Efficient allocation of field sampling utilizing forest resource maps

Juha Heikkinen
Finnish Forest Research Institute
juha.heikkinen@metla.fi

Workshop on *Statistical Issues in Forest Management*
May 2–4, 2011, Université Laval, Québec, Canada

Introduction

- Multi-source methods can provide estimates (predictions) of forest resources at any given location (e.g., presentation of McRoberts)
- Methods have also been developed to update these predictions, essentially by updating the field plot data (ground truth) and using up-to-date maps and remote sensing material (Tomppo et al. 2009, sec. 3.1)

Aim

Hence, given previous NFI and current auxiliary material, current value of given forest variable can be predicted at any given location.

How to utilize this potential in planning the design of a new NFI?
More specifically, in this study based on Finnish NFI

- use predictions of stem volumes by ‘species’ (pine, spruce, birch, other broadleaved)
- to select locations of field sample plots
- so that the precision of the resulting mean volume estimates improves

These volumes also correlated with many other variables of interest.

VERY MUCH WORK IN PROGRESS

General idea

Double (two-phase) cluster sampling; sample plots clustered into approx. one day's work (Tomppo's presentation)

- 1st phase: a dense grid of (probably overlapping) clusters; volumes predicted for each plot
- 2nd phase: finite population (sub)sampling (of clusters) from the 1st phase sample, utilizing the predicted (mean) volumes in selection

Alternative approach: stratification

- Divide whole inventory region into strata using a map of predicted volumes
- Sample each stratum independently of others
- With proportional allocation, guarantees that each stratum is appropriately represented in the sample
- Also enables denser sampling in more interesting / more variable strata

Disadvantages

- Practical complications when combined with clustered design
- Loss of information due to aggregation of volume predictions

But: can also use stratification in the 2nd phase of double sampling.

Test material

Data from the 10th NFI of Finland (2004-8); one sampling density region 7.8 mill. ha (land), 6.1 mill. ha forested

Surrogate of 1st phase sample:
4502 sample plots completely within mineral soil forest land, distributed to 1345 clusters

- a subset of a total of 12299 plots on 1815 clusters with centre point on mineral soil forest,
- subsetting mainly due to direct availability of plot-level multi-source predictions, and divided plots

Predictions based on leave-plot-out cross-validation with k-nn based on sample plot pixels.

- somewhat optimistic w.r.t. real situation, because updating error not included
- on the other hand, plot-level predictions could perhaps be improved (stabilized) by including information on pixels neighbouring the plots.

Note that in the real sampling situation, the plot density (in 2nd phase sample) will be approx. same as in NFI10, hence greater than in the 1st phase sample of this study.

- But the results will be required for much smaller areas (forestry centre regions), and
- ideally, based on annual data.

In first tests, reported here, 2nd phase sample size 67 clusters (5% of 1345), while the number of NFI10 clusters per year in forestry centre regions varied from 80 to 140.

Reference values

'True mean volumes' = mean volumes in the whole test material acting as a surrogate of the 1st phase sample

$$m^{\text{sp}} = \sum_{i=1}^{N} y_i^{\text{sp}} / N = \frac{\sum_{c=1}^{M} Y_c^{\text{sp}}}{\sum_{c=1}^{M} N_c},$$

where

N $= 4502$ is the number of plots,

y_i^{sp} mean volume, m^3/ha, of 'species' sp in plot i,

sp $\in \{\text{total}, \text{pine}, \text{spruce}, \text{birch}, \text{other}\}$,

M $= 1345$ is the number of cluster,

Y_c^{sp} $= \sum_{i \in c} y_i^{\text{sp}}$ (sum over plots in cluster c), and

N_c number of plots in cluster c.

Reference values

sp	total	pine	spruce	birch	other
m^{sp}	115.2	57.6	37.6	16.7	3.4

Leave-plot-out predictions $\hat{y}_i^{\text{sp}}$ available for each plot

sp	total	pine	spruce	birch	other
$\text{Cor}(\hat{y}_i^{\text{sp}}, y_i^{\text{sp}})$	0.69	0.58	0.73	0.35	0.10
bias^{sp}	0.8	0.2	0.3	0.1	0.2
RMSPE^{sp}	70.6	55.5	50.4	30.2	16.6

$$\text{bias}^{\text{sp}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i^{\text{sp}} - y_i^{\text{sp}} = \hat{m}^{\text{sp}} - m^{\text{sp}}$$

$$\text{RMSPE}^{\text{sp}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i^{\text{sp}} - y_i^{\text{sp}})^2}$$

Simple random sampling

Select $m = 67$ of the $M = 1345$ clusters at random

$\Longrightarrow$ sample s from which the estimated volumes are

$$\hat{m}^{\text{sp}}_{\text{SRS}} = \frac{\sum_{c \in s} Y_c^{\text{sp}}}{\sum_{c \in s} N_c}$$

Relative root mean squared error in $T = 1000$ repeated sample selections

$$\text{RRMSE}_{\text{SRS}} = \frac{\sqrt{\sum_{t=1}^{T} (\hat{m}^{\text{sp}}_{\text{SRS},t} - m^{\text{sp}})^2 / T}}{m^{\text{sp}}}$$

sp	total	pine	spruce	birch	other
$\text{RRMSE}_{\text{SRS}}$, %	6.4	8.7	15.6	13.5	31.1

Methods

$RRMSE_{SRS}$ acts as a baseline, or current practise:
predicted volumes were not used in selecting the sample.

Refined formulation of the objective:
How much can RRMSE be reduced by utilizing predicted mean volumes $\sum_{i \in c} \hat{y}_i^{sp} / N_c$ in the selection of clusters to be included in the 2nd phase sample

- Sampling with probability proportional to prediction (PPP)
- Stratification (of 1st phase sample as opposed to whole population, which was discussed earlier) according to the predicted values (double sampling for stratification)
- Balanced sampling (bal), e.g., to force 2nd phase sample means of predictions to equal their population 1st phase sample means

Among these methods, balanced sampling not as well known as the others (e.g. Gregoire and Valentine 2008); this presentation aims to promote it.

Ratio estimation

Reduction in RRMSE is also compared to that obtained by utilizing the predicted volumes in the estimation phase via ratio estimation.

In case of SRS (omitting superscript sp),

$$\hat{m}_{\text{rat}} = \hat{m}_{\text{SRS}} \left(\frac{\sum_{i=1}^{N} \hat{y}_i / N}{\sum_{c \in s} \hat{Y}_c / \sum_{c \in s} N_c} \right)$$

adjusting the plain SRS-estimator by the ratio of the known population (1st phase sample) mean of the predictions and its estimate based on the same (2nd phase) sample as $\hat{m}_{\text{SRS}}$.

Double sampling for stratification

1st phase sample divided into (homogeneous) strata using the predicted volumes; each stratum sampled independently of the others (Gregoire and Valentine 2008, sec 5.6).

- Particularly useful, if certain strata of special interest or more diverse than others: Sample those more intensively
- Must choose number of strata, stratum limits
- Not (yet) included in this study

Sampling with probability proportional to prediction

A special case of unequal probability sampling designs Gregoire and Valentine (2008, sec 3.3), cluster c included in the 2nd phase sample with probability

$$\pi_c = \frac{m\hat{Y}_c^{\text{total}}/N_c}{\sum_{c'=1}^{M}\hat{Y}_{c'}^{\text{total}}/N_{c'}}.$$

Design-unbiased Horvitz-Thompson (HT) estimator for unequal probability sampling

$$\hat{m}_{\text{PPP}}^{\text{sp}} = \frac{1}{N}\sum_{c\in s}\frac{Y_c^{\text{sp}}}{\pi_c}$$

Motivation for PPP

- for perfect predictions, $\hat{Y}_c^{\text{total}} = Y_c^{\text{total}}$, $\hat{m}_{\text{PPP}}^{\text{sp}}$ would be constant (except for variation in N_c)
- for good predictions, precision should be good (variance small)

PPP ctd.

It can also be argued that PPP is useful, when (loosely speaking) variance of y_i proportional to $\hat{y}_i$; analogue to more intensive sampling in more diverse strata.

Note, however, a severe limitation: Predictions of only one variable can be utilized; here total volume was chosen.

In the current application (clustered sampling), fixed number of clusters were sampled (using the cube method, to be introduced next), but adjustment by $N / \sum_{c \in S} \frac{N_c}{\pi_c}$ was made for variable number of plots.

Of course, SRS is a special case with constant π_c's and HT reducing to the sample mean. From now on, HT refers to the design-unbiased estimator, not utilizing the predictions in the estimation phase, as opposed to rat.

Balanced sampling

Very general method, where for any given auxiliary variables $x^{(1)},\ldots,x^{(K)}$ and inclusion probabilities π_c, sample s is selected so that, for all $k = 1,\ldots,K$,

$$\sum_{c\in s} \frac{x_c^{(k)}}{\pi_c} = \sum_{c=1}^{M} x_c^{(k)}$$

This can be obtained with the cube method (Deville and Tillé 2004), which has been implemented in R-package sampling (Tillé and Matei 2011).

Balanced sampling

yields (see Nedyalkova and Tillé 2008, for details)

- stratified sample with proportional allocation, when $x^{(k)}$'s are stratum indicators
- PPP with fixed sample size, when $K = 1$ and $x_c^{(1)} = \pi_c$
- a resolution of model-based and model-assisted paradigms
- can be combined with stratification (Chauvet 2009)

Note that systematic sampling, often applied in NFI, yields balancing of coordinates.

In this study, species-specific volume predictions were used as $x^{(k)}$'s.

Results

estimator	sampling	RRMSE, %				
		total	pine	spruce	birch	other
HT	SRS	6.4	8.7	15.6	13.5	31.1
	PPP	7.4	9.8	14.7	14.9	34.2
	SRS+bal	5.0	7.4	11.8	12.9	30.8
	PPP+bal	7.1	8.5	12.7	14.4	32.8
RAT	SRS	4.3	6.6	9.2	12.7	33.2
	PPP	4.4	6.5	8.8	13.8	34.3
	SRS+bal	4.3	6.6	8.9	12.4	31.7
	PPP+bal	4.4	6.6	8.6	14.2	34.1

Main results

- Predictions were utilized more efficiently in estimation than in sampling phase
- With ratio estimation, no effect of sampling design
- PPP seems unreliable, even with balancing
- simple balancing improves efficiency
- differences between species related to correlation between true and predicted volumes

Discussion

Poor performance of PPP probably caused by instability of ratios $y/\hat{y}$; large RMSPE.
PPP highly dependent on the quality of plot-level predictions; ratio estimation and balanced sampling based on sample and population means of predictions.

Balanced sampling very flexible; can be combined, e.g., with stratified and unequal probability designs.

Cannot be combined with systematic sampling, but geographic spread could be ensured by spatial stratification.

An advantage of balanced vs. systematic sampling is availability of approximately design-unbiased variance estimator (Deville and Tillé 2005).

Discussion ctd.

Simple random sampling aims at balance, but does not do it very well (Valliant et al. 2000, sec. 3.4.1)

Utilization of volume predictions in the estimation phase more efficient, but leads to different estimators for different variables, which may sometimes be problematic.

Conclusions (so far)

- PPP can not be recommended, when large RMSPE
- balanced sampling
  - can be justified from both design- and model-based perspective
  - flexible
  - simple to implement
  - can improve efficiency
  - does not seem to cause any harm

Future

- aim for more complete test set (including peatlands etc.)
- use (ideally) predictions based on NFI9 field plots updated to 2003 plus maps and satellite images that were available then
- check usefulness of map/image data from a larger window around plots
- try stratification
- test variance approximation for balanced sampling

References

Chauvet, G. 2009. Stratified balanced sampling. Survey Methodology **35**:115–119.

Deville, J.-C., and Y. Tillé. 2004. Efficient balanced sampling: The cube method. Biometrika **91**:893–912.

Deville, J.-C., and Y. Tillé. 2005. Variance approximation under balanced sampling. J. Statist. Plan. Infer. **128**:411–425.

Gregoire, T. G., and H. T. Valentine. 2008. Sampling techniques for natural and environmental resources. Chapman & Hall/CRC, Boca Raton.

Nedyalkova, D., and Y. Tillé. 2008. Optimal sampling and estimation strategies under the linear model. Biometrika **95**:521–537.

Tillé, Y., and A. Matei. 2011. sampling: Survey Sampling. Available at `http://CRAN.R-project.org/package=sampling`. R package version 2.4.

Tomppo, E., M. Haakana, M. Katila, K. Mäkisara, and J. Peräsaari. 2009. The Multi-source National Forest Inventory of Finland - methods and results 2005. Working paper 111, Finnish Forest Research Institute. Available at `www.metla.fi/julkaisut/workingpapers/2009/mwp111.htm`

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. Finite population sampling and inference: a prediction approach. Wiley, New York.