Using auxiliary data to allocate sampling - experiences from forest inventories

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

Research seminar in statistics
November 6, 2013, University of Helsinki

Contents

Broader than ordered (balanced sampling), and also broder than title, to cover (partly) open research questions in NFI sampling.

National Forest Inventory (NFI)

Produces information on national and regional

- forest resources - volume, growth and quality of growing stock,
- forest health,
- biodiversity of forests, etc.

since 1920's (NFI1 1921–24, NFI11 2009–13) for, e.g.,

- forest policy making at national and international levels,
- regional and national forest management planning,
- assessing sustainability of forestry,
- evaluation of greenhouse gas sinks and emissions.

NFI data is also important and widely used research material.

▸ www.metla.fi/ohjelma/vmi/info-en.htm

Main results

- Areas,
- stem volumes and biomasses by tree compartments (stem, branches, etc.), and
- increments

by various stratifications based on type of trees and forests, for example,

- area of pine-dominated stands older than 160y, or
- total volume of saw-timber on mineral soil forests by tree species.

and

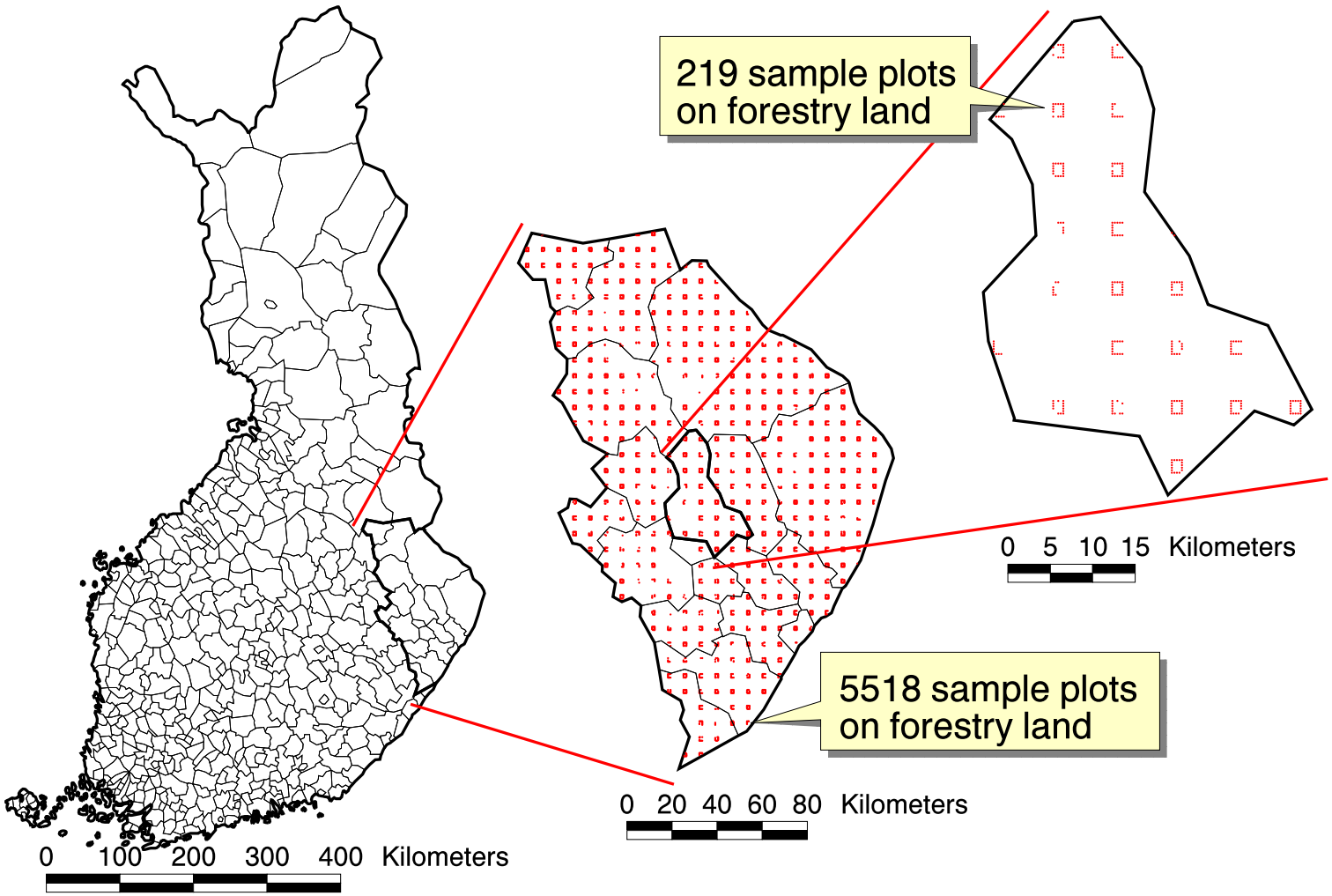
- by Forestry Centre regions ▸ www.metla.fi/metinfo/vmi/vmi-taulukot.htm
- by municipalities ▸ Tomppo et al. (2013)
- as thematic maps ▸ www.metla.fi/ohjelma/vmi/vmi-moni-en.htm

Estimators

Based on extensive field measurements (over 10,000 sample plots per year)

- National and Forestry Centre results are basically sample means of field observations
- for municipality (or map pixel) x, synthetic estimator (Tomppo et al. 2013)
  - weighted mean of field plot measurements
  - including plots outside x
  - weights derived using auxiliary data: satellite images, digital maps etc.

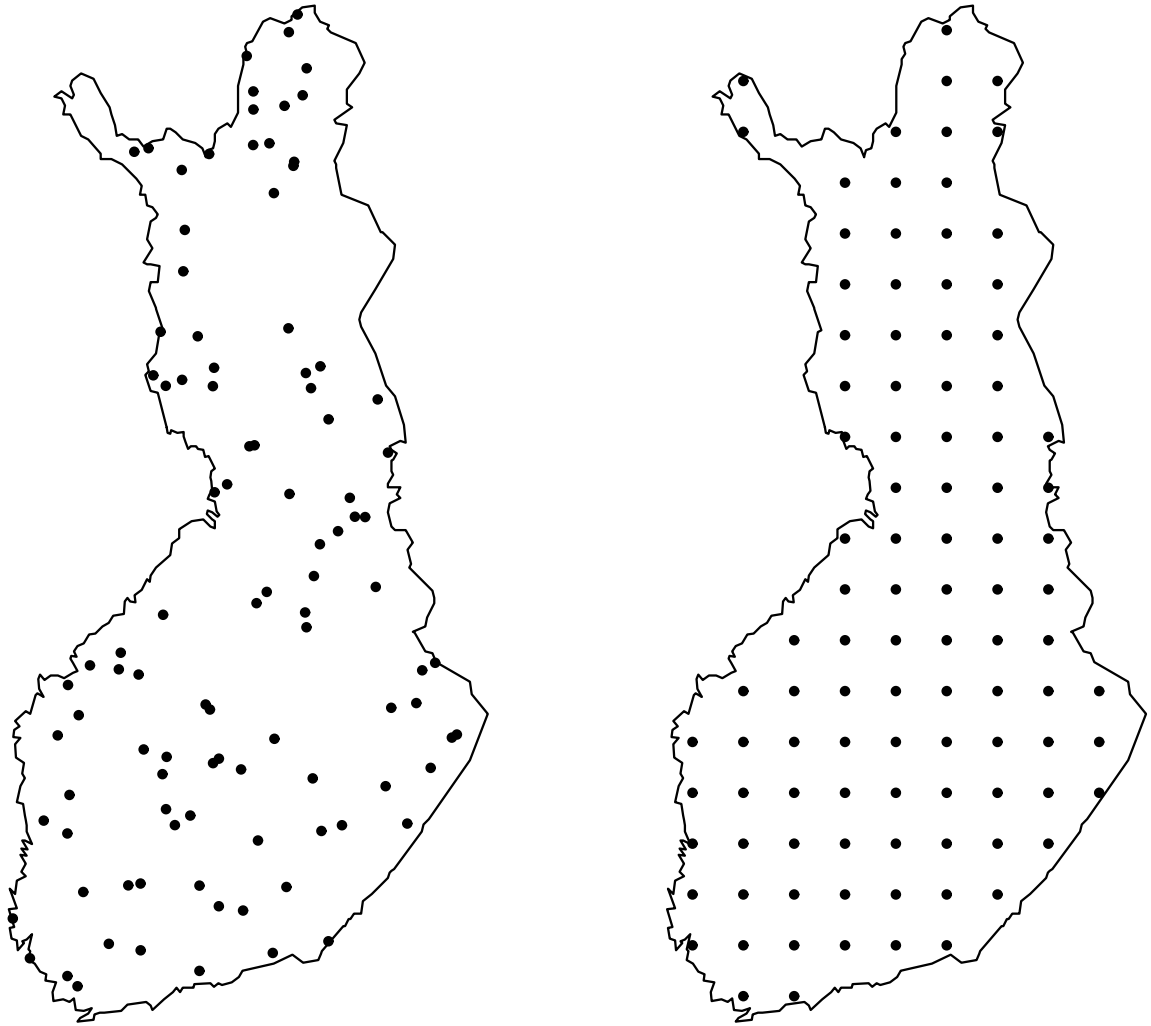
Field sample plots in NFI9



National → Forestry Centre → municipality

cluster ≈ one day's fieldwork

Simple random vs. systematic sample

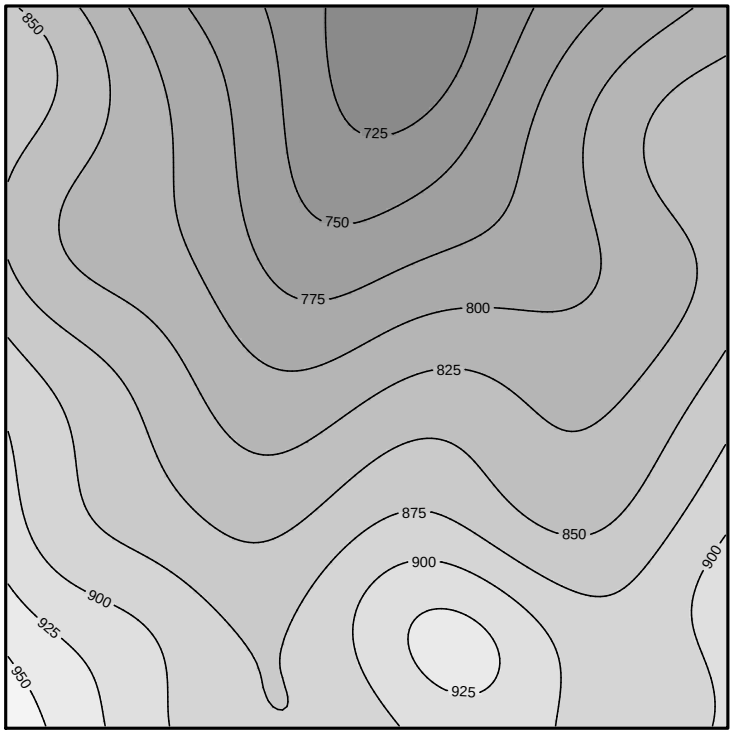


Field plot clusters treated as sampling units in uncertainty assessment.

More generally

Estimate mean of spatial surface over a region (denoted by A)

$$Z(A) = \frac{1}{|A|} \int_A Z(s)\, ds$$



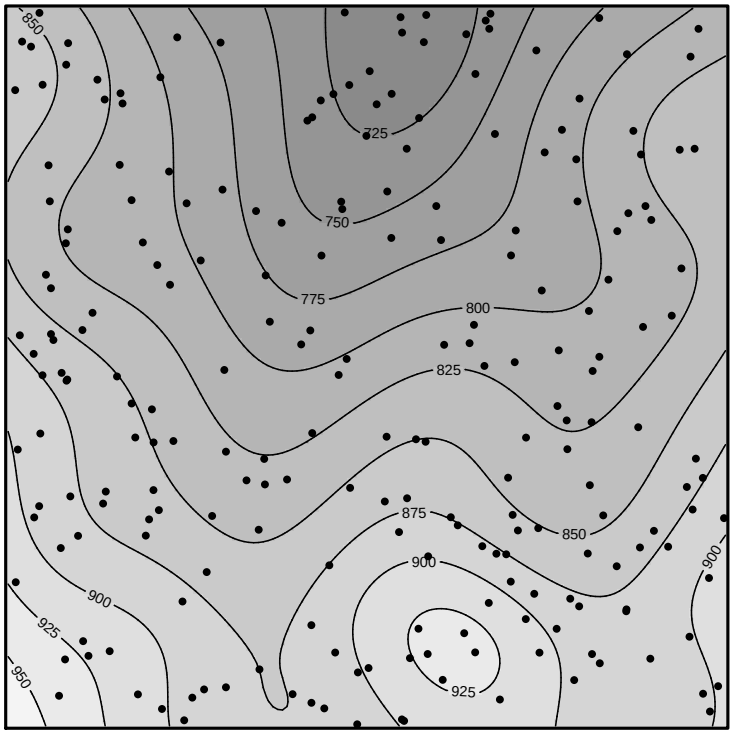
Example data: elevation map based on dataset `topo` of R-package `MASS` (top left in Venables and Ripley 2002, Fig. 15.8)

More generally

Estimate mean of spatial surface over a region (denoted by A)

$$Z(A) = \frac{1}{|A|} \int_A Z(s)\, ds$$

from measurements $Z(s_i)$ taken at arbitrarily located sample points $s = \{s_1, s_2, \ldots, s_n\} \subset A$.



Example data: elevation map based on dataset `topo` of R-package `MASS` (top left in Venables and Ripley 2002, Fig. 15.8)

Uncertainty assessment

For systematic samples from smooth surfaces, SRS-variance

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} [Z(s_i) - \bar{z}]^2,$$

often strongly over-estimates the squared error $[\bar{z} - Z(A)]^2$.
An old idea (see Papritz and Webster 1995, for a review): use local balanced differences, replacing $Z(s_i) - \bar{z}$'s by

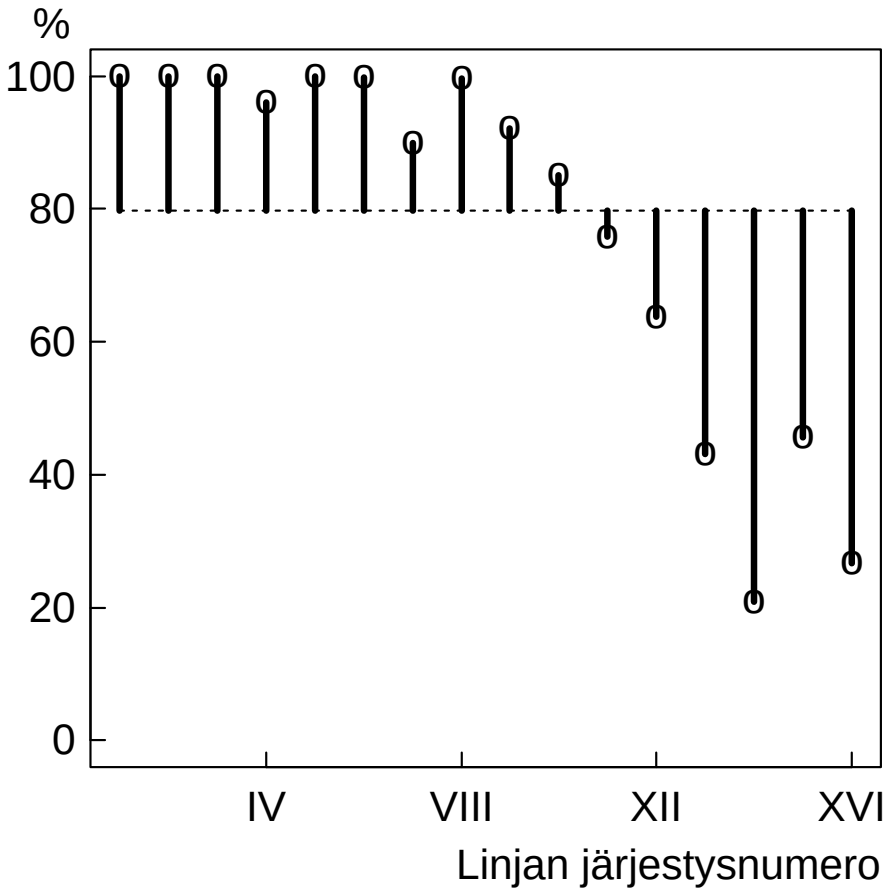
$$d_g = \sum_{j=1}^{m_g} a_{g,j} Z(s_{g,j}),$$

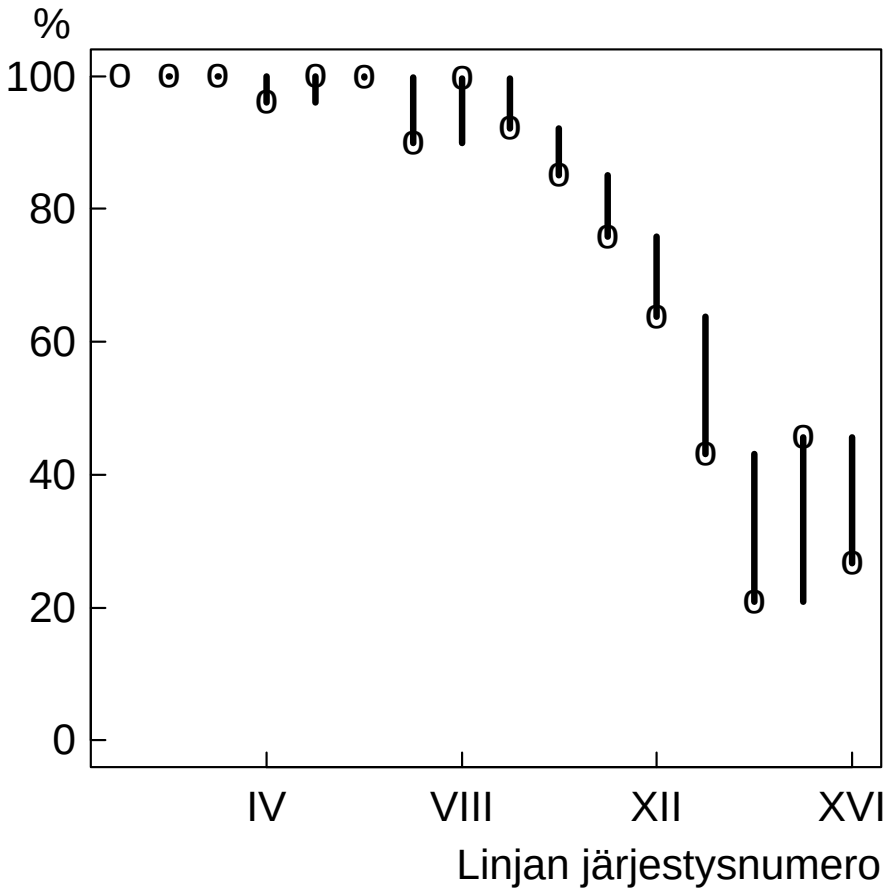
where g is a group of m_g nearby sample points $s_{g,j}$, $\sum a_{g,j} = 0$, and $\sum a_{g,j}^2 = 1$.

Simple example in 1d

g pair of subsequent sample points, say, s_i, s_{i-1}, and

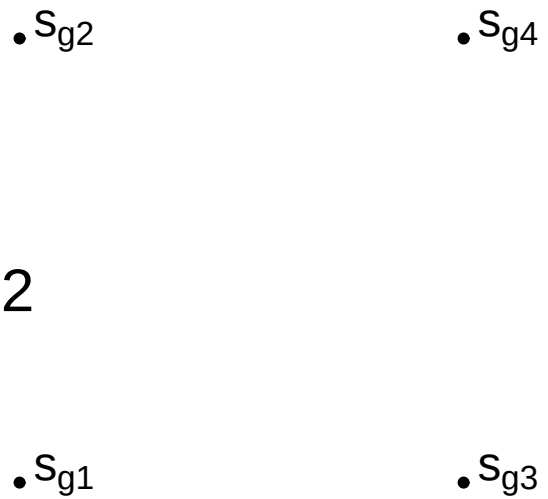
$$d_g = [Z(s_i) - Z(s_{i-1})]/\sqrt{2}.$$





Matérn's generalization to 2d

For a systematic sample with rectangular lattice of sample points:

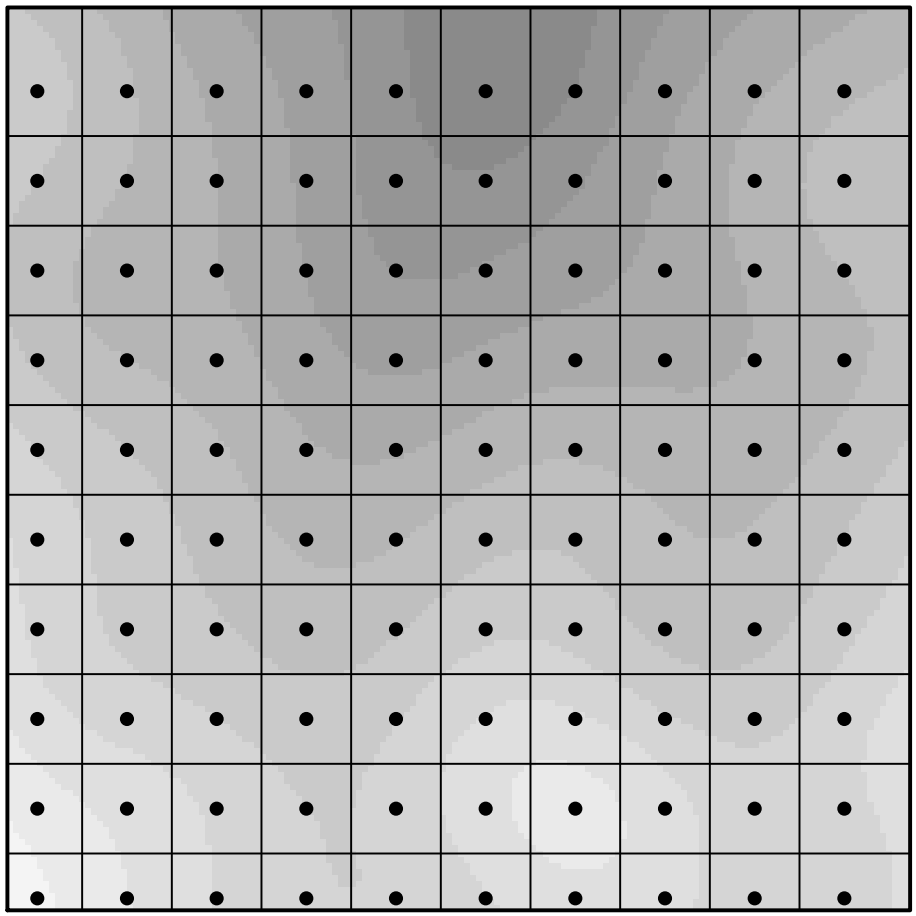


$$d_g = \{Z(s_{g1}) - Z(s_{g2}) - Z(s_{g3}) + Z(s_{g4})\}/2$$
$$= \{Z(s_{g1}) + Z(s_{g4})\}/2 - \{Z(s_{g2}) + Z(s_{g3})\}/2$$

Can prove under very general assumptions that still should over-estimate the error of sample mean.

Pitfalls for systematic sampling and its uncertainty assessments

- Periodicity of true response surface often mentioned but rarely problem in practice.
- Non-centricity potentially more serious, if strong trend over *A*.

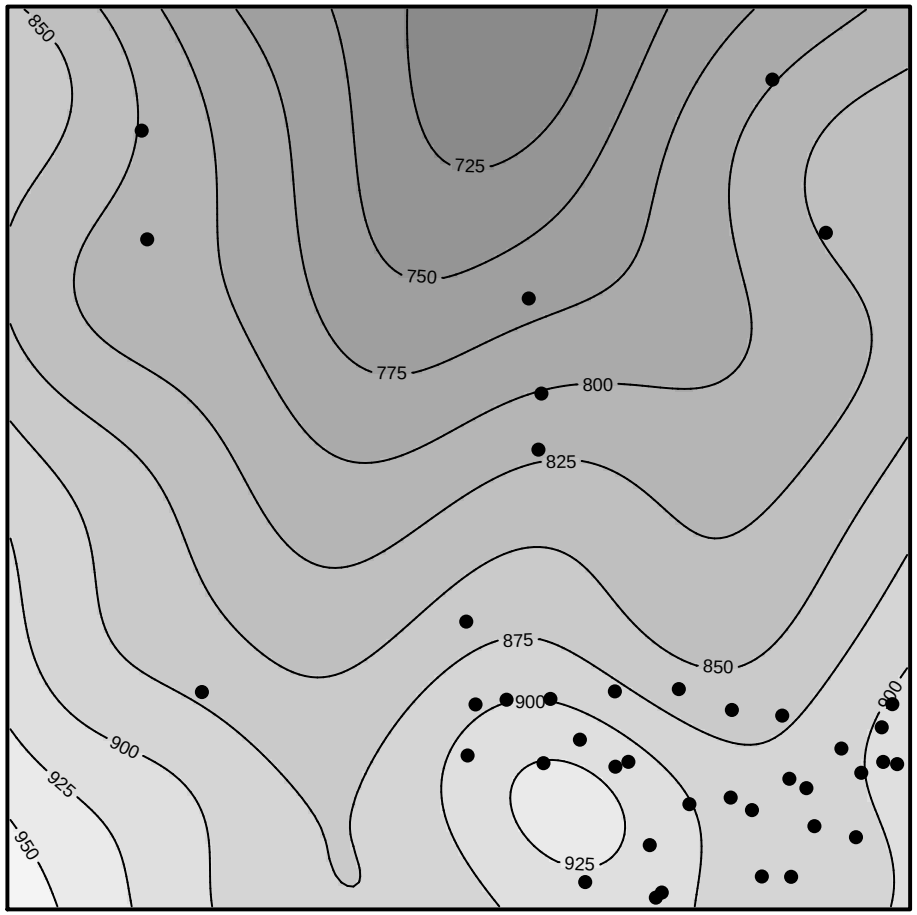


Why polygonal declustering?

Sample mean

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} Z(s_i)$$

not always a good estimator of $Z(A)$.
For example, if (some) sample points are clustered.

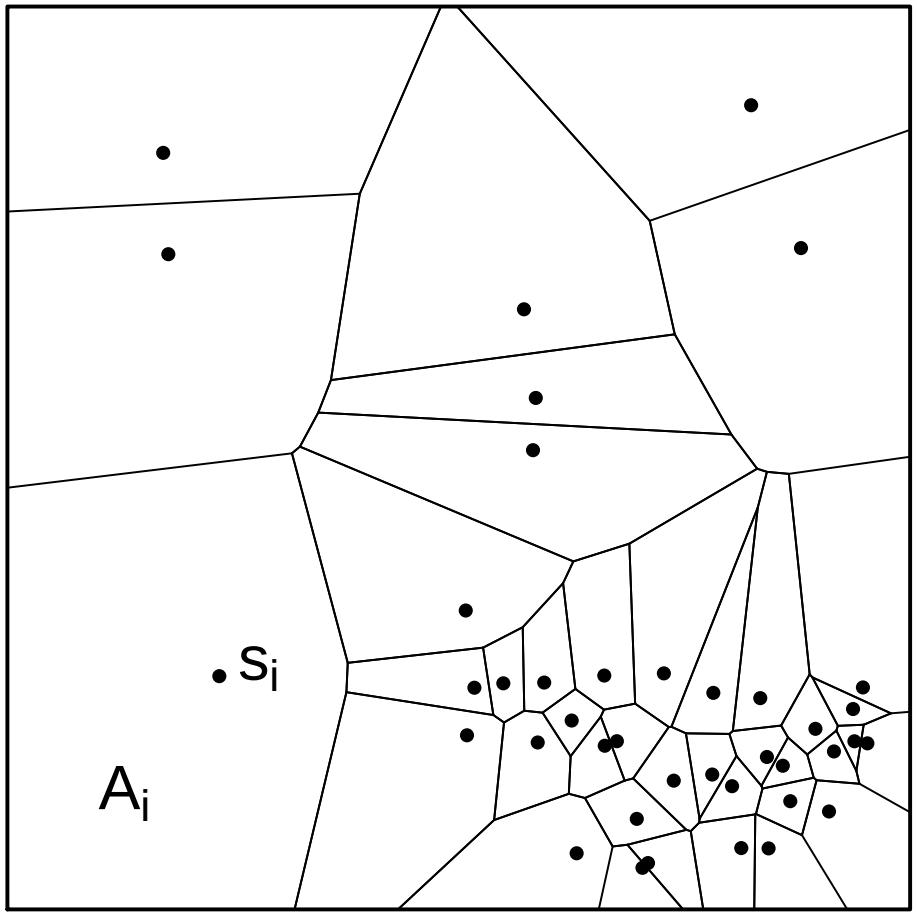


Polygonal declustering

$Z(A)$ estimated by weighted sample mean

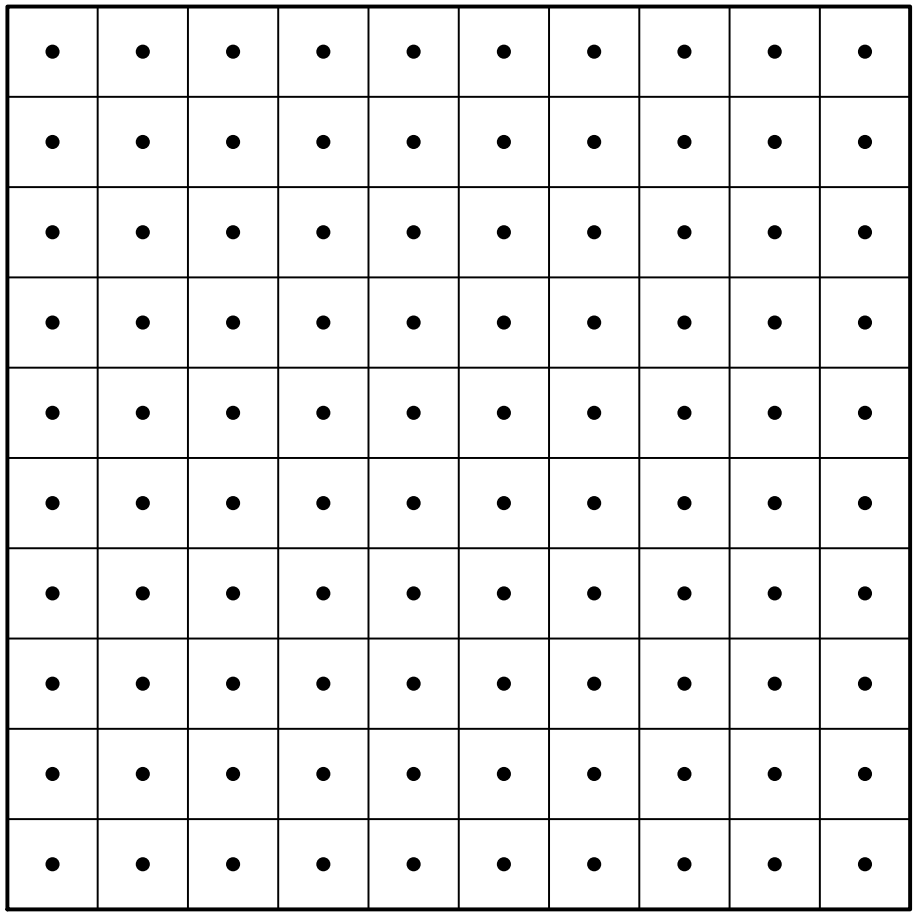
$$\widehat{Z(A)} = \frac{1}{|A|} \sum_{i=1}^{n} |A_i| Z(s_i),$$

where the weight $|A_i|$ is the area of Voronoi polygon around s_i.
Isolated points receive more weight than those clustered with others.



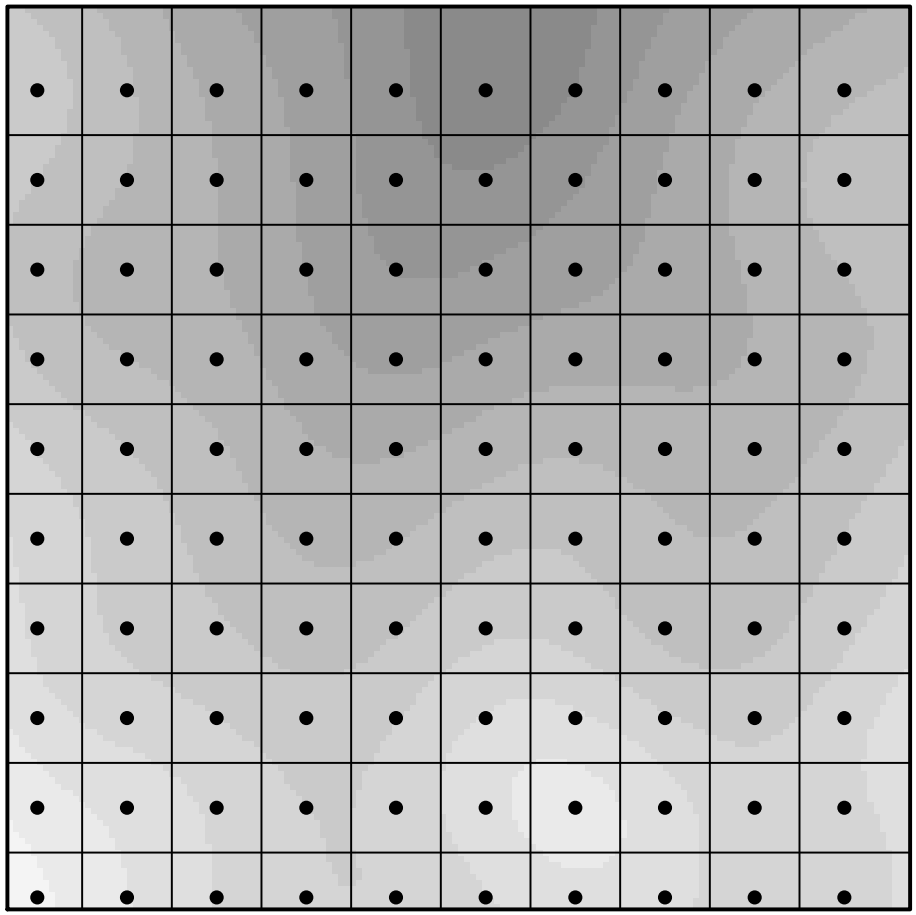
Systematic centric sample

Special case, where $|A_i| = |A|/n$, $\forall i$, so that polygonal declustering estimator is identical to the sample mean.



Systematic non-centric sample

Sample mean no longer good, if strong trend over *A*.
Polygonal declustering leads to edge correction similar to that introduced by Yates (1948) for response curves in 1d (see, e.g., Cochran 1977, 8.7).



Relevance in NFI?

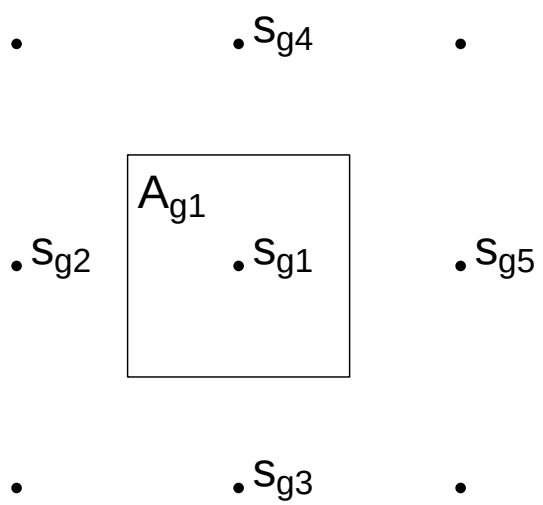
- Originally motivated by shift from regionally proceeding NFI to a panel system
- Systematic grid from 5 years rotation, but first results already after 3 years
- So how to formulate local balanced differences to non-gridded clusters?
- Differences between each cluster from the (weighted) mean of its neigbours
- More details and experiments with `topo` data in Heikkinen (2009)
- Here just 2 more slides to show the idea.

Idea in case of systematic sample

$$d_g = c\left\{Z(s_{g1}) - \frac{1}{4}[Z(s_{g2}) + Z(s_{g3}) + Z(s_{g4}) + Z(s_{g5})]\right\},$$

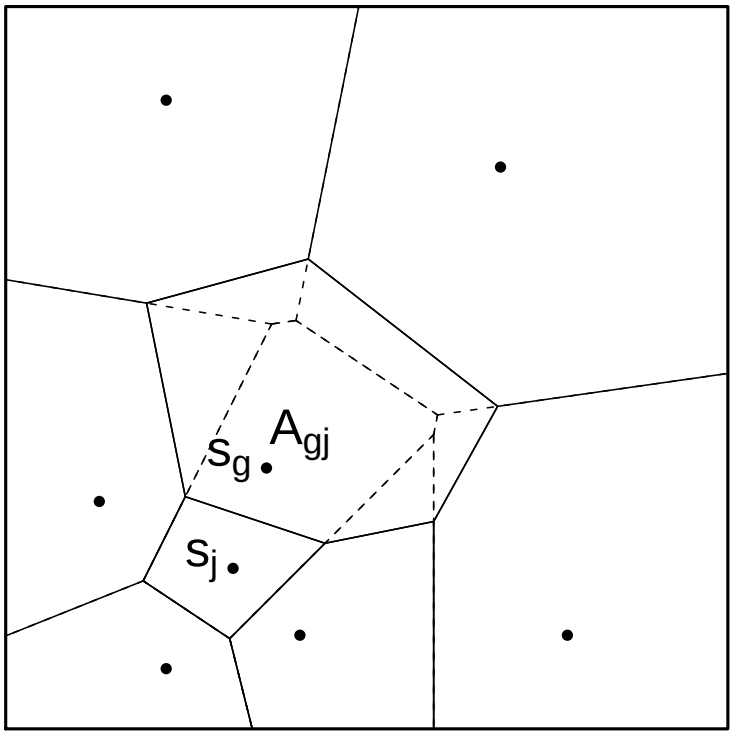
where $c = 2/\sqrt{5}$ chosen so that $\sum a_{g,j}^2 = 1$

d_g^2 should be an overestimate of $[Z(s_{g1}) - Z(A_{g1})]^2$, because points $s_{g2}, \ldots, s_{g5}$ further away from s_{g1} than any point of A_{g1}.



Generalization to irregularly distributed sample points

$$d_g = c\left[Z(s_g) - \sum_{j\sim g} \frac{|A_{g,j}|}{|A_g|} Z(s_j)\right],$$



where

$j \sim g$ if A_g and A_j are adjacent

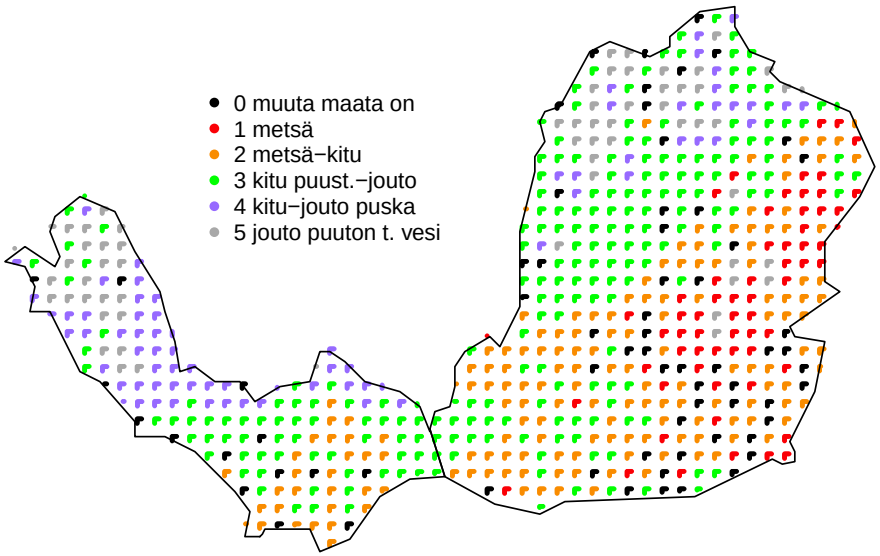
$A_{g,j}$ is that part of A_g, which belongs to A_j in the tessellation of $s \setminus s_g$.

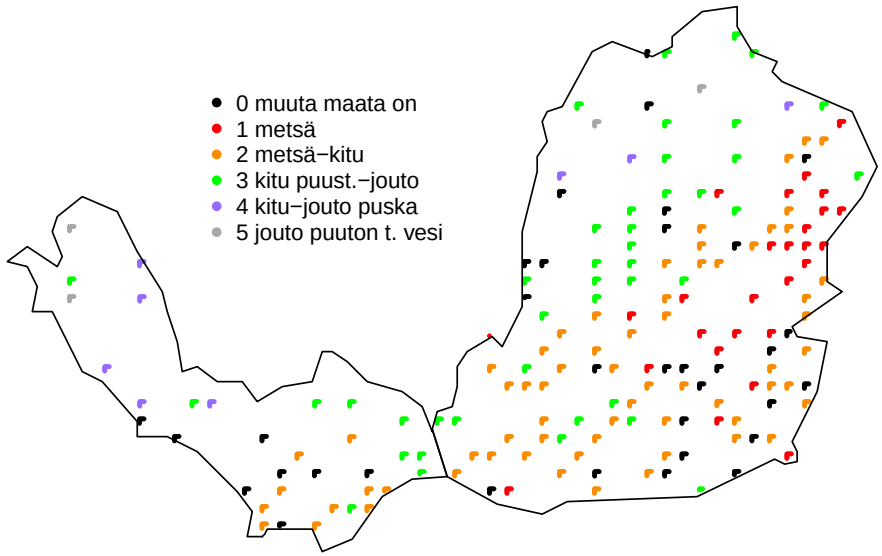
plus some modification (not yet implemented) to guarantee that

$\|s_g - s_j\| > \|s_g - t\| \quad \forall t \in A_{g,j}$

Northernmost Finland

- Large areas of sparse forest cover
- User earlier MS-NFI maps to stratify according to forest cover
- Sample more forested strata more intensively
- Direct stratified sampling would lead to high variation in number of plots per cluster
- Instead, stratify 1st phase sample, a dense square grid of clusters (Tomppo et al. 2011, 2.1.3)





More efficient utilization of auxiliary information

- Methods available to update MS-NFI predictions, essentially by updating the field plot data and using up-to-date maps and remote sensing material (Tomppo et al. 2013, 3.1)
- 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

- 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.

General idea

Double (two-phase) cluster sampling as in Northernmost Finland; sample plots clustered into approx. one day's work

1st phase: a dense grid of clusters; volumes predicted for each plot

2nd phase: subsampling of clusters from the 1st phase sample, utilizing the predicted volumes in selection

Test material

- Data from the 10th NFI of Finland (2004-8); region of 7.8 mill. ha (land)
- Surrogate of 1st phase sample: 4502 sample plots completely within mineral soil forest land, distributed to 1345 clusters
- Predictions based on leave-plot-out cross-validation.
- 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.
- 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

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

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)
- Balanced sampling (bal), e.g., to force 2nd phase sample means of predictions to equal their population 1st phase sample means

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 $\hat{m}_{\text{SRS}}$.

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 $k = 1,\ldots,K$,

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

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

Stratified sampling and PPP as special cases

- Let

  $\pi_c \equiv m/M$ (constant) and
  $x_c^{(k)}$ be stratum indicators: Each cluster belongs to one of K strata, and

$$x_c^{(k)} = \begin{cases} 1 & \text{if cluster } c \text{ belongs to stratum } k \\ 0 & \text{otherwise} \end{cases}$$

  Then balancing equations $\Leftrightarrow$ # stratum k clusters in sample $\approx \frac{m}{M} \times$ # stratum k clusters in population.

- Let

  $K = 1$
  $x_c^{(1)} = \pi_c$

  Then balancing equations $\Leftrightarrow$ Sample size $\approx \sum_{c=1}^{M} \pi_c =$ constant.

Balanced sampling

- protects against bias in model-based inference (Valliant et al. 2000, 3.2.1)
- can provide a resolution of model-based and model-assisted paradigms (Nedyalkova and Tillé 2008)
- can be combined with stratification (Chauvet 2009)

Note that centric systematic sampling 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, 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.

Summary & current state

- Sampling issues very important in forest inventories; currently also various international projects
- Case Northernmost Lapland straight-forward; manuscript in preparation
- Other cases interesting ideas and preliminary experiments:
  - achievable gain from polygonal declustering in typical NFI
  - properties of local balanced estimator of uncertainty in polygonal declustering
  - utilization of continuous auxiliary variables

  Currently lack of resources hinders research on these issues. Good components for a PhD project, if anybody interested.

References

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

Cochran, W. G. 1977. Sampling techniques. 3rd edition. Wiley, New York.

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.

Heikkinen, J. 2009 Assessing the uncertainty of the polygonal declustering estimator of a spatial mean 2nd Nordic-Baltic Biometric Conference, 10-12 June 2009, Tartu, Estonia. Available at `www-1.ms.ut.ee/NBBC09/confbook.pdf`.

Matérn, B. 1960. Spatial variation. Meddelanden från Statens Skogsforskningsinstitut **49.5**:1–144. Also appeared as number 36 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1986.

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

Papritz, A., and R. Webster. 1995. Estimating temporal change in soil monitoring II. Sampling from simulated fields. European Journal of Soil Science **46**:13–27.

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. et al. 2011. Designing and conducting a forest inventory — case 9th National Forest Inventory of Finland. Springer, Dordrecht.

Tomppo, E., M. Katila, K. Mäkisara, and J. Peräsaari. 2013. The Multi-source National Forest Inventory of Finland - methods and results 2009. Working paper 273, Finnish Forest Research Institute. Available at `www.metla.fi/julkaisut/workingpapers/2013/mwp273.pdf`

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

Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. 4th edition. Springer, New York.

Yates, F. 1948. Systematic sampling. Phil. Trans. R. Soc. **A, 241**:345–377.