Design-based treatment of nonresponse in large-scale forest inventories: an application to the Italian National Inventory of Forests and Forest Carbon Sinks

Daniela Maffei

Ai miei genitori e ad Agnese.

Contents

INTRODUCTION

The objective of this thesis is to obtain reduced bias and to estimate the variance of timber volume estimates achieved in forest inventories in presence of item nonresponse due to the unrecorded values of points located in roughly or difficult terrain. In most situations, these points cannot be reached by foresters or, even reached, the recording activities cannot be performed. Since the points can be reached or not (i.e. with probability 1 or 0) there is no random mechanism generating nonresponse. Thus, nonresponse can be handled in a complete design-based framework. The thesis considers a three-phase sample strategy for forest survey (which has been the strategy adopted in the last Italian National Forest Inventories) and proposes a calibration approach to account for nonresponse.

It is structured in four chapters:

the first chapter has an introductory nature and deal with unit nonresponse and its design-based treatment by means of calibration approach. At least to our knowledge, the theoretical results achieved of the chapter are unreported

in literature. Theoretical findings are validated by a simulation study.

The second chapter describes the three-phase sampling strategy for forest surveys, neglecting nonresponse.

The third chapter deals with the application of the nonresponse treatment procedure outlined in the first chapter in the framework of the three phase sampling strategy considered in the second chapter. Properties of the resulting estimator are investigated by a simulation study.

The fourth chapter contains the application of nonresponse treatment to the Italian National Forest Inventory for the data regarding the administrative district of Trentino.

The study was funded by a PhD scolarship offered by the Agricoltural Research Council (CRA) and the Italian Ministry of Agricolture, Food and Forestry Policies. The author thanks Dr Patrizia Gasparini and Dr Flora De Natale from CRA Research Unit for Forest Monitoring and Planning that suggested to work on the topic of missing data in forest inventories and contributed to the discussion.

Chapter 1

DESIGN-BASED TREATMENT OF NONRESPONSE

1.1 Preliminaries and notations

Let U be a population of N units, let y_j be the value of a positive survey variable Y for the j-th unit and suppose that the population total, say

$$T_y = \sum_{j \in \mathsf{U}} y_j$$

be the interest quantity to be estimated on the basis of a sample $\mathsf{S} \subset \mathsf{U}$ of size n, selected from the population by means of a fixed-size scheme inducing first- and second-order inclusion probabilities π_j and π_{jh} $(h > j = 1, \ldots, N)$. Suppose also that $\pi_{jh} > 0$ for any $h > j = 1, \ldots, N$. If the y_js are recorded for each $j \in \mathsf{S}$ (complete response), then the Horvitz-Thompson (HT) estimator

$$\hat{T}_{HT} = \sum_{j \in \mathsf{S}} \frac{y_j}{\pi_j}$$

is design-unbiased with design-variance

$$\mathrm{V}(\hat{T}_{HT}) = \sum_{h>j\in \mathrm{U}} (\pi_{\mathrm{j}}\pi_h - \pi_{jh}) \left(\frac{y_j}{\pi_j} - \frac{y_h}{\pi_h}\right)^2$$

which can be unbiasedly estimated by the well-known Sen-Yates-Grundy (SYG) variance estimator

$$\hat{V}(\hat{T}_{HT}) = \sum_{h>j\in \mathrm{S}} \frac{\pi_{\mathrm{j}}\pi_h - \pi_{jh}}{\pi_{jh}} \left(\frac{y_j}{\pi_j} - \frac{y_h}{\pi_h}\right)^2$$

On the other hand, if the y_js are recorded only for $j \in \mathsf{R} \subset \mathsf{S}$ (partial response), then the R-based HT estimator

$$\hat{T}_R = \sum_{j\in \mathrm{R}} \frac{y_j}{\pi_j} \tag{1.1}$$

turns out to be invariably smaller than $\hat{T}_{HT}$ and hence negatively biased. In order to compensate the decrease of $\hat{T}_{HT}$ due to non response, the weights $1/\pi_j$ attached to each y_j should be suitably increased.

Widely applied methods to account for non response, usually referred to as traditional approach to weighting (Särndal and Lundstrom, 2005, section 6.1), view the response set R as the result of a two-phase sampling: in the first phase S is selected from U by means of a sampling scheme, while in the second phase R is realized as a subset of S assuming the existence of a response mechanism for which every unit j has its individual response probability θ_j, with $0 < \theta_j \leq 1$ for each j=1....N. Then, a realistic model is formulated in which the unknown θ_js depend on some auxiliary variables.

The θ_js subsequently are estimated on the basis of the auxiliary information available at sample level (info S) and/or at population level (info U). The estimation theory built around the idea that units j is equipped with a design-based inclusion probability π_j and an unknown model-based response probability θ_j has been termed as quasi-randomization theory by Oh and Scheuren (1983). In this framework, the double-expansion estimator

$$\hat{T}_{DEX} = \sum_{j \in R} \frac{y_j}{\pi_j \hat{\theta}_j}$$

(where θ_j denote the model-based estimate of θ_j) inflates (1.1) accounting for two-phase selection. As pointed out by Kott (1994), the procedure requires that all the θ_js be strictly positive, while this requirement is often unrealistic because most populations contained units that do not respond under any circumstances. Moreover, Särndal and Lundström (2005, p.52) point out that the knowledge about response behaviour is usually limited, so it is difficult to defend any proposed model adopted to estimate the θ_js as being more realistic than any alternative. Apart from these two (relevant) drawbacks, there are surveys in which response of unit j cannot be viewed as the outcome of a dichotomous experiment with unknown probabilities (just as a toss of an unfair coin). Indeed, there may be situations in which the response pattern is fixed, in the sense that the population is partitioned into two strata: the respondent stratum, say U_R of size N_R and the nonrespondent stratum $\mathsf{U} - \mathsf{U}_R$

of size $N - N_R$. In this case responses are fixed characteristics like the values of the interest variable and we are obliged to adopt a design-based treatment of nonresponse. The design-based treatment of nonresponse is mandatory in many environmental surveys such as forest inventories, when a population of plots scattered over the study area are sampled. In some circumstances, it may occur that some selected plots located in roughly terrains cannot be reached by the forester team and so the timber volume contained within the plots is missed. Obviously, in this framework there is no random experiment, since the plot can be reached or not. When nonresponse is a fixed characteristic, the quasi-randomization approach cannot be adopted and *imputation techniques* or *calibration approach* should be used.

1.2 Imputation and calibration

Imputation is a procedure in which missing values are replaced by substitutes. In turn, substitutes can be constructed in a wide variety of ways. More precisely, all the values y_j that are missing are imputed by guess values, say $\hat{y}_j$, in such a way that the completed sample data are given by y_j for $j \in \mathsf{R}$ and $\hat{y}_j$ for $j \in \mathsf{S} - \mathsf{R}$. Accordingly, the HT estimator on the completed data,

usually referred to as the imputed estimator, turns out to be

$$\hat{T}_{IMP} = \sum_{j \in R} \frac{y_j}{\pi_j} + \sum_{j \in S-R} \frac{\hat{y_j}}{\pi_j}$$

As pointed out by Särndal and Lundström (2005, p.52), imputed values are artificial and as such affected by errors. Accordingly, imputation errors may be treated as measurement errors, as when an erroneous value is recorded for a sampled unit. Commonly used techniques of imputation are *regression imputation*, *nearest neighbour imputation* and *hot deck imputation* (for a review see e.g. Särndal and Lundström, 2005, section 12.7). Without entering on these techniques, but just from a general point of view, it should be once again pointed out that since knowledge about response behaviour is usually limited, it is difficult to defend any proposed method of imputation as being more realistic than any alternative.

As an alternative to imputation, the *calibration approach* or, more precisely, the *calibration approach to weighting* may be adopted. In order to describe the method, details and notations about auxiliary information are needed. Denote by x_{jk} the value of an auxiliary variable X_k for the j-th unit in the population. Now, suppose that the values of K_U auxiliary variables, say X_1^*,.....,$X_{K_U}^*$, are known for each unit of the population (info U), i.e. the K_U-vectors $\mathbf{x}_j^*$=$[x_{j1}^*, ..., x_{jK_U}^*]^T$ are known for each j=1,.....,N. Accordingly,

the vector of population totals for the K_U variables, say

$$\mathbf{T}^* = \sum_{j \in U} \mathbf{x}_j^*$$

is also known. Moreover, suppose that the values of K_S auxiliary variables, say X_1°,...,$X_{K_S}^\circ$, are known for each unit in the sample (info S), i.e. the K_S-vectors $\mathbf{x}_j^\circ$=$[x_{j1}^\circ, ..., x_{jK_S}^\circ]^T$ are known for each $j \in \mathsf{S}$. In this case, the vector of the HT estimates of the population totals computed on S, say

$$\hat{\mathbf{T}}^\circ = \sum_{j \in S} \frac{\mathbf{x}_j^\circ}{\pi_j}$$

is known. Obviously $\hat{\mathbf{T}}^\circ$ constitutes an unbiased estimator of the vector of population totals of the K_S variables, say

$$\mathbf{T}^\circ = \sum_{j \in U} \mathbf{x}_j^\circ$$

Now, for simplicity of notation, the two sets of variables can be joined into a unique set of $K = K_U + K_S$ variables simply by using the K-vector $\mathbf{x}_j = \begin{bmatrix} \mathbf{x}_j^* \\ \\ \mathbf{x}_j^\circ \end{bmatrix}$ as well as the K-vector $\hat{\mathbf{T}} = \begin{bmatrix} \mathbf{T}^* \\ \mathbf{T}^\circ \end{bmatrix}$. In the parlance of Särndal and Lundström (2005, Table 6.1), the joined information is referred to as the Info US while the vector $\hat{\mathbf{T}}$ is referred to as the *information input* owing to its basic role in the subsequent calibration. It is worth noting that while the first K_U components of $\hat{\mathbf{T}}$ are known constants, the remaining K_S components are random variables depending on S. Obviously the design-based expectation of $\hat{\mathbf{T}}$ turns out to be $\mathrm{E}(\hat{\mathbf{T}}) = \mathbf{T}$, where $\mathbf{T} = \begin{bmatrix} \mathbf{T}^* \\ \mathbf{T}^\circ \end{bmatrix}$. Turn now to the calibration

approach. In order to compensate the reduction of $\hat{T}_{HT}$ due to non response, the weights $1/\pi_j$ in (1.1) are changed into new weights, say w_j, in such a way to satisfy the so called *calibration equation*

$$\sum_{j \in R} w_j \mathbf{x}_j = \hat{\mathbf{T}} \tag{1.2}$$

Once the w_js are determined from (1.2), the resulting estimator of T_y

$$\hat{T}_{CAL} = \sum_{j \in R} w_j y_j \tag{1.3}$$

is referred to as the *calibration estimator*. The rationale behind (1.3) is quite obvious: if the w_js are able to guess $\hat{\mathbf{T}}$ without errors, then they should be suitable also for estimating T_y, providing that a close relationship exists between the interest and the auxiliary variables. Even if no superiority of calibration with respect to imputation can be generally claimed, the calibration approach seems to be more convincing than imputation because even at its best, i.e. when all the imputed values $\hat{y}_j$ for $j \in \mathsf{S} - \mathsf{R}$ are guessed without errors, imputation cannot improve upon the performance of the HT estimator. On the other hand, as will be proved in the next section, if a perfect linear relationship exist between interest and auxiliary variables, the calibration approach estimates T_y without error. Accordingly, $\hat{T}_{CAL}$ is likely to perform well for suitable choices of auxiliary variables.

1.3 Mathematical properties of calibration estimator

The results presented in this section are taken from Särndal and Lundström (2005, Chapter 6). The properties derived from these results are called of mathematical nature since they just stem from the analytical structure of estimators of type (1.3) and hold whenever the scheme adopted to select the sample S from U. Rewriting the w_js as modifications of the HT weights, i.e. $w_j = v_j/\pi_j$, and from the fact that (even if it is not mandatory) the w_js should constitute enlargements of the π_js, a suitable structure for the v_js as enlargement factors (i.e. v_j>1) may be

$$v_j = 1 + \mathbf{c}^T \mathbf{x}_j, \quad j \in \mathsf{R} \tag{1.4}$$

From (1.4), the calibration equation can be rewritten as

$$\sum_{j \in \mathrm{R}} \frac{(1 + \mathbf{c}^{\mathrm{T}} \mathbf{x}_j)}{\pi_j} \mathbf{x}_j = \hat{\mathbf{T}} \tag{1.5}$$

in such a way that, solving with respect to $\mathbf{c}$, equation(1.5) is satisfied for

$$\hat{\mathbf{c}} = \left(\sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right)^{-1} (\hat{\mathbf{T}} - \hat{\mathbf{T}}_R) \tag{1.6}$$

where

$$\hat{\mathbf{T}}_R = \sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j}{\pi_j}$$

denotes the K-vector of the R-based HT estimates performed on the auxiliary variables, and providing that the matrix to be inverted is positive definite.

Finally, on the basis of (1.6), the calibration estimator of type (1.3) turns out to be

$$\hat{T}_{CAL} = \sum_{j \in \mathrm{R}} \frac{(1 + \hat{\mathbf{c}}^{\mathrm{T}} \mathbf{x}_j)}{\pi_j} y_j \tag{1.7}$$

As previously emphasized, from (1.7) it can be proven that if there is a perfect linear relationship between the interest and auxiliary variables, i.e. $y_j = \mathbf{b}^{\mathrm{T}} \mathbf{x}_j$ for all $j = 1, \ldots, N$, then $\hat{T}_{CAL} = T_y$ (e.g. Särndal and Lundström, 2005, p. 61). Moreover, after trivial algebra, the calibration weights in (1.7) can be decomposed as $w_j = w_{MAINj} + w_{REMj}$ for any $j \in \mathsf{R}$, where

$$w_{MAINj} = \frac{1}{\pi_j} \mathbf{x}_j^{\mathrm{T}} \left(\sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right)^{-1} \hat{\mathbf{T}} \tag{1.8}$$

is referred to as the *main component* of w_j while

$$w_{REMj} = \frac{1}{\pi_j} \left\{ 1 - \mathbf{x}_j^{\mathrm{T}} \left(\sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right)^{-1} \hat{\mathbf{T}}_R \right\}$$

is referred to as the *remainder* (e.g. Särndal and Lundström, 2005, section 6.8). Indeed, it can be proven that the main components alone ensure calibration, in the sense that

$$\sum_{j \in \mathrm{R}} w_{MAINj} \mathbf{x}_j = \hat{\mathbf{T}}$$

and hence

$$\sum_{j \in \mathrm{R}} w_{REMj} \mathbf{x}_j = \mathbf{0}$$

Accordingly, it is generally true that the remainder terms do not contribute to calibration. Moreover, it can also be proved that if there exists a K-vector

$\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x}_j = 1 \quad \forall\ j = 1, \ldots, N \tag{1.9}$$

than $w_{REMj} = 0$ for any $j \in \mathsf{R}$ in such a way that the calibration weights coincide with their main components. Now, if an auxiliary variable (say the first without loss of generality) is invariably equal to 1, i.e. if $\mathbf{x}_j = [1, x_{j2}, \ldots, x_{jK}]^{\mathrm{T}}$ for any $j = 1, \ldots, N$, relation (1.9) is trivially true for $\boldsymbol{\lambda} = [1, 0, \ldots, 0]^{\mathrm{T}}$. Accordingly, henceforth the auxiliary variable invariably equal to 1 will be tacitly included in the set of the K auxiliary variables in order to use the simplified version (1.8) of the calibration weights. It is worth noting that in this case the calibration weights of type (1.8) guess the population size without error, i.e.

$$\sum_{j \in \mathrm{R}} w_j = N$$

Using expression (1.8) into (1.3), the calibration estimator reduces to

$$\hat{T}_{CAL} = \hat{\mathbf{b}}_R^{\mathrm{T}} \hat{\mathbf{T}} \tag{1.10}$$

where

$$\hat{\mathbf{b}}_R = \left(\sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right)^{-1} \sum_{j \in \mathrm{R}} \frac{y_j \mathbf{x}_j}{\pi_j} \tag{1.11}$$

Expression (1.10) will constitute the definitive formulation for the calibration estimator adopted throughout the work.

1.4 Design-based bias of calibration estimator

In order to treat nonrespose as a fixed characteristic a dummy variable, say R, is considered such that $r_j = 1$ for $j \in \mathsf{U}_R$ while $r_j = 0$ for $j \in \mathsf{U} - \mathsf{U}_R$. Hence, the vector $\hat{\mathbf{b}}_R$ in (1.10) can be rewritten as

$$\hat{\mathbf{b}}_R = \left[\hat{b}_{R1} \ \ldots \ \hat{b}_{RK}\right]^{\mathrm{T}} = \left(\sum_{j\in\mathrm{S}} \frac{r_j \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}}{\pi_j}\right)^{-1} \sum_{j\in\mathrm{S}} \frac{r_j y_j \mathbf{x}_j}{\pi_j} \tag{1.12}$$

In this way, $\hat{T}_{CAL}$ depends on the selection of the sole sample S while nonresponse is accounted for in the r_js ($j \in \mathsf{S}$).

Now, denote by

$$\hat{\mathbf{A}}_R = \begin{bmatrix} \hat{a}_{R11} \ \ldots \ \hat{a}_{R1K} \\ \ldots\ldots\ldots\ldots\ldots \\ \hat{a}_{RK1} \ \ldots \ \hat{a}_{RKK} \end{bmatrix} = \sum_{j\in\mathrm{S}} \frac{r_j \mathbf{x}_j \mathbf{x}_j^T}{\pi_j}$$

and

$$\hat{\mathbf{a}}_R = [\hat{a}_{R1} \ \ldots \ \hat{a}_{RK}]^{\mathrm{T}} = \sum_{j\in\mathrm{S}} \frac{r_j y_j \mathbf{x}_j}{\pi_j}$$

the two HT-like estimators in $\hat{\mathbf{b}}_R$. Then, it is at once apparent that

$$\mathrm{E}(\hat{\mathbf{A}}_R) = \mathrm{E}\left(\sum_{j\in\mathrm{S}} \frac{r_j \mathbf{x}_j \mathbf{x}_j^T}{\pi_j}\right) = \sum_{j\in\mathrm{U}} r_j \mathbf{x}_j \mathbf{x}_j^T = \sum_{j\in\mathrm{U}_R} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} = \begin{bmatrix} a_{R11} \ \ldots \ a_{R1K} \\ \ldots\ldots\ldots\ldots\ldots \\ a_{RK1} \ \ldots \ a_{RKK} \end{bmatrix} = \mathbf{A}_R$$

and

$$\mathrm{E}(\hat{\mathbf{a}}_R) = \mathrm{E}\left(\sum_{j\in\mathrm{S}} \frac{r_j y_j \mathbf{x}_j}{\pi_j}\right) = \sum_{j\in\mathrm{U}} r_j y_j \mathbf{x}_j = \sum_{j\in\mathrm{U}_R} y_j \mathbf{x}_j = [a_{R1} \ \ldots \ a_{RK}]^{\mathrm{T}} = \mathbf{a}_R$$

Thus, keeping in mind that $\mathrm{E}(\hat{\mathbf{T}}) = \mathbf{T}$, $\hat{T}_{CAL}$ can be rewritten as a function of the three HT estimators on which it depends, say

$$\hat{T}_{CAL} = f(\hat{\mathbf{a}}_R, \hat{\mathbf{A}}_R, \hat{\mathbf{T}}) = \hat{\mathbf{a}}_R^{\mathrm{T}} \hat{\mathbf{A}}_R^{-1} \hat{\mathbf{T}} = \sum_{k=1}^{K} \sum_{l=1}^{K} \hat{a}_{Rk} \hat{T}_l \hat{a}_R^{kl} \tag{1.13}$$

where $\hat{a}_R^{kl}$ denotes the kl-element of $\hat{\mathbf{A}}_R^{-1}$ and $\hat{T}_k$ may be one component of $\hat{\mathbf{T}}$. Differentiating (1.13) with respect to all the variables involved, it follows that

$$\frac{\partial f}{\partial \hat{a}_{Rk}} = \sum_{l=1}^{K} \hat{T}_l \hat{a}_R^{kl}\,, \quad k = 1, \ldots, K$$

$$\frac{\partial f}{\partial \hat{a}_{Rkl}} = \hat{\mathbf{a}}_R^{\mathrm{T}} \hat{\mathbf{A}}_R^{-1} \mathbf{E}_{kl} \hat{\mathbf{A}}_R^{-1} \hat{\mathbf{T}}\ , \quad k, l = 1, \ldots, K$$

$$\frac{\partial f}{\partial \hat{T}_l} = \sum_{k=1}^{K} \hat{a}_{Rk} \hat{a}_R^{kl}\,, \quad l = 1, \ldots, K$$

where $\mathbf{E}_{kl}$ is a K-square matrix of 0s, with a 1 in position kl. Now, evaluating these partial derivatives at the expected points $\hat{\mathbf{a}}_R = \mathbf{a}_R$, $\hat{\mathbf{A}}_R = \mathbf{A}_R$ and $\hat{\mathbf{T}} = \mathbf{T}$, the first-order Taylor series approximation of $\hat{T}_{CAL}$ gives rise to

$$\hat{T}_{CAL} \approx \sum_{k=1}^{K} \sum_{l=1}^{K} a_{Rk} T_l a_R^{kl} + \sum_{k=1}^{K} \sum_{l=1}^{K} (\hat{a}_{Rk} - a_{Rk}) T_l a_R^{kl}$$

$$- \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{E}_{kl} \mathbf{A}_R^{-1} \mathbf{T} \, (\hat{a}_{Rkl} - a_{Rkl}) + \sum_{k=1}^{K} \sum_{l=1}^{K} \hat{a}_{Rk} (\hat{T}_l - T_l) \hat{a}_R^{kl}$$

$$= \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} + (\hat{\mathbf{a}}_R - \mathbf{a}_R)^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \, (\hat{\mathbf{A}}_R - \mathbf{A}_R) \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{a}_R^{T} \mathbf{A}_R^{-1} (\hat{\mathbf{T}} - \mathbf{T})$$

$$= \hat{\mathbf{a}}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \, \hat{\mathbf{A}}_R \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \hat{\mathbf{T}} \tag{1.14}$$

where a_R^{kl} denotes the kl-element of $\mathbf{A}_R^{-1}$. From (1.14)

$$E(\hat{T}_{CAL}) \approx \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} = \mathbf{b}_R^{\mathrm{T}} \mathbf{T} \tag{1.15}$$

where

$$\mathbf{b}_R = [b_{R1} \ \ldots \ b_{RK}]^{\mathrm{T}} = \mathbf{A}_R^{-1}\mathbf{a}_R = \left(\sum_{j\in \mathrm{U}_R} \mathbf{x}_j\mathbf{x}_j^T\right)^{-1} \sum_{j\in \mathrm{U}_R} y_j\mathbf{x}_j$$

is the coefficient vector of the least-square hyperplane fitted from the respondent population scatter $\{(\mathbf{x}_j, y_j),\ j \in \mathsf{U}_R\}$. In this sense, $\hat{\mathbf{b}}_R$ may be viewed as an approximately unbiased estimator of $\mathbf{b}_R$, obtained from the respondent sample R. In order to obtain some insight into the possible bias of $\hat{T}_{CAL}$, expression (1.13) can be suitably rewritten as

$$\begin{aligned} E(\hat{T}_{CAL}) \approx \mathbf{b}_R^{\mathrm{T}}\mathbf{T} = \mathbf{b}_R^{\mathrm{T}} \sum_{j\in \mathrm{U}} \mathbf{x}_j &= \sum_{j\in \mathrm{U}} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j \\ &= \sum_{j\in \mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j + \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j \\ &= \sum_{j\in \mathrm{U}_R} (y_j - e_{Rj}) + \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j \\ = \sum_{j\in \mathrm{U}_R} y_j + \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j &= T_{Ry} + \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j \end{aligned} \tag{1.16}$$

where the e_{Rj} denotes the 0-sum residuals from the least-square fitting performed on the respondent population scatter, i.e. $e_{Rj} = y_j - \mathbf{b}_R^T\mathbf{x}_j$ for $j \in \mathsf{U}_R$ and T_{Ry} obviously denotes the total of the interest variable in the respondent population. In accordance with (1.14), the approximate design-based bias of $\hat{T}_{CAL}$ turns out to be

$$\mathrm{AB}(\hat{T}_{CAL}) \approx \mathrm{E}(\hat{T}_{CAL}) - T_y = T_{Ry} + \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j - T_y = \sum_{j\in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}}\mathbf{x}_j - \sum_{j\in \mathrm{U}-\mathrm{U}_R} y_j$$

$$= \sum_{j \in \mathrm{U}-\mathrm{U}_R} \mathbf{b}_R^{\mathrm{T}} \mathbf{x}_j - \sum_{j \in \mathrm{U}-\mathrm{U}_R} (\mathbf{b}_{NR}^{\mathrm{T}} \mathbf{x}_j + e_{NRj}) = \sum_{j \in \mathrm{U}-\mathrm{U}_R} (\mathbf{b}_R - \mathbf{b}_{NR})^{\mathrm{T}} \mathbf{x}_j \quad (1.17)$$

where the e_{NRj} denotes the 0-sum residuals from the regression performed on the nonrespondent population scatter, i.e. $e_{NRj} = y_j - \mathbf{b}_{NR}^T \mathbf{x}_j$ for $j \in \mathsf{U} - \mathsf{U}_R$ and

$$\mathbf{b}_{NR} = [b_{NR1} \ \ldots \ b_{NRK}]^{\mathrm{T}} = \left(\sum_{j \in \mathrm{U}-\mathrm{U}_R} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j \in \mathrm{U}-\mathrm{U}_R} y_j \mathbf{x}_j$$

As pointed out by Särndal and Lundström (2005, p.99) the approximate bias does not depend from the design. Moreover, it is at once apparent from (1.17) that the bias of $\hat{T}_{CAL}$ strictly depends on the difference between the least-squares hyperplanes fitted from the respondent and nonrespondent population scatters. Unbiasedness is achieved when the two hyperplanes are identical, i.e. the linear relationship among interest and auxiliary variables is similar for respondent and nonrespondent units. Expression(1.17) can also be obtained from the more general bias expression derived by Särndal and Lundström (2005, Proposition 9.1) under the so called *nonresponse model approach*, under which inference is made with respect to the joint distribution induced by the sampling design and the nonresponse mechanism, supposing a response probability θ_j for each unit and that units respond independently of one another. If the r_js are supposed to be degenerate random variables equal to 1 for $j \in \mathsf{U}_R$ and 0 otherwise, then the Särndal-Lundström bias

expression coincides with(1.17).

1.5 Design-based variance of calibration estimator

In order to derive an approximate expression for the variance of $\hat{T}_{CAL}$, it should be noticed that an additive constant is present in (1.14) owing to the fact that the first K_U components of $\hat{\mathbf{T}}$ are the true population totals of $X_1^*, \ldots, X_{K_U}^*$ (info U). Accordingly, (1.14) can be more suitably expressed as

$$\hat{T}_{CAL} \approx \hat{\mathbf{a}}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \hat{\mathbf{A}}_R \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{b}_R^{*\mathrm{T}} \mathbf{T}^* + \mathbf{b}_R^{\circ\mathrm{T}} \hat{\mathbf{T}}^\circ =$$

$$= \left(\sum_{j \in \mathrm{S}} \frac{r_j y_j \mathbf{x}_j}{\pi_j} \right)^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \left(\sum_{j \in \mathrm{S}} \frac{r_j \mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right) \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{b}_R^{\circ\mathrm{T}} \left(\sum_{j \in \mathrm{S}} \frac{r_j \mathbf{x}_j^\circ}{\pi_j} \right) + const =$$

$$= \sum_{j \in \mathrm{S}} \frac{r_j \left(y_j \mathbf{x}_j^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{x}_j \mathbf{x}_j^T \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{b}_R^{\circ\mathrm{T}} \mathbf{x}_j^\circ \right)}{\pi_j} + const =$$

$$= \sum_{j \in \mathrm{S}} \frac{r_j \left(e_{Rj} \mathbf{x}_j^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} + y_j - e_{Rj}^\circ \right)}{\pi_j} + const = \sum_{j \in \mathrm{S}} \frac{r_j u_j}{\pi_j} + const \qquad (1.18)$$

where for any $j \in \mathsf{U}_R$ $u_j = e_{Rj} \mathbf{x}_j^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} + y_j - e_{Rj}^\circ$ are the so called *influence values* (e.g. Davison and Hinkley, 1997), e_{Rj} denotes the 0-sum residuals from the least-square fitting performed on the respondent population scatter, i.e. $e_{Rj} = y_j - \mathbf{b}_R^{\mathrm{T}} \mathbf{x}_j$ for $j \in \mathsf{U}_R$, $\mathbf{b}_R^\circ$ denotes the last M components of $\mathbf{b}_R$ and $e_{Rj}^\circ = y_j - \mathbf{b}_R^{\circ\mathrm{T}} \mathbf{x}_j^\circ$ for $j \in \mathsf{U}_R$ are the non-0-sum residuals from

the fitting obtained neglecting the Info-U-variable coefficients of $\mathbf{b}_R$. Since $r_j u_j$ equals 0 for any $j \in \mathsf{U}_{NR}$, there is no need to define the u_js outside the nonresponse stratum U_R. Up to a constant term, the approximation (1.18) to $\hat{T}_{CAL}$ may be viewed as the HT estimator of the total of the $r_j u_j$s over U (which coincides with the total of the u_js over U_R which, in turn, coincides with $\sum_{j\in \mathrm{U}_R} y_j - \sum_{j\in \mathrm{U}_R} \overset{\circ}{e}_{Rj}$). Accordingly, the approximate variance of $\hat{T}_{CAL}$ turns out to be (e.g. Särndal et al, 1992, p.175)

$$\mathrm{AV}(\hat{T}_{CAL}) \approx \sum_{h>j\in \mathrm{U}} (\pi_j \pi_h - \pi_{jh}) \left(\frac{r_j u_j}{\pi_j} - \frac{r_h u_h}{\pi_h} \right)^2 \tag{1.19}$$

Moreover, adopting the alternative expression for the variance of the HT estimators, expression (1.19) can be rewritten as

$$\begin{aligned} \mathrm{AV}(\hat{T}_{CAL}) &\approx \sum_{j\in \mathrm{U}} \frac{1-\pi_j}{\pi_j} r_j u_j^2 + 2 \sum_{h>j\in U} \frac{\pi_{jh} - \pi_j \pi_h}{\pi_j \pi_h} r_j r_h u_j u_h \\ &= \sum_{j\in \mathrm{U}_R} \frac{1-\pi_j}{\pi_j} u_j^2 + 2 \sum_{h>j\in \mathrm{U}_R} \frac{\pi_{jh} - \pi_j \pi_h}{\pi_j \pi_h} u_j u_h \end{aligned} \tag{1.20}$$

where

$$u_j = y_j \mathbf{x}_j^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} - \mathbf{a}_R^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{x}_j \mathbf{x}_j^T \mathbf{A}_R^{-1} \mathbf{T} + \mathbf{b}_R^{\circ \mathrm{T}} \mathbf{x}_j^{\circ}, j \in \mathsf{U}_R$$

From the previous expressions it readily follows that the design-based approximate variability of $\hat{T}_{CAL}$ jointly depends on: *i*) the ability of the whole set of K auxiliary variables to predict the interest variable; *ii*) the ability of the K_S Info-S variables to predict the interest variable neglecting the contribution of the K_U Info-U variables; *iii*) the estimation of the total of the

interest variable in the U_R domain. Accordingly, if the K auxiliary variables predict the interest variable without errors, the variability depends on *ii*) and *iii*) only, while if the K_S Info-S variables suffice to predict the interest variable without error (i.e. if the first K_U components of $\mathbf{b}_R$ are equal to 0) then the variability depends only on the ability of the design to estimate the total of the interest variable in the U_R domain. Moreover, expression (1.20) is more effective than (1.19) in emphasizing how the design-based variability of $\hat{T}_{CAL}$ is mainly determined by the respondent population U_R, depending on U by means of $\mathbf{T}$ only.

On the basis of (1.19), the Sen-Yates-Grundy variance estimator is given by

$$\mathrm{V}^2_{\mathrm{SYG}}(\hat{T}_{CAL}) = \sum_{h>j\in \mathrm{S}} \frac{\pi_j\pi_h - \pi_{jh}}{\pi_{jh}} \left(\frac{r_j\hat{u}_j}{\pi_j} - \frac{r_h\hat{u}_h}{\pi_h} \right)^2 \tag{1.21}$$

where $\hat{u}_j = \hat{e}_{Rj}\mathbf{x}_j^{\mathrm{T}}\hat{\mathbf{A}}_R^{-1}\hat{\mathbf{T}} + \hat{\mathbf{b}}_R^{\circ\mathrm{T}}\mathbf{x}_j^{\circ}$ are the empirical influence values computed for each, $j \in \mathsf{R}$, $\hat{e}_{Rj} = y_j - \hat{\mathbf{b}}_R^{\mathrm{T}}\mathbf{x}_j$ are the residual achieved from the least-square fitting performed on the respondent point scatter $\left\{ \left(\mathbf{x}_j, y_j \right), j \in \mathsf{R} \right\}$ and $\hat{\mathbf{b}}_R^{\circ}$ denotes the last K_S components of $\hat{\mathbf{b}}_R$.

While on the basis of (1.20) the HT variance estimator is given by

$$\begin{aligned} \mathrm{V}^2_{\mathrm{HT}} &= \sum_{j\in \mathrm{S}} \frac{1-\pi_j}{\pi_j^2} r_j\hat{u}_j^2 + 2 \sum_{h>j\in \mathrm{S}} \frac{\pi_{jh} - \pi_j\pi_h}{\pi_j\pi_h\pi_{jh}} r_j r_h \hat{u}_j\hat{u}_h \\ &= \sum_{j\in \mathrm{R}} \frac{1-\pi_j}{\pi_j^2} \hat{u}_j^2 + 2 \sum_{h>j\in \mathrm{R}} \frac{\pi_{jh} - \pi_j\pi_h}{\pi_j\pi_h\pi_{jh}} \hat{u}_j\hat{u}_h \end{aligned} \tag{1.22}$$

where now $\hat{u}_j = y_j \mathbf{x}_j^{\mathrm{T}} \hat{\mathbf{A}}_R^{-1} \hat{\mathbf{T}} - \hat{\mathbf{a}}_R^{\mathrm{T}} \hat{\mathbf{A}}_R^{-1} \mathbf{x}_j \mathbf{x}_j^T \hat{\mathbf{A}}_R^{-1} \hat{\mathbf{T}} + \hat{\mathbf{b}}_R^{\circ \mathrm{T}} \mathbf{x}_j^{\circ}$ are the empirical influence values computed for each $j \in \mathsf{R}$. Alternatively, the jackknife variance estimator by Berger and Skinner (2005) can be used. The jackknife estimator is analogous (1.22) but with the empirical influence values which are numerically approximated instead of obtained by analytic differentiation. Quoting from Berger and Skinner (2005), denote by

$$v_{(j)} = \left(1 - \frac{1}{\hat{N}\pi_j}\right) \left\{\hat{T}_{CAL} - \hat{T}_{CAL(j)}\right\}$$

where

$$\hat{N} = \sum_{j \in \mathrm{S}} \frac{1}{\pi_j}, \ \hat{T}_{CAL(j)} = \hat{\mathbf{b}}_{R(j)}^{\mathrm{T}} \hat{\mathbf{T}}_{(j)},$$

$$\hat{\mathbf{b}}_{R(j)} = \left(\sum_{h \in \mathrm{S}_{-j}} \frac{r_h \mathbf{x}_h \mathbf{x}_h^T}{\pi_h}\right)^{-1} \sum_{h \in \mathrm{S}_{-j}} \frac{r_h y_h \mathbf{x}_h}{\pi_h}, \hat{\mathbf{T}}_{(j)} = \left[T_1, \ldots, T_{K_U}, \hat{T}_{1(j)}, \ldots, \hat{T}_{K_S(j)}\right]^{\mathrm{T}},$$

$$\hat{T}_{k(j)}^{\circ} = \frac{N}{\hat{N}} \sum_{h \in \mathrm{S}_{-j}} \frac{x_{h,k}^{\circ}}{\pi_h}, \quad for \ k = 1, \ldots, K_S$$

and finally S_{-j} consists of the sample S with the j-th unit deleted. Accordingly, the jackknife estimator for the variance of $\hat{T}_{CAL}$ turns out to be

$$V_{jack}^2 = \sum_{j \in \mathrm{S}} (1 - \pi_{\mathrm{j}}) v_{(j)}^2 + 2 \sum_{h > j \in \mathrm{S}} \frac{\pi_{jh} - \pi_{\mathrm{j}} \pi_h}{\pi_{jh}} v_{(j)} v_{(h)} \tag{1.23}$$

1.6 Searching for the effective auxiliary information

Särndal and Lundström (2005, p. 98) point out as the bias of any nonresponse-adjusted estimator should be the main concern, emphasizing that variance is of minor importance since "*if an estimator is greatly biased, it is poor consolation that its variance is low*". Since $\hat{T}_{CAL}$ estimates T_y without error, the search for auxiliary information which is likely to be effective from a design-based point of view should be guided by the following criterion, referred to as *Principle 2* in the parlance of Särndal and Lundström (2005, p. 110): "*the auxiliary vector should explain the main study variables*". In this framework, a good indicator of the capacity of the $\mathbf{x}_j$s to predict the y_js should obviously given by the fraction of the Y-variance explained by the selected variables $X_1, \ldots, X_K$, i.e.

$$\eta^2 = 1 - \frac{\sum_{j \in \mathrm{U}} (y_j - \mathbf{b}^{\mathrm{T}} \mathbf{x}_j)^2}{\sum_{j \in \mathrm{U}} (y_j - \bar{Y})^2} \tag{1.24}$$

where

$$\mathbf{b} = \left(\sum_{j \in \mathrm{U}} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \sum_{j \in \mathrm{U}} y_j \mathbf{x}_j$$

is now the coefficient vector of the least-square hyperplane fitted from the whole population scatter $\{(\mathbf{x}_j, y_j),\, j \in \mathsf{U}\}$ and $\bar{Y} = T_y/N$ is the population mean. Unfortunately, Principle 2 does not seem a suitable solution, at least in a complete design-based approach. Indeed, η^2 is unknown so that we are

forced to estimate it by means of information provided by the respondent sample R. A very natural solution is given by Särndal and Lundström (2005, p. 122) who propose to assess the effectiveness of the selected auxiliary variables on the basis of

$$\hat{\eta}_R^2 = 1 - \frac{\sum_{j\in \mathrm{R}} \frac{1}{\pi_j}(y_j - \hat{\mathbf{b}}_R^{\mathrm{T}}\mathbf{x}_j)^2}{\sum_{j\in \mathrm{R}} \frac{1}{\pi_j}(y_j - \hat{\bar{Y}}_R)^2} \tag{1.25}$$

where $\hat{\bar{Y}}_R = \hat{T}_R/\hat{N}_R$ and

$$\hat{N}_R = \sum_{j\in \mathrm{R}} \frac{1}{\pi_j}$$

In order to derive the design-based properties of (1.25) as an estimator of (1.24), it is convenient to rewrite (1.25) in a more suitable form. After trivial algebra we have

$$\hat{\eta}_R^2 = 1 - \frac{\hat{Q}_R - \hat{\mathbf{a}}_R^{\mathrm{T}}\hat{\mathbf{A}}_R^{-1}\hat{\mathbf{a}}_R}{\hat{Q}_R - \frac{\hat{T}_R^2}{\hat{N}_R}} \tag{1.26}$$

where

$$\hat{Q}_R = \sum_{j\in \mathrm{S}} \frac{r_j y_j^2}{\pi_j}$$

and $\hat{T}_R$ and $\hat{N}_R$ are suitably expressed in term of S as

$$\hat{T}_R = \sum_{j\in \mathrm{S}} \frac{r_j y_j}{\pi_j}$$

$$\hat{N}_R = \sum_{j\in \mathrm{S}} \frac{r_j}{\pi_j}$$

Obviously, it is at once apparent that

$$\mathrm{E}(\hat{Q}_R) = E\left(\sum_{j\in \mathrm{S}} \frac{r_j y_j^2}{\pi_j}\right) = \sum_{j\in \mathrm{U}} r_j y_j^2 = \sum_{j\in \mathrm{U}_R} y_j^2 = Q_R$$

as well as

$$\mathrm{E}(\hat{T}_R) = E\left(\sum_{j\in \mathrm{S}} \frac{r_j y_j}{\pi_j}\right) = \sum_{j\in \mathrm{U}} r_j y_j = \sum_{j\in \mathrm{U}_R} y_j = T_R$$

$$\mathrm{E}(\hat{N}_R) = E\left(\sum_{j\in \mathrm{S}} \frac{r_j}{\pi_j}\right) = \sum_{j\in \mathrm{U}} r_j = N_R$$

Thus, $\hat{\eta}^2_R$ can be rewritten as a function of the five HT estimators on which it depends, say $\hat{\eta}^2_R = f(\hat{Q}_R, \hat{\mathbf{a}}_R, \hat{\mathbf{A}}_R, \hat{T}_R, \hat{N}_R)$, in such a way that differentiating f with respect to all the variables involved, evaluating the partial derivatives at the expected points $\hat{Q}_R = Q_R$, $\hat{\mathbf{a}}_R = \mathbf{a}_R$ $\hat{\mathbf{A}}_R = \mathbf{A}_R$, $\hat{T}_R = T_R$ and $\hat{N}_R = N_R$, and then considering the expectation of the first-order Taylor series approximation of $\hat{\eta}^2_R$, it follows that

$$\mathrm{E}(\hat{\eta}^2_{\mathrm{R}}) \approx \mathrm{f}(\mathrm{Q_R}, \mathbf{a}_{\mathrm{R}}, \mathbf{A}_{\mathrm{R}}, \mathrm{T_R}, \mathrm{N_R}) = 1 - \frac{\mathrm{Q_R} - \mathbf{a}^{\mathrm{T}}_{\mathrm{R}}\mathbf{A}^{-1}_{\mathrm{R}}\mathbf{a}_{\mathrm{R}}}{\mathrm{Q_R} - \frac{\mathrm{T}^2_{\mathrm{R}}}{\mathrm{N_R}}}$$

$$= 1 - \frac{\sum_{j\in \mathrm{U}_R}(y_j - \mathbf{b}^{\mathrm{T}}_R \mathbf{x}_j)^2}{\sum_{j\in \mathrm{U}_R}(y_j - \bar{Y}_R)^2} = \eta^2_R$$

where $\bar{Y}_R = T_R/N_R$.

Practically speaking $\hat{\eta}^2_R$ provides an approximately unbiased estimator of the Y-variance explained by the selected variables, not in the whole population U, but only in the respondent population U_R. Paradoxically, the procedure based on $\hat{\eta}^2_R$ may provide reliable choices of the auxiliary variables only if the linear relationship among interest and auxiliary variables is similar for respondent and nonrespondent units, a situation which alone ensures approximate unbiasedness. In order to search for auxiliary variables which behave

similarly for respondent and norespondent units, a promising, even if trivial, procedure should be based on the comparison of ranges in respondent and nonrespondent populations (Info U) or samples (Info S). Indeed, if the values of an auxiliary variable in the respondent population (sample) tend to be much greater or lower than the values of the same variable in the nonresponse counterpart, then it is quite difficult that the same linear relationship may hold for both cases. As a very simple example, consider the slope of terrain as an auxiliary variable adopted to predict the timber volume in forest inventories. Slopes (in percentage) in the sites/plots placed in flat terrain which can be easily reached by foresters (respondent population) usually range from 0 to 40% while they range from about 40 to 60% (with some values reaching 80%) for those plots placed in steep terrains and not suitable for surveying (nonrespondent population). Thus, it is quite unlike that the same linear relationship may be valid to predict timber volume as linear function of slope in the whole range 0-80%. From these considerations, it seems that the choice of auxiliary information in design-based calibration approach should be guided by practical considerations about the nature of the variables and their relationship with the interest variable rather than by rigid quantitative indicators which, being necessarily computed from the respondent sample, can reflect the actual situation only for the respondent population. Finally, even if obvious, it is also worth noting that the selection of highly correlated

auxiliary variables should be avoided and only one of them should be used. Indeed, the use of highly correlated variables deteriorates the estimation of the regression coefficients without providing relevant additional information.

1.7 Simulation study

Empirical investigations were used to throw light in: a) the capability of the approximate expressions for the bias and the variance to guess the actual values; b) the design-based accuracy of the calibration estimator in terms of amount of nonresponse, sampling effort, effectiveness of the auxiliary variables to predict the interest variable, differences in the behaviour of the auxiliary variable between respondent and norespondent units and multicollinearity among auxiliary variables; c) the capability of variance estimators to evaluate the accuracy of the calibration estimator and to give confidence interval with coverage near to the nominal level. To this purpose a population of $N = 1,000$ individuals was considered, partitioned into respondent and nonrespondent stratum. The size of respondent stratum was presumed to be $N_R = 300,\ 600,\ 900$ corresponding to respondent percentages (say RP) of 30%, 60%, 90%. Then, two auxiliary variables X_1^* and X_2^* were supposed to be known for each population unit (Info U). For each unit $j \in \mathsf{U}$, the values x_1^* and x_2^* were generated from a bivariate normal distribution with

expectations $\mu_1 = \mu_2 = 1$ and variances $\sigma_1^2 = \sigma_2^2 = 1$. Moreover, in order to take into account different degrees of multicollinearity (MC), three levels of correlation were presumed between X_1^* and X_2^*: 0, 0.5, 0.9. Then, for each unit of the respondent stratum the interest variable Y was achieved from the relation

$$y_j = 1 + 0.5x_{j1}^* + 0.5x_{j2}^* + \varepsilon_j \tag{1.27}$$

where ε_j was an error term generated from a centred normal distribution. On the other hand, as to nonrespondent stratum, three similarity levels (SL) with relation(1.27) were considered: a first situation (say SL1) in which the y_js were generated by the same relation adopted in respondent stratum, a second situation (say SL2) in which the coefficients attached to x_1^* and x_2^* were two times those adopted in respondent stratum, a last situation (say SL3) in which the coefficients were four times those adopted in respondent stratum. Finally, the variances of the error terms in respondent and nonrespondent stratum, say δ_R^2 and δ_{NR}^2 was chosen in such a way to achieves a fraction of explained variance (say FEV) equal to 0.3, 0.6 and 0.9 for both the respondent and nonrespondent population point scatters. Then, from the possible combinations of RP, SL, FEV and MC, a final set of 81 populations was achieved. In each population, 10,000 samples of size $n = 50$ (corresponding to sampling fraction of 5) were selected by means of simple random

sampling without replacement (SRSWOR). For each selected sample the following quantities were computed: $\hat{T}_{CAL}$, V^2_{SYG} and V^2_{jack} (note that under SRSWR, V^2_{SYG} and V^2_{HT} coincide). From the variance estimates the corresponding estimates of the relative standard error $RSE_{SYG} = V_{SYG}/\hat{T}_{CAL}$ and $RSE_{jack} = V_{jack}/\hat{T}_{CAL}$ were also computed together with the confidence intervals $\hat{T}_{CAL}\pm 1.96 V_{SYG}$ and $\hat{T}_{CAL}\pm 1.96 V_{jack}$. Then, from the resulting Monte Carlo distributions, the relative bias, coefficient of variation and relative root mean squared error of $\hat{T}_{CAL}$, say RB-CAL, CV-CAL and RRMSE-CAL were empirically evaluated together with the expectations of RSE_{SYG} and RSE_{jack}, say ERSE-SYG and ERSE-JACK and the coverage of the confidence intervals, say CVRG-SYG and CVRG-JACK, achieved as the percentage of times the intervals included the true total. Moreover, for each population the approximate bias and variance of $\hat{T}_{CAL}$, say ARB-CAL and ACV-CAL were analytically computed by means of equations (1.17) and (1.19), together with the coefficient of variation which would be achieved by the HT estimator in the case of complete response, say CV-HT. This quantity was included as a bench-mark with which the accuracy of $\hat{T}_{CAL}$ can be compared. For each population, Tables 1.1-1.5 reports the percent values of ARB-CAL, RB-CAL, ACV-CAL, CV-CAL, RRMSE-CAL, CV-HT, ERSE-SYG, ERSE-JACK, CVRG-SYG and CVRG-JACK. The simulation results motivates the following comments:

- when the relationship among interest and auxiliary variable is similar in respondent and nonrespondent sub-populations (SL1), underestimation due to nonresponse turns out to be negligible but it markedly increases as differences in the relationships are present (SL2 and SL3); downward bias also increases with the amount of nonresponse but it seems to be poorly influenced by FEV and MC factors (see Table 1.1);

- the approximate bias expression (1.17) turns out to be quite accurate; for RP equal to 30%, the approximate relative bias shows differences with the actual relative bias always smaller than 5 percent points except for FEV equal to 0.3 and MC equal to 0.9, when the differences are of about 10 percent points; the accuracy of the approximation quickly increases as RP increases with differences which become negligible when RP reaches 90% (see Table 1.1);

- even if the approximate variance expression (1.19) invariably provides underestimation of the actual variance, it turns out to be satisfactory: the differences between the approximate and actual coefficient of variations are always smaller then 3 percent points for RP equal to 30% and become negligible as RP increases (see Table 1.2);

- when the relationship among interest and auxiliary variable is similar in respondent and nonrespondent sub-populations (SL1), nonresponse deteriorates the performance of calibration estimation with respect to the complete-

sample HT estimator only in presence of a massive amount of nonresponse (RP=30%), while for small amount of nonresponse the calibration procedure even provide improvement with the complete sample performance; on the other side, when the difference between the relationships in respondent and nonrespondent sub-populatioins become marked, the presence of substantial bias deteriorates the performance of calibration estimator with relative errors 3-5 times greater that those achieved with complete samples (see Table 1.3);

- the SYG/HT estimator always provides underestimation of the relative standard error as opposite to jackknife which proves to be invariably conservative; both downward and upward bias tend to reduce as RP and FEV increase: for RP equal to 90% and FEV equal to 0.9 both the estimators are practically unbiased (see Table 1.4);

- the coverage of confidence intervals well approximate the nominal level only when the relationship among interest and auxiliary variable is similar in respondent and nonrespondent sub-populations (SL1) and for respondent percentages of 60 and 90%; in the other cases (SL2 and SL3), the presence of bias skews the confidence intervals entailing disastrous losses of coverage; intervals achieved using the jackknife variance estimator invariably perform better than those achieved using SYG/HT estimator (see Table 1.5).

Table 1.1

			RP = 30%		RP = 60%		RP = 90%	
SL	η^2	ρ	ARB-CAL	RB-CAL	ARB-CAL	RB-CAL	ARB-CAL	RB-CAL
SL1	0.3	0	0	-2	0	3	0	0
		0.5	0	3	0	4	0	1
		0.9	0	-7	0	1	0	1
	0.6	0	0	-2	0	1	0	0
		0.5	0	-2	0	-1	0	0
		0.9	0	1	0	1	0	0
	0.9	0	0	1	0	0	0	0
		0.5	0	0	0	0	0	0
		0.9	0	1	0	0	0	0
SL2	0.3	0	-26	-26	-17	-19	-5	-4
		0.5	-24	-30	-17	-20	-4	-2
		0.9	-25	-28	-16	-17	-5	-3
	0.6	0	-27	-27	-16	-15	-4	-5
		0.5	-27	-26	-16	-17	-5	-4
		0.9	-26	-27	-16	-16	-6	-5
	0.9	0	-26	-26	-16	-17	-5	-5
		0.5	-24	-25	-17	-17	-5	-5
		0.9	-24	-24	-17	-17	-4	-5
SL3	0.3	0	-49	-50	-36	-39	-14	-14
		0.5	-50	-50	-37	-37	-14	-12
		0.9	-57	-46	-37	-41	-14	-14
	0.6	0	-50	-51	-37	-38	-12	-12
		0.5	-51	-50	-37	-37	-12	-14
		0.9	-50	-55	-41	-39	-13	-15
	0.9	0	-51	-50	-38	-39	-13	-13
		0.5	-51	-51	-37	-37	-12	-12
		0.9	-51	-52	-40	-39	-13	-13

Table 1.2

			RP = 30%		RP = 60%		RP = 90%	
SL	η^2	ρ	ACV-CAL	CV-CAL	ACV-CAL	CV-CAL	ACV-CAL	CV-CAL
SL1	0.3	0	12	14	10	10	8	8
		0.5	16	19	12	13	9	10
		0.9	18	21	13	14	10	11
	0.6	0	7	8	5	5	4	4
		0.5	9	10	6	7	5	5
		0.9	10	12	7	8	6	6
	0.9	0	3	3	2	2	2	2
		0.5	3	4	3	3	2	2
		0.9	4	5	3	3	2	2
SL2	0.3	0	10	11	8	8	7	8
		0.5	13	15	10	11	10	10
		0.9	14	15	11	11	9	10
	0.6	0	5	6	5	5	4	4
		0.5	7	8	5	6	5	5
		0.9	7	8	6	6	5	5
	0.9	0	2	2	2	2	2	2
		0.5	3	3	2	2	2	2
		0.9	3	4	2	3	2	2
SL3	0.3	0	6	7	6	6	6	7
		0.5	7	8	7	7	9	9
		0.9	10	12	8	8	9	9
	0.6	0	4	4	3	3	4	4
		0.5	5	5	4	4	5	5
		0.9	5	5	4	5	5	5
	0.9	0	2	2	1	1	2	2
		0.5	2	2	2	2	2	2
		0.9	2	2	2	2	2	2

Table 1.3

			RP = 30%		RP = 60%		RP = 90%	
SL	η^2	ρ	RRMSE-CAL	CV-HT	RRMSE-CAL	CV-HT	RRMSE-CAL	CV-HT
SL1	0.3	0	14	9	10	9	8	9
		0.5	19	11	14	11	10	11
		0.9	22	13	14	12	11	12
	0.6	0	8	6	5	6	4	6
		0.5	10	8	7	8	5	8
		0.9	12	9	8	9	6	9
	0.9	0	3	5	2	5	2	5
		0.5	4	6	3	6	2	6
		0.9	5	7	3	7	2	7
SL2	0.3	0	28	11	21	11	9	10
		0.5	34	14	23	14	10	13
		0.9	32	16	21	15	10	12
	0.6	0	27	9	16	8	7	7
		0.5	27	10	18	10	7	9
		0.9	28	12	17	12	7	10
	0.9	0	26	7	17	7	5	6
		0.5	25	9	18	8	5	7
		0.9	24	10	17	10	5	8
SL3	0.3	0	51	15	40	16	16	13
		0.5	51	18	38	18	15	16
		0.9	47	22	42	21	17	18
	0.6	0	51	12	38	12	13	9
		0.5	51	13	37	14	15	11
		0.9	56	14	39	16	16	13
	0.9	0	50	10	39	11	13	9
		0.5	51	12	37	13	12	10
		0.9	52	13	39	13	13	11

Table 1.4

			RP = 30%			RP = 60%			RP = 90%		
SL	η^2	ρ	ERSE-SYG		ERSE-JACK	ERSE-SYG		ERSE-JACK	ERSE-SYG		ERSE-JACK
SL1	0.3	0	12	(14)	16	9	(10)	10	8	(8)	9
		0.5	16	(19)	20	12	(13)	13	9	(10)	10
		0.9	20	(21)	26	13	(14)	14	10	(11)	11
	0.6	0	7	(8)	9	5	(5)	5	4	(4)	4
		0.5	9	(10)	12	6	(7)	7	5	(5)	5
		0.9	10	(12)	13	7	(8)	8	6	(6)	6
	0.9	0	3	(3)	4	2	(2)	2	2	(2)	2
		0.5	3	(4)	4	3	(3)	3	2	(2)	2
		0.9	4	(5)	5	3	(3)	3	2	(2)	2
SL2	0.3	0	13	(11)	17	10	(8)	11	8	(8)	8
		0.5	18	(15)	24	13	(11)	14	10	(10)	10
		0.9	18	(15)	24	13	(11)	14	10	(10)	10
	0.6	0	7	(6)	9	5	(5)	6	4	(4)	4
		0.5	9	(8)	11	6	(6)	7	5	(5)	6
		0.9	9	(8)	12	7	(6)	8	5	(5)	6
	0.9	0	3	(2)	4	2	(2)	2	2	(2)	2
		0.5	3	(3)	4	3	(2)	3	2	(2)	2
		0.9	4	(4)	5	3	(3)	3	2	(2)	2
SL3	0.3	0	12	(7)	16	10	(6)	11	8	(7)	8
		0.5	15	(8)	20	11	(7)	12	10	(9)	10
		0.9	19	(12)	24	13	(8)	14	11	(9)	11
	0.6	0	7	(4)	9	5	(3)	6	4	(4)	4
		0.5	9	(5)	12	6	(4)	7	5	(5)	6
		0.9	10	(5)	13	7	(5)	8	6	(5)	6
	0.9	0	3	(2)	4	2	(1)	2	2	(2)	2
		0.5	4	(2)	5	3	(2)	3	2	(2)	2
		0.9	4	(2)	5	3	(2)	3	2	(2)	2

Table 1.5

			RP = 30%		RP = 60%		RP = 90%	
SL	η^2	ρ	CVRG-SYG	CVRG-JACK	CVRG-SYG	CVRG-JACK	CVRG-SYG	CVRG-JACK
SL1	0.3	0	88	95	92	94	93	94
		0.5	88	94	92	94	93	95
		0.9	86	93	93	95	94	95
	0.6	0	88	94	92	94	93	95
		0.5	88	94	92	94	94	95
		0.9	88	94	92	94	94	95
	0.9	0	89	95	93	95	93	95
		0.5	88	94	92	94	94	95
		0.9	87	94	93	95	94	95
SL2	0.3	0	25	39	34	41	90	92
		0.5	33	51	47	55	93	95
		0.9	40	56	62	68	92	93
	0.6	0	2	6	9	13	72	76
		0.5	4	11	11	15	85	87
		0.9	6	15	25	32	82	84
	0.9	0	0	0	0	0	18	21
		0.5	0	0	0	0	31	35
		0.9	0	0	0	0	40	44
SL3	0.3	0	0	0	0	0	39	43
		0.5	0	1	0	0	69	72
		0.9	2	7	0	0	65	69
	0.6	0	0	0	0	0	8	10
		0.5	0	0	0	0	14	16
		0.9	0	0	0	0	16	18
	0.9	0	0	0	0	0	0	0
		0.5	0	0	0	0	0	0
		0.9	0	0	0	0	0	0

1.8 Final remarks

The design-properties of the calibration estimation are approximated considering the unit nonresponse as a fixed characteristic, just like the values of interest and auxiliary variables, a situation which is likely to occur in environmental surveys. On the basis of the approximate expression of the variance, three variance estimators were attempted using the SYG, HT and jackknife criteria. Obviously all the estimators are likely to provide reliable accuracy evaluations and confidence intervals only when nonresponse bias is small. The results of simulation study largely confirm these considerations. In presence of a considerable bias, which is mainly generated when different relationships among interest and auxiliary variables hold in respondent and nonrespondent sub-populations, any inference (estimation, estimation of accuracy and confidence interval construction) turns out to be completely unreliable. Thus, attention should be paid in the selection of auxiliary variables which should be chosen not on the basis of their capability to explain the interest variable (which can only be checked on the respondent population) but rather on the basis of the stability of their relationship with the interest variable in respondent and non-respondent sub-populations. In this framework, then the choice of auxiliary variables should be mainly guided by practical considerations and previous experience. Under small bias, sim-

ulation results prove the effectiveness of nonresponse calibration weighting (NCW) under small amount of nonresponse. Then, conditional to small biases, NCW approach seems to be especially appealing in environmental surveys, where nonresponse percents are likely to be smaller than 5%.

1.9 Practical considerations

As already pointed out in section 1.4, the calibration estimation can be performed providing that the cross-product matrix $\hat{\mathbf{A}}_R$ is positive definite. If the K selected variables are of quantitative nature there is no problem, apart for the trivial case in which the selected variables are such that $\mathbf{d}^T\mathbf{x}_j = 0$ for any $j \in \mathsf{U}$ (info U) or any $j \in \mathsf{S}$ (info S) and for one or more constant vectors $\mathbf{d}$. Obviously, in this case there is at least a redundant variable to discard since it linearly depends on the remaining $K - 1$ variables. As to the choice of quantitative variables, it is also worth noting that, even if not mandatory, the selection of highly correlated auxiliary variables should be avoided and only one of them should be used. Indeed, as is well known in linear regression, the use of highly correlated variables deteriorates the estimation of the regression coefficients without providing relevant additional information. In many situations information is also provided by auxiliary variables of categorical nature. In this case, for any categorical auxiliary

variable X_k which may takes C_k possible categories, the information of each unit j can be straightforwardly accounted by means of a C_k-vector of 0s, say $\mathbf{x}_{j(k)}$, with a 1 in the position corresponding to the category characterizing the unit j. But now a problem arises when there are two (or more) categorical auxiliary variables, say X_k and X_l. Indeed, in this case the vectors $\mathbf{x}_{j(k)}$ and $\mathbf{x}_{j(l)}$ are such that $\mathbf{1}^{\mathrm{T}}\mathbf{x}_{j(k)} - \mathbf{1}^{\mathrm{T}}\mathbf{x}_{j(l)} = 0$ for any $j \in \mathsf{U}$ or any $j \in \mathsf{S}$. Thus, a linear dependence is introduced among the auxiliary variable. In this case, the problem can be easily overcome by using $(C_k - 1)$-vectors instead of C_k-vectors, deleting one of the C_k categories of X_k. Obviously for a given X_k, the missed category must be the same for any $j \in \mathsf{U}$ or any $j \in \mathsf{S}$. Taking in mind these considerations, once K auxiliary variables have been identified as possible source of information, the following procedure may be adopted for determining the best set of auxiliary variables: the index $\hat{\eta}^2_R$ is at first computed on the whole set of variables, and then it is computed discarding one variable at time, two variables at time and so until sets of unique variables are obtained. The greatest value of $\hat{\eta}^2_R$ identifies the more suitable set. Obviously, when a categorical variable is discarded all the $(C_k - 1)$ variables related to it are jointly discarded. Moreover, the variable equal to 1 for any unit in the population is always present in all the sets generated by the procedure, in order to ensure the simplified form of calibration weights. Finally, when two or more sets of auxiliary variables give rise to very similar values

of $\hat{\eta}_R^2$, the smaller set should be preferred in order to provide more stable estimates of regression coefficients.

Chapter 2

A THREE-PHASE SAMPLING STRATEGY FOR FOREST SURVEYS

2.1 Preliminaries

Most forest surveys performed over large scale (e.g. national forest inventories) involve several phases of sampling. Usually, satellite imagery or aerial photo-based information is collected by means of an intensive first-phase sampling while the subsequent phases are performed by ground inspections, in order to combine aerial and field data. This chapter deals with some theoretical aspects of a three-phase sampling strategy adopted in the recent Italian National Forest Inventory (IFNC). The first phase is performed by means of a systematic search over the area to be surveyed, in which the area is partitioned into regular polygons of the same size and points are randomly thrown, one per polygon. In the second phase, the land cover class

of first-phase points is classified by very high resolution remotely sensed imagery available for the whole area. Points aerially classified in a non forest class are discarded while a sample of first-phase points classified as forest is selected in accordance with a probabilistic sampling scheme. Second-phase points are visited on the ground recording their actual land cover class and, for those points actually lying in forest, several characteristics of qualitative and quantitative nature are recorded such as forest category, type of ownership (public or private), altitude, slope, exposure, etc. Finally, in the third phase, a sample of the second-phase forest points is selected in accordance with a probabilistic sampling scheme and a biophysical attribute (e.g. timber volume, biomass, basal area) is recorded for all the trees lying within the plots of pre-fixed size centred at those points. While the second phase suffices to estimate the areal extent of land cover classes as well as of forest categories, the third phase is necessary to estimate the totals of the biophysical attribute in each forest category. The statistical properties of the second-phase estimators of areal extents as well as of the third-phase estimator of totals of the biophysical attribute in each forest category are derived in a multivariate setting by Fattorini et al (2006), when the second- and third-phase points are selected by means of stratified sampling with simple random sampling without replacement (SRSWOR) performed within each stratum. Non-response is actually absent in the second-phase, where most qualitative and quanti-

tative attributes to be recorded in this phase can be actually recorded even when points cannot be reached by foresters. Then, the second-phase estimation of the areal extents of land cover classes and of forest categories is not affected by non response and is here neglected. Rather, we here focus on the third-phase estimation of totals for a biophysical attribute, which instead necessitates that all third-phase points be reached in order to perform all the recording activities within the plots centred on them. Moreover, in order to better focus on non response issues, as opposite to Fattorini et al (2006), we leave unspecified the second- and third phases sampling schemes and we deal with the estimation of the total of a biophysical attribute for the whole forest class rather than for each forest category. Accordingly, the results of this section are of univariate nature and hold for any second- and third-phase fixed-size sampling scheme. Henceafter, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, ... denote areas while $|\mathcal{A}|$, $|\mathcal{B}|$, $|\mathcal{C}|$, $|\mathcal{D}|$, ... denote their corresponding sizes.

2.2 One-phase estimation

Consider a delineated study area $\mathcal{A}$ partitioned into two land cover classes: forest and non forest. Denote by $\mathcal{F} \subset \mathcal{A}$ the forest portion of $\mathcal{A}$ and by F the population of forest trees within $\mathcal{F}$. Denote by y_i the value of an attribute such as above-ground biomass, wood volume or basal area for the i-th forest

tree and suppose that the population total

$$T = \sum_{i \in F} y_i$$

be the interest quantities to be estimated. 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 (as in the present case), 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, also giving (on p. 327) a list of references from the early 90s in which such seminal intuition was first developed. In this setting, the interest parameter T can be expressed as an integral over the study area and the spatial design for selecting points (from which plots 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 may lead to uneven coverage of the study area. As pointed out by Cordy and Thompson (1995) and Stevens (2006), so-called *uniform random sampling* may be unsatisfactory since some subregions may be sparsely sampled whereas others are intensively sampled. To avoid uneven coverage of the study area, systematic schemes can be adopted. However, pure systematic sampling based on a regular grid of points with

a random start (commonly adopted in large-scale forest inventories) may be unsuitable in the presence of some spatial regularity, leading to substantial losses of efficiency. Accordingly, spatially stratified schemes based on a regular tessellation of the study area and the random placement of a point in each tessellation unit have been theoretically preferred by statisticians. One such scheme, usually referred to as *tessellation stratified sampling* (TSS) has a long standing in statistical literature (see e.g. Overton and Stehman, 1993, Cordy and Thompson, 1995, Stevens 1997). The scheme has been proposed by Fattorini et al. (2006) in the first phase of the three-phase sampling strategy constructed for estimating totals of forest attributes on large scale. By TSS scheme, the area is covered by a region, say $\mathcal{R} \supset \mathcal{A}$, constituted by M non-overlapping regular polygons, say $\mathcal{Q}_1, \ldots, \mathcal{Q}_M$, of equal size and such that $\mathcal{Q}_j \bigcap \mathcal{A} \neq \emptyset$ for all $j = 1, \ldots, N$. Then, for each polygon j, a point, say $\mathbf{p}_j$, is randomly thrown within the polygon, in such a way that a discrete population of M points, say $\mathsf{U}_1 = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ is achieved. If each point of U_1 was visited on the ground and a plot of fixed size a was delineated around the point, then for each point j a sample of forest trees, say P_j, was obtained. If the interest attribute was measured and recorded for all the trees of P_j, the HT-like estimator of T at point j turned out to be

$$\hat{T}_j = \frac{|\mathcal{R}|}{a} \sum_{i \in \mathrm{P}_j} y_i \ , \ \ j = 1, \ldots, M$$

Then, it is well-known from Monte Carlo integration (e.g. Gregoire and Valentine, 2008, chapter 10) that the arithmetic mean of the $\hat{T}_j$s, say

$$\hat{\bar{T}}_1 = \frac{1}{M}\sum_{j=1}^{M}\hat{T}_j \tag{2.1}$$

constituted a one-phase unbiased estimator for T. Moreover, owing to the independence of the $\hat{T}_j$s, the variance of (2.1) turned out to be

$$\mathrm{V}_1(\hat{\bar{T}}_1) = \frac{1}{M^2}\sum_{j=1}^{M}\mathrm{V}_1(\hat{T}_j) \tag{2.2}$$

while the quantity

$$V_1^2 = \frac{1}{M(M-1)}\sum_{j=1}^{M}(\hat{T}_j - \hat{\bar{T}}_1)^2 \tag{2.3}$$

constituted a conservative estimator of (2.2), in the sense that $\mathrm{E}_1(V_1^2) \geq \mathrm{V}_1(\hat{\bar{T}}_1)$ (e.g. Wolter, 1985, Theorem 2.4.1). 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 U_1 which can be selected by means of TSS. For the estimation of (2.3) in the subsequent phases, this quantity can be suitably rewritten as

$$V_1^2 = \frac{1}{M^2}\sum_{j=1}^{M}\hat{T}_j^2 - \frac{2}{M^2(M-1)}\sum_{h>j=1}^{M}\hat{T}_j\hat{T}_h \tag{2.4}$$

It is worth noting that some edge effects might be present owing to forest trees positioned near the internal edge of the study region, which will have inclusion probabilities smaller than $a/|\mathcal{R}|$. A long list of correction methods has been proposed in order to avoid the negative bias induced by edge

effects (see e.g. Gregoire and Valentine, 2008, section 7.5). Fortunately, in this framework, TSS performs like the correction method usually referred to as the *buffer method* (e.g. Gregoire and Valentine, 2008, section 7.5.1), which entails allowing sample points to fall outside the boundary of $\mathcal{A}$, but within some larger region that includes $\mathcal{A}$. For this reason, under TSS the presence of forest trees in $\mathcal{A}$ whose inclusion zone overlaps the boundary of the enlarged region $\mathcal{R} \supset \mathcal{A}$ should become negligible. Moreover, it should be noticed that in large scale forest surveys (e.g. national or regional forest inventories) edge problems are present in the plots near the study area's borderlines i.e. mountains ridges, rivers, lakes, roads 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 estimator. As to the theoretical properties of estimator of type (2.1) arising from TSS scheme, if the study area can be exactly tessellated by polygons, TSS turns out to be invariably more efficient than uniform random sampling. This result had been reached, *mutatis mutandi*, in Monte Carlo integration (see e.g. Haber, 1966) but it does not hold when the study areas have irregular shapes, owing to the necessity of introducing an enlarged covering region $\mathcal{R}$ to perform TSS scheme. However, TSS displays variances decreasing with $M^{-\alpha}$, for $1 < \alpha \leq 3$ (Barabesi, 2003, Barabesi and Marcheselli 2003, 2008) while uniform random sampling provides variances decreasing with M^{-1}. Accord-

ingly, for large M, TSS gives rise to relevant gains in precision with respect to the uniform random scheme. It is worth noting that such a plethora of theoretical results cannot be proved under the pure systematic scheme. Indeed, in the presence of some spatial regularity, the systematic scheme may be even worse than the random scheme.

2.3 Two-phase estimation

Unfortunately, owing to the costs and time involved, in real situations plot sampling cannot be performed for each first-phase point, but rather for a portion of these points selected in the second phase. Accordingly, the first-phase survey is only hypothetical and its treatment has had the sole aim of constructing the theoretical basis for the analysis of the subsequent phases. As to the second phase, it is worth noting that the collection U_1 can be partitioned into two sub-sets: the sub-set U_{1F} of the points aerially classified in the forest class and the sub-set $\mathsf{U}_1 - \mathsf{U}_{1F}$ of the remaining points aerially classified in other land cover classes. Obviously, since the plots centred at the points of $\mathsf{U}_1 - \mathsf{U}_{1F}$ should lie completely or partially outside forest, no forest trees are likely to be found in these plots. Hence, it is convenient to suppose $\hat{T}_j = 0$ for any $j \in \mathsf{U}_1 - \mathsf{U}_{1F}$, in such a way that the sampling effort can be completely devoted to U_{1F} without detrimental effect on the estimation of

T. Under the last assumption, estimators (2.1) and (2.4) are equivalent to

$$\hat{\bar{T}}_1 = \frac{1}{M} \sum_{j \in \mathrm{U}_{1F}} \hat{T}_j \tag{2.5}$$

and

$$V_1^2 = \frac{1}{M^2} \sum_{j \in \mathrm{U}_{1F}} \hat{T}_j^2 - \frac{2}{M^2(M-1)} \sum_{h>j \in \mathrm{U}_{1F}} \hat{T}_j \hat{T}_h \tag{2.6}$$

Now, it is at once apparent that, conditional on the population of forest points U_{1F} (which, in turn, is univocally determined by U_1), expression (2.5) constitutes an unknown finite population total which can be estimate by a second phase of sampling. Accordingly, denote by $\mathsf{U}_2 \subset \mathsf{U}_{1F}$ the sample of size N selected from U_{1F} by means of a fixed-size scheme inducing first- and second-order inclusion probabilities τ_j and τ_{jh} ($h > j \in \mathsf{U}_{1F}$). Suppose also that $\tau_{jh} > 0$ for any $h > j \in \mathsf{U}_{1F}$. If the $\hat{T}_j$s were recorded for each $j \in \mathsf{U}_2$, then the double-expansion estimator (Särndal et al., 1992, chapter 9)

$$\hat{\bar{T}}_2 = \frac{1}{M} \sum_{j \in \mathsf{U}_2} \frac{\hat{T}_j}{\tau_j} \tag{2.7}$$

turned out to be unbiased with sampling variance

$$\begin{aligned} \mathrm{V}_{12}(\hat{\bar{T}}_2) &= \mathrm{E}_1 \left\{ \mathrm{V}_2(\hat{\bar{T}}_2 | \mathsf{U}_1) \right\} + \mathrm{V}_1 \left\{ \mathrm{E}_2(\hat{\bar{T}}_2 | \mathsf{U}_1) \right\} \\ &= \mathrm{E}_1 \left\{ \frac{1}{M^2} \sum_{h>j \in \mathsf{U}_{1F}} (\tau_j \tau_h - \tau_{jh}) \left(\frac{\hat{T}_j}{\tau_j} - \frac{\hat{T}_h}{\tau_h} \right)^2 \right\} + \mathrm{V}_1(\hat{\bar{T}}_1) \end{aligned} \tag{2.8}$$

where now $\mathrm{E}_2(\cdot|\mathsf{U}_1)$ and $\mathrm{V}_2(\cdot|\mathsf{U}_1)$ denote expectation and variance with re-

spect to the second phase of sampling, i.e. with respect to all the possible samples U_2 which can be selected by the second-phase scheme, conditional to the set of points U_1 selected in the first phase, while E_{12} and V_{12} denote expectation and variance with respect to both the sampling phases. Moreover, in not too difficult to prove that a conservative estimator for (2.8) was given by

$$V_2^2 = \frac{1}{M^2}\left\{\sum_{j\in\mathsf{U}_2}\frac{1-\tau_j}{\tau_j}\hat{T}_j^2 + 2\sum_{h>j\in\mathsf{U}_2}\frac{\tau_{jh}-\tau_j\tau_h}{\tau_j\tau_h}\frac{\hat{T}_j\hat{T}_h}{\tau_{jh}} + \sum_{j\in\mathsf{U}_2}\frac{\hat{T}_j^2}{\tau_j} - \frac{2}{M-1}\sum_{h>j\in\mathsf{U}_2}\frac{\hat{T}_j\hat{T}_h}{\tau_{jh}}\right\} =$$

$$\frac{1}{M^2}\left\{\sum_{j\in\mathsf{U}_2}\frac{\hat{T}_j^2}{\tau_j^2} + 2\sum_{h>j\in\mathsf{U}_2}\left(\frac{\tau_{jh}-\tau_j\tau_h}{\tau_j\tau_h} - \frac{1}{M-1}\right)\frac{\hat{T}_j\hat{T}_h}{\tau_{jh}}\right\} =$$

$$\frac{1}{M^2}\left\{\sum_{j\in\mathsf{U}_2}\frac{\hat{T}_j^2}{\tau_j^2} + 2\sum_{h>j\in\mathsf{U}_2}\frac{\hat{T}_j\hat{T}_h}{\tau_j\tau_h} - 2\frac{M}{M-1}\sum_{h>j\in\mathsf{U}_2}\frac{\hat{T}_j\hat{T}_h}{\tau_{jh}}\right\} \qquad (2.9)$$

in the sense that $\mathrm{E}_{12}(V_2^2) \geq \mathrm{V}_{12}(\hat{\bar{T}}_2)$. Usually, as suggested by Fattorini et al. (2006), in the second phase the population U_{1F} of the points aerially classified as forest is partitioned into L strata, say $\mathsf{U}_{1F(1)}, \ldots, \mathsf{U}_{1F(L)}$, of sizes $M_{F(1)}, \ldots, M_{F(L)}$, once again by using the very high resolution remotely sensed imagery available for the whole area. Then samples of points, say $\mathsf{U}_{2(1)}, \ldots, \mathsf{U}_{2(L)}$, of size $N_1, \ldots, N_L$ (with $N_1 + \ldots + N_L = N$) are selected from each stratum by means of SRSWOR. Accordingly, in this case the first- and second-order inclusion probabilities to be used into expressions (2.7) and

(2.9) are

$$\tau_j = N_l/M_{F(l)},\ j \in \mathsf{U}_{1F(l)}$$

and

$$\tau_{jh} = \begin{cases} \frac{N_l(N_l-1)}{M_{F(l)}(M_{F(l)}-1)} & j,\ h \in \mathsf{U}_{1F(l)} \\ \frac{N_l}{M_{F(l)}}\frac{N_{l'}}{M_{F(l')}} & j \in \mathsf{U}_{1F(l)},\ h \in \mathsf{U}_{1F(l')} \end{cases}$$

2.4 Three-phase estimation

Once again, owing to the costs and time involved, in real situations plot sampling cannot be performed for each second-phase point, but rather for a portion of these points selected in the third phase. Usually, all the N second-phase are visited on the ground, but only some qualitative and quantitative attribute are recorded without performing all the recording activities within plots. Accordingly, also the second-phase survey is only hypothetical and its treatment has had the sole aim of constructing the theoretical basis for the analysis of the three-phase estimators. As to the third phase, once all the second-phase points are visited, the second-phase sample U_2 may be partitioned into two sub-samples: the sub-sample $\mathsf{U}_{2F} = \{\mathbf{p}_j : \mathbf{p}_j \in \mathcal{F}\}$ of the N_F points actually lying in the forest class and the sub-sample $\mathsf{U}_2 - \mathsf{U}_{2F}$ of the $N - N_F$ points erroneously classified in the forest class by means of

aerial imagery but actually lying outside. Once again, no forest trees are likely to be found in the plots centred at the points of $\mathsf{U}_2 - \mathsf{U}_{2F}$. Hence, it is convenient to suppose $\hat{T}_j = 0$ for any $j \in \mathsf{U}_2 - \mathsf{U}_{2F}$, in such a way that the sampling effort can be completely devoted to U_{2F} without detrimental effect on the estimation of T. Under the last assumption, estimators (2.7) and (2.9) are equivalent to

$$\hat{\bar{T}}_2 = \frac{1}{M} \sum_{j \in \mathsf{U}_{2F}} \frac{\hat{T}_j}{\tau_j} \tag{2.10}$$

and

$$V_2^2 = \frac{1}{M^2} \left\{ \sum_{j \in \mathsf{U}_{2F}} \frac{\hat{T}_j^2}{\tau_j^2} + 2 \sum_{h > j \in \mathsf{U}_{2F}} \frac{\hat{T}_j \hat{T}_h}{\tau_j \tau_h} - 2 \frac{M}{M-1} \sum_{h > j \in \mathsf{U}_{2F}} \frac{\hat{T}_j \hat{T}_h}{\tau_{jh}} \right\} \tag{2.11}$$

Also in this case, it is at once apparent that, conditional on the second-phase sample of forest points U_{2F} (which, in turn, is univocally determined by U_2), expression (2.10) constitutes an unknown finite population total which can be estimate by a third phase of sampling. Accordingly, denote by $\mathsf{S} \subset \mathsf{U}_{2F}$ the sample of size n selected from U_{2F} by means of a fixed-size scheme inducing first- and second-order inclusion probabilities π_j and π_{jh} ($h > j = 1, \ldots, N_F$). Suppose also that $\tau_{jh} > 0$ for any $h > j = 1, \ldots, N_F$. If all the selected points were reached on the ground and the $\hat{T}_j$s were recorded for each $j \in \mathsf{U}$ (complete response), the triple-expansion estimator

$$\hat{\bar{T}}_3 = \frac{1}{M} \sum_{j \in \mathrm{S}} \frac{\hat{T}_j}{\tau_j \pi_j} \tag{2.12}$$

turned out to be unbiased with sampling variance

$$\mathrm{V}_{123}(\hat{\bar{T}}_3) = \mathrm{E}_{12}\left\{\mathrm{V}_3(\hat{\bar{T}}_3|\mathsf{U}_1,\mathsf{U}_2)\right\} + \mathrm{V}_{12}\left\{\mathrm{E}_3(\hat{\bar{T}}_3|\mathsf{U}_1,\mathsf{U}_2)\right.$$

$$= \mathrm{E}_{12}\left\{\frac{1}{M^2}\sum_{h>j\in U_{2F}}(\pi_j\pi_h - \pi_{jh})\left(\frac{\hat{T}_j}{\tau_j\pi_j} - \frac{\hat{T}_h}{\tau_h\pi_h}\right)^2\right\} + \mathrm{V}_{12}(\hat{\bar{T}}_2) \qquad (2.13)$$

where now $\mathrm{E}_3(\cdot|\mathsf{U}_1,\mathsf{U}_2)$ and $\mathrm{V}_3(\cdot|\mathsf{U}_1,\mathsf{U}_2)$ denote expectation and variance with respect to the third phase of sampling, i.e. with respect to all the possible samples S which can be selected by the third-phase scheme, conditional to the set of points U_1 selected in the first phase and to the sample U_2 selected in the second phase, while E_{123} and V_{123} denote expectation and variance with respect to all the three sampling phases. Moreover, a conservative estimator for (2.13) was given by

$$V_3^2 = \frac{1}{M^2}\left\{\sum_{j\in\mathrm{S}}\frac{1-\pi_j}{\pi_j^2}\frac{\hat{T}_j^2}{\tau_j^2} + 2\sum_{h>j\in\mathrm{S}}\frac{\pi_{jh}-\pi_j\pi_h}{\pi_j\pi_h}\frac{\hat{T}_j\hat{T}_h}{\tau_j\tau_h\pi_{jh}}\right\}$$

$$+\frac{1}{M^2}\left\{\sum_{j\in\mathrm{S}}\frac{\hat{T}_j^2}{\tau_j^2\pi_j} + 2\sum_{h>j\in\mathrm{S}}\frac{\hat{T}_j\hat{T}_h}{\tau_j\tau_h\pi_{jh}} - 2\frac{M}{M-1}\sum_{h>j\in\mathrm{S}}\frac{\hat{T}_j\hat{T}_h}{\tau_{jh}\pi_{jh}}\right\} \qquad (2.14)$$

in the sense that $\mathrm{E}_{123}(V_3^2) \geq \mathrm{V}_{123}(\hat{\bar{T}}_3)$. For the subsequent investigations it should be noticed that while the first term in (2.14) is an unbiased estimator of the first term in (2.13) (i.e. the portion of variance due to the third sampling phase), the second term in (2.14) is a conservative estimator of the second term in (2.13). Usually, as suggested by Fattorini et al. (2006), samples

$\mathsf{U}_{2F(l)}, \ldots, \mathsf{U}_{2F(L)}$ of size $N_{F(1)}, \ldots, N_{F(L)}$ (with $N_{F(1)} + \ldots + N_{F(L)} = N_F$) are obtained from the second-phase samples $\mathsf{U}_{2(1)}, \ldots, \mathsf{U}_{2(L)}$ by discarding those points erroneously classified in the forest class by means of aerial imagery but actually lying outside. Then, each second-phase sample $\mathsf{U}_{2F(l)}$ of size $N_{F(l)}$ is further partitioned into G strata, say $\mathsf{U}_{2F(l,1)}, \ldots, \mathsf{U}_{2F(l,G)}$, of sizes $N_{F(l,1)}, \ldots, N_{F(l,G)}$ (with $N_{F(l,1)} + \ldots + N_{F(l,G)} = N_{F(l)}$) and third-phase samples of points, say $\mathsf{S}_{F(l,1)}, \ldots, \mathsf{S}_{F(l,G)}$, of size $n_{(l,1)}, \ldots, n_{(l,G)}$ are selected from each stratum by means of SRSWOR ($l = 1, \ldots, L$). The G strata usually correspond to the G forest categories considered in the surveys. Accordingly, in this case the first- and second-order inclusion probabilities to be used into expressions (2.12) and (2.14) are

$$\pi_j = n_{(l,g)}/N_{F(l,g)}, \;\; j \in \mathsf{U}_{2F(l,g)}$$

and

$$\pi_{jh} = \begin{cases} \frac{n_{(l,g)}(n_{(l,g)}-1)}{N_{F(l,g)}(N_{F(l,g)}-1)} & j, \; h \in \mathsf{U}_{2F(\mathrm{l},g)} \\ \frac{n_{(l,g)}}{N_{F(l,g)}} \frac{n_{(l',g')}}{N_{F(l',g')}} & j \in \mathsf{U}_{F(l,g)}, \; h \in \mathsf{U}_{2F(l',g')} \end{cases}$$

Chapter 3

DESIGN-BASED NON RESPONSE TREATMENT IN FOREST SURVEYS

3.1 Recording and non response in three-phase forest surveys

Consider a three-phase forest survey such as that delineated in the previous section. After the first TSS phase, the M selected points are aerially classified: non forest points are discarded and a sample of N points, referred to as U_2, is selected from those points classified in the forest class. All the N points of U_2 are visited on the ground, points erroneously classified in the forest class by means of aerial imagery are discarded, while a set of qualitative and quantitative attributes are recorded for the N_F points actually lying in the forest class, previously referred to as U_{2F}. Usually, the recording of these variables does not necessitate the points to be reached in the ground.

For example some qualitative attributes such as forest category, type of ownership, temperature, altitude, crown cover, can be recorded some distance away from the point or even from the map. For these reasons, non response is supposed to be absent in the second sampling phase. Accordingly, denote by K_U the number of variables, say $X_1^*, \ldots, X_{K_U}^*$, recorded at each point of U_{2F} in such a way that, for each $j \in \mathsf{U}_{2F}$ a K_U vector $\mathbf{x}_j^*$ is available.

In the third phase, a sample S of size n is selected from U_{2F}. In this case all the points in S necessitate to be reached in order to delineate the plot of a pre-fixed radius and to record the interest variable (e.g. timber volume) for all the trees with certain characteristics lying within the plot. As already pointed out in Chapter 1, some third-phase points located in roughly terrain cannot be reached or, even if reached, the recording activities within the plot cannot be performed by foresters. In this case the sample S is split into two sub-samples: the respondent sample R in which the $\hat{T}_j$s are available for each $j \in \mathsf{R}$, and the non respondent sample S-R, in which the $\hat{T}_j$s are missing. Notwithstanding this, some quantitative and qualitative attributes such the management system (usually coppice or high forests) can be recorded for all the third-phase points without reaching the points. Accordingly, denote by K_S the number of variables, say $X_1^\circ, \ldots, X_{K_S}^\circ$, recorded at each point of S in such a way that, for each $j \in \mathsf{S}$ a K_S vector $\mathbf{x}_j^\circ$ is available.

3.2 Calibration approach under three-phase non response

In order to handle the third-phase non response in forest surveys, it is convenient to condition on the second-phase points selected in the second phase. Thus in this framework, the N_F points of U_{2F} play the role of the population U of Chapter 1, which in turn can be partitioned in two strata: the stratum $\mathsf{U}_{2F(R)}$ of the $N_{F(R)}$ points which can be reached and/or travelled in the field (respondent stratum) and the stratum $\mathsf{U}_{2F} - \mathsf{U}_{2F(R)}$ of the $N_F - N_{F(R)}$ points which cannot be reached and/or travelled in the field (non respondent stratum), while the quantities $\hat{T}_j/\tau_j$ for each $j \in \mathrm{U}_{2F}$ play the role of the y_js and the quantity

$$\hat{\bar{T}}_2 = \frac{1}{M} \sum_{j \in \mathrm{U}_{2F}} \frac{\hat{T}_j}{\tau_j} \tag{3.1}$$

plays the role of the interest parameters. Obviously, (3.1) will be estimated from the respondent sample of third-phase points $\mathsf{R} \subset \mathsf{S}$ by means of the calibration approach described in Chapter 1. To this purpose, the vectors $\mathbf{x}_j^*$ for $j \in \mathsf{U}_{2F}$ and the corresponding K_U vector of totals

$$\mathbf{T}^* = \sum_{j \in \mathrm{U}_{2F}} \mathbf{x}_j^*$$

play the role of Info U while the vectors $\mathbf{x}_j^\circ$ for $j \in \mathsf{S}$ and the corresponding K_S vector of HT estimates

$$\hat{\mathbf{T}}^\circ = \sum_{j \in \mathrm{S}} \frac{\mathbf{x}_j^\circ}{\pi_j}$$

play the role of Info S. Obviously, $\hat{\mathbf{T}}^{\circ}$ constitutes un unbiased estimator of the K_S vector of totals $\mathbf{T}^{\circ}$ conditional on U_{2F}, in the sense that $\mathrm{E}_3(\hat{\mathbf{T}}^{\circ}|\mathrm{U}_{2F}) = \mathbf{T}^{\circ}$ where

$$\mathbf{T}^{\circ} = \sum_{j \in \mathrm{U}_{2F}} \mathbf{x}_j^{\circ}$$

Also in this case, for simplicity of notation the two sets of variables are joined into a unique set of $K = K_U + K_S$ variables by using the K-vector $\mathbf{x}_j = \begin{bmatrix} \mathbf{x}_j^{*} \\ \mathbf{x}_j^{\circ} \end{bmatrix}$ as well as the K-vector $\hat{\mathbf{T}} = \begin{bmatrix} \mathbf{T}^{*} \\ \hat{\mathbf{T}}^{\circ} \end{bmatrix}$, in such a way that we can write $\mathrm{E}_3(\hat{\mathbf{T}}|\mathrm{U}_{2F}) = \mathbf{T}$ where $\mathbf{T} = \begin{bmatrix} \mathbf{T}^{*} \\ \mathbf{T}^{\circ} \end{bmatrix}$. It is worth noting that the variable which equals one for each $j \in \mathsf{U}_{2F}$ is included among the K_U auxiliary variables constituting the Info U in such a way that the simplified calibration weights of type (1.8) can be adopted.

Then, in accordance with expressions (1.10) and (1.11), the third-phase calibration estimator of $\hat{\bar{T}}_2$ turns out to be

$$\hat{\bar{T}}_{3CAL} = \frac{1}{M} \hat{\mathbf{b}}_R^{\mathrm{T}} \hat{\mathbf{T}} \tag{3.2}$$

where

$$\hat{\mathbf{b}}_R = \left(\sum_{j \in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi_j} \right)^{-1} \sum_{j \in \mathrm{R}} \frac{\hat{T}_j \mathbf{x}_j}{\tau_j \pi_j} \tag{3.3}$$

As pointed out in Section 1.4, if the hyperplane fitted from respondent population point scatter $\left\{ (\mathbf{x}_j, \hat{T}_j/\tau_j),\ j \in \mathsf{U}_{2FR} \right\}$ and that fitted from nonrespondent population point scatter $\left\{ (\mathbf{x}_j, \hat{T}_j/\tau_j),\ j \in \mathsf{U}_{2F} - \mathsf{U}_{2FR} \right\}$ are identical,

then $\hat{\bar{T}}_{3CAL}$ constitutes an approximately unbiased estimator of $\hat{\bar{T}}_2$ conditional on U_{2F}, in the sense that $\mathrm{E}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_{2F}) \approx \hat{\bar{T}}_2$. On the other hand, in accordance with Section 1.5, the approximate expression for the conditional variance of $\hat{\bar{T}}_{3CAL}$ turns out to be

$$\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathrm{U}_{2F}) \approx \frac{1}{M^2} \sum_{h>j \in \mathrm{U}_{2F}} (\pi_j \pi_h - \pi_{jh}) \left(\frac{r_j u_j}{\pi_j} - \frac{r_h u_h}{\pi_h} \right)^2 \tag{3.4}$$

where $u_j = e_{Rj} \mathbf{x}_j^{\mathrm{T}} \mathbf{A}_R^{-1} \mathbf{T} + (\hat{T}_j/\tau_j) - e^{\circ}_{Rj}$, e_{Rj} denotes the 0-sum residuals from the least-square fitting performed on the respondent population scatter, i.e. $e_{Rj} = \hat{T}_j/\tau_j - \mathbf{b}_R^T \mathbf{x}_j$ for $j \in \mathsf{U}_R$, $\mathbf{b}_R^{\circ}$ denotes the last K_S components of $\mathbf{b}_R$ and $e^{\circ}_{Rj} = \hat{T}_j/\tau_j - \mathbf{b}_R^{\circ \mathrm{T}} \mathbf{x}_j^{\circ}$ for $j \in \mathsf{U}_{2F(R)}$ are the non-0-sum residuals from the fitting obtained neglecting the Info-U-variable coefficients of $\mathbf{b}_R$. In turn, $\mathbf{b}_R = \mathbf{A}_R^{-1} \mathbf{a}_R$ and

$$\mathbf{A}_R = \sum_{j \in \mathrm{U}_{2F(R)}} \mathbf{x}_j \mathbf{x}_j^T$$

$$\mathbf{a}_R = \sum_{j \in \mathrm{U}_{2F(R)}} \frac{\hat{T}_j \mathbf{x}_j}{\tau_j}$$

Alternatively, expression (3.4) can be rewritten as

$$\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathrm{U}_{2F}) \approx \frac{1}{M^2} \left(\sum_{j \in \mathrm{U}_{2F(R)}} \frac{1-\pi_{\mathrm{j}}}{\pi_{\mathrm{j}}} u_j^2 + \sum_{h>j \in \mathrm{U}_{2F(R)}} \frac{\pi_{jh} - \pi_{\mathrm{j}} \pi_h}{\pi_{\mathrm{j}} \pi_h} u_j u_h \right) \tag{3.5}$$

Since the population U_{2F} is univocally determined by the population U_2 of the points selected in the second phase, which in turn depends on the population U_1 of the points selected in the first phase, there is no difference

in writing $\mathrm{E}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_1,\mathsf{U}_2)$ and $\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_1,\mathsf{U}_2)$ instead of $\mathrm{E}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_{2F})$ and $\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_{2F})$ in accordance with the notation adopted in Chapter 2. Thus, it is at once apparent that $\hat{\bar{T}}_{3CAL}$ is unconditionally unbiased, in the sense that $\mathrm{E}_{123}(\hat{\bar{T}}_{3CAL}) = T$ providing that it is conditionally unbiased, i.e. $\mathrm{E}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_1,\mathsf{U}_2) = \hat{\bar{T}}_2$. Moreover the unconditional variance of $\hat{\bar{T}}_{3CAL}$ turns out to be

$$\mathrm{V}_{123}(\hat{\bar{T}}_3) = \mathrm{E}_{12}\left\{\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_1,\mathsf{U}_2)\right\} + \mathrm{V}_{12}(\hat{\bar{T}}_2) \tag{3.6}$$

where $\mathrm{V}_3(\hat{\bar{T}}_{3CAL}|\mathsf{U}_1,\mathsf{U}_2)$ is given by expressions (3.4) or (3.5) while the second term is given by expression (2.8).

3.3 Variance estimation

Now a problem arises in estimating the unconditional variance of type (3.6). Obviously, in accordance with section 1.5,

$$V^2_{3SYG} = \frac{1}{M^2}\sum_{h>j\in\mathrm{S}} \frac{\pi_j\pi_h - \pi_{jh}}{\pi_{jh}}\left(\frac{r_j\hat{u}_j}{\pi_j} - \frac{r_h\hat{u}_h}{\pi_h}\right)^2 \tag{3.7}$$

or

$$V^2_{3HT} = \frac{1}{M^2}\left(\sum_{j\in\mathrm{R}}\frac{1-\pi_j}{\pi_j^2}\hat{u}_j^2 + 2\sum_{h>j\in\mathrm{R}}\frac{\pi_{jh}-\pi_\mathrm{j}\pi_h}{\pi_\mathrm{j}\pi_h\pi_{jh}}\hat{u}_j\hat{u}_h\right) \tag{3.8}$$

both constitute unbiased estimators of the first term of (3.6), where $\hat{u}_j = \hat{e}_{Rj}\mathbf{x}_j^{\mathrm{T}}\hat{\mathbf{A}}_R^{-1}\hat{\mathbf{T}} + \hat{\mathbf{b}}_R^{\circ\mathrm{T}}\mathbf{x}_j^{\circ}$ are the empirical influence values computed for each $j \in \mathsf{R}$, $\hat{e}_{Rj} = \hat{T}_j/\tau_j - \hat{\mathbf{b}}_R^T\mathbf{x}_j$ are the residual achieved from the least-square

fitting performed on the respondent point scatter $\left\{(\mathbf{x}_j, \hat{T}_j/\tau_j),\, j \in \mathsf{R}\right\}$ and $\hat{\mathbf{b}}_R^{\circ}$ denotes the last K_S components of $\hat{\mathbf{b}}_R$, while $\hat{\mathbf{b}}_R = \hat{\mathbf{A}}_R^{-1}\hat{\mathbf{a}}_R$ and

$$\hat{\mathbf{A}}_R = \sum_{j\in \mathrm{R}} \frac{\mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}}{\tau_j \pi_j}$$

$$\hat{\mathbf{a}}_R = \sum_{j\in \mathrm{R}} \frac{\hat{T}_j \mathbf{x}_j}{\tau_j \pi_j}$$

Moreover,

$$V_{3jack}^2 = \sum_{j\in \mathrm{S}} (1-\pi_{\mathrm{j}}) v_{(j)}^2 + 2 \sum_{h>j\in \mathrm{S}} \frac{\pi_{jh} - \pi_{\mathrm{j}}\pi_h}{\pi_{jh}} v_{(j)} v_{(h)} \tag{3.9}$$

constitutes the jackknife estimator of the first term of (3.6) where in this case

$$v_{(j)} = \left(1 - \frac{1}{\hat{N}\pi_j}\right)\left\{\hat{T}_{3CAL} - \hat{T}_{3CAL(j)}\right\}$$

$$\hat{N}_F = \sum_{j\in \mathrm{S}} \frac{1}{\pi_j}$$

$$\hat{T}_{3CAL(j)} = \frac{1}{M}\hat{\mathbf{b}}_{R(j)}^{\mathrm{T}} \hat{\mathbf{T}}_{(j)}$$

$$\hat{\mathbf{b}}_{R(j)} = \left(\sum_{h\in \mathrm{S}_{-j}} \frac{r_h \mathbf{x}_h \mathbf{x}_h^{\mathrm{T}}}{\pi_h}\right)^{-1} \sum_{h\in \mathrm{S}_{-j}} \frac{r_h \hat{T}_h \mathbf{x}_h}{\tau_h \pi_h}$$

$$\hat{\mathbf{T}}_{(j)} = \left[T_1^*, \ldots, T_{K_U}^*, \hat{T}_{1(j)}^{\circ}, \ldots, \hat{T}_{K_S(j)}^{\circ}\right]^{\mathrm{T}}$$

$$\hat{T}_{k(j)}^{\circ} = \frac{N_F}{\hat{N}_F} \sum_{h\in \mathrm{S}_{-j}} \frac{x_{h,k}^{\circ}}{\pi_h} \quad , \quad k = 1, \ldots, K_S$$

and finally S_{-j} consists of the sample S with the j-th unit deleted.

Unfortunately, as to the second term of (3.6), the second term of (2.14), say

$$V_{12}^2 = \frac{1}{M^2}\left\{\sum_{j\in \mathrm{S}} \frac{\hat{T}_j^2}{\tau_j^2 \pi_j^2} + 2 \sum_{h>j\in \mathrm{S}} \frac{\hat{T}_j \hat{T}_h}{\tau_j \tau_h \pi_{jh}} - 2\frac{M}{M-1} \sum_{h>j\in \mathrm{S}} \frac{\hat{T}_j \hat{T}_h}{\tau_{jh}\pi_{jh}}\right\} \tag{3.10}$$

which would constitute a conservative estimator of the second term of (3.6) under complete response is actually unknown. Indeed, owing to non response, the $\hat{T}_j$s are known only for $j \in \mathsf{R}$. Obviously, the estimator restricted to respondent sample R, say

$$V^2_{12R} = \frac{1}{M^2}\left\{\sum_{j\in \mathrm{R}} \frac{\hat{T}_j^2}{\tau_j^2\pi_j^2} + 2\sum_{h>j\in \mathrm{R}} \frac{\hat{T}_j\hat{T}_h}{\tau_j\tau_h\pi_{jh}} - 2\frac{M}{M-1}\sum_{h>j\in \mathrm{R}} \frac{\hat{T}_j\hat{T}_h}{\tau_{jh}\pi_{jh}}\right\} \quad (3.11)$$

cannot be adopted since all the three terms involved in (3.11) would constitute negatively biased estimator of the population counterparts. Accordingly, the three quantities involved in (3.11) would necessitate a calibration approach like the one adopted for estimating T, thus rendering the estimation procedure quite cumbersome. However, it should be noticed that the bias of $\hat{\hat{T}}_{3CAL}$ is negligible and hence the estimation of variance makes sense only if the linear relation among $\mathbf{x}_j$s and $(\hat{T}_j/\tau_j)$s is similar in respondent and non response populations. Accordingly, variance estimation can be straightforwardly performed by estimating/imputing the missing values via the linear model fitted from the respondent sample R. In other word, for each $j \in \mathsf{S\text{-}R}$ the missing $(\hat{T}_j/\tau_j)$s are imputed via $\hat{\mathbf{b}}_R^{\mathrm{T}}\mathbf{x}_j$, in such a way that the quantity (3.10) can be readily computed.

Then, in accordance with the estimator adopted for the first term of (3.6), the following variance estimator are available

$$V^2_{SYG} = V^2_{3SYG} + \hat{V}^2_{12(IMP)} \quad (3.12)$$

$$V_{HT}^2 = V_{3HT}^2 + \hat{V}_{12(IMP)}^2 \tag{3.13}$$

$$V_{jack}^2 = V_{3jack}^2 + \hat{V}_{12(IMP)}^2 \tag{3.14}$$

where $\hat{V}_{12(IMP)}^2$ is given by(3.10) with $\hat{T}_j/\tau_j = \hat{\mathbf{b}}_R^{\mathrm{T}}\mathbf{x}_j$ when $j \in \mathsf{S} - \mathsf{R}$.

3.4 Simulation study

In order to evaluate the effectiveness of the calibration procedure in three-phase forest inventories a simulation study was performed. A quadrat study area $\mathcal{A}$ of side 20 km was presumed for a total size $|\mathcal{A}| = 40,000$ ha. The altitude on each point $\mathbf{p} \in \mathcal{A}$ of the study area, say $h(\mathbf{p})$ was obtained from the mixture of 18 probability density functions of bivariate normal distributions with different means vectors and variance-covariance matrices in such a way that altitude values ranged from 0 to 2,000 m.. The resulting surface is represented in Figure 1. The forest portion $\mathcal{F}$ within $\mathcal{A}$ was presumed to depend on the altitude: forest of type 1 was supposed to be settled on zones where the altitude ranged from 300 to 1,000 m for the portion of the study area below the line $p_2 = 2 - 1.65p_1$ and on zones where the altitude ranged from 300 to 600 m for the portion between the lines $p_2 = 2 - 1.65p_1$ and $p_2 = 1.5 - 3.5p_1$; forest of type 2 was supposed to be settled on zones where the altitude ranged from 600 to 1,000 m for the portion of the study

area between the lines $p_2 = 2 - 1.65p_1$ and $p_2 = 1.5 - 3.5p_1$ and on zones where the altitude ranged from 300 to 600 m for the portion over the line $p_2 = 1.5 - 3.5p_1$. Forest was presumed to be absent below 300 m and over 1,000 m. The resulting sizes for the forests of type 1 and 2 were of 6,248 and 7,748 ha respectively, for a total size of 13,996 ha corresponding to about the 35% of the study area. The resulting forest portions within the study area are represented in Figure 2.

For each point $\mathbf{p} \in \mathcal{A}$ the slope, say $z(\mathbf{p})$, was presumed to be a linear function of the altitude perturbed by a periodic function. More precisely it was presumed

$$z(\mathbf{p}) = H\left\{32h(\mathbf{p}) + 1.5h(\mathbf{p})\varphi(\mathbf{p})\right\} \tag{3.15}$$

where

$$\varphi(\mathbf{p}) = \frac{\alpha(p_1) + \alpha(p_2) + \alpha(p_1p_2)}{6}$$

$$\alpha(p) = \sin\left(\frac{\pi}{2} - \pi p\right) + \cos(\pi p)$$

and H was a normalizing constant ensuring that $z(\mathbf{p})$ ranged from 0 to 80 degrees. Figure 3 shows the relationship achieved via (3.15) between slope and altitude for a network of 10,000 points on the study area.

Within the forest area a population of trees was randomly settled in accordance with varying densities. Type 1 forests with altitudes ranging 300 to 600 m were presumed to have the highest density of 1,200 trees per ha, thus

achieving a total of 5,227,200 trees; type 2 forests with altitudes ranging from 300 to 600 m and type 1 forests with altitudes ranging from 600 to 1,000 m was presumed to have a density of 1,000 per ha, with a total of 4,276,000 trees; type 2 forests with altitudes ranging from 600 to 1,000 m were presumed to have the smallest density of 800 trees per ha, with a total of 4,291,200 trees. Accordingly a population of 13,794,400 was randomly settled within the forest zones. The population abundance was presumed to be the parameter under estimation, which obviously coincided with the total T of an interest variable equal to 1 for each tree in the population.

Finally, in order to simulate nonresponse, it was presumed that points with slope greater than 40% cannot be reached by foresters or, if reached, the recording activities within the plots centered at these points cannot be performed. Figure 4 maps the forest zones with different densities and the zones which cannot be reached. As these zones contains 459,353 trees, about 3.3% of the forest trees cannot be sampled.

From the resulting scenario, the selection of 1,000 third phase samples was simulated. At first, in accordance with the protocol of TSS sampling, the study area was partitioned into M=1,600 quadrats of size 25 ha. Then, in each simulation run, a point is randomly selected within each quadrat, thus achieving the set U_1 of first-phase points. Points of U_1 are classified in accordance with their position, and only the set of M_F forest points U_{1F} was con-

sidered for the subsequent sampling phases. Then, a set U_2 of $N = 0.25M_F$ was selected from U_1 by means of simple random sample without replacement. Accordingly, $\tau_j = 0.25$ and $\tau_{jh} = 0.25\,(N-1)\,/(M_F-1)$ for each $h > j \in \mathsf{U}_1$. As no error in aerial imaging classification was presumed, the second-phase sample U_2 coincided with the sample U_{2F} achieved after the second-phase points were visited on the ground, since no points aerially classified as forest point was discarded. Obviously, in this case $N = N_F$. For each second phase points, the forest type (type 1 or 2) was adopted as stratification variables, determining two strata, say $\mathsf{U}_{2(1)}$ and $\mathsf{U}_{2(2)}$ of size $N_{(1)}$ and $N_{(2)}$ while the altitude was adopted as the auxiliary variable to be used as Info U in the third-phase calibration approach. Finally, in the third phase two samples $\mathsf{S}_{(1)}$ and $\mathsf{S}_{(2)}$ of size $n_{(1)} = 0.3N_{(1)}$ and $n_{(2)} = 0.3N_{(2)}$ respectively, were selected from $\mathsf{U}_{2(1)}$ and $\mathsf{U}_{2(2)}$ by means of simple random sampling without replacement.

Accordingly, $\pi_j = 0.3$ and $\tau_{jh} = 0.3(n_{(1)}-1)/(N_{(1)}-1)$ for each $h > j \in \mathsf{U}_{2(1)}$, $\tau_{jh} = 0.3(n_{(2)}-1)/(N_{(2)}-1)$ for each $h > j \in \mathsf{U}_{2(2)}$ and $\tau_{jh} = 0.09$ for each $j \in \mathsf{U}_{2(1)}$ and $h \in \mathsf{U}_{2(2)}$. Finally, the set of third-phase points having slope greater than 40% were discarded, and the remaining points constituted the respondent sample R. Thus, for each point $j \in \mathsf{R}$, the number of trees within the circle of radius 13.8 mt (size 600 mt^2) centred at the point, say $\#(\mathrm{P_j})$,

was considered and the quantity

$$\hat{T}_j = \frac{|\mathcal{A}|}{a}\#(\mathrm{P_j})$$

was computed.

Then, for each simulation run, the calibration estimate $\hat{\bar{T}}_{3CAL}$ of the population abundance was achieved by means of (3.4), adopting $\mathbf{x}_j = [1, x_j]^{\mathrm{T}}$ for $j \in \mathsf{U}_2$ as auxiliary information, where x_j denotes the altitude for unit $j \in \mathsf{U}_2$. Moreover, the three alternative variance estimates V^2_{SYG}, V^2_{HT} and V^2_{jack} were obtained by using (3.12), (3.13) and (3.14). From the variance estimates the corresponding estimates of the relative standard error $RSE_{SYG} = V_{SYG}/\hat{\bar{T}}_{3CAL}$, $RSE_{HT} = V_{HT}/\hat{\bar{T}}_{3CAL}$ and $RSE_{jack} = V_{jack}/\hat{\bar{T}}_{3CAL}$ were also computed together with the confidence intervals $\hat{\bar{T}}_{3CAL} \pm 1.96V_{SYG}$, $\hat{\bar{T}}_{3CAL} \pm 1.96V_{HT}$, and $\hat{\bar{T}}_{3CAL} \pm 1.96V_{jack}$. As benchmarks, the complete-sample HT estimate achieved if all the points in S were visited, say

$$\hat{\bar{T}}_{3R} = \frac{1}{M}\sum_{j\in \mathrm{S}} \frac{\hat{T}_j}{\tau_j \pi_j}$$

was also computed together with the HT estimate based on the sole sample R, say

$$\hat{\bar{T}}_{3R} = \frac{1}{M}\sum_{j\in \mathrm{R}} \frac{\hat{T}_j}{\tau_j \pi_j}$$

Then, from the resulting Monte Carlo distributions of these quantities, the relative bias of $\hat{\bar{T}}_{3CAL}$ turns out to be negligible, being equal to -0.46% with

a relative root mean squared error of 3.04%. Interestingly, the calibration estimator performed equivalently even with the complete sample HT estimator which was unbiased with a relative root mean squared error of 2.96%. On the other hand, the HT estimator based on the sole sample R shows a considerable downward relative bias of -4.30% and a relative root mean squared error of 3.95%. As to the variance estimators, they reveal highly conservative, as the averages of the relative standard error estimators are 4.86% (SYG), 4.88% (HT) and 4.36% (jack) against a true value of 3.04%, while the coverage of the corresponding confidence interval were invariably greater than 99% against a nominal level of 95%.

Simulation results give positive insights on the effectiveness of the calibration estimator in reducing the bias induced by nonresponse also providing performance comparable with that achieved from complete sample as well as conservative evaluations of the accuracy.

Figure 3.1: Representation of the altitude surface of the study area (m).

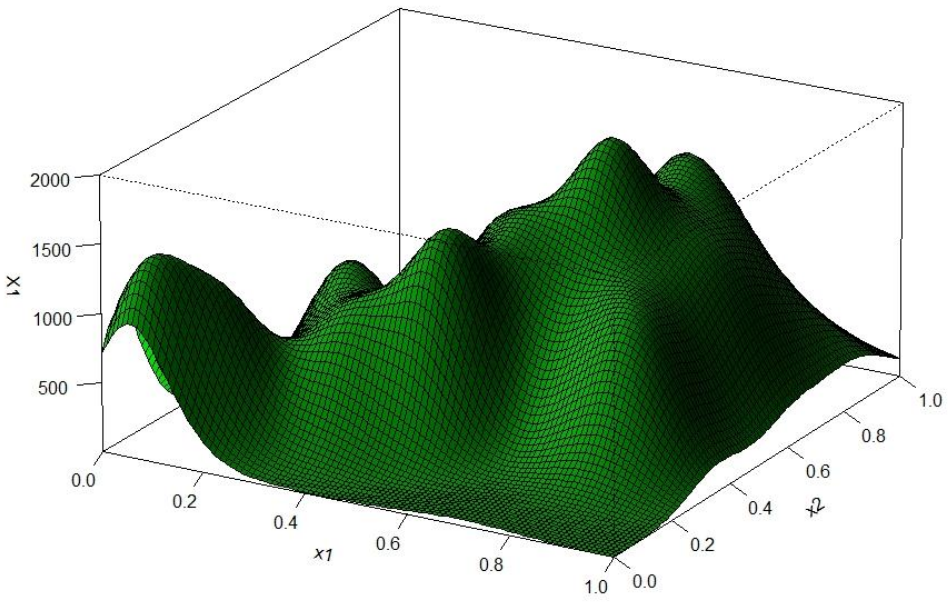


Figure 3.2: Forest portions of type 1 and 2 within the study area.

Figure 3.3: Relationship between slope (Z) and altitude (H) observed in network of 10,000 points onto the study area.

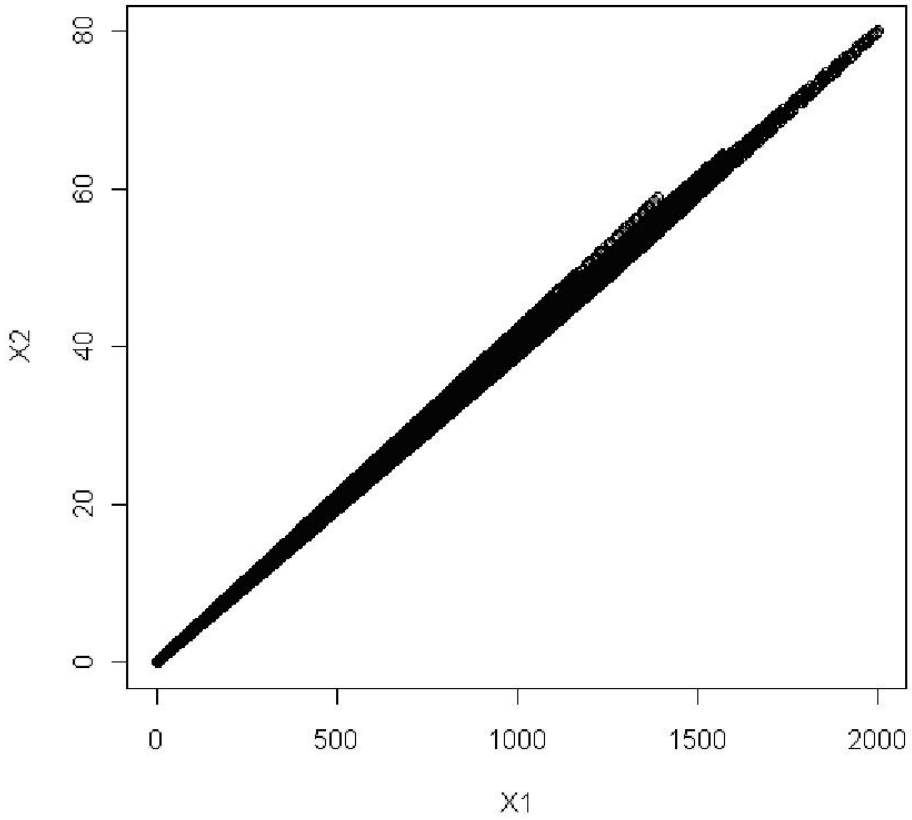
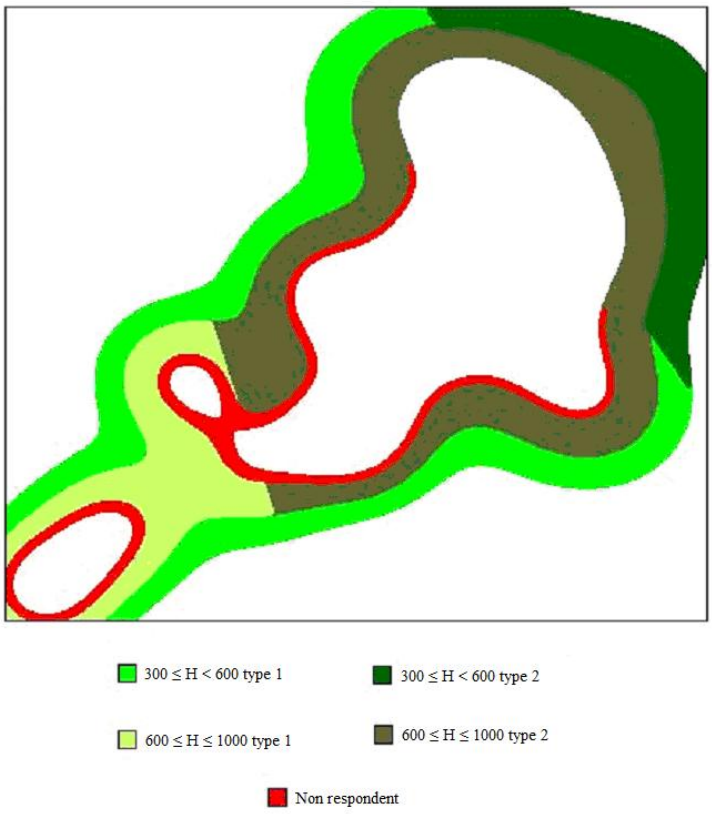


Figure 3.4: Map of forest zone with different densities and of zones with slopes greater than 40% which cannot be sampled.



Chapter 4

A CASE STUDY: THE ESTIMATION OF TIMBER VOLUME IN THE FORESTS OF TRENTINO (NORTH ITALY)

4.1 The Italian National Forest Inventory

The Italian National Forest Inventory (INFC) was a three-phase sample survey of forest resources on the whole Italy. The inventory operations started in 2003 and were concluded in 2006. In the first phase of sampling, the Italian territory of size $30,132,846$ *ha* was covered by a grid of 306,831 quadrats of size 100 *ha*. The grid was constructed by starting from a point of coordinates (4,750,000 N, 2,354,000 E) from the geographic reference system Gauss Boaga Rome 40. Then, a random point was selected in each quadrat, giving rise to a population of 306,831 first-phase points. In the second phase, on the

basis of aerial photos, the population of first phase points was partitioned into 12 strata and a sample of 26,480 points was selected from the three forest strata using simple random sampling without replacement within each forest stratum. Subsequently each points selected in the second phase was visited and points erroneously classified in forest strata were discarded, while forest points were classified by ground inspection into one of the 18 forest categories on the basis of the characteristics of the surrounding areas. More precisely, a point was aerially assigned to a forest category if it established that it laid within a homogeneous area of the same type of size greater than 0.5 ha . The 18 forest categories constituted the strata from which third-phase points were selected by means of simple random sampling without replacement, with a sampling fraction of about 30%. For those third-phase sample points which were reached by forester teams, a circular plot of radius 13.8 mt (size 600 mt^2) was centered at the point. Then, the diameters of all the trees in the plot with a circumference greater than 7.85 cm were recorded and the corresponding biomasses will be predicted on the basis of previously constructed equations linking diameters to biomass. For more detail see Fattorini et al (2006) and the web site http://www.infc.it.

4.2 Inventory data from Trentino administrative district

On the basis of the scheme adopted for performing INFC, the administrative district of Trentino (North Italy) was covered by a grid of $14,115$ quadrats. Thus, the population U_1 of first-phase points randomly selected within each quadrat was constituted by $M = 14,115$ units. In the second phase these points were aerially classified in forest and non forest strata. Among them a total of $M_F = 4,266$ (30.2%) points was classified in the three forest strata. These points constituted the population U_{1F} from which the sample of second phase points was selected. Table 1 reports structure of this population and the sizes of the sample selected from each stratum by means of simple random sampling without replacement and the corresponding inclusion probabilities.

Table 4.1: Stratification of first-phase forest points together with second-phase sample sizes and inclusion probabilities within strata

Stratum	stratum size($M_{F(l)}$)	sample size (N_l)	inclusion probability(τ_j)
FORMFOR	4106	1137	0.2769
FORMRAD	12	10	0.8333
INCLASS	148	30	0.2027
Total	4266	1177	

After the second phase of sampling, the population of second-phase points U_2 was constituted by $N = 1,177$ points. All these points were visited on

the ground. Among them 139 (12%) points were discarded since erroneously classified in forest classes, while the remaining N_F points were classified into 17 forest categories and constituted the population U_{2F} from which third-phase points were selected. Table 2 reports structure of this population and the sizes of the sample selected from each stratum by means of simple random sampling without replacement and the corresponding inclusion probabilities.

Table 4.2: Stratification of second-phase forest points together with third-phase sample sizes and inclusion probabilities within strata

Stratum	stratum size ($N_{F(l)}$)	sample size (n_l)	inclusion probability (π_j)
FORMFOR b01	164	37	0.2256
FORMFOR b02	378	96	0.2540
FORMFOR b03	44	18	0.4091
FORMFOR b04	59	18	0.3051
FORMFOR b05	17	13	0.7647
FORMFOR b08	170	39	0.2294
FORMFOR b09	16	10	0.6250
FORMFOR b11	5	4	0.8000
FORMFOR b12	113	21	0.1858
FORMFOR b13	9	9	1.0000
FORMFOR b14	42	17	0.4048
FORMFOR b15	1	1	1.0000
FORMFOR b24	7	6	0.8571
FORMRAD b12	1	1	1.0000
INCLASS b01	8	3	0.3750
INCLASS b02	2	2	1.0000
INCLASS b08	2	2	1.0000
Total	1038	297	

It is worth noting that the values of altitude were available for each of the 1038 points of U_{2F} and as such they could be adopted as Info U in the

calibration procedure. After the third phase of sampling, the sample S was constituted by $n = 297$ points. All these points were re-visited on the ground. Among them a set R of 293 (98.7%) points were reached by foresters while the remaining 4 points were not reached or, if reached, were judged unfeasible for performing the recording operation within plots. Accordingly, the $\hat{T}_j$s (which in this case were achieved by 235,350 times the total of timber volumes recorded within the plots) were missing for each $j \in \mathsf{R}$.

Figure 4.1 reports the frequency distribution of the total timber volumes within respondent plots while Figure 4.2 reports the frequency distribution of altitude in respondent and nonrespondent sample.

4.3 Calibration estimation of timber volume

The altitude ranges are quite similar in respondent and nonrespondent sample, thus, in accordance with the considerations reported in Chapter 1, it seems feasible to use the altitude as calibration variable, besides the dummy variable equal to 1 for each $j \in \mathsf{U}_{2F}$. While the HT estimate based on the sole sample of respondent give rise to a value of $105,371,784m^3$, the calibration estimator turns out to be 1.01 times greater, giving rise to an estimate value of $106,695,917m^3$. As to the estimated accuracy of the calibration procedure, results are quite satisfactory, as the estimates of relative standard

error adopting Sen-Yates-Grundy, Horvitz-Thompson and jackknife estimator to estimate the variance due to the first two phases are $R\hat{S}E_{SYG} = 5.90\%$, $R\hat{S}E_{HT} = 4.59\%$ and $R\hat{S}E_{JACK} = 4.68\%$, respectively.

Figure 4.1: Frequency distribution of total timber volume in the respondent sample.

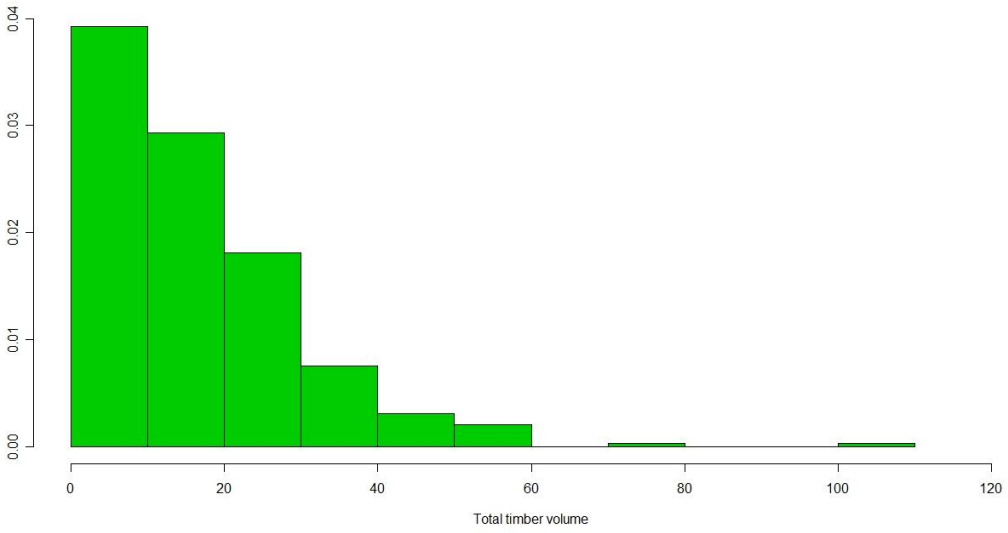
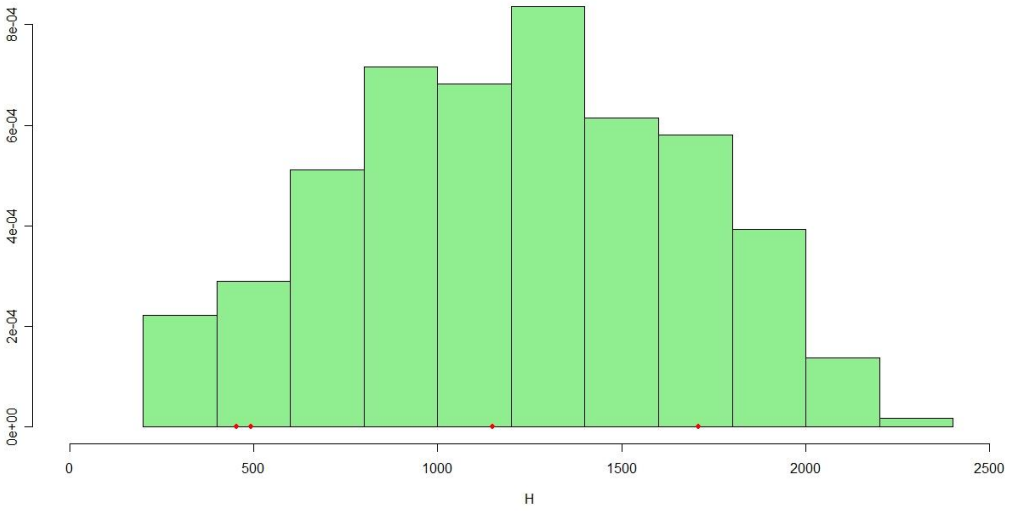


Figure 4.2: Frequency distribution of altitude in respondent sample (nonrespondent sample are represented by red points).



Bibliography

[1] Barabesi L. (2003). A Monte Carlo integration approach to Horvitz-Thompson estimation in replicated environmental designs. *Metron* **61**, 355-374.

[2] Barabesi L., Marcheselli M. (2003). A modified Monte Carlo integration. *International Mathematical Journal* **3**, 555-565.

[3] Barabesi L., Marcheselli M. (2008). Improved strategies for coverage estimation by using replicated line-intercept sampling. *Environmental and Ecological Statistics* **15**, 139-215.

[4] Berger Y.G., Skinner C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society* B, **67**, 79-89.

[5] Cordy C.B., Thompson C.M. (1995). An application of deterministic variogram to design-based variance estimation. *Mathematical Geology* **27**, 173-205.

[6] Davison A., Hinkley D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge, Cambridge University Press.

[7] Durrant G.B. (2005). *Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review*. NCRM Methods Review Papers NCRM/002, ESRC National Centre for Research Methods, University of Southampton (UK).

[8] Fattorini L., Marcheselli M., Pisani C. (2006). A three-phase sampling strategy for large-scale multiresources forest inventories. *Journal of Agricoltural, Biological and Environmental Statistics* **11**, 1-21.

[9] Gregoire T.G., Valentine H.T. (2008). *Sampling Strategies for Natural Resources and the Evironment*. Chapman & Hall, New York.

[10] Haber, S. (1966). A modified Monte-Carlo quadrature. *Mathematical Computations* **20**, 361-368.

[11] Haziza D., Thompson K.J., Yung W. (2010). The effect of nonresponse adjustments on variance estimation. *Survey Methodology*, **36**, 35-43.

[12] Kott P.S. (1994). A note on handling nonresponse in sample survey. *Journal of the American Statistical Association*, **89**, 693-696.

[13] Little R.J.A., Rubin D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York, Wiley.

[14] Oh H.L., Scheuren F.J. (1983). Weighting adjustment for unit nonresponse. In WG Madow, I Olkin, DB Rubin (eds). *Incomplete Data in Sample Surveys*, Vol 2. New York, Academic Press, pp.143-184.

[15] Overton W.S., Stehman S.V. (1993). Properties of designs for sampling continuous spatial resources from a triangular grid. *Communications in Statistics -Theory and Methods* **22**, 2641-2660.

[16] Särndal C.E., Lundström S. (2005). *Estimation in Survey with Nonresponse*. New York, Wiley.

[17] Särndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. New York, Springer-Verlag.

[18] Stevens D.L. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics* **8**, 167-195.

[19] Stevens D.L. (2006). Spatial properties of design-based versus model-based approaches to environmental sampling. In: Caetano M., Paino M. (eds.). *Proceedings of 7th International Symposium on Spatial Accuracy Assessment of Natural Resources and Environmental Sciences. Lisbona* pp.119-125.

[20] Wolter, K.M., 1985. *Introduction to variance estimation*. Springer-Verlag, New York.

Ringraziamenti

Se non avessi avuto una famiglia forte e solida come la mia non avrei potuto né intraprendere, né continuare il mio percorso di studi. Ai miei genitori e ad Agnese vanno la mia riconoscenza e il mio affetto per essere stati sempre e comunque presenti con attenzione, facendo di me la loro priorità, nonostante la lontananza, nonostante tutto.

Ringrazio anche i mie zii Rosetta e Mimmo e miei cugini Luca e Mino per avermi seguita e incoraggiata con infinito affetto.

Un grazie sentito ai docenti che ho avuto modo di conoscere durante gli anni di studio. In ordine di comparsa, ma non di importanza, Rosa Capobianco e Stefano Maria Pagnotta, che ho avuto modo di apprezzare durante la stesura della mia tesi di laurea, Anna Clara Monti, senza la quale non avrei pensato di intraprendere il "viaggio del Dottorato", tutti i docenti del Dipartimento di Statistica "G. Parenti" di Firenze, che hanno tenuto le lezioni dei corsi del Dottorato e Timothy Gregoire, che mi ha ospitata con gentilezza nella Yale School of Forestry & Environmental Studies durante il mio soggiorno negli

Stati Uniti.

Grazie soprattutto al mio Relatore Lorenzo Fattorini per avermi ospitata nell'Università di Siena, seguita con pazienza, costanza e disponibilità e per avermi aperto una finestra sull'affascinante mondo della statistica ambientale.

Ringrazio di cuore Sara Franceschi per l'inestimabile aiuto nelle simulazioni e per l'amicizia dimostratami insieme a Federica Baffetta durante il mio periodo di permanenza a Siena.

Un pensiero va anche alle persone che avrebbero potuto esserci.

Infine, ringrazio i colleghi di Dottorato per il tempo trascorso insieme, i vecchi e i nuovi amici, per i tanti momenti e pensieri condivisi e tutte le persone straordinarie che ho avuto modo di conoscere negli ultimi tre anni.