Design-based treatment of missing data in two-phase forest inventories by using canopy heights from laser scanning

P. Corona[(1)], G. Chirici[(2)], S. Franceschi[(3)], D. Maffei[(4)],
M. Marcheselli[(3)], C. Pisani[(3)], L. Fattorini[(3)]

(1) Agricoltural Research Council (CRA), Forestry Research Centre, Arezzo (Italy)

(2) Department of Sciences and Technologies for the Environment, University of Molise, Contrada Fonte Lappone, Pesche, Isernia (Italy)

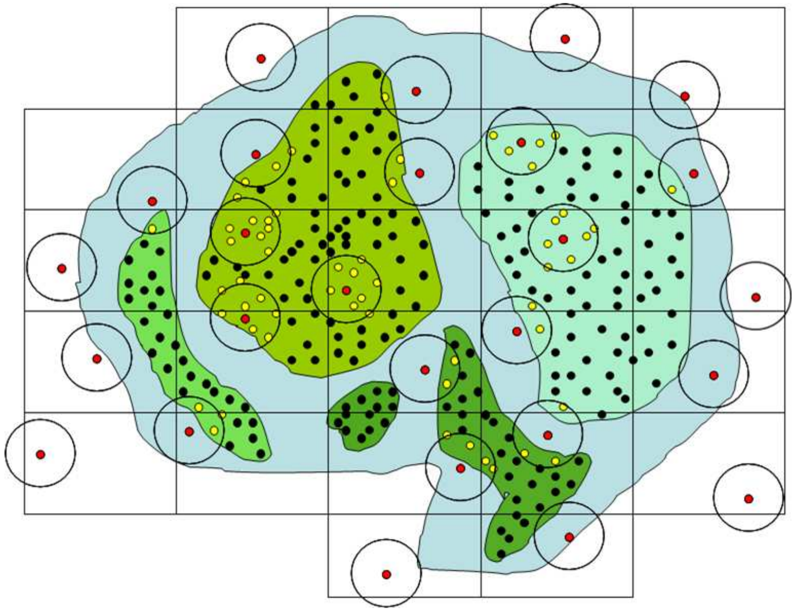
(3) Department of Economics and Statistics, University of Siena (Italy)

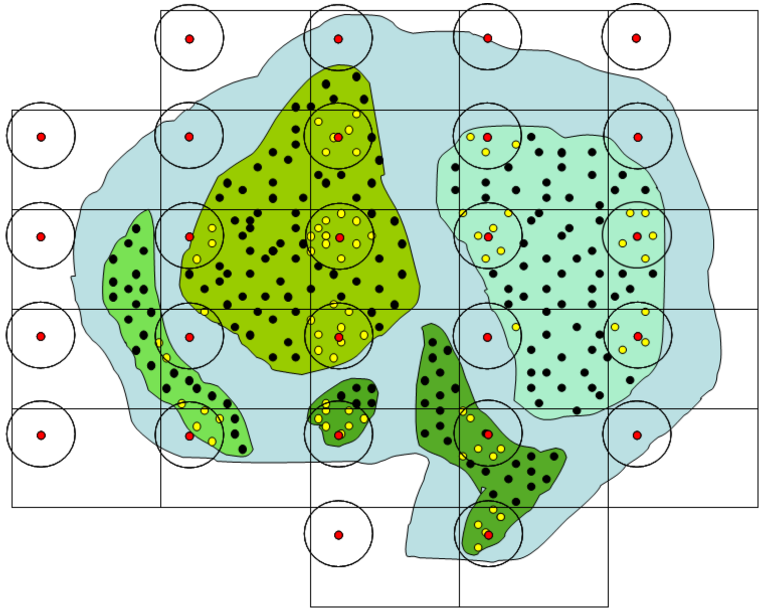
(4) Department of Statistics "G. Parenti", University of Florence (Italy)

Most forest surveys performed over large scale, such as national forest inventories involve two phases of sampling:

Most forest surveys performed over large scale, such as national forest inventories involve two phases of sampling:

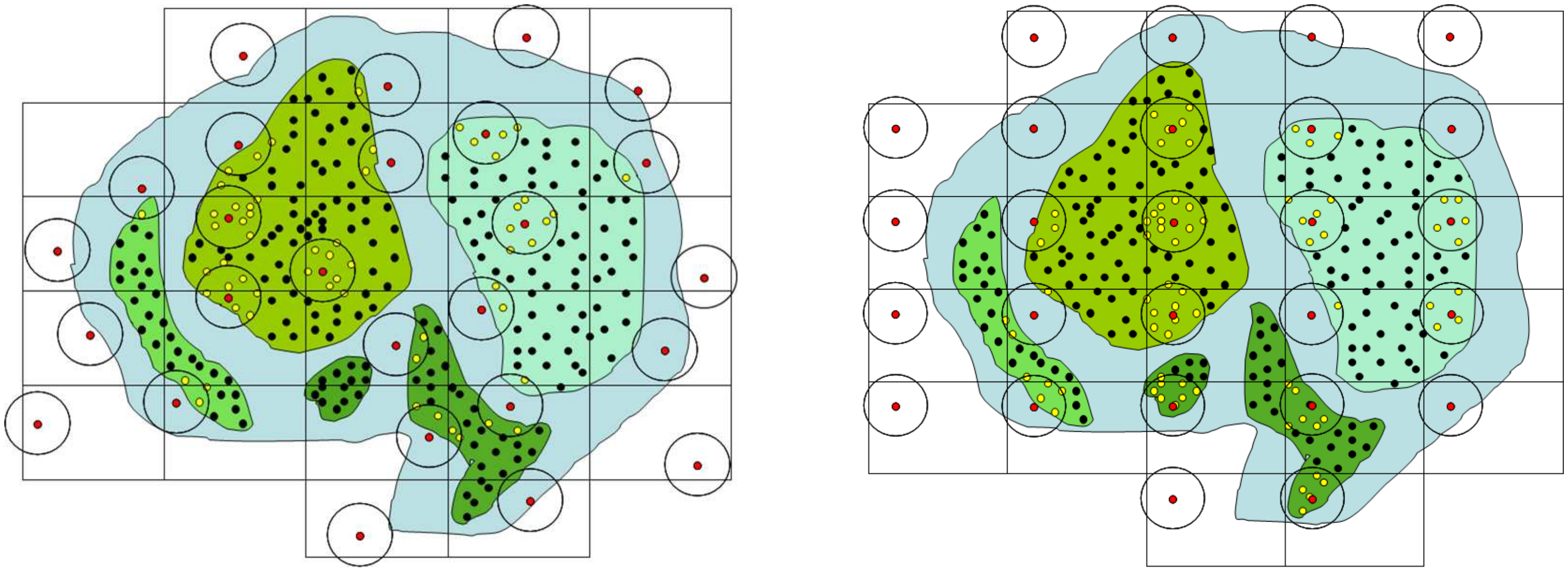
- in the first phase the area is partitioned into regular polygons of the same size and points are randomly or systematically thrown in each polygon.





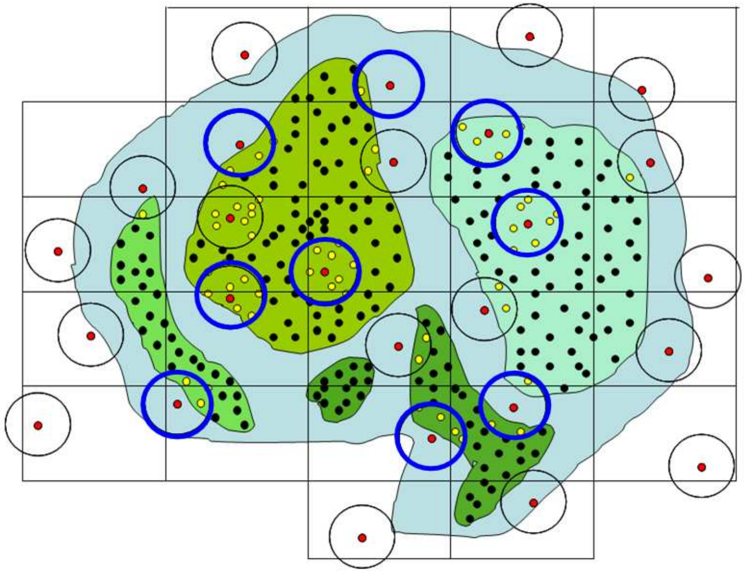
Most forest surveys performed over large scale, such as national forest inventories involve two phases of sampling:

- in the first phase the area is partitioned into regular polygons of the same size and points are randomly or systematically thrown in each polygon.

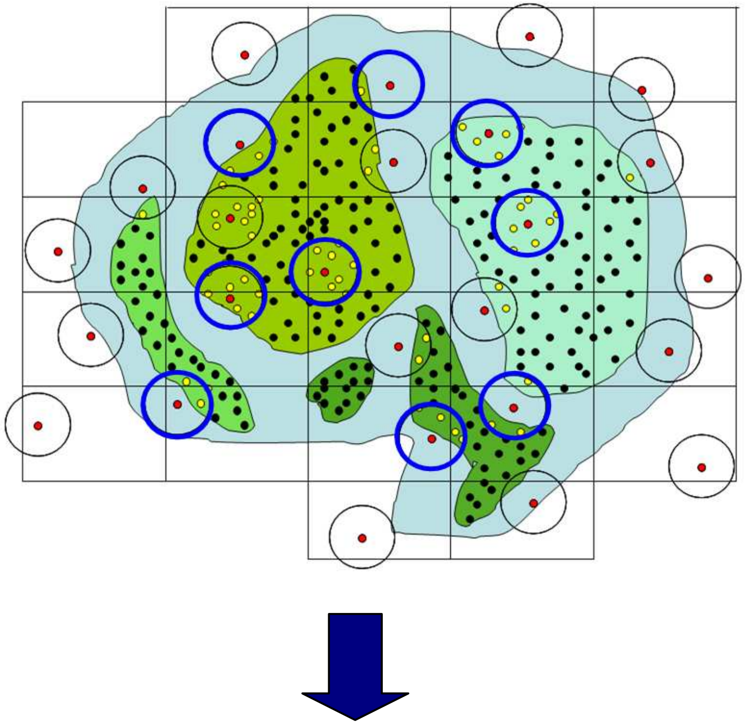


remote sensing information (photo-interpreted land cover class, elevation, thematic mapping spectral bands etc.) is recorded for each first-phase point

- in the second phase, a sample of first-phase points is selected in accordance with a probabilistic sampling scheme



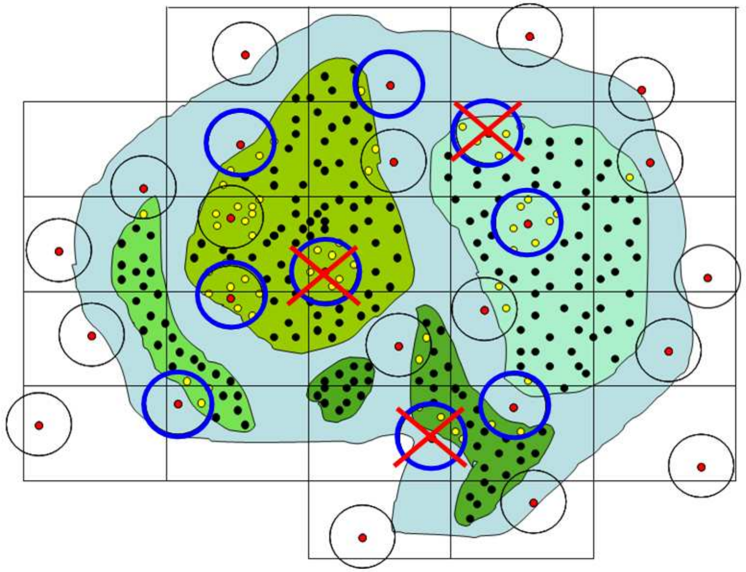
- in the second phase, a sample of first-phase points is selected in accordance with a probabilistic sampling scheme



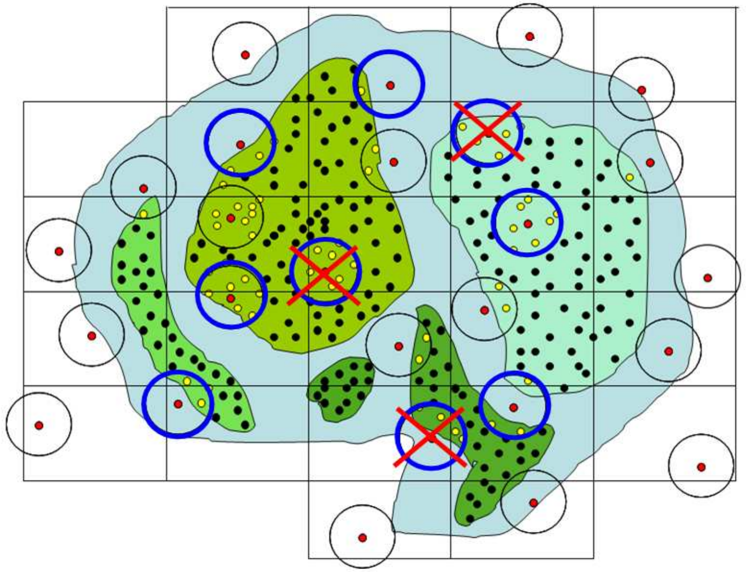
the second-phase points are visited on the ground in order to record several variables (actual land use class, forest category, total wood volume, tree basal area and biomass etc.) within plots of prefixed size centred at these points

Nonresponse mostly occurs when some second-phase points are located in difficult terrains and cannot be reached by the survey crew or, even if reached, the steep slope of the terrain does not allow the recording activities within the plot

Nonresponse mostly occurs when some second-phase points are located in difficult terrains and cannot be reached by the survey crew or, even if reached, the steep slope of the terrain does not allow the recording activities within the plot



Nonresponse mostly occurs when some second-phase points are located in difficult terrains and cannot be reached by the survey crew or, even if reached, the steep slope of the terrain does not allow the recording activities within the plot



- some attempts for treating nonresponse in forest inventory are proposed by McRoberts (2003) and Scott et al. (2004)

- from a more general point of view, a vast statistical literature deals with the problem of nonresponse

Nonresponse Propensity Weighting:

a random response mechanism is assumed in such a way that each population unit has its own (invariably positive) response probability and responds independently to the others

Nonresponse Propensity Weighting:

a random response mechanism is assumed in such a way that each population unit has its own (invariably positive) response probability and responds independently to the others

In forest inventories unit responses cannot be viewed as outcomes of dichotomous and independent experiments:

- if some points cannot be reached, no random experiment can be claimed as these points will never be reached
- neighbouring points, lying in terrains with the same characteristics, tend to have the same response pattern

Nonresponse Propensity Weighting:

a random response mechanism is assumed in such a way that each population unit has its own (invariably positive) response probability and responds independently to the others

In forest inventories unit responses cannot be viewed as outcomes of dichotomous and independent experiments:

- if some points cannot be reached, no random experiment can be claimed as these points will never be reached

- neighbouring points, lying in terrains with the same characteristics, tend to have the same response pattern

the use of nonresponse propensity weighting in forest surveys does not seem to be logically defensible

Imputation Techniques (regression imputation, nearest neighbour imputation, hot deck imputation and multiple imputation):

missing values are replaced by substitutes, the imputed values, which are usually obtained by means of a prediction model presuming a relationship among the interest variable and a set of variables and estimation is performed on the completed data

Imputation Techniques (regression imputation, nearest neighbour imputation, hot deck imputation and multiple imputation):

missing values are replaced by substitutes, the imputed values, which are usually obtained by means of a prediction model presuming a relationship among the interest variable and a set of variables and estimation is performed on the completed data

But prediction models cannot be validated in the set of nonrespondents

it is difficult to scientifically defend any proposed method/model of imputation

As both nonresponse propensity weighting and imputation techniques are not enough convincing for treating nonresponse in forest inventories, a complete design-based treatment of nonresponse, viewing interest attributes and nonresponse as fixed characteristics, is considered

As both nonresponse propensity weighting and imputation techniques are not enough convincing for treating nonresponse in forest inventories, a complete design-based treatment of nonresponse, viewing interest attributes and nonresponse as fixed characteristics, is considered

Fattorini et al. (2013) propose the use of a technique recently referred to as *nonresponse calibration weighting* (Haziza et al., 2010).

Nonresponse Calibration Weighting (NCW):

the weights originally attached to each respondent unit are modified into new weights able to estimate the population means of a set of auxiliary variables without error

Rationale if a relationship exists between the interest and the auxiliary variables, the calibration weights should also be suitable for estimating the population mean of the interest variable

Nonresponse Calibration Weighting (NCW):

the weights originally attached to each respondent unit are modified into new weights able to estimate the population means of a set of auxiliary variables without error

Rationale if a relationship exists between the interest and the auxiliary variables, the calibration weights should also be suitable for estimating the population mean of the interest variable

Remarks

- NCW does not need to refer explicitly to any model, allowing for a straightforward design-based treatment

Nonresponse Calibration Weighting (NCW):

the weights originally attached to each respondent unit are modified into new weights able to estimate the population means of a set of auxiliary variables without error

Rationale if a relationship exists between the interest and the auxiliary variables, the calibration weights should also be suitable for estimating the population mean of the interest variable

Remarks

- NCW does not need to refer explicitly to any model, allowing for a straightforward design-based treatment
- NCW both reduces nonresponse bias and ensures consistency when the relationships among interest and auxiliary variables are similar in respondents and nonrespondents

Nonresponse Calibration Weighting (NCW):

the weights originally attached to each respondent unit are modified into new weights able to estimate the population means of a set of auxiliary variables without error

Rationale if a relationship exists between the interest and the auxiliary variables, the calibration weights should also be suitable for estimating the population mean of the interest variable

Remarks

- NCW does not need to refer explicitly to any model, allowing for a straightforward design-based treatment
- NCW both reduces nonresponse bias and ensures consistency when the relationships among interest and auxiliary variables are similar in respondents and nonrespondents
- NCW can even increase the accuracy of estimation with respect to the complete-sample estimation when a close linear relationship exists among interest and auxiliary variables

Second-phase calibration estimator

T parameter of interest: total of an attribute (wood volume, basal area, etc.)

R second-phase respondent sample (points allowing recording activities)

$\mathbf{x}_j = \left[x_{j1}, \ldots, x_{jK} \right]^{\mathrm{T}}$ values of K auxiliary variables recorded on the j-th point

$$\hat{\hat{T}}_{2CAL} = \hat{\mathbf{b}}_R^{\mathrm{T}} \overline{\mathbf{X}}$$

where

$\overline{\mathbf{X}}$ is the mean vector of the auxiliary variables in the first-phase sample

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

π_j first-order inclusion probabilities

$\hat{T}_j$ total estimate in the j-th plot

How to select auxiliary information:

the auxiliary variables should be chosen in such a way that their relationship (linear or not, intense or not) with the interest variable is similar in respondents and nonrespondents

How to select auxiliary information:

the auxiliary variables should be chosen in such a way that their relationship (linear or not, intense or not) with the interest variable is similar in respondents and nonrespondents

In recent years airborn laser scanning is increasingly being applied in forest inventories, providing measurements of the height of upper canopy for the surveyed area:

- canopy height data are often available at low or even no cost

- a close relationship has been proven between the timber volume (or tree biomass) of the inventory plots and the canopy height model (CHM) data obtained from ALS surveys (e.g. Corona and Fattorini, 2008, Gregoire et al., 2011, Corona et al., 2012)

- the relationship is likely to hold irrespective of the fact that plots can be reached or not

How to select auxiliary information:

the auxiliary variables should be chosen in such a way that their relationship (linear or not, intense or not) with the interest variable is similar in respondents and nonrespondents

In recent years airborn laser scanning is increasingly being applied in forest inventories, providing measurements of the height of upper canopy for the surveyed area:

- canopy height data are often available at low or even no cost

- a close relationship has been proven between the timber volume (or tree biomass) of the inventory plots and the canopy height model (CHM) data obtained from ALS surveys (e.g. Corona and Fattorini, 2008, Gregoire et al., 2011, Corona et al., 2012)

- the relationship is likely to hold irrespective of the fact that plots can be reached or not

the exploitation of CHM data as auxiliary variable under the NCW approach seems to be a suitable estimation strategy

Simulation Study

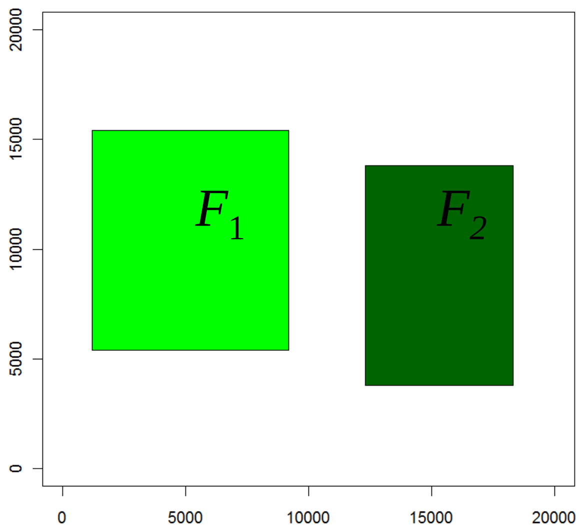
In order to evaluate the use of CHM data as auxiliary variable under the NCW approach a simulation study was performed

The artificial population:

- a quadrat study region of side 20 km was assumed

- the forest portion was : - the 35% of the whole area

  - constituted by two rectangles corresponding to two different forest categories (size 8000ha and 6000ha)



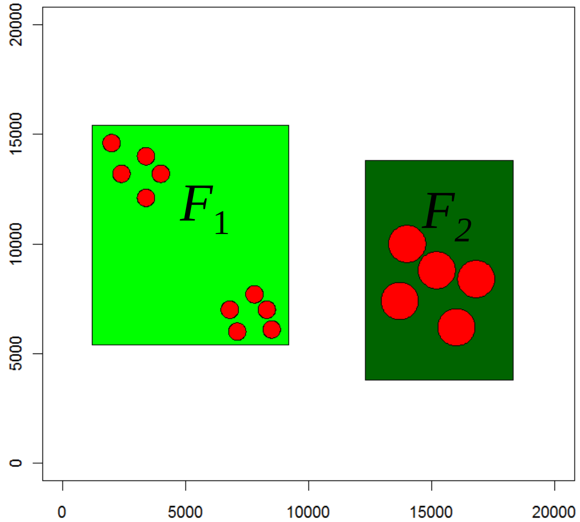
Simulation Study

In order to evaluate the use of CHM data as auxiliary variable under the NCW approach a simulation study was performed

The artificial population:

- a quadrat study region of side 20 km was assumed
- the forest portion was : - the 35% of the whole area

  - constituted by two rectangles corresponding to two different forest categories (size 8000ha and 6000ha)



Non respondent pattern:

- F_1 nonresponse area of 5%
- F_2 nonresponse area of 15%

The generation of CHM data and volume values:

- F_1 and F_2 were partitioned into a discrete population of 80 and 60 millions of pixels of size 1 mq labelled by a couple of integers identifying their position in the study region

- for each pixel the canopy height within was obtained from the mixture of 20 bivariate normal probability density functions with different mean vectors and variance-covariance matrices

- in order to represent a forest coverage of about 40% in F_1 and a forest coverage of about 90% in F_2 about 60% and 10% of the heights were respectively set to 0 by means of a mathematical function

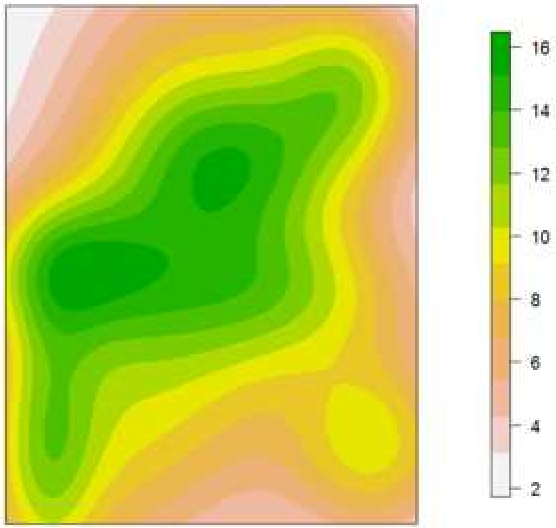
The generation of CHM data and volume values:

- F_1 and F_2 were partitioned into a discrete population of 80 and 60 millions of pixels of size 1 mq labelled by a couple of integers identifying their position in the study region

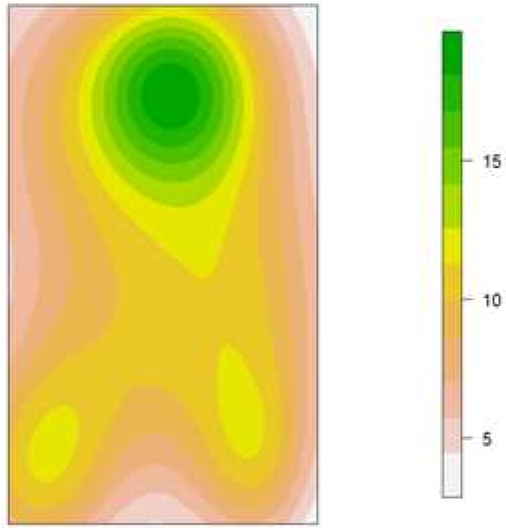
- for each pixel the canopy height within was obtained from the mixture of 20 bivariate normal probability density functions with different mean vectors and variance-covariance matrices

- in order to represent a forest coverage of about 40% in F_1 and a forest coverage of about 90% in F_2 about 60% and 10% of the heights were respectively set to 0 by means of a mathematical function

- in accordance with the results of some empirical investigations (Bortolot and Wynne, 2005), for each pixel the tree volume was presumed to be a linear function of the squared canopy height, perturbed by a periodic function



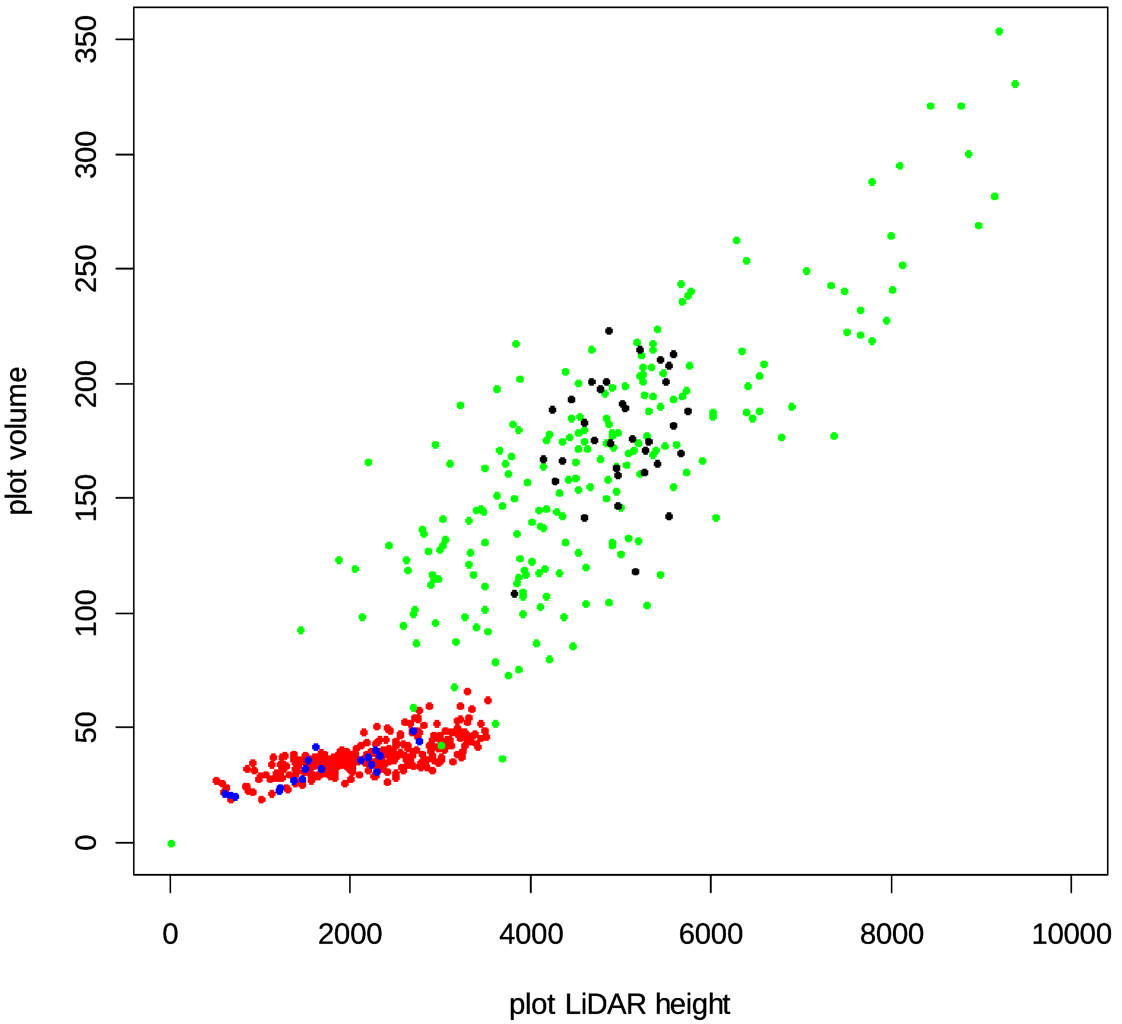
Region F_1



Region F_2

	F_1	F_2
Total canopy height	319 999 999.7 m	540 000 000.0 m
Total Volume	5 639 462.9 m^3	19 042 212.3 m^3
Maximum canopy height at pixel level	16.47 m	19.66 m

Canopy height vs volume at plot level in a sample of first-phase points



Survey simulation:

- 1,000 two-phase inventories were simulated

At each run:

- the study region was partitioned into quadrats of size 25 ha
- the first phase sampling was implemented by randomly (TSS) or systematically (SGS) selecting a point within each quadrat

Survey simulation:

- 1,000 two-phase inventories were simulated

At each run:

- the study region was partitioned into quadrats of size 25 ha
- the first phase sampling was implemented by randomly (TSS) or systematically (SGS) selecting a point within each quadrat

- in the second phase, the first-phase points were partitioned on the basis of their position into 3 strata:
  - the stratum of points falling outside the two forest regions
  - the stratum of points falling in F_1
  - the stratum of points falling in F_2

Survey simulation:

- 1,000 two-phase inventories were simulated

At each run:

- the study region was partitioned into quadrats of size 25 ha
- the first phase sampling was implemented by randomly (TSS) or systematically (SGS) selecting a point within each quadrat

- in the second phase, the first-phase points were partitioned on the basis of their position into 3 strata:
  - the stratum of points falling outside the two forest regions
  - the stratum of points falling in F_1
  - the stratum of points falling in F_2

  From each forest stratum a sample of second-phase points was selected by means of SWSROR (sampling fraction equals to 0.25)

Survey simulation:

- 1,000 two-phase inventories were simulated

At each run:

- the study region was partitioned into quadrats of size 25 ha
- the first phase sampling was implemented by randomly (TSS) or systematically (SGS) selecting a point within each quadrat

- in the second phase, the first-phase points were partitioned on the basis of their position into 3 strata:
  - the stratum of points falling outside the two forest regions
  - the stratum of points falling in F_1
  - the stratum of points falling in F_2

From each forest stratum a sample of second-phase points was selected by means of SWSROR (sampling fraction equals to 0.25)

Finally, the respondent sample was derived discarding the points falling within the nonresponse areas and for each respondent point a circular plot - centred at the point - of radius 13 mt was considered

Survey simulation – The auxiliary variables

Two alternative choices of the auxiliary variables were considered:

1) Two auxiliary variables: the “intercept” and the total canopy height of the pixels in the selected plot

Survey simulation – The auxiliary variables

Two alternative choices of the auxiliary variables were considered:

1) Two auxiliary variables: the "intercept" and the total canopy height of the pixels in the selected plot

$$\Downarrow$$

$$\hat{\hat{T}}_{2CAL}$$

Survey simulation – The auxiliary variables

Two alternative choices of the auxiliary variables were considered:

1) Two auxiliary variables: the “intercept” and the total canopy height of the pixels in the selected plot

$$\Downarrow$$

$$\hat{\hat{T}}_{2CAL}$$

2) Four auxiliary variables: the “intercept” and the total canopy height of the pixels in the selected plot taking into account the different forest categories

Survey simulation – The auxiliary variables

Two alternative choices of the auxiliary variables were considered:

1) Two auxiliary variables: the "intercept" and the total canopy height of the pixels in the selected plot

$$\hat{\hat{T}}_{2CAL}$$

2) Four auxiliary variables: the "intercept" and the total canopy height of the pixels in the selected plot taking into account the different forest categories

$$\hat{\hat{T}}_{2CAL,F_1F_2}$$

Simulation results

As benchmarks, the following two estimators were also computed:

- $\hat{\hat{T}}_{2HT}$ the complete-sample estimator achieved if all the second-phase points were visited

- $\hat{\hat{T}}_{2R}$ the estimator based on the sole sample of respondents

The following performance indicators were considered:

- relative bias (RB)

- relative root mean squared error (RRMSE)

- expectation of the relative standard error estimators (ERSEE)

- coverage of the confidence intervals at the nominal level of 95% (COV95)

Simulation results

	TSS		SGS	
	RB	RRMSE	RB	RRMSE
$\hat{\hat{T}}_{2HT}$	0.1%	2.9%	-0.1%	2.8%
$\hat{\hat{T}}_{2R}$	-12.8%	13.5%	-13.2%	13.7%
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	-1.3%	2.8%
$\hat{\hat{T}}_{2CAL,F_1F_2}$	-0.2%	2.2%	-0.2%	2.0%

Simulation results

	TSS		SGS	
	RB	RRMSE	RB	RRMSE
$\hat{\hat{T}}_{2HT}$	0.1%	2.9%	-0.1%	2.8%
$\hat{\hat{T}}_{2R}$	-12.8%	13.5%	-13.2%	13.7%
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	-1.3%	2.8%
$\hat{\hat{T}}_{2CAL,F_1F_2}$	-0.2%	2.2%	-0.2%	2.0%

✓ the estimator based on the sole respondent sample shows a considerable downward relative bias

Simulation results

	TSS		SGS	
	RB	RRMSE	RB	RRMSE
$\hat{\hat{T}}_{2HT}$	0.1%	2.9%	-0.1%	2.8%
$\hat{\hat{T}}_{2R}$	-12.8%	13.5%	-13.2%	13.7%
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	-1.3%	2.8%
$\hat{\hat{T}}_{2CAL,F_1F_2}$	-0.2%	2.2%	-0.2%	2.0%

✓ the estimator based on the sole respondent sample shows a considerable downward relative bias

✓ $\hat{\hat{T}}_{2CAL}$ shows a remarkable reduction in the relative bias compared with the sole respondent estimator

Simulation results

	TSS		SGS	
	RB	RRMSE	RB	RRMSE
$\hat{\hat{T}}_{2HT}$	0.1%	2.9%	-0.1%	2.8%
$\hat{\hat{T}}_{2R}$	-12.8%	13.5%	-13.2%	13.7%
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	-1.3%	2.8%
$\hat{\hat{T}}_{2CAL,F_1F_2}$	-0.2%	2.2%	-0.2%	2.0%

✓ the estimator based on the sole respondent sample shows a considerable downward relative bias

✓ $\hat{\hat{T}}_{2CAL}$ shows a remarkable reduction in the relative bias compared with the sole respondent estimator

✓ the relative bias of the calibration estimator $\hat{\hat{T}}_{2CAL,F_1F_2}$ turns out to be negligible, with a performance comparable with that of the complete sample estimator

Simulation results

TSS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\hat{T}}_{2CAL,F1F2}$	-0.2%	2.2%	5.1%	5.0%	5.1%	1.00	1.00	1.00

SGS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\hat{T}}_{2CAL}$	-1.3%	2.8%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\hat{T}}_{2CAL,F1F2}$	-0.2%	2.0%	5.0%	5.0%	5.1%	1.00	1.00	1.00

Simulation results

TSS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\hat{T}}_{2CAL}$	-1.3%	3.0%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\hat{T}}_{2CAL,F1F2}$	-0.2%	2.2%	5.1%	5.0%	5.1%	1.00	1.00	1.00

SGS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\hat{T}}_{2CAL}$	-1.3%	2.8%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\hat{T}}_{2CAL,F1F2}$	-0.2%	2.0%	5.0%	5.0%	5.1%	1.00	1.00	1.00

✓as to the variance estimators, they reveal to be highly conservative

Simulation results

TSS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\tilde{T}}_{2CAL}$	-1.3%	3.0%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\tilde{T}}_{2CAL,F1F2}$	-0.2%	2.2%	5.1%	5.0%	5.1%	1.00	1.00	1.00

SGS first-phase sampling scheme

			ERSEE			COV95		
	RB	RRMSE	SYG	HT	Jack	SYG	HT	Jack
$\hat{\tilde{T}}_{2CAL}$	-1.3%	2.8%	5.3%	5.5%	5.4%	1.00	1.00	1.00
$\hat{\tilde{T}}_{2CAL,F1F2}$	-0.2%	2.0%	5.0%	5.0%	5.1%	1.00	1.00	1.00

✓as to the variance estimators, they reveal to be highly conservative

✓ coverage of the confidence interval were invariably greater than the nominal value

References

Barabesi, L., Franceschi, S. 2011. Sampling properties of spatial total estimators under tessellation stratified designs. Environmetrics 22: 271–278.

Bortolot, Z.J., Wynne, R.H. 2005. Estimating forest biomass using small footprint LiDAR data: An individual tree-based approach that incorporates training data. ISPRS Journal of Photogrammetry & Remote Sensing 59: 342– 360.

Corona, P., Fattorini, L. 2008. Area-based LiDAR-assisted estimation of forest standing volume. Canadian Journal of Forest Research 38: 2911-2916.

Corona, P., Cartisano, R., Salvati, R., Chirici, G., Floris, A., Di Martino, P., Marchetti, M., Scrinzi, G., Clementel, F., Travaglini, D., and Torresan, C. 2012. Airborne Laser Scanning to support forest resource management under alpine, temperate and Mediterranean environments in Italy. European Journal of Remote Sensing 45: 27-37.

Fattorini, L., Franceschi, S., Maffei, D. 2013. Design-based treatment of unit nonresponse in environmental surveys using calibration weighting. Submitted

Gregoire, T.G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., Holm, S. 2011. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research 41: 83-95

Haziza, D., Thompson, K.J., Yung, W. 2010. The effect of nonresponse adjustments on variance estimation. Survey Methodology 36: 35-43

Mc Roberts, R.E. 2003. Compensating for missing plot observations in forest inventory estimation. Canadian Journal of Forest Research 33: 1990-1997

References

Mc Roberts, R.E. 2003. Compensating for missing plot observations in forest inventory estimation. Canadian Journal of Forest Research 33: 1990-1997

Opsomer, J.D., Breidt, F.J., Gretchen, G.M., Kauermann, G. 2007. Model-assisted estimation of forest resources with generalized additive models. Journal of the American Statistical Association 102: 400-409

Scott, C.T., Bechtold, W.A., Reams, G.A., Smith, W.D., Hansen, M.H., Moisen, G.G. 2004. Sample-based estimators utilized by the Forest Inventory and Analysis National Information Management System. In: *The Enhanced Forest Inventory and Analysis Programmational Sampling Design and Estimation Procedures*, W.A. Bechtold and P.L. Patterson (eds). Asheville (NC), Department of Agricolture Forest Service, Southern Research Station: 43-68

Survey simulation – The use of the auxiliary variable

Denote by h_j the CHM height within the j-th plot (sum of the heights within the pixels belonging to the plot)

Two alternative choices:

1)

in such a way that $\overline{\mathbf{X}}$ is the mean vector whose components are:

- fraction of first-phase points falling in forest regions
- average CHM height of the first-phase points falling in forest regions

$\hat{\hat{T}}_{2CAL}$

2) $\mathbf{x}_j = \left[I_{F_1}(j), I_{F_2}(j), I_{F_1}(j)h_j, I_{F_2}(j)h_j \right]^{\mathrm{T}}$

in such a way that $\overline{\mathbf{X}}$ is the mean vector whose components are:

- fraction of first-phase points falling in F_1
- fraction of first-phase points falling in F_2
- average CHM height of the first-phase points falling in F_1
- average CHM height of the first-phase points falling in F_2

$\hat{\hat{T}}_{2CAL,F_1F_2}$