An extension of the weight share method when using a continuous sampling frame
Application to the French national forest inventory

Guillaume Chauvet
ENSAI/IRMAR

Based on joint work with Olivier Bouriaud, Philippe Brion, Trinh Duong and Minna Pulkkinen

IASS Webinar
27/09/2023

The French National Forest Inventory

The French National Forest Inventory (NFI) follows a design-based sampling protocol, in order to produce useful and relevant information for the data production on the French forest.

The French NFI was created in 1958 to assess French forest resources. The methodology was changed in November 2004, and currently makes use of sample points on a grid defined for a 10-year period, from which one tenth is dealt with each year (Hervé, 2017).

The French NFI collects dendrometric, ecological and floristic information. The sampled data are used to create forest maps by administrative county through interpreting aerial photographs. The survey can also take additional data on request (dead wood, forest health, ...)

Overview of the talk

During this presentation, I will talk about:

1. Some examples of sampling strategies used by NFIs.
2. The weight share method, which is useful to relate the sampled units (points) to the surveyed units (trees).
3. Application to the French NFI.
4. Some perspectives for future work.

Sampling designs for forest inventories

Objectives

We are interested in a finite target population U^B, which is typically a population of trees in forest inventories. Let y_k^B denote the attribute of interest for $k \in U^B$. We wish to estimate the total

$$\tau_y^B = \sum_{k \in U^B} y_k^B, \quad (1)$$

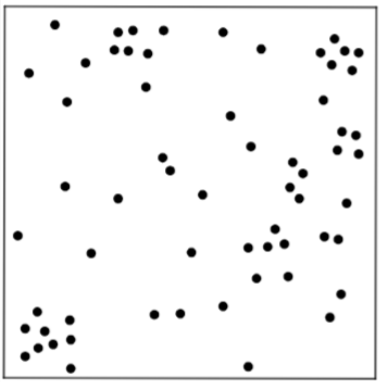
which may be the total volume of wood, for example.

Some form of indirect sampling is used. Let $\mathcal{U}^A$ denote a continuous territory containing all the units in U^B. A typical inventory design consists in:

1. selecting a large 1st-phase sample of points in $\mathcal{U}^A$ (continuous sampling design),
2. classifying the points according to the land cover (photo-interpretation),
3. selecting a smaller, 2nd-phase sample using the 1st-phase auxiliary information,
4. using fixed-shape supports from these points to survey the units in U^B.

Step 1: some possible 1st-phase sampling designs

Uniform random sampling

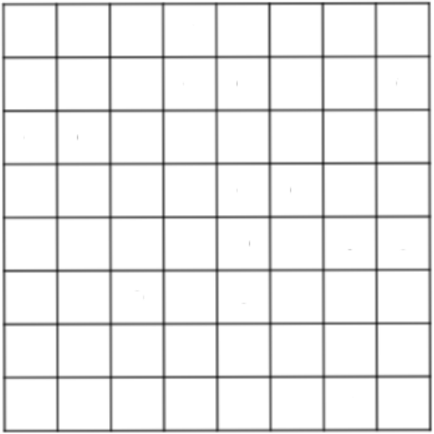


The 1st phase sample may be obtained by independent uniform random sampling inside the territory $\mathcal{U}^A$.

Not useful/used in practice: some areas may be covered by several points, while others are not surveyed.
$\Rightarrow$ poor spatial balance

Step 1: some possible 1st-phase sampling designs

Grid sampling ≡ Tesselation sampling schemes



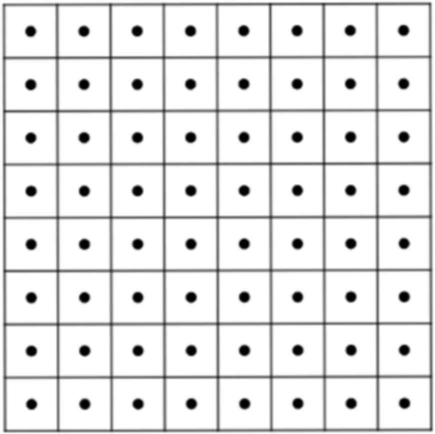
Grid sampling is very common in forest inventories.

A sample of cells is selected (possibly all), and a sample of points is selected inside each selected cell (usually one).

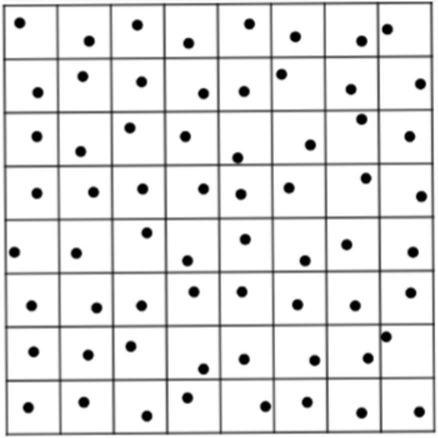
Grid sampling is very useful to achieve sample coordination both in space and in time (Pisani, Di Biase, Marcheselli, 2023).

Step 1: some possible 1st-phase sampling designs

Spatially systematic aligned sample

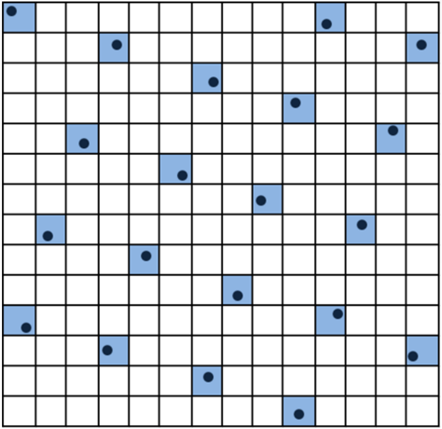


Spatially systematic unaligned sample



Step 1: some possible 1st-phase sampling designs

French annual sample: a two-stage design

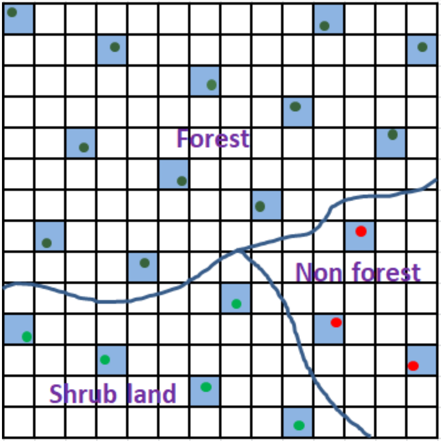


A sample of cells is first selected, by using some form of systematic sampling with equal probabilities.

One point is randomly selected inside each cell.
⇒ First-phase sample S_{1p}^{A}.

The cells are randomly partitioned into 10 rotation groups (negative coordination). All the cells are surveyed in ten years.

Steps 2-3: photo-interpretation and 2nd phase sampling

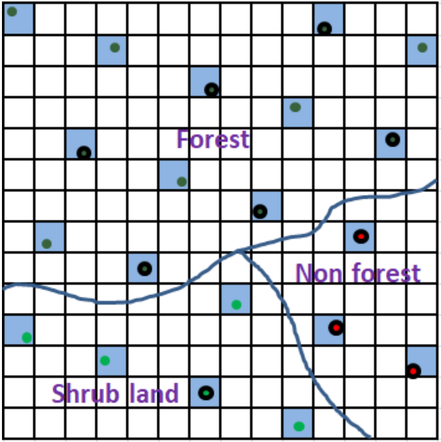


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest).

The 1st phase sample is stratified, with $\neq$ sub-sampling intensities inside strata.

For France, $f_{2g} = 1/2$ for forest, $f_{2g} = 1/4$ for shrub land, and $f_{2g} = 1$ for non-forest (but no visit on the field in this last case).

Steps 2-3: photo-interpretation and 2nd phase sampling

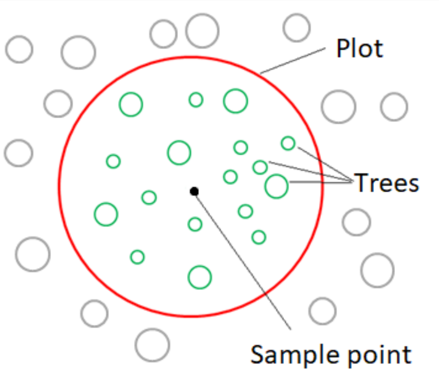


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest).

The 1st phase sample is stratified, with $\neq$ sub-sampling intensities inside strata.

For France, $f_{2g} = 1/2$ for forest, $f_{2g} = 1/4$ for shrub land, and $f_{2g} = 1$ for non-forest (but no visit on the field in this last case).
$\Rightarrow$ Second-phase sample S_{2p}^{A}.

Step 4: use of fixed shape supports



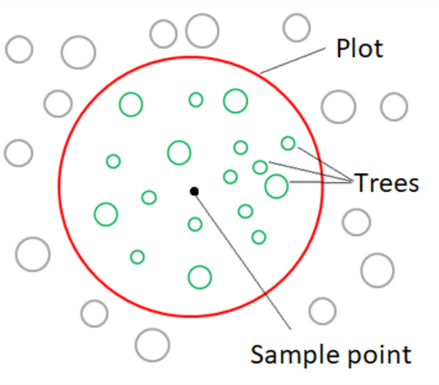
A plot with fixed radius r is centered at the sampled point, and the trees within are surveyed.

For the French NFI, 3 plot radiuses:

Plot radius	Tree's circonference at 1.3m
6m	23.5-70.5cm (ST)
9m	70.5-117.5cm (MT)
15m	$\geq$ 117.5cm (LT)

Other techniques are possible (Bitterlich sampling, not covered here).

Step 4: use of fixed shape supports

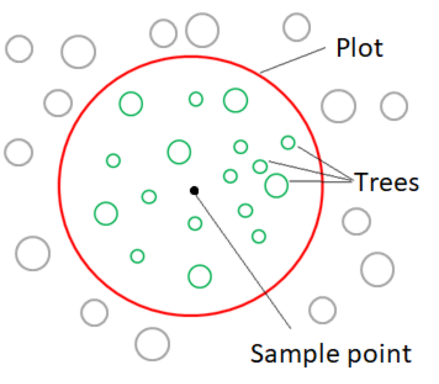


Remark: there can be a third phase of sampling (not covered here).

Cheap attributes (e.g., basal diameter) are collected on the whole 2nd phase sample, while expensive attributes (e.g., volume) are collected on a sub-sample only, and imputed on the complementary.

This is the case in the French NFI.

Step 4: use of fixed shape supports



The trees within the plot(s) are surveyed, if they belong to the corresponding circonference class.

In summary:

- a sample of points S^A is selected in a continuous territory $\mathcal{U}^A$,
- a sample of trees S^B is surveyed on the field.

How to obtain estimators for the population of trees?

The weight share method

Principle

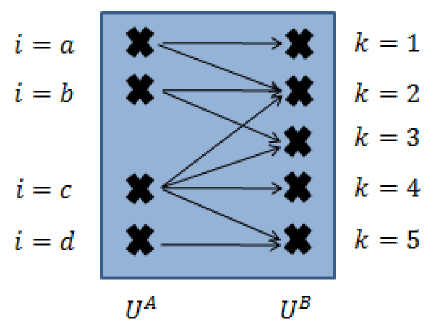
The weight share method (Deville and Lavallée, 2006) is very useful when we wish to perform estimations on a population U^B, by using another population for which a sampling frame is available.

In particular, this is the key tool to produce cross-sectional estimations in longitudinal surveys when households present at time $t+1$ are catched via individuals sampled at time t, and followed over time (e.g., Ardilly et Lavallée, 2007).

I will first present the principles of the method with a discrete sampling frame U^A, and then an extension of the method in case of a continuous sampling frame $\mathcal{U}^A$.

The discrete-discrete case

Weight share method: discrete-discrete case



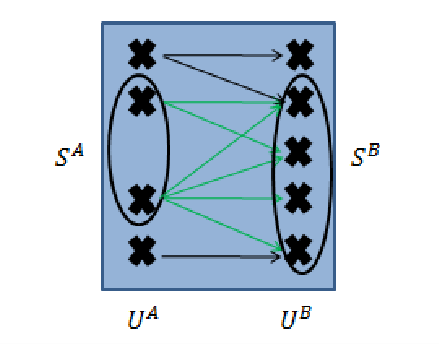
We are interested in a discrete pop. U^B, with a var. of interest y_k^B for which we estimate the total

$$\tau_y^B = \sum_{k \in U^B} y_k^B .$$

No sampling frame for U^B, but linked to a pop. U^A with a sampling frame. Let us note

$$L^{AB}(i,k) = \begin{cases} 1 & \text{if } i \text{ and } k \\ & \text{are linked,} \\ 0 & \text{otherwise.} \end{cases}$$

Weight share method: discrete-discrete case



A sample S^A is selected in U^A.

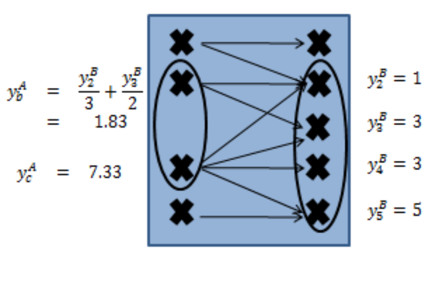
All the units in U^B linked to some unit $i \in S^A$ are surveyed:

$$S^B = \Big\{ k \in U^B;\ L^{AB}(i,k) = 1 \text{ for at least one } i \in S^A \Big\}.$$

It is supposed that any $k \in U^B$ has at least one link to U^A (no coverage bias):

$$N_{+k}^{AB} = \sum_{i \in U^A} L^{AB}(i,k) > 0.$$

Weight share method: discrete-discrete case



Duality principle: the variable y_k^B may be transformed into a variable y_i^A defined on U^A such that:

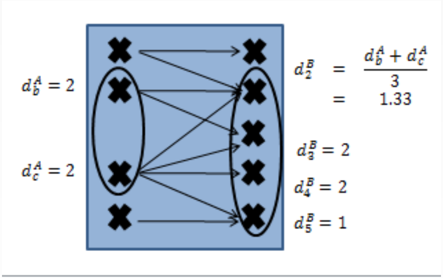
$$\tau_y^B = \sum_{k \in U^B} y_k^B = \sum_{i \in U^A} y_i^A.$$

More precisely

$$y_i^A = \sum_{k \in U^B} \frac{L^{AB}(i,k) y_k^B}{N_{+k}^{AB}},$$

i.e. each y_k^B is equally shared among the linked units in U^A.

Weight share method: discrete-discrete case



Let d_i^A denote the sampling weight of $i \in S^A$. Reverting the duality principle:

$$\hat{\tau}_y^B = \sum_{i \in S^A} d_i^A y_i^A = \sum_{k \in S^B} d_k^B y_k^B$$

with

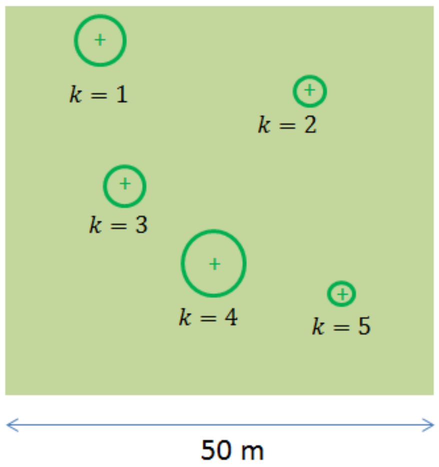
$$d_k^B = \frac{1}{N_{+k}^{AB}} \sum_{i \in S^A} L^{AB}(i,k) d_i^A$$

$$\text{(Weight share method)}$$

Note that for each $k \in S^B$, the total number of links N_{+k}^{AB} needs to be known.

The continuous-discrete case

Weight share method: continuous-discrete case



We are interested in a discrete pop. U^B (trees), with a var. of interest y_k^B for which we estimate the total

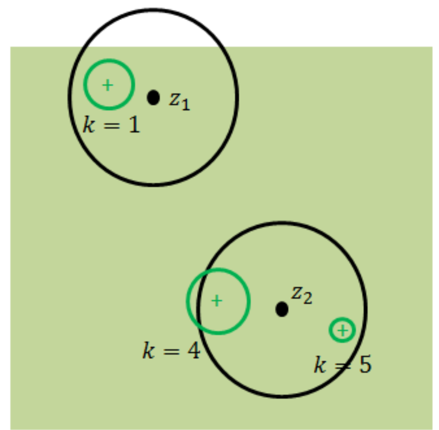
$$\tau_y^B = \sum_{k \in U^B} y_k^B.$$

No sampling frame for U^B, but linked to a continuous pop. $\mathcal{U}^A$ (territory). Let us note

$$L^{AB}(x,k) = \begin{cases} 1 & \text{if } x_k \in C(x,r), \\ 0 & \text{otherwise.} \end{cases}$$

Remark : link function depending on the sampling protocol (ST, MT, LT).

Weight share method: continuous-discrete case



A sample $S^A = \{z_1, \dots, z_n\}$ of n points is selected in $\mathcal{U}^A$.

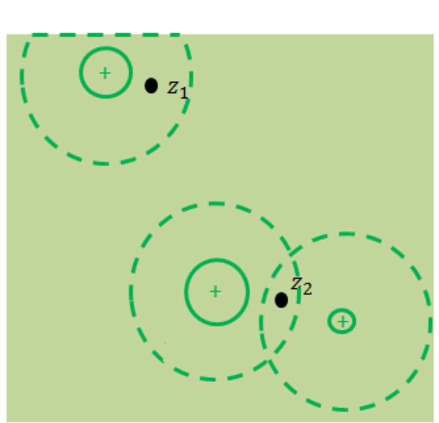
All the trees in U^B linked to the sampled points (i.e., inside the associated plots) are surveyed:

$$S^B = \Big\{k \in U^B;\ L^{AB}(x,k) = 1 \text{ for at least one } x \in S^A\Big\}.$$

We suppose that $\forall k \in U^B$, the surface of the inclusion area is > 0:

$$N^{AB}_{+k} = \int_{x \in \mathcal{U}^A} L^{AB}(x,k)dx > 0.$$

Weight share method: continuous-discrete case



A sample $S^A = \{z_1, \ldots, z_n\}$ of n points is selected in $\mathcal{U}^A$.

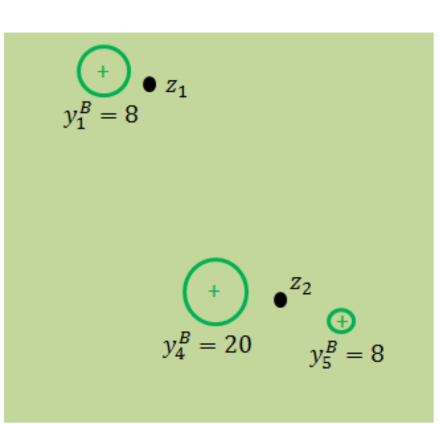
All the trees in U^B linked to the sampled points (i.e., inside the associated plots) are surveyed:

$$S^B = \Big\{k \in U^B;\ L^{AB}(x,k) = 1 \text{ for at least one } x \in S^A\Big\}.$$

We suppose that $\forall k \in U^B$, the surface of the inclusion area is > 0:

$$N^{AB}_{+k} = \int_{x \in \mathcal{U}^A} L^{AB}(x,k)dx > 0.$$

Weight share method: continuous-discrete case



Duality principle: the variable y_k^B may be transformed into a local variable $y^A(x)$ defined on $\mathcal{U}^A$ such that:

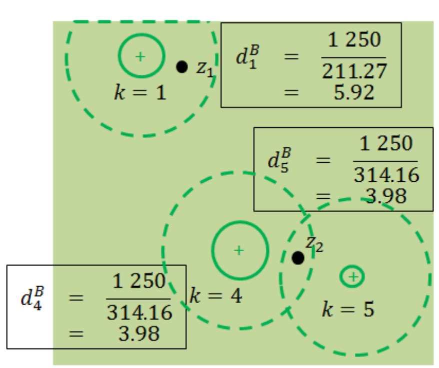
$$\tau_y^B = \sum_{k \in U^B} y_k^B = \int_{x \in \mathcal{U}^A} y^A(x) dx.$$

More precisely

$$y^A(x) = \sum_{k \in U^B} \frac{L^{AB}(x,k) y_k^B}{N_{+k}^{AB}},$$

i.e. each y_k^B is equally shared among the points inside its inclusion area.

Weight share method: continuous-discrete case



Let $d^A(x)$ denote the sampling weight of $x \in S^A$. Reverting the duality principle:

$$\hat{\tau}_y^B = \sum_{x \in \mathcal{S}^A} d^A(x) y^A(x) = \sum_{k \in S^B} d_k^B y_k^B$$

with

$$d_k^B = \frac{1}{N_{+k}^{AB}} \sum_{x \in S^A} L^{AB}(x,k) d^A(x).$$

(Weight share method)

Note that for each $k \in S^B$, the surface N_{+k}^{AB} of its inclusion area needs to be known.

Discussion

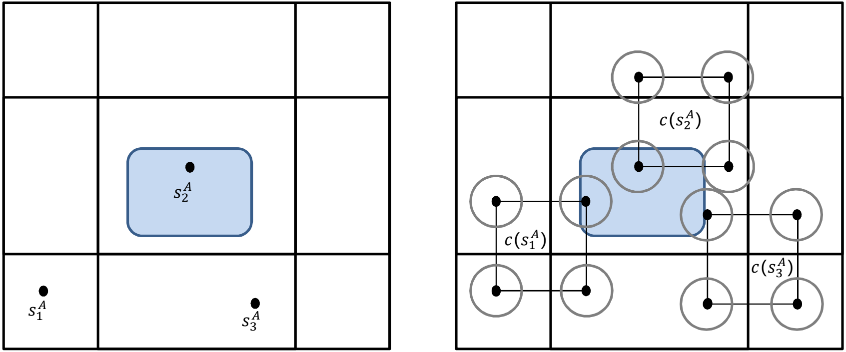
The approach has already been considered in the literature (infinite population approach):

- Stevens and Urqhart (2000) propose a very similar approach. Important but difficult paper, which seems overlooked in the literature.
- Mandallaz (2007) introduces the local density, which is an estimator making implicit use of the synthetic variable.
- Gregoire and Valentine (2007) describe the attribute density $\equiv$ particular case of the synthetic variable.

Purpose of Chauvet, Brion and Bouriaud (2023): make the approach explicit, to state the estimation weights, the estimators of totals and associated variance estimators.

The weight share method also enables to deal with more complicated sampling strategies, like cluster/trakt sampling.

An example of cluster sampling



Continuous Horvitz-Thompson estimation

The estimation of τ_y^B may be performed by continuous Horvitz-Thompson (HT) estimation (Cordy, 1993) in the population $\mathcal{U}^A$:

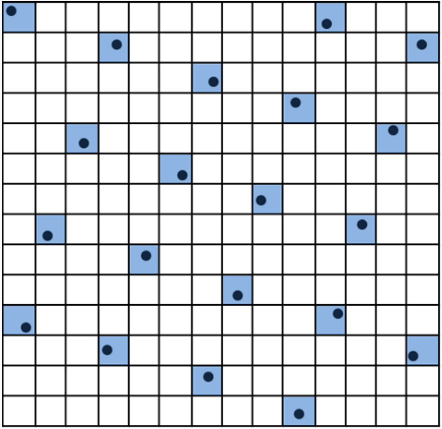
$$
\begin{aligned}
\hat{\tau}_y^B &= \sum_{x \in S^A} d^A(x) y^A(x) = \sum_{x \in S^A} \frac{y^A(x)}{\pi^A(x)}, \\
\hat{V}_{YG}\left(\hat{\tau}_y^B\right) &= \frac{1}{2} \sum_{x \neq x' \in S^A} \left\{ \frac{\pi^A(x)\pi^A(x') - \pi^A(x,x')}{\pi^A(x,x')} \right\} \left\{ \frac{y^A(x)}{\pi^A(x)} - \frac{y^A(x')}{\pi^A(x')} \right\}^2
\end{aligned}
$$

with $\pi^A(x)$ the inclusion density and $\pi^A(x,x')$ the joint inclusion density for some points $x \neq x' \in \mathcal{U}^A$.

The estimator $\hat{\tau}_y^B$ is fine, but $\hat{V}_{YG}\left(\hat{\tau}_y^A\right)$ is inconsistent for most sampling designs used in NFIs (one point per cell selected). We need to consider (approximate) (conservative) variance estimators.

Application to the French NFI

French annual sample: a first-phase two-stage design



Sample of n_I cells first selected among the N_I cells, with equal probabilities.
One point randomly selected inside each cell of area A_c.

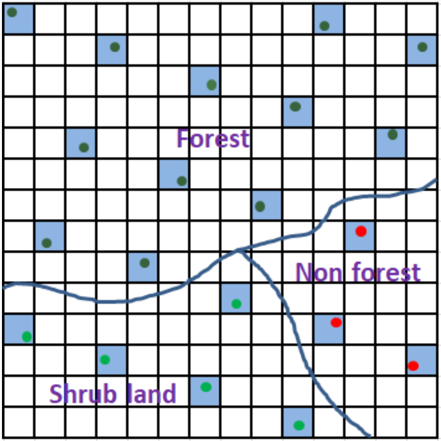
The first-phase inclusion density/HT estimator are

$$\pi_{1p}(x) = \frac{n_I}{N_I} \times \frac{1}{A_c} = \frac{n_{1p}}{A_F}$$

for any $x \in \mathcal{U}^A$,

$$\hat{\tau}_{y,1p} = \frac{A_F}{n_{1p}} \sum_{x \in S_{1p}^A} y^A(x).$$

French annual sample: second-phase sampling design

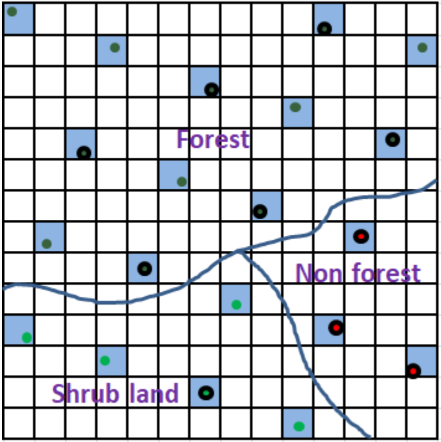


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest). Sub-sampling fraction f_{2g} in the category g.

Second-phase inclusion density:

$$\pi_{2p}(x) = \pi_{1p}(x)\ f_{2g} \text{ for } x \in S^{A}_{1p,g}.$$

French annual sample: second-phase sampling design



The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest). Sub-sampling fraction f_{2g} in the category g.

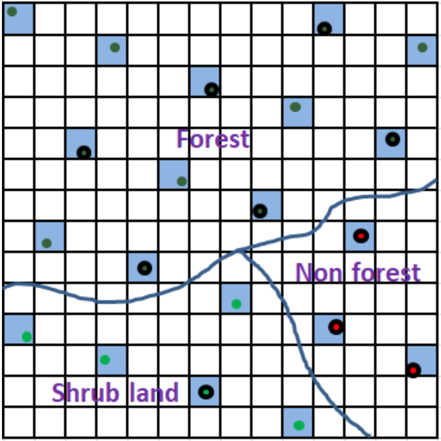
Second-phase inclusion density:

$$\pi_{2p}(x) = \pi_{1p}(x)\ f_{2g} \text{ for } x \in S^A_{1p,g}.$$

Expansion estimator:

$$\hat{\tau}_{y,2p} = \frac{A_F}{n_{1p}} \sum_{g=1}^{G} \frac{1}{f_{2g}} \sum_{x \in S^b_{2p,g}} y^A(x).$$

French annual sample: second-phase sampling design



The estimator

$$\hat{\tau}_{y,2p} = \frac{A_F}{n_{1p}} \sum_{g=1}^{G} \frac{1}{f_{2g}} \sum_{x \in S^b_{2p,g}} y^A(x)$$

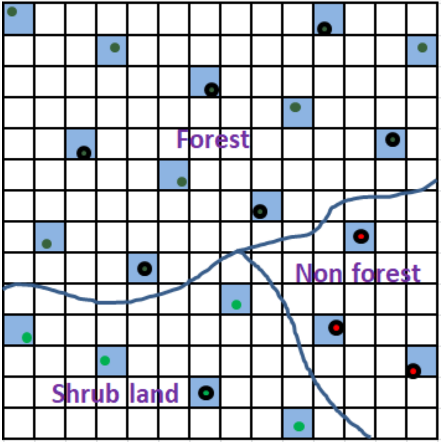
is post-stratified using 1st phase information:

$$\hat{\tau}_{y,post} = \frac{A_F}{n_{1p}} \sum_{g=1}^{G} \frac{n_{1p,g}}{n_{2p,g}} \sum_{x \in S^b_{2p,g}} y^A(x).$$

For example:

$$n_{1p,Forest} = 12,$$
$$n_{2p,Forest} = 6.$$

French annual sample: variance estimation



Unbiased variance estimation is not possible (one point per cell selected).
We compute a variance estimator with two components.

One accounts for the two-stage first-phase design. Expected to improve on the classical uniform random sampling approximation.

One accounts for the second-phase design and poststratification (Duong, Bouriaud and Chauvet, 202X).

Current/future work

Current work

Evaluating the variance estimation method by Monte Carlo simulations. The simulation results (not presented here) seem quite good.

Bootstrap variance estimation would be a nice addition, does not seem to be so much considered in NFIs. Actually simpler than in finite population sampling, since you do not need to account for finite population corrections in continuous population sampling.

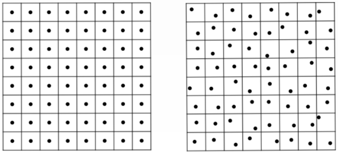
Comparing the sampling strategy used in the French NFI with other techniques, and in particular the spatially balanced sampling of Grafström and Matei (2018) which makes use of the local pivotal method.
Canopy height from LiDAR data (if available at a national scale) could be an auxiliary variable for the sampling design.

More theoretical work

Several papers on the asymptotic properties of estimators for environmental surveys. Barabesi, Franceschi and Marcheselli (2012) studied the consistency and asymptotic normality of sampling designs used in NFIs, but the assumptions on regularity on the local variable do not seem compatible with the synthetic variable obtained under the weight share method.

Recent works of Fattorini, Marcheselli, Pisani and Pratelli (2017, 2020). Consistency of estimators under NFI-like sampling strategies, including systematic unaligned, under weaker assumptions for the local variable.

Asymptotic normality needs to be studied under these weaker assumptions (Chauvet and Pulkkinen, 202X). Also, systematic aligned sampling is difficult to study unless we are willing to use some superpopulation model.



Future work for the french NFI

Accounting for the real features of the sampling design:

- Lack of randomness in the second phase selection. It may lead to a small bias in the post-stratified estimator, but it needs to be checked theoretically/empirically.
- Post-stratification is a bit more intricate than presented. The post-stratification is performed by départements (NUTS3), with possible collapsing of land cover post-strata inside départements.
  Ongoing work of Minna Pulkkinen to compare post-stratification strategies, and we think of the bootstrap to compare different post-stratification strategies.
- Accounting for the third phase of sampling. Some variables of interest (e.g. volume) are imputed for a part of the 2nd phase sample. Not currently clear what the 3rd phase sampling strategy is.

Other future work

Longitudinal estimations in the French NFI.
The points selected/surveyed a given year are revisited 5 years after for longitudinal estimation.
More intricate than for usual longitudinal estimation in social/household surveys, since both the continuous population $\mathcal{U}^A$ and the population of trees U^B are subject to changes in time.

Small area estimation, of course. The NFI would like to make better use of satellite data (LiDAR, GEDI), see Saarela et al (2018).

Thanks for your attention
Any questions?

References

Ardilly, P. and Lavallée, P. (2007). Weighting in rotating samples: the SILC survey in France. Survey Methodology, 33(2), pp. 131-137.

Barabesi, L., Franceschi, S., and Marcheselli, M. (2012). Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation. The Annals of Applied Statistics, 6(1), pp. 210-228.

Chauvet, G., Bouriaud, O. and Brion, P. (2023). An extension of the weight share method when using a continuous sampling frame. Survey Methodology, forthcoming.

Cordy, C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Statistics and Probability Letters, 18(5), pp. 353-362. Deville, J-C. and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. Survey Methodology, 32(2), pp. 165-176.

Fattorini, L., Marcheselli, M., Pisani, C., and Pratelli, L. (2017). Design-based asymptotics for two-phase sampling strategies in environmental surveys. Biometrika, 104(1), pp. 195-205.

Fattorini, L., Marcheselli, M., Pisani, C., and Pratelli, L. (2020). Design-based consistency of the Horvitz-Thompson estimator under spatial sampling with applications to environmental surveys. Spatial Statistics, 35.

Grafström, A., and Matei, A. (2018). Spatially balanced sampling of continuous populations. Scandinavian Journal of Statistics, 45(3), pp. 792-805.

Gregoire, T. G., and Valentine, H. T. (2007). Sampling strategies for natural resources and the environment. CRC Press.

Herv'e, J-C. (2017). National forest inventories, assessment of wood availability and use - France. Synthesis of EU Cost action 1001, Springer, pp. 385-404.

Mandallaz, D. (2007). Sampling techniques for forest inventories. CRC Press.

Pisani, C. and Di Biase, R.M. and Marcheselli, M. (2023). Achieving spatial balance without tears: the tessellation sampling schemes. 8th Italian Conference on Survey Methodology, Rende.

Saarela, S., Holm, S., Healey, S. P., Andersen, H.-E., Petersson, H., Prentius, W., Patterson, P. L., Naesset, E., Gregoire, T. G., and Stahl, G. (2018). Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data. Remote Sensing, 10 (11).

Stevens, D.L. and Urquhart, N. S. (2000). Response designs and support regions in sampling continuous domains. Environmetrics, 11(1), pp. 13-41.