Forest Inventories: Multi-Phase Sampling Strategies for Estimating Forest and Non-Forest Resources Over Large Areas

Lorenzo Fattorini

Abstract The multi-phase sampling strategies adopted in large-scale forest inventories to estimate extents and totals of forest attributes for several forest categories and administrative districts are considered. The use of forest inventories in the assessment of non-forest resources is also discussed.

1 Introduction

The analysis of forest ecology and the wise management of forest resources requires accurate monitoring of forestlands at regular time intervals. In turn, the monitoring process involves sample surveys to estimate extents and standing volumes for a wide set of forest categories and districts. These surveys are usually referred to as *forest inventories*. The early forest inventories were performed toward the end of 15th century in the form of censuses of oaks in the Republic of Venice (Corona, 2000). Occasional censuses of forest resources were also performed in Tuscany till the 18th century. During the 19th century methods for inventorying forests were rapidly improved and the possibility of reducing costs by adopting sampling methods were recognized. The first forest inventory at country level, henceforth referred to as *national forest inventory* (NFI), was performed in Finland from 1922 to 1924 and from now repeated every ten years. The first Italian NFI was performed from 1982 to 1985, while the second started in 2003 and concluded at the end of 2006 (see http://www.infc.it). The present paper deals with NFIs as multi-phase sampling strategies performed at large scale, usually accomplished on the basis of an intensive first phase carried out on satellite imagery or aerial photos and subsequent phases performed by ground inspections in order to

[1] Lorenzo Fattorini, Department of Economics and Statistics, University of Siena; lorenzo.fattorini@unisi.it

combine aerial and field data. The aerial information acquired in the first phase is also considered to investigate non forest resources, such as woodlots, tree-rows and isolated trees outside the forest. Throughout the paper HT denotes the Horvitz-Thompson estimator.

2 Preliminaries and notations

Consider a delineated study area A partitioned into two land cover classes: forest and non forest. Denote by $F \subset A$ the forest portion of A and by U the population of forest trees within F. Suppose that F is partitioned into K sub-portions $F_1,\ldots,F_K$ corresponding to K forest categories (e.g oak, pine, larch, etc) or K spatial districts or combinations of the twos, in such a way that U is correspondingly partitioned into K sub-populations $U_1,\ldots,U_K$ of trees within the K sub-portions. Generally speaking, a forest inventory is a sampling strategy to estimate the total of a forest attribute Y (e.g. tree volume, biomass, basal area, etc), say

$$T_k = \sum_{j \in U_k} y_j$$

for each $k = 1,\ldots,K$, where y_j denotes the amount of Y corresponding to the j-th tree of U_k.

3 First phase

Gregoire and Valentine (2008, chapter 10) provide an excellent introductory chapter on the issue of sampling discrete objects (forest trees in the present case) scattered over a region by means of plots, lines or points. The authors emphasize that these designs may be conveniently re-formulated as spatial designs for sampling the continuous populations of points constituting the study area. In this setting, the interest parameter T_k can be expressed as an integral over the study area and the spatial design for selecting points (from which plots, lines or points are centred) may be viewed as a two-dimensional Monte Carlo integration, thus focusing on the problem of how to effectively select these points. Despite its simplicity, the completely random placement of sample points, usually referred to as the *uniform random sampling* (URS), may lead to uneven coverage of the study area, since some parts of the area may be sparsely sampled whereas others are intensively sampled. To avoid the drawback, stratified or systematic schemes can be adopted. A stratified scheme, usually referred to as the *tessellation stratified sampling* (TSS) is performed as it follows: the area A is covered by a region, say $R \supset A$, constituted by N non-overlapping regular polygons of equal size, say $R_1,\ldots,R_N$, and such that $R_i \cap A \neq \emptyset$ for all $i = 1,\ldots,N$. Then, for each polygon i, a point is randomly thrown within the polygon. Alternatively, a systematic scheme, usually referred to as the *systematic grid sampling* (SGS) can be used: in this case a point is randomly selected in one polygon (e.g. the first) and then repeated in the remaining $N-1$. If each point is visited on the ground and a plot of fixed size *a* is delineated around the point, then for each polygon *i*, a sample of trees from U_k, say S_{ik}, is obtained. If the interest attribute *Y* is recorded for all the trees of S_{ik}, the HT-like estimator of T_k from plot *i* turns out to be

$$\hat{T}_{ik} = \frac{|R|}{a} \sum_{j \in S_{ik}} y_j$$

where $|R|$ denotes the size of *R* and $a/|R|$ would be the inclusion probability of any tree in *U* if the points were randomly selected onto *R* by the URS scheme. It is well-known from Monte Carlo integration (Gregoire and Valentine, 2008, chapter 10) that the arithmetic mean of the $\hat{T}_{ik}$ s, say

$$\hat{\bar{T}}_{1k} = \frac{1}{N} \sum_{i=1}^{N} \hat{T}_{ik} \tag{1}$$

constituted the first-phase unbiased estimator for T_k under URS, STS and SGS schemes. Moreover, under STS, owing to the independence of the $\hat{T}_{ik}$ s, the variance of (1) could be conservatively estimated by

$$V_{1k}^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\hat{T}_{ik} - \hat{\bar{T}}_{1k})^2 \tag{2}$$

in the sense that $\mathrm{E}_1(V_{1k}^2) \geq \mathrm{V}_1(\hat{\bar{T}}_{1k})$ (Wolter, 1985, Theorem 2.4.1), while nothing can be said about the properties of (2) under SGS, as in this case the estimation of variance required more refined procedures (Opsomer et al, 2007; Fewster, 2011). Obviously, in this framework E_1 and V_1 denote expectation and variance with respect to the first phase of sampling, i.e. with respect to all the possible sets of *N* plots which can be placed onto *R* by means of TSS or SGS. For the variance estimation in the subsequent phases, (2) can be more suitably rewritten as

$$V_{1k}^2 = \frac{1}{N^2} \sum_{i=1}^{N} \hat{T}_{ik}^2 - \frac{2}{N^2(N-1)} \sum_{h>i=1}^{N} \hat{T}_i \hat{T}_h \tag{3}$$

It is worth noting that some edge effects might be present owing to forest trees positioned near the edge of the study region, which will have inclusion probabilities smaller than the inner trees. A long list of correction methods has been proposed in order to avoid the negative bias induced by edge effects (Gregoire and Valentine, 2008, section 7.5). Fortunately, in this framework, the TSS and SGS schemes, selecting points onto the enlarged region *R*, perform like the correction method usually referred to as the *buffer method* (Gregoire and Valentine, 2008, section 7.5.1), which entails allowing the *N* points to fall outside the boundary of *A*, but within some larger region that includes *A*. For this reason, under TSS and SGS the presence of forest trees in *A* whose inclusion zone overlaps the boundary of the enlarged region *R* is likely to be negligible. Moreover, it should be noticed that in NFIs edges coincides with the country's borderlines i.e. mountains ridges, rivers, sea in which the presence of forest trees is very unlike to occur. Thus, edge effects can be ignored throughout the paper with no detrimental effect on the bias of the estimators.

As to the theoretical properties of estimators of type (1) arising from TSS and SGS schemes, under very mild conditions both schemes display $o(N^{-1})$ variances (Barabesi, 2003; Barabesi and Marcheselli 2003; Barabesi and Franceschi, 2011) while URS provides $O(N^{-1})$ variances. Accordingly, for large *N*, tessellation gives rise to relevant gains in precision with respect to URS. Most of the NFIs adopted the SGS scheme, while TSS has been recently applied in the last Italian NFI (Fattorini et al, 2006).

4 Subsequent phases

Owing to costs and time involved, in real situations the *N* plots selected in the first phase cannot be visited, but rather only a portion of these points, selected in a second phase of sampling, is visited on the ground. Actually, the first-phase is only hypothetical and its treatment has the sole aim of constructing the estimators arising from the subsequent phases. In other words, conditional on the first phase, the set of HT-like estimates $\hat{T}_{1k},\ldots,\hat{T}_{Nk}$ constitutes an unknown population and its mean (1) is the object parameter to be estimated in a second phase of sampling. As to the second phase, the collection of the *N* points selected in the first phase, say P, is partitioned into the sub-set P_F of the points lying in the forest area *F*, and the sub-set $P-P_F$ of the remaining points lying outside. It is worth noting that the partition can be usually performed by satellite imagery of aerial photos, without no field work. Obviously, since the plots centred at the points of $P-P_F$ should lie completely or partially outside forest, no or very few forest trees are likely to be found in these plots. Hence, it is customary to assume $\hat{T}_{ik}=0$ for any $i\in P-P_F$, in such a way that the sampling effort can be completely devoted to P_F without detrimental effects on the estimation of T_k. Under the last assumption, the first-phase estimators (2) and (3) can be rewritten as

$$\hat{\hat{T}}_{1k}=\frac{1}{N}\sum_{i\in P_F}\hat{T}_{ik} \tag{4}$$

and

$$V_{1k}^2=\frac{1}{N^2}\sum_{i\in P_F}\hat{T}_{ik}^2-\frac{2}{N^2(N-1)}\sum_{h>i\in P_F}\hat{T}_{ik}\hat{T}_{hk} \tag{5}$$

respectively. It is apparent that, conditional on the population of forest points P_F (which is univocally determined by P), expression (4) constitutes an unknown finite population mean which can be estimate by a second phase of sampling. Accordingly, denote by $Q\subset P_F$ the sample of size *n* selected from P_F by means of a fixed-size scheme inducing first- and second-order inclusion probabilities π_i and π_{ih} ($h>i\in P_F$). Suppose also that $\pi_{ih}>0$ for any $h>i\in P_F$. If the $\hat{T}_{ik}$ s are recorded for each $i\in Q$, then the double-expansion estimator (Särndal et al., 1992, chapter 9)

$$\hat{\hat{T}}_{2k} = \frac{1}{N}\sum_{i\in Q}\frac{\hat{T}_{ik}}{\pi_i} \tag{6}$$

turns out to be unbiased with sampling variance

$$\begin{aligned}\mathrm{V}_{12}(\hat{\hat{T}}_{2k}) &= \mathrm{E}_1\left\{\mathrm{V}_2(\hat{\hat{T}}_{2k} \mid P)\right\} + \mathrm{V}_1\left\{\mathrm{E}_2(\hat{\hat{T}}_{2k} \mid P)\right\} \\ &= \mathrm{E}_1\left\{\frac{1}{N^2}\sum_{h>i\in P_F}(\pi_i\pi_h - \pi_{ih})\left(\frac{\hat{T}_{ik}}{\pi_i} - \frac{\hat{T}_{hk}}{\pi_h}\right)^2\right\} + \mathrm{V}_1(\hat{\hat{T}}_{1k})\end{aligned} \tag{7}$$

where now $\mathrm{E}_2(\bullet \mid P)$ and $\mathrm{V}_2(\bullet \mid P)$ denote expectation and variance with respect to the second phase of sampling, i.e. with respect to all the possible samples *Q* which can be selected by the second-phase scheme, conditional to the set of points P selected in the first phase, while E_{12} and V_{12} denote expectation and variance with respect to both the sampling phases. Moreover, it can be proven that under TSS a conservative estimator for (7) is given by

$$V_{2k}^2 = \frac{1}{N^2}\left\{\sum_{i\in Q}\frac{\hat{T}_{ik}^2}{\pi_{ik}^2} + 2\sum_{h>i\in Q}\frac{\hat{T}_i\hat{T}_h}{\pi_i\pi_h} - 2\frac{N}{N-1}\sum_{h>i\in Q}\frac{\hat{T}_i\hat{T}_h}{\pi_{ih}}\right\} \tag{8}$$

in the sense that $\mathrm{E}_{12}(V_{2k}^2) \geq \mathrm{V}_{12}(\hat{\hat{T}}_{2k})$, while estimation is more complex under SGS (Opsomer et al, 2007). Most of NFIs involve only two phase of sampling as opposite to the recent Italian NFI in which three phases are adopted. Actually, in the Italian case, the second-phase points are visited only to record the forest category in order to estimates the extents of these categories. Indeed, it can be readily proven that the two-phase estimator for the extent of F_k and its variance estimator are obtained from (6) and (8) respectively, when $\hat{T}_{ik} = 1$ if the *i*-th points falls in F_k and 0 otherwise. Total of forest attributes are instead estimated from a third-phase sample of points selected from the second-phase sample *Q*. The expressions of the third-phase estimators are obviously more cumbersome and are not reported for brevity. Details on third-phase estimators are given by Fattorini et al (2006).

5 Estimation of non-forest resources

During the FAO Expert Consultation on Global Forest Resources Assessment 2000 (Kotka - Finland 1996), the importance of trees outside forests (TOF) and the need for complete and detailed information about these stands were underlined for the first time. NFIs are currently requested to broaden their scopes to include the assessment of TOF attributes. (Kleinn, 2000, 2002). TOF include small woodlots, three rows, urban forests and isolated trees and play a basic role in biodiversity conservation and carbon sequestration. The main objective of TOF inventories is the estimation of totals and/or averages of some physical attribute of the units (e.g. size and length). Probably, an efficient solution would require the use of *ad hoc* sampling schemes for each of the target parameters. However, in order to save time and resources, it may be appealing to perform the estimation in the first-phase of NFIs, as most physical attributes can be recorded from the aerial information collected during the inventories without any field work. Let W be the population of M woodlots, or trees rows or urban forests in the study area, say $w_1,\ldots,w_M$ and denote by y_j the value of a physical attribute of woodlot j which can be recorded from aerial imagery. Suppose that the population total

$$T_W = \sum_{j\in W} y_j$$

and/or the population mean $\bar{Y}_W = T_W / M$ are the parameters to be estimated. To this purpose, denote by G the set of distinct woodlots, tree rows or urban forests which contain at least one of the N first-phase points and let m be the (random) size of G. As proven by Baffetta et al (2011b, Appendix 2), under TSS the quantity

$$\hat{T}_{1W} = \frac{|R|}{N}\sum_{j\in G}\frac{y_j}{|w_j|} \qquad (9)$$

can be viewed as an approximation of both the HT and Hansen-Hurvitz estimators of T_W and as such it turns out to be approximately unbiased. It is worth noting that (9) avoids the troublesome quantification of the portion of the selected units lying in adjacent quadrats, as would be requested by the genuine HT estimator. Moreover,

$$V_{1W}^2 = \frac{1}{N(N-1)}\left\{|R|^2\sum_{j\in G}\left(\frac{y_j}{|w_j|}\right)^2 - N\hat{T}_{1W}^2\right\}$$

is proven to be a conservative estimator for the variance of $\hat{T}_{1W}$, while $\hat{T}_{1W} \pm 1.96 V_{1W}$ provides an approximately conservative confidence interval with nominal coverage of 95%. For y_j invariably equal to 1, T_W coincides with the population abundance M and (9) provides an abundance estimator, say $\hat{M}_1$. Thus a very natural estimator of $\bar{Y}_W = T_W / M$ is given by the ratio $\hat{\bar{Y}}_{1W} = \hat{T}_{1W} / \hat{M}_1$, which is approximately unbiased with variance estimator

$$V_{1\bar{Y}}^2 = \frac{|R|^2}{\hat{M}_1^2 N(N-1)}\sum_{j\in G}\left(\frac{y_j - \hat{\bar{Y}}_{1W}}{|w_j|}\right)^2$$

The validity of these estimators are empirically checked by a simulation study (Baffetta et al, 2011b) and applied to estimate the average size and the abundance of urban forests from the sample of 430 forests selected throughout Italy in the first phase of the last Italian NFI (Corona et al, 2012). Finally, as to isolated trees, their abundance can be estimated from the aerial information acquired during NFIs, even if a further aerial sampling phase is necessary in this case. Baffetta et al (2011a) propose the use of a second-phase in which the N first-phase points are partitioned into strata by using aerial imagery. Usually the strata coincide with land cover classes easily identifiable from the imagery. Then a second-phase sample of points is selected from each strata by simple random sampling without replacement, a circle of fixed size is centred in the second-phase points and the number of isolated trees within is counted once again from the aerial imagery. As the presence of isolated trees is more likely in agricultural land (cropland and grassland), the agricultural strata should be more intensively sampled. Moreover, as isolated trees are rare and widely scattered over territories, a suitable choice should be circles of about 100-200 *m* radius which are much larger than those usually adopted when surveying within forests (10-20 *m* radius). Accordingly, if *W* now denotes the population of *M* isolated trees over the study area, if $P_1,\ldots,P_L$ denote the *L* strata in which the population *P* of the *N* first-phase points is partitioned, $N_1,\ldots,N_L$ denote the stratum sizes, $Q_1,\ldots,Q_L$ denote the samples of points selected from each stratum and $n_1,\ldots,n_L$ the sample sizes, the two-phase aerial estimator of *M* turns out to be (Baffetta et al, 2011a)

$$\hat{M}_2 = \frac{|R|}{b}\sum_{l=1}^{L} p_l \overline{m}_l \qquad (10)$$

where *b* is the size of circles, $p_l = N_l / N$, m_i denotes the number of isolated trees aerially counted within the plot *i* and $\overline{m}_l$ is the average of the m_i s for $i \in Q_l$. Estimator (10) is unbiased with variance which can be unbiasedly estimated by

$$V_{1M}^2 = \frac{|R|^2}{b^2(N-1)}\left\{\sum_{l=1}^{L} p_l (N_l - 1)\frac{s_l^2}{n_l} + \sum_{l=1}^{L} p_l (\overline{m}_l - \hat{M}_2)^2\right\}$$

where s_l^2 is the sample variance of the m_i s for $i \in S_l$. Obviously, if totals or averages of some biophysical attributes such as tree volume and biomass are of interest, subsequent sampling phases must be performed on the field. Corona and Fattorini (2006) propose the use of cluster sampling to survey tree rows, while Corona et al (2011) propose the use of sector sampling to survey woodlots. A third stratified sampling phase is suggested by Baffetta et al (2011a) for field surveys of isolated trees.

6 Conclusions

As already pointed out, NFIs usually require estimates of population totals and related quantities for various regions defined by political subdivisions, for types of forest as well as for other domains such as ownership categories and silvicultural types and for combinations of them. Practically speaking, thousands of estimates are produced as the output of a NFI. In this framework, statisticians have neither time nor resources to select *ad hoc* estimators for each survey variable and the only practical way is to adopt pure design-based approaches in which sample weights are the inverse of the inclusion probabilities determined by the sampling design. However, as pointed out by Opsomer et al (2007), there is an increasing availability of various inexpensive auxiliary data derived from remote sensing sources which represent a great opportunity to improve the accuracy of estimates. Remote sensing auxiliary information has been (partially) used at design level in the last Italian NFI to stratify the first phase points in accordance with land cover classes (Fattorini et al, 2006). Alternatively, Opsomer et al (2007) propose the use of auxiliary information at estimation level, incorporating multivariate super-population models in the framework of model-assisted estimation. Interestingly, the use of auxiliary information at estimation level makes possible the treatment of missing data, which may take rise in the field phases when some of the selected points located in difficult terrains cannot be reached by field crews.

References

1. Baffetta, F., Corona, P., Fattorini, L.: Assessing the attributes of scattered trees outside the forest by a multi-phase sampling strategy. Forestry **84**, 315-325 (2011a)
2. Baffetta, F., Fattorini, L., Corona, P.: Estimation of small woodlot and tree row attributes in large scale forest inventories. Environ. Ecol. Stat. **18**, 147-167 (2011b)
3. Barabesi, L.: A Monte Carlo integration approach to Horvitz-Thompson estimation in replicated environmental designs. Metron **61**, 355-374 (2003)
4. Barabesi L, Marcheselli M: A modified Monte-Carlo integration. Int. Math. J. **3**, 555-565 (2003)
5. Barabesi L, Franceschi S: Sampling properties of spatial total estimators under tessellation stratified designs. Environmetrics **22**, 271-278 (2011)
6. Corona, P.: Introduzione al Rilevamento Campionario delle Risorse Forestali. Edizioni CUSL, Firenze (2000)
7. Corona, P., Fattorini, L.: The assessment of tree row attributes by stratified two-stage sampling. Eur. J. Forest Res. **125**, 57-66 (2006)
8. Corona, P., Fattorini, L., Franceschi, S: Two-stage sector sampling for estimating small woodlot attributes. Can. J. For. Res. **41**, 1819-1826 (2011)
9. Corona, P., Agrimi, M., Baffetta, F., Barbati, A., Chiriacò, M.V., Fattorini, L., Pompei, E., Valentini, R., Mattioli, W.: Extending large-scale forest inventories to assess urban forests. Environ. Monit. Assess. **184**, 1409-1422 (2012)
10. Fattorini, L., Marcheselli, M., Pisani, C.: A three-phase sampling strategy for large-scale multiresource forest inventories. J. Agric. Biol. Environ. Statist. **11**, 1-21 (2006)
11. Fewster, R.M.: Variance estimation for systematic designs in spatial surveys. Biometrics **67**, 1518-1531 (2011)
12. Gregoire, T.G., Valentine, H.T.: Sampling Strategies for Natural Resources and the Environment. Chapmam & Hall, Boca Raton, FL (2008)
13. Kleinn, C.: On large area inventory and assessment of trees outside forests. Unasylva **51**, 3-10 (2000)
14. Kleinn, C.: New technologies and methodologies for national forest inventories. Unasylva **53**, 10-15 (2002)
15. Opsomer, J.D., Breidt, F.G., Moisen, G.G., Kauermann, G.: Model-assisted estimation of forest resources with generalized additive models. J. Amer. Stat. Assoc. **102**, 400-416 (2007)
16. Särndal, C.-E., Swensson, B., Wretman, J.: Model-Assisted Survey Sampling. Springer-Verlag, New York (1992)
17. Wolter, K.M.: Introduction to Variance Estimation. Springer-Verlag, New York (1985)