Techniques avancées d'échantillonnage

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

Master de Statistique Publique
25/09/2024

Principaux objectifs du cours

- Rappels sur l'échantillonnage et l'estimation en population finie
- Compléments sur les méthodes d'échantillonnage à probabilités inégales
- Méthodes d'échantillonnage équilibré et applications
- Méthodes d'échantillonnage spatial

Nous utiliserons :

- le package R `sampling` pour l'échantillonnage,
- le package R `gustave` pour l'estimation de variance (créé et maintenu par Martin Chevalier et Khaled Larbi, Insee).

```
#Appel des packages
> library(sampling)
> help(package="sampling")
> library(gustave)
> help(package="gustave")
```

Bases de sondage

Nous utiliserons deux bases de sondage disponibles avec le package `sampling`.

La base de sondage `belgianmunicipalities` fournit des informations sur les 589 communes de Belgique au 01/07/2004, ainsi que des informations financières datées de 2001.
La base de sondage `MU284` fournit des informations sur les 284 communes de Suède datées de 1985.

```
#Récupération de deux bases de données du package
> data("belgianmunicipalities")
> attach(belgianmunicipalities)
> data("MU284")
> attach(MU284)
```

Variables de "belgianmunicipalities"

Commune	Municipality name
INS	INS Code INS
Province	Province number
Arrondiss	Administrative division number
Men04	Number of men on July 1, 2004
Women04	Number of women on July 1, 2004
Tot04	Total population on July 1, 2004
Men03	Number of men on July 1, 2003
Women03	Number of women on July 1, 2003
Tot03	Total population on July 1, 2003
Diffmen	Men04 minus Men03
Diffwom	Women04 minus Women03
DiffTOT	Tot04 minus Tot03
TaxableIncome	Total taxable income in euros in 2001
Totaltaxation	Total taxation in euros in 2001
Averageincome	Average of the income-tax return in euros in 2001
Medianincome	median of the income-tax return in euros in 2001.

Variables de "MU284"

LABEL	Identifier number from 1 to 284
P85	1985 population (in thousands)
P75	1975 population (in thousands)
RMT85	Revenues from 1985 municipal taxation (in millions of kronor)
CS82	Number of Conservative seats in municipal council
SS82	Number of Social-Democratic seats in municipal council
S82	Total number of seats in municipal council
ME84	Number of municipal employees in 1984
REV84	Real estate values according to 1984 assessment (in millions of kronor)
REG	Geographic region indicator
CL	Cluster indicator (a cluster consists of a set of neighboring)

Variables de "commune" (hors package)

IDENT	Variable identifiant l'adresse
NLOG	Nombre de logements de l'adresse
ACTIFS	Nombre d'actifs
INACTIFS	Nombre d'inactifs
NATFN	Nombre de français de naissance
NATHE	Nombre d'étrangers hors Union Européenne
NATUE	Nombre d'étrangers de l'Union Européenne
NATFA	Nombre de français par acquisition
HOMMES	Nombre d'hommes
FEMMES	Nombre de femmes
_0019	Nombre de personnes de moins de 20 ans
_2039	Nombre de personnes de 20 à 39 ans
_4059	Nombre de personnes de 40 à 59 ans
_6074	Nombre de personnes de 60 à 74 ans
_7599	Nombre de personnes de 75 ans et plus
H0019	Nombre d'hommes de moins de 20 ans
...	...
F7599	Nombre de femmes de plus de 75 ans

1. Rappel sur les méthodes d'échantillonnage
   - Principes généraux
   - Etude par simulations
   - Calcul de variance
   - Modèle de travail
2. Méthodes d'échantillonnage à probabilités inégales
   - Tirage systématique
   - Méthode du pivot
   - Tirage de Poisson
   - Tirage réjectif
3. Echantillonnage équilibré
   - Principe
   - La méthode du Cube
   - Le Recensement
4. Echantillonnage spatial
   - Echantillonnage spatial en population finie
   - Sampling in a continuous population

Rappel sur les méthodes d'échantillonnage

Principes généraux

Notations

Nous nous plaçons dans le cadre d'une population finie U d'*unités statistiques* supposées identifiables par un label. Nous noterons

$$U = \{1, \ldots, k, \ldots, N\}$$

où N désigne la taille de la population U, qui n'est pas forcément connue.

Nous nous intéressons à une *variable d'intérêt* y prenant la valeur y_k sur $k \in U$. Nous souhaitons disposer d'indicateurs pour la population U :

- total : $t_y = \sum_{k \in U} y_k$,
  Ex : Nombre total d'actifs dans la population française
- total sur un *domaine* U_d : $t_{yd} = \sum_{k \in U_d} y_k$,
  Ex : Nombre total d'actifs dans l'Aire Urbaine de Rennes
- ratio de deux totaux : $R = t_y / t_x$.
  Ex : Taux de chômage en Bretagne

Plan de sondage

La variable d'intérêt est mesurée sur un échantillon aléatoire S obtenu selon un *plan de sondage* p. Il s'agit d'une loi de probabilité sur les parties de U :

$$\forall s \subset U \quad p(s) \geq 0 \text{ et } \sum_{s \subset U} p(s) = 1. \tag{1}$$

En pratique, nous utilisons un algorithme de tirage pour sélectionner S, et le plan de sondage n'est pas complètement spécifié. Deux quantités sont importantes pour calculer des estimateurs et mesurer leur précision :

- les probabilités d'inclusion d'ordre 1

$$\pi_k \equiv Pr(k \in S),$$

  sont utilisées pour le calcul des estimateurs ponctuels,
- les probabilités d'inclusion d'ordre 2

$$\pi_{kl} \equiv Pr(k, l \in S),$$

  sont utilisées pour le calcul des estimateurs de variance.

Nous noterons $n(S)$ la taille de l'échantillon S, qui peut être aléatoire.

Exemples

Exemple 1 : Les enquêtes-ménages de l'Insee visent à décrire les conditions de vie des ménages (emploi, logement, patrimoine, ...). Les ménages enquêtés sont sélectionnés dans un échantillon de zones appelé l'*Echantillon-Maître*.

Exemple 2 : Les enquêtes-entreprises sont réalisées à l'aide d'une base de sondage (répertoire SIRUS) et de sources externes.

Exemple 3 : Les inventaires forestiers nationaux sont réalisés en sélectionnant un échantillon de points sur le territoire. Des motifs sont ensuite construits autour de ces points (e.g., des placettes circulaires) pour sélectionner les arbres qui dont l'objet de mesures sur le terrain.

Mesures de précision

La qualité d'un estimateur $\hat{\theta}$ est évaluée par :

- son biais

$$B_p(\hat{\theta}) = E_p(\hat{\theta} - \theta) = \sum_{s \subset U} p(s)\{\hat{\theta}(s) - \theta\},$$

- sa variance

$$V_p(\hat{\theta}) = E_p\left\{\hat{\theta} - E_p(\hat{\theta})\right\}^2,$$

- ou encore son Erreur Quadratique Moyenne (EQM)

$$EQM_p(\hat{\theta}) = E_p(\hat{\theta} - \theta)^2 = B_p(\hat{\theta})^2 + V_p(\hat{\theta}).$$

Etude par simulations

Etude par simulations

Il est possible de vérifier les propriétés théoriques d'un estimateur (biais, variance, EQM) en utilisant une base de sondage sur laquelle les variables d'intérêt sont connues sur toute la population.

La première possibilité consiste à lister tous les échantillons s sélectionnables, avec leur probabilité de sélection. Il est alors possible de calculer :

$$\text{Le biais} \qquad B_p(\hat{\theta}) = \sum_{s \subset U} p(s) \left\{ \hat{\theta}(s) - \theta \right\},$$

$$\text{La variance} \qquad V_p(\hat{\theta}) = \sum_{s \subset U} p(s) \left\{ \hat{\theta}(s) - \sum_{s' \subset U} p(s') \hat{\theta}(s') \right\}^2,$$

$$\text{L'EQM} \qquad EQM_p(\hat{\theta}) = \sum_{s \subset U} p(s) \left\{ \hat{\theta}(s) - \theta \right\}^2.$$

Cette méthode n'est possible que sur de petites populations pour laquelle il n'est pas trop coûteux de lister l'ensemble des échantillons.

Simulations de Monte-Carlo

Une autre possibilité consiste à répéter un grand nombre de fois B, indépendamment, le tirage d'échantillons S_b selon le plan de sondage $p(\cdot)$, pour obtenir des répliques i.i.d de $\hat{\theta}$ notées $\hat{\theta}_b,\ b = 1, \ldots, B$.

Rappelons que pour un échantillon $(X_1, \ldots, X_n)$ i.i.d., nous avons

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \to_{Pr} E(X) \quad \text{et} \quad s_X^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \to_{Pr} V(X).$$

Nous avons donc pour la simulation de Monte Carlo :

$$\bar{\hat{\theta}}_B \equiv \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b \to_{Pr} E_p(\hat{\theta}) \quad \text{et} \quad s_{\hat{\theta}}^2 \equiv \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\hat{\theta}}_B\right)^2 \to_{Pr} V_p(\hat{\theta}).$$

Si le nombre B de simulations est grand, nous pouvons obtenir une bonne approximation par simulations de $E_p(\hat{\theta})$ et $V_p(\hat{\theta})$.

Estimateur de Horvitz-Thompson

Probabilités d'inclusion d'ordre 1

La probabilité pour l'unité k d'être sélectionnée dans l'échantillon est notée

$$\pi_k = Pr(k \in S)$$

Les valeurs de ces probabilités sont fixées avant le tirage.

En l'absence d'information auxiliaire, les unités sont tirées à probas égales

$$\pi_k = \frac{n}{N}.$$

Exemple : tirage de Bernoulli, sondage aléatoire simple.

Si une variable auxiliaire x_k est connue pour tout $k \in U$, nous pouvons utiliser des probabilités d'inclusion proportionnelles à la taille

$$\pi_k = n\frac{x_k}{\sum_{l\in U} x_l}. \quad (2)$$

Exemple : tirage systématique, méthode du pivot, tirage de Poisson, tirage réjectif.

Recalcul des probabilités d'inclusion

Si certaines unités sont particulièrement grosses (au sens de x), certaines probabilités d'inclusion peuvent être supérieures à 1. Dans ce cas les unités correspondantes sont sélectionnées d'office, et les probabilités d'inclusion des autres unités sont recalculées.

```
#Calcul de probas d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
> summary(pi_50)
[1]  Min.  1st Qu. Median   Mean 3rd Qu.   Max.
[1]0.05693 0.07675 0.08375 0.08489 0.09113 0.14076

> n=400
> pi_400=inclusionprobabilities(averageincome,n)
> summary(pi_400)
[1]  Min.  1st Qu. Median   Mean 3rd Qu.   Max.
[1]0.4556 0.6142  0.6702 0.6791 0.7293 1.0000
```

L'estimateur de Horvitz-Thompson

La connaissance des probabilités π_k permet une estimation sans biais d'un total sous le plan de sondage. Le total t_y est estimé sans biais par l'estimateur de Horvitz-Thompson (HT)

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \tag{3}$$

si tous les π_k sont > 0, en notant $I_k = 1(k \in S)$ l'indicatrice d'appartenance à l'échantillon.

C'est un estimateur pondéré, où les *poids de sondage* $d_k = 1/\pi_k$ ne dépendent pas de la variable d'intérêt.

Si certaines probabilités d'inclusion sont nulles, nous sommes en présence d'un biais de couverture.

L'estimateur de Horvitz-Thompson (2)

La connaissance des probabilités π_k permet une estimation sans biais d'un total sous le plan de sondage. Le total t_y est estimé sans biais par l'estimateur de Horvitz-Thompson (HT)

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \tag{4}$$

si tous les π_k sont > 0, en notant $I_k = 1(k \in S)$ l'indicatrice d'appartenance à l'échantillon.

```
#Tirage d'un échantillon selon un plan réjectif
>ech=UPmaxentropy(pi_50)
#Estimation de HT du total de TaxableIncome
>y=TaxableIncome
>est_ht=HTestimator(y[ech==1],pi_50[ech==1])
>est_ht
[1,] 1.092e+11
```

Calcul de variance

Probabilités d'inclusion d'ordre 2

La probabilité pour deux unités distinctes k et l d'être sélectionnées conjointement dans l'échantillon est notée

$$\pi_{kl} = Pr(k, l \in S).$$

Ces probabilités π_{kl} ne sont pas choisies avant le tirage : elles dépendent des probabilités d'inclusion π_k, et du plan de sondage utilisé. Elles interviennent dans le calcul des estimateurs de variance.

Ces probabilités sont souvent difficiles à calculer exactement, sauf pour certains plans de sondage particuliers. Même si elles sont calculables, on préfère souvent utiliser des estimateurs de variance simplifiés n'utilisant que les probabilités d'inclusion d'ordre 1.

Probabilités d'inclusion d'ordre 2 (2)

Le package `sampling` permet de calculer la matrice des probabités d'inclusion d'ordre deux pour 5 plans de sondage à probabilités inégales:

- le tirage réjectif ou tirage de Poisson conditionnel,
- la méthode de Midzuno,
- le tirage de Rao-Sampford,
- le tirage systématique,
- la méthode de Tillé.

```
#Calcul de probas d'inclusion d'ordre 2 pour le réjectif
pikl_rej_50=UPmaxentropypi2(pi_50)
#Calcul de probas d'inclusion d'ordre 2 pour Rao-Sampford
pikl_sam_50=UPsampfordpi2(pi_50)
#Calcul de probas d'inclusion d'ordre 2 pour le systématique
pikl_sys_50=UPsystematicpi2(pi_50)
```

Estimateur de variance de Horvitz-Thompson

Pour un plan de sondage quelconque, la variance de l'estimateur de HT est donnée par

$$V_p(\hat{t}_{y\pi}) = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl} \quad \text{avec} \quad \Delta_{kl} = \pi_{kl} - \pi_k \pi_l. \tag{5}$$

Cette variance peut être estimée sans biais par

$$v_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}} \tag{6}$$

si tous les π_{kl} sont strictement positifs.

Principe : un couple (k, l) d'individus de l'échantillon représente $1/\pi_{kl}$ couples de la population.

Estimateur de variance de Horvitz-Thompson (2)

```
#Tirage d'un échantillon selon un plan réjectif
>ech=UPmaxentropy(pi_50)
#Estimation de HT du total de TaxableIncome
>y=TaxableIncome
>est_ht=HTestimator(y[ech==1],pi_50[ech==1])
#Estimation de variance de HT (PACKAGE SAMPLING)
>vest_ht=varHT(y[ech==1],pikl_rej_50[ech==1,ech==1],1)
>options("scipen"=-100,digits="4")
>est_ht
[1,] 1.092e+11
>vest_ht
[1] 2.518e+20
```

Estimateur de variance de Yates-Grundy

Pour un plan de sondage *de taille fixe*, la variance peut se réécrire

$$V_p(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}. \quad (7)$$

Cette variance peut être estimée sans biais par

$$v_{YG}(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k\neq l\in S}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}} \quad (8)$$

si tous les π_{kl} sont strictement positifs. Il s'agit de l'estimateur de variance de Yates-Grundy.

Si le plan de sondage vérifie les *conditions de Yates-Grundy* :
$\forall k \neq l \in U \quad \Delta_{kl} \leq 0$, cet estimateur de variance est toujours à valeurs positives.

Estimateur de variance de Yates-Grundy

```
#Tirage d'un échantillon selon un plan réjectif
>ech=UPmaxentropy(pi_50)
#Estimation de HT du total de TaxableIncome
>y=TaxableIncome
>est_ht=HTestimator(y[ech==1],pi_50[ech==1])
#Estimation de variance de YG (PACKAGE SAMPLING)
>vest_yg=varHT(y[ech==1],pikl_rej_50[ech==1,ech==1],2)
>vest_yg
[1] 2.804e+20
#Estimation de variance de YG (PACKAGE GUSTAVE)
>vest_yg_gus=varSYG(y[ech==1],pikl_rej_50[ech==1,ech==1])
>vest_yg_gus
[1] 2.804e+20
```

Tous les algorithmes de tirage à probabilités inégales dans `sampling` sont de taille fixe, sauf le tirage de Poisson (fonction `UPpoisson`).

Intervalle de confiance

En l'absence de biais de couverture, l'estimateur de HT $\hat{t}_{y\pi}$ estime sans biais t_y. Un intervalle de confiance pour t_y de niveau $1-\alpha$ est donné par :

$$IC_{1-\alpha}(t_y) = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}}\sqrt{v(\hat{t}_{y\pi})}\right]$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite $\mathcal{N}(0,1)$.

L'intervalle de confiance est asymptotiquement valide :

- si l'estimateur $\hat{t}_{y\pi}$ centré-réduit est asympt. normalement distribué :

$$\frac{\hat{t}_{y\pi}-t_y}{\sqrt{V_p(\hat{t}_{y\pi})}} \longrightarrow_{\mathcal{L}} \mathcal{N}(0,1).$$

- si l'estimateur de variance $v(\hat{t}_{y\pi})$ est consistant :

$$\frac{v(\hat{t}_{y\pi})}{V_p(\hat{t}_{y\pi})} \longrightarrow_{Pr} 1.$$

Modèle de travail

Principe

Au stade de l'échantillonnage, nous supposons disponible un q-vecteur $\mathbf{x}_k$ de variables auxiliaires connues pour chaque unité $k \in U$.

Cette information va nous servir à construire un plan de sondage. Ce qui nous motive est une relation supposée entre la variable d'intérêt y_k et les variables auxiliaires $\mathbf{x}_k$, que nous appelons le *modèle de travail* :

$$y_k = \mathbf{x}_k^\top \beta + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma_k^2. \end{array} \right.$$

Quel que soit le plan de sondage choisi, l'estimateur de HT sera sans biais en l'absence de biais de couverture. Si l'information auxiliaire $\mathbf{x}_k$ peut être utilisée pour définir le plan de sondage, alors la variance sera réduite si ce modèle reflète (au moins partiellement) la relation entre y_k et $\mathbf{x}_k$.

Calcul de variance pour un plan de taille fixe

En écrivant la variable d'intérêt selon le modèle de travail

$$y_k = \beta\pi_k + \epsilon_k,$$

nous avons pour un plan de taille fixe

$$V_p(\hat{t}_{y\pi}) = V_p(\hat{t}_{\epsilon\pi}).$$

La variance sera donc faible si les résidus ϵ_k sont petits, i.e. si la variable y_k est approximativement proportionnelle à π_k.

Un bon choix consiste à utiliser des probabilités d'inclusion proportionnelles à une mesure de taille x_k (nombre d'employés d'une entreprise, nombre de résidences principales d'une commune). Nous obtenons la formule des probabilités d'inclusion proportionnelles à la taille :

$$\pi_k = n\frac{x_k}{\sum_{l\in U} x_l}.$$

En résumé

On utilise un plan de sondage $p(\cdot)$ respectant des probabilités d'inclusion d'ordre 1 choisies, ce qui permet de calculer pour le total t_y son estimateur de Horvitz-Thompson

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

La variance sera faible si y_k et π_k sont approximativement proportionnels. Pour un plan de taille fixe, cette variance est estimée par

$$v_{YG}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k \neq l \in S} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\Delta_{kl}}{\pi_{kl}}$$

En utilisant une approximation normale pour $\hat{t}_{y\pi}$, on obtient l'intervalle de confiance

$$\left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{v\left[\hat{t}_{y\pi}\right]} \right] \quad \text{où} \quad v(\hat{t}_{y\pi}) \equiv \begin{cases} v_{HT}(\hat{t}_{y\pi}) & \text{pds quelconque,} \\ v_{YG}(\hat{t}_{y\pi}) & \text{pds de taille fixe.} \end{cases}$$

Cas du sondage aléatoire simple

Sondage aléatoire simple (SRS)

Il s'agit du plan qui donne la même probabilité à tous les échantillons de taille n d'être sélectionnés. L'estimateur de Horvitz-Thompson du total est donné par

$$\hat{t}_{y\pi} = N\ \bar{y} \quad \text{avec} \quad \bar{y} = \frac{1}{n}\sum_{k\in S} y_k. \tag{9}$$

Sa variance s'obtient à partir de la formule de Sen-Yates-Grundy :

$$V_p[\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} S_y^2 \quad \text{avec} \quad S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \mu_y)^2. \tag{10}$$

On l'estime sans biais par

$$v_{YG}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 \quad \text{avec} \quad s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y})^2. \tag{11}$$

Le package `sampling` contient 2 algorithmes permettant de réaliser un SRS.

Algorithme de sélection pour un SRS: méthode draw by draw

La première méthode consiste à utiliser un algorithme de sélection unité par unité.

```
> n=100
> Npop=589
> ech_srs=srswor(n,Npop)
```

Algorithme 1 Méthode de sélection draw by draw

1. Pour $k = 1, \ldots, n$, sélectionner une unité dans U à probabilités égales parmi les unités qui n'ont pas déjà été tirées.

Algorithme de sélection pour un SRS: sélection-rejet

La population est parcourue séquentiellement, en tirant chaque unité avec la probabilité conditionnelle au nombre d'unités déjà tirées (Fan, Muller et Rezucha, 1962).

```
> n=100
> Npop=589
> ech_srs=srswor1(n,Npop)
```

Algorithme 2 Méthode de sélection-rejet

1. On initialise $j = 0$.
2. Pour $k = 1, \ldots, N$, faire :
   - Avec une probabilité

$$\begin{aligned} \text{ProbCond} &= \frac{n-j}{N-(k-1)} \\ &= \frac{\text{Nb d'unités restant à tirer}}{\text{Nb d'unités restantes}}, \end{aligned}$$

   on sélectionne l'unité k et $j = j + 1$.

Méthode de sélection-rejet : exemple

Individu	u_k	j	ProbCond	I_k
1	0.65	0	3/8=0.38	0
2	0.98	0	3/7=0.43	0
3	0.86	0	3/6=0.50	0
4	0.82	0	3/5=0.60	0
5	0.27	0	3/4=0.75	1
6	0.50	1	2/3=0.67	1
7	0.96	2	1/2=0.50	0
8	0.13	2	1/1=1.00	1

Estimation pour un SRS

```
#Tirage d'un échantillon aléatoire simple
>n <- 100
>Npop <- 589
>ech=srswor1(n,Npop)
#Estimation
>pi <- rep(n/Npop,Npop)
>y=TaxableIncome
>est_ht=HTestimator(y[ech==1],pi_50[ech==1])
>est_ht
[1,] 2.517e+11
#Estimation de variance pour un SRS (PACKAGE SAMPLING)
>vest_srs=varest(y[ech==1],,pi[ech==1],)
>vest_srs
[1] 1.16e+20
#Estimation de variance pour un SRS (PACKAGE GUSTAVE)
>vest_srs_gus=var_srs(y[ech==1],pi[ech==1])
>vest_srs_gus
[1] 1.16e+20
```

Exercice

Nous considérons la population `belgianmunicipalities` et les variables d'intérêt `Tot04` et `TaxableIncome`. Nous souhaitons estimer le total de ces deux variables en utilisant un sondage aléatoire simple de taille $n = 100$.

Mettre en place une étude par simulations pour vérifier que :

1. L'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ (équation 9) est sans biais pour le total t_y,
2. L'estimateur de variance $v_{YG}(\hat{t}_{y\pi})$ (équation 11) est sans biais pour la vraie variance $V_p(\hat{t}_{y\pi})$,
3. L'intervalle de confiance estimé (cf diapo 29) possède un taux de couverture de 95 %

Vous utiliserez au moins $B = 10,000$ simulations.

Exercice

Initialisation des paramètres

```
#Une étude par simulations pour le SRS
#Initialisation des paramètres
> n=100
> Npop=589
> pi=rep(n/Npop,589)
> sim=10000

#Pile des simulations format (Est,EstVar,Binf,Bsup)
> pile_Tot04=array(0,c(sim,4))
> pile_TaxableIncome=array(0,c(sim,4))

#Estimateur de HT
> ht=numeric(2)
#Estimateur de variance
> ev=numeric(2)
#Intervalle de confiance
> ic=numeric(4)
```

Exercice

Boucle de Monte-Carlo

```
#Etude par simulations : boucle de Monte-Carlo
> for(i in 1:sim)
 {
  cat("Simulation ",i,"\n")
  #Selection de l'échantillon
  ech=...
  #Estimation et estimation de variance
  ht[1]=...
  ...
  ev[2]=...
  #Intervalle de confiance
  ic[1]=...
  ...
  ic[4]=...
  #Empilement
  pile_Tot04[i,]=cbind(ht[1],ev[1],ic[1],ic[2])
  pile_TaxableIncome[i,]=cbind(ht[2],ev[2],ic[3],ic[4])
  }
```

Exercice

Comparaison totaux-espérance de Monte-Carlo de l'est.

```
#Comparaison vrais totaux et estimateurs de HT
>tot=c(sum(Tot04),sum(TaxableIncome))
>Emc_ht=c(mean(pile_Tot04[,1]),mean(pile_TaxableIncome[,1]))
>cat("Vrais totaux \n")
>tot
[1] 1.042e+07 1.211e+11
>cat("Espérance Monte Carlo \n")
>Emc_ht
[1] 1.042e+07 1.211e+11
```

Exercice

Comparaison variance -espérance de Monte-Carlo de l'estimateur de variance

```
#Comparaison Variance et estimateur de variance
>pikl_srs <- UPmaxentropypi2(pi)
>var_srs=numeric(2)
>var_srs[1] <- t(Tot04/pi)%*%(pikl_srs-pi%*%t(pi))
                            %*%(Tot04/pi)
>var_srs[2] <- t(TaxableIncome/pi)%*%(pikl_srs-pi%*%t(pi))
                                    %*%(TaxableIncome/pi)
>Emc_ev=c(mean(pile_Tot04[,2]),mean(pile_TaxableIncome[,2]))
>cat("Vraies variance \n")
>var_srs
[1] 2.247e+12 2.977e+20
>cat("Espérance Monte Carlo \n")
>Emc_ev
[1] 2.249e+12 2.975e+20
```

Exercice

Taux de couverture

```
#Taux de couverture intervalle de confiance
>inside_Tot04=(pile_Tot04[,3]<tot[1]) *
    (pile_Tot04[,4]>tot[1])
>inside_TaxableIncome=(pile_TaxableIncome[,3]<tot[2]) *
    (pile_TaxableIncome[,4]>tot[2])
>tc=c(mean(inside_Tot04),mean(inside_TaxableIncome))
>cat("taux de couverture Monte Carlo \n")
options("scipen"=100,digits="3")
>tc
[1] 0.865 0.870

#Coefficient de variation des var. d'intérêt
>cv_Tot04 <- (sd(Tot04) / mean(Tot04)) * 100
>cv_Tot04
[1] 158
>cv_TaxInc <- (sd(TaxableIncome) / mean(TaxableIncome)) * 100
>cv_TaxInc
[1] 156
```

Méthodes d'échantillonnage à probabilités inégales

Introduction

La stratification est une méthode simple permettant de réduire la variance des estimateurs. Si les strates sont homogènes, le sondage aléatoire simple stratifié constitue une stratégie efficace d'échantillonnage (fonction `strata` du package `sampling`).

En pratique, il peut subsister une forte hétérogénéité dans les strates. C'est notamment le cas pour un premier degré d'échantillonnage, e.g. lors de la sélection d'un échantillon de communes pour une enquête auprès des ménages. Dans ce cas, nous pouvons rechercher une stratégie d'échantillonnage plus efficace en individualisant les probabilités de sélection π_k.

Nous devons ensuite faire le choix d'un *algorithme de tirage*, i.e. d'une méthode pratique de sélection respectant les probabilités d'inclusion choisies.

Algorithmes de tirage

Il existe en pratique des dizaines d'algorithmes de tirage permettant de respecter un jeu de probabilités d'inclusion fixé (voir Tillé, 2011), et le package `sampling` permet d'implémenter plusieurs d'entre elles.

Nous présentons rapidement les différentes méthodes d'échantillonnage proposées dans le package `sampling`, et nous étudierons plus en détail quatre d'entre elles.

Remarque importante : la méthode d'échantillonnage sans remise à probabilités inégales programmée dans la fonction de base `sample` est fausse.

Fonction de base sample

```
> sample(x,[n],size,replace = FALSE, prob = NULL)
```

- `x` : vecteur dans lequel sélectionner, ou entier positif.
- `size` : taille d'échantillon (entier positif).
- `replace`: échantillonnage sans remise (FALSE) ou avec remise (TRUE). **L'option FALSE donne une méthode biaisée d'échantillonnage à probabilités inégales**.
- `prob` : vecteur de probabilités, les probabilités d'inclusion sont proportionnelles à `prob` (NULL pour un tirage à probabilités égales).

Fonction de base sample

```
> sample(1:10,6,replace = FALSE)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers selon un SRS : Ok.

```
> prob <- c(1,1,1,1,1,2,2,2,2,2)
> sample(1:10,6,replace = TRUE,prob)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers. Tirage avec remise à probabilités inégales : Ok.

```
> prob <- c(1,1,1,1,1,2,2,2,2,2)
> sample(1:10,6,replace = FALSE,prob)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers. Tirage sans remise à probabilités inégales : méthode de tirage fausse.

Algorithmes du package sampling étudiés

- Tirage systématique `UPsystematic`
- Tirage systématique randomisé `UPrandomsystematic`
- Tirage du pivot `UPpivotal`
- Tirage du pivot randomisé `UPrandompivotal`
- Tirage de Poisson `UPpoisson`
- Tirage de Poisson conditionnel `UPmaxentropy`

Autres algorithmes du package sampling

- Méthode de Brewer (`UPbrewer`), méthode de Sampford (`UPsampford`), échantillonnage ordonné (`UPopips`)
  $\Rightarrow$ proches du tirage de Poisson conditionnel
- Tirage à support minimal (`UPminimalsupport`), méthode de Midzuno (`UPmidzuno`), méthode de Tillé (`UPtille`)
  $\Rightarrow$ peu utilisées en pratique
- Tirage multinomial (`UPmultinomial`)
  $\Rightarrow$ equivalente à la fonction `sample` avec l'option `replace=TRUE`.

Algorithmes étudiés

Le *tirage systématique* et la *méthode du pivot* (Deville et Tillé, 1998) tiennent compte de l'ordre des unités de la population.
Si cet ordre est informatif, cela peut permettre de diminuer la variance de l'estimateur de HT.

Le *tirage de Poisson* et le *tirage de Poisson conditionnel/réjectif* (Hajek, 1964) ne tiennent pas compte de l'ordre des unités de la population.
Ce sont des méthodes de tirage beaucoup plus aléatoires que les deux méthodes précédentes.
L'avantage (et l'inconvénient) est que la variance ne dépend pas de l'ordre des unités dans le fichier.

Tirage systématique

Principe

C'est une méthode simple et très rapide permettant de sélectionner un échantillon à probabilités inégales et de taille fixe.
C'est la méthode la plus utilisée en pratique, même pour un tirage à probabilités égales.

Principe :

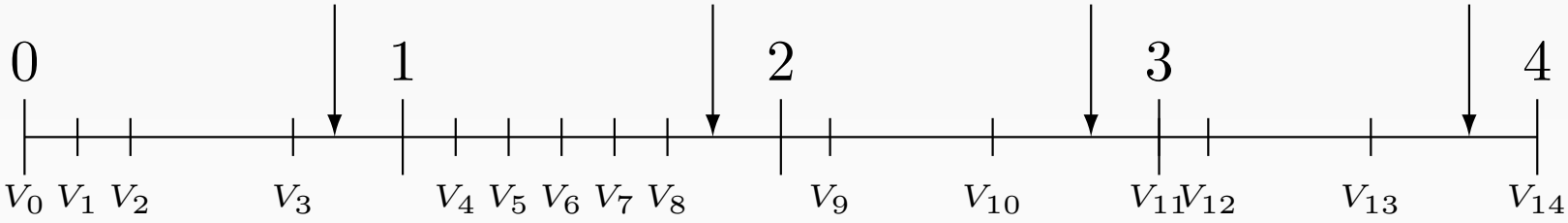
- Les unités de la population sont représentées sur un segment de longueur n. Chaque unité k est représentée par un segment de longueur π_k.
- Nous générons un nombre aléatoire $u \sim U[0,1]$, puis les nombres $u_i = u + (i-1),\ i = 1, \ldots, n-1$.
- Une unité est sélectionnée si un de ces nombres aléatoire tombe dans son segment.

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
#Tirage systématique
> ech_sys=UPsystematic(pi_50)
```

Exemple

Population U de taille $N = 14$ avec $n = 4$:

- $\pi_1 = \pi_2 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_{12} = 1/7$,
- $\pi_3 = \pi_4 = \pi_9 = \pi_{10} = \pi_{11} = \pi_{13} = \pi_{14} = 3/7$.



$$
\begin{aligned}
u = 0.82 \in [V_3, V_4] &\Rightarrow \text{l'unité 4 est sélectionnée,} \\
1 + u = 1.82 \in [V_8, V_9] &\Rightarrow \text{l'unité 9 est sélectionnée,} \\
2 + u = 2.82 \in [V_{10}, V_{11}] &\Rightarrow \text{l'unité 11 est sélectionnée,} \\
3 + u = 3.82 \in [V_{13}, V_{14}] &\Rightarrow \text{l'unité 14 est sélectionnée.}
\end{aligned}
$$

Probabilités d'inclusion

Les probabilités d'inclusion π_k sont exactement respectées. Les probabilités d'inclusion d'ordre deux sont calculables (Tillé, 2011, p. 126), mais beaucoup d'entre elles sont nulles. Par conséquent, il n'existe pas d'estimateur sans biais de variance pour l'estimateur HT.

```
#Probabilités d'inclusion d'ordre 2
> pikl_sys=UPsystematicpi2(pi_50)
> pikl_sys[1:6,1:6]
      [,1]   [,2]   [,3]   [,4]   [,5]  [,6]
[1,] 0.114 0.0000 0.0000 0.0000 0.0000 0.000
[2,] 0.000 0.0747 0.0000 0.0000 0.0000 0.000
[3,] 0.000 0.0000 0.0997 0.0000 0.0000 0.000
[4,] 0.000 0.0000 0.0000 0.0741 0.0000 0.000
[5,] 0.000 0.0000 0.0000 0.0000 0.0901 0.000
[6,] 0.000 0.0000 0.0000 0.0000 0.0000 0.103
```

Cas de probabilités d'inclusion égales

Dans le cas de probabilités d'inclusion égales, la méthode est généralement plus efficace que le SRS si la population est triée avant le tirage selon une variable auxiliaire x_k corrélée avec la variable d'intérêt.

```
#Corrélation entre Tot04 et TaxableIncome
> y=TaxableIncome
> cor(Tot04,y)
[1] 0.988

#Tri de la population selon la variable Tot04
> permutation <- order(Tot04)
> Tot04_rank <- Tot04[permutation]
> y_rank <- y[permutation]

#Paramètres de l'échantillonnage (probabilités égales)
> n <- 50
> Npop <- 589
> pi0_50 <- rep(n/Npop,Npop)
```

Cas de probabilités d'inclusion égales

Comparaison entre SRS et tirage systématique

```
#Probabilités d'inclusion d'ordre 2 pour un SRS
> pikl_srs <- UPsampfordpi2(pi0_50)
#Variance exacte sous un SRS
> var_srs <-    t(y_rank/pi0_50)
            %*%(pikl_srs-pi0_50%*%t(pi0_50))
            %*%(y_rank/pi0_50)
#Probabilités d'inclusion d'ordre 2 pour le SYS
> pikl_sys <- UPsystematicpi2(pi0_50)
#Variance exacte sous un SYS
> var_sys <-    t(y_rank/pi0_50)
             %*%(pikl_sys-pi0_50%*%t(pi0_50))
             %*%(y_rank/pi0_50)

> options("scipen"=-100,digits="3")
> var_srs
[1,] 6.56e+20
> var_sys
[1,] 3.08e+20
```

Estimateur de variance

Beaucoup d'estimateurs de variance ont été proposés dans la littérature pour le tirage systématique, voir par exemple Iachan (1982).

Dans le cas d'un tirage à probabilités égales, on peut notamment citer :

- l'estimateur de variance du sondage aléatoire simple

$$v_{SRS}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2.$$

  Estimateur de variance conservatif en cas d'effet de stratification.

- l'estimateur de variance des différences successives

$$v_{DIFF}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} \times \frac{1}{n} \sum_{i=1}^{n/2} \left\{ y_{(2i)} - y_{(2i-1)} \right\}^2,$$

  avec $y_{(i)}$ la $i^{\text{ème}}$ unité échantillonnée au sens de l'ordre initial du fichier. C'est l'estimateur de variance correspondant à une stratification en $n/2$ strates, avec tirage de 2 éléments dans chacune.

Méthode du pivot

Principe de la méthode (Deville et Tillé, 1998)

Basée sur des duels. A l'étape 1, les unités 1 et 2 s'affrontent :

- si $\pi_1 + \pi_2 \leq 1$, une unité est éliminée et l'autre survit avec la probabilité cumulée :

$$(\pi_1, \pi_2) = \begin{cases} (\pi_1 + \pi_2, 0) & \text{avec proba } \frac{\pi_1}{\pi_1+\pi_2}, \\ (0, \pi_1 + \pi_2) & \text{avec proba } \frac{\pi_2}{\pi_1+\pi_2}. \end{cases}$$

- si $\pi_1 + \pi_2 > 1$, une unité est tirée et l'autre survit avec la probabilité résiduelle :

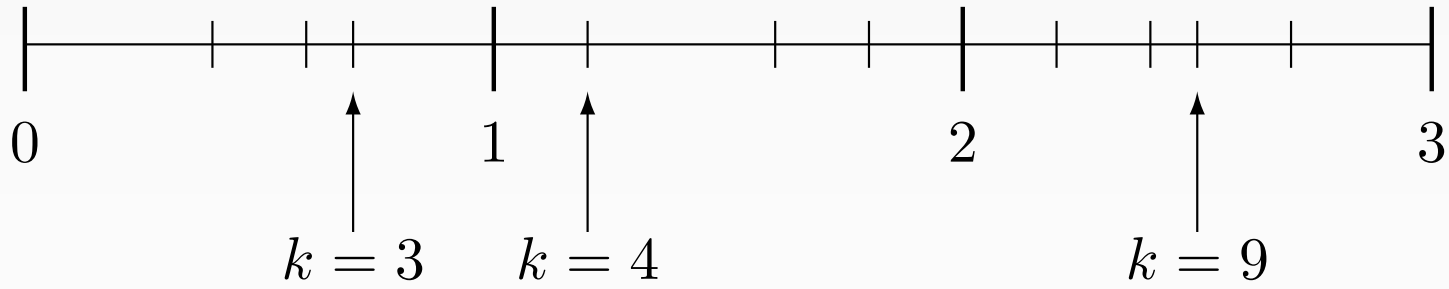
$$(\pi_1, \pi_2) = \begin{cases} (1, \pi_1 + \pi_2 - 1) & \text{avec proba } \frac{1-\pi_2}{2-\pi_1-\pi_2}, \\ (\pi_1 + \pi_2 - 1, 1) & \text{avec proba } \frac{1-\pi_1}{2-\pi_1-\pi_2}. \end{cases}$$

A l'étape t, le survivant affronte l'unité $t+1$ selon le même principe.
A l'étape $N-1$, un échantillon de n unités a été tiré, en respectant les probabilités d'inclusion souhaitées.

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

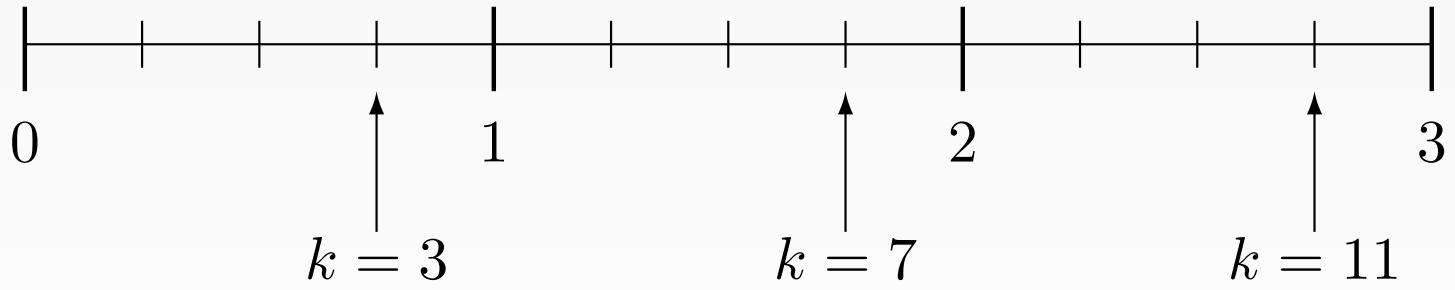
$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



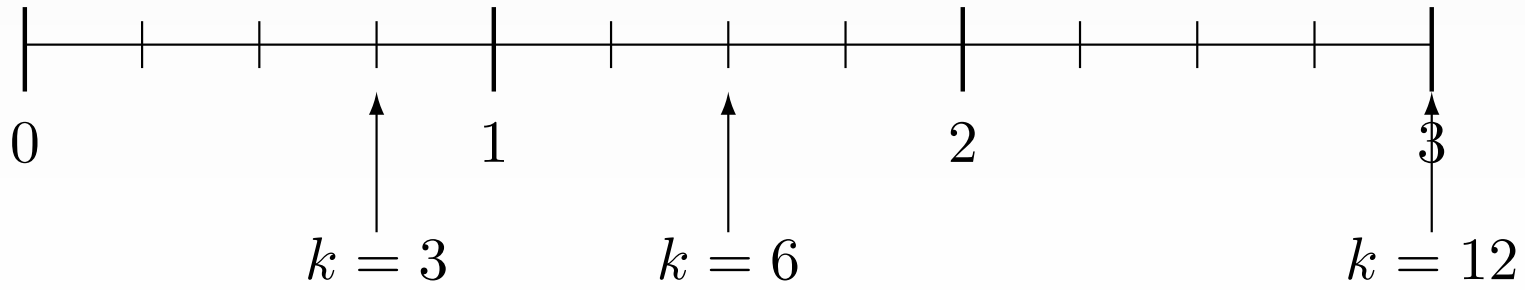
C'est une méthode simple, séquentielle, qui respecte les probas π_k.
Tirage d'une unité par *microstrate* $\Rightarrow$ effet de stratification.
Evite la sélection d'unités contigues $\Rightarrow$ "well-spread sample" (Grafström et al., 2012).
Plus aléatoire que le tirage systématique $\Rightarrow$ bonnes propriétés statistiques.
Cas particulier de la méthode du cube (Deville et Tillé, 2004).

Comparaison pour un tirage à probas égales

Pop. U de taille $N = 12$, avec $n = 3$ et $\pi_k = 3/12$.
Tirage systématique : tirage à la même position dans chaque strate.



Méthode du pivot : tirage indépendant dans chaque strate.



Probabilités d'inclusion

Il et possible de montrer que les probabilités d'inclusion π_k sont exactement respectées. Les probabilités d'inclusion d'ordre deux sont calculables (non disponible dans `sampling`), mais les expressions sont complexes (Deville, 1998; Chauvet, 2012).

Cette méthode de tirage est plus aléatoire que le tirage systématique. Il est possible d'obtenir des propriétés statistiques importantes pour l'estimateur de Horvitz-Thompson (consistance, TCL).

Cette méthode reste peu aléatoire car elle est très contrainte (tirage d'une unité exactement par microstrate). Cela entraîne une baisse de la variance si l'ordre de la population est informatif de la variable d'intérêt.

Estimateur de variance

Beaucoup de couples d'unités présentent des $\pi_{kl} = 0$. Il n'existe donc pas d'estimateur sans biais de variance.

Il est possible d'utiliser :

- l'estimateur de variance pour un tirage avec remise:

$$v_{mult}(\hat{t}_{y\pi}) = \frac{n}{n-1} \sum_{k \in S} \left(\frac{y_k}{\pi_k} - \frac{\hat{t}_{y\pi}}{n} \right)^2 .$$

  C'est un estimateur de variance (généralement très) conservatif.

- un estimateur de variance utilisant des différences successives

$$v_{DIFF}(\hat{t}_{y\pi}) = \sum_{i=1}^{n/2} \left\{ \frac{y_{(2i)}}{\pi_{(2i)}} - \frac{y_{(2i-1)}}{\pi_{(2i-1)}} \right\}^2 ,$$

  avec $y_{(i)}$ la $i^{\text{ème}}$ unité échantillonnée au sens de l'algorithme.
  C'est un estimateur de variance (un peu moins) conservatif (Chauvet et Le Gleut, 2019).

Mise en oeuvre sous R

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)

#Tirage du pivot et estimation du total de TaxableIncome
> ech_piv=UPpivotal(pi_50)
> y=TaxableIncome
> HTestimator(y[ech_piv==1],pi_50[ech_piv==1])
[1,] 1.27e+11

#Tirage du pivot randomisé et estimation
> ech_rpiv=UPrandompivotal(pi_50)
> HTestimator(y[ech_rpiv==1],pi_50[ech_rpiv==1])
[1,] 9.08e+10
```

Tirage de Poisson

Principe

C'est un principe de piles ou faces indépendants, avec une pièce et un lancer différents pour chaque unité.

- Etape 1 : on génère $u_1 \sim U[0,1]$. Si $u_1 \leq \pi_1$, l'unité 1 est retenue dans l'échantillon.
- Etape 2 : on génère $u_2 \sim U[0,1]$ indépendamment de u_1. Si $u_2 \leq \pi_2$, l'unité 2 est retenue dans l'échantillon.
- ...
- Etape N : on génère $u_N \sim U[0,1]$ indépendamment de $u_1, \ldots, u_{N-1}$. Si $u_N \leq \pi_N$, l'unité N est retenue dans l'échantillon.

En utilisant les propriétés d'une loi $U[0,1]$ et l'indépendance des tirages :

$$\begin{aligned} \mathbb{P}(k \in S) &= \mathbb{P}(u_k \leq \pi_k) = F_U(\pi_k) = \pi_k, \\ \pi_{kl} &= \pi_k \, \pi_l \text{ si } k \neq l. \end{aligned}$$

Dans le cas d'un tirage à probabilités égales, on parle de plan de Bernoulli.

Estimateur de Horvitz-Thompson

La variance s'obtient à partir de l'expression générale de HT :

$$V_{pois}(\hat{t}_{y\pi}) = \sum_{k \in U} \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k),$$

qui est estimée sans biais par

$$v(\hat{t}_{y\pi}) = \sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k).$$

En particulier, cela implique que la taille d'échantillon est aléatoire :

$$V_{pois}\{n(S)\} = \sum_{k \in U} \pi_k (1 - \pi_k).$$

Utilisation

Le tirage de Poisson présente une grande variance d'échantillonnage. Il est cependant utilisé pour certaines enquêtes auprès des entreprises, car il permet de simplifier la coordination du tirage de plusieurs échantillons.

On parle de coordination :

- négative quand on tire plusieurs échantillons afin qu'ils soient aussi disjoints que possible,
- positive quand on tire plusieurs échantillons afin qu'ils se recouvrent autant que possible.

Le tirage de Poisson est également utilisé dans un contexte de *non-réponse*, pour modéliser le mécanisme de réponse totale dans l'échantillon S complet.

Mise en oeuvre sous R

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)

#Tirage de Poisson et estimation du total de TaxableIncome
> ech_poi=UPpoisson(pi_50)
> y=TaxableIncome
> HTestimator(y[ech_poi==1],pi_50[ech_poi==1])
[1,] 1.220165e+11

#Estimation de variance de HT
> pikl_poi_50=pi_50 %*% t(pi_50) +diag(pi_50-pi_50*pi_50)
> varHT(y[ech_poi==1],pikl_poi_50[ech_poi==1,ech_poi==1],1)
[1] 6.1382e+20

#Estimation de variance (package GUSTAVE)
y_mat <- matrix(y, ncol = 1)
var_pois(y_mat[ech_poi==1, , drop = FALSE],pi_50[ech_poi==1])
[1] 6.1382e+20
```

Tirage réjectif ou tirage de Poisson conditionnel

Principe

Nous cherchons à obtenir un plan de sondage :

- avec les avantages du tirage de Poisson : une grande entropie

$$\mathcal{L}(p) = -\sum_{s \subset U} p(s) \ln\{p(s)\},$$

  qui assure que l'échantillonnage n'est pas sensible à l'ordre des données,
- sans ses inconvénients : taille d'échantillon aléatoire.

Le plan de sondage réjectif est obtenu :

- en tirant un échantillon selon un plan de Poisson de probabilités d'inclusion $p_k,\ k \in U$, avec $\sum_{k \in U} p_k = n$;
- en rejetant l'échantillon tant qu'il n'est pas de la taille voulue n.

Plan de sondage

Nous notons :

- $p(\cdot)$ le plan de Poisson et S_p l'échantillon aléatoire correspondant,
- $p_r(\cdot)$ le plan réjectif associé, et S_r l'échantillon associé.

Pour tout $s \subset U$, nous avons

$$p_r(s) \equiv Pr(S_p = s | n(S_p) = n).$$

Les probabilités d'inclusion π_k du plan $p_r(\cdot)$ ne sont pas égales aux probabilités d'inclusion p_k du plan $p(\cdot)$.

Pour pouvoir calculer l'estimateur de Horvitz-Thompson, il faut pouvoir calculer les probabilités d'inclusion effectives π_k.

Exemple

Soit une population U de taille 5. Nous utilisons un plan de Poisson $p(\cdot)$ avec les probabilités d'inclusion

$$p_1 = p_2 = \frac{1}{2} \qquad p_3 = p_4 = p_5 = \frac{1}{3}.$$

Nous mettons en oeuvre ce plan de Poisson en ne retenant que les échantillons de taille $\sum_{k \in U} p_k = 2$. Le plan réjectif obtenu a pour probabilités d'inclusion

$$\pi_1 = \pi_2 = \frac{10}{19} \qquad \pi_3 = \pi_4 = \pi_5 = \frac{6}{19}.$$

Mise en oeuvre d'un plan réjectif

Le tirage peut être réalisé à l'aide de la fonction `UPmaxentropy` du package `sampling`. Dans le cas particulier de probabilités d'inclusion égales, la méthode est équivalente au sondage aléatoire simple sans remise.

Comme le tirage est de taille fixe par construction, il est possible d'utiliser la formule de Yates-Grundy :

$$V_p(\hat{t}_{y\pi}) = \frac{1}{2} \sum_{k \neq l \in U} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 (\pi_k \pi_l - \pi_{kl}),$$

et l'estimateur de variance de Yates-Grundy correspondant.

La matrice des probabilités d'inclusion d'ordre 2 peut être déterminée à l'aide de la fonction `UPmaxentropypi2`.

Mise en oeuvre d'un plan réjectif (suite)

Il est également possible d'utiliser une approximation uniforme des π_{kl} (Hajek, 1964). Elle conduit à l'estimateur de variance (Deville, 1993) :

$$\begin{aligned} v_{dev}(\hat{t}_{y\pi}) &= \frac{1}{1-\sum_{k\in S}(a_k)^2}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k}-\sum_{l\in S}a_l\frac{y_l}{\pi_l}\right)^2 \\ \text{avec } a_l &= \frac{1-\pi_l}{\sum_{m\in S}1-\pi_m}. \end{aligned}$$

Cet estimateur est couramment utilisé dans les enquêtes Insee (Caron et al., 1998). Il est calculable :

- avec la fonction `varest` du package "sampling",
- avec la fonction `varDT` du package "gustave".

Mise en oeuvre sous R

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
> options("scipen"=-100,digits="5")

#Tirage réjectif et estimation du total de TaxableIncome
> ech_rej=UPmaxentropy(pi_50)
> y=TaxableIncome
> est_ht=HTestimator(y[ech_rej==1],pi_50[ech_rej==1])
> est_ht
[1,] 1.3068e+11
```

Mise en oeuvre sous R

Estimation de variance

```
#Estimateur de variance de HT
> pikl_rej_50=UPmaxentropypi2(pi_50)
> varHT(y[ech_rej==1],pikl_rej_50[ech_rej==1,ech_rej==1],1)
[1] 1.1896e+21

#Estimateur de variance de YG
> varHT(y[ech_rej==1],pikl_rej_50[ech_rej==1,ech_rej==1],2)
[1] 1.1907e+21

#Estimateur de variance de Deville (package SAMPLING)
> varest(y[ech_r==1],,pi_50[ech_r==1],)
[1] 1.1897e+21

#Estimateur de variance de Deville (package GUSTAVE)
> varDT(y[ech_rej==1],pi_50[ech_rej==1])
[1] 1.1897e+21
```

Echantillonnage équilibré

Principe

Choix du plan de sondage

Le plan de sondage est choisi de façon à minimiser la variance des estimateurs, tout en respectant des contraintes de coût.

- stratification, tirage à probabilités inégales
  $\Rightarrow$ réduction de la variance
- tirage multidegrés
  $\Rightarrow$ réduction des coûts

La précision du plan repose sur des **propriétés d'équilibrage** : l'échantillon est sélectionné de façon à respecter une information connue.

Exemples :

- respect de structures âge-sexe (méthode des quotas),
- répartition par effectif salarié (stratification),
- taille fixe d'échantillon (tirage systématique, méthode du pivot, tirage réjectif).

Echantillonnage équilibré

De façon générale, supposons que des variables $\mathbf{x}_k$ sont disponibles au moment de l'échantillonnage pour chaque individu k de la population.

Un échantillon s est dit équilibré sur les totaux $t_{\mathbf{x}}$ si

$$\hat{t}_{\mathbf{x}\pi}(s) = t_{\mathbf{x}}.$$

Le total $t_{\mathbf{x}}$ est donc parfaitement estimé.

Par extension, un plan de sondage est dit équilibré sur les totaux $t_{\mathbf{x}}$ si seuls les échantillons équilibrés sur $\mathbf{x}$ ont une probabilité non nulle d'être sélectionnés.

Exemples d'équations d'équilibrage

Supposons que $x_k = \pi_k$. L'équation d'équilibrage implique que

$$\begin{aligned} \sum_{k \in s} \frac{x_k}{\pi_k} &= \sum_{k \in s} \frac{\pi_k}{\pi_k} = n(s) \\ &= \sum_{k \in U} \pi_k = E_p\left[n\left(S\right)\right]. \end{aligned}$$

Le plan de sondage est donc de taille fixe.

Supposons que $x_k = 1$. L'équation d'équilibrage implique que

$$\begin{aligned} \sum_{k \in s} \frac{x_k}{\pi_k} &= \sum_{k \in s} \frac{1}{\pi_k} = \widehat{N}_\pi \\ &= \sum_{k \in U} 1 = N. \end{aligned}$$

La taille de la population est donc parfaitement estimée.

Exemples de plans de sondage équilibrés

Les plans de sondage à probabilités inégales de taille fixe sont équilibrés sur la variable $x_k = \pi_k$.

Le sondage aléatoire simple est équilibré sur la variable $x_k = 1$
$\Rightarrow$ plan de taille fixe + taille de la population parfaitement estimée.

Le sondage aléatoire simple stratifié est équilibré sur le vecteur

$$x_k = \{1(k \in U_1), \ldots, 1(k \in U_H)\}$$

Conséquences ?

Motivation

Sous le modèle de travail

$$y_k = \mathbf{x}_k^\top \beta + \epsilon_k,$$

l'estimateur de HT peut être réécrit sous la forme

$$\hat{t}_{y\pi} = \{\hat{t}_{\mathbf{x}\pi}\}^\top \beta + \hat{t}_{\epsilon\pi}.$$

Principe :

- Le **respect** des probabilités d'inclusion permet d'obtenir une estimation sans biais.
- La **restriction du support** du plan de sondage aux échantillons équilibrés permet d'annuler la variabilité du 1er terme.
- Le **choix** des probabilités d'inclusion permet de limiter la variabilité du 2nd terme.

La variance n'est plus donnée que par les résidus du modèle de travail.

Remarques

Remarque 1 : en pratique, la base de sondage contient toujours au moins deux variables : la probabilité d'inclusion π_k et la variable constante. Par rapport au tirage de taille fixe à probabilités inégales, cela revient à ajouter une constante dans le modèle de régression

$$y_k = \beta + \alpha\ x_k + \epsilon_k.$$

Remarque 2 : la non-réponse totale va détruire l'équilibrage. L'échantillonnage équilibré est donc particulièrement intéressant pour un premier degré de tirage ou quand on anticipe une faible non-réponse:

- tirage des Unités Primaires de l'Echantillon Maître,
- tirage des Groupes de Rotation du Recensement.

La méthode du Cube

Représentation du Cube

Deville et Tillé (2004) ont proposé un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec un jeu de probabilités d'inclusion $\pi = (\pi_1, \ldots, \pi_N)$ quelconque.

Un échantillon s est vu comme un sommet $(s_1, \ldots, s_N) \in \{0,1\}^N$ du N-cube $\mathrm{C} = [0,1]^N$. Les équations d'équilibrage définissent l'espace des contraintes :

$$\sum_{k \in S} \frac{\mathbf{x_k}}{\pi_k} = \sum_{k \in U} \mathbf{x_k}$$

$$\Longleftrightarrow \sum_{k \in U} \frac{\mathbf{x_k}}{\pi_k}(I_k - \pi_k) = 0$$

$$\Longleftrightarrow A \times (I - \pi) = 0 \text{ avec } A = (\mathbf{x_k}/\pi_k)_{k \in U}$$

$$\Longleftrightarrow I \in \pi + Ker(A).$$

L'algorithme consiste à arrondir aléatoirement des composantes du vecteur π par une marche aléatoire dans l'espace des contraintes.

Etape de base de l'algorithme

Nous initialisons avec $\pi^{(0)} = \pi$.
A l'étape t, soit $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$ avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t)\ u(t) & \text{avec proba. } \lambda_2(t)/(\lambda_1(t)+\lambda_2(t)) \\ -\lambda_2(t)\ u(t) & \text{avec proba. } \lambda_1(t)/(\lambda_1(t)+\lambda_2(t)) \end{cases},$$

où

- $\lambda_1(t), \lambda_2(t) > 0$
  $\rightarrow$ assure qu'au moins une unité est sélectionnée ou définitivement rejetée.
- $u(t) \in Ker(A)$ est un vecteur (non aléatoire)
  $\rightarrow$ assure que les équations d'équilibrage sont exactement respectées
- le choix aléatoire assure que les probabilités d'inclusion sont exactement respectées.

La méthode du pivot est un cas particulier de la méthode du Cube, obtenue avec $\mathbf{x}_k = \pi_k$ (échantillonnage de taille fixe).

La phase d'atterrisage

L'algorithme précédent est appelé la **phase de vol**. A l'issue de cet algorithme :

- Le statut (tiré/non tiré) est connu pour au moins $N - p$ individus.
- Les contraintes d'équilibrage et les probabilités d'inclusion sont exactement respectées.
- En revanche, il n'est plus possible de finir l'échantillonnage en respectant ces deux contraintes.

La phase de vol est complétée par une **phase d'atterrissage**. Elle permet de statuer sur les individus restant en respectant **exactement** les probabilités d'inclusion, et en respectant **approximativement** les équations d'équilibrage.

Phase d'atterrissage : relâchement des contraintes

La $1^{\text{ère}}$ possibilité consiste à relâcher les contraintes une par une.

Nous introduisons donc un degré de liberté à chaque fois, ce qui permet de poursuivre l'échantillonnage.

C'est l'option la plus générale, au sens où elle permet de travailler sur un nombre quelconque de variables d'équilibrage. Mais les premières variables relâchées peuvent être mal équilibrées.

Phase d'atterrissage : échantillon optimal

La $2^{\text{ème}}$ possibilité consiste à définir un plan de sondage sur les unités restantes :

- respectant les probabilités d'inclusion de départ,
- minimisant (en moyenne) l'écart à l'équilibre, à l'aide d'un critère de type

$$\min E \left\| \hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}} \right\|^2 .$$

Cette option permet d'obtenir un bon équilibrage global. Elle nécessite de définir entièrement un plan de sondage sur une population de p individus. C'est possible si le nombre de contraintes est faible, mais impraticable si p est grand
($p = 19 \Rightarrow 500\ 000$ échantillons possibles environ)

Fonction "samplecube"

Extrait de la documentation de "sampling"

Selects a balanced sample (a vector of 0 and 1) or an almost balanced sample. Firstly, the flight phase is applied. Next, if needed, the landing phase is applied on the result of the flight phase.

```
samplecube(X,pik,order=1,comment=TRUE,method=1)
```

Arguments:

- X: matrix of auxiliary variables on which the sample must be balanced.
- pik: vector of inclusion probabilities.
- order
  - 1: the data are randomly arranged,
  - 2: no change in data order,
  - 3: the data are sorted in decreasing order.

Fonction "samplecube"

Extrait de la documentation de "sampling"

Selects a balanced sample (a vector of 0 and 1) or an almost balanced sample. Firstly, the flight phase is applied. Next, if needed, the landing phase is applied on the result of the flight phase.

```
samplecube(X,pik,order=1,comment=TRUE,method=1)
```

Arguments (continued):

- comment: a comment is written during the execution if comment is TRUE.
- method
  1: for a landing phase by linear programming,
  2: for a landing phase by suppression of variables.

Exemple extrait de la documentation de "sampling"

```
> data(MU284)
# Computation of the inclusion probabilities
> pik=inclusionprobabilities(MU284$P75,50)
# Definition of the matrix of balancing variables
> X=cbind(MU284$P75,MU284$CS82,
          MU284$SS82,MU284$S82,MU284$ME84)
# Computation of the Horvitz-Thompson estimator for a
    balanced sample
> s=samplecube(X,pik,1,TRUE)

BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 5e+00 variables and 284
    units
The size of the inclusion probability vector is 284
The sum of the inclusion probability vector is 5e+01
The inclusion probability vector has 281 non-integer elements
Step 1
```

Exemple extrait de la documentation de "sampling"

```
BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain 5 non integer
    probabilities
The sum of these probabilities is 3e+00
This sum is integer
The linear program will consider 10 possible samples
The mean cost is 1.561e-02
The smallest cost is 7.617e-04
The largest cost is 3.63e-02
The cost of the selected sample is 7.617e-04

QUALITY OF BALANCING
     TOTALS HorvitzThompson_estimators Relative_deviation
1 8.182e+03                 8.182e+03       -5.558e-14
2 2.583e+03                 2.589e+03        2.248e-01
3 6.301e+03                 6.354e+03        8.423e-01
4 1.350e+04                 1.361e+04        7.909e-01
5 5.052e+05                 5.051e+05       -3.000e-02
```

Estimation de variance

Estimation de variance

Il est théoriquement possible d'utiliser l'estimateur de variance de HT, ou celui de YG si la probabilité d'inclusion fait partie des contraintes d'équilibrage.

En pratique, les probabilités d'inclusion d'ordre 2 sont presque impossibles à calculer, en dehors de cas particuliers (e.g., méthode du pivot).

Il est possible d'obtenir une approximation par simulations de la matrice des probabilités d'inclusion d'ordre 2 (e.g., Breidt et Chauvet, 2011), mais un très grand nombre de simulations est nécessaire pour obtenir un estimateur numériquement stable.

Approximation de variance de Deville et Tillé

Deville et Tillé (2005) ont proposé une classe d'estimateurs de variance, sous les hypothèses suivantes :

1. le plan de sondage est **exactement équilibré**,
2. le plan de sondage est à **entropie maximale**, parmi les plans équilibrés sur les mêmes variables $\mathbf{x}_k$, avec les mêmes probabilités d'inclusion π_k.

La condition 1 (équilibrage exact) n'est généralement pas vérifiée en raison de la phase d'atterrissage. L'approximation de variance de Deville et Tillé (2005) prend essentiellement en compte la variance due à la phase de vol.

La condition 2 (entropie maximale) n'est pas nécessairement réalisée, notamment si l'algorithme du Cube est appliqué sur un fichier trié préalablement selon une variable auxiliaire.
Pour qu'elle soit approximativement vérifiée, il est possible de trier aléatoirement les unités de la population avant d'appliquer la méthode du Cube (option `order=1` de la fonction "samplecube").

Approximation de variance de Deville et Tillé

Deville et Tillé (2005) montrent que la variance est approx. celle d'un tirage de Poisson, pour les résidus de la régression de y sur $\mathbf{x}$:

$$
\begin{aligned}
V_{app}(\hat{t}_{y\pi}) &= \frac{N}{N-p}\sum_{k\in U}\pi_k(1-\pi_k)\left(\frac{E_k}{\pi_k}\right)^2, \\
\text{avec } E_k &= y_k - \mathbf{x}_k^\top \mathbf{B} \\
\text{et } \mathbf{B} &= \left\{\sum_{k\in U}\pi_k(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{\mathbf{x}_k^\top}{\pi_k}\right\}^{-1}\sum_{k\in U}\pi_k(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{y_k}{\pi_k}.
\end{aligned}
$$

Par substitution, nous obtenons l'estimateur de variance :

$$
\begin{aligned}
\hat{V}_{DT}(\hat{t}_{y\pi}) &= \frac{n}{n-p}\sum_{k\in S}(1-\pi_k)\left(\frac{e_k}{\pi_k}\right)^2, \\
\text{avec } e_k &= y_k - \mathbf{x}_k^\top \hat{\mathbf{B}}_\pi \\
\text{et } \hat{\mathbf{B}}_\pi &= \left(\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{\mathbf{x}_k^\top}{\pi_k}\right)^{-1}\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{y_k}{\pi_k}.
\end{aligned}
$$

Mise en oeuvre sous R

```
# Matrice des variables d'équilibrage
> X=cbind(MU284$P75,MU284$CS82,MU284$SS82,MU284$S82)

# Estimateur de Horvitz-Thompson
> s=samplecube(X,pik,1,TRUE)
> y <- MU284$RMT85
> HTestimator(y[s==1],pik[s==1])
[1,] 68783

# Estimation de variance DT : package GUSTAVE
> varDT(y[s==1],pik[s==1],X[s==1,])
[1] 487627
```

Application

La base de sondage est le fichier commune. Nous nous intéressons à l'estimation des variables : nombre d'actifs (variable ACTIFS), nombre d'inactifs (variable INACTIFS), nombre d'étrangers de l'Union européenne (variable NATUE).

1- Sélectionner un échantillon de taille 50, à probabilités égales, équilibré sur la variable : probabilité d'inclusion. Commenter les sorties.
A quelle contrainte correspond l'équilibrage sur la probabilité d'inclusion ?

2- Procéder aux estimations demandées, et donner l'estimateur de variance de Deville-Tillé associé.

Application

3- Sélectionner un échantillon de taille 50, à probabilités égales, équilibré sur les variables :

- Probabilité d'inclusion,
- Nombre de logements.

A quelles contraintes correspond l'équilibrage sur ces deux variables?

4) Procéder aux estimations demandées. Comparer avec les estimations de variance obtenues à la question 2.

Application

5- Sélectionner un échantillon de taille 100, à probabilités proportionnelles au nombre de logements, équilibré sur les variables :

- Nombre de logements,
- Variable constante égale à 1,
- Variables croisées âge-sexe : `F0019` , ..., `H7599` (10 variables).

A quelle contrainte correspond l'équilibrage sur la variable constante égale à 1? Pourquoi ne pas équilibrer sur la probabilité d'inclusion ?

6- Commenter les sorties, et procéder aux estimations demandées.

Le Recensement

Principe

La méthode du Cube a été utilisée pour les Enquêtes Annuelles de Recensement. Le plan de sondage utilisé (Godinot, 2005) distingue :

- les grandes communes ($10\ 000$ habitants ou plus au RP 1999)
  $\Rightarrow$ au sein de chacune, sélection et enquête auprès d'un échantillon d'adresses.
- les petites communes (moins de $10\ 000$ habitants)
  $\Rightarrow$ au sein de chaque région, échantillonnage de petites communes dont toutes les adresses sont enquêtées.

Les échantillons du Nouveau Recensement ont été sélectionnés selon des principes de coordination négative : les échantillons sont non chevauchants d'une année sur l'autre.

Principe de constitution des groupes de rotation

Soit U dans laquelle on dispose d'un vecteur $\mathbf{x}_k$ de variables auxiliaires. On tire S_1 avec des probabilités d'inclusion $\pi_{1k} \equiv \pi$, en équilibrant sur le vecteur $\mathbf{x}_k$.
Alors l'échantillon $U \setminus S_1$ est :

- tiré avec des probabilités d'inclusion $1 - \pi$,
- équilibré sur les variables $\mathbf{x}_k$.

Nous tirons dans $U \setminus S_1$ un échantillon S_2 avec des probabilités d'inclusion conditionnelles $\pi_{2k|\bar{1}} \equiv \dfrac{\pi}{1-\pi}$, en équilibrant sur le vecteur $\mathbf{x}_k$.
Alors non conditionnellement, l'échantillon S_2 est :

- tiré avec des probabilités d'inclusion π,
- équilibré sur les variables $\mathbf{x}_k$.

Le cas des petites communes

Les résultats précédents sont utilisés pour partitionner aléatoirement, au sein de chaque région, les petites communes en 5 groupes de rotation. Ils sont tirés à probabilités égales, en équilibrant sur des variables socio-démographiques et la population par département.

Les groupes de rotation sont sélectionnés successivement :

- le GR S_1 est tiré dans U avec des probas $\pi \equiv \frac{1}{5}$,
- le GR S_2 est tiré dans $U \setminus S_1$ avec des probas $\frac{\pi}{1-\pi} \equiv \frac{1}{4}$,
- le GR S_3 est tiré dans $U \setminus \{S_1 \cup S_2\}$ avec des probas $\frac{\pi}{1-2\pi} \equiv \frac{1}{3}$,
- le GR S_4 est tiré dans $U \setminus \{S_1 \cup S_2 \cup S_3\}$ avec des probas $\frac{\pi}{1-3\pi} \equiv \frac{1}{2}$,
- le GR S_5 est donné par le reste de la population.

Une année donnée, toutes les adresses d'un groupe de rotation de petites communes sont enquêtées. Nous avons donc l'exhaustivité sur un cycle de 5 ans, mais un décalage temporel dans les données collectées.

Application : découpage de "commune" en 4 groupes de rotation

Partitionner aléatoirement la table commune en 4 échantillons de taille 250, sélectionnés à probabilités égales et équilibrés sur les variables : probabilité d'inclusion, nombre de Logements, nombre d'hommes, nombre de femmes.

```
#Tirage du 1er groupe de rotation
> n=250
> Npop=1000
> pi=rep(n/Npop,Npop)

> FEM=f0019+f2039+f4059+f6074+f7599
> HOM=h0019+h2039+h4059+h6074+h7599
> X=cbind(pi,FEM,HOM,NLOG)
> ech1=samplecube(X,pi,1,TRUE)
> ident_ech1=ident[ech1==1]
```

Application (suite)

```
#Tirage du second groupe de rotation
> n=250
> Npop=750
> pi=rep(n/Npop,Npop)
> ident_reste=ident[ech1==0]
> Xreste=X[ech1==0,]
> ech2=samplecube(Xreste,pi,1,TRUE)
> ident_ech2=ident_reste[ech2==1]
```

Application (fin)

```
#Tirage des groupes de rotation 3 et 4
> n=250
> Npop=500
> pi=rep(n/Npop,Npop)
> ident_reste=ident_reste[ech2==0]
> Xreste=Xreste[ech2==0,]
> ech3=samplecube(Xreste,pi,1,TRUE)
> ident_ech3=ident_reste[ech3==1]
> ident_ech4=ident_reste[ech3==0]
```

Le cas des grandes communes

Au sein de chaque grande commune, les adresses sont (schématiquement) réparties en trois strates :

- Les grandes adresses (plus de 60 logements),
- Les adresses neuves,
- Les autres adresses.

Chacune de ces strates fait l'objet d'un plan de sondage spécifique.

Les grandes adresses sont réparties (aléatoirement ou non) en 5 groupes de rotation, et un groupe est enquêté exhaustivement chaque année (idem pour les adresses neuves).

Le cas des grandes communes (suite)

Les autres adresses sont partitionnées aléatoirement en 5 groupes de rotation, sélectionnés à probabilités égales, en équilibrant sur des variables socio-démographiques et le nombre de logements par IRIS.

La technique est la même que pour les petites communes. Une année donnée, 40% (environ) des adresses d'un groupe de rotation sont sélectionnées et enquêtées.

En résumé, le plan de sondage est ici stratifié par grande commune et type d'adresse. Selon la strate, l'échantillon annuel est sélectionné en une ou deux phases de tirage, aléatoirement ou non.

Echantillonnage spatial

Echantillonnage spatial en population finie

Application à l'échantillonnage spatial

Dans un contexte spatial, première loi de géographie de Tobler : "Everything is related to everything else, but near things are more related than distant things".

Modèle de travail de type (voir Grafström and Tillé, 2013) :

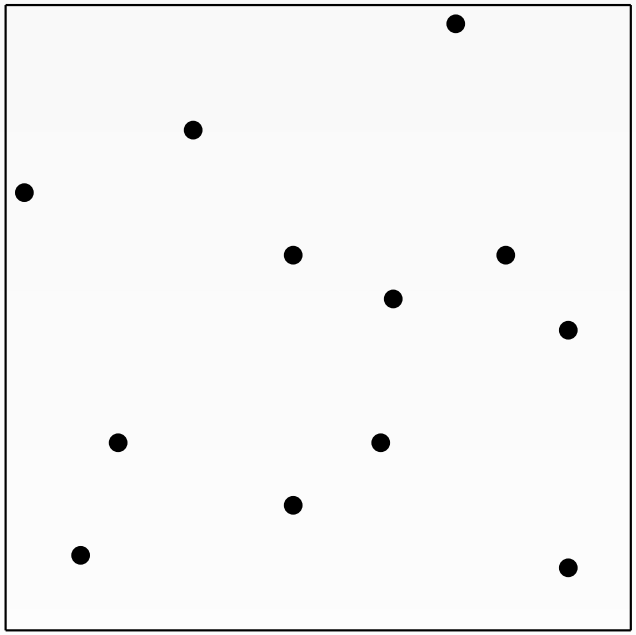
$$\begin{aligned} y_k &= \beta\pi_k + \epsilon_k, \\ E_m(\epsilon_k) = 0 \quad &\text{et} \quad Cov_m(\epsilon_k, \epsilon_l) = \sigma_k\sigma_l\rho^{d(k,l)}. \end{aligned}$$

$\Rightarrow$ il vaut mieux éviter de tirer des unités contigues, qui portent une information similaire.
$\Rightarrow$ il est préférable de bien répartir l'échantillon dans l'espace.

Il est possible d'incorporer plus d'information auxiliaire dans le plan de sondage, ce qui permet d'avoir des stratégies plus efficaces (Grafström and Tillé, 2013; Le Gleut, 2017).

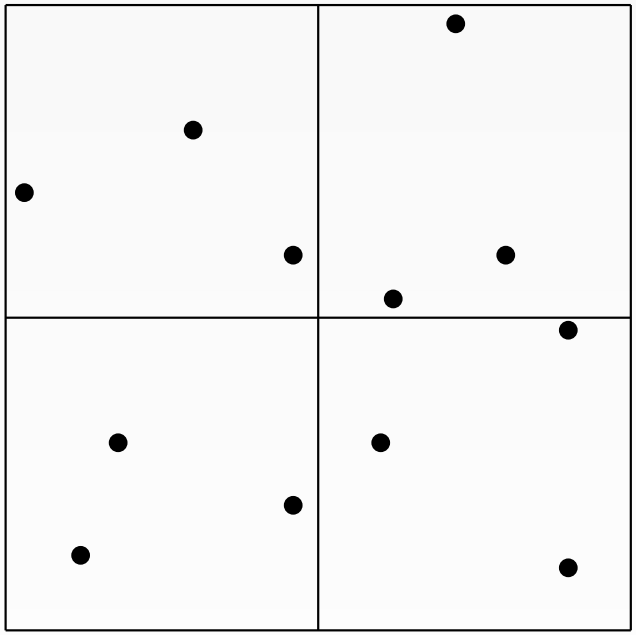
Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)



Generalized Random Tesselation Sampling (GRTS)

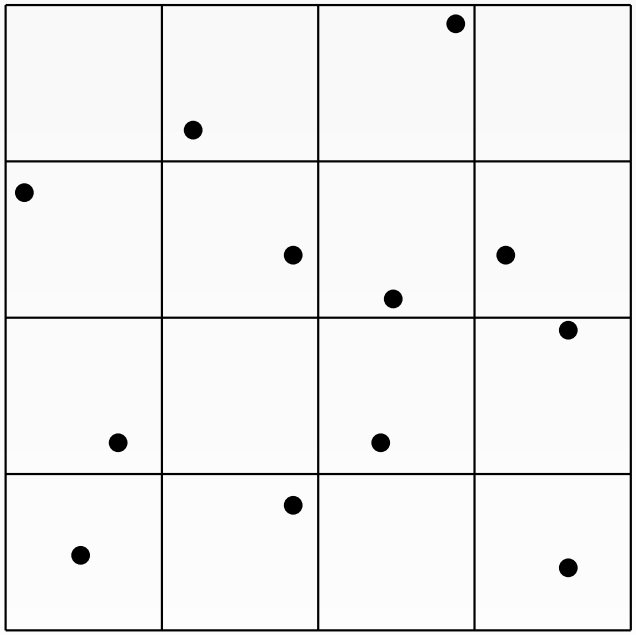
Stevens and Olsen (2004)



Tesselation de la zone selon une grille régulière, avec des "adresses".

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)

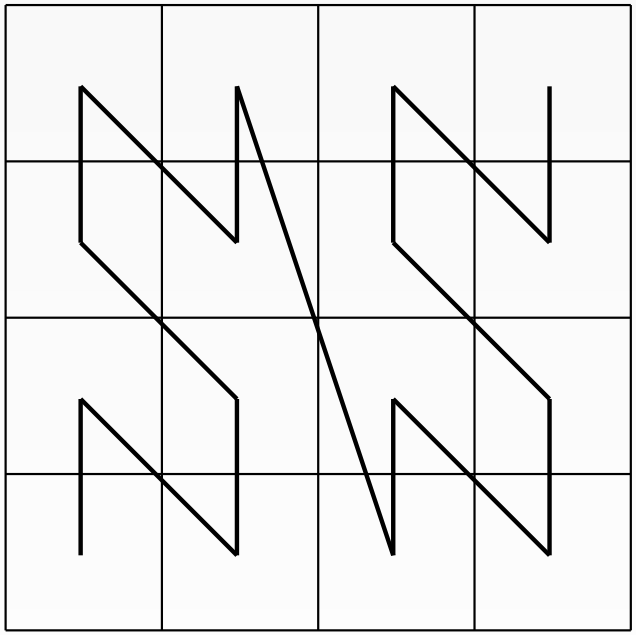


Tesselation de la zone selon une grille régulière, avec des "adresses".

Les adresses sont triées sur une ligne.

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)

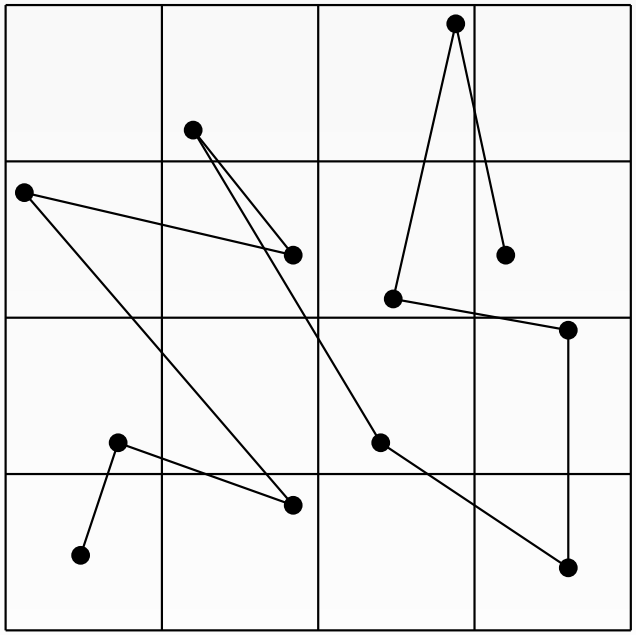


Tesselation de la zone selon une grille régulière, avec des "adresses".

Les adresses sont triées sur une ligne.

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)

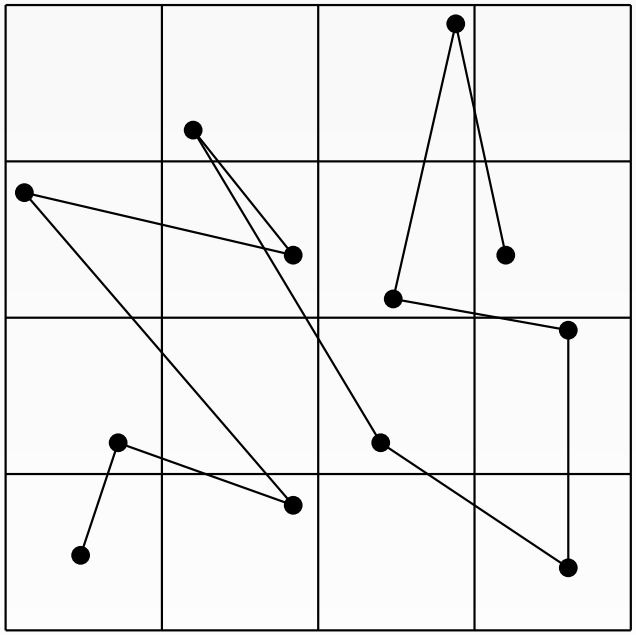


Tesselation de la zone selon une grille régulière, avec des "adresses".

Les adresses sont triées sur une ligne.

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)



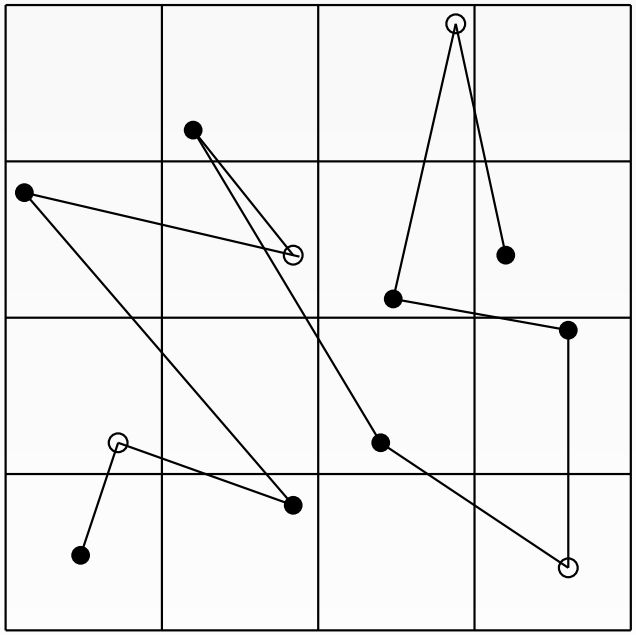
Tesselation de la zone selon une grille régulière, avec des "adresses".

Les adresses sont triées sur une ligne.

L'échantillon est sélectionné par tirage systématique.

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)



Tesselation de la zone selon une grille régulière, avec des "adresses".

Les adresses sont triées sur une ligne.

L'échantillon est sélectionné par tirage systématique.

Pivotal Tesselation Sampling (PTS)

La méthode GRTS donne des échantillons bien équilibrés dans l'espace (Stevens and Olsen, 2004), mais avec un tirage systématique les propriétés statistiques des estimateurs sont difficiles à établir, même si les unités sont partiellement randomisées le long de la ligne.

Une possibilité consiste à utiliser la méthode de tesselation, mais en remplaçant le tirage systématique par la méthode du pivot (Chauvet et Le Gleut, 2019). L'échantillon est toujours bien réparti dans l'espace, et avec de bonnes propriétés statistiques.

Il est également possible d'utiliser la méthode du pivot avec d'autre plans de sondage spatiaux qui utilisent une forme de tri des unités (voir par exemple Dickson and Tillé, 2015).

Méthode du pivot local

Les méthodes précédentes (GRTS, PTS) consistent à projeter un espace de dimension $d \geq 2$ dans un espace de dimension 1, dans lequel un algorithme d'échantillonnage usuel peut être appliqué (tirage systématique ou méthode du pivot).

Grafström et al. (2012) ont proposé une autre méthode appelée le pivot local, qui ne nécessite pas de se projeter dans une dimension plus petite.

L'idée consiste à utiliser la méthode du pivot, en choisissant à chaque étape des unités très proches pour leur appliquer l'étape de base. La méthode permet donc d'éviter la sélection d'unités contigues.

Méthode du pivot local : version 1

1. Une unité i est choisie aléatoirement.
2. L'unité j qui est le plus proche voisin de i est choisie (tirage aléatoire en cas d'égalité).
3. L'étape de base de la méthode du pivot est appliquée à i et j si i est également un plus proche voisin de j. Sinon, retour à l'étape 1.
4. Si toutes les probabilités des unités sont arrondies aléatoirement à 0 ou 1, l'algorithme s'arrête. Sinon, retour à l'étape 1.

Le nombre d'opérations pour sélectionner un échantillon selon cette méthode est de l'ordre de N^3.

Méthode du pivot local : version 2

1. Une unité i est choisie aléatoirement.
2. L'unité j qui est le plus proche voisin de i est choisie (tirage aléatoire en cas d'égalité).
3. L'étape de base de la méthode du pivot est appliquée à i et j ~~si i est également un plus proche voisin de j~~. Sinon, retour à l'étape 1.
4. Si toutes les probabilités des unités sont arrondies aléatoirement à 0 ou 1, l'algorithme s'arrête. Sinon, retour à l'étape 1.

Le nombre d'opérations pour sélectionner un échantillon selon cette méthode est de l'ordre de N^2.

Echantillonnage spatial doublement équilibré

Grafström et Tillé (2013) ont proposé une méthode d'échantillonnage spatial doublement équilibrée :

- l'échantillon est tiré de façon à être bien réparti dans l'espace (premier équilibrage),
- l'échantillon est tiré selon la méthode du Cube pour être équilibré sur p variables de contrôle (second équilibrage).

Cette méthode est notamment utilisée par l'Insee pour le tirage du nouvel échantillon maître NAUTILE (Costa et al., 2018).

Echantillonnage spatial doublement équilibré

Le principe est similaire à celui du pivot local :

1. Une unité i est choisie aléatoirement. Le sous-ensemble des p unités les plus proches de i est utilisé.
2. Calcul du barycentre du nuage de points, et recherche des $p+1$ points les plus proches. L'opération est réitérée tant que la somme des carrés des distances au barycentre diminue.
3. L'étape de base de la méthode du cube est appliquée aux $p+1$ points retenus, en équilibrant sur $\mathbf{x}_k$.
4. Si toutes les probabilités des unités sont arrondies aléatoirement à 0 ou 1, l'algorithme s'arrête. Sinon, retour à l'étape 1.

Sampling in a continuous population

Notations

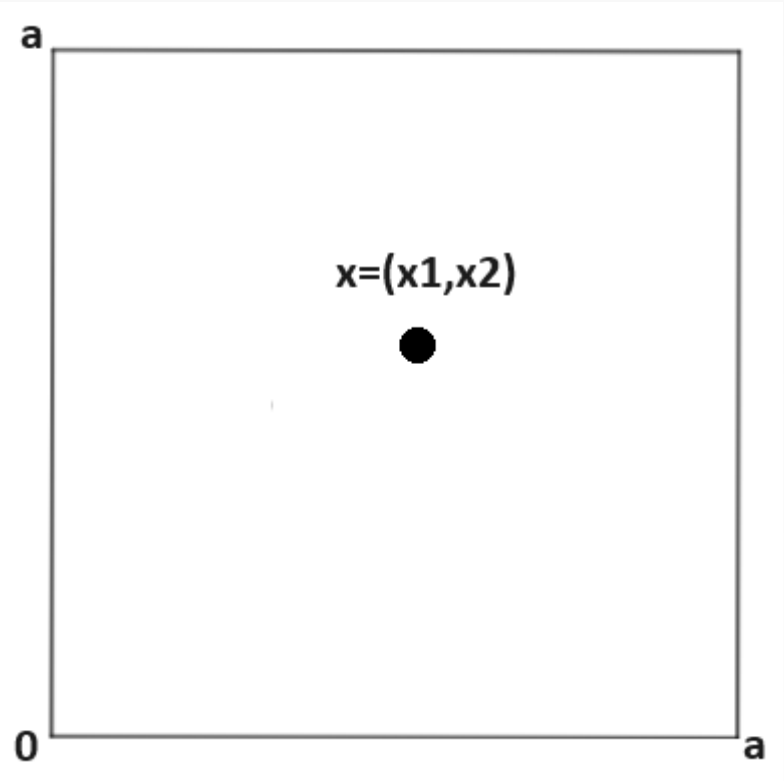
Suppose that we are no more interested in a finite population U, but in a continuous territory $\mathcal{U}^A \subset \mathbb{R}^2$. The *area* of the territory is denoted as A.

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 $f_i(\cdot)$:

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

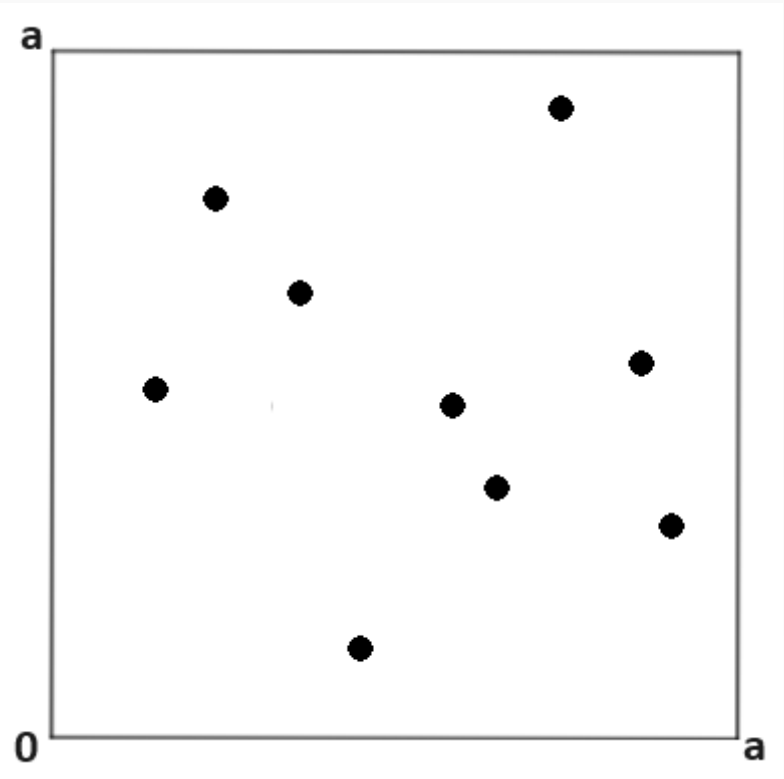
The inclusion density function is defined as

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

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

$$f(x_1, \ldots, x_n) = \frac{1}{A^n} \prod_{i=1}^{n} 1\left(x_i \in \mathcal{U}^A\right),$$

$$f_i(x_i) = \frac{1}{A} 1\left(x_i \in \mathcal{U}^A\right),$$

$$\begin{aligned} \pi(x) &= \sum_{i=1}^{n} f_i(x) \\ &= \frac{n}{A} \text{ for } x \in \mathcal{U}^A, \end{aligned}$$

$$\int_{\mathcal{U}^D} \pi(x) dx = n \, \frac{A^D}{A}.$$

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

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

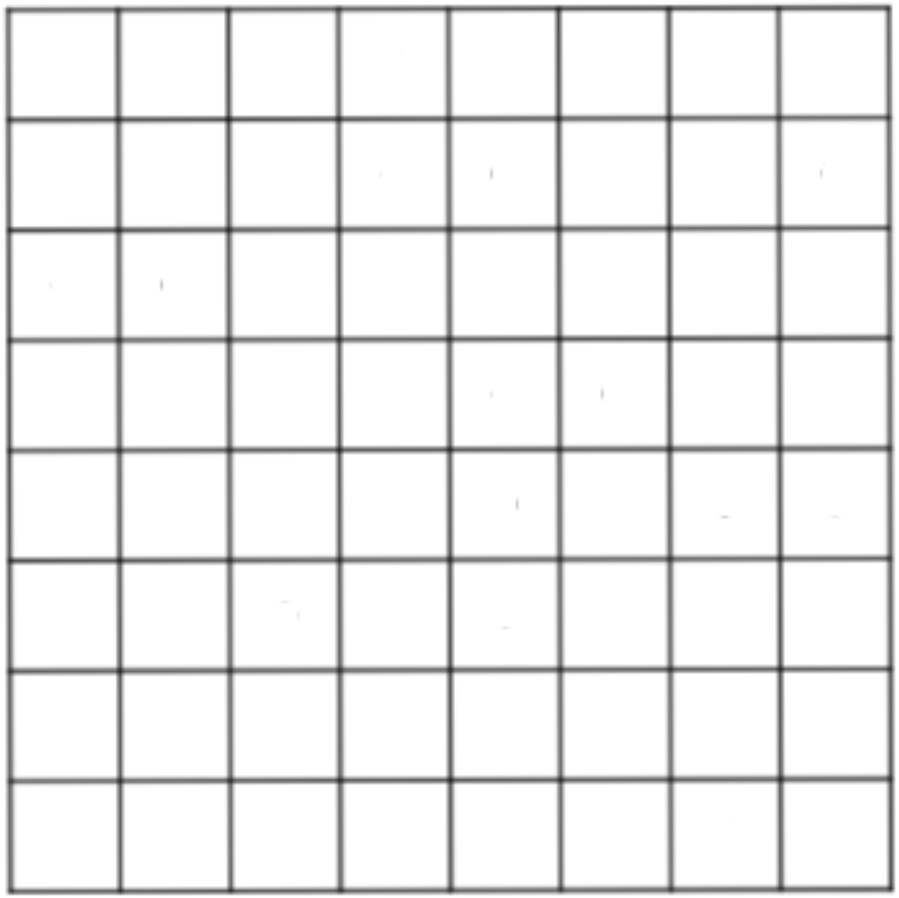
The Horvitz-Thompson estimator is

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

An unbiased variance estimator is

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

Grid sampling



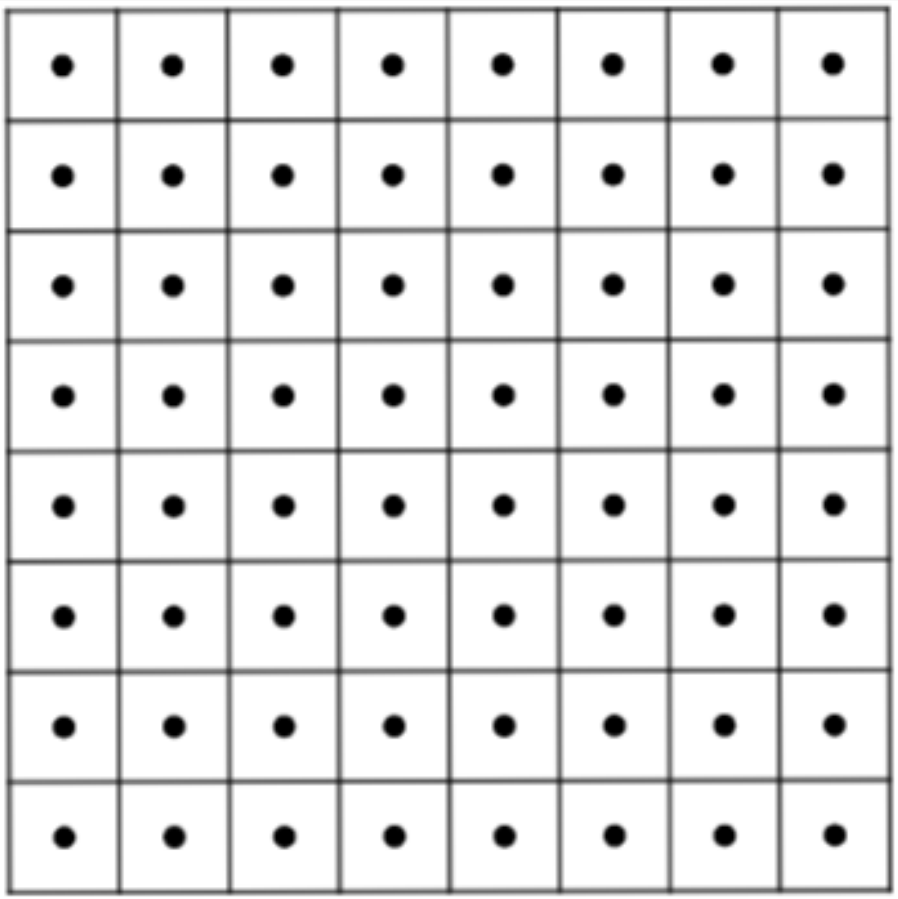
In practice, uniform sampling is hardly ever used. Some areas may be covered by several points, while others may not be surveyed.
$\Rightarrow$ poor spatial balance.

Grid sampling is very common in forest inventories.

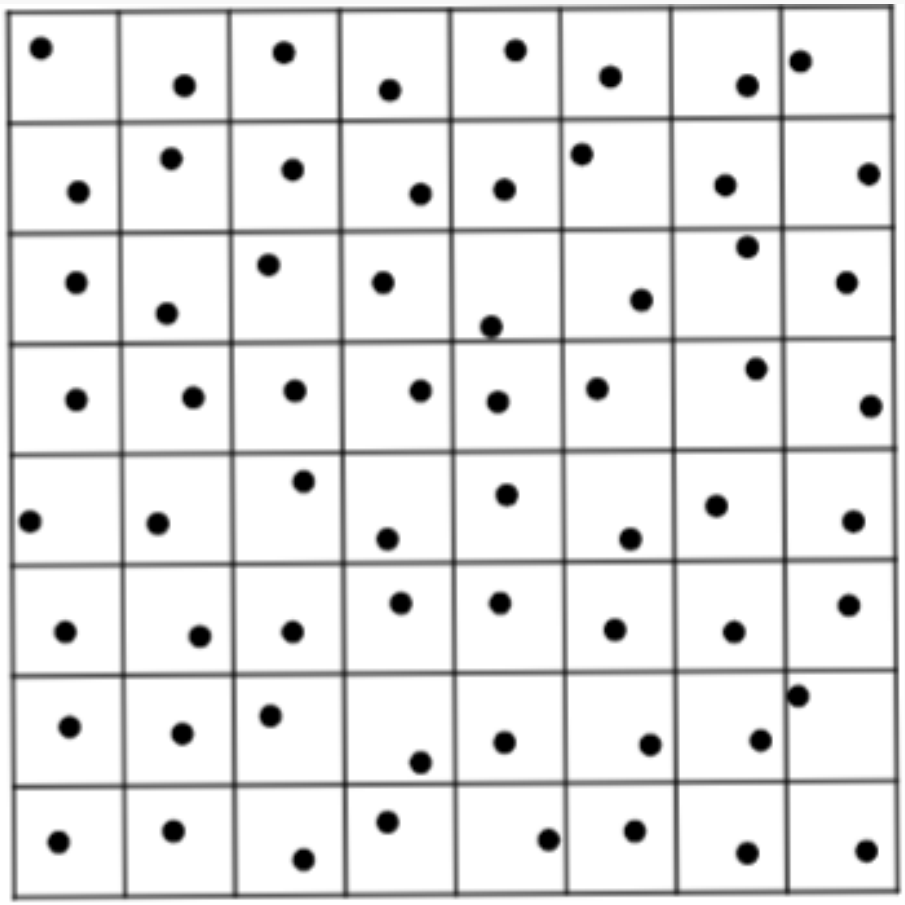
A sample of cells is selected (possibly all), and a sample of points is selected inside each selected cell (usually one).

Some possible grid sampling designs

Spatially systematic aligned sample



Spatially systematic unaligned sample



Some possible grid sampling designs (2)

Like with uniform sampling, the two previous sampling designs lead to a constant inclusion density function:

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

The Horvitz-Thompson estimator is

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

$$\text{with } \bar{\rho} = \frac{1}{n} \sum_{i=1}^{n} \rho(s_i) \text{ the sample mean.}$$

Unbiased variance estimation is not possible.
It is common practice to treat this design as uniform sampling for variance estimation, which usually results in an overestimation of the variance (conservative approach).

Forest inventories

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, ...)

Objectives

We are interested in a finite population U of trees. Let y_k denote the attribute of interest for $k \in U$. We wish to estimate the total

$$t_y = \sum_{k \in U} y_k, \tag{12}$$

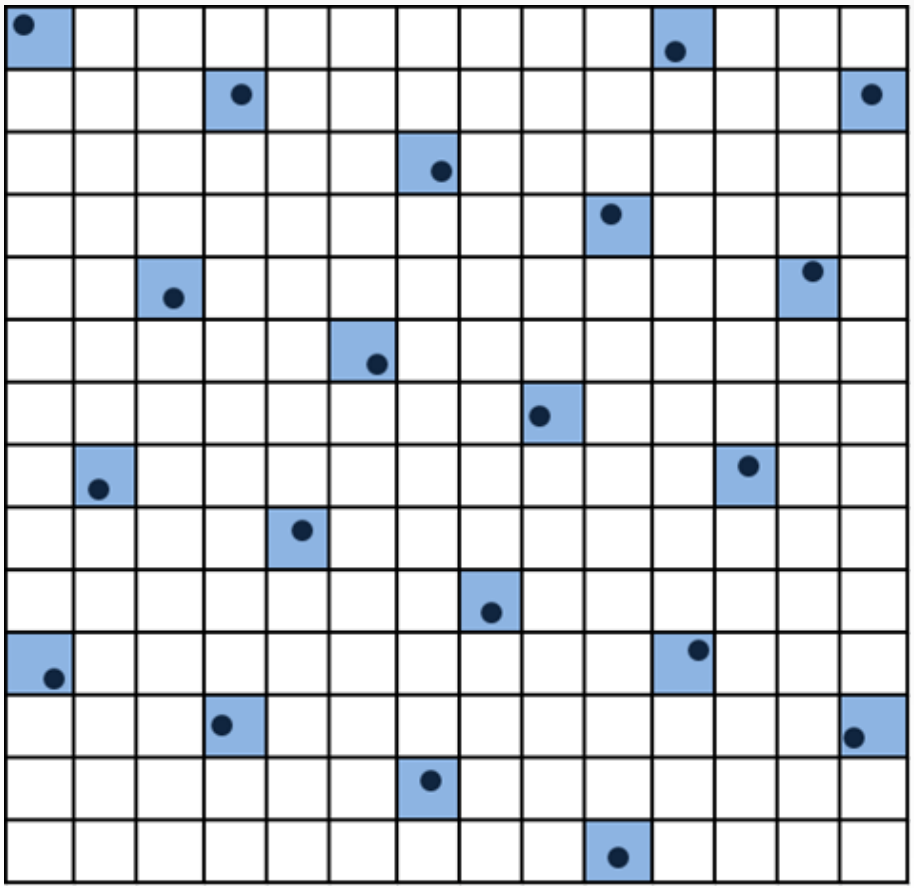
which may be the total volume of wood, for example.

Some form of indirect sampling is used. Let $\mathcal{U}^A$ denote a continuous territory containing all the units in U. 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 (finite sampling design),
4. using fixed-shape supports from these points to survey the units in U.

Step 1: 1st-phase sampling

French annual sample: a two-stage design

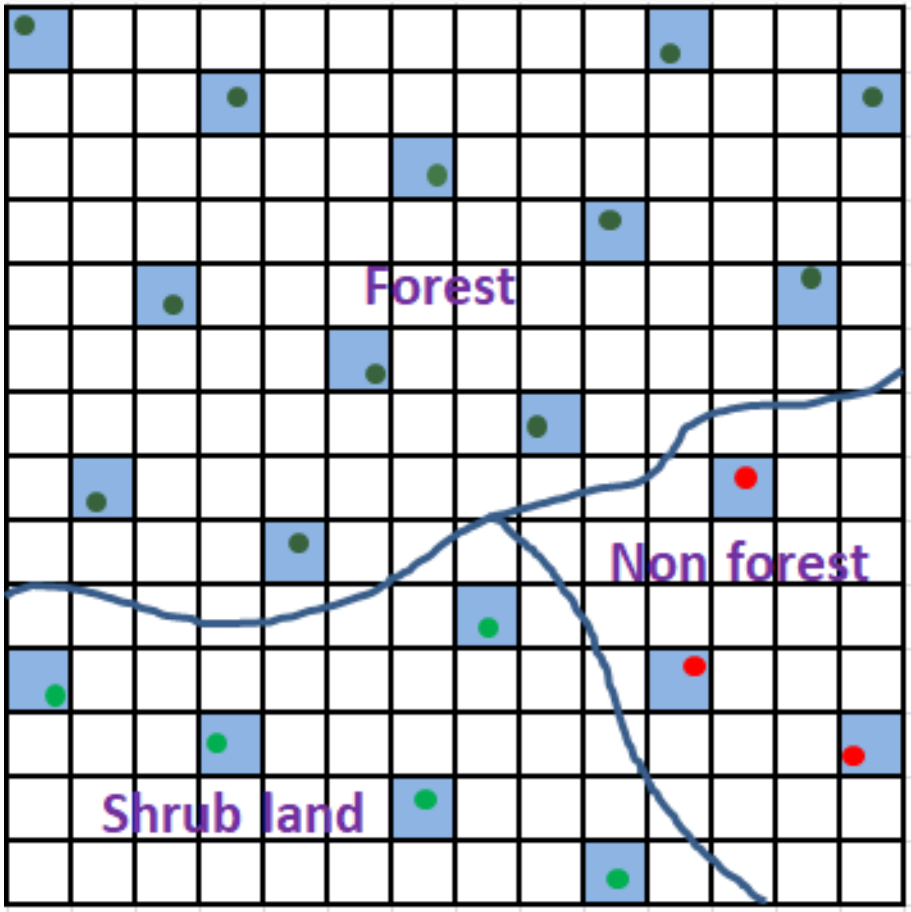


A sample of cells is first selected, by using some form of *systematic sampling* with equal probabilities.

One point is randomly selected inside each cell.
$\Rightarrow$ First-phase sample of points 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

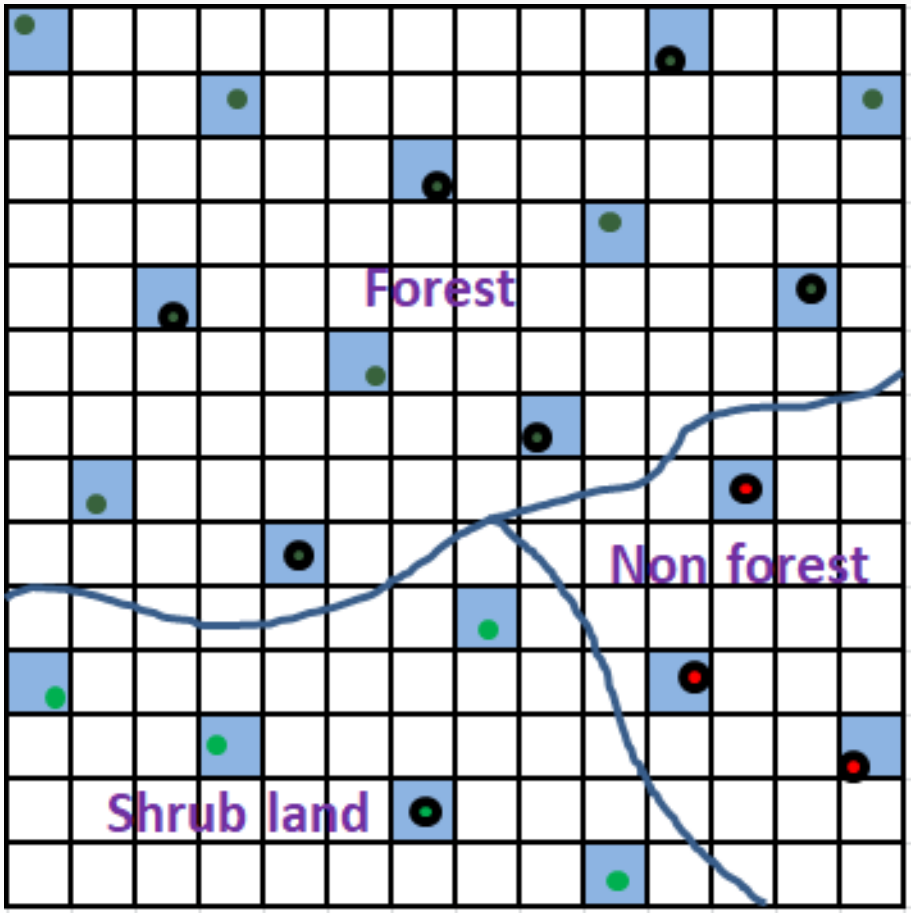


The 1st phase points are classified according to the land cover (forest, shrub land, non forest), using photo-interpretation.

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,
$f_{2g} = 1$ for non-forest,
(no visit on the field in this last case).

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

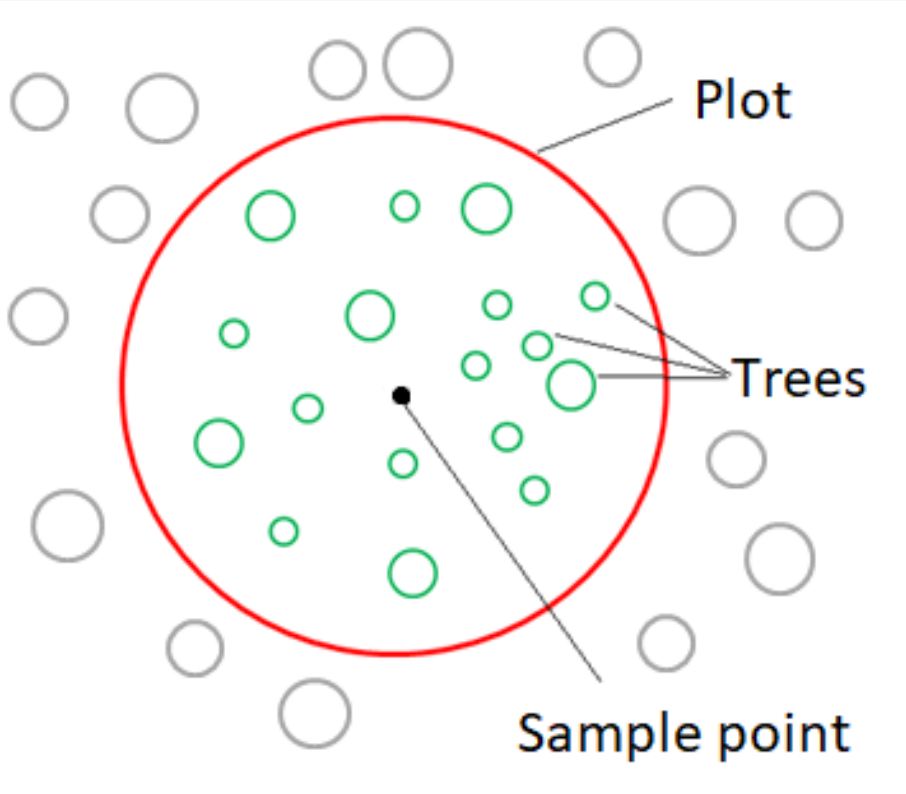


The 1st phase points are classified according to the land cover (forest, shrub land, non forest), using photo-interpretation.

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

Step 4: use of fixed shape supports

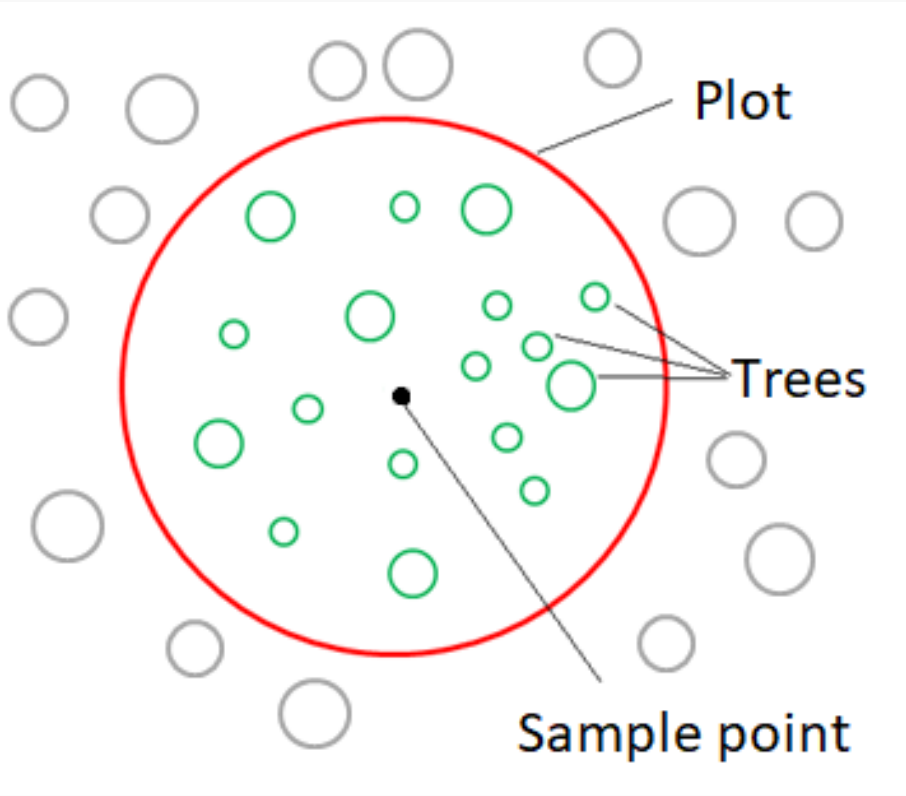


A plot with fixed radius r is centered at the sampled point, and the trees within are surveyed.

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)

Step 4: use of fixed shape supports

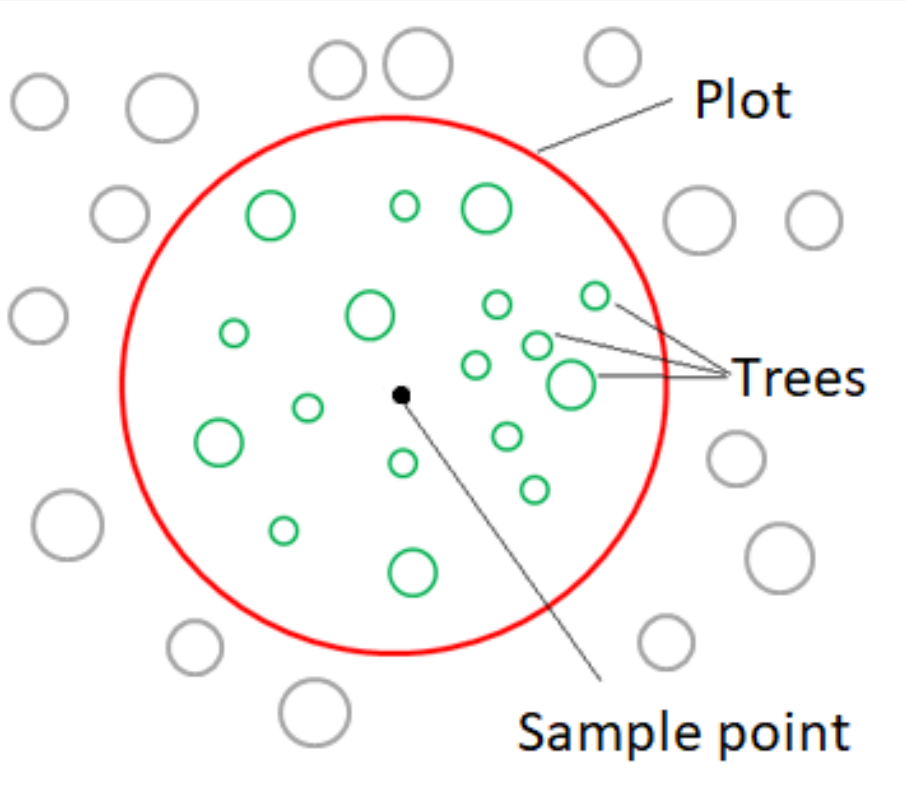


Remark: there can be a third phase of sampling (not covered here).

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



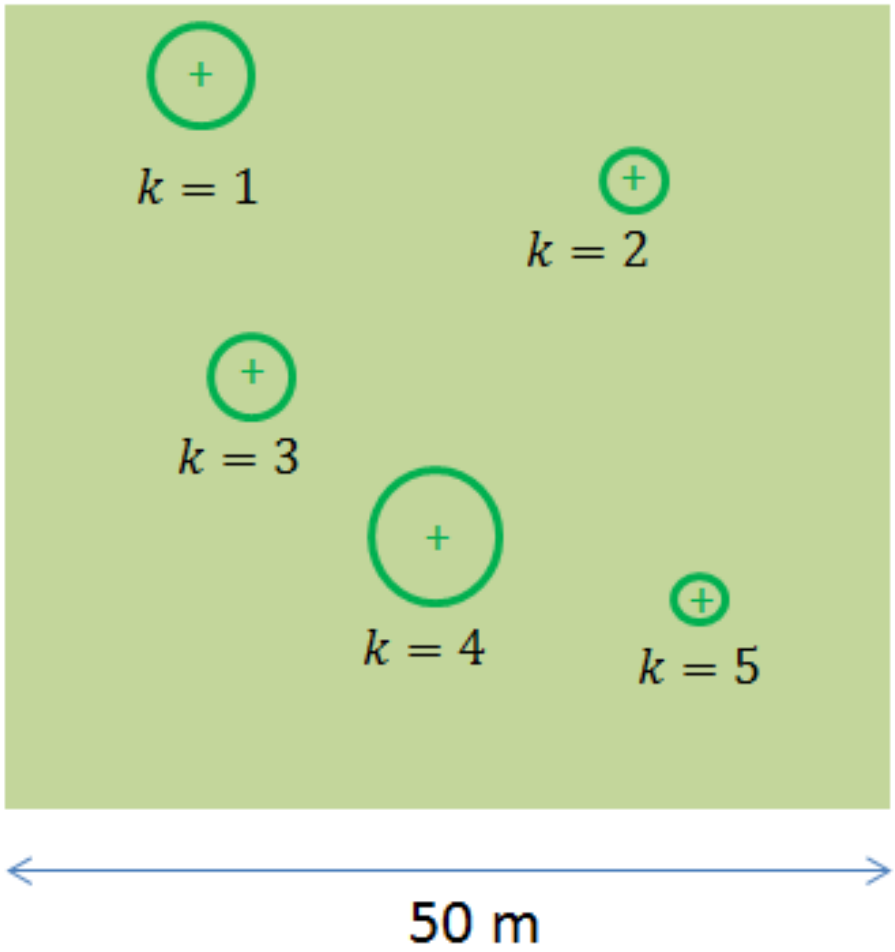
The trees within the plot(s) are surveyed, if they belong to the corresponding circonference class.

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?

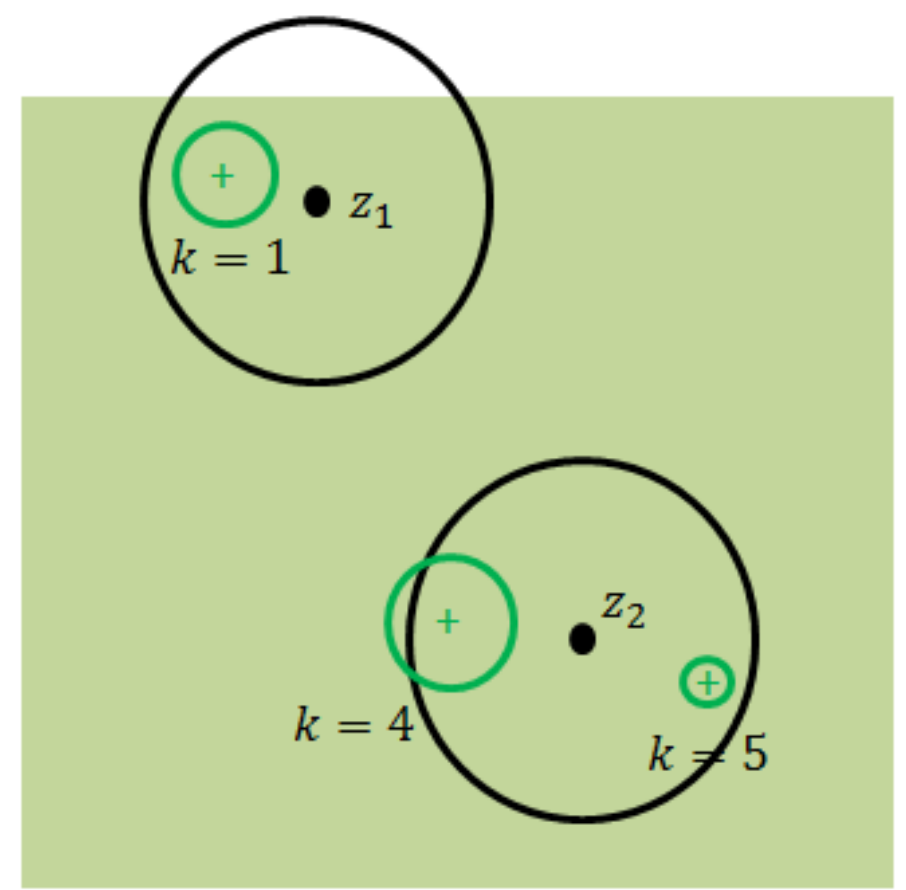
Weight share method



The weight share method (Deville and Lavallée, 2006; Chauvet et al., 2023) enables to use the weights of the sampled points to give estimation weights to the sampled trees.

We illustrate the principle on a toy example. Suppose that the population of interest if a square of length 50 m ($A = 2,500$ m^2) containing only 5 trees.

Weight share method (2)



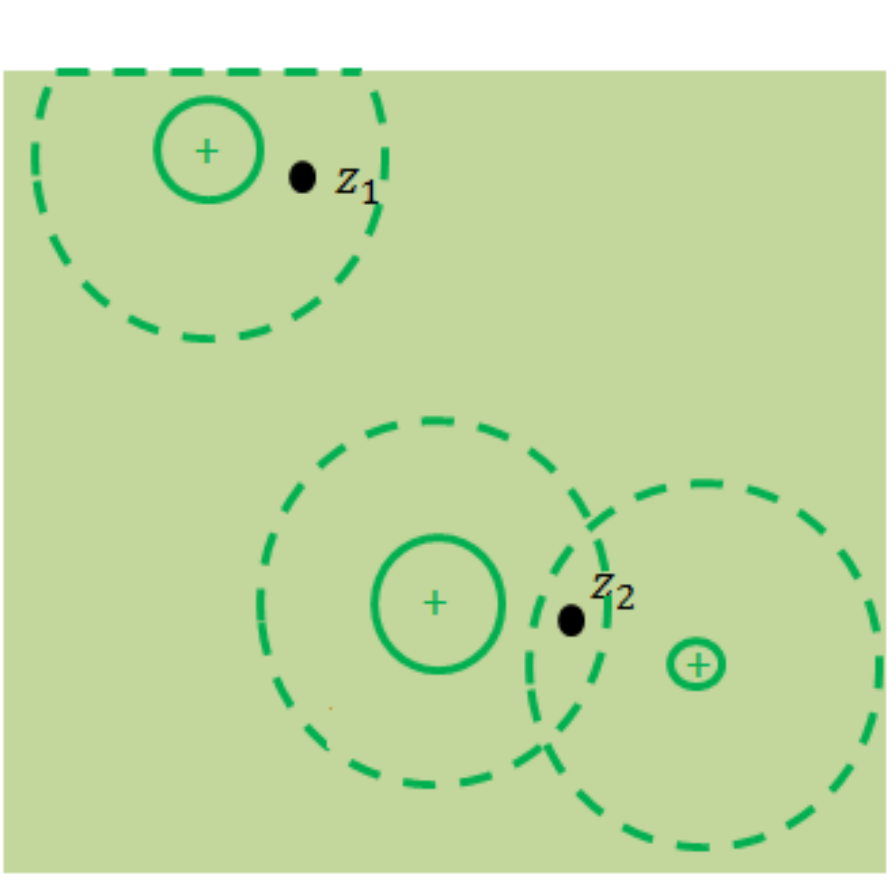
A sample $S^A = \{z_1, z_2\}$ of 2 points is selected in $\mathcal{U}^A$ with a constant inclusion density.

We have

$$d(z_1) = d(z_2) \quad = \quad \frac{A}{n} = 1,250.$$

All the trees in U inside the plots of radius r centered on z_1, z_2 are surveyed (namely, $k = 1, 4$ or 5).

Weight share method (3)



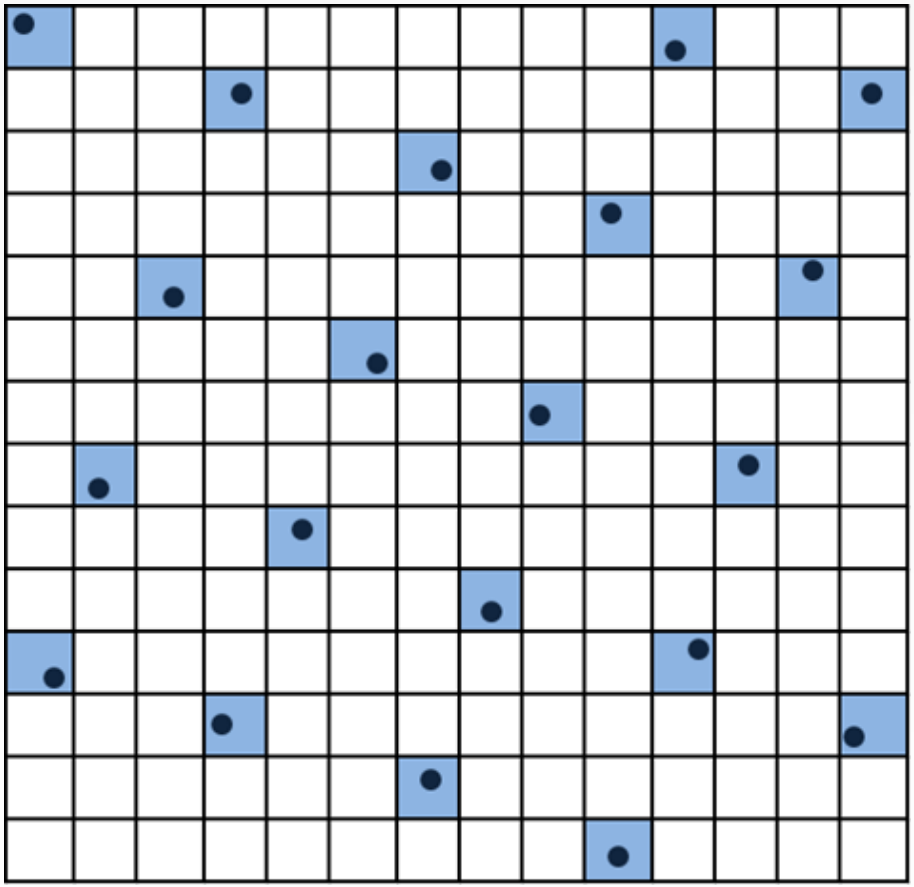
The *inclusion area* of a tree k is the sub-territory which leads to the selection of k is a point is sampled inside.

The weight of the trees are given by:

$$
\begin{aligned}
d_k &= \frac{\sum_{i=1}^{n} d(z_i) 1\{z_i \in \text{i.a. of } k\}}{\text{Surface of the i.a. of } k} \\
&= \begin{cases} \frac{1{,}250}{211.27} = 5.92 & \text{for } k = 1, \\ \frac{1{,}250}{314.16} = 3.98 & \text{for } k = 4, \\ \frac{1{,}250}{314.16} = 3.98 & \text{for } k = 5. \end{cases}
\end{aligned}
$$

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.

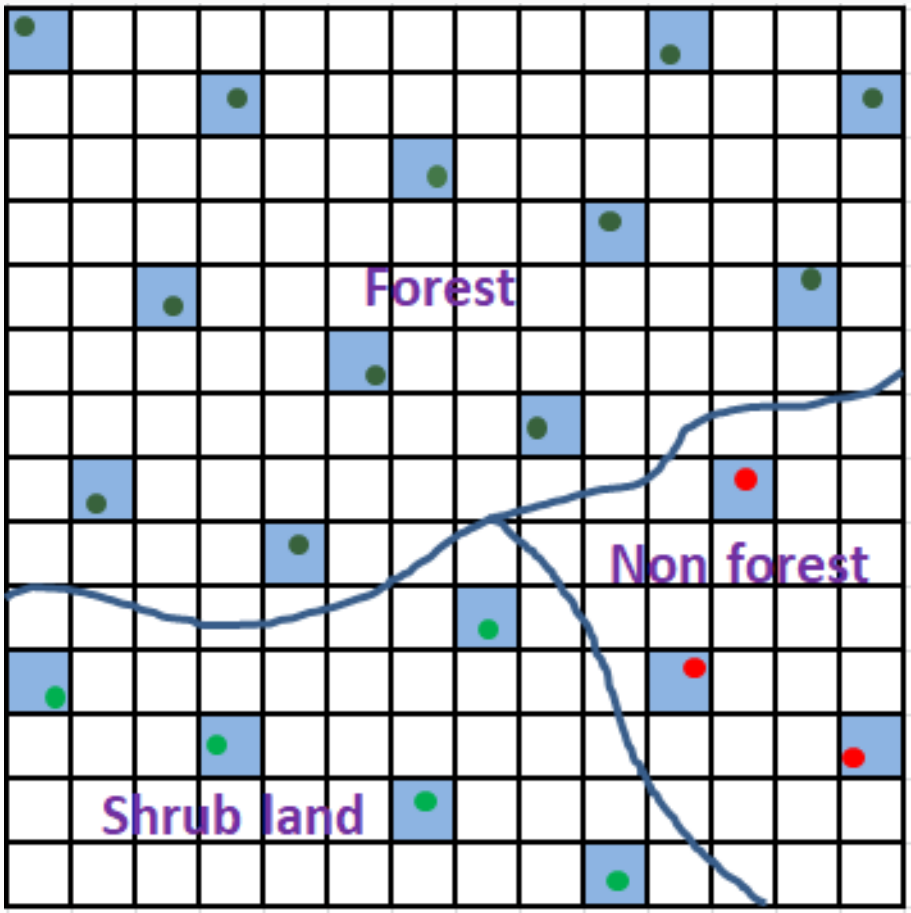
The first-phase inclusion density/HT estimator are

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

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

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

French annual sample: second-phase sampling design

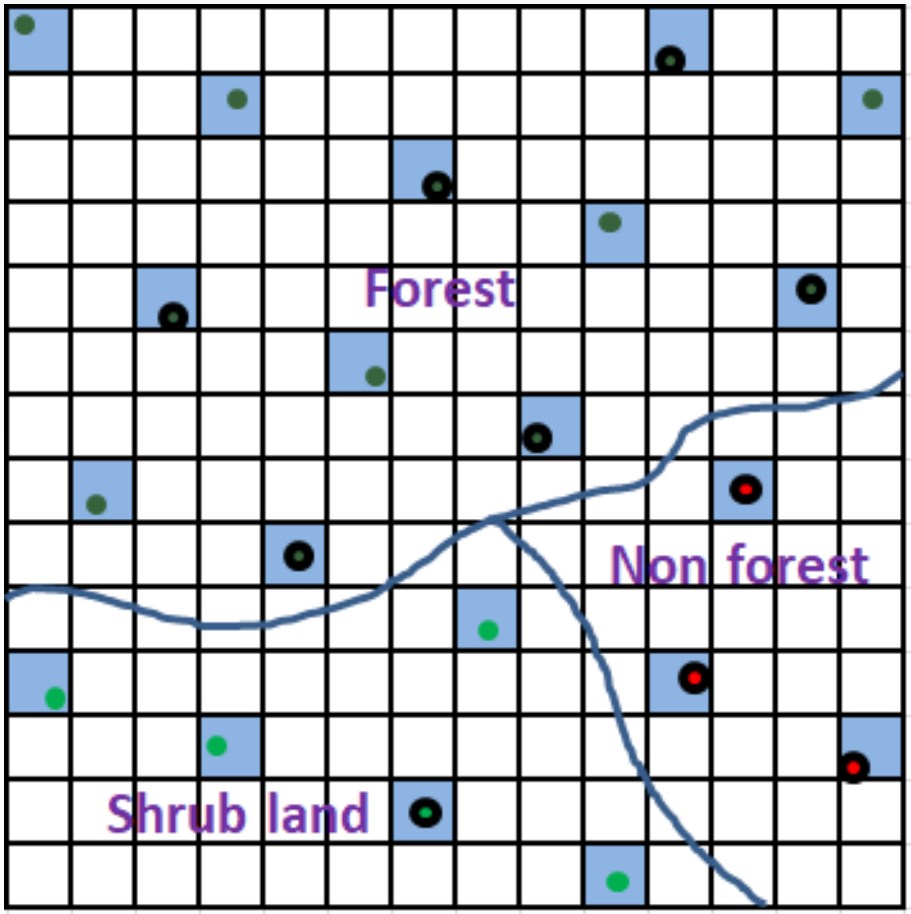


The 1st phase points are classified according to the land cover (e.g., forest, shrub land, non forest). Sub-sampling fraction f_{2g} in the category g.

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.

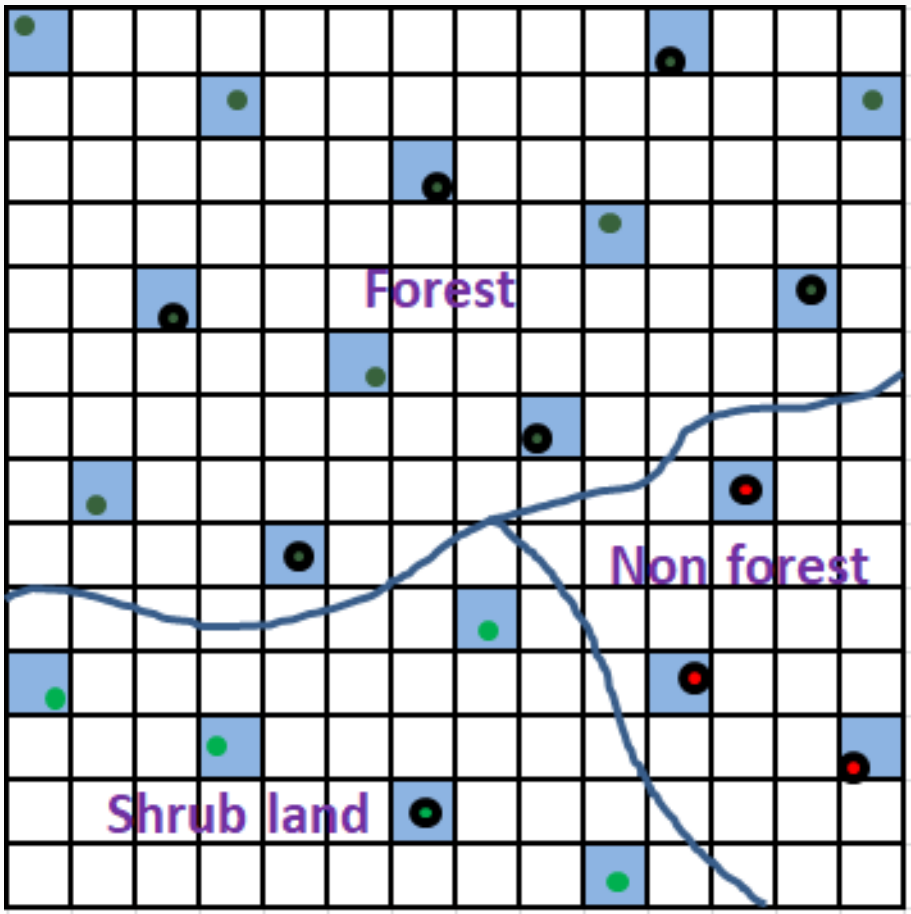
Second-phase inclusion density:

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

Expansion estimator:

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

French annual sample: second-phase sampling design



The estimator

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

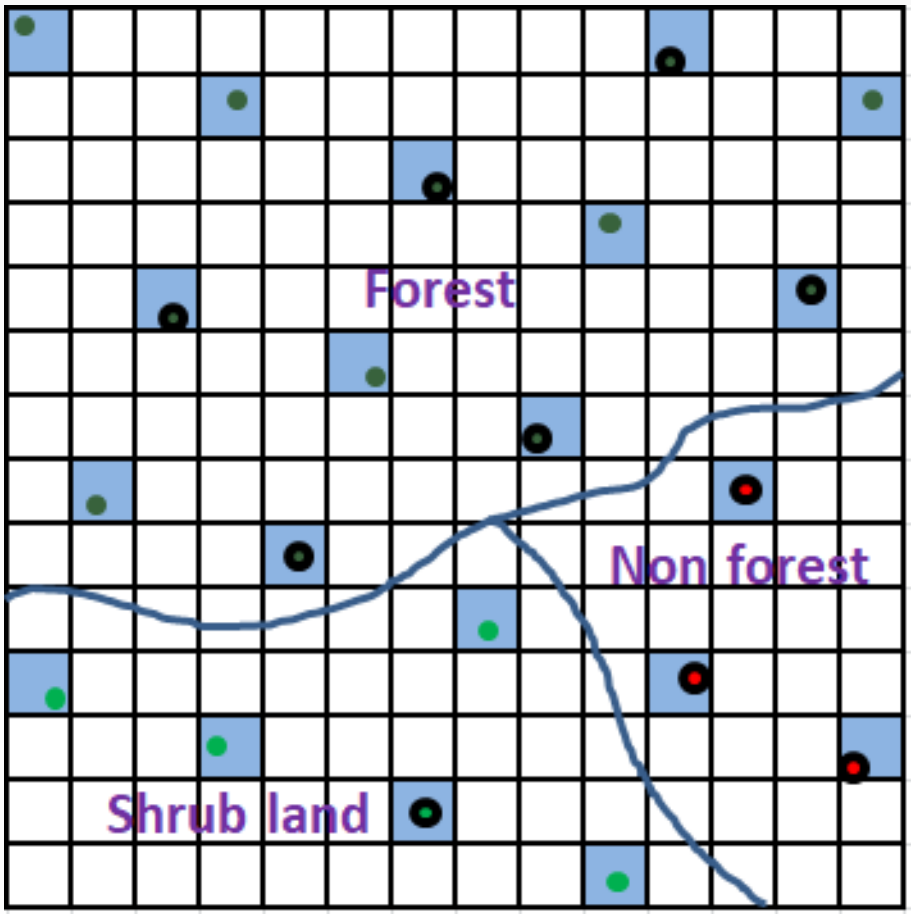
is post-stratified using 1st phase information:

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

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, 202X).

Bibliographie (1)

Echantillonnage en population finie

- Ardilly, P. (2006), *Les techniques de Sondage*, Technip.
- Breidt, F.J., Chauvet, G. (2011), *Improved variance estimation for balanced samples drawn via the Cube method*, JSPI, 141.
- Chauvet, G. (2012). *On a characterization of ordered pivotal sampling*. Bernoulli, 18(4), pp. 1320-1340.
- Chauvet, G., Le Gleut, R. (2019). *Inference under pivotal sampling: properties, variance estimation and application to tesselation for spatial sampling*. SJS.
- Chauvet, G., Tillé, Y. (2006). *A fast algorithm for balanced sampling*. Comp. Stat., 21, 53-62.
- Costa, L., Guillo, C., Paliod, N., Merly-Alpa, T., Vincent, L., Chevalier, M., Deroyon, T. (2018). *Le tirage coordonné du nouvel Échantillon-Maître Nautile avec l'échantillon de l'enquête Emploi en continu*. JMS.

Bibliographie (2)

Echantillonnage en population finie

- Deville, J-C. , Tillé, Y (2004). *Efficient balanced sampling: the cube method*. Biometrika, 91, 893-912.
- Deville, J-C. , Tillé, Y (2005). *Variance approximation under balanced sampling*. JSPI, 128, 569-591.
- Dickson, M.M., and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. Comp. Stat., 31, 1359-1372.
- Grafström, A., Lundström, N.L.P., and Schelin, L. (2012). *Spatially Balanced Sampling through the Pivotal Method*, Biometrics, 68, 514-520.
- Grafström, A., Tillé (2013). *Doubly balanced spatial sampling with spreading and restitution of auxiliary totals*. Environmetrics, 24(2), 120-131.
- Hajek, J. (1964). *Asymptotic theory of rejective sampling with varying probabilities from a finite population*. AoS, 1491-1523.
- Stevens, D.L., and Olsen, A.R. (2004). *Spatially balanced sampling of natural resources*. Journal of the American Statistical Association, 99(465), 262-278.
- Tillé, Y. (2011), *Sampling Algorithms*, Springer-Verlag.

Bibliographie (3)

Echantillonnage en population continue

- Chauvet, G., Bouriaud, O. and Brion, P. (2023). An extension of the weight share method when using a continuous sampling frame. Surv. Methodol.
- Cordy, C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Stat. Probabil. Lett., 18(5), pp. 353-362.
- Deville, J-C. and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. Surv. Methodol., 32(2), pp. 165-176.
- Duong, K.T., Bouriaud, O., Chauvet. G. (2024). A new sampling framework for spatial surveys with application to the French NFI. In revision.
- 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.
- Mandallaz, D. (2007). Sampling techniques for forest inventories. CRC Press.