$
JU	949	11.6	2.3	383.00	382.75	380.94	380.36	382.75
				(7.02)	(7.55)	(7.46)	(7.70)	(7.54)
SP	1103	11.1	2.3	411.37	408.76	408.94	406.68	408.68
				(8.12)	(8.39)	(8.38)	(8.55)	(8.40)
PA	1012	12.2	2.6	464.38	461.53	460.75	459.12	461.42
				(9.88)	(10.24)	(10.25)	(10.40)	(10.26)
AL	1676	11.4	2.0	330.20	328.86	326.23	324.10	328.67
				(5.97)	(6.31)	(6.26)	(6.48)	(6.40)
SA	662	11.2	1.3	252.94	248.93	245.67	242.09	246.88
				(7.15)	(7.55)	(7.33)	(7.83)	(8.05)
CH	5402	11.5	2.2	371.72	369.70	368.06	366.09	369.51
				(3.56)	(3.72)	(3.70)	(3.80)	(3.75)

Swiss National Forest Inventory, 2003.

Legend:
n_2: number of plots, m_1: average number of first-stage trees, m_2 average number of second-stage trees, $\hat{Y}_k$ and $\hat{Y}_{2,p}$ point estimates given by (2.5) and(4.4), in parentheses the estimated standard errors according to (3.16), and obtained via the delete-one jackknife for $\hat{Y}_{2,p}$.

Table A.2: **Point Estimates and Standard Errors, in (), for Timber Volume in $\frac{m^3}{ha}$.**

Half Sample Size

Region	n_2	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{2,p}$
JU	448	377.25	376.41	373.92	371.03	376.44
		(9.49)	(10.44)	(10.09)	(10.22)	(10.39)
SP	532	409.70	400.81	402.43	398.75	400.44
		(11.74)	(12.05)	(12.04)	(12.20)	(12.05)
PA	512	444.19	442.58	442.46	441.64	442.51
		(12.55)	(13.25)	(13.25)	(13.60)	(13.30)
AL	806	335.66	330.81	329.15	325.57	330.04
		(8.95)	(9.30)	(9.31)	(9.73)	(9.43)
SA	342	255.29	252.24	246.57	241.95	250.55
		(9.14)	(10.15)	(9.96)	(10.58)	(11.16)
CH	2640	368.27	364.15	362.79	359.71	363.74
		(4.92)	(5.17)	(5.15)	(5.30)	(5.22)

Swiss National Forest Inventory, 2003.

Table A.3: **Point Estimates and Standard Errors, in (), for Timber Volume in $\frac{m^3}{ha}$.**

One Quarter Sample Size

Region	n_2	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{2,p}$
JU	230	369.12	367.38	365.21	361.41	367.43
		(12.85)	(14.31)	(14.01)	(14.08)	(14.25)
SP	269	407.27	400.58	402.91	397.93	400.46
		(16.74)	(17.02)	(17.26)	(17.29)	(17.00)
PA	270	452.13	444.66	443.89	441.31	444.09
		(16.62)	(17.04)	(17.27)	(17.85)	(17.14)
AL	405	337.01	329.41	329.27	326.25	328.10
		(12.82)	(13.33)	(13.50)	(14.24)	(13.55)
SA	167	267.20	267.67	258.10	255.18	267.95
		(14.52)	(15.78)	(15.55)	(16.13)	(17.13)
CH	1341	371.10	365.71	364.42	360.97	365.11
		(6.92)	(7.19)	(7.25)	(7.46)	(7.26)

Swiss National Forest Inventory, 2003.

Table A.4: **Point Estimates and Standard Errors, in (), for Timber Volume in $\frac{m^3}{ha}$.**

One Eight Sample Size

Region	n_2	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{2,p}$
JU	112	378.26	372.80	370.03	363.25	372.92
		(19.89)	(20.75)	(20.61)	(20.62)	(20.70)
SP	130	395.29	389.63	391.24	383.93	389.28
		(25.47)	(26.08)	(26.72)	(26.63)	(26.11)
PA	133	464.56	467.77	467.57	470.56	467.99
		(24.06)	(26.37)	(26.80)	(28.05)	(26.58)
AL	217	341.60	333.65	334.12	330.06	332.28
		(18.72)	(19.64)	(20.09)	(21.46)	(19.97)
SA	83	290.62	284.29	273.97	272.16	280.72
		(20.98)	(22.80)	(22.59)	(22.98)	(24.87)
CH	675	375.98	371.28	369.98	366.50	370.70
		(10.20)	(10.77)	(10.96)	(11.36)	(10.90)

Swiss National Forest Inventory, 2003.

Table A.5: **Point Estimates and Standard Errors, in (), for Timber Volume in $\frac{m^3}{ha}$.**

One Sixteenth Sample Size

Region	n_2	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_{2,p}$
JU	61	400.56	398.51	395.90	390.29	398.51
		(26.76)	(27.93)	(27.87)	(28.17)	(27.92)
SP	66	414.07	397.12	402.41	396.40	396.40
		(38.84)	(38.58)	(38.38)	(38.19)	(38.36)
PA	62	445.50	443.05	440.15	437.33	442.83
		(32.86)	(35.21)	(35.80)	(36.98)	(35.56)
AL	113	336.98	324.54	324.30	321.00	321.60
		(25.98)	(25.89)	(26.74)	(29.98)	(26.31)
SA	45	317.10	312.66	296.60	300.78	310.52
		(33.59)	(36.28)	(36.02)	(36.36)	(38.86)
CH	347	379.63	370.98	368.85	365.69	369.73
		(14.35)	(14.67)	(14.86)	(15.58)	(14.81)

Swiss National Forest Inventory, 2003.

Table A.6: **Estimated Influence of Second Stage Variance as Proportion of Total Variance of $\hat{Y}_1^*$ For Different Sample Sizes**

Region	*Full*	1/2	1/4	1/8	1/16
JU	0.052	0.066	0.078	0.053	0.041
SP	0.026	0.037	0.036	0.030	0.025
PA	0.028	0.043	0.038	0.037	0.035
AL	0.025	0.031	0.029	0.025	0.022
SA	0.025	0.036	0.030	0.031	0.023
CH	0.028	0.038	0.037	0.031	0.027

Swiss National Forest Inventory, 2003.

Table A.7: **Ratio of Empirical Variances to Asymptotic Variances Described in (3.12) For Different Sample Sizes.**

Estimator	Region	*Full*	1/2	1/4	1/8	1/16
$\hat{Y}_2$	JU	0.9965	0.9548	0.9809	1.0038	1.0105
$\hat{Y}_3$	...	1.0636	0.9778	0.9901	1.0051	1.0320
$\hat{Y}_2$	SP	1.0098	1.0117	1.0407	1.0623	1.0013
$\hat{Y}_3$	...	1.0524	1.0375	1.0448	1.0550	0.9918
$\hat{Y}_2$	PA	1.0132	1.0131	1.0405	1.0437	1.0466
$\hat{Y}_3$	...	1.0420	1.0680	1.1116	1.1436	1.1167
$\hat{Y}_2$	AL	0.9982	1.0165	1.0382	1.0593	1.0806
$\hat{Y}_3$	...	1.0696	1.1088	1.1555	1.2086	1.3583
$\hat{Y}_2$	SA	0.9586	0.9783	0.9825	0.9942	0.9948
$\hat{Y}_3$	...	1.0918	1.1038	1.0577	1.0289	1.0136
$\hat{Y}_2$	CH	1.0044	1.0077	1.0310	1.0471	1.0395
$\hat{Y}_3$	...	1.0585	1.0674	1.0912	1.1249	1.1429

Swiss National Forest Inventory, 2003.

Legend:
Ratios are derived from $\frac{\hat{\mathbb{V}}\hat{Y}_2(x)}{\hat{\mathbb{V}}\hat{Y}_2(x)asymptotic}$ and $\frac{\hat{\mathbb{V}}\hat{Y}_3(x)}{\hat{\mathbb{V}}\hat{Y}_3(x)asymptotic}$ respectively where the asymptotic variance formula is described in (3.12).

Table A.8: **Empirical Relative Differences of Variance Estimates For $\hat{Y}_2$ and $\hat{Y}_3$ Compared to $\hat{Y}_1$, with Anticipated Gain in ().**

Estimator	Region	*Full*	1/2	1/4	1/8	1/16
$\hat{Y}_2$	JU	−2.37%	−6.58%	−4.14%	−1.35%	−0.43%
$\hat{Y}_3$	...	4.20%	−4.32%	−3.24%	−1.23%	1.68%
	...	(−2.03%)	(−2.16%)	(−2.27%)	(−1.73%)	(−1.46%)
$\hat{Y}_2$	SP	−0.21%	−0.04%	2.81%	4.97%	−1.06%
$\hat{Y}_3$	...	3.99%	2.51%	3.21%	4.25%	−2.00%
	...	(−1.19%)	(−1.19%)	(−1.21%)	(−1.19%)	(1.19%)
$\hat{Y}_2$	PA	0.20%	−0.01%	2.74%	3.26%	3.39%
$\hat{Y}_3$	...	3.04%	5.40%	9.77%	13.14%	10.32%
	...	(−1.11%)	(−1.31%)	(−1.25%)	(−1.07%)	(−1.21%)
$\hat{Y}_2$	AL	−1.60%	0.34%	2.55%	4.62%	6.69%
$\hat{Y}_3$	...	5.44%	9.45%	14.13%	19.38%	34.10%
	...	(−1.42%)	(−1.29%)	(−1.22%)	(−1.23%)	(−1.27%)
$\hat{Y}_2$	SA	−5.55%	−3.67%	−3.00%	−1.82%	−1.43%
$\hat{Y}_3$	...	7.57%	8.69%	4.44%	1.61%	0.43%
	...	(−1.47%)	(−1.53%)	(−1.26%)	(−1.25%)	(−0.91%)
$\hat{Y}_2$	CH	−0.90%	−0.62%	1.72%	3.44%	2.68%
$\hat{Y}_3$	...	4.43%	5.26%	7.66%	11.12%	12.89%
	...	(−1.34%)	(−1.38%)	(−1.34%)	(−1.22%)	(−1.23%)

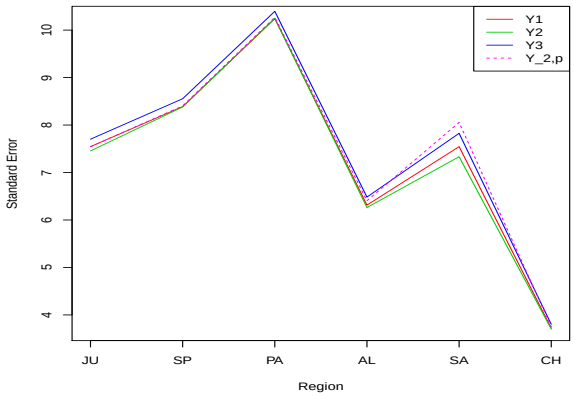
Swiss National Forest Inventory, 2003.

Note:
Anticipated gains are calculated using $\frac{(\hat{\mathbb{V}}_{asy}Y(x) - \hat{\mathbb{V}}Y_1(x))}{\hat{\mathbb{V}}Y_1(x)}$ where $\hat{\mathbb{V}}_{asy}Y(x)$ is calculated in accordance to the asymptotic variance formula described in (3.12). The anticipated gain is therefore the same for both $Y_k(x)$ for $k = 2, 3$ since the 2nd order approximation of $\mathbb{V}_{2,3}Y_k(x)$ yielded the same result. The relative difference for a variance estimate is given by $\frac{(\hat{\mathbb{V}}Y_k(x) - \hat{\mathbb{V}}Y_1(x))}{\hat{\mathbb{V}}Y_1(x)}$.

Appendix B

Figures

Figure B.1: **Visual Comparison of Estimators**



Note: $\hat{Y}_{2,p}$ is represented with a dotted line because it was calculated using the jackknife. In survey statistics, the jackknife is typically applied to the largest sampling units, in this case entire plots. Since the stabilization factors in estimators $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$ are unaffected when a plot is deleted from sample, there is no advantage to calculating jackknife estimates because, on the plot-level, the statistic is simply the mean. For more information regarding the implementation of the jackknife in design based survey data refer to Wolter (2003).

Figure B.2: **Overview of All Plots**

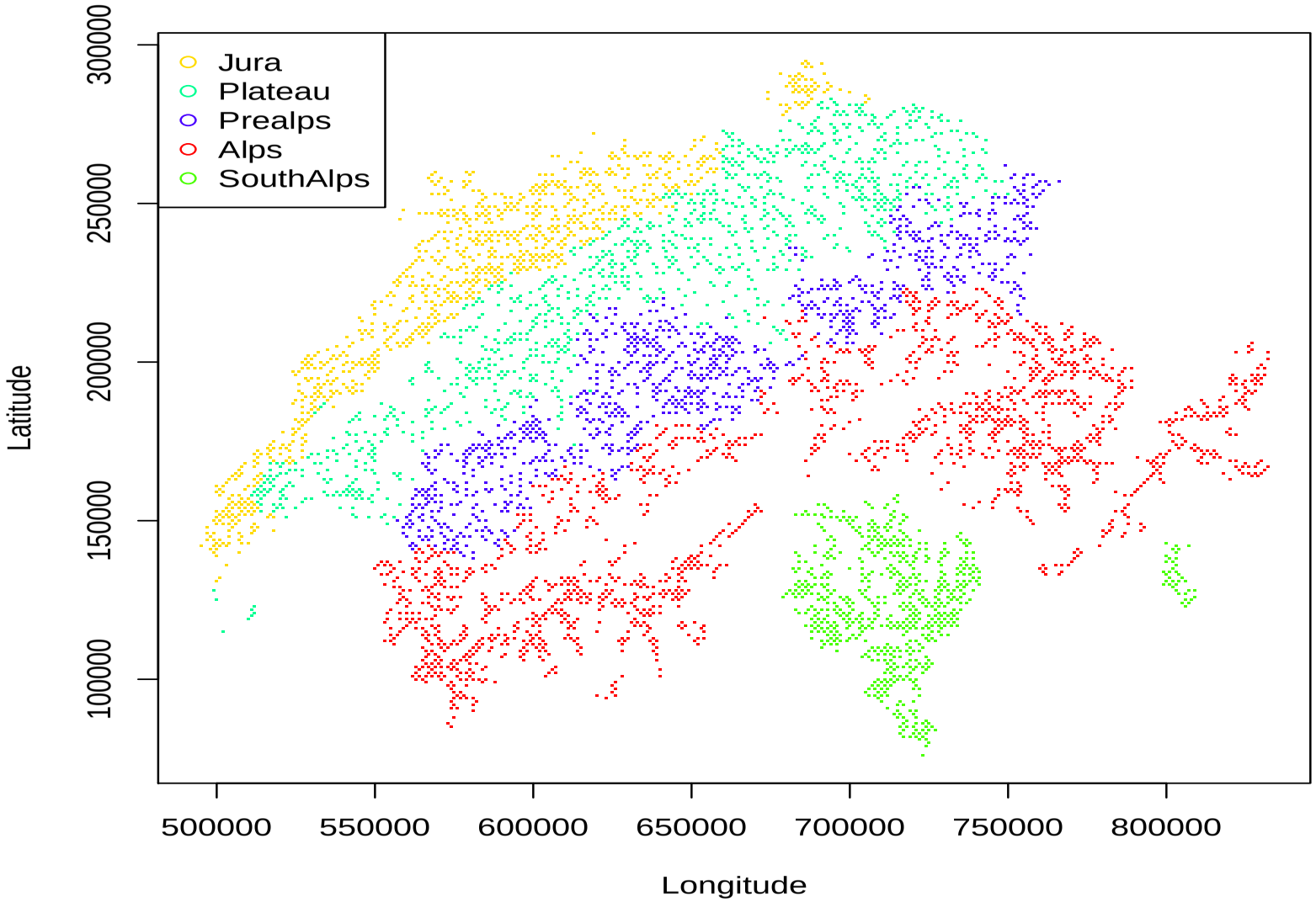


Figure B.3: **Example of Diagnostic Set of Plots Used to Choose Subsamples**

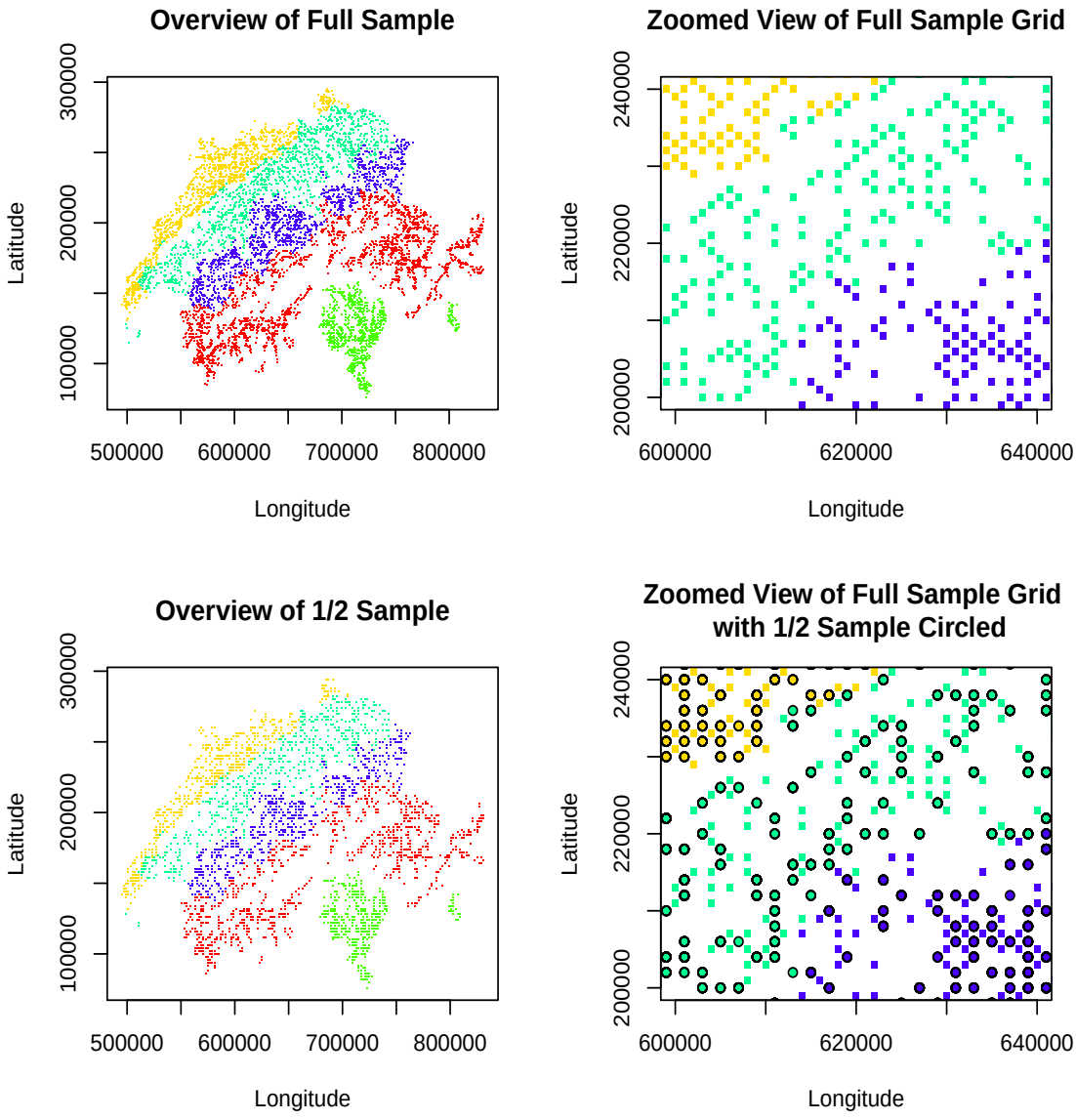


Figure B.4: **Arbitrary Zoomed View of All Subgrids**

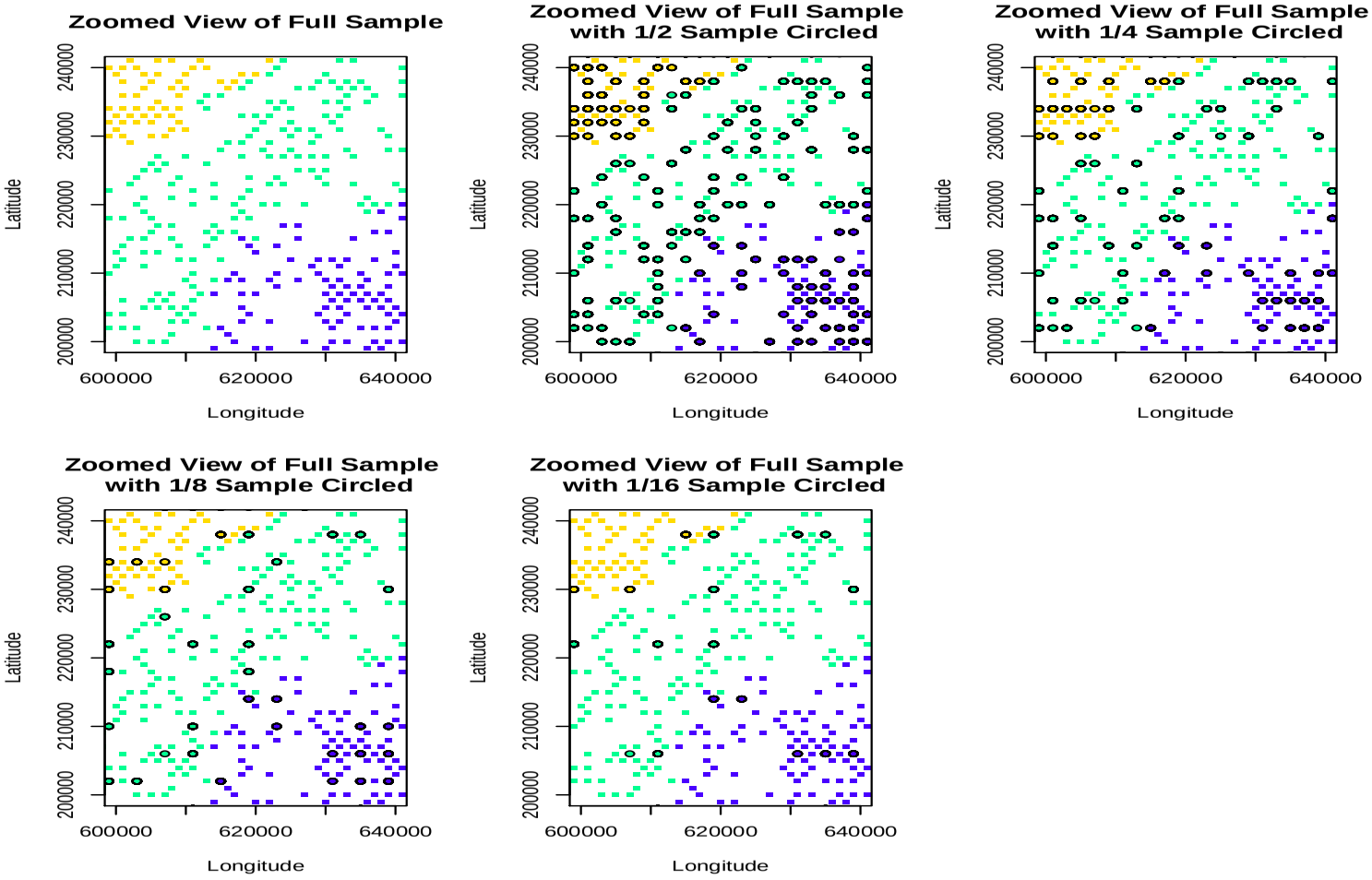


Figure B.5: **Comparison of Geographic Overviews of All Subgrids**

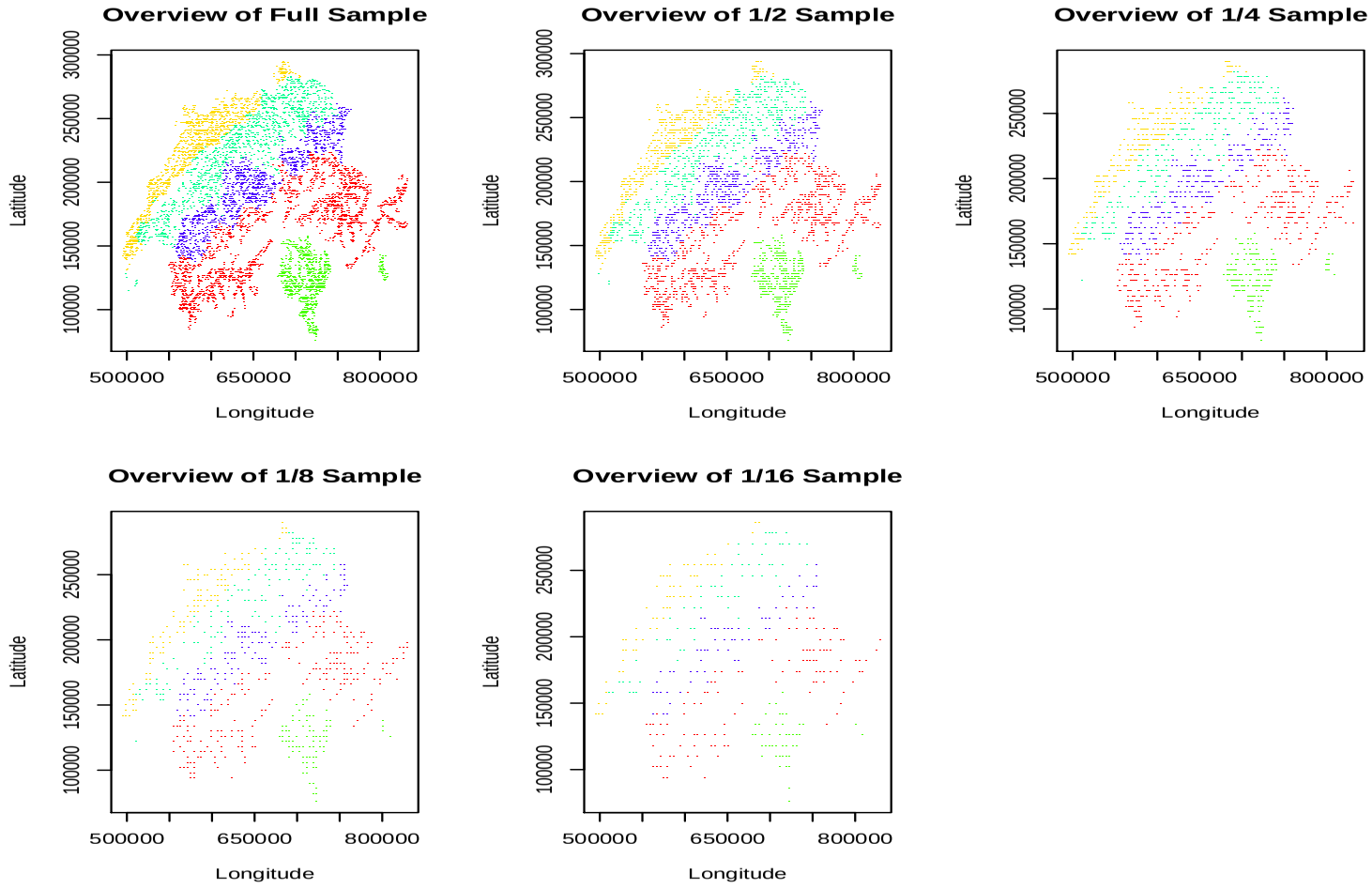


Figure B.6: **Comparison of Estimators Under Varying Sample Size**

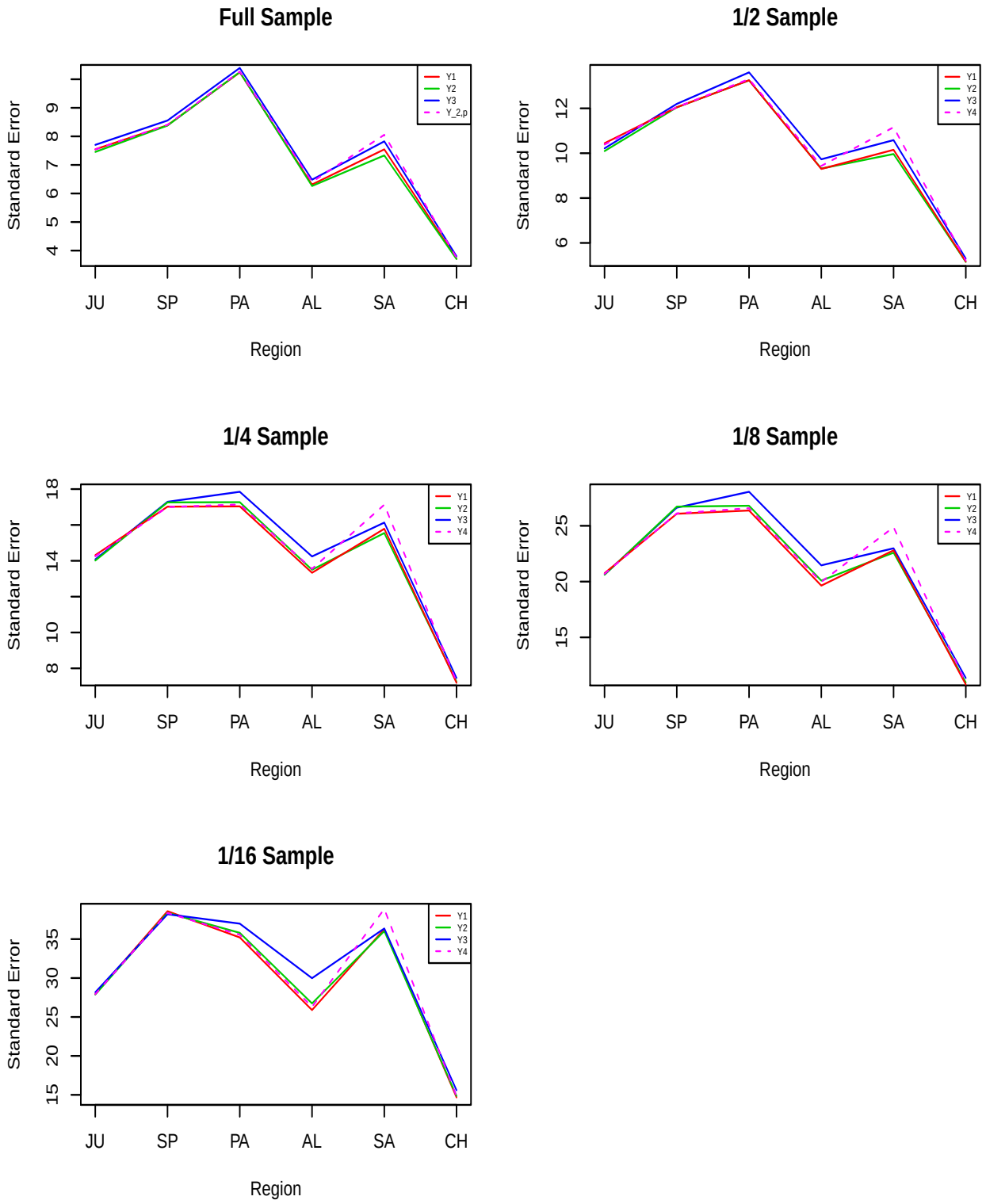


Figure B.7: **Visual Comparison of Stabilization Factors**

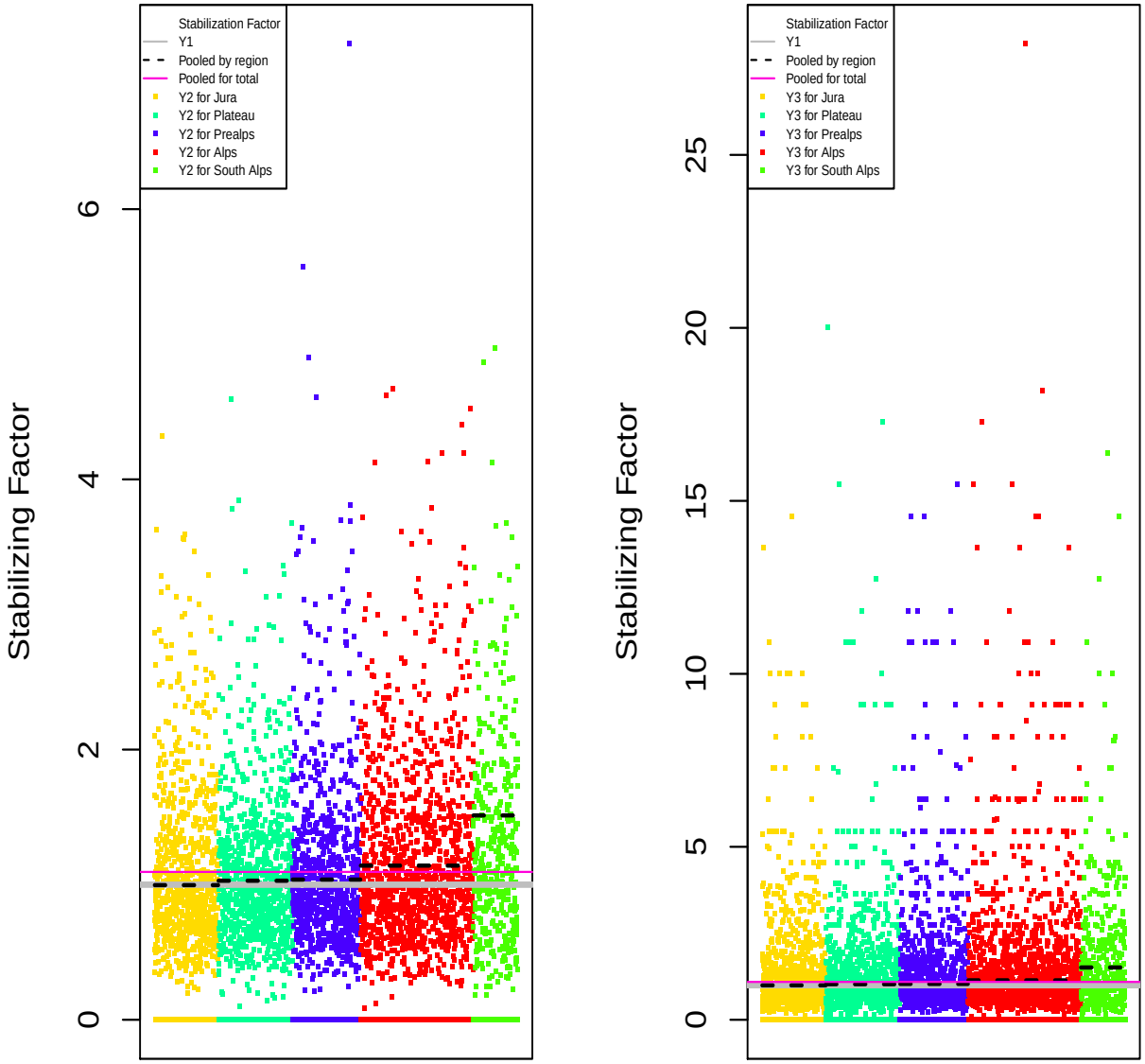
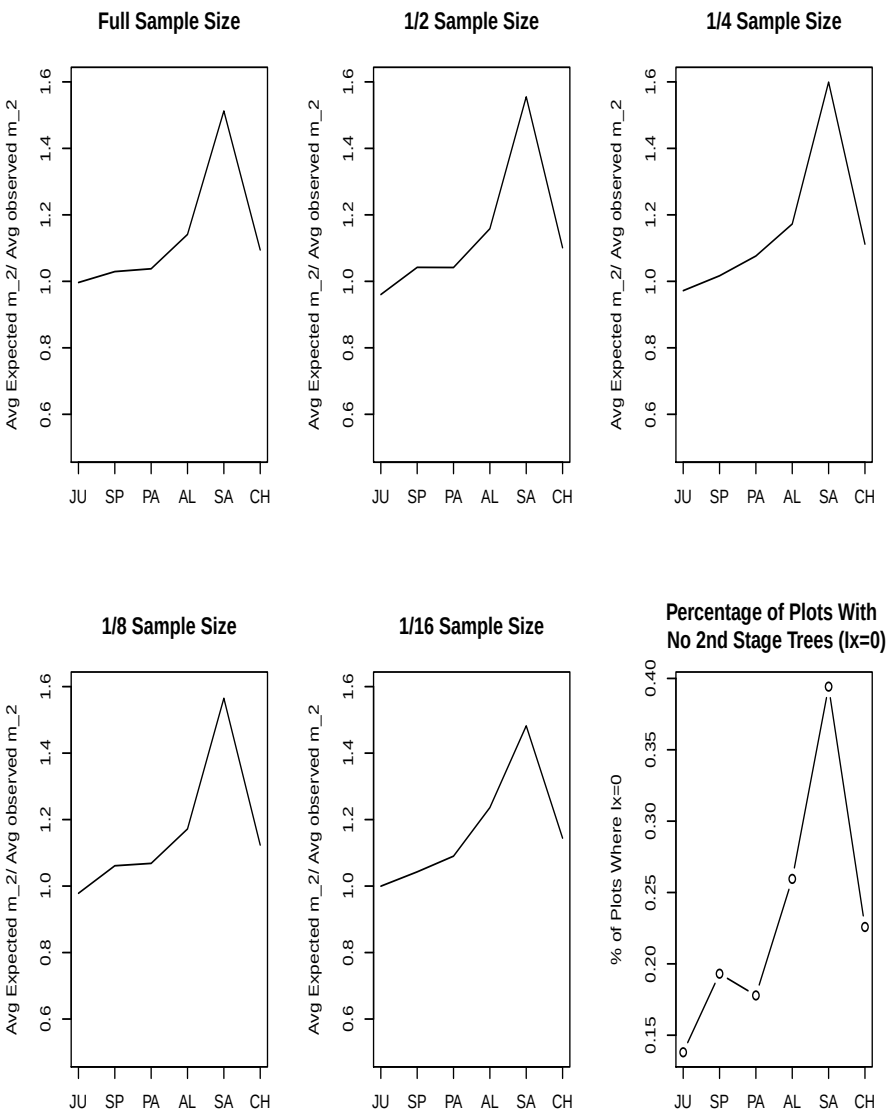


Figure B.8: **Stabilization Factors by Region for Pooled Estimator and Percentage of Plots With No 2nd Stage Trees In Last Panel**



Appendix C

R Functions

```
library(survey)
library(bootstrap)
library(boot)
library(wle)

################################
## FUNCTIONS section begins ##
################################

#################################################
###### 1 Phase 1 Stage Density Estimate Y0 ######
#################################################

Y_OneStage_X <- function(data){
sum(data$FI*data$VMRD)
}

YOneStage <- function(dat){
plot <- factor(dat$PLOT_ID)
loc.density.est <- by(dat,plot, Y_OneStage_X)
n <-  nrow(loc.density.est)
density_est <- mean(loc.density.est)
density_var <- (1/(n*(n-1)))*sum((loc.density.est-density_est)^2)

out <- list(density_est=density_est, n=n, density_var=density_var)  # n is number of plots
return(out)
}

###############################################
###### Generalized Density Estimate Y1 ######
###############################################

Y1_X <- function(data){
sum(data$FI*data$VMRD) + sum(data$FI*data$VPPS)
}

Y1 <- function(dat){
plot <- factor(dat$PLOT_ID)
```

```
loc.density.est <- by(dat,plot,Y1_X)
n <-  nrow(loc.density.est)
density_est <- mean(loc.density.est)
density_var <- (1/(n*(n-1)))*sum((loc.density.est-density_est)^2)

out <- list(density_est=density_est, n=n, density_var=density_var)  # n is number of plots
return(out)
}

##########################################
####### Density Estimate Using Y2 ########
##########################################

Y2_X <- function(data){
ifelse(sum(data$J_i)==0,
stab_factor2 <- 0 ,
stab_factor2 <- sum(data$p_i_2)/sum(data$J_i))
#set residual to zero if Ix=0
sum(data$FI*data$VMRD) + (sum(data$FI*data$VPPS)*(stab_factor2))
}

Y2 <- function(dat){
loc.density.est <- by(dat,factor(dat$PLOT_ID),Y2_X)
n <-  nrow(loc.density.est)
density_est <- mean(loc.density.est)
density_var <- (1/(n*(n-1)))*sum((loc.density.est-density_est)^2)

out <- list(density_est=density_est, n=n, density_var=density_var)  # n is number of plots
return(out)
}

##########################################
####### Density Estimate Using Y3 ########
##########################################

Y3_X <- function(data){
ifelse(sum(data$J_i)==0,
stab_factor3 <- 0 ,
stab_factor3 <- sum(data$I_i)/sum(data$J_i/data$p_i_2))
#set residual to zero if Ix=0
sum(data$FI*data$VMRD) + (sum(data$FI*data$VPPS)*(stab_factor3))
}

Y3 <- function(dat){
loc.density.est <- by(dat,factor(dat$PLOT_ID),Y3_X)
n <-  nrow(loc.density.est)
density_est <- mean(loc.density.est)
density_var <- (1/(n*(n-1)))*sum((loc.density.est-density_est)^2)

out <- list(density_est=density_est, n=n, density_var=density_var)
return(out)
}

##################################################
```

```
##### Density Using Pooled Estimate with Y4  ######
#################################################

Y4 <- function(dat){
n2 <- length(levels(factor(dat$PLOT_ID)))
stab_factor4 <- sum(dat$T_i*dat$p_i_2) / sum(dat$S_i)
density_est <- (sum(dat$FI*dat$T_i*dat$VMRD)/n2)
+ (sum(dat$FI*dat$VPPS*dat$J_i)/n2)*stab_factor4

out <- list(density_est=density_est)
return(out)
}

########################################
###### 2nd Stage Variance Estimate ######
########################################

V_x <- function(data){
R_i <- ifelse(is.na(data$R_i), R_i <-0, R_i <- data$R_i)   # Change R_i with NA to 0
sum(((data$FI*R_i)^2)*(1-data$p_i_2)/(data$p_i_2)^2)
# Defined according to Mandallaz page 72
}



#############################################################################
###### 1st Stage and 2nd Stage components of total Variance Estimate      ######
###### NOTE: Input Y1 as "func".  Defined according to Mandallaz (4.33) on p. 72 #
#############################################################################

V_Y_x <- function(data,func=Y1){
loc.within.plot.var.est <- by(data,factor(data$PLOT_ID),V_x)
n2 <- nrow(loc.within.plot.var.est)
run_desired_function <- func(data)

expected_second_stage_component <- mean(loc.within.plot.var.est)
# estimate of E[V(x)] which is (1/n2)*sum(V(x))

first_stage_component <- (n2*run_desired_function$density_var)
- mean(loc.within.plot.var.est)  # estimate of V[Y(x)]

second_stage_percentage_of_total_var <-
(expected_second_stage_component/n2)/run_desired_function$density_var
#refer (Mandallaz (4.29))

out <- list(expected_second_stage_component = expected_second_stage_component,
first_stage_component = first_stage_component,
second_stage_percentage_of_total_var =
second_stage_percentage_of_total_var, n2=n2)
return(out)
}


#############################################################################
```

```
### Jackknife APPROACH adapted to estimators 1,2,3 for faster computing ###
### Forest_jack works too but will recalculate all plot densities again ###
### for every iteration of the cross-validation.                        ###
#############  NOTE: Input function is Y2_X not Y2 etc. ####################
################################################################################

################################################################################
### Generic Jackknife Function For Swiss National Forest Inventory SLOW!! ###
################################################################################
Forest_jack <- function(data,func,trace=TRUE){
index <- levels(factor(data$PLOT_ID))
n <- length(index)
d_minus_i <- rep(NA,n)
d <- func(data)$density_est
j <- 1
if(trace) {cat(" i = ")}
for(i in 1:n){
if(trace) {cat(ifelse(j %% (n %/% 10) == 1,paste(j, ""), "."))}
d_minus_i[j] <- func(data[data$PLOT_ID!=index[i],])$density_est
j <- j+1
}
dji <- (n*d)-((n-1)*d_minus_i)
jack_est <- mean(dji)
jack_var <- (1/n)*(1/(n-1))*sum((dji-jack_est)^2)
jack_bias <- jack_est-d

out <- list(jack_est=jack_est, jack_var=jack_var, jack_bias=jack_bias)
return(out)
}

#Fast jackknife function streamilned for estimators YOneStage, Y1, Y2 and Y3#
fast_jack <- function(dat,func){
plot <- factor(dat$PLOT_ID)
loc.density.est <- by(dat,plot,func)
n2 <-  nrow(loc.density.est)
d <- mean(loc.density.est)
d_minus_i <- rep(NA,n2)
for(i in 1:n2){  d_minus_i[i] <- mean(loc.density.est[-i])  }
dji <- (n2*d)-((n2-1)*d_minus_i)
jack_est <- mean(dji)
jack_var <- (1/n2)*(1/(n2-1))*sum((dji-jack_est)^2)
jack_bias <- jack_est-d

out <- list(jack_est=jack_est, jack_var=jack_var, jack_bias=jack_bias)
return(out)
}

############################################################
############  Test for different sampling sizes ############
#### NOTE:                                          #####
###  This function returns point ests, std errs, and  #####
###  jackknife bias estimates                         #####
############################################################
```

```
#input subsample data set and this function will call all functions
sample_size_test <- function(data){

cat("calculating Y0... ")

YOneStage_tot <- YOneStage(data)
YOneStage_reg <- by(data,factor(data $REG), YOneStage)

## fast_jack verification of above results
YOneStage_jack <- fast_jack(data, Y_OneStage_X)

#Jackknife by Region
YOneStage_reg_jack <- by(data, factor(data$REG), fast_jack, func=Y_OneStage_X)

cat("calculating Y1... ")

Y1_tot <- Y1(data)
Y1_reg <- by(data,factor(data $REG),Y1)

## fast_jack verification of above results
Y1_jack <- fast_jack(data,Y1_X)

#Jackknife by Region
Y1_reg_jack <- by(data, factor(data$REG), fast_jack, func=Y1_X)

cat("calculating Y2... ")

Y2_tot <- Y2(data)
Y2_reg <- by(data,factor(data$REG),Y2)
Y2_jack <- fast_jack(data,Y2_X)
Y2_reg_jack <- by(data, factor(data$REG), fast_jack, func=Y2_X)

cat("calculating Y3... ")

Y3_tot <- Y3(data)
Y3_reg <- by(data,factor(data$REG),Y3)
Y3_jack <- fast_jack(data,Y3_X)
Y3_reg_jack <- by(data, factor(data$REG), fast_jack, func=Y3_X)

cat("calculating Y4... ")

Y4_tot <- Y4(data)
Y4_reg <- by(data,factor(data$REG),Y4)
cat("calculating slow Forest jackknife routine for total... ")
Y4_jack <- Forest_jack(data,Y4,trace=FALSE)
Y4_jack_est <- Y4_jack$jack_est
Y4_jack_var <- Y4_jack$jack_var
Y4_jack_sd <- sqrt(Y4_jack$jack_var)
cat("calculating slow Forest jackknife routine by region... ")
Y4_reg_jack <- by(data, factor(data$REG), Forest_jack, func=Y4, trace=FALSE)

############ Aggregate Results ############
#gather estimate results in table
est_table <- matrix(c(
```

```
YOneStage_reg$Jura$density_est,     Y1_reg$Jura$density_est,
Y2_reg$Jura$density_est,      Y3_reg$Jura$density_est,      Y4_reg[1],
YOneStage_reg$Plateau$density_est,   Y1_reg$Plateau$density_est,
Y2_reg$Plateau$density_est,   Y3_reg$Plateau$density_est,   Y4_reg[2],
YOneStage_reg$Prealps$density_est,Y1_reg$Prealps$density_est,
Y2_reg$Prealps$density_est,   Y3_reg$Prealps$density_est,   Y4_reg[3],
YOneStage_reg$Alps$density_est,Y1_reg$Alps$density_est,
Y2_reg$Alps$density_est,      Y3_reg$Alps$density_est,      Y4_reg[4],
YOneStage_reg$SouthAlps$density_est,Y1_reg$SouthAlps$density_est,
Y2_reg$SouthAlps$density_est, Y3_reg$SouthAlps$density_est, Y4_reg[5],
YOneStage_tot$density_est,     Y1_tot$density_est,
Y2_tot$density_est,           Y3_tot$density_est,           Y4_tot$density_est)
         ,nrow=6,ncol=5,byrow=TRUE)


#make vector of Attach N to matrix  (note that n here is number of plots)
N_vector <- c(Y1_reg$Jura$n, Y1_reg$Plateau$n, Y1_reg$Prealps$n, Y1_reg$Alps$n,
Y1_reg$SouthAlps$n, Y1_tot$n)
est_table <- cbind(N_vector,est_table)


#input estimates in data.frame and name rows/columns
point_estimates <- as.data.frame(est_table)
row.names(point_estimates) <-
c("Jura","Plateau","Prealps","Alps","South Alps","Switzerland")
names(point_estimates) <- c("N", "Y0", "Y1", "Y2", "Y3", "Y4")

#VIEW POINT ESTIMATES
print(point_estimates)


#gather std. error results in table
make_tab_var <- matrix(c(
YOneStage_reg$Jura$density_var,Y1_reg$Jura$density_var,
Y2_reg$Jura$density_var,        Y3_reg$Jura$density_var,
Y4_reg_jack$Jura$jack_var,  YOneStage_reg$Plateau$density_var,
Y1_reg$Plateau$density_var,   Y2_reg$Plateau$density_var,
Y3_reg$Plateau$density_var,   Y4_reg_jack$Plateau$jack_var,
YOneStage_reg$Prealps$density_var, Y1_reg$Prealps$density_var,
Y2_reg$Prealps$density_var,   Y3_reg$Prealps$density_var,
Y4_reg_jack$Prealps$jack_var, YOneStage_reg$Alps$density_var,
Y1_reg$Alps$density_var,      Y2_reg$Alps$density_var,
Y3_reg$Alps$density_var,      Y4_reg_jack$Alps$jack_var,
YOneStage_reg$SouthAlps$density_var, Y1_reg$SouthAlps$density_var,
Y2_reg$SouthAlps$density_var, Y3_reg$SouthAlps$density_var,
Y4_reg_jack$SouthAlps$jack_var, YOneStage_tot$density_var,
Y1_tot$density_var,           Y2_tot$density_var,
Y3_tot$density_var,           Y4_jack_sd^2)
       ,nrow=6,ncol=5,byrow=TRUE)

#make std. error
std_err_table <- sqrt(make_tab_var)

#input std. err in data.frame and name rows/columns
```

```
standard_errors <- as.data.frame(std_err_table)
row.names(standard_errors) <-
c("Jura","Plateau","Prealps","Alps","South Alps","Switzerland")
names(standard_errors) <-
c("std(Y0)", "std(Y1)", "std(Y2)", "std(Y3)", "std(Y4)_jk")


#VIEW STANDARD ERROR OF POINT ESTIMATES
print(standard_errors)


#gather bias results in table
bias_table <- matrix(c(
YOneStage_reg_jack$Jura$jack_bias,Y1_reg_jack$Jura$jack_bias,
Y2_reg_jack$Jura$jack_bias,     Y3_reg_jack$Jura$jack_bias,
Y4_reg_jack$Jura$jack_bias, YOneStage_reg_jack$Plateau$jack_bias,
Y1_reg_jack$Plateau$jack_bias,  Y2_reg_jack$Plateau$jack_bias,
Y3_reg_jack$Plateau$jack_bias,  Y4_reg_jack$Plateau$jack_bias,
YOneStage_reg_jack$Prealps$jack_bias,Y1_reg_jack$Prealps$jack_bias,
Y2_reg_jack$Prealps$jack_bias,  Y3_reg_jack$Prealps$jack_bias,
Y4_reg_jack$Prealps$jack_bias,YOneStage_reg_jack$Alps$jack_bias,
Y1_reg_jack$Alps$jack_bias,     Y2_reg_jack$Alps$jack_bias,
Y3_reg_jack$Alps$jack_bias,     Y4_reg_jack$Alps$jack_bias,
YOneStage_reg_jack$SouthAlps$jack_bias,Y1_reg_jack$SouthAlps$jack_bias,
Y2_reg_jack$SouthAlps$jack_bias,Y3_reg_jack$SouthAlps$jack_bias,
Y4_reg_jack$SouthAlps$jack_bias,YOneStage_jack$jack_bias,
Y1_jack$jack_bias,              Y2_jack$jack_bias,
Y3_jack$jack_bias,              Y4_jack$jack_bias)
         ,nrow=6,ncol=5,byrow=TRUE)


#input bias in data.frame and name rows/columns
bias_estimates <- as.data.frame(bias_table)
row.names(bias_estimates) <-
c("Jura","Plateau","Prealps","Alps","South Alps","Switzerland")
names(bias_estimates) <-
c("bias(Y0)", "bias(Y1)", "bias(Y2)", "bias(Y3)", "bias(Y4)")


#VIEW BIAS ESTIMATES OF POINT ESTIMATES
print(bias_estimates)


out <- list(point_estimates=point_estimates, standard_errors=standard_errors,
bias_estimates=bias_estimates)
return(out)
}


##############################################################################
###### The function displays the mean second stage variance component as  ######
###### percentage of empirical total variance.                            ######
###### It is not clear whether running V_Y_x with other estimators than   ######
###### Y1 makes any sense since V_x will be fixed according to the definition ##
```

```
##############################################################################

second_stage_percentage_of_total_variance_table <- function(data){
Y1_percent <- V_Y_x(data,Y1)
Y2_percent <- V_Y_x(data,Y2)
Y3_percent <- V_Y_x(data,Y3)

Y1_reg_percent <- by(data, factor(data$REG), V_Y_x, func=Y1)
Y2_reg_percent <- by(data, factor(data$REG), V_Y_x, func=Y2)
Y3_reg_percent <- by(data, factor(data$REG), V_Y_x, func=Y3)


percent_matrix <- matrix(c(
Y1_reg_percent$Jura$second_stage_percentage_of_total_var,
Y2_reg_percent$Jura$second_stage_percentage_of_total_var,
Y3_reg_percent$Jura$second_stage_percentage_of_total_var,
Y1_reg_percent$Plateau$second_stage_percentage_of_total_var,
Y2_reg_percent$Plateau$second_stage_percentage_of_total_var,
Y3_reg_percent$Plateau$second_stage_percentage_of_total_var,
Y1_reg_percent$Prealps$second_stage_percentage_of_total_var,
Y2_reg_percent$Prealps$second_stage_percentage_of_total_var,
Y3_reg_percent$Prealps$second_stage_percentage_of_total_var,
Y1_reg_percent$Alps$second_stage_percentage_of_total_var,
Y2_reg_percent$Alps$second_stage_percentage_of_total_var,
Y3_reg_percent$Alps$second_stage_percentage_of_total_var,
Y1_reg_percent$SouthAlps$second_stage_percentage_of_total_var,
Y2_reg_percent$SouthAlps$second_stage_percentage_of_total_var,
Y3_reg_percent$SouthAlps$second_stage_percentage_of_total_var,
Y1_percent$second_stage_percentage_of_total_var,
Y2_percent$second_stage_percentage_of_total_var,
Y3_percent$second_stage_percentage_of_total_var)
, nrow=6, ncol=3, byrow=TRUE)

#input percents in data.frame and name rows/columns
second_stage_variance_proportion_of_total_variance_table
<- as.data.frame(percent_matrix)
row.names(second_stage_variance_proportion_of_total_variance_table)
<- c("Jura","Plateau","Prealps","Alps","South Alps","Switzerland")
names(second_stage_variance_proportion_of_total_variance_table) <- c("Y1", "Y2", "Y3")


#VIEW PERCENT ESTIMATES OF POINT ESTIMATES
print(second_stage_variance_proportion_of_total_variance_table)

out <- list(second_stage_variance_proportion_of_total_variance_table =
second_stage_variance_proportion_of_total_variance_table)
return(out)
}

##########################################################################
### Asymptotic Variance function                                       ###
### NOTE: This function can be used with estimator Y2                  ###
##########################################################################
```

```
asy_variance_first <- function(data,func){
m2 <-  mean(by(data$J_i,factor(data$PLOT_ID),sum))
# mean number of 2nd stage trees observed per plot
E2Po_x <-  exp(-m2)
# Expected value of exp(-sum(Ii(x)*pi))
R_i <-  ifelse(is.na(data$R_i), R_i <-0, R_i <- data$R_i)
#Change R_i with NA to 0
V2_Y_x <-  V_Y_x(data,Y1)
#Return object with average 2nd stage variance
#and first stage variance
R_x_vector <-  sum((data$FI)*R_i/data$P_i_2)
#weighted up sum of residuals
Rbar <-  mean(R_x_vector)
ER_x_squared <- mean(R_x_vector^2)-V2_Y_x$expected_second_stage_component
asy_var <- (V2_Y_x$first_stage_component
+ (1-E2Po_x)* V2_Y_x$expected_second_stage_component
+ E2Po_x*ER_x_squared
- E2Po_x*Rbar^2)/V2_Y_x$n2

out <- list(asy_var = asy_var, m2 = m2, E2Po_x = E2Po_x)
return(out)
}

#####################################################################
## Second order aymptotic variance calculation for Y2 and Y3 ##
#####################################################################
asy_variance_second <- function(data,func){
m2 <-  mean(by(data$J_i,factor(data$PLOT_ID),sum))
# mean number of 2nd stage trees observed per plot
m1 <-  mean(by(data$I_i,factor(data$PLOT_ID),sum))
# mean number of 1st stage trees observed per plot
E2Po_x <-  exp(-m2)
# Expected value of exp(-sum(Ii(x)*pi))
R_i <-  ifelse(is.na(data$R_i), R_i <-0, R_i <- data$R_i)
#Change R_i with NA to 0
V2_Y_x <-  V_Y_x(data,Y1)
#Return object with average 2nd stage variance
#and first stage variance
data$residual <- (data$FI)*data$R_i/data$p_i_2
#temporarily append vector of weighted up residuals
R_x_vector <-  by(data$residual,factor(data$PLOT_ID),sum, na.rm=TRUE)
#residual estimate by plot
Rbar <-  mean(R_x_vector)
#Average residual estimate
ER_x_squared <- mean(R_x_vector^2)-V2_Y_x$expected_second_stage_component
asy_bias_estimate <- exp(-m2)*Rbar
#first and second order asymptotic bias estimate of Y2 and Y3
asy_var <- (V2_Y_x$first_stage_component + (E2Po_x^2)*(ER_x_squared-Rbar^2)
+ ((1-E2Po_x)^2)*V2_Y_x$expected_second_stage_component
- ((1-E2Po_x)^2)*((1/m2)-(1/m1))*ER_x_squared
+ (1-E2Po_x)*E2Po_x*ER_x_squared)/V2_Y_x$n2

out <- list(asy_var = asy_var, m2 = m2, m1 = m1, E2Po_x = E2Po_x,
asy_bias_estimate = asy_bias_estimate)
```

```
return(out)
}

##############################################################################
## Compare different methods for estimating avg 2nd stage sample size m2 ##
##############################################################################
m2_calculator <- function(data){
expected_m2 <- mean(by(data$p_i_2,factor(data$PLOT_ID),sum))
#expected avg number 2nd stage trees per plot
observed_m2 <- mean(by(data$J_i,factor(data$PLOT_ID),sum))
#observed avg number 2nd stage trees per plot

out <- list(expected_m2 = expected_m2, observed_m2 = observed_m2)
return(out)
}

##############################################################################################
## This function outputs a table displaying the relative differences between      ##
## variance estimates                                                             ##
##  i.e.comparing var(Y2)-var(Y1)/var(Y1) and the asymptotic variance estimate    ##
##############################################################################################
relative_difference_table <- function(data, fun){
asy_var_second_for_fun <- asy_variance_second(data, fun)
asy_var_second_reg_for_fun <- by(data,factor(data$REG), asy_variance_second, func=fun)
fun_reg <- by(data,factor(data$REG),fun)
Y1_reg <- by(data,factor(data$REG),Y1)
fun_tot <- fun(data)
Y1_tot <- Y1(data)

#Relative difference of empirical variance to empirical variance
empirical_improvement_fun <-
round(matrix(c((fun_reg$Jura$density_var-Y1_reg$Jura$density_var)
/Y1_reg$Jura$density_var,
(fun_reg$Plateau$density_var-Y1_reg$Plateau$density_var)/Y1_reg$Plateau$density_var,
(fun_reg$Prealps$density_var-Y1_reg$Prealps$density_var)/Y1_reg$Prealps$density_var,
(fun_reg$Alps$density_var-Y1_reg$Alps$density_var)/Y1_reg$Alps$density_var,
(fun_reg$SouthAlps$density_var-Y1_reg$SouthAlps$density_var)
/Y1_reg$SouthAlps$density_var,
(fun_tot$density_var-Y1_tot$density_var)/Y1_tot$density_var)
,nrow=6,ncol=1),4)

#Relative difference of asymptotic variance (23) to empirical variance
estimate_improvement_fun <-
round(matrix(c((asy_var_second_reg_for_fun$Jura$asy_var-Y1_reg$Jura$density_var)
/Y1_reg$Jura$density_var,
(asy_var_second_reg_for_fun$Plateau$asy_var-Y1_reg$Plateau$density_var)
/Y1_reg$Plateau$density_var,
(asy_var_second_reg_for_fun$Prealps$asy_var-Y1_reg$Prealps$density_var)
/Y1_reg$Prealps$density_var,
(asy_var_second_reg_for_fun$Alps$asy_var-Y1_reg$Alps$density_var)
/Y1_reg$Alps$density_var,
(asy_var_second_reg_for_fun$SouthAlps$asy_var-Y1_reg$SouthAlps$density_var)
/Y1_reg$SouthAlps$density_var,
(asy_var_second_for_fun$asy_var-Y1_tot$density_var)/Y1_tot$density_var)
```

```
,nrow=6,ncol=1),4)

#Table of Relative Differences
rel_difference_fun <- as.data.frame(cbind(estimate_improvement_fun,
empirical_improvement_fun))
names(rel_difference_fun) <- list("estimated","empirical")
row.names(rel_difference_fun) <- list("JU","SP","PA","AL","SA","CH")
rel_difference_fun
}


##############################################################################################
#Intermediary function for Ix_0 function used to find no 2ns stage trees exist in Plot #
##############################################################################################
ind_prod <- function(d){
prod(d$Ix)
}

##############################################################################################
## This function returns a percentage of plots with no 2nd stage selected in data file ##
##############################################################################################
Ix_0 <- function(data){
test_dat2 <- data[,c(1,8)]
test_dat2$Ix <- ifelse(is.na(test_dat2$WV),1,0)
#if NA then 1, if number then 0

ind <- by(test_dat2,factor(test_dat2$PLOT_ID),ind_prod)
table(ind==1)[2]/ (table(ind==1)[1]+table(ind==1)[2])
#True means all WV values are NA in given plot
}

#########################################################################
## Combined with a by statement the following two functions       ##
## calculate stabilization factors for Y2 and Y3                 ##
#########################################################################
stabilizing_factor2 <- function(data){
stab_fact <-(sum(data$p_i_2)/sum(data$J_i))
stab_fact[stab_fact==Inf]<-0
stab_fact
}

stabilizing_factor3 <- function(data){
stab_fact <-(sum(data$I_i)/(sum((data$I_i*data$J_i)/data$p_i_2)))
stab_fact[stab_fact==Inf]<-0
stab_fact
}

stabilizing_factor_pooled <- function(data){
stab_fact <-(sum(data$T_i*data$p_i_2)/sum(data$S_i))
stab_fact[stab_fact==Inf]<-0
stab_fact
}

#Average expected 2nd stage sample size over Average observed 2nd stage sample
```

```
compare2ndstagesamples <- function(data){
reg <- by(data, factor(data$REG), m2_calculator)
JU <- reg$Jura$expected_m2 / reg$Jura$observed_m2
SP <- reg$Plateau$expected_m2 / reg$Plateau$observed_m2
PA <- reg$Prealps$expected_m2 / reg$Prealps$observed_m2
AL <- reg$Alps$expected_m2 / reg$Alps$observed_m2
SA <- reg$SouthAlps$expected_m2 / reg$SouthAlps$observed_m2
CH <- m2_calculator(data)$expected_m2 / m2_calculator(data)$observed_m2
c(JU,SP,PA,AL,SA,CH)
}


##################################
#                                #
# ############################## #
# # FUNCTIONS section ends #### #
# ############################## #
#                                #
##################################


####################################################################
# ************************************************************** #
# ************************************************************** #
# ************************************************************** #
####################################################################



############################################################
## Start Section: LOAD DATA SET AND CREATE VARIABLES ##
############################################################

#Master Data Set
SNFI3 <- read.csv2("/Users/alexandermassey/Desktop/Masters Thesis/lfi3.txt",
dec=".", strip.white=TRUE)

#Designation for Master Data Set with numerical variable codes
NUMERIC_SNFI3 <- read.csv2("/Users/alexandermassey/Desktop/Masters Thesis/lfi3.txt",
dec=".", strip.white=TRUE)

#Numeric region variable for color coding plots
SNFI3$NUM_REG <- NUMERIC_SNFI3$REG

#Data set containing the X, Y coordinates for each PLOT_ID
COORDINATES <- read.csv2("/Users/alexandermassey/Desktop/Masters Thesis/koordinaten.txt",
dec=".", strip.white=TRUE)


#Dump coordinates for plots unused in SNFI3
COORDINATES <- COORDINATES[COORDINATES$PLOT_ID %in% SNFI3$PLOT_ID,]
SNFI3$X <- NA
SNFI3$Y <- NA
```

```
#second_stage_tree_counter vector counts how many trees are in each plot
second_stage_tree_counter <- by(SNFI3[order(SNFI3$PLOT_ID),],factor(SNFI3$PLOT_ID),nrow)
COORD_SORT <- COORDINATES[order(COORDINATES$PLOT_ID),]

#Create matching lengthed coordinate variable
X_Coord_sorted <- rep(COORD_SORT$X, second_stage_tree_counter)
Y_Coord_sorted <- rep(COORD_SORT$Y, second_stage_tree_counter)

#Add Coordinates to SNFI3
SNFI3[order(SNFI3$PLOT_ID),]$X <- X_Coord_sorted
SNFI3[order(SNFI3$PLOT_ID),]$Y <- Y_Coord_sorted



SNFI3$PLOT_ID <- as.factor(SNFI3$PLOT_ID)
SNFI3$REG <- as.factor(SNFI3$REG)
SNFI3$H <- as.numeric(SNFI3$H)
SNFI3$ST <- as.factor(SNFI3$ST)
SNFI3$SPEC <- as.factor(SNFI3$SPEC)
SNFI3$SG <- as.factor(SNFI3$SG)
SNFI3$D13 <- as.numeric(SNFI3$D13)
SNFI3$WV <- as.numeric(SNFI3$WV)
SNFI3$VMRD <- as.numeric(SNFI3$VMRD)
SNFI3$FI <- as.numeric(SNFI3$FI)
SNFI3$VPPS <- as.numeric(SNFI3$VPPS)

#Assign Levels
levels(SNFI3$REG) <- c("Jura", "Plateau", "Prealps", "Alps", "SouthAlps")
levels(SNFI3$SPEC) <-
c("Spruce","Fir","Pine","Larch","Pinus cembra","other conifers","beech","aple",
"ash","oak","chestnut","other deciduous trees")
levels(SNFI3$SG) <- c("deciduous", "conifer")



str(SNFI3)

### NEW VARIABLE CREATION ###

#Create Residual Variable
SNFI3$R_i <- SNFI3$WV - SNFI3$VMRD

#Create 1nd stage indicator variable
SNFI3$I_i <- rep(1,nrow(SNFI3))

#Create 1nd stage count variable for pooled estimate (equal to I_i in systematic sampling)
SNFI3$T_i <- SNFI3$I_i

#Create 2nd stage indicator variable
SNFI3$J_i <- ifelse(is.na(SNFI3$R_i),0,1)

#Create 2nd stage count variable for pooled est (equal to J_i in syst sampling)
SNFI3$S_i <- SNFI3$J_i
```

```
#Create 2nd stage inclusion probability p_i
SNFI3$p_i <- (SNFI3$R_i)/SNFI3$VPPS

SNFI3$p_i_measurable <- 0.91

SNFI3$p_i_selection <- ifelse(0.000015*(SNFI3$D13^2)*SNFI3$FI >= 1,
SNFI3$p_i_selection <- 1,
SNFI3$p_i_selection <- 0.000015*(SNFI3$D13^2)*SNFI3$FI)

SNFI3$p_i_sector <- ifelse(SNFI3$D13 < 60, SNFI3$p_i_sector <- 150/400, 1)

#Create 2nd stage inclusion probability p_i
SNFI3$p_i <- (SNFI3$R_i)/SNFI3$VPPS

# THIS IS THE GO-TO variable for the 2nd stage prob because it is defined for all trees
SNFI3$p_i_2 <- SNFI3$p_i_selection*SNFI3$p_i_measurable*SNFI3$p_i_sector


#SORT SNFI3 BY REGION AND THEN PLOT_ID
ii <- order(SNFI3$REG, SNFI3$PLOT_ID)
SNFI3 <- SNFI3[ii,]
NUMERIC_SNFI3 <- NUMERIC_SNFI3[ii,]

#Create Region Subfiles
Jura <- SNFI3[SNFI3$REG=="Jura",]
Plateau <- SNFI3[SNFI3$REG=="Plateau",]
Prealps <- SNFI3[SNFI3$REG=="Prealps",]
Alps <- SNFI3[SNFI3$REG=="Alps",]
SouthAlps <- SNFI3[SNFI3$REG=="SouthAlps",]

#Create Subsamples files of SNFI3 that preserve geographic structure
#NOTE: If X coordinate / 1000 is odd then Y/1000 is even and vice versa

SNFI3_X <- (SNFI3[!duplicated(SNFI3$X),]$X / 1000)  #List of all X coordinates
SNFI3_Y <- (SNFI3[!duplicated(SNFI3$Y),]$Y / 1000)  #List of all Y coordinates

SNFI3_X <- SNFI3_X[order(SNFI3_X)]  #Order X Coordinates
SNFI3_Y <- SNFI3_Y[order(SNFI3_Y)]  #Order Y Coordinates


###
#Select every other even X Coordinate (74.33% of original SNFI3 sample)
SNFI3_Every_4th <- (SNFI3$X / 1000) %in% SNFI3_X[as.logical(SNFI3_X %% 4)]
SNFI3_3Quarter_sample <- SNFI3[SNFI3_Every_4th,]
length(SNFI3_3Quarter_sample$PLOT_ID)/length(SNFI3$PLOT_ID)
# percentage of original plots

###
#Select even X Coordinates (corresponds to 48.47% of original SNFI3 sample)
SNFI3_Every_2nd <- (SNFI3$X / 1000) %in% SNFI3_X[as.logical(SNFI3_X %% 2)]
SNFI3_half_sample <- SNFI3[SNFI3_Every_2nd,]
length(SNFI3_half_sample$PLOT_ID)/length(SNFI3$PLOT_ID)  # percentage of original plots
```

```
###
#Select every other Y Coordinate from SNFI3_half_sample (24.94799% of sample)
SNFI3_half_sample_Y <- (SNFI3_half_sample[!duplicated(SNFI3_half_sample$Y),]$Y / 1000)
#List Y coord of half_sample
SNFI3_half_sample_Y <- SNFI3_half_sample_Y[order(SNFI3_half_sample_Y)]
#Order Y Coordinates
SNFI3_half_half <-
(SNFI3_half_sample$Y / 1000)
%in%
SNFI3_half_sample_Y[seq(from = 1, to = length(SNFI3_half_sample_Y), by = 2)]
SNFI3_half_half_sample <- SNFI3_half_sample[SNFI3_half_half,]
length(SNFI3_half_half_sample$PLOT_ID)/length(SNFI3$PLOT_ID)  # percentage of original plots

###
#Select every other X Coordinate from SNFI3_half_half_sample (0.1269252% of sample)
SNFI3_half_half_sample_X <-
(SNFI3_half_half_sample[!duplicated(SNFI3_half_half_sample$X),]$X / 1000)
#List X coord of half_half_sample
SNFI3_half_half_sample_X <- SNFI3_half_half_sample_X[order(SNFI3_half_half_sample_X)]
#Order Y Coordinates
SNFI3_one_eighth <-
(SNFI3_half_half_sample$X / 1000)
%in%
SNFI3_half_half_sample_X[seq(from = 1, to = length(SNFI3_half_half_sample_X), by = 2)]
SNFI3_one_eighth_sample <- SNFI3_half_half_sample[SNFI3_one_eighth,]
length(SNFI3_one_eighth_sample$PLOT_ID)/length(SNFI3$PLOT_ID)  # percentage of original plots

###
#Select every other Y Coordinate from SNFI3_one_eighth_sample (0.06375292% of sample)
SNFI3_one_eighth_sample_Y <-
(SNFI3_one_eighth_sample[!duplicated(SNFI3_one_eighth_sample$Y),]$Y / 1000)
#List Y coord of one_eighth_sample
SNFI3_one_eighth_sample_Y <- SNFI3_one_eighth_sample_Y[order(SNFI3_one_eighth_sample_Y)]
#Order Y Coordinates
SNFI3_one_sixteenth <-
(SNFI3_one_eighth_sample$Y / 1000)
%in%
SNFI3_one_eighth_sample_Y[seq(from = 1, to = length(SNFI3_one_eighth_sample_X), by = 2)]
SNFI3_one_sixteenth_sample <- SNFI3_one_eighth_sample[SNFI3_one_sixteenth,]
length(SNFI3_one_sixteenth_sample$PLOT_ID)/length(SNFI3$PLOT_ID)
# percentage of original plots




#########################################################
## End Section: LOAD DATA SET AND CREATE VARIABLES ##
#########################################################
```