



Report

Note on the design-based estimation of densities, totals and goodness of fit in extensive inventories

Author(s):

Mandallaz, Daniel

Publication Date:

2014

Permanent Link:

<https://doi.org/10.3929/ethz-a-010151635> →

Rights / License:

[In Copyright - Non-Commercial Use Permitted](#) →

This page was generated automatically upon download from the [ETH Zurich Research Collection](#). For more information please consult the [Terms of use](#).

Note on the design-based estimation of densities,
totals and goodness of fit in extensive inventories

Daniel Mandallaz ¹

Chair of Land Use Engineering

ETH Zurich

CH 8092 Zurich, Switzerland

April 2014

¹Tel. ++41(0)44 6323186 e-mail daniel.mandallaz@env.ethz.ch

Abstract

This note intends to clarify some per se elementary considerations about the estimation of densities, totals and goodness of fit in extensive inventories, which, rather surprisingly, do not seem to be common knowledge in the community of forest inventors. The main difficulty is due to the fact that the forest area F is only part of a given territory \tilde{F} , e.g. the country or a region thereof, and that its exact delineation and surface area $\lambda(F)$ are usually unknown, in contrast to \tilde{F} . A common practice to bypass this problem is to estimate a spatial mean density for the variable of interest, e.g. timber volume, over \tilde{F} , and not over the forested area only, and to then multiply this mean density with the total surface area $\lambda(\tilde{F})$. This note shows that this approach can be misleading for the assessment of goodness of fit R^2 in two-phase sampling and inefficient for the estimation of totals if $p_F = \frac{\lambda(F)}{\lambda(\tilde{F})}$ is small and the R^2 is large within F . The same holds for small-area estimation.

1 Introduction

In this note we use the Monte-Carlo approach (infinite population models), as described in Mandallaz (2008).

We consider the forested area F as a subset of \tilde{F} which can be viewed as the entire country for which the sum $Y_F = \sum_{i=1}^N Y_i$ of the response variable Y_i for a well defined population of N trees, e.g. the total volume of the N trees, and the spatial mean $\bar{Y}_F = \frac{1}{\lambda(F)} \sum_{i=1}^N Y_i$ must be estimated. The surface areas of F and \tilde{F} are denoted by $\lambda(F)$ and $\lambda(\tilde{F})$ respectively. We emphasize the fact that we consider only trees (and their values Y_i) in the forested area F , trees outside F are irrelevant.

We now briefly discuss the issues in the simple case of one-phase purely terrestrial inventory.

2 One-phase sampling

We consider the set $s_2 = \{x_1, x_2, \dots, x_{n_2}\}$ of n_2 points independently uniformly distributed in \tilde{F} , at point x_i we have the local density $Y(x_i)$ which by definition is 0 whenever the point is outside the forested area, i.e. $x_i \notin F$. Note however that it could be zero also for some $x \in F$. The local density $Y(x)$ could be e.g. the timber volume in $\frac{m^3}{ha}$ obtained by an inventory using concentric plots or angle count. We have implicitly assumed that the field crew has an unequivocal decision rule to decide for any point x whether it belongs to F or not (actually F is defined as the set of such x which will be classified by observation on the grounds as being in the forested area): this is of course not as clear-cut in practice particularly at the boundaries of F .

The Monte-Carlo approach leads to the following expected value of $Y(x)$ (which in the

design-based approach is a random variable because point x is random):

$$[1] \quad \mathbb{E}_{x \in F} Y(x) = \frac{1}{\lambda(F)} \int_F Y(x) dx = \bar{Y}_F = \frac{1}{\lambda(F)} \sum_{i=1}^N Y_i, \quad Y_F = \lambda(F) \bar{Y}_F$$

The variance of the local density over F is then

$$[2] \quad \mathbb{V}_{x \in F}(Y(x)) = \frac{1}{\lambda(F)} \int_F (Y(x) - \bar{Y}_F)^2 dx$$

The following statement is simple but important for conditional inference: if the n_2 points x_i are uniformly independently distributed in \tilde{F} then given that $n_{2F} \leq n_2$ points are in fact in F these n_{2F} points are uniformly independently distributed in F (see Appendix A).

We first derive the relationships between means and variance over \tilde{F} and F . One has by the definitions

$$[3] \quad \bar{Y}_{\tilde{F}} = \mathbb{E}_{x \in \tilde{F}} Y(x) = \frac{1}{\lambda(\tilde{F})} \int_{\tilde{F}} Y(x) dx = p_F \frac{1}{\lambda(F)} \int_F Y(x) dx = p_F \bar{Y}_F$$

where $p_F = \frac{\lambda(F)}{\lambda(\tilde{F})}$.

For the variance one splits the integrals according to $\int_{\tilde{F}} = \int_{\tilde{F} \setminus F} + \int_F$ and writes

$$Y(x) - \bar{Y}_{\tilde{F}} = Y(x) - \bar{Y}_F + \bar{Y}_F - \bar{Y}_{\tilde{F}}$$

Expanding the square one obtains after some tedious but simple algebra the following important relationship between the variances over \tilde{F} and F

$$[4] \quad \mathbb{V}_{x \in \tilde{F}}(Y(x)) = (1 - p_F) p_F \bar{Y}_F^2 + p_F \mathbb{V}_{x \in F}(Y(x))$$

One estimates the mean over \tilde{F} by

$$[5] \quad \hat{Y}_{\tilde{F}} = \frac{\sum_{i=1}^{n_2} Y(x_i)}{n_2}$$

with theoretical and estimated variances given by

$$[6] \quad \begin{aligned} \mathbb{V}(\hat{Y}_{\tilde{F}}) &= \frac{1}{n_2} \mathbb{V}_{x \in \tilde{F}}(Y(x)) \\ \hat{\mathbb{V}}(\hat{Y}_{\tilde{F}}) &= \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y(x_i) - \hat{Y}_{\tilde{F}})^2 \end{aligned}$$

This leads to the following unbiased estimates for the total

$$[7] \quad \begin{aligned} \hat{T}_{\tilde{F}} &= \lambda(\tilde{F}) \hat{Y}_{\tilde{F}} \\ \mathbb{V}(\hat{T}_{\tilde{F}}) &= \lambda^2(\tilde{F}) \mathbb{V}(\hat{Y}_{\tilde{F}}) \\ \hat{\mathbb{V}}(\hat{T}_{\tilde{F}}) &= \lambda^2(\tilde{F}) \hat{\mathbb{V}}(\hat{Y}_{\tilde{F}}) \end{aligned}$$

One can also restrict the sample to the points in F and considers the conditionally unbiased estimators

$$[8] \quad \begin{aligned} \hat{Y}_F &= \frac{\sum_{x \in s_{2F}} Y(x)}{n_{2F}} \\ \mathbb{V}(\hat{Y}_F) &= \frac{1}{n_{2F}} \mathbb{V}_{x \in F}(Y(x)) \\ \hat{\mathbb{V}}(\hat{Y}_F) &= \frac{1}{n_{2F}} \frac{1}{n_{2F} - 1} \sum_{x \in s_{2F}} (Y(x) - \hat{Y}_F)^2 \end{aligned}$$

where we have set $s_{2F} = s_2 \cap F$ and $n_{2F} = \sum_{x \in s_2} I_F(x)$ is the number of points falling in F .

In [8] the variance are conditional, i.e. given the observed number of points in F , which are independently and uniformly distributed in F . There are very good mathematical

reasons to condition the inference on n_{2F} (random sample sizes as ancillary statistics, see Cox and Hinkley (1974)).

In the ideal case where $\lambda(F)$ is known one can consider the following estimators for the total

$$\begin{aligned}
 \hat{T}_F &= \lambda(F) \hat{Y}_F \\
 \mathbb{V}(\hat{T}_F) &= \lambda^2(F) \mathbb{V}(\hat{Y}_F) \\
 \hat{\mathbb{V}}(\hat{T}_F) &= \lambda^2(F) \hat{\mathbb{V}}(\hat{Y}_F)
 \end{aligned}
 \tag{9}$$

To calculate the **relative efficiency ρ of \hat{T}_F** with respect to the best possible but usually not available estimator \hat{T}_F we **replace n_{2F} by its expected value $n_2 p_F$** and use the obvious identity $\lambda(F) = p_F \lambda(\tilde{F})$, we obtain from [4]

$$\rho = \frac{\mathbb{V}(\hat{T}_{\tilde{F}})}{\mathbb{V}(\hat{T}_F)} = 1 + \frac{(1 - p_F)}{CV_F^2}
 \tag{10}$$

where $CV_F^2 = \frac{\mathbb{V}_{x \in F}(Y(x))}{\bar{Y}_F^2} =: \frac{V_F}{\bar{Y}_F^2}$ is the **squared coefficient of variation of $Y(x)$ over F** .

Hence, if the surface area of the forest is known, \hat{T}_F is better.

If $\lambda(F)$ is unknown an obvious estimator is to take (in the absence of further information)

$$\hat{\lambda}(F) = \frac{n_{2F}}{n_2} \lambda(\tilde{F})$$

and to estimate the total by

$$\hat{T}_F^* = \hat{\lambda}(F) \hat{Y}_F = \frac{\lambda(\tilde{F})}{n_2} n_{2F} \hat{Y}_F$$

To calculate the variance we use the well known variance decomposition (see Mandallaz (2008), pp 221-223)

$$\mathbb{V}(\hat{T}_F^*) = \mathbb{V}_{n_{2F}}(\mathbb{E}(\hat{T}_F^* \mid n_{2F})) + \mathbb{E}_{n_{2F}}(\mathbb{V}(\hat{T}_F^* \mid n_{2F}))$$

Because $\mathbb{E}(\hat{Y}_F \mid n_{2F}) = \bar{Y}_F$ and n_{2F} is binomial with parameters n_2 and p_F the first term gives $\frac{1}{n_2} \lambda^2(\tilde{F}) p_F (1 - p_F) \bar{Y}_F^2$. For the second term we know that the conditional variance $\mathbb{V}(\hat{Y}_F \mid n_{2F})$ is equal to $\frac{V_F}{n_{2F}}$. Finally one obtains

$$[11] \quad \mathbb{V}(\hat{T}_F^*) = \frac{\lambda^2(\tilde{F})}{n_2} (p_F (1 - p_F) \bar{Y}_F^2 + p_F V_F)$$

Using [4], [6] and [7] this leads to

$$[12] \quad \mathbb{V}(\hat{T}_F^*) = \mathbb{V}(\hat{T}_{\tilde{F}})$$

As a matter of fact this result is trivial because, by definition $Y(x) = 0$ for all $x \in \tilde{F} \setminus F$ and therefore $\hat{Y}_{\tilde{F}} = \frac{n_{2F}}{n_2} \hat{Y}_F$ so that

$$\hat{T}_{\tilde{F}} = \hat{T}_F^*$$

Estimating the surface area of the forest only with the proportion of points falling into it does not improve the estimation of the total.

Without external information the best strategy is therefore to choose the set \tilde{F} as the smallest possible set containing F under practical constraints.

In practice one uses one single systematic grid (with random start and possibly also random orientation) so that no design-based variance estimation exists. There is however empirical and up to a certain degree theoretical evidence that treating a systematic grid as a random sample in the sense defined above is acceptable in extensive inventories: the

point estimates are practically unbiased and the variance is usually slightly overestimated. If $\lambda_2(c_o)$ denotes the surface area of the fundamental cell defining the grid it can be shown (Mandallaz (2008), pp 92-93) that $\hat{T} = \lambda_2(c_o) \sum_k Y(x)$ (where k runs over all cells intersecting F) is an unbiased estimate of the total.

We now investigate the problem in the context of two-phase sampling.

3 Two-phase sampling

We follow the description given in Mandallaz (2008) (Chapter 5, in the framework of external models: i.e. ignoring the sampling error in the estimated regression coefficients). We have a sample s_1 of n_1 point independently uniformly distributed in \tilde{F} out of which n_2 are selected by simple random sampling without replacement that are also independently uniformly distributed in \tilde{F} . Restricting the samples to F we have $s_{1F} = s_1 \cap F$ and $s_{2F} = s_2 \cap F$ of n_{1F} and n_{2F} points uniformly distributed in F , and s_{2F} is obtained from s_{1F} by simple random sampling without replacement. This is intuitively plausible and a formal proof is given in Appendix B.

At all points $x \in s_{1F}$ we have the vector of auxiliary information $\mathbf{Z}(x) \in \mathcal{R}^p$ and at all terrestrial points $x \in s_{2F}$ we have the local density $Y(x)$. **By definition we set $\mathbf{Z}(x) = \mathbf{0}$ and $Y(x) = 0$ at all points $x \in \tilde{F} \setminus F$.** We consider the linear model (for details and proofs see Mandallaz (2012))

$$[13] \quad Y(x) = \mathbf{Z}^t(x)\boldsymbol{\beta} + R(x)$$

In the **model-dependent approach** the point x is fixed and $R(x)$ is a random variable with zero mean and a given covariance structure. In the **design-based approach** $Y(x), \mathbf{Z}(x), R(x)$ are random variables because x is random. The true regression coeffi-

cient β is by definition the theoretical least squares estimate minimizing

$$\int_F R^2(x)dx = \int_F (Y(x) - \mathbf{Z}^t(x)\beta)^2 dx$$

It satisfies the normal equation

$$[14] \quad \left(\int_F \mathbf{Z}(x)\mathbf{Z}^t(x)dx \right) \beta = \int_F Y(x)\mathbf{Z}(x)dx$$

and the orthogonality relationship

$$[15] \quad \int_F R(x)\mathbf{Z}(x)dx = \mathbf{0}$$

We shall assume that $\mathbf{Z}(x)$ contains the intercept term 1, or, more generally, that the intercept can be expressed as a linear combination of the component of $\mathbf{Z}(x)$, which then insures that the mean residual is zero, i.e. $\int_F R(x)dx = 0$ (note that the intercept component is also set to 0 for $x \notin F$).

Let

$$[16] \quad \begin{aligned} \mathbf{U}(x) &= Y(x)\mathbf{Z}(x) \\ \mathbf{A}_{s_2} &= \frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x)\mathbf{Z}^t(x) \\ \mathbf{U}_{s_2} &= \frac{1}{n_2} \sum_{x \in s_2} \mathbf{U}(x) \end{aligned}$$

The theoretical and empirical regression vector parameters can then be written as

$$[17] \quad \begin{aligned} \beta &= \mathbf{A}^{-1}\mathbf{U} \\ \hat{\beta}_{s_2} &= \mathbf{A}_{s_2}^{-1}\mathbf{U}_{s_2} \end{aligned}$$

where we have set $\mathbf{A} = \mathbb{E}_{x \in F} \mathbf{Z}(x) \mathbf{Z}^t(x)$ and $\mathbf{U} = \mathbb{E}_{x \in F} \mathbf{Z}(x) Y(x)$. $\hat{\boldsymbol{\beta}}_{s_2}$ is asymptotically design-unbiased for $\boldsymbol{\beta}$.

The empirical predictions and residuals are $\hat{Y}(x) = \mathbf{Z}^t(x) \hat{\boldsymbol{\beta}}_{s_2}$ and $\hat{R}(x) = Y(x) - \hat{Y}(x)$, which are both zero outside F . The asymptotic design-based variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{s_2}$ is

$$[18] \quad \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \mathbf{A}_{s_2}^{-1} \left(\frac{1}{n_2^2} \sum_{x \in s_2} \hat{R}^2(x) \mathbf{Z}(x) \mathbf{Z}(x)^t \right) \mathbf{A}_{s_2}^{-1}$$

Note that the statistically relevant quantities $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{s_2}, \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}$ do not depend on the particular choice of $\tilde{F} \supset F, s_2 \supset s_{2F}$.

We shall consider the following two versions of the regression estimators

$$[19] \quad \begin{aligned} \hat{Y}_{reg, F} &= \frac{1}{n_{1F}} \sum_{x \in s_{1F}} \hat{Y}(x) + \frac{1}{n_{2F}} \sum_{x \in s_{2F}} \hat{R}(x) \\ \hat{Y}_{reg, \tilde{F}} &= \frac{1}{n_1} \sum_{x \in s_1} \hat{Y}(x) + \frac{1}{n_2} \sum_{x \in s_2} \hat{R}(x) \end{aligned}$$

In the following we shall work under the external model assumption, i.e. we neglect the error in $\hat{\boldsymbol{\beta}}_{s_2}$ and set formally $\hat{\boldsymbol{\beta}}_{s_2} \equiv \boldsymbol{\beta}$, which is asymptotically acceptable. One obtains the asymptotic theoretical variance

$$[20] \quad \begin{aligned} \mathbb{V}(\hat{Y}_{reg, \tilde{F}}) &= \frac{1}{n_1} \mathbb{V}_{x \in \tilde{F}}(Y(x)) + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \mathbb{V}_{x \in \tilde{F}}(\hat{R}(x)) \\ &= \frac{1}{n_1} \mathbb{V}_{x \in \tilde{F}}(\hat{Y}(x)) + \frac{1}{n_2} \mathbb{V}_{x \in \tilde{F}}(\hat{R}(x)) \end{aligned}$$

(where we have kept the notation $\hat{Y}(x) = \mathbf{Z}(x)^t \boldsymbol{\beta}$ and $\hat{R}(x) = Y(x) - \mathbf{Z}(x)^t \boldsymbol{\beta}$). This can be estimated by taking a sample copy or, better, by using the asymptotically equivalent g-weight technique based on $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}$ (see Mandallaz (2013a, 2012)). Likewise, one has the

conditional variance

$$\begin{aligned}
\mathbb{V}(\hat{Y}_{reg,F}) &= \frac{1}{n_{1F}} \mathbb{V}_{x \in F}(Y(x)) + \left(1 - \frac{n_{2F}}{n_{1F}}\right) \frac{1}{n_{2F}} \mathbb{V}_{x \in F} \hat{R}(x) \\
[21] \qquad &= \frac{1}{n_{1F}} \mathbb{V}_{x \in F}(\hat{Y}(x)) + \frac{1}{n_{2F}} \mathbb{V}_{x \in F} \hat{R}(x)
\end{aligned}$$

The coefficient of determination (goodness of fit) of the model over F is defined as

$$[22] \qquad R_F^2 = \frac{\mathbb{V}_{x \in F}(\hat{Y}(x))}{\mathbb{V}_{x \in F}(Y(x))} = \frac{\int_F (\hat{Y}(x) - \bar{Y}_F)^2 dx}{\int_F (Y(x) - \bar{Y}_F)^2 dx}$$

Since the design-based covariance between the predictions $\hat{Y}(x)$ and the mean of the residuals $\hat{R}(x)$ is zero over F we obtain

$$\mathbb{V}_{x \in F}(\hat{R}(x)) = (1 - R_F^2) \mathbb{V}_{x \in F}(Y(x))$$

For \tilde{F} we have

$$[23] \qquad R_{\tilde{F}}^2 = \frac{\mathbb{V}_{x \in \tilde{F}}(\hat{Y}(x))}{\mathbb{V}_{x \in \tilde{F}}(Y(x))} = \frac{\int_{\tilde{F}} (\hat{Y}(x) - \bar{Y}_{\tilde{F}})^2 dx}{\int_{\tilde{F}} (Y(x) - \bar{Y}_{\tilde{F}})^2 dx}$$

where $\bar{Y}_{\tilde{F}} = \frac{1}{\lambda(\tilde{F})} \int_{\tilde{F}} Y(x) dx = p_F \bar{Y}_F$ with $p_F = \frac{\lambda(F)}{\lambda(\tilde{F})}$.

As the mean residual is zero over F , we have $\bar{\hat{Y}}_F = \bar{Y}_F$ and by [4] we then get

$$\begin{aligned}
\mathbb{V}_{x \in \tilde{F}}(Y(x)) &= (1 - p_F) p_F \bar{Y}_F^2 + p_F \mathbb{V}_{x \in F}(Y(x)) \\
[24] \qquad \mathbb{V}_{x \in \tilde{F}}(\hat{Y}(x)) &= (1 - p_F) p_F \bar{Y}_F^2 + p_F \mathbb{V}_{x \in F}(\hat{Y}(x))
\end{aligned}$$

and consequently the following relationship between the two coefficients of determination

$$[25] \qquad R_F^2 = \frac{(1 - p_F) \bar{Y}_F^2 + R_F^2 \mathbb{V}_{x \in F}(Y(x))}{(1 - p_F) \bar{Y}_F^2 + \mathbb{V}_{x \in F}(Y(x))} \geq R_F^2$$

with equality only if $p_F = 1$. Hence the coefficient of determination or the goodness of fit is always too optimistic by considering the model over \tilde{F} instead of F .

With the coefficient of variation CV_F defined by $CV_F^2 = \frac{\mathbb{V}_{x \in F}(Y(x))}{\bar{Y}_F^2}$ we can write

$$[26] \quad R_{\tilde{F}}^2 = \frac{(1 - p_F) + R_F^2 CV_F^2}{(1 - p_F) + CV_F^2}$$

Figures 1 and 2 below display the relationship between $R_{\tilde{F}}^2$ and R_F^2 for some specific values of p_F and $CV_F = 0.5$. Clearly the $R_{\tilde{F}}^2$ can severely overestimate the goodness of fit as given by R_F^2 when the proportion of the forested surface area p_F and R_F^2 are small. The overestimation is a decreasing function of p_F . Hence, a very poor model over F can be viewed as excellent over \tilde{F} , simply because the trivial predictions outside the forest will be the most frequent ones. It is rather unfortunate that some publications dealing with inventories using remote sensing data give $R_{\tilde{F}}^2$ instead of R_F^2 .

Using $\mathbb{V}_{x \in F} \hat{R}(x) = (1 - R_F^2) V_{x \in F} Y(x)$ we get with [24]

$$[27] \quad \mathbb{V}(\hat{Y}_{reg, \tilde{F}}) = \frac{1}{n_1} p_F (1 - p_F) \bar{Y}_F^2 + \frac{p_F \mathbb{V}_{x \in F} Y(x)}{n_2} \left(1 - \left(1 - \frac{n_2}{n_1}\right) R_F^2\right)$$

The total $T = \sum_{i=1}^N Y_i$ can be estimated with $\hat{T}_{reg, \tilde{F}} = \lambda(\tilde{F}) \hat{Y}_{reg, \tilde{F}}$ with variance

$$[28] \quad \mathbb{V}(\hat{T}_{reg, \tilde{F}}) = \lambda^2(\tilde{F}) \left(\frac{1}{n_1} p_F (1 - p_F) \bar{Y}_F^2 + \frac{p_F \mathbb{V}_{x \in F} Y(x)}{n_2} \left(1 - \left(1 - \frac{n_2}{n_1}\right) R_F^2\right) \right)$$

When $\lambda(F)$ is known we can use $\hat{T}_{F, reg} = \lambda(F) Y_{reg, F}$, with variance

$$[29] \quad \mathbb{V}(\hat{T}_{F, reg}) = \lambda^2(F) \mathbb{V}(Y_{reg, F})$$

The variance $\mathbb{V}(\hat{Y}_{reg,F})$ in [21] can be rewritten as

$$\mathbb{V}(\hat{Y}_{reg,F}) = \frac{\mathbb{V}_{x \in F}(Y(x))}{n_{2F}} \left(1 - \left(1 - \frac{n_{2F}}{n_{1F}}\right) R_F^2\right)$$

With $\lambda(F) = p_F \lambda(\tilde{F})$ and the approximations $n_{k,F} = n_k p_F$, $k = 1, 2$ the above results lead after some algebra to the relative efficiency

$$[30] \quad \rho = \frac{\mathbb{V}(\hat{T}_{\tilde{F},reg})}{\mathbb{V}(\hat{T}_{F,reg})} = 1 + \frac{\frac{n_2}{n_1}(1 - p_F)}{CV_F^2(1 - (1 - \frac{n_2}{n_1})R_F^2)}$$

One has obviously $\rho = 1$ for $p_F = 1$ and $\lim_{n_1 \rightarrow \infty} \rho = 1$. Also, for $n_2 = n_1$ we get [10] again from one-phase sampling.

Figures 3 and 4 below display the coefficient ρ as a function of R_F^2 for specific values of p_F and CV_F .

The loss in efficiency can be substantial for small p_F . The efficiency loss is increasing with R_F^2 , so that improving the model is counterbalanced by using \tilde{F} to large with respect to F . Increasing the sample size n_1 of the remote sensing data improves the relative efficiency. By analogy with one-phase sampling one can estimate the unknown $\lambda(F)$ from the large sample with $\hat{\lambda}(F) = \lambda(\tilde{F}) \frac{n_{1F}}{n_1}$ to obtain the following estimate of the total

$$[31] \quad \hat{T}_{F,reg}^* = \hat{\lambda}(F) \hat{Y}_{F,reg}$$

We have implicitly assumed that the classification of a point $x \in s_1$ to s_{1F} is error-free. This is certainly true for obvious areas (above vegetation limit, lakes, large cities), probably correct most of the time with GIS external sources (Land Use Register) and sometimes not as clear-cut: a logistic model based on s_{2F} could be used to choose the most probable allocation. One can conjecture that the proportion of wrong allocations (wrongly forest and wrongly non-forest) will cancel out on average so that $\hat{\lambda}(F)$ remains

approximately unbiased.

We use the decompositions

$$\mathbb{V}(\hat{T}_{F,reg}^*) = \mathbb{V}_{n_{1F}, n_{2F}}(\mathbb{E}(\hat{T}_{F,reg}^* \mid n_{1F}, n_{2F})) + \mathbb{E}_{n_{1F}, n_{2F}}(\mathbb{V}(\hat{T}_{F,reg}^* \mid n_{1F}, n_{2F}))$$

$$\mathbb{E}_{n_{1F}, n_{2F}}(\mathbb{V}(\hat{T}_{F,reg}^* \mid n_{1F}, n_{2F})) = \mathbb{E}_{n_{1F}}\left(\mathbb{E}_{n_{2F}}(\mathbb{V}(\hat{T}_{F,reg}^* \mid n_{1F}, n_{2F}))\right)$$

The first term gives

$$\lambda^2(\tilde{F}) \frac{1}{n_1^2} \mathbb{V}(n_{1F}) \bar{Y}_F^2 = \lambda^2(\tilde{F}) \frac{1}{n_1} p_F (1 - p_F) \bar{Y}_F^2$$

and for the second term we use the fact that $\mathbb{E}(n_{2F} \mid n_{1F}) = \frac{n_2}{n_1} n_{1F}$ (see Appendix) and the approximations $\mathbb{E}(\frac{1}{n_{2F}} \mid n_{1F}) \approx \frac{1}{\mathbb{E}(n_{2F} \mid n_{1F})}$ to get

$$[32] \quad \mathbb{V}(\hat{T}_{F,reg}^*) \approx \lambda^2(\tilde{F}) \left(\frac{1}{n_1} p_F (1 - p_F) \bar{Y}_F^2 + \frac{1}{n_2} p_F \mathbb{V}_{x \in F}(Y(x)) (1 - (1 - \frac{n_2}{n_1})) R_F^2 \right)$$

comparing with [28] we see that we have the asymptotic equivalence

$$[33] \quad \mathbb{V}(\hat{T}_{F,reg}^*) \approx \mathbb{V}(\hat{T}_{\tilde{F},reg})$$

As in one-phase sampling, **estimating $\lambda(F)$ with the inventory data**, in this case from the first phase, **does not improve the estimation of the total**. If $n_1 \rightarrow \infty$ then $\hat{\lambda}(F) = \lambda(F)$, i.e. we know exactly the surface of the forested area, then, as we have seen, $\rho = 1$.

The **same result holds for small area estimation** for any domain $\tilde{G} \subset \tilde{F}$ and with $G = \tilde{G} \cap F$, provided one works with the model $\mathbf{Z}(x)$ which is $\mathbf{Z}(x)$ extended with the indicator variable $I_{\tilde{G}}$ to ensure zero-mean residual over F and G (see Mandallaz (2012) and Mandallaz (2013a)). One obtains (with the further approximation $\frac{n_{2,G}}{n_{1,G}} \approx \frac{n_2}{n_1}$) for the relative efficiency

$$[34] \quad \rho_G = \frac{\mathbb{V}(\hat{T}_{\tilde{G},reg})}{\mathbb{V}(\hat{T}_{G,reg})} = 1 + \frac{\frac{n_2}{n_1}(1 - p_G)}{CV_G^2(1 - (1 - \frac{n_2}{n_1})R_G^2)}$$

where $p_G = \frac{\lambda(G)}{\lambda(\tilde{G})}$, $CV_G = \frac{\mathbb{V}_{x \in G}(Y(x))}{\bar{Y}_G^2}$ and $R_G^2 = \frac{\mathbb{V}_{x \in G}(\hat{Y}(x))}{\mathbb{V}_{x \in G}(Y(x))}$.

For completeness and illustration we briefly consider the case of post-stratification. We set $\tilde{F} = F_0 \cup_{j=1}^L F_j := F_0 \cup F$, F_0 is the stratum ‘non-forest’ and the forest F itself is partitioned in L strata, with true means \bar{Y}_j and within strata variance V_j ($j = 1, 2, \dots, L$). Note that $\bar{Y}_0 = 0$ and $V_0 = 0$ by definition. We follow the description given in section 5.2.1 of Mandallaz (2008). The strata means are estimated by $\hat{\bar{Y}}_j = \frac{1}{n_{2,j}} \sum_{x \in s_{2,j}} Y(x)$, where $s_{2,j} = s_2 \cap F_j$ and $n_{l,j}$ is the number of points of s_l falling into F_j , $l = 1, 2$. Again, $\hat{\bar{Y}}_0 = 0$. The post-stratified estimate with respect to \tilde{F} is

$$[35] \quad \hat{Y}_{post, \tilde{F}} = \sum_{j=0}^L \hat{p}_j \hat{\bar{Y}}_j$$

where $\hat{p}_j = \frac{n_{1,j}}{n_1}$ is the estimate of the proportion $\tilde{p}_j = \frac{\lambda(F_j)}{\lambda(\tilde{F})}$. The asymptotic variance is given by

$$[36] \quad \mathbb{V}(\hat{Y}_{post, \tilde{F}}) = \frac{1}{n_1} \sum_{j=0}^L \tilde{p}_j (\bar{Y}_j - \bar{Y}_{\tilde{F}})^2 + \frac{1}{n_2} \sum_{j=0}^L \tilde{p}_j V_j$$

The post-stratified estimate with respect to F is

$$[37] \quad \hat{Y}_{post, F} = \sum_{j=1}^L \hat{p}_j \hat{\bar{Y}}_j$$

where $\hat{p}_j = \frac{n_{1,j}}{n_{1,F}}$ is the estimate of the proportion $p_j = \frac{\lambda(F_j)}{\lambda(F)}$. The asymptotic conditional variance is given by

$$[38] \quad \mathbb{V}(\hat{Y}_{post,F}) = \frac{1}{n_{1,F}} \sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2 + \frac{1}{n_{2,F}} \sum_{j=1}^L p_j V_j$$

Setting $1 - p_F = p_0 = \frac{\lambda(\tilde{F} \setminus F)}{\lambda(\tilde{F})}$ and noting that $\bar{Y}_{\tilde{F}} = p_F \bar{Y}_F$, $\tilde{p}_j = p_F p_j$ one obtains with the approximations $n_{l,F} \approx n_l p_F$, $l = 1, 2$ and some tedious but simple algebra the relative efficiency

$$[39] \quad \rho = 1 + \frac{\frac{n_2}{n_1} (1 - p_F) \bar{Y}_F^2}{\sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2 + \sum_{j=1}^L p_j V_j}$$

Since we also have

$$\mathbb{V}_{x \in F}(Y(x)) = \sum_{j=1}^L p_j V_j + \sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2$$

$$\mathbb{V}(\hat{Y}(x)) = \sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2$$

we see that in this case

$$R_F^2 = \frac{\sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2}{\sum_{j=1}^L p_j V_j + \sum_{j=1}^L p_j (\bar{Y}_j - \bar{Y}_F)^2}$$

It can be checked that [39] is indeed equivalent to [30] in this special case.

In any case, the efficiency increases as the non-forest stratum gets smaller.

In practice one uses systematic grids and, as in one-phase sampling, it is possible to get (see Mandallaz (2008) pp. 92-93) directly an unbiased estimate of the total by considering

$$[40] \quad \hat{T}_{syst,reg} = \lambda(c_1) \sum_k \hat{Y}(x) + \lambda(c_2) \sum_l \hat{R}(x_l)$$

where $\lambda(c_1)$ is the surface area of the fundamental cell defining the first-phase grid and $\lambda(c_2) > \lambda(c_1)$ is the surface area of the fundamental cell defining the sub-grid of the terrestrial second-phase, k and l run over all cells intersecting F . The first term in [40] can be rewritten as $\lambda(c_1)n_{1F}\frac{1}{n_{1F}}\sum_{x\in s_{1F}}\hat{Y}(x)$ and likewise the second term as $\lambda(c_2)n_{2F}\frac{1}{n_{2F}}\sum_{x\in s_{2F}}\hat{R}(x)$. Both $\lambda(c_1)n_{1F}$ and $\lambda(c_2)n_{2F}$ estimate $\lambda(F)$ so that $\hat{T}_{syst,reg}$ is closely related to $\hat{T}_{F,reg}^*$. However $\hat{T}_{syst,reg}$ has the disturbing feature that $\lambda(F)$ is estimated once from the large sample and once from the small sample, so that it is not as reliable as $\hat{\lambda}(F) = \frac{n_{1F}}{n_1}\lambda(\tilde{F})$. Again, no design-based variance estimates are available under systematic sampling and one has to treat them as random samples according to the approach described above.

A simple alternative approach would be to consider the relative error of the density, i.e. $\frac{\sqrt{\hat{V}(\hat{Y}_{F,reg})}}{\hat{Y}_{F,reg}}$, which must be the same as the relative error of the total as given by $\hat{T}_{syst,reg}$.

In this purely design-based approach we have seen that one should attempt to filter out most of the non-forested areas in order to increase the efficiency and to assess correctly the quality of the models. In the model-dependent approach, such as e.g. geostatistical Kriging techniques, it is obvious that model building and estimation should take place only within the forested area F , even if an approximate delineation of polygons defining F suffices (see Mandallaz (2008), chapter 7).

4 Three-phase sampling

The **null phase** draws a very large sample s_0 of n_0 points $x_i \in s_0$ ($i = 1, 2, \dots, n_0$) that are independently and uniformly distributed within the forest area F . At each of those points auxiliary information is collected, very often coding information of qualitative nature (e.g. following the interpretation of aerial photographs) or quantitative (e.g. timber volume estimates based on LiDAR measurements). We shall assume that the auxiliary

information at point x is described by the column vector $\mathbf{Z}^{(1)}(x) \in \mathbb{R}^p$. We shall call this component **pseudo-exhaustive** by analogy with the terminology used in Mandallaz (2013a), which deals with the case $n_0 = \infty$ (i.e. $\mathbf{Z}^{(1)}(x)$ is in this case the **exhaustive** component). The **first phase** draws a large sample $s_1 \subset s_0$ of $n_1 \ll n_0$ points by simple random sampling in s_0 . Note that the points $x \in s_1$ are also uniformly independently distributed in F . For each point in the first phase a further component $\mathbf{Z}^{(2)}(x) \in \mathbb{R}^q$ of the auxiliary information is available and hence also the vector $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)t}(x), \mathbf{Z}^{(2)t}(x)) \in \mathbb{R}^{p+q}$ (the upper index t denotes the transposition operator). The **second phase** draws a small sample $s_2 \subset s_1$ of n_2 points from s_1 by simple random sampling and consists of the terrestrial inventory.

We shall work with the following linear models (see Mandallaz (2013c) for more details)

1. The large model M

$$[41] \quad Y(x) = \mathbf{Z}^t(x)\boldsymbol{\beta} + R(x) = \mathbf{Z}^{(1)t}(x)\boldsymbol{\beta}^{(1)} + \mathbf{Z}^{(2)t}(x)\boldsymbol{\beta}^{(2)} + R(x)$$

with $\boldsymbol{\beta}^t = (\boldsymbol{\beta}^{(1)t}, \boldsymbol{\beta}^{(2)t})$ and the predictions $\hat{Y}(x) = \mathbf{Z}^t(x)\boldsymbol{\beta}$.

The intercept term is contained in $\mathbf{Z}^{(1)}(x)$ or it is a linear combination of its components.

The theoretical regression parameter $\boldsymbol{\beta}$ minimizes $\int_F (Y(x) - \mathbf{Z}^t(x)\boldsymbol{\beta})^2 dx$, it satisfies the normal equation $(\int_F \mathbf{Z}(x)\mathbf{Z}^t(x)dx)\boldsymbol{\beta} = \int_F Y(x)\mathbf{Z}(x)dx$ and the orthogonality relationship $\int_F R(x)\mathbf{Z}(x)dx = \mathbf{0}$, in particular the zero mean residual property $\frac{1}{\lambda(F)} \int_F R(x)dx = 0$.

2. The reduced model M_1

$$[42] \quad Y(x) = \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha} + R_1(x) =$$

The theoretical regression parameter $\boldsymbol{\alpha}$ minimizes $\int_F (Y(x) - \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha})^2 dx$. It satisfies the normal equation $(\int_F \mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)t}(x)dx)\boldsymbol{\alpha} = \int_F Y(x)\mathbf{Z}^{(1)}(x)dx$ and the orthogonality relationship $\int_F R_1(x)\mathbf{Z}^{(1)}(x)dx = \mathbf{0}$, in particular the zero mean residual property $\frac{1}{\lambda(F)} \int_F R_1(x)dx = 0$.

We consider the following design-based least squares estimators of the regression coefficients of the reduced model and large models, which are solutions of sample copies of the normal equations

$$[43] \quad \hat{\boldsymbol{\alpha}}_2 = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)t}(x) \right)^{-1} \frac{1}{n_2} \sum_{x \in s_2} Y(x)\mathbf{Z}^{(1)}(x)$$

Likewise for the large large model we set

$$[44] \quad \hat{\boldsymbol{\beta}}_2 = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x)\mathbf{Z}^t(x) \right)^{-1} \frac{1}{n_2} \sum_{x \in s_2} Y(x)\mathbf{Z}(x)$$

The generalized regression estimators in three-phase sampling are defined by

$$\begin{aligned} \hat{Y}_{\tilde{F},g3reg} &= \frac{1}{n_0} \sum_{x \in s_0} \hat{Y}_1(x) + \frac{1}{n_1} \sum_{x \in s_1} (\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_2} \sum_{x \in s_2} (Y(x) - \hat{Y}(x)) \\ &= (\hat{\mathbf{Z}}_{\tilde{F},0}^{(1)} - \hat{\mathbf{Z}}_{\tilde{F},1}^{(1)})^t \hat{\boldsymbol{\alpha}}_2 + \hat{\mathbf{Z}}_{\tilde{F},1}^t \hat{\boldsymbol{\beta}}_2 \\ \hat{Y}_{F,g3reg} &= \frac{1}{n_{0,F}} \sum_{x \in s_{0,F}} \hat{Y}_1(x) + \frac{1}{n_{1,F}} \sum_{x \in s_{1,F}} (\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_{2,F}} \sum_{x \in s_{2,F}} (Y(x) - \hat{Y}(x)) \\ [45] \quad &= (\hat{\mathbf{Z}}_{F,0}^{(1)} - \hat{\mathbf{Z}}_{F,1}^{(1)})^t \hat{\boldsymbol{\alpha}}_2 + \hat{\mathbf{Z}}_{F,1}^t \hat{\boldsymbol{\beta}}_2 \end{aligned}$$

Under the external model assumption it follows that $\hat{Y}_{\tilde{F},g3reg}$ is design unbiased with design-based variance

$$[46] \quad \mathbb{V}_{0,1,2}(\hat{Y}_{\tilde{F},g3reg}) = \frac{1}{n_0} \mathbb{V}_{x \in \tilde{F}}(Y(x)) + \left(1 - \frac{n_1}{n_0}\right) \frac{1}{n_1} \mathbb{V}_{x \in \tilde{F}}(R_1(x)) + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \mathbb{V}_{x \in \tilde{F}}(R(x))$$

Likewise one gets for the conditional variance

[47]

$$\mathbb{V}_{0,1,2}(\hat{Y}_{F,g3reg}) = \frac{1}{n_{0,F}} \mathbb{V}_{x \in F}(Y(x)) + (1 - \frac{n_{1,F}}{n_{0,F}}) \frac{1}{n_{1,F}} \mathbb{V}_{x \in F}(R_1(x)) + (1 - \frac{n_{2,F}}{n_{1,F}}) \frac{1}{n_{2,F}} \mathbb{V}_{x \in F}(R(x))$$

As in two-phase sampling one can consider the following estimates of total $\hat{T}_{\tilde{F},g3reg} = \lambda(\tilde{F})\hat{Y}_{\tilde{F},g3reg}$ and $\hat{T}_{F,g3reg} = \lambda(F)\hat{Y}_{F,g3reg}$.

Using mutatis mutandis all the previous results, in particular [4], one obtains after tedious but elementary algebra the following expression for the relative efficiency

$$[48] \quad \rho = \frac{\mathbb{V}(\hat{T}_{\tilde{F},g3reg})}{\mathbb{V}(\hat{T}_{F,g3reg})} = 1 + \frac{\frac{n_2}{n_0}(1 - p_F)}{CV_F^2 \left(\frac{n_2}{n_0} + \frac{n_2}{n_1} (1 - \frac{n_1}{n_0}) (1 - R_{F,1}^2) + (1 - \frac{n_2}{n_1}) (1 - R_F^2) \right)}$$

with the coefficients of determination over F $R_{F,1}^2 = \frac{\mathbb{V}_{x \in F}(\hat{Y}_1(x))}{\mathbb{V}_{x \in F}(Y(x))}$ and $R_F^2 = \frac{\mathbb{V}_{x \in F}(\hat{Y}(x))}{\mathbb{V}_{x \in F}(Y(x))}$ for the small and large models respectively.

It is very instructive to consider the following special case in which the almost ‘exhaustive’ component is the indicator variable of the forested area, that is $\mathbf{Z}^{(1)}(x) = I_F(x)$ (i.e. the intercept term within F) and the second component is $\mathbf{Z}^{(2)}(x)$, so that $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)}(x), \mathbf{Z}^{(2)}(x))$.

The reduced model yields $\hat{\alpha} = \hat{Y}_F = \frac{1}{n_{2F}} \sum_{x \in s_{2,F}} Y(x)$ and the full model yields $\hat{\beta}$. In this case one obtains the very intuitive result

$$[49] \quad \hat{Y}_{\tilde{F},g3reg} = (\hat{p}_{0,F} - \hat{p}_{1,F})\hat{Y}_F + \hat{p}_{1,F}\hat{Y}_{F,reg}$$

where we have set $\hat{p}_{0,F} = \frac{n_{0,F}}{n_0}$, $\hat{p}_{1,F} = \frac{n_{1,F}}{n_1}$, $\hat{Y}_{F,reg} = \hat{\mathbf{Z}}_{1,F}^t \hat{\beta}$. In other words, if one has a better delineation of the forested area, based on a large sample, than one can correct the classical two-phase regression estimator. Furthermore, in this case $R_{F,1}^2 = 0$ (the predictions are constant over F !) and the relative efficiency is found to be

$$[50] \quad \rho = \frac{\mathbb{V}(\hat{T}_{\tilde{F},g3reg})}{\mathbb{V}(\hat{T}_{F,g3reg})} = 1 + \frac{\frac{n_2}{n_0}(1 - p_F)}{CV_{\tilde{F}}^2(1 - (1 - \frac{n_2}{n_1})R_F^2)}$$

Hence, increasing p_F and n_0 will improve the relative efficiency ρ , which is closer to 1 than under two-phase sampling.

As in two-phase sampling one can estimate $\lambda(F)$ by $\hat{\lambda}(F) = \lambda(\tilde{F})\frac{n_{0,F}}{n_0}$ and consider the following estimate of total

$$[51] \quad \hat{T}_{F,g3reg}^* = \hat{\lambda}(F)\hat{Y}_{F,g3reg}$$

With the same arguments as in two-phase sampling we obtain after tedious but simple calculations the asymptotic equivalence

$$[52] \quad \mathbb{V}(\hat{T}_{F,g3reg}^*) \approx \mathbb{V}(\hat{T}_{\tilde{F},g3reg})$$

Hence, estimating the totals by considering $Y(x)$ over \tilde{F} and setting $Y(x) = 0$ for $x \notin F$ (likewise for the auxiliary information) amounts essentially to estimate \bar{Y}_F and multiply it by the estimate of $\lambda(F)$ obtained from the largest available sample, i.e. $\hat{\lambda}(F)_k = \lambda(\tilde{F})\frac{n_{k,F}}{n_k}$ for $k = 0, 1, 2$.

It is clear that doing so ignores prior information about the obviously non-forested areas, which gives a very intuitive understanding of the previous results concerning the efficiency loss of this procedure.

5 Numerical example based on the Swiss National Inventory

In the third (2004-2006) Swiss National Inventory one has $p_F \approx 0.3$ and R_F^2 lies between 0.3 and 0.4 depending on the models, $\frac{n_2}{n_1} = \frac{1}{8}$ and the coefficient of variation is $CV_F \approx 0.80$. In this example one gets e.g. $R_{\tilde{F}}^2 \approx 0.71$ instead of $R_F^2 = 0.4$, which is really misleading. For the relative efficiency given by [30] one obtains $\rho \approx 1.21$ for two-phase sampling with $p_F = 0.3, R_F^2 = 0.4$. One can improve the efficiency of SNI by filtering out the obvious non-forested areas (land above the vegetation limit 2200m, lakes, glaciers, urban areas, rocks, etc.). By constructing a polygon F^* containing F but much smaller than Switzerland (\tilde{F}) one can easily achieve $p_F = 0.5$ instead of 0.3 and $\rho = 1.12$ instead of 1.21. Note that improving the model with, say, $R_F^2 = 0.6$ one has $\rho = 1.30$ for $p_F = 0.3$ and $\rho = 1.15$ for $p_F = 0.5$. The relative efficiency is worse but the resulting variance would be smaller. In other words, improving the prediction models is not really meaningful without increasing p_F , which is obvious by looking at Figures 3 and 4.

6 Generalization

The previous results will also hold for cluster sampling (which is widely used in national inventories) and for two-stage sampling at the plot level, for small-area estimation as developed in Mandallaz et al. (2013); Mandallaz (2014) (for proofs and details see Mandallaz (2013c,b)).

7 Conclusions

The present note shows clearly that one must assess the quality of the models over the forested area F only. Also, in order to reduce the loss of efficiency, one should filter

out the obvious non-forested areas and attempt to delineate polygons coming as close as possible to the boundary of F under practical constraints, and if possible also increase the sample size of the first phase collecting remote sensing auxiliary information or use a three-phase sampling scheme to get a better delineation of the forested area F . It is also worth mentioning that improving the quality of predictions models based on remote sensing, which is challenging, is counterbalanced, in term of efficiency, by not filtering out the obvious non forested areas, which is comparison much easier to do.

References

- Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London.
- Feller, W. (1970). *An introduction to probability theory and its applications, Volume I*. Wiley, New York.
- Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Chapman and Hall, Boca Raton FL.
- Mandallaz, D. (2012). Design-based properties of small-area estimators in forest inventory with two phase sampling. Technical report, ETH Zurich, Department of Environmental Systems Science, <http://e-collection.library.ethz.ch>.
- Mandallaz, D. (2013a). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Can. J. For. Res.*, **43**:pp. 441–449.
- Mandallaz, D. (2013b). 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, <http://e-collection.library.ethz.ch>.
- Mandallaz, D. (2013c). Regression estimators in forest inventories with two-phase sampling and partially exhaustive information with application to small-area estimation. Technical report, ETH Zurich, Department of Environmental Systems Science, <http://e-collection.library.ethz.ch>.
- Mandallaz, D. (2014). A three-phase sampling extension of the generalized regression estimator with partially exhaustive information. *Can. J. For. Res.*, **44**:pp. 383–388.
- Mandallaz, D., Breschan, J., and Hill, A. (2013). New regression estimators in forest inventory with two-phase sampling and partially exhaustive information: a design-

based Monte Carlo approach with applications to small-area estimation. *Can. J. For. Res.*, **43**:pp. 1023–1031.

A Inheritance of uniform distribution

We first show that given any two points x and y independently uniformly distributed in \tilde{F} then, given that they are in F , they are independently uniformly distributed in F . Consider two arbitrary sets A and B in the plane, then

$$\mathbb{P}(x \in A, y \in B \mid x \in F, y \in F) = \frac{\mathbb{P}(x \in A \cap F, y \in B \cap F)}{\mathbb{P}(x \in F, y \in F)}$$

which is equal to

$$\frac{\mathbb{P}(x \in A \cap F) \mathbb{P}(y \in B \cap F)}{\mathbb{P}(x \in A) \mathbb{P}(y \in B)}$$

and to

$$\frac{\frac{\lambda(A \cap F)}{\lambda(\tilde{F})} \frac{\lambda(B \cap F)}{\lambda(\tilde{F})}}{\frac{\lambda(F)}{\lambda(\tilde{F})} \frac{\lambda(F)}{\lambda(\tilde{F})}} = \mathbb{P}(x \in A \mid x \in F) \mathbb{P}(y \in B \mid x \in F)$$

It is clear that the proof is valid for $n_1 > 2$ points. In other words, if points are independently uniformly distributed in \tilde{F} they inherit this property if they are actually in $F \subset \tilde{F}$.

B Inheritance of simple random sampling

We have to show that if the n_2 points in s_2 are selected from the n_1 points in s_1 by simple random sampling without replacement (irrespective of their locations) then the same holds for s_{2F} from s_{1F} for given n_{1F}, n_{2F} .

There are

$$k(n_{2F} \mid n_1, n_2, n_{1F}) = \binom{n_{1F}}{n_{2F}} \binom{n_1 - n_{1F}}{n_2 - n_{2F}}$$

possible samples with this given configuration and its probability is given by the hypergeometric distribution

$$\mathbb{P}(n_{2F} \mid n_1, n_2, n_{1F}) = \frac{k(n_{2F} \mid n_1, n_2, n_{1F})}{\binom{n_1}{n_2}}$$

with expected value $\mathbb{E}(n_{2F} \mid n_1, n_2, n_{1F}) = \frac{n_2}{n_1} n_{1F}$ (Feller (1970) pp 43-47)). If we now require that one given element in F is selected in s_2 we are left with only

$$\binom{n_{1F}-1}{n_{2F}-1} \binom{n_1-1-(n_{1F}-1)}{n_2-1-(n_{2F}-1)} = \binom{n_{1F}-1}{n_{2F}-1} \binom{n_1-n_{1F}}{n_2-n_{2F}}$$

possible configurations. Hence, the inclusion probability that this given element is selected is the ratio of these two numbers, which is easily seen to be equal to $\frac{n_{2F}}{n_{1F}}$. For two elements we have to subtract -2 instead of -1 in the above expressions and we obtain the pairwise conditional inclusion probabilities $\frac{n_{2F}(n_{2F}-1)}{n_{1F}(n_{1F}-1)}$, i.e. simple random sampling without replacement.

Figure 1: $R_F^2 = Y$ v. $R_F^2 = X$ for $CV_F = 0.8, p_F = 0.3, 0.4, 0.5, 0.8, 1.0$

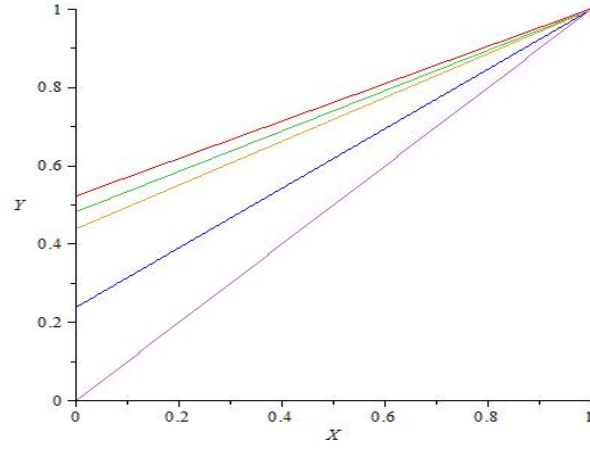


Figure 2: $R_F^2 = Y$ v. $R_F^2 = X$ for $CV_F = 0.5, p_F = 0.3, 0.4, 0.5, 0.8, 1.0$

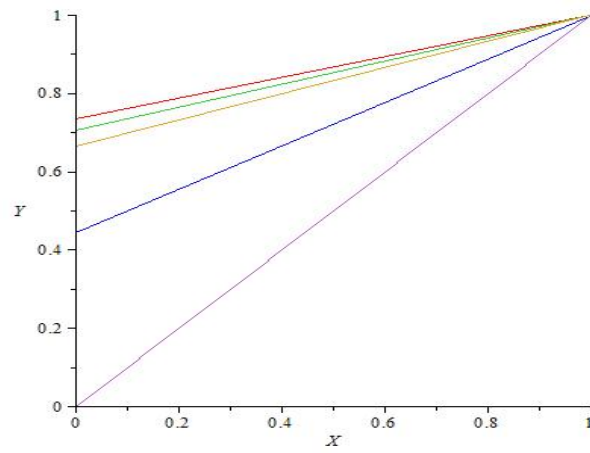


Figure 3: $\rho = Y$ v. $R_F^2 = X$ for $CV_F = 0.8, p_F = 0.3, 0.4, 0.5, 0.8, 1.0$

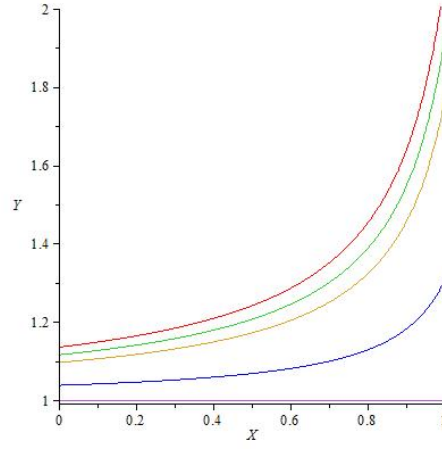


Figure 4: $\rho = Y$ v. $R_F^2 = X$ $p_F = 0.3, 0.4, 0.5, 0.8, 1.0$ and $CV_F = 0.50$

