Expert opinion on the documentation of the parameter estimation for the second cycle of the Czech National Forest Inventory

ACNIL UHUL/2476/2016/KM (Version of June 26, 2016)

Adrian Lanz, adrian.lanz@wsl.ch

12 October 2016

Adrian Lanz is forestry engineer and statistician. He is working for the Swiss national forest inventory program, a joint project of the Swiss Federal Institute for Forest, Snow and Landscape Research WSL and the Swiss Federal Agency for the Environment FOEN. He is consulting the project in statistics since many years, specifically in the related protocols for field data collection and the estimation procedures. Adrian Lanz has also been co-leading research and knowledge exchange projects which have been initiated by the European national forest inventory network ENFIN. From these activities, he disposes of a profound first-hand expertise about national forest inventory techniques applied in Europe.

Contents

1 Overview

I read the document on parameter estimation in the second cycle of the Czech national forest inventory with great pleasure, and I hold the work in high regard. The methods of making design-based inference from the sample of terrestrial data are depicted in a not only comprehensive but also very detailed and yet still focused way. The foundations of probability sampling in the context of a forest inventory with systematically distributed sample points are systematically presented, with a logical sectioning of the contents into a main part and a series of appendices of in-depth material.

The document provides a comprehensive outline and justification of the estimation procedures used in the inventory, all of highest relevance for public statistics. The design-based inference approach, which is the topic of the report and which is based on a randomised selection of the sample, is *per se* objective, so that the probabilistic conclusions about state and change of the population are real and generally accepted.

The foundations of probability based inference in forest inventory are not as clear-cut as one might expect. Although forest inventory operations have a very long tradition, it is not much more than some 50 years ago when the ground-breaking work on generalised estimation principles have been published, the estimators known today as Horvitz-Thompson estimators [Horvitz and Thompson, 1952]. In forest inventory, various topics, such as the angle count (relascope) sampling of trees, the correction for plots near the forest boundary or the inference from plots located on a regular grid, have been intensively discussed within forest inventory specialists and statisticians. Model-based methods from other fields, such as the theory of randomised variables (geo-statistics, kriging) developed and applied in mining, have been applied in forest inventory, and specific ad-hoc solutions, such as Matérn's method based on spatial differences [Matérn, 1986], have been developed in the context of forest inventory.

But it was not until some 25 years ago, when the foundations for design-based inference for forest inventory have been fully described. The foundations are known under the name *infinite population approach to forest*

inventory and the basic work has been published independently and approximately at the same time by Mandallaz [1991] in Europe and Eriksson [1995] in the USA. The same theory is today also known as Monte-Carlo approach, e.g. in the textbook of Gregoire and Valentine [2008], which belongs, together with the monograph of Mandallaz [2008], to the most relevant and in-depth textbooks on forest inventory. While the infinite population approach to forest inventory has been developed for an uniform distribution of sample points over the universe, which is mostly relevant in forest inventory practise, [Cordy, 1993] generalised the Horvitz-Thompson estimators to the infinite universe. In the document, the author depicts these fundamentals to establish and justify the parameter estimation method for the Czech national forest inventory in a very deepened and comprehensive analysis of the up-to-date knowledge on the subject. The report as such a valuable document on the foundations of inference in any forest inventory.

A very valuable part of the document covers numerical case studies, which provide insight into the comportment of certain estimators under the sampling design and populations of interest in the Czech national forest inventory. Such studies are relevant and important as they provide the confidence and insight into the published statistics, which is expected by the user.

Over all, I like to express my highest esteem to the enormous work done by the Czech colleagues and to the responsible authorities which rendered this essential engagement in the core subject of any national forest inventory possible. There is no doubt that the proposed estimators are best adapted to the conditions of the national forest inventory, and that the report fully documents the justification for the chosen estimators.

In the following I will discuss and evaluate some specific aspects of the report section by section. The intention is not at all to point on flaws in the report, but just to contribute from my expertise to the representation of the topics and to optional views and solutions in some of the problems.

2 Detailed comments

2.1 Local density

The crucial trick of the infinite population approach may not be depicted strong enough, i.e. the replacement of the actual population of interest, the finite number of trees in forest, by a new population, the local density

function defined on all, the infinite number, of points in forest (or some convenient area sampling frame). Personally, I currently refer to the article of Stevens and Urquhart [2000] in this subject, but Eriksson and Mandallaz are of course also clear on this. The local density function has fixed value at each point in the frame, and the expectation mentioned in (2) exists only under random (uniform) sampling of the function $y(x)$ in D, which is not explicitly mentioned.

Some readers are not familiar or confused with the meaning if the integration over an area. The definition of Lebesgue integration or a short explanation of the meaning of the integration may be useful.

In the section about the concentric fixed sized circular plot design, one may explicitly give the nominal extrapolation factors for the trees. The sampling actually combines (adds) the sampling of one population (trees with DBH between 7 cm and 12 cm) with the sampling of an other population (trees with DBH above 12 cm). The plot configurations are different, but we can simply add the space of the two local density function into a (new) local density function for which we know that the Lebesgue integral is equal to the population parameter of interest, i.e. the total of the single stem volume over all trees on forest land. We may even use angle count sampling for one of the sub-populations. We just have to make sure that both local densities are equal to the respective population parameters of interest, if Lebesgue integrated over the entire sampling frame (and the sampling frame should be identical for both sub-populations, at least virtually).

For the two-stage sampling procedure for trees with height measurements, we use basically the same arguments. The reference to Mandallaz's and Swiss NFI's approach to second stage tree sampling is not really appropriate, as explained in the report. And actually, Swiss NFI did change the procedure, too. Interestingly are the estimators the same under both approaches, but with an of course different meaning of second stage "inclusion probabilities".

In this light, one might question the need for an additional set of trees for diameter to height relationship calibration. There may be no problem in calibrating an internal model. We are no in context of model-assisted estimation. All local densities are fixed and additive to the population parameter of interest. We just calibrate first stage volume predictions to "good" local densities, with the quality of predictions influencing the precision of estimates. In this sense is the study on the optimal BAF for second

stage tree selection remains, of course, very informative and valuable.

Subtitles would be valuable in this large subsection 2.1.

2.2 Sampling with tracts

The tract sampling procedure of the Czech national forest inventory is, indeed, a bit tricky. One requires that

$$E\langle \hat{Y}_D \rangle = \lambda_c \sum_{x \in D^+} E\langle Y_{\mathrm{tct}}(x) \rangle = \lambda_c E\langle n_{D^+} \rangle E\langle Y_{\mathrm{tct}}(x) \rangle = \lambda_{D^+} E\langle Y_{\mathrm{tct}}(x) \rangle$$

is true, and analysing the expectation of the local density of $Y_{\mathrm{tct}}(x)$ (21) on realises that

$$\begin{aligned} E\langle Y_{\mathrm{tct}}(x) \rangle = \Bigg(& \frac{P\langle x_1 \in D \,|x_1 \sim U(D^+) \rangle}{2} \\ & + \frac{P\langle x_2 \in D \,|x_1 \notin D \wedge x_1 \sim U(D^+) \rangle}{2} \\ & + \frac{P\langle x_2 \in D \,|x_1 \in D \wedge x_1 \sim U(D^+) \rangle}{2} \Bigg) \, E\langle Y(x) \rangle \end{aligned}$$

can not be assessed easily. Because of the random orientation of the vector defining the second plot centre of tracts, the expectation depends on the geometry of the domain D and the tract. But approximately,the equality $E\langle Y_{\mathrm{tct}}(x) \rangle \approx \frac{\lambda_D}{\lambda_{D^+}} \bar{Y}_D$ can be assumed to hold. The first term is known ($\frac{\lambda_D}{2\lambda_{D^+}}$), the second term is around $\frac{1}{3}\frac{\lambda_{D^+-D}}{2\lambda_{D^+}}$, and the last term around $\frac{\lambda_{D^-}}{2\lambda_{D^+}} + \frac{2}{3}\frac{\lambda_{D-D^-}}{2\lambda_{D^+}}$, where D^- denotes the sub-region within D within which a primary tract plot centre generates with certainty a secondary plot centre located within D. The measures $\frac{1}{3}$ and $\frac{2}{3}$ are a bit arbitrary and should express, here, that less (and more) than half of the secondary plot centres can be expected within D.

A good alternative in practice, and closer to the original idea of Mandallaz, would be to use only tracts with at least one plot centre within the domain D in the calculations. The estimator would be

$$\hat{Y}_D = \lambda_C \, n_D \, \hat{\tilde{Y}}_D = \lambda_C \, n_D \, \frac{\sum_{x \in D^{\oplus}} M(x_i) \, \tilde{Y}(x_i)}{\sum_{x \in D^{\oplus}} M(x_i)} = \lambda_C \, n_D \, \hat{R}_{M(x)\tilde{Y}(x),\, M(x)}$$

in which $\tilde{Y}(x)$ is the arithmetic mean of the local densities of plot centres located inside domain D, only. The estimator is approximately design-unbiased with expectation

$$E\langle\hat{Y}_D\rangle = E\langle\lambda_C\, n_D\rangle\, E\langle\hat{R}_{M(x)\tilde{Y}(x),\, M(x)}\rangle = \lambda_D\bar{Y}_D = Y_D\ .$$

$x \in D^{\oplus}$ indicates tracts with at least on plot centre within domain D and n_D is the number of tracts with first plot centre x_1 located in domain D. Therefore, $\hat{R}_{M(x)\tilde{Y}(x),\, M(x)}$ is simply the mean of plots with centres in D and n_D is usually a bit smaller than the number of tracts used for the sums in the nominator and denominator of the ratio.

The estimator is a ratio of two sums and the appropriate variance estimators (26) can be used. The advantages of this estimator are that (a) less tracts are involved in the calculations, that (b) these clusters are easily detected in the database, that (c) there is no need to calculate the surface area of the buffer, and that, therefore, (d) the estimator can be expected to be a bit more precise (lower expected variance).

As a variance estimator, one might use the usual formula for the variance of the product two random variables (Goodmann)

$$\hat{V}\langle\hat{Y}_D\rangle = \lambda_c^2\left(\hat{Y}_D^2\,\hat{V}\langle n_D\rangle + n_D^2\,\hat{V}\langle\hat{\tilde{Y}}_D\rangle - \hat{V}\langle n_D\rangle\,\hat{V}\langle\hat{\tilde{Y}}_D\rangle\right)$$

where the last term is usually very small compared to the other terms, and therefore is often neglected. $\hat{V}\langle\hat{\tilde{Y}}_D\rangle$ can be estimated with the usual estimator for a ratio of two random variables with $M(x)\tilde{Y}(x)$ in the nominator and $M(x)$ in the denominator, and $\hat{V}\langle n_D\rangle$ is known from the simulations that have been carried out.

With this alternative estimator, the variance of the ration of two target variables would have to be estimated differently, taking the random size of the tracts into account see Mandallaz [2008].

The actual idea of the estimator proposed for the Czech national forest inventory is to ignore the tract inventory in the sense, that tracts with reference points inside D^+ are considered only. Doing this alone, the tract local density is not exactly uniformly distributed over D^+, but by compensating through the acceptance of secondary points located outside D^+ the inclusion probabilities are compensated and folded back in D^+. Therefore, the estimators are correct, under the usual assumptions.

2.3 Estimation of a ratio

In section 4, the special the estimation of ratios is introduced. The specific reference to attribute domains is correct, but may not be needed. On the one hand, because other types of ratios are very often used in result tables, and on the other hand , because the estimation of a ratio is equal when the entire sample of (tract) points is used in the calculations.

Ratios, and specifically proportions and percentages, are very often required by users and are calculated and estimated at various levels. At tree level, e.g. for the percentage of growing stock in broadleaves compared to the total growing stock, at plot level, e.g. the percentage of oak stands on the total forest land, or across time points, e.g. the percentage of gross increment in rowing stock on the growing stock in the previous inventory. The calculations are mostly easily done with appropriate indicator variables at plot and tree level, but estimation and interpretation of the estimates are not always clear-cut and often need instruction for correct interpretation.

To come back to the proposed estimators in section 4: The point (ratio) estimator may be without change in the estimate extended to the entire sample of tracts in D^+, whereas there would be a small change in the estimates for the variance, if the entire sample in D^+ was used. The underestimation of the variance would be of order $\frac{n_{D^+}(n_{A^+}-1)}{n_{A^+}(n_{D^+}-1)}$, which is often small in practise. Assuming an attribute domain of size $n_{A^+} = 20$, the underestimation of the variance would be of 5 %. In the Swiss NFI, we accept this underestimation in favour of always identical sample sizes in the calculations (in fact, the estimation procedure always calculates spatial means and totals in the same run).

2.4 Appendices

All details are very valuable and very welcome in such a report. Everything is virtually correct and relevant, without typing and other mistakes.

I give only a few very detailed remarks.

- (44) Varying probability density is not introduced here, so that (44) may be directly given with $\frac{1}{\lambda_D}$ (just before the integral sign, instead of the $f(x)$).
- (54) And the (relevant) conclusion is: $\hat{V}\langle\hat{Y}_{\mathrm{URS}}\rangle = \lambda_D^2 \frac{\hat{S}^2}{n_D(n_D-1)} =$

$\lambda_D^2 \hat{V}\langle\hat{\bar{Y}}_{\mathrm{URS}}\rangle$, where $\hat{S}^2$ is the empirical variance between the local densities at point level.

- B: The explanation of the local linearisation is perfect. The only a bit confusion aspect in the illustration might be the notation. The point estimators are always as given in (56) or (64). The replacement by (57) or (69) are only needed for the derivation of the variation, so that this replacement may be denoted by e.g. $\rightarrow$ instead of $\approx$ (also in (106)). After (77), it should say, that the expectation of $\hat{\tilde{Z}}$ is zero, not the (sample) mean.

- G: The term certainty may be better replaced by e.g. validity, because certainty does not really exist in probability sampling. Confidence intervals are always probability statements.

- (115) is called the confidence level (or coverage probability) of the confidence interval at the 1-α level.

- About bias: But it should be mentioned that the bias ratio can still be very high in a large scale inventory, because the sample size is high! Say the stem volume functions underestimate the volume by 3 % (which may not be unrealistic), then, the bias ratio may have a value of 3, assuming a standard error of 3 %.

3 Conclusions

The report comprehensively describes the estimation procedures for the second cycle of the Czech national forest inventory. The mathematical foundation of the estimators is logically outlined and the relevant literature is mentioned and cited. The estimators are genuine and innovative, and perfectly adapted to targets. The two-stage estimation of growing stock (and biomass) is a valuable security procedure against bias in the estimation of these core variables, and the additional assessment and use of the sample size variance through simulations guarantees for both, precise and geographically additive estimates. Thus, the estimators are correct and efficient, as well as based on the most up-to-date probability sampling principles and algorithms so that the estimates of the Czech national forest inventory can be considered objective and trustable without reservation.

References

C. Cordy. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Probability and Statistics Letters*, 18:353–362, 1993.

M. Eriksson. Design-based approaches to horizontal-point-sampling. *Forest Science*, 41(4):890–907, 1995.

T. G. Gregoire and H. T. Valentine. *Sampling Strategies for natuaral Resources and the Environment*. Number 1 in Applied environmental statistics. Chapman & Hall/CRC, Boca Raton, FL, U.S.A., 2008.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

D. Mandallaz. A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. Technical report, Eidgenössische Technische Hochschule Zürich, Professur Forsteinrichtung und Waldwachstum, Zürich, Schweiz, 1991. Doctoral Thesis no. 9378, Swiss Federal Institute of Technology, Zürich. Available at: `http://e-collection.ethbib.ethz.ch`.

D. Mandallaz. *Sampling Techniques for Forest Inventories*. Applied Environmental Statistics. Chapman & Hall/CRC, Boca Raton, FL, U.S.A., 2008. ISBN 978-1-58488-976-2.

B. Matérn. *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer Verlag, Berlin, Deutschland, 2 edition, 1986. ISBN 3-540-96365-0. First edition published in Meddelanden från Statens Skogsforskningsinstitut, Volume 49, No. 5, 1960.

D. L. Stevens, Jr. and N. S. Urquhart. Response designs and support regions in sampling continuous domains. *ENVIRONMENTRICS*, 11(1):13–41, 2000.