Enquêtes longitudinales
Compléments sur la méthode de partage des poids

Guillaume Chauvet

08/10/2025

1 - An extension of the weight share method when using a continuous sampling frame

Application to the French national forest inventory

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, ...)

Sampling in a continuous population

Notations

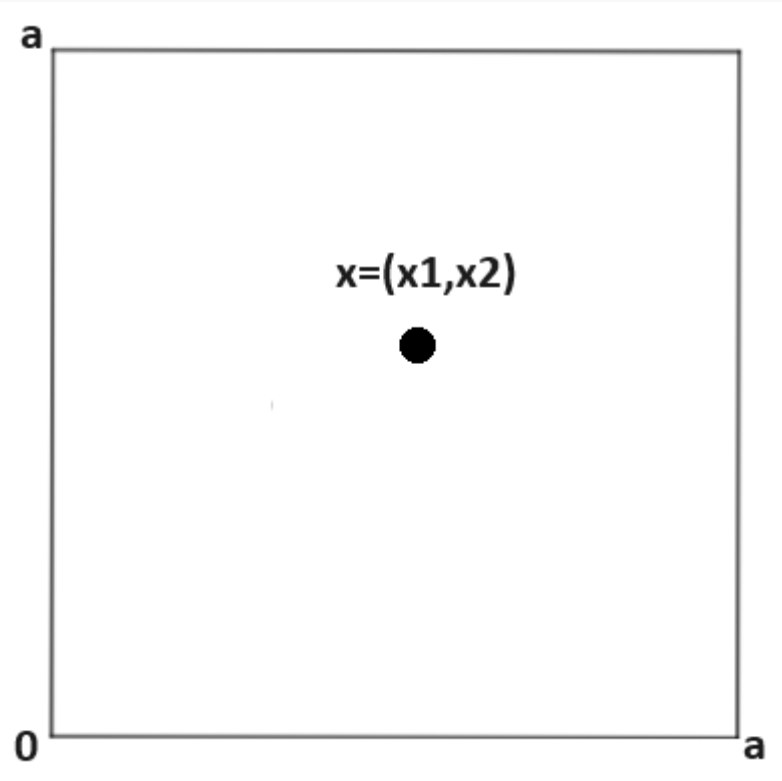
We are interested in a continuous territory $\mathcal{U}^A \subset \mathbb{R}^2$. The *area* of the territory is denoted as A.

We are interested in a *variable of interest* $\rho(\cdot)$, taking the value $\rho(x)$ for point $x \in \mathcal{U}^A$. The variable $\rho(\cdot)$ is also seen as deterministic.

We wish to estimate parameters over the population $\mathcal{U}^A$, like the integral of the variable $\rho(\cdot)$:

$$\tau_\rho = \int_{\mathcal{U}^A} \rho(x) dx.$$

A toy example on a square of length a



$$\begin{aligned} \rho_1(x) &= 1 \\ \int_{\mathcal{U}^A} \rho_1(x)dx &= A = a^2 \end{aligned}$$

$$\begin{aligned} \rho_2(x) &= x_1 \\ \int_{\mathcal{U}^A} \rho_2(x)dx &= \frac{a^3}{2} \end{aligned}$$

$$\begin{aligned} \rho_3(x) &= x_1 x_2 \\ \int_{\mathcal{U}^A} \rho_3(x)dx &= \frac{a^4}{4}. \end{aligned}$$

Sampling design

A random sample $S = \{s_1, \ldots, s_n\}$ of n locations is selected by means of a *continuous sampling design*. We assume the existence of the joint probability density function (PDF) of the sample locations:

$$f(x_1, \ldots, x_n).$$

We also suppose the existence of the marginal PDF $f_i(\cdot)$:

$$f_i(x) = \int f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n.$$

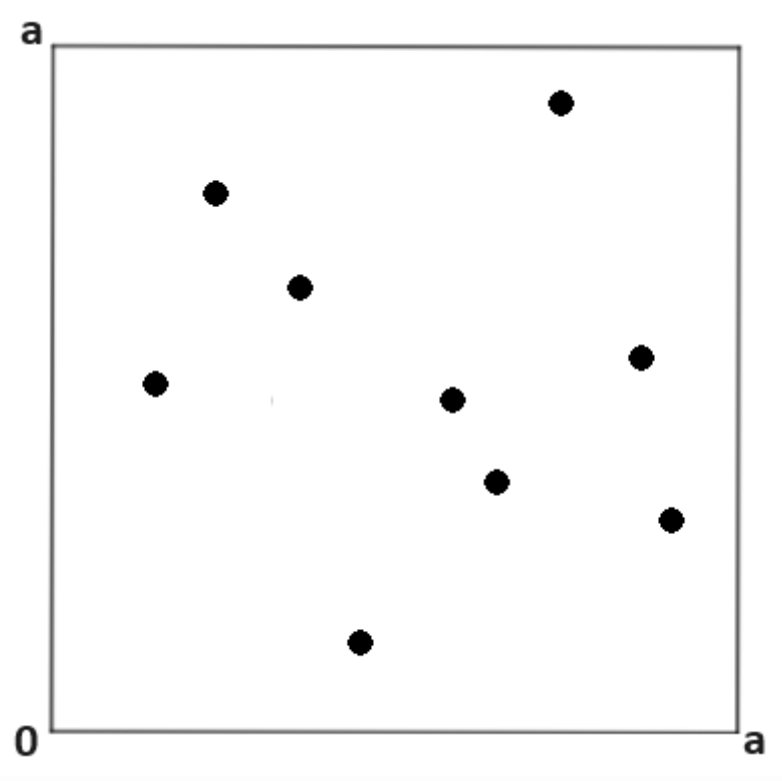
The inclusion density function is defined as

$$\pi(x) = \sum_{i=1}^{n} f_i(x).$$

For any $\mathcal{U}^D \subset \mathcal{U}^A$, $\int_{\mathcal{U}^D} \pi(x)dx$ is the average number of points which is selected in $\mathcal{U}^D$.

Unit square of length a

Uniform sampling of size n



We have

$$
\begin{aligned}
f(x_1, \ldots, x_n) &= \frac{1}{A^n} \prod_{i=1}^{n} 1\left(x_i \in \mathcal{U}^A\right) \\
f_i(x_i) &= \frac{1}{A} 1\left(x_i \in \mathcal{U}^A\right), \\
\pi(x) &= \sum_{i=1}^{n} f_i(x) \\
&= \frac{n}{A} \text{ for } x \in \mathcal{U}^A, \\
\int_{\mathcal{U}^D} \pi(x) dx &= n \frac{A^D}{A}.
\end{aligned}
$$

Horvitz-Thompson estimation

For the estimation of the population integral $\tau_\rho = \int_{\mathcal{U}^A} \rho(x)dx$, we consider the Horvitz-Thompson estimator

$$\hat{\tau}_{\rho\pi} = \sum_{i=1}^{n} \frac{\rho(s_i)}{\pi(s_i)} = \sum_{i=1}^{n} d(s_i)\rho(s_i),$$

with $d(s_i)$ the *sampling weights*.

This is a weighted estimator, which is unbiased for τ_ρ provided that $\pi(x) > 0$ almost everywhere on $\mathcal{U}^A$ (no coverage bias).

Under this condition, it is unbiased for **any variable of interest** collected during the survey.

Uniform sampling

Under uniform sampling of size n, the joint PDF is the same for all the points inside the territory. We obtain

$$f_i(x) = \frac{1}{A} \quad \text{and} \quad \pi(x) = \frac{n}{A} \text{ for any } x \in \mathcal{U}^A.$$

The Horvitz-Thompson estimator is

$$\begin{aligned} \hat{\tau}_{\rho\pi} &= \sum_{i=1}^{n} \frac{\rho(s_i)}{\pi(s_i)} = A\ \bar{\rho}, \\ \text{with } \bar{\rho} &= \frac{1}{n} \sum_{i=1}^{n} \rho(s_i) \text{ the sample mean.} \end{aligned}$$

An unbiased variance estimator is

$$\begin{aligned} \hat{V}(\hat{\tau}_{\rho\pi}) &= A^2 \frac{s_\rho^2}{n} \\ \text{with } s_\rho^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \{\rho(s_i) - \bar{\rho}\}^2 \text{ the sample dispersion.} \end{aligned}$$

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

$$\tau_y^B = \sum_{k \in U^B} y_k^B, \tag{1}$$

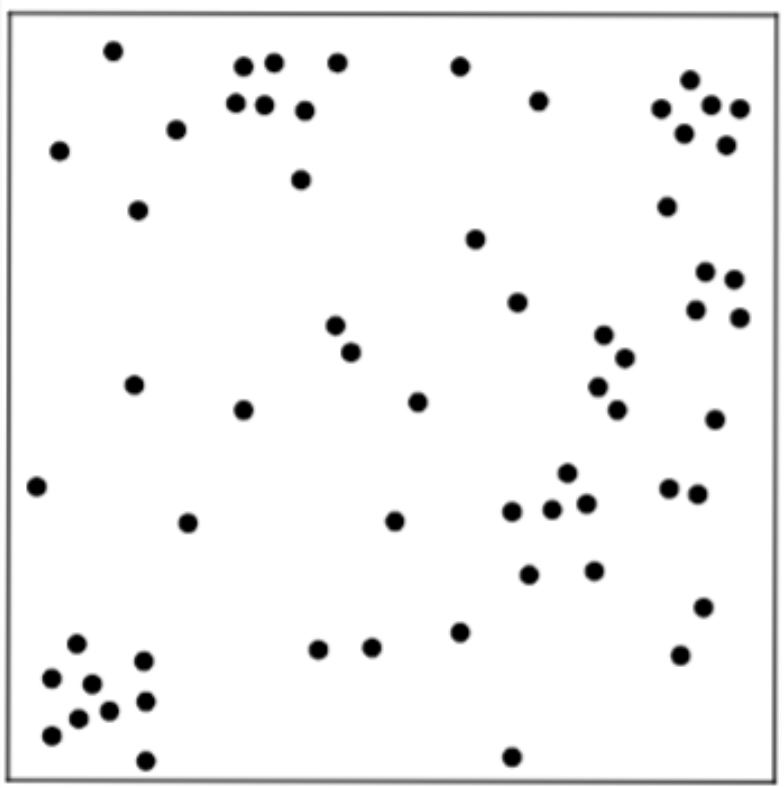
which may be the total volume of wood, for example.

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

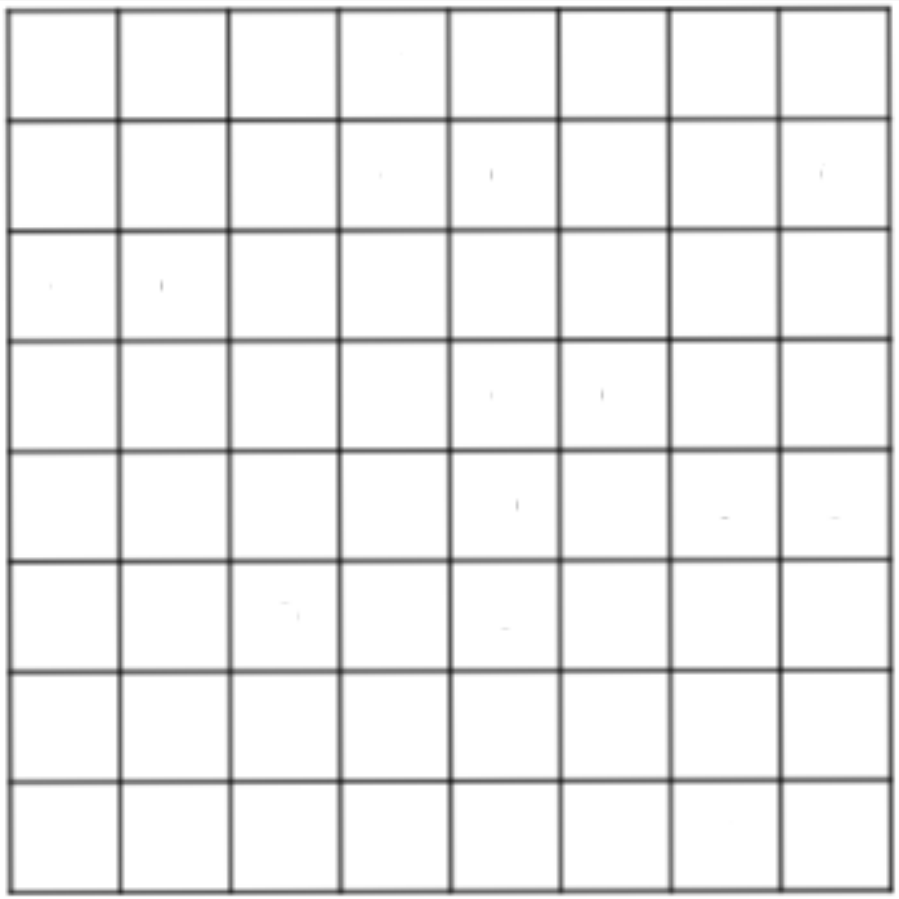


The 1st phase sample may be obtained by independent uniform random sampling inside the territory $\mathcal{U}^A$.

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



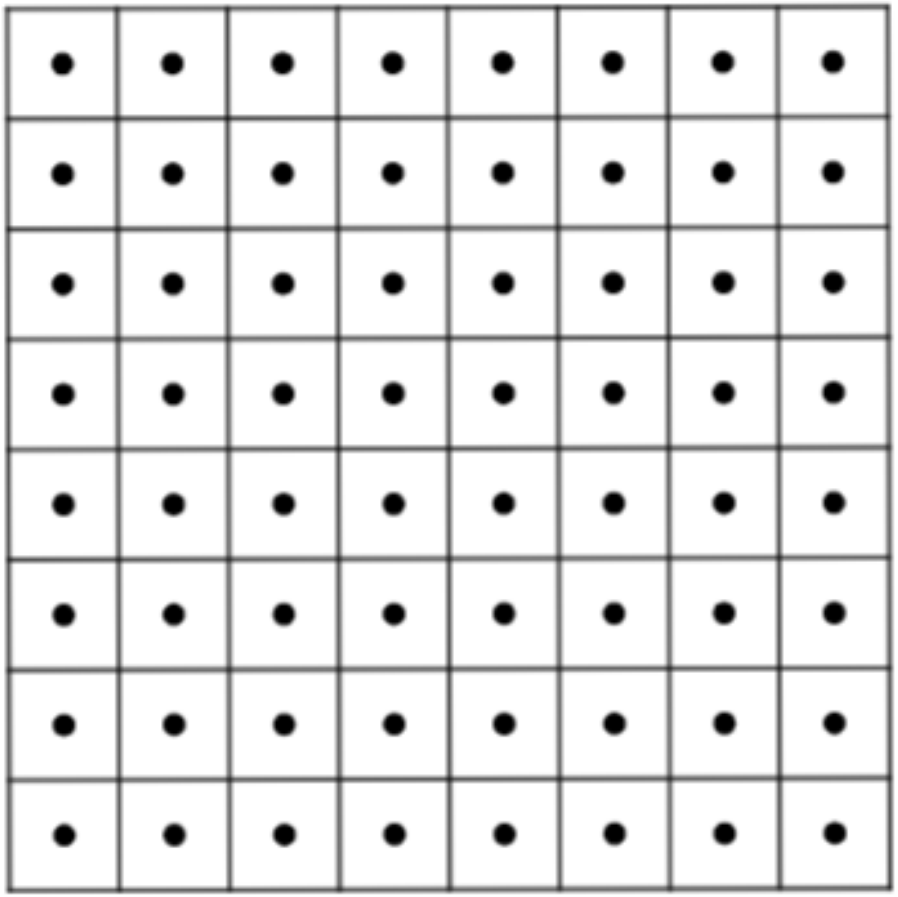
Grid sampling is very common in forest inventories.

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

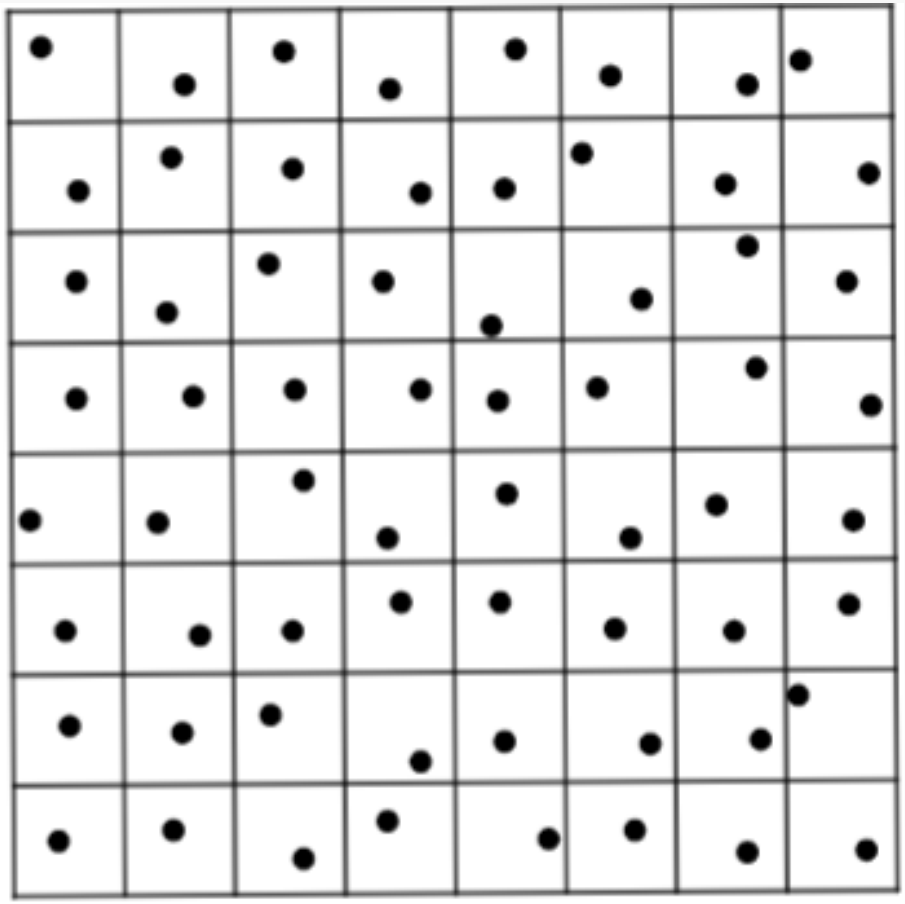
Grid sampling is very useful to achieve sample coordination both in space and in time (Pisani, Di Biase, Marcheselli, 2023).

Some possible 1st-phase sampling designs in NFIs

Spatially systematic aligned sample (Austrian NFI)

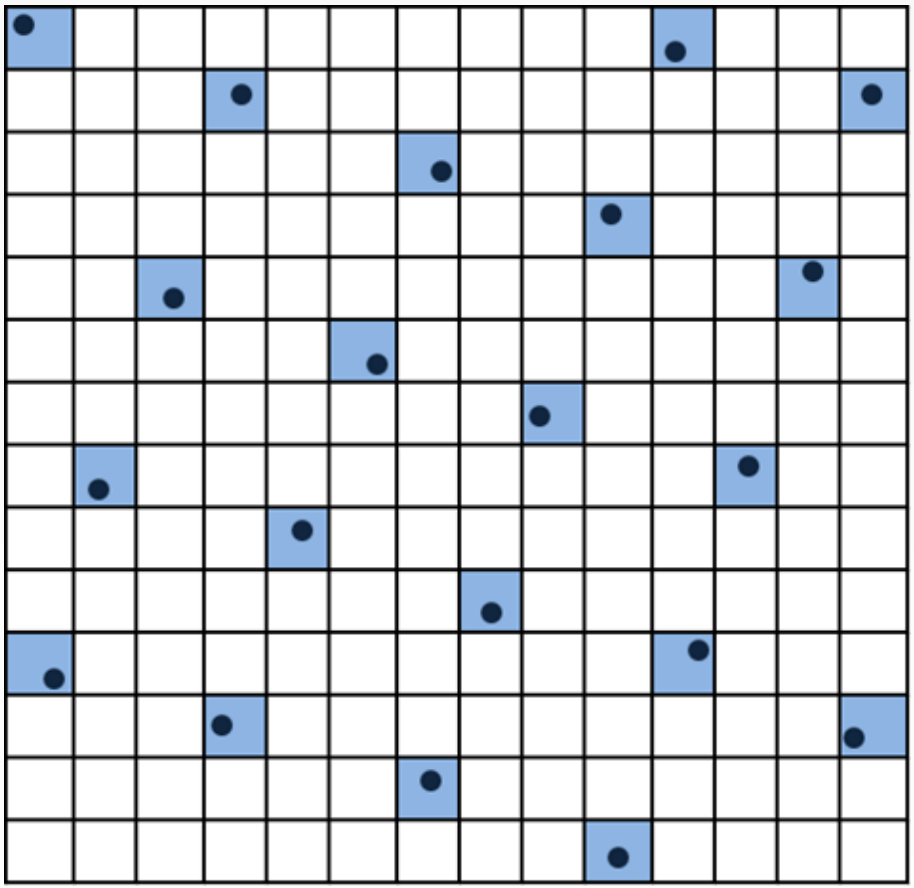


Spatially systematic unaligned sample (Polish and Italian NFI)



Step 1: some possible 1st-phase sampling designs

French annual sample: a two-stage design

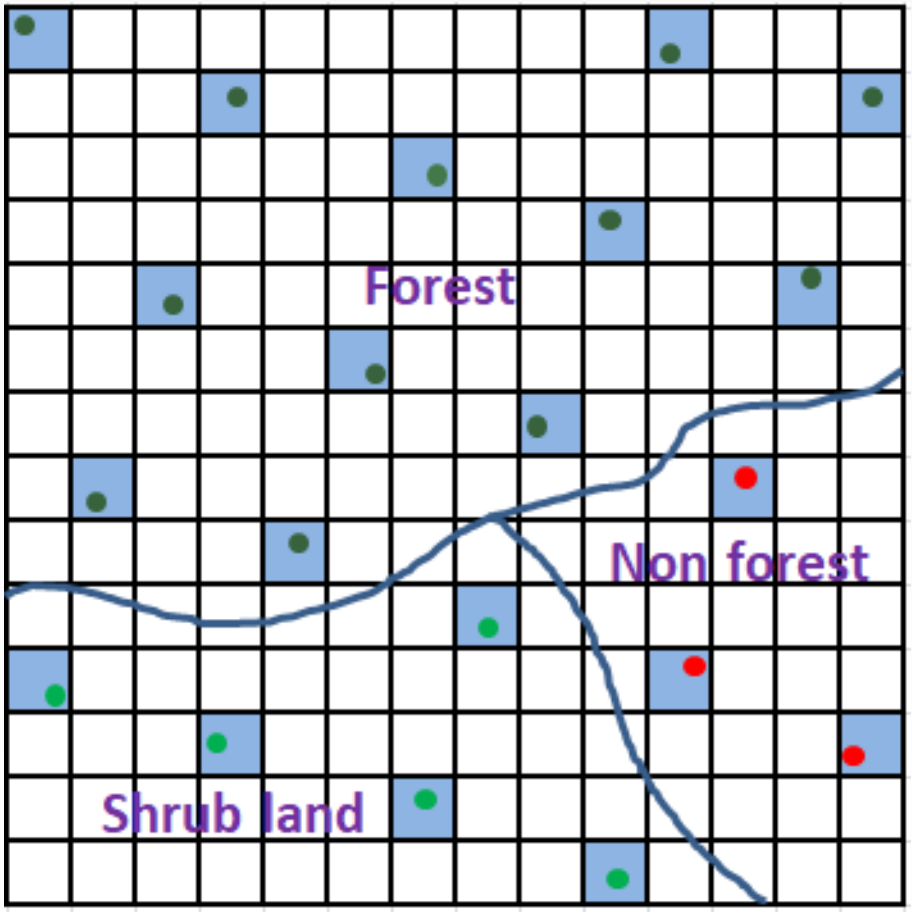


A sample of cells is first selected, by using some form of systematic sampling with equal probabilities.

One point is randomly selected inside each cell.
$\Rightarrow$ 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

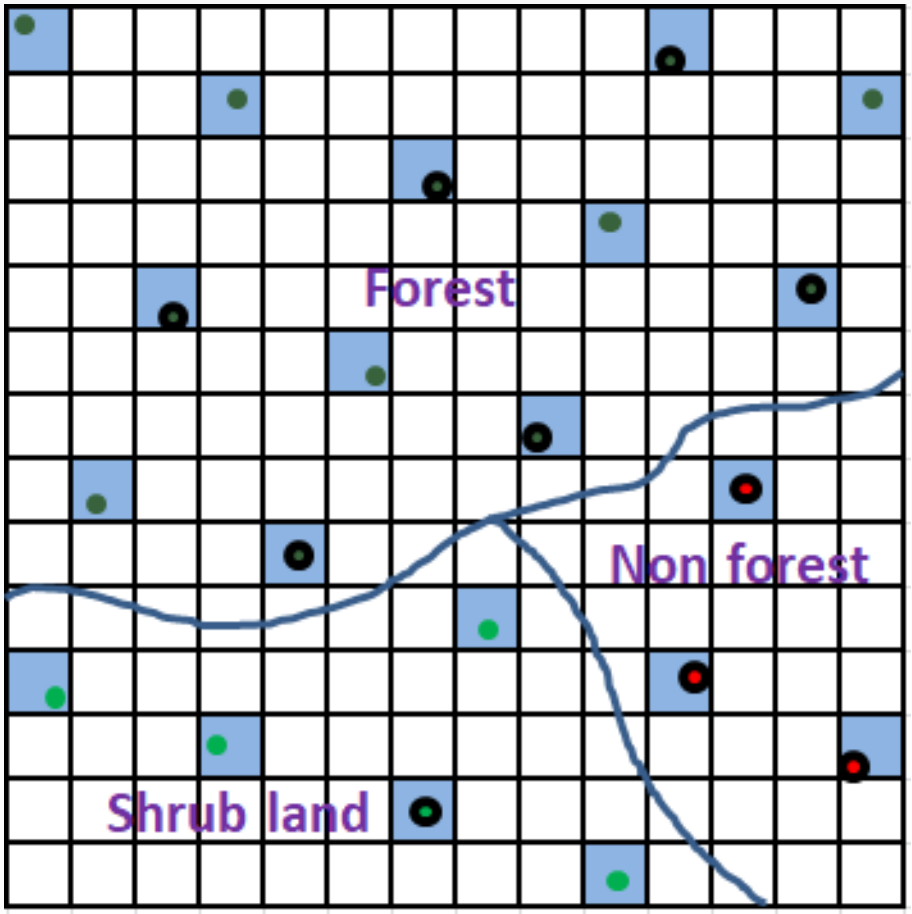


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest).

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

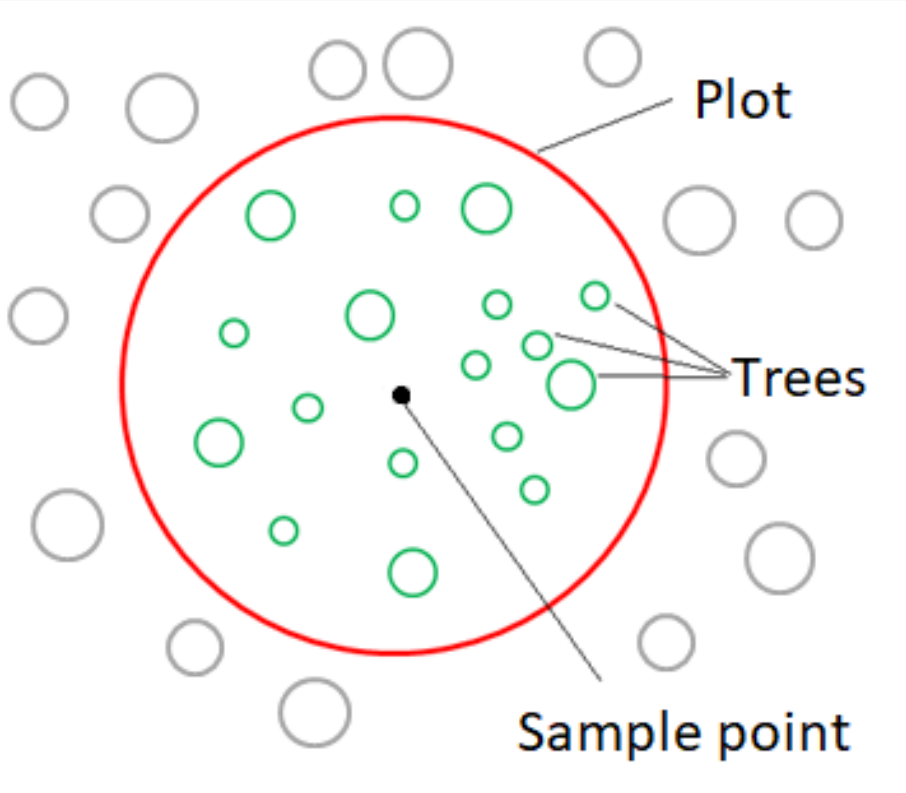


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest).

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.

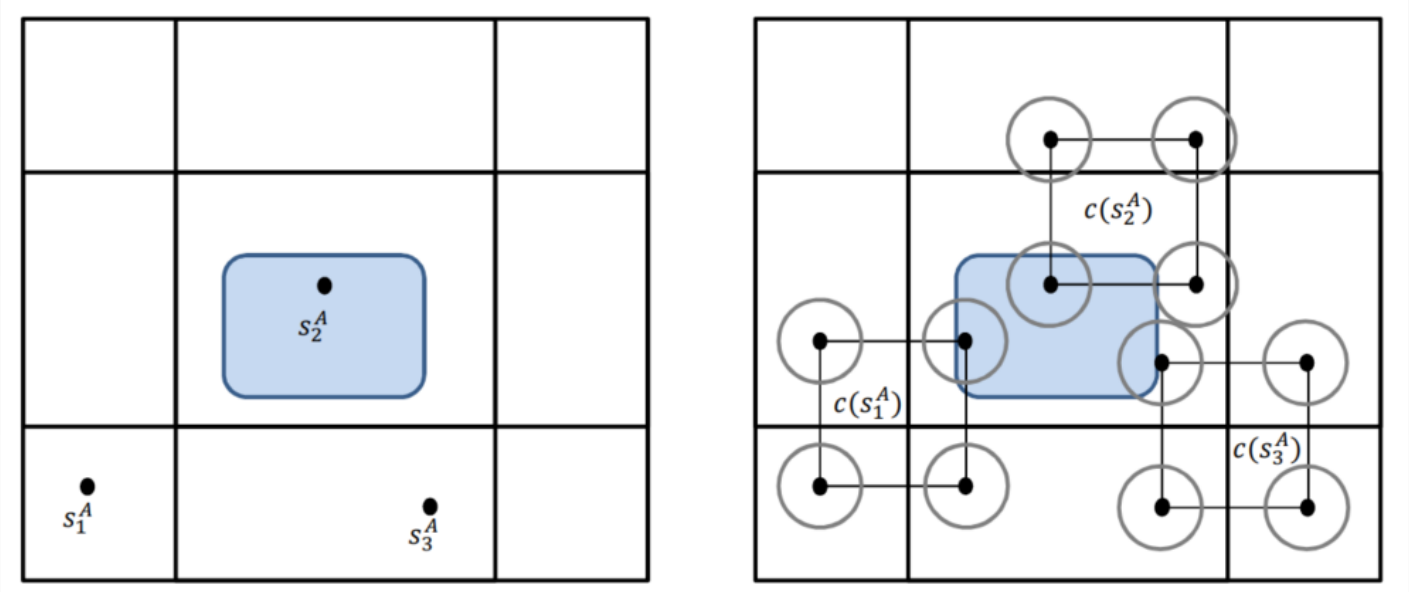
For the French NFI, 3 plot radiuses:

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.

Step 4: use of fixed shape supports

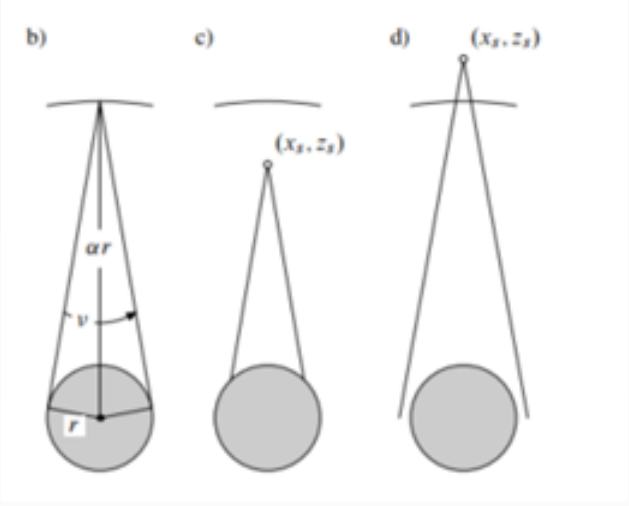
Cluster sampling



Left panel: a sample of points is selected (3 in the example) with the associated plot.
Right panel: satellite plots are selected according to some standard configuration (a square in the example).
Cost reduction in large territories (e.g., Finnish NFI; Tomppo et al., 2011).

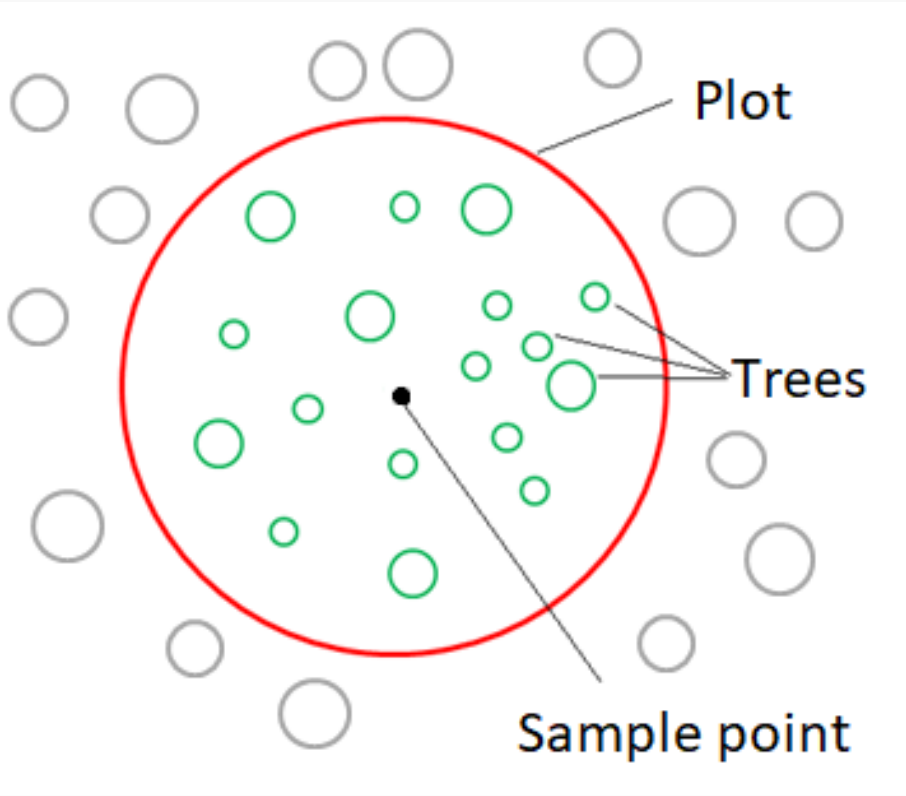
Step 4: use of fixed shape supports

Bitterlich sampling: figure from Gregoire and Valentine (2007, p. 251)



An angle α is used at the sample point.
If the width of the tree fills the field of view on the angle gauge, the tree is selected (panels b and c).
A tree is selected with probability proportional to the basal area.
Used in the Austrian and German NFIs (Gschwantner et al., 2016).

Step 4: use of fixed shape supports

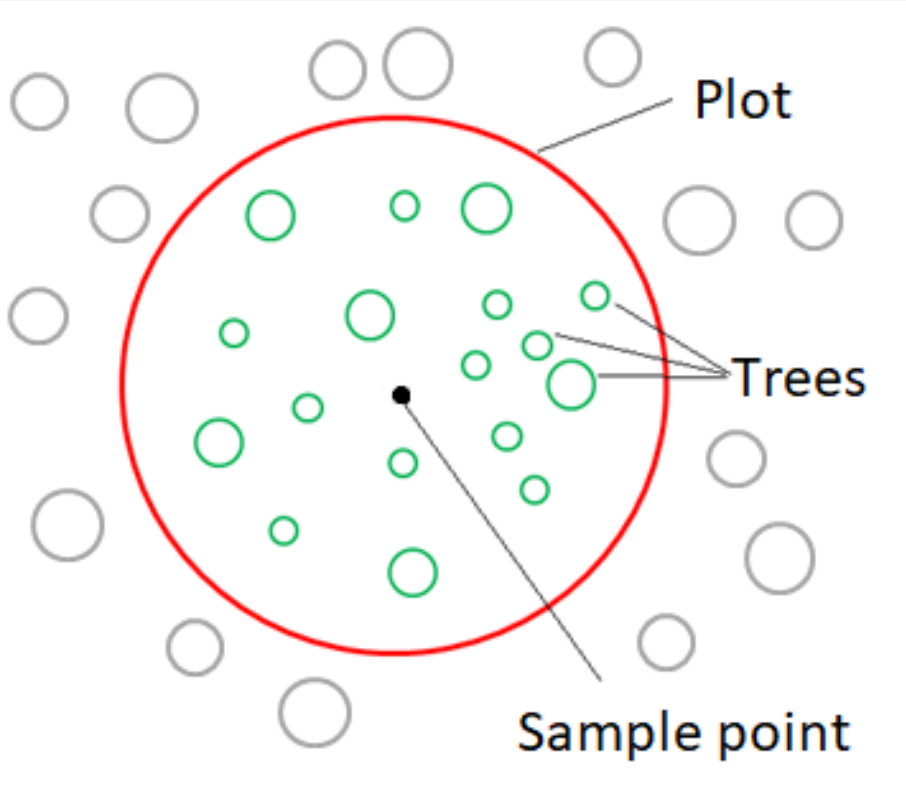


Remark: there can be a third phase of sampling (not covered here).

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

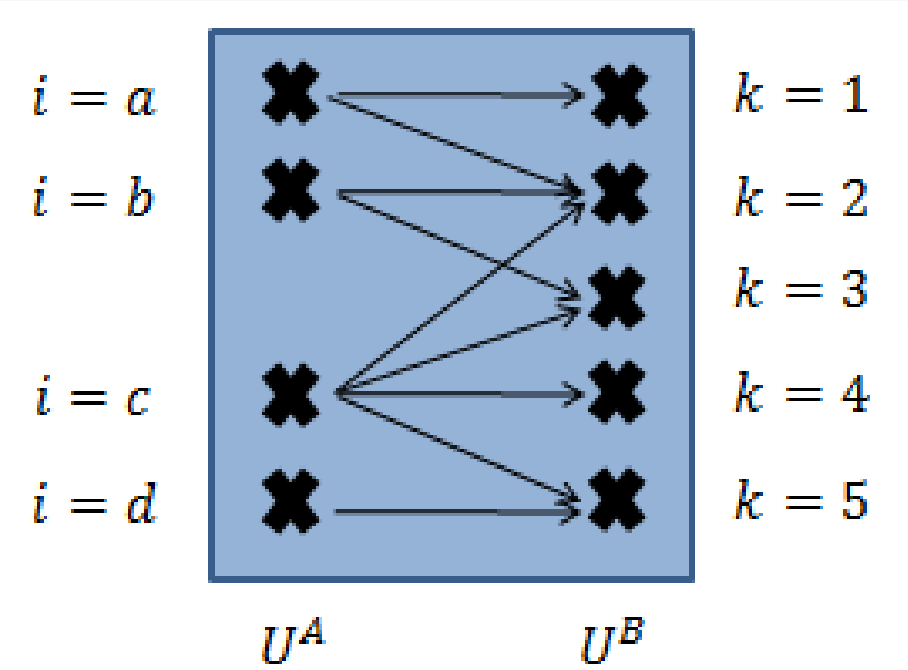
The weight share method (Deville and Lavallée, 2006) is very useful when we wish to perform estimations on a population U^B, by using another population for which a sampling frame is available.

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



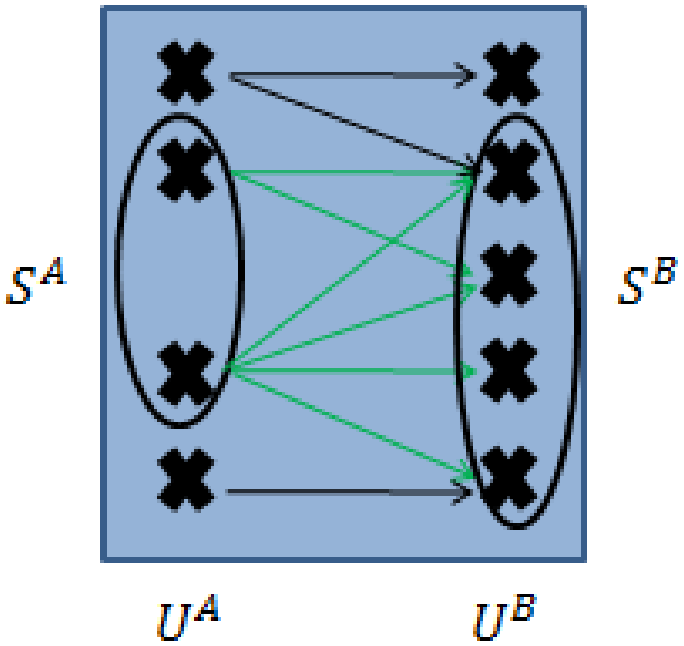
We are interested in a discrete pop. U^B, with a var. of interest y_k^B for which we estimate the total

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

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.

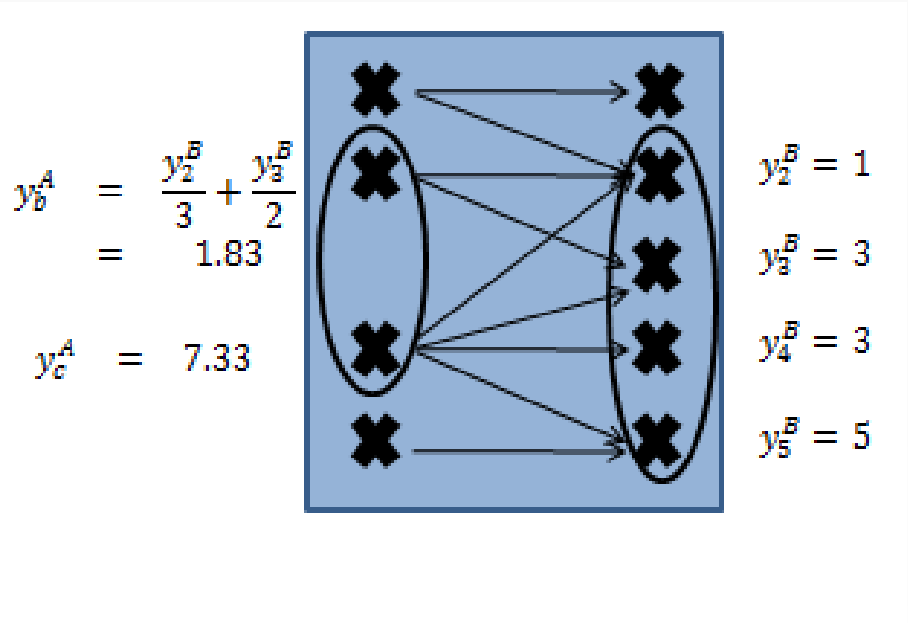
All the units in U^B linked to some unit $i \in S^A$ are surveyed:

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

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

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

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:

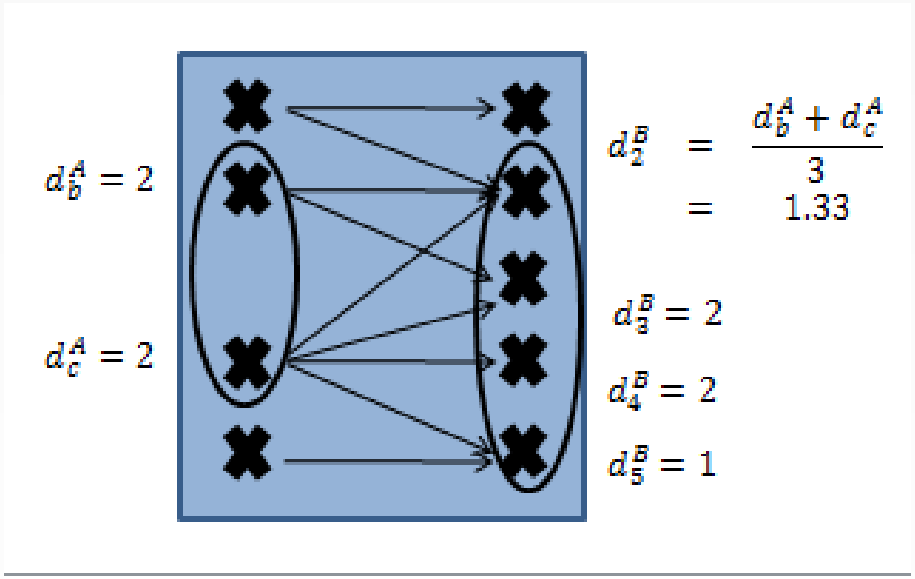
$$\tau_y^B = \sum_{k \in U^B} y_k^B = \sum_{i \in U^A} y_i^A.$$

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

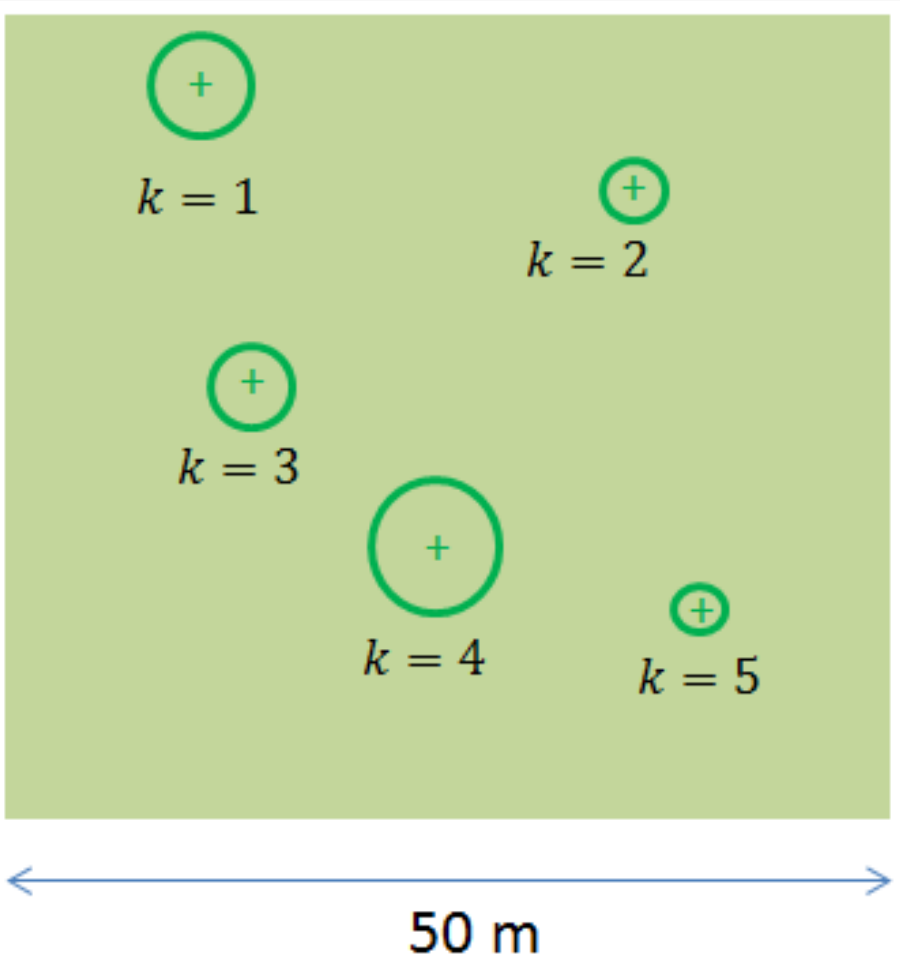
$$d_k^B = \frac{1}{N_{+k}^{AB}} \sum_{i \in S^A} L^{AB}(i,k) d_i^A$$

(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

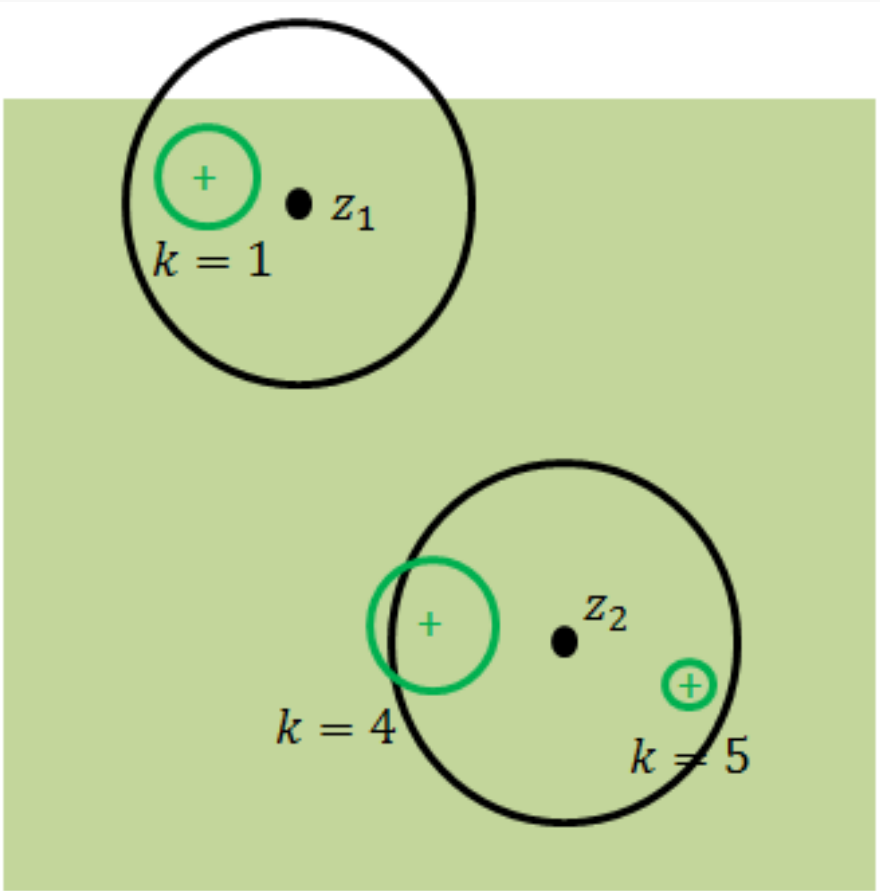
$$\tau_y^B = \sum_{k \in U^B} y_k^B.$$

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

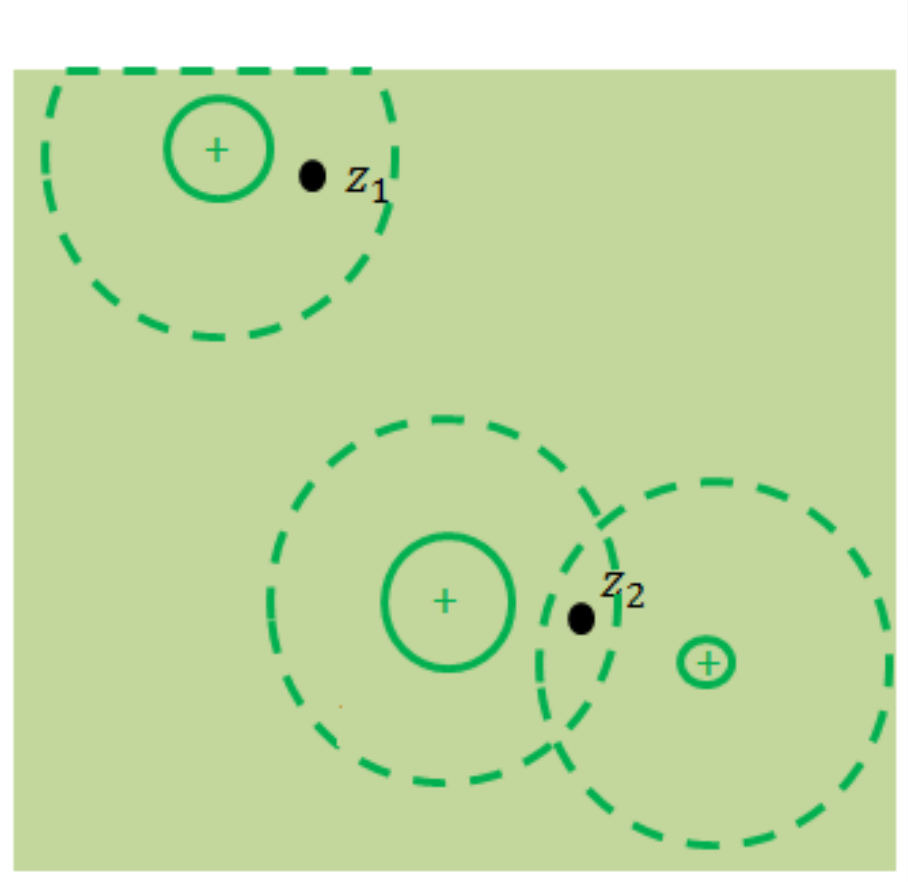
All the trees in U^B linked to the sampled points (i.e., inside the associated plots) are surveyed:

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

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

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

Weight share method: continuous-discrete case



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

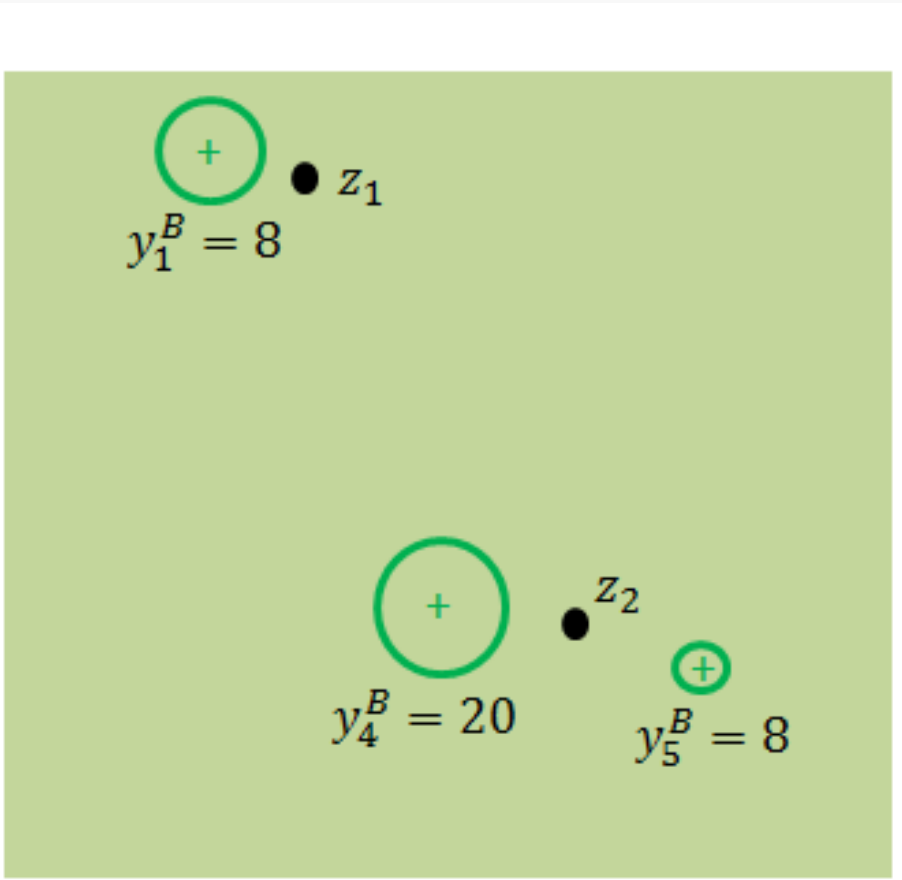
All the trees in U^B linked to the sampled points (i.e., inside the associated plots) are surveyed:

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

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

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

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:

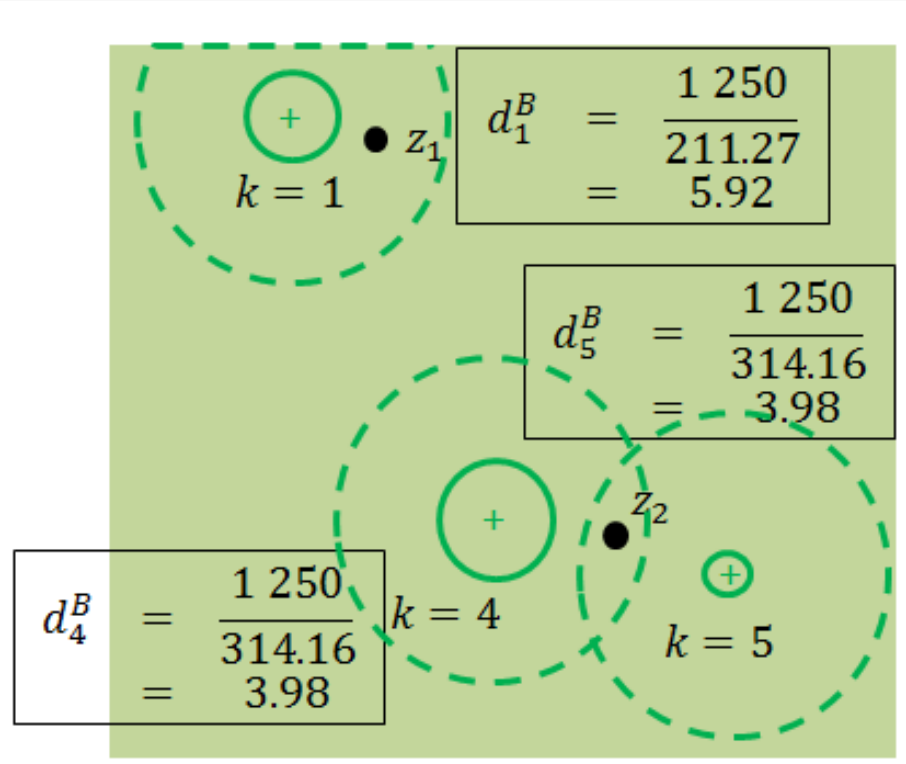
$$\tau_y^B = \sum_{k \in U^B} y_k^B = \int_{x \in \mathcal{U}^A} y^A(x) dx.$$

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:

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

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.

Continuous Horvitz-Thompson estimation

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

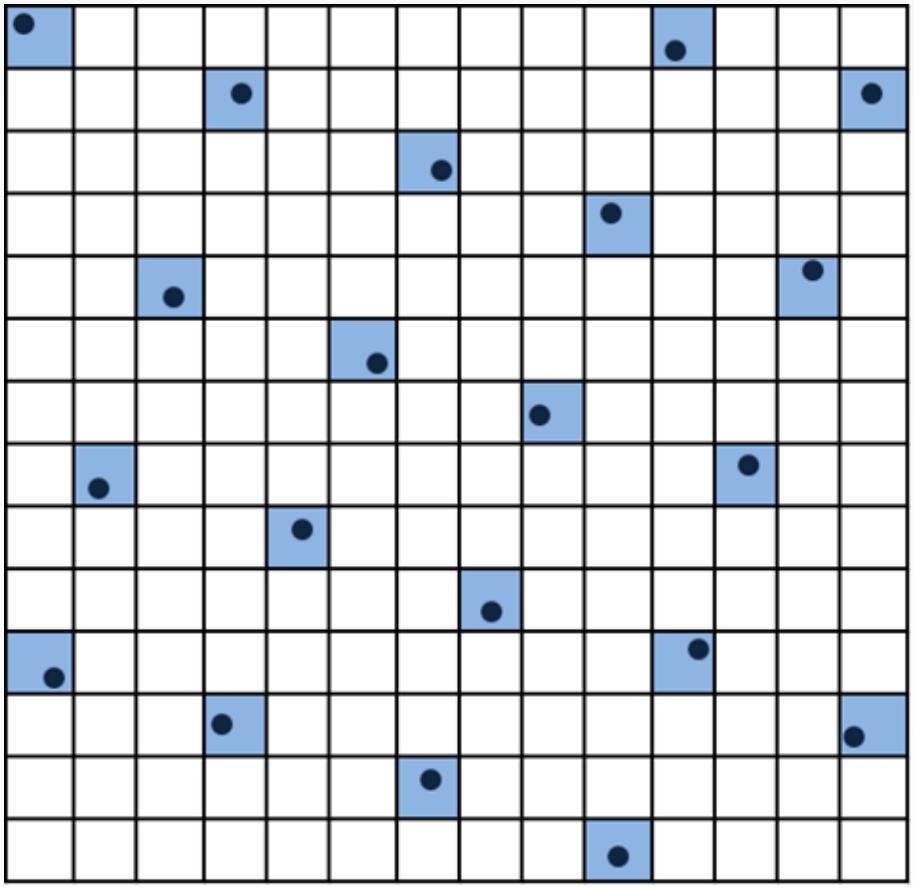
$$\begin{aligned}
\hat{\tau}_y^B &= \sum_{x \in S^A} d^A(x) y^A(x) = \sum_{x \in S^A} \frac{y^A(x)}{\pi^A(x)}, \\
\hat{V}_{YG}\left(\hat{\tau}_y^B\right) &= \frac{1}{2} \sum_{x \neq x' \in S^A} \left\{ \frac{\pi^A(x)\pi^A(x') - \pi^A(x,x')}{\pi^A(x,x')} \right\} \left\{ \frac{y^A(x)}{\pi^A(x)} - \frac{y^A(x')}{\pi^A(x')}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

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



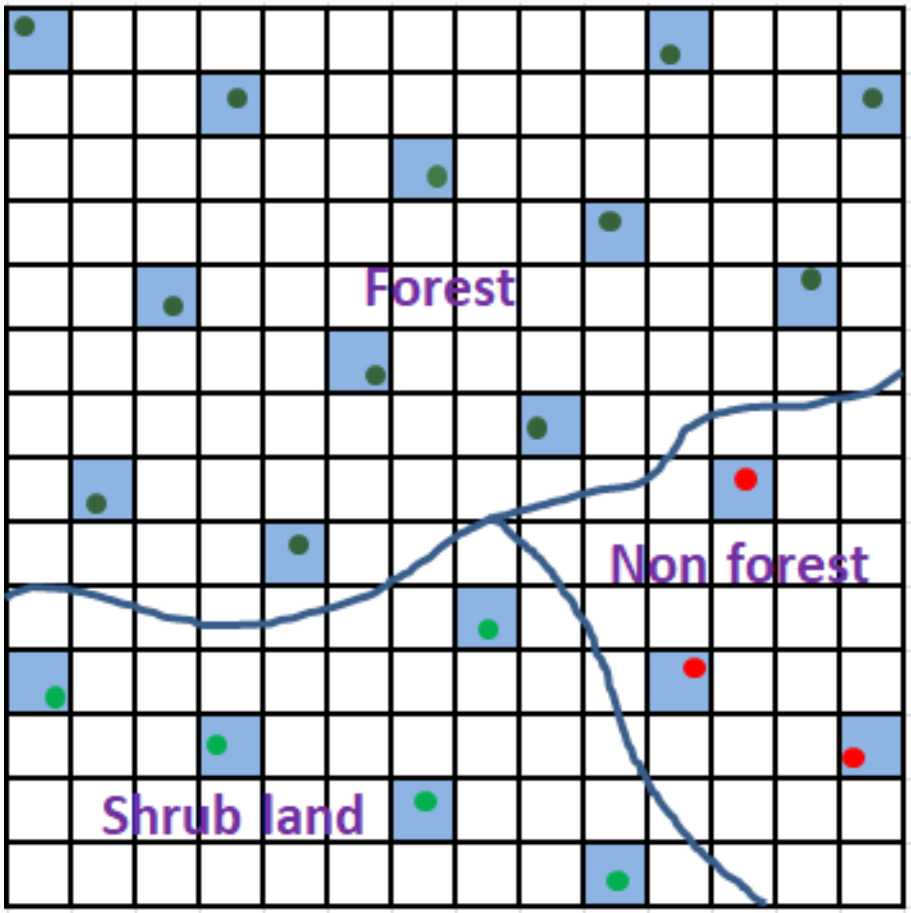
Sample of n_I cells first selected among the N_I cells, with equal probabilities.
One point randomly selected inside each cell of area A_c.

The first-phase inclusion density/HT estimator are

$$\pi_{1p}(x) = \frac{n_I}{N_I} \times \frac{1}{A_c} = \frac{n_{1p}}{A_F} \quad \text{for any } x \in \mathcal{U}^A,$$

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

French annual sample: second-phase sampling design

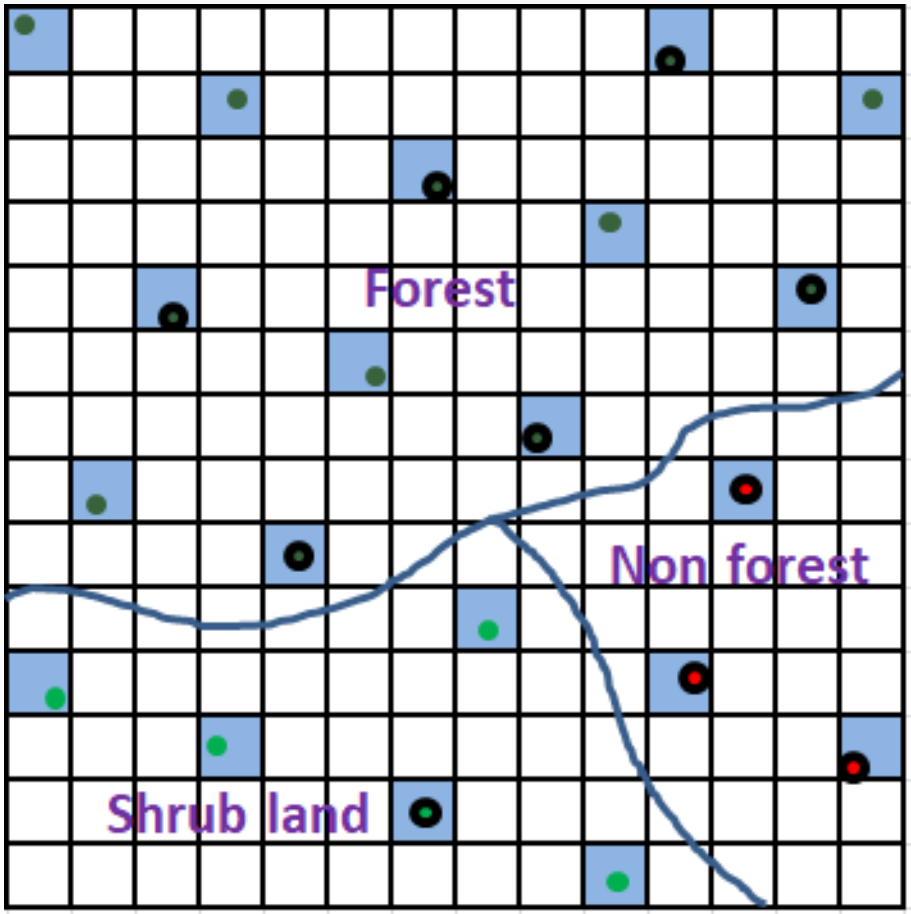


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest). Sub-sampling fraction f_{2g} in the category g.

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.

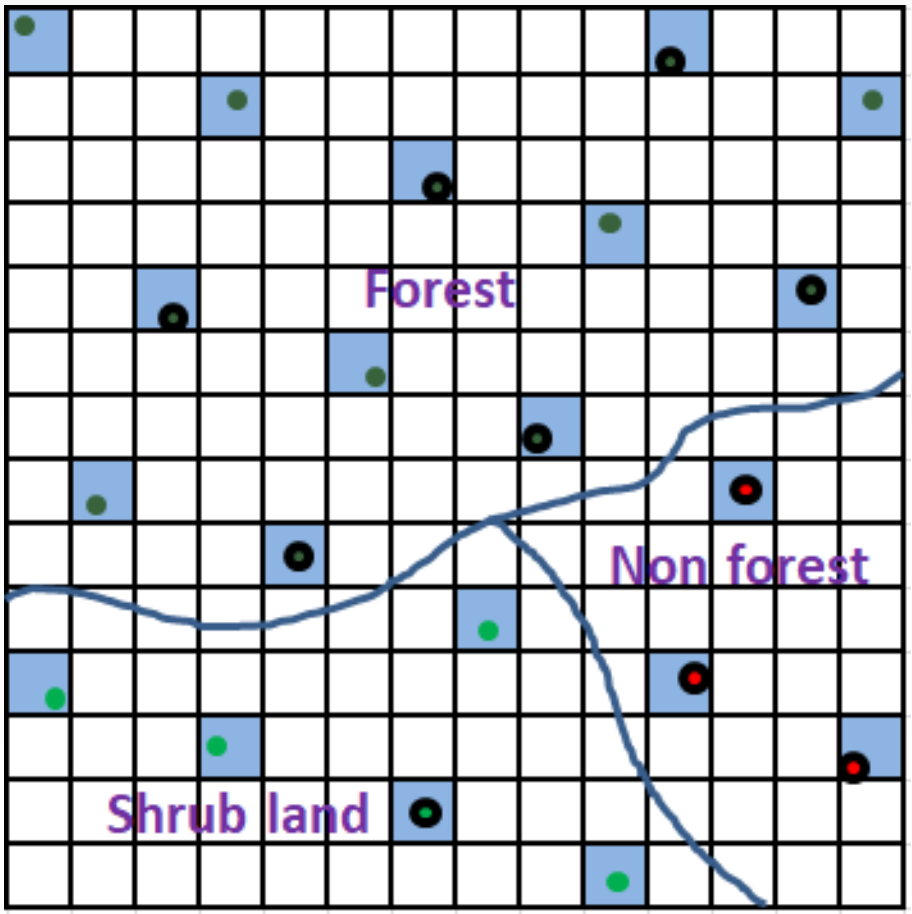
Second-phase inclusion density:

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

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

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

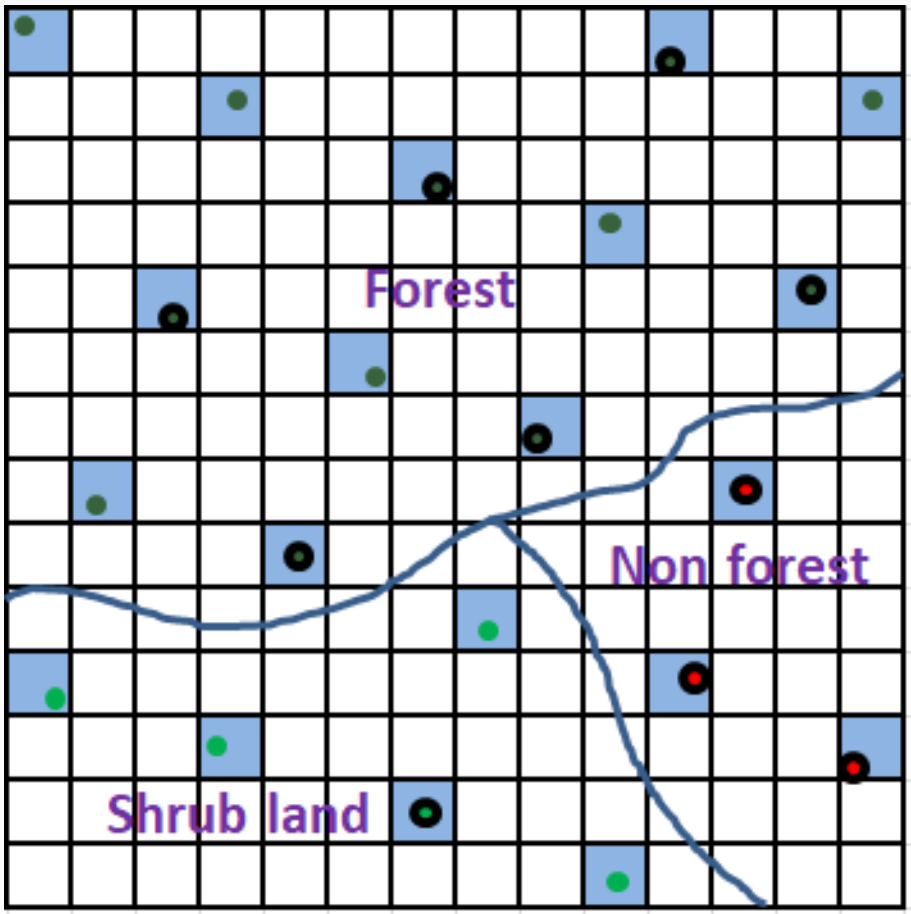
is post-stratified using 1st phase information:

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

For example:

$$\begin{aligned} n_{1p,Shrub} &= 5, \\ n_{2p,Shrub} &= 1. \end{aligned}$$

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, 2025).

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.
Hervé, 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.

2 - Bootstrap pour l'enquête Histoire de Vie et Patrimoine

Qu'est-ce que le bootstrap?

Technique computationnelle (Efron, 1979) qui permet (en théorie) d'estimer la distribution d'un estimateur, en reproduisant de façon répétée le mécanisme de sélection et la procédure d'estimation utilisée.

Dans le cadre des enquêtes, l'objectif est souvent plus simplement de produire un estimateur de variance. Le bootstrap consiste alors :

- à répliquer B fois la création des poids d'extrapolation par rééchantillonnage + réestimation,
- à obtenir ainsi un jeu de B poids bootstrap,
- à les utiliser pour calculer les statistiques bootstrappées.
  Leur dispersion est alors utilisée comme estimateur de variance.

Procédure très simple d'un point de vue utilisateur.

Cadre général

Population finie $U = \{1, \ldots, N\}$. On utilise un plan de sondage $p(\cdot)$ respectant des probas d'inclusion $\pi_k > 0$ pour tirer un échant. aléatoire S. Supposons de plus S sélectionné selon un plan à plusieurs degrés, avec sélection d'un échantillon S_I de n_I unités primaires (UP).

On s'intéresse à un paramètre $\theta = f(t_y)$, où la fonction $f(\cdot)$ est connue mais le p-vecteur des totaux $t_y = \sum_{k \in U} y_k$ est inconnu. On l'estime par substitution

$$\begin{aligned} \hat{\theta} &= f(\hat{t}_{y\pi}) \\ \text{avec } \hat{t}_{y\pi} &= \sum_{k \in S} d_k y_k = \sum_{u_i \in S_I} \sum_{k \in S_i} d_{I|i} d_{k|i} y_k. \end{aligned}$$

Bootstrap avec remise des unités primaires

Cas particulier de la méthode de Rao, Wu et Yue (1992), mais qui semble la plus utilisée en pratique (Rust and Rao, 1996; Yeo et al., 1999).

La méthode de bootstrap procéde ainsi :

1. On sélectionne un échantillon avec remise et à probabilités égales de $n_I - 1$ unités dans S_I.
2. On donne à l'UP u_i le poids bootstrap de sondage :

$$d^*_{Ii} = \frac{n_I}{n_I - 1} \times \{\text{Nb de rééchantillonnages de } u_i\} \times d_{Ii}$$

3. On reproduit la chaîne d'estimation selon les deux principes suivants :
   - Les étapes d'échant./NR post 1er degré ne sont pas bootstrappées.
   - Les étapes d'estim. (correction de la NR, calage) sont bootstrappées.

Bootstrap avec remise des unités primaires

Cette méthode donne un estimateur sans biais de la variance si les UP sont sélectionnées avec remise: cf Bessonneau et al. (2021), où nous expliquons quel est l'estimateur de variance de référence que l'on cherche à reproduire.

Quand les UP sont sélectionnées sans remise, cette méthode :

- surestime la variance du premier degré,
- estime correctement la variance due aux étapes suivantes de tirage et de non-réponse.

Jeu de données

Nous allons utiliser un jeu de données artificiel pour expliquer le fonctionnement de la méthode du rescaled bootstrap. Il est constitué d'un échantillon de $n_{hh} = 10$ ménages tiré dans une population U_{hh} de taille $N_{hh} = 100$. L'échantillon est noté

$$S_{hh} = \{A, B, C, D, E, F, G, H, I, J\}.$$

Nous commencerons par considérer le cas où seul l'échantillonnage des ménages est pris en compte. Nous prendrons ensuite en compte la non-réponse et le calage des estimateurs.

Cas d'un sondage aléatoire simple sans remise

Supposons que S_{hh} est tiré selon un sondage aléatoire simple sans remise. Nous sommes dans le cas particulier d'un seul degré de tirage (un ménage = une unité primaire). Nous avons

$$d_k = \frac{N_{hh}}{n_{hh}} = 10 \quad \text{pour tout } k \in S_{hh}.$$

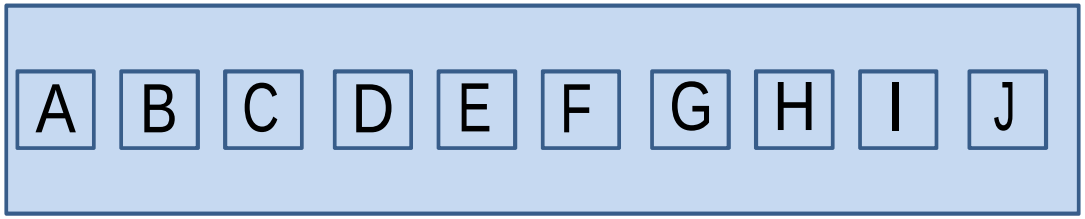
La méthode de bootstrap consiste :

1. à sélectionner un échantillon avec remise et à probabilités égales de $n_{hh} - 1 = 9$ unités dans S,
2. à donner à chaque ménage k le poids bootstrap de sondage :

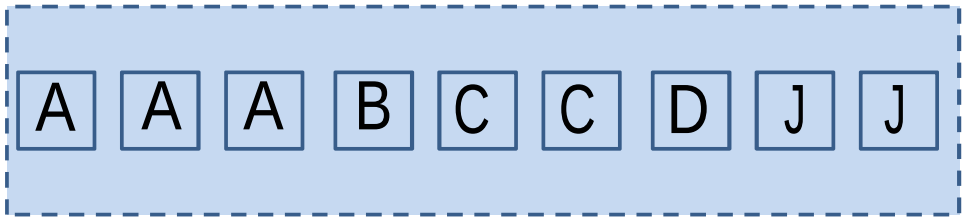
$$d_k^* = \underbrace{\frac{n_{hh}}{n_{hh}-1} \times \{\text{Nb de rééchantillonnages de } k\}}_{G_k} \times d_k$$

Sondage aléatoire simple sans remise

Un exemple de rééchantillon bootstrap



$$d_k = \frac{100}{10} = 10$$



$$d_{A*} = \frac{10}{9} * 3 * 10 \qquad d_{B*} = d_{D*} = \frac{10}{9} * 1 * 10$$

$$d_{C*} = d_{J*} = \frac{10}{9} * 2 * 10$$

$$d_{E*} = d_{F*} = d_{G*} = d_{H*} = 0$$

Cas d'un tirage à probabilités inégales

Supposons que S_{hh} est sélectionné selon un tirage à probabilités inégales à un seul degré. Nous sommes toujours dans le cas particulier d'un seul degré de tirage (un ménage = une unité primaire). Supposons par exemple que :

$$d_A = d_B = d_C = d_D = 4 \quad \text{et} \quad d_E = d_F = d_G = d_H = d_I = d_J = 16.$$

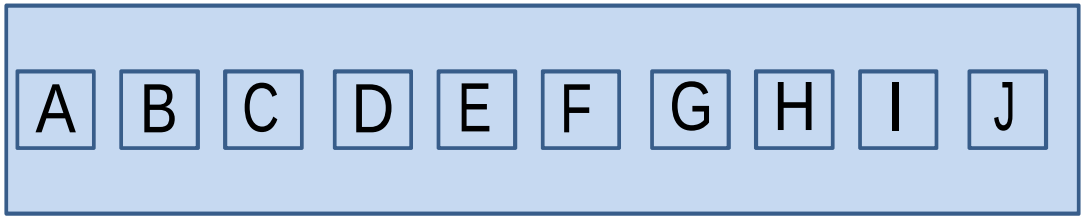
La méthode de bootstrap consiste :

1. à sélectionner un échantillon avec remise et à probabilités égales de $n_{hh} - 1 = 9$ unités dans S,
2. à donner à chaque ménage k le poids bootstrap de sondage :

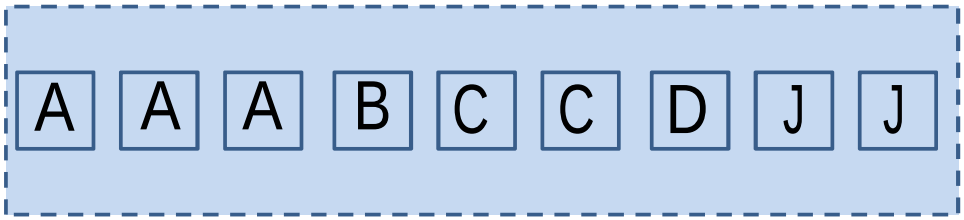
$$d_k^* = \underbrace{\frac{n_{hh}}{n_{hh}-1} \times \{\text{Nb de rééchantillonnages de } k\}}_{G_k} \times d_k$$

Tirage à probabilités inégales

Un exemple de rééchantillon bootstrap

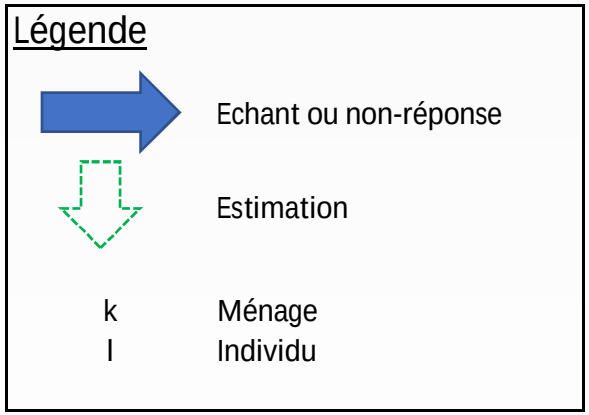


$$d_A = d_B = d_C = d_D = 4$$
$$d_E = d_F = d_G = d_H = d_I = d_J = 16$$

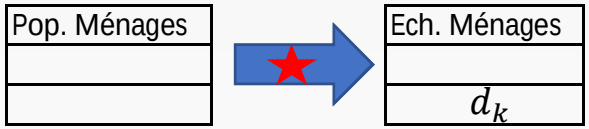


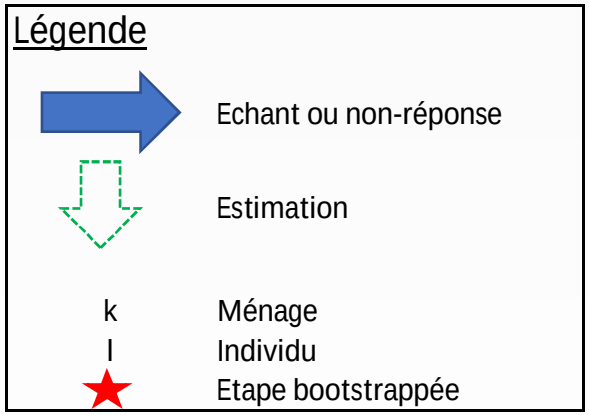
$$d_{A*} = \frac{10}{9} * 3 * 4 \qquad d_{B*} = d_{D*} = \frac{10}{9} * 1 * 4$$
$$d_{C*} = \frac{10}{9} * 2 * 4 \qquad d_{J*} = \frac{10}{9} * 2 * 16$$
$$d_{E*} = d_{F*} = d_{G*} = d_{H*} = 0$$

Echantillonnage de ménages



Echantillonnage de ménages : bootstrap





Algorithmes pour l'estimation de variance bootstrap

Algorithme 1 Etape de bootstrap pour le tirage des ménages

- Etape 1: calcul du poids bootstrap G_k
- Etape 2: prise en compte du tirage de ménages en calculant pour tout $k \in S_{hh}$, le poids bootstrap de tirage $d_{k*} = G_k d_k$.

Algorithme 2 Estimation de variance bootstrap pour le tirage des ménages

1. Répéter B fois les étapes 1 et 2 de l'Algorithme 1. Calculer :

$$\hat{\theta}^{*1} = f(\hat{t}_{y\pi}^{*1}), \cdots, \hat{\theta}^{*B} = f(\hat{t}_{y\pi}^{*B}).$$

2. L'estimateur de variance bootstrap de $\hat{\theta}$ est

$$v_{boot}^{B}\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\theta}^{*b} - \frac{1}{B}\sum_{b'=1}^{B}\hat{\theta}^{*b'}\right\}^2.$$

Prise en compte de la non-réponse et du calage

Dans la suite, nous traitons seulement le cas où S_{hh} est tiré selon un sondage aléatoire simple sans remise :

$$d_k = \frac{N_{hh}}{n_{hh}} = 10 \quad \text{pour tout } k \in S_{hh}.$$

Nous considérons également :

- que l'échantillon souffre de non-réponse totale, corrigé selon la méthode des Groupes Homogènes de Réponse (GHR),
- que les poids corrigés de la non-réponse sont calés sur des totaux auxiliaires $t_x = \sum_{k \in U_{hh}} x_k$.

Non-réponse totale

En raison de la non-réponse totale, seul un sous-échantillon $S_{r,hh}$ de ménages répondants est observé. Soit r_k l'indicatrice de réponse du ménage k.

Nous supposons l'utilisation du modèle des GHR, selon lequel l'échantillon peut être divisé en C groupes $S_{hh}^1, \ldots, S_{hh}^C$, au sein desquels la probabilité de réponse p_{hh}^c est constante.

L'estimateur corrigé de la non-réponse totale est

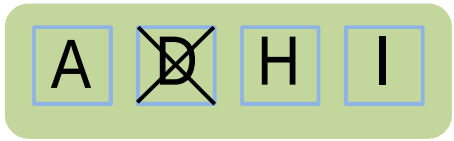
$$\hat{t}_{yr} = \sum_{k \in S_{r,hh}} \frac{d_k y_k}{\hat{p}_{hh}^{c(k)}} \equiv \sum_{k \in S_{r,hh}} d_{rk} y_k \quad \text{avec} \quad \hat{p}_{hh}^c = \frac{\sum_{k \in S_{hh}^c} \omega_k r_k}{\sum_{k \in S_{hh}^c} \omega_k}.$$

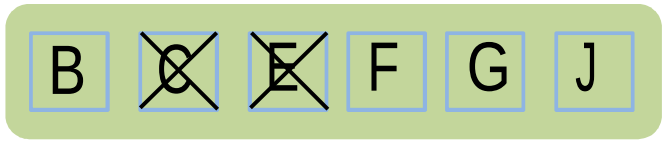
Le choix $\omega_k = 1$ est le plus classique (fréquence de réponse non pondérée dans les GHR).

Echantillon de ménages

Traitement de la non-réponse totale

$$d_k = \frac{100}{10} = 10$$





$$\boldsymbol{RHG\ c = 1}$$
$$\hat{p}_{hh}^{1} = \frac{3}{4}$$
$$d_{rk} = \frac{40}{3}$$

$$\boldsymbol{RHG\ c = 2}$$
$$\hat{p}_{hh}^{2} = \frac{4}{6}$$
$$d_{rk} = \frac{30}{2}$$

Calage

Les poids corrigés de la non-réponse totale sont calés sur un vecteur t_x de totaux auxiliaires. Nous avons

$$w_k = d_{rk} F(\lambda^\top x_k) \quad \text{avec} \quad \sum_{k \in S_{r,hh}} d_{rk} F(\lambda^\top x_k) x_k = t_x,$$

et $F(\cdot)$ une fonction de calage.

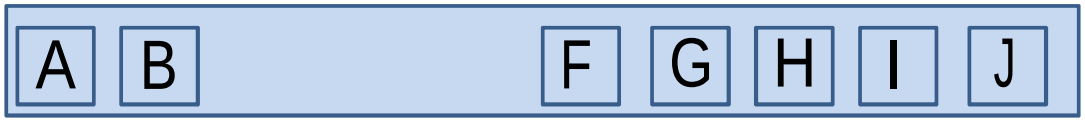
Par exemple, la méthode linéaire correspond à $F(x) = 1 + x$ et conduit à l'estimateur GREG

$$\hat{t}_{yw} = \hat{t}_{yr} + \hat{b}_r^\top \left(t_x - \hat{t}_{xr} \right)$$

$$\text{avec } \hat{b}_r = \left(\sum_{k \in S_{r,hh}} d_{rk} x_k x_k^\top \right)^{-1} \sum_{k \in S_{r,hh}} d_{rk} x_k y_k.$$

Echantillon de ménages

Calage



$$d_{rA} = d_{rH} = d_{rI} = \frac{40}{3} \qquad d_{rB} = d_{rF} = d_{rG} = d_{rJ} = \frac{30}{2}$$

Vecteur de variables auxiliaires $x_k = \left(x_{1k}, \dots, x_{pk}\right)^T$

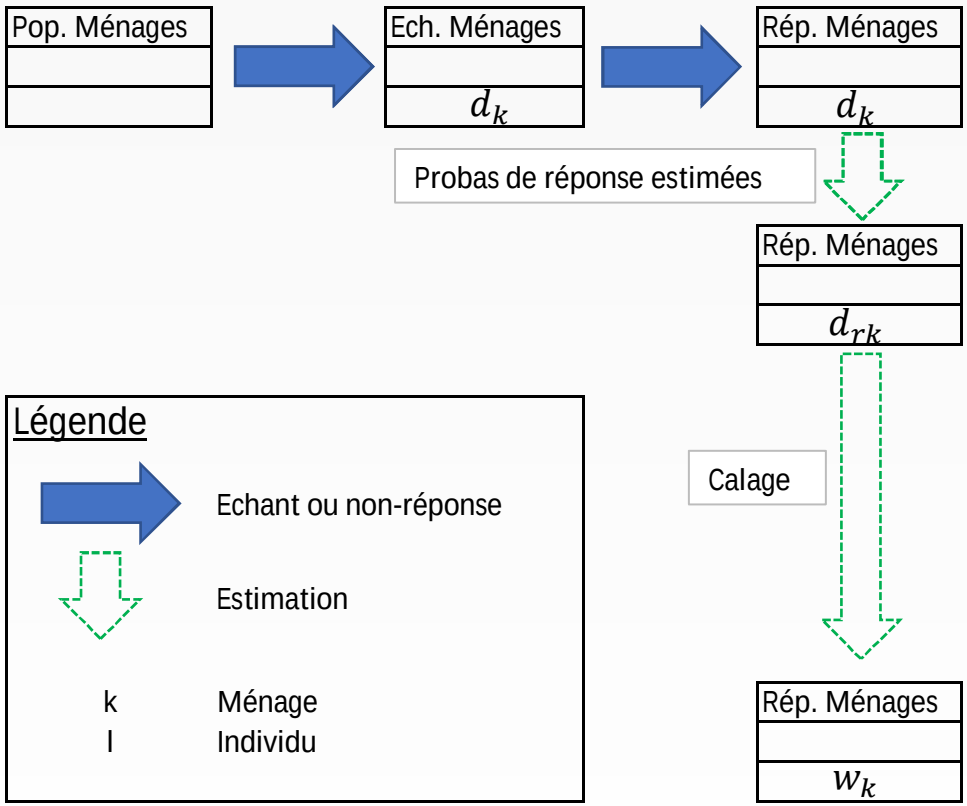
Total connu $t_x = \left(t_{x1}, \dots, t_{xp}\right)^T$

Calcul des poids calés $w_k = d_{rk} F\left(x_k^T \lambda\right)$

avec $\sum_{k \in S_{r,hh}} d_{rk} F\left(x_k^T \lambda\right) x_k = t_x$

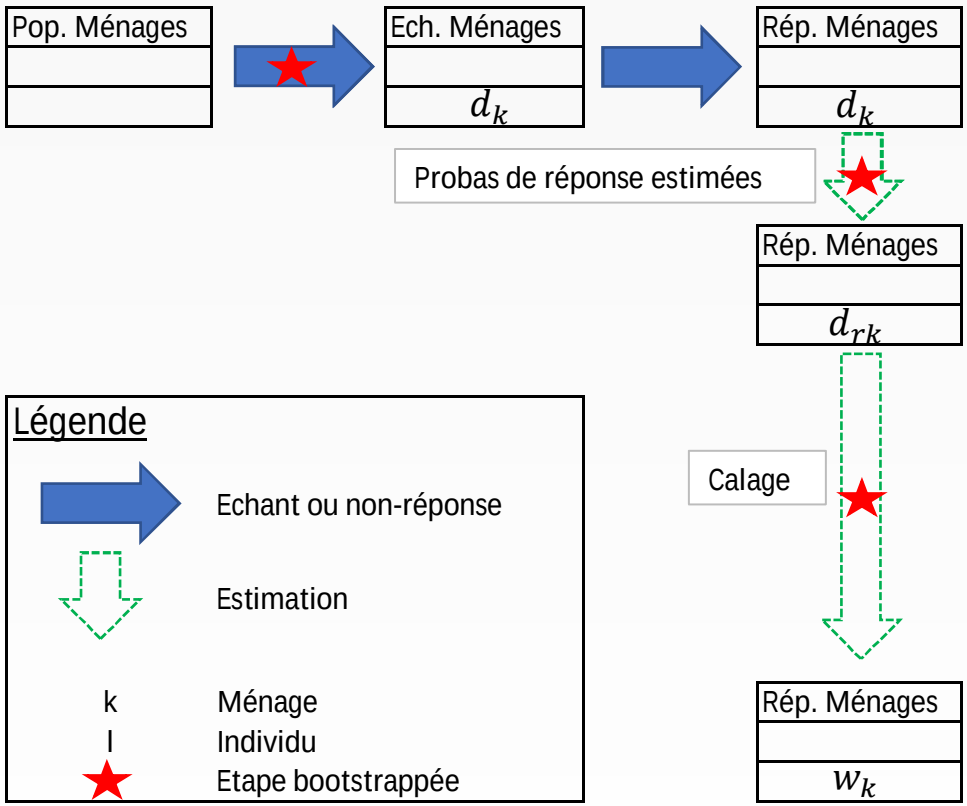
Echantillonnage de ménages

Avec non-réponse et calage



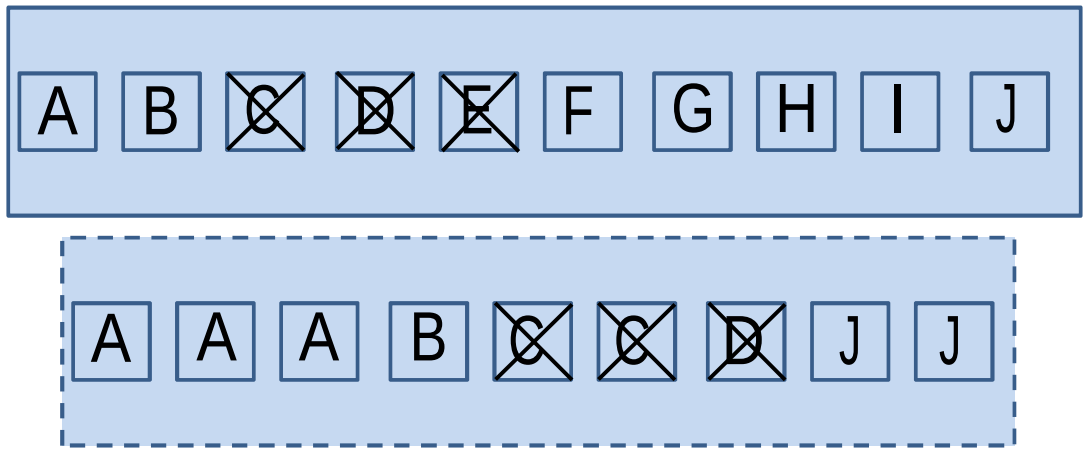
Echantillonnage de ménages : bootstrap

Avec non-réponse et calage



Echantillon de ménages : bootstrap

Poids de sondage bootstrappés

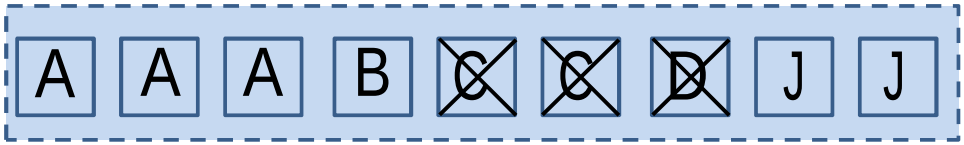


$$d_k = 10$$

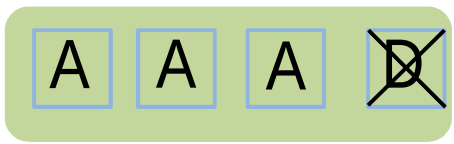
$$d_{A*} = \frac{300}{9} \qquad d_{B*} = d_{D*} = \frac{100}{9} \qquad d_{C*} = d_{J*} = \frac{200}{9}$$

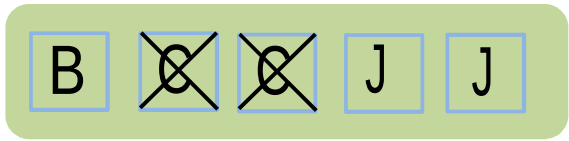
Echantillon de ménages : bootstrap

Poids corrigés de la NR bootstrappés



$$d_{A*} = \frac{300}{9} \qquad d_{B*} = d_{D*} = \frac{100}{9} \qquad d_{C*} = d_{J*} = \frac{200}{9}$$





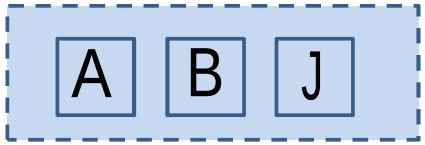
RHG c = 1

$$\hat{p}_{hh}^{1*} = \frac{3}{4}$$

RHG c = 2

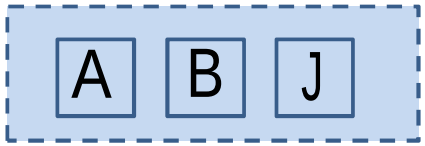
$$\hat{p}_{hh}^{2*} = \frac{3}{5}$$

$$d_{rA*} = \frac{400}{9} \qquad d_{rB*} = \frac{500}{27} \qquad d_{rJ*} = \frac{1000}{27}$$



Echantillon de ménages : bootstrap

Poids calés bootstrappés



$$d_{rA*} = \frac{400}{9} \qquad d_{rB*} = \frac{500}{27} \qquad d_{rJ*} = \frac{1000}{27}$$

Vecteur de variables auxiliaires $x_k = \left(x_{1k}, \dots, x_{pk}\right)^T$

Total connu $t_x = \left(t_{x1}, \dots, t_{xp}\right)^T$

Calcul des poids calés $w_{k*} = d_{rk*} F\left(x_k^T \lambda_*\right)$

avec $\sum_{k \in S_{r,hh*}} d_{rk*} F\left(x_k^T \lambda_*\right) x_k = t_x$

Enquête Histoire de Vie et Patrimoine

Enquête Histoire de Vie et Patrimoine

L'enquête a pour objectif de décrire les actifs financiers, immobiliers et professionnels des ménages français. Elle s'inscrit depuis 2010 dans un cadre européen : partie française du Household Finance and Consumption Survey (HFCS).

L'enquête a lieu tous les trois ans, avec une réinterrogation sur plusieurs vagues d'une partie des ménages (panel rotatif). Le dispositif d'enquête a été panélisé entre 2014 et 2017 : 1ère réinterrogation en 2017-18, 4ème et dernière réinterrogation en 2023 des premiers individus panélisés de 2014.

Le réseau HFC demande que l'estimation de variance puisse être faite par bootstrap. Ce travail (en cours) vise à expliquer comment le rescaled bootstrap (Rao and Wu, 1988) peut être utilisé pour des estimations transversales de l'enquête HVP. La méthode fait l'objet d'une double programmation SAS/R.

Le plan de sondage en deux mots

Pour une estimation transversale en $t+3$ (disons en 2023), l'échantillon HVP est obtenu à partir de 4 sous-échantillons sélectionnées lors de 4 vagues différentes $t, \ldots, t+3$. Chaque sous-échantillon est interrogé 4 fois.

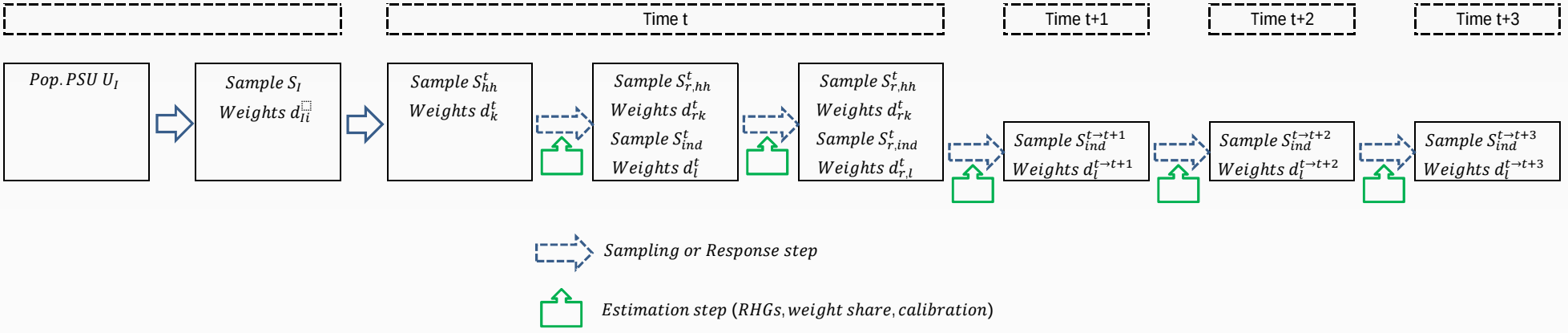
Le sous-échant. au temps t est sélectionné et enquêté ainsi :

- Un échantillon de logements est sélectionné dans l'échantillon-maître OCTOPUSSE (tirage à 2 degrés).
- Tous les ménages et les individus de ces logements sont sélectionnés au temps t.
- Les individus sont suivis et ré-enquêtés 3 fois. Aux temps $t+1$, $t+2$, $t+3$, leur ménage est enquêté. Leurs cohabitants sont également enquêtés, mais pas suivis dans le temps.

On procède de la même façon pour les sous-échantillons sélectionnés aux temps $t+1$, $t+2$ et $t+3$. Lors de l'estimation transversale en $t+3$, les ménages associés aux 4 sous-échantillons d'individus sont réunis en utilisant la méthode de partage des poids (Deville and Lavallée, 2006).

Sous-échantillon sélectionné au temps t

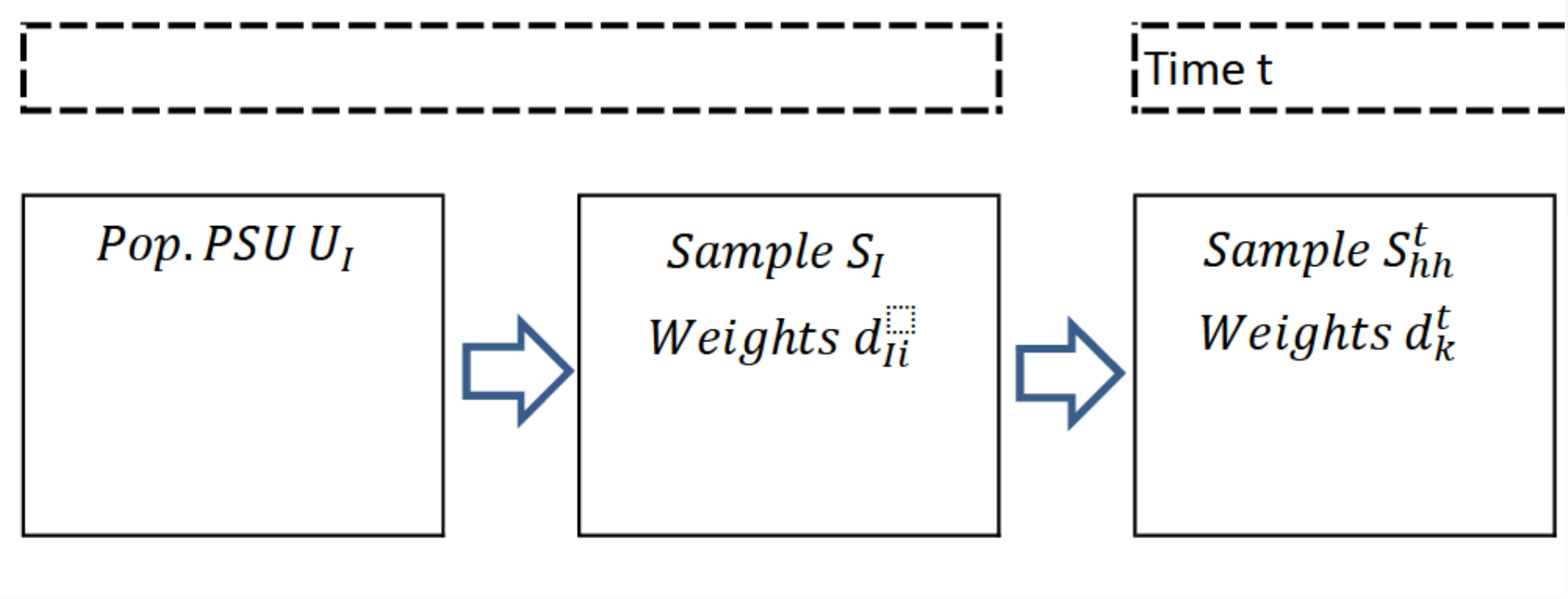
Chaîne de traitements entre t et $t+3$



Flèche bleue horizontale : une étape d'échantillonnage (traits pleins) ou de non-réponse (pointillés).
Flèche verte verticale : une étape d'estimation.
Ce sera important pour modéliser le bootstrap.

Sous-échantillon sélectionné au temps t

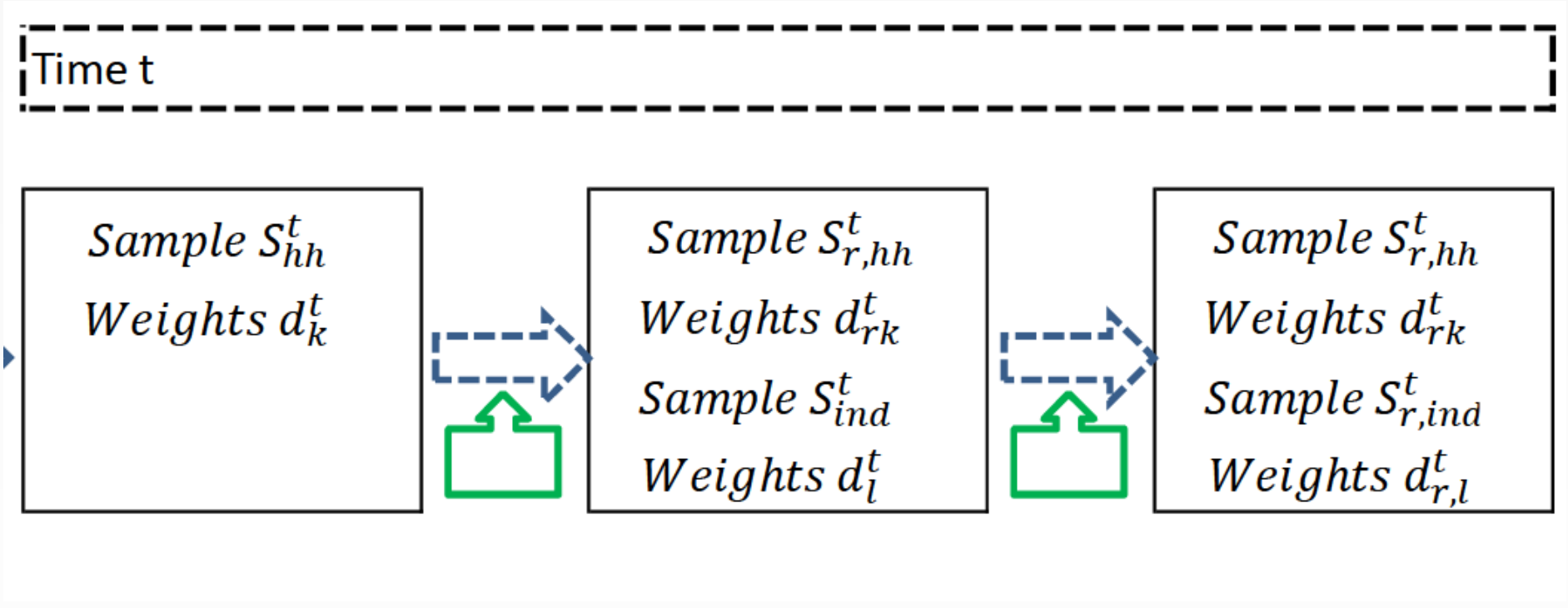
Echantillonnage au temps t



Tirage de l'échantillon d'unités primaires S_I.
Tirage du sous-échantillon de ménages S_{hh}^t (hh=household).

Sous-échantillon sélectionné au temps t

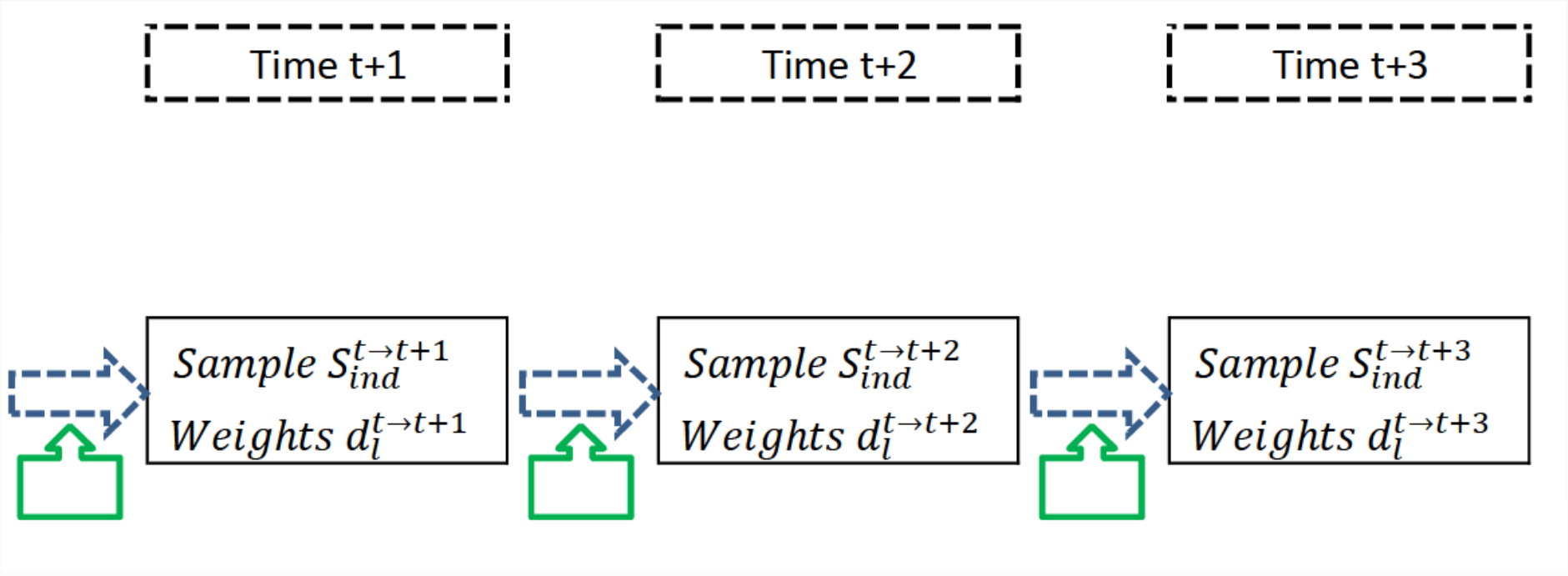
Correction de la non-réponse au temps t



Correction de la non-réponse dans l'échant. de ménages répondants $S^t_{r,hh}$.
Correction de la non-réponse dans l'échant. d'individus répondants $S^t_{r,ind}$.

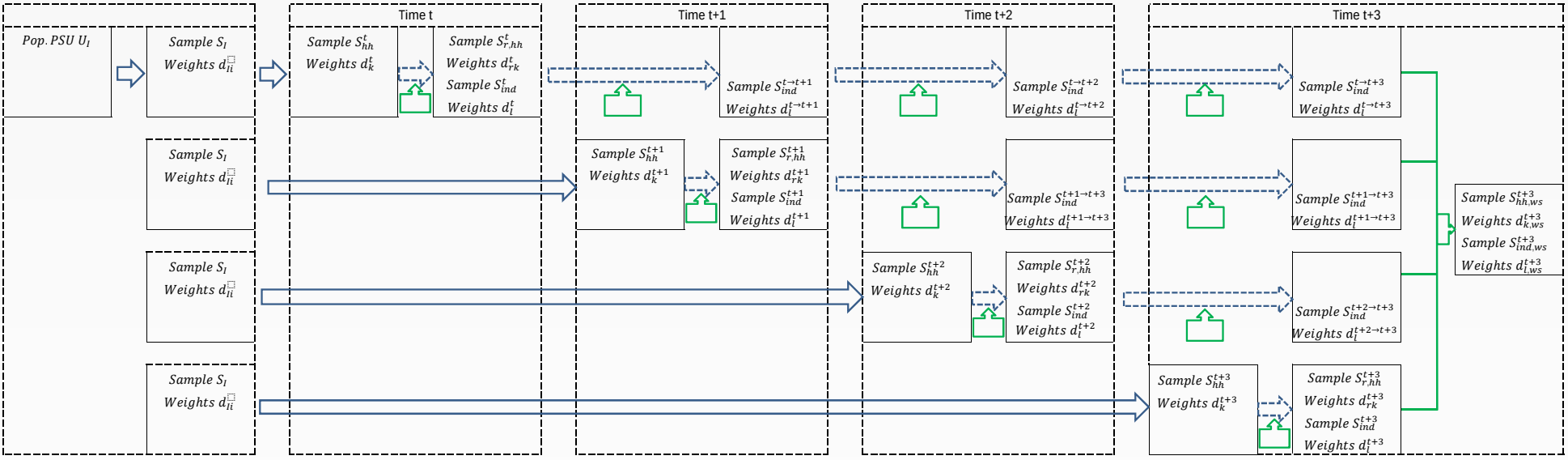
Sous-échantillon sélectionné au temps t

Suivi aux temps $t+1$ à $t+3$



Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+1}$.
Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+2}$.
Correction de l'attrition dans l'échant. d'individus répondants $S_{ind}^{t\to t+3}$.

Processus global pour une estimation transversale en $t+3$



Au temps $t+3$, on réunit quatre sous-échantillons :

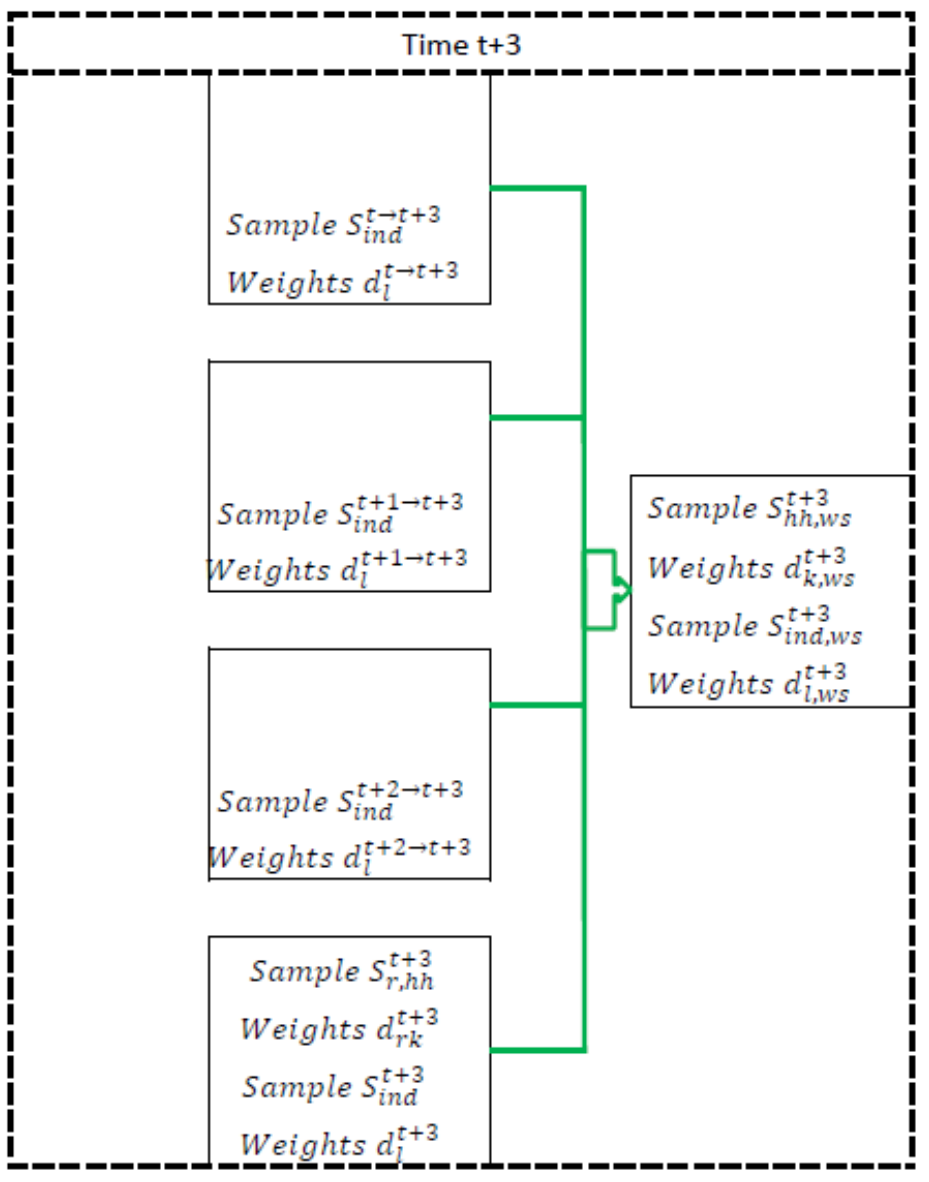
$S_{ind}^{t\to t+3}$ suivi depuis le temps t,

$S_{ind}^{t+1\to t+3}$ suivi depuis le temps $t+1$,

$S_{ind}^{t+2\to t+3}$ suivi depuis le temps $t+2$,

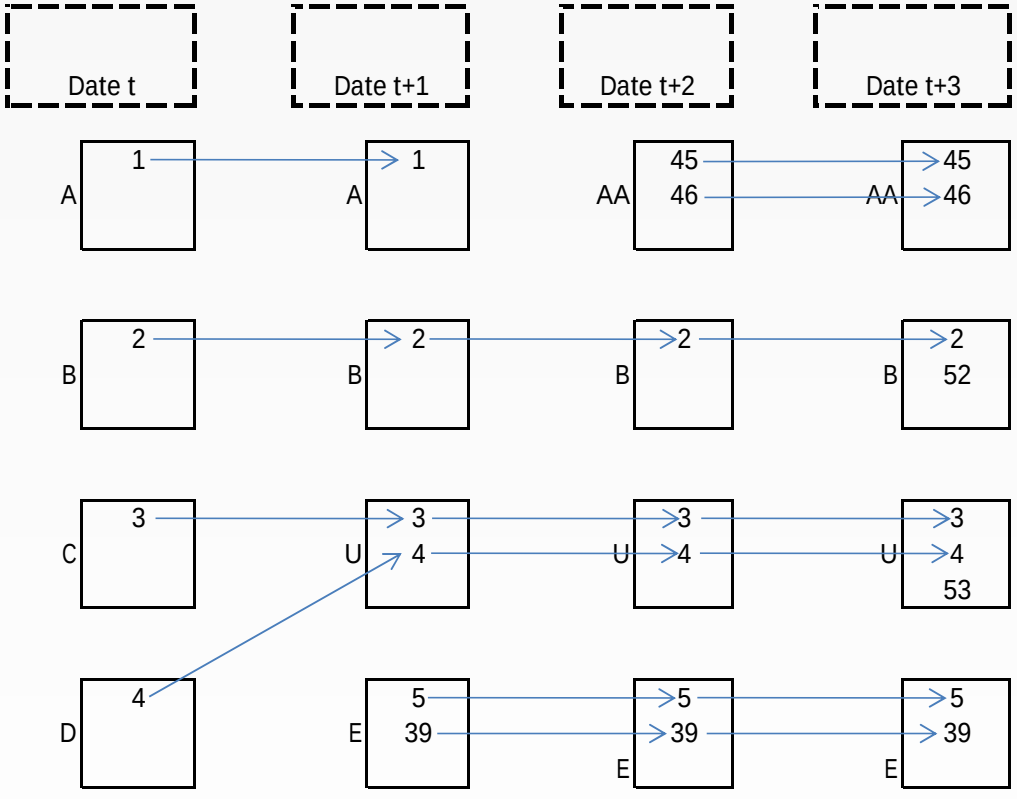
S_{ind}^{t+3} sélectionné au temps $t+3$.

Partage des poids pour une estimation transversale en $t+3$



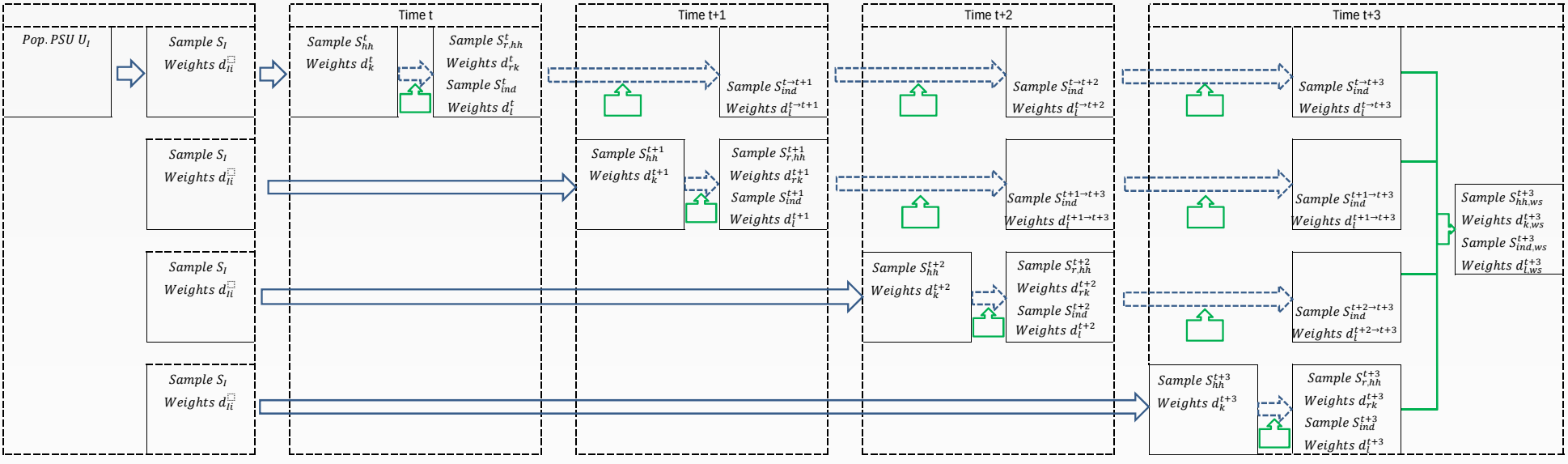
Partage des poids pour une estimation transversale en $t + 3$

Un petit exemple

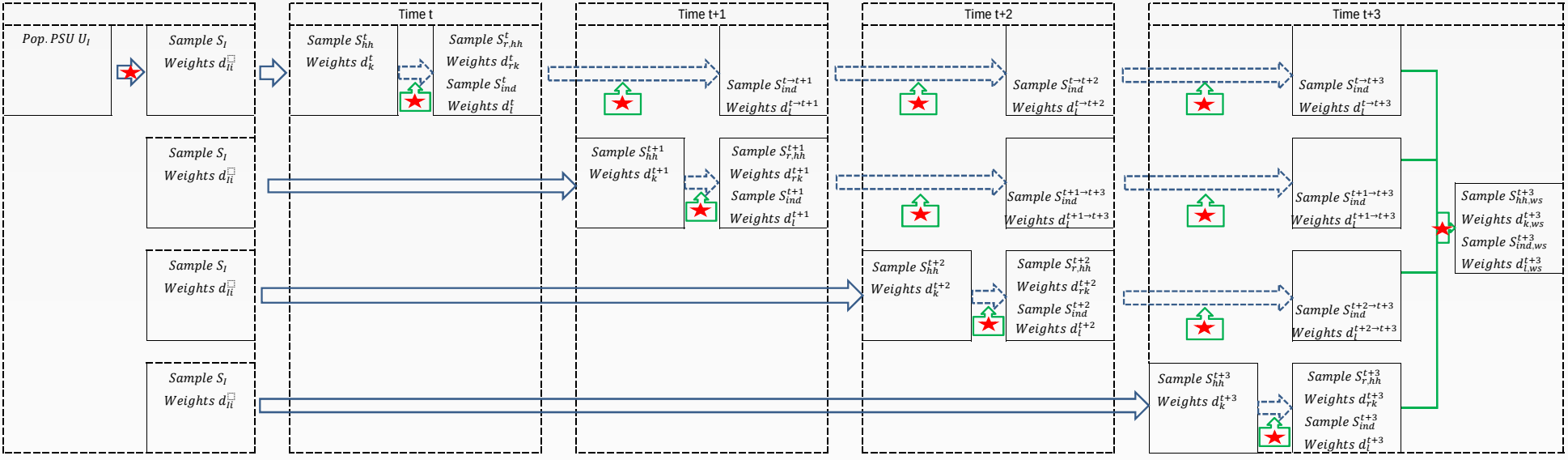


Household	Number of links between the household and the stacked pop. of individuals
	2
	2
AA	4
	4
	1
B	5
	4
	4
	1
U	9
	3
	3
E	6

Processus global pour une estimation transversale en $t+3$



Bootstrap pour une estimation transversale en $t+3$



A chaque itération :

1. On bootstrappe le tirage des unités primaires, mais pas les autres étapes d'échantillonnage/non-réponse.
2. On bootstrappe toute l'estimation $\Rightarrow$ estim. transversal bootstrappé.
3. On répète 1 et 2 un grand nb de fois $\Rightarrow$ estimation de variance.

Etude par simulations

Etude par simulations

Population U_I de $N_I = 2,000$ UP contenant (en moyenne) 40 ménages. Au temps t, pop. U_{hh}^t de 80 520 ménages contenant (en moyenne) 2 individus; pop. U_{ind}^t de 161 027 individus.

Les populations évoluent de la façon suivante. Aux temps subséquents :

- 4 % des individus sortent du champ de l'enquête,
- 5 % d'individus entrent dans le champ dans les ménages existants,
- 1 % de nouveaux ménages sont créés.

Dans la population des ménages au temps $t + 3$, deux variables auxiliaires z_1 et z_2 sont générées selon des distributions Gamma. Pour chaque UP i et ménage j, trois variables y_{ij}^m, $m = 1, \ldots, 3$ sont générées selon le modèle

$$y_{ij}^m = 5 + 2z_{1,ij} + 2z_{2,ij} + \epsilon_i + \sigma_m \epsilon_{ij}. \quad (2)$$

Etude par simulations

Plan de sondage

Un échantillon S_l de $n_l \in \{20, 50, 100, 200\}$ UP est tiré selon la méthode de Rao-Sampford. Au temps t:

- Un sous-échantillon de $n_0 = 10$ ménages est sélectionné, et tous les individus du ménage sont sélectionnés et suivis dans le temps.
- Non-réponse d'environ 20% à chaque temps (groupes de réponse homogènes)

Même principe pour les sous-échantillons sélectionnés aux temps $t+1$, $t+2$, $t+3$. Les quatre sous-échantillons sont réunis au temps $t+3$, et la méthode de partage des poids est utilisée pour produire une estimation transversale sur U_{hh}^{t+3}.

L'échantillonnage est répété $D_1 = 2,000$ fois pour obtenir une approximation Monte Carlo de la variance. Nous calculons également $D_2 = 100$ fois un estimateur de variance bootstrap avec $B = 50$ itérations bootstrap (faible, travail préliminaire).

Résultats : biais relatif de l'estimateur bootstrap de variance

Table: Biais relatif (en pourcentage) de l'estim. de variance bootstrap

	Taille de l'échantillon d'UP n_I			
Var. d'intérêt	20	50	100	200
y_1	-1.6%	-1.2%	$+6.0\%$	$+5.0\%$
y_2	$+1.9\%$	-3.6%	$+4.6\%$	$+2.8\%$
y_3	$+0.9\%$	$+4.5\%$	$+6.7\%$	$+16.3\%$