Sélectionner dextrement l'échantillon

Yves Tillé
Université de Neuchâtel

Colloque Francophone sur les Sondages

Gatineau, octobre 2016

Table of contents

Échantillonner

Échantillonner

- Échantillonner, c'est sélectionner une partie (un échantillon) pour extrapoler son analyse au tout.
- Il existe plusieurs principes d'extrapolation : basée sur la plan/basée le modèle/assistée par le modèle.
- Et si on essayait d'utiliser tous les principes ensemble.

Plan d'expérience vs plans d'échantillonnage.

Plans

Les plans d'expérience et plans d'échantillonnage n'ont pas les mêmes buts.

- Plan d'expérience : on veut estimer les paramètres d'un modèle.
- Plan de sondage : on veut estimer des caractéristiques de la population.

Donc la structure de la population est capitale.

Inférence basée sur le plan de sondage

Inférence basée sur le plan de sondage

- Population : $U = \{1, \ldots, k, \ldots, N\}$.
- Échantillon $s \subset U$.
  Exemple $U = \{1, 2, 3, 4, 5\}$, échantillon $s = \{2, 3, 5\}$ ou $\mathbf{s} = (0, 1, 1, 0, 1)^\top$.
- Plan de sondage $p(s) \geq 0$ et $\sum_{s \subset U} p(s) = 1$.
- Échantillon aléatoire S, $\Pr(S = s) = p(s)$, for all $s \subset U$.
- Probabilités d'inclusion $\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s)$.
- Probabilités d'inclusion jointe $\pi_{k\ell} = \Pr(\{k, \ell\} \in S) = \sum_{s \supset \{k,\ell\}} p(s)$.
- Total $Y = \sum_{k \in U} y_k$. Moyenne $\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k$.

Inférence basée sur le plan

- Estimateur de Nairin-Horvitz-Thompson (NHT) : $\widehat{\overline{Y}} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}$.
- $\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k\pi_\ell & \text{if } k \neq \ell \\ \pi_k(1-\pi_k) & \text{if } k = \ell. \end{cases}$
- Variance de l'estimateur NHT : $\mathrm{var}_p\left(\widehat{\overline{Y}}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{\ell\in U}\frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell}$.

  $\mathrm{var}_p\left(\widehat{\overline{Y}}\right) = -\frac{1}{2N^2}\sum_{k\in U}\sum_{\substack{\ell\in U\\ k\neq\ell}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2\Delta_{k\ell}$ (taille fixe).
- Estimations

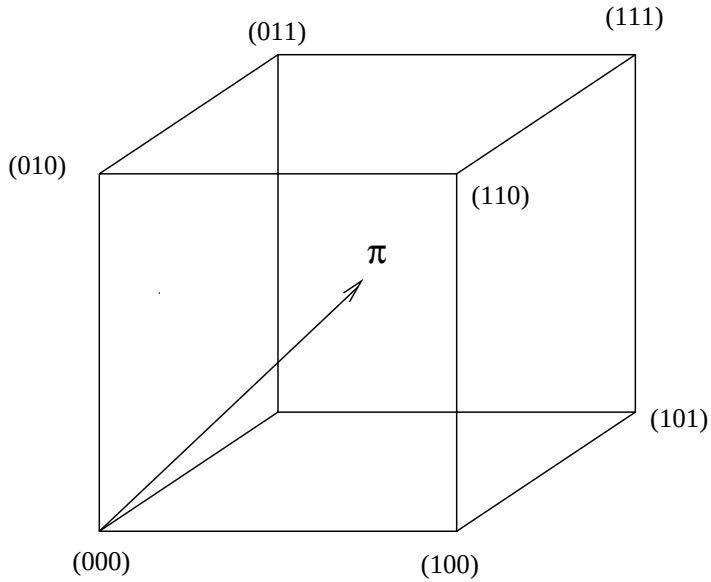
  $\widehat{\mathrm{var}}\left(\widehat{\overline{Y}}\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta_{k\ell}}{\pi_{k\ell}}$,

  $\widehat{\mathrm{var}}\left(\widehat{\overline{Y}}\right) = -\frac{1}{2N^2}\sum_{k\in S}\sum_{\substack{\ell\in S\\ k\neq\ell}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2\frac{\Delta_{k\ell}}{\pi_{k\ell}}$ (taille fixe).

Plan

Plans

- Un plan de sondage n'est rien d'autre qu'un vecteur aléatoire discret positif.
- Chaque composante représente le nombre de fois qu'une unité est sélectionnée dans l'échantillon.
- Un plan sans remise est une loi de probabilités sur les sommets d'un cube.



Plan de Bernoulli

Plan de Bernoulli

- Plans de Bernoulli avec des probabilités π.

$$p(s) = \pi^{\#s}(1-\pi)^{N-\#s} \ , \ \text{for all } s \in \mathcal{S},$$

où n_s est la taille de l'échantillon s.

- L'échantillon $(s_1, \ldots, s_k, \ldots, s_N)$ est un vecteur de variables de Bernoulli i.i.d. de paramètre π.
- La taille de l'échantillon $n_S \sim \mathrm{Bin}(N, \pi)$.

Plans de Poisson

Plan de Poisson

- Plan de Poisson

$$p(s) = \left\{ \prod_{k \in S} \pi_k \right\} \left\{ \prod_{k \notin S} (1 - \pi_k) \right\} \text{ , pour tout } s \in \mathcal{S}.$$

- L'échantillon $(s_1, \ldots, s_k, \ldots, s_N)$ un vecteur de variable i.i.d. de Bernoulli avec les paramètres π_k.
- Probabilités d'inclusion π_k.
- La taille de l'échantillon suit une loi Poisson-binomiale (Hodges Jr. & Le Cam, 1960; Stein, 1990; Chen, 1993).

Plan simple sans remise

Plan simple sans remise

- Plan simple sans remise

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{pour tout } s \text{ de taille } n \\ 0 & \text{sinon} \end{cases}$$

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

et

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)},$$

pour tout $k \neq \ell \in U$.

Plans à probabilités inégales

Plans à probabilités inégales

- Plans à probabilités inégales π_k de taille fixe.
- Plusieurs dizaines de méthodes (voir Brewer & Hanif, 1983; Tillé, 2006).
- Plan de Poisson conditionnel (CPS) ou plan à entropie maximale (MES) :

$$p(s) = \frac{\sum_{k \in S} \exp \lambda_k}{\sum_{s \in \mathcal{S}_n} \sum_{k \in S} \exp \lambda_k},$$

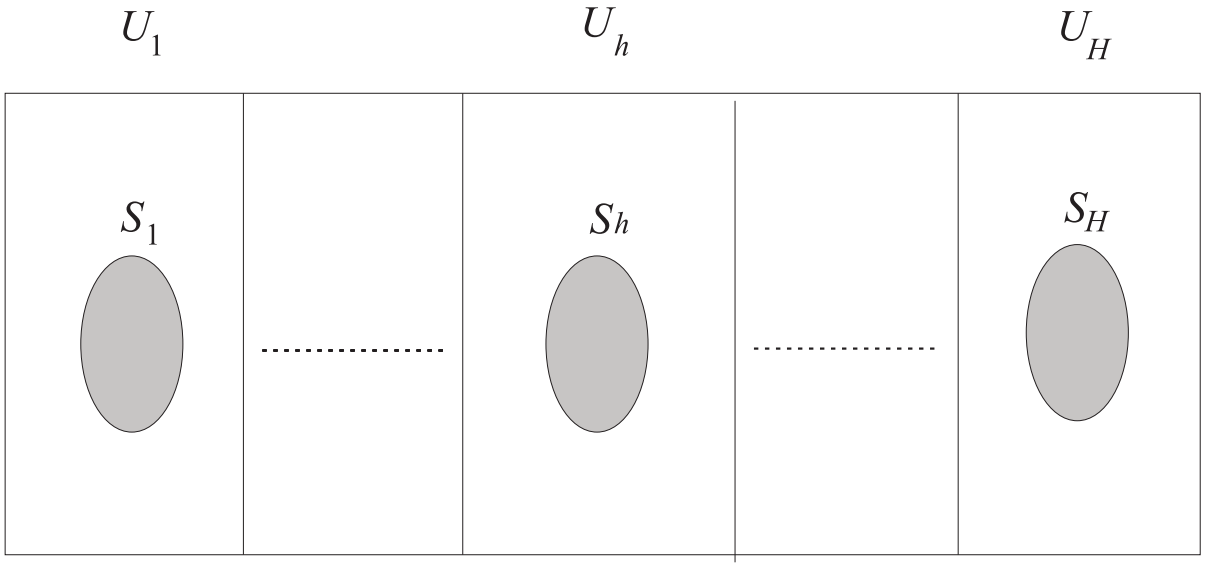
où $\mathcal{S}_n$ est l'ensemble des échantillons de taille fixe n et les λ_k sont obtenus en résolvant

$$\sum_{s \in \{s \in \mathcal{S}_n | s \ni k\}} p(s) = \pi_k. \tag{1}$$

- Implémentation complexe (Chen, Dempster & Liu, 1994; Deville, 2000; Tillé & Matei, 2015).
- L'échantillon est un vecteur aléatoire qui appartient à la famille exponentielle. Il existe une bijection entre les espérances π_k et les paramètres λ_k.

Stratification

Stratification



Stratification avec H strates $U_1, \ldots, U_H$.

$$p(s) = \begin{cases} \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} & \text{for all } s \text{ tel que } \#(U_h \cap s) = n_h. \\ 0 & \text{sinon.} \end{cases}$$

Stratification

Stratification (suite)

- Les probabilités d'inclusion $\pi_k = n_h/N_h$.
- Tous les échantillons de probabilités non-nulles ont la même probabilité.
- Deux allocations de base
  - Allocation proportionnelle $n_h = \dfrac{nN_h}{N}$.
  - Allocation optimale de Neyman $n_h = \dfrac{nN_hV_h}{\sum_{\ell=1}^{H} N_\ell V_\ell}$,

où V_h est l'écart-type de la variable d'intérêt dans la strate h.

Plans équilibrés

Bamboleo

Plans équilibrés

- 'Procédure rejective' : On génère des échantillons jusqu'à l'obtention d'un échantillon équilibré.

$$p\left(s \left| \sum_{j=1}^{p} \left| \frac{\widehat{X}_j - X_j}{X_j} \right| < c \right.\right),$$

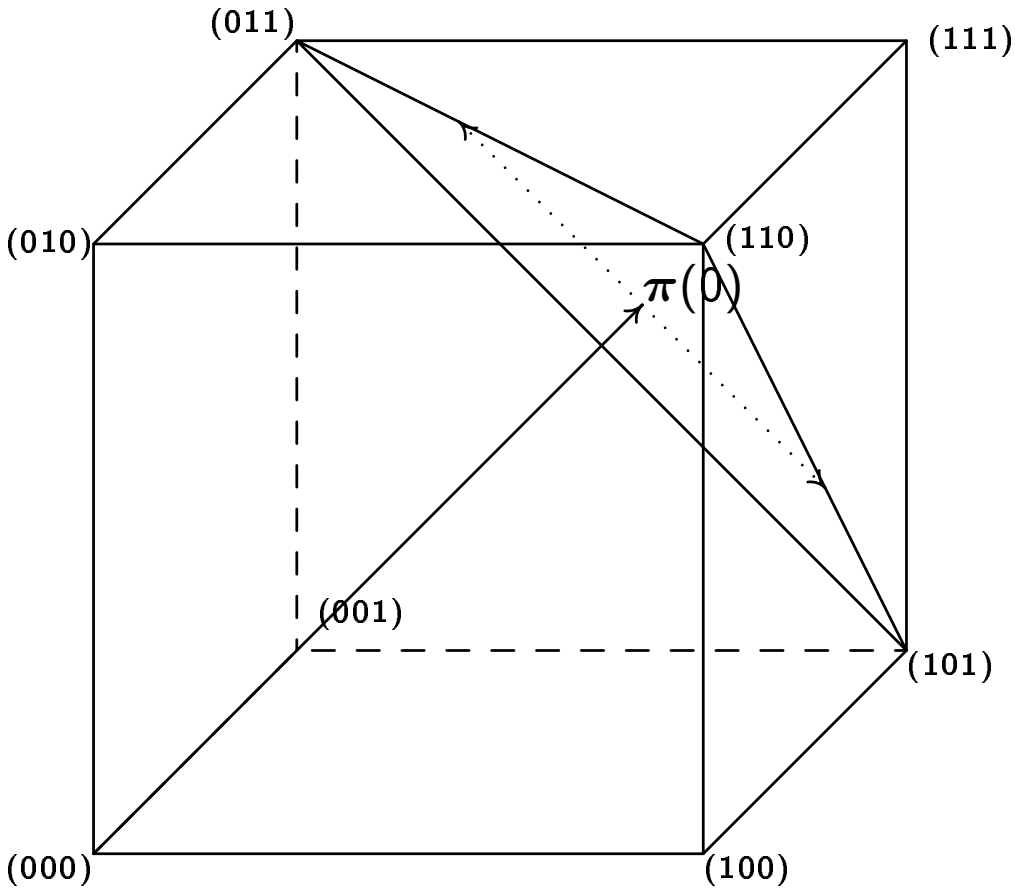
où

$$X_j = \sum_{k \in U} x_{kj} \text{ et } \widehat{X}_j = \sum_{k \in S} \frac{x_{kj}}{\pi_k},$$

et c est une valeur petite. (Hájek, 1981; Legg & Yu, 2010).

- Problème : les unités extrêmes ont des probabilités d'inclusion réduites. (Fuller, 2009; Chauvet, Haziza & Lesage, 2015)

Idée de la méthode du cube



Plan équilibré

- Méthode du cube. Sélectionne un plan équilibré tel que

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k.$$

  (Deville & Tillé, 2004; Tillé & Favre, 2004; Deville & Tillé, 2005; Tillé & Favre, 2005; Chauvet & Tillé, 2006; Chauvet, Bonnéry & Deville, 2011; Tillé, 2011; Breidt & Chauvet, 2011).
- Les probabilités d'inclusion sont exactement satisfaites.
- Les contraintes d'équilibrage ne peuvent en général pas être exactement satisfaites.
- Deux phases : phase de vol et phase d'atterrissage.
- Implémentation en langage R (Tillé & Matei, 2015; Grafström & Lisic, 2016) et en SAS (Rousseau & Tardieu, 2004; Chauvet & Tillé, 2005).

Inférence basée sur un modèle Brewer (1963); Royall (1970a,b, 1971)

Inférence basée sur un modèle

- Modèle $y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$ avec $\mathrm{var}(\varepsilon_k) = \sigma_k^2$ et $\mathrm{cov}(\varepsilon_k, \varepsilon_\ell) = 0$.
- Estimateur BLU (sous le modèle) minimise $\mathrm{E}(\widehat{Y}_{BLU} - Y)^2$
- $\widehat{Y}_{BLU} = \sum_{k \in S} y_k + \sum_{k \notin S} \widehat{y}_k$.
- $\widehat{y}_k = \mathbf{x}_k \widehat{\boldsymbol{\beta}}$, où $\widehat{\boldsymbol{\beta}} = \left(\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k^\top}{\sigma_k^2}$.
- $\mathrm{E}_M(\widehat{Y}_{BLU}) = Y$ (sans biais sous le modèle).
- $\mathrm{var}_M(\widehat{Y}_{BLU} - Y) = \sum_{k \notin S} \mathbf{x}_k^\top \left(\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \notin S} \mathbf{x}_k + \sum_{k \notin S} \sigma_k^2$.

Cas particulier

Inférence basée sur le modèle

- $y_k = \beta + \varepsilon_k$ avec $\mathrm{var}(\varepsilon_k) = \sigma^2$ et $\mathrm{cov}(\varepsilon_k, \varepsilon_\ell) = 0$.
- estimateur BLU $\widehat{Y}_{BLU} = \sum_{k \in S} y_k + \sum_{k \notin S} \bar{y} = N\bar{y}$, où $\bar{y} = \frac{1}{n} \sum_{k \in S} y_k$.
- $\mathrm{var}_M(\bar{y}) = \frac{N-n}{Nn} \sigma^2$

Inférence basée sur le plan

- Plan simple : estimateur HT $\widehat{Y}_\pi = N\bar{y}$.
- $\mathrm{var}_p(\bar{y}) = \frac{N-n}{Nn} S_y^2$.
- $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y})^2$, $\mathrm{E}_M(S_y^2) = \sigma^2$.

Inférence basée sur le modèle ou sur le plan

Même inférence

- Les deux approches peuvent conduire à la même inférence.
- Correspondance plan/modèle ?
- Exemple : Stratification avec un modèle d'analyse de la variance.
- Peut-on spécifier quand les deux approches coïncident ?

Coïncidence des approches

Définition

Hétéroscédasticité pleinement explicable

(i) il existe un vecteur $\boldsymbol{\lambda} \in \mathbb{R}^q$ tel que $\boldsymbol{\lambda}^\top \mathbf{x}_k = \sigma_k^2$;

(ii) il existe un vecteur $\boldsymbol{\theta} \in \mathbb{R}^q$ tel que $\boldsymbol{\theta}^\top \mathbf{x}_k = \sigma_k$.

Réinterprétation par Nedyalkova & Tillé (2008) d'un résultat de Royall (1992) et de Valliant, Dorfman & Royall (2000, pp. 98-100).

Résultat

Si le modèle de superpopulation a une hétéroscédasticité pleinement explicable et si le plan est équilibré avec des probabilités d'inclusion proportionnelles aux σ_k, alors l'estimateur BLU $\widehat{Y}_{\mathrm{BLU}}$ sous le modèle est égal à l'estimateur HT $\widehat{Y}_\pi$.

Coïncidence des approches

Coïncidence des approches

- Faire coïncider les approches rend l'estimation doublement robuste.
- Il faut modéliser la population et ensuite concevoir le plan en fonction de ce modèle.
- La variable d'hétéroscédasticité doit faire partie du modèle.

Principes d'échantillonnage

- Randomisation,
- Restriction,
- Surreprésentation.

Randomisation

1er principe : Randomisation

- Sélection de l'échantillon de la manière la plus aléatoire
- Maximisation de l'entropie

$$I(p) = -\sum_{s \subset U} p(s) \log p(s).$$

- Malheureusement, ce n'est pas toujours possible.
- Important pour l'estimation de variance (Berger, 1996, 1998a,b,c).

Restriction

2ème principe : Restriction

- Éviter les mauvais échantillons.
- Sélectionner seulement des échantillons avec des caractéristiques particulières.
- Plus généralement, sélection d'échantillon équilibrés

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k.$$

- Cas particuliers : taille fixe, stratification.

Surreprésentation

3ème principe : Surreprésentation

- Surreprésenter là où la dispersion est plus large.
- Exemple : Allocation optimale de Neyman est un cas particulier.

Remarque : le plan optimal n'existe pas

- La variance : forme quadratique $\mathrm{var}(\widehat{Y}) = \check{\mathbf{y}}^\top \boldsymbol{\Delta} \check{\mathbf{y}}$, où

$$\check{\mathbf{y}}^\top = \left(\frac{y_1}{\pi_1} \dots \frac{y_N}{\pi_N} \right).$$

- Avec les même probabilités d'inclusion, si il existe une variable y pour laquelle

$$\mathrm{var}_1(\widehat{Y}) = \check{\mathbf{y}}^\top \boldsymbol{\Delta}_1 \check{\mathbf{y}} < \mathrm{var}_2(\widehat{Y}) = \check{\mathbf{y}}^\top \boldsymbol{\Delta}_2 \check{\mathbf{y}},$$

  il existe une autre variable z telle que

$$\mathrm{var}_1(\widehat{Z}) = \check{\mathbf{z}}^\top \boldsymbol{\Delta}_1 \check{\mathbf{z}} > \mathrm{var}_2(\widehat{Z}) = \check{\mathbf{z}}^\top \boldsymbol{\Delta}_2 \check{\mathbf{z}}$$

  (Tillé, 2004).

- Interprétation : Si deux plans de sondage ont les mêmes probabilités d'inclusion, les sommes des valeurs propres de $\boldsymbol{\Delta}_1$ et $\boldsymbol{\Delta}_2$ sont égales.
- Si la variance est diminuée dans une direction, elle est nécessairement augmentée dans une autre direction.

Remarque sur l'optimalité

Comparaison plan de Bernoulli avec plan simple $\pi = n/N$ (même probabilités d'inclusion)

Plan de Bernoulli

- $\mathbf{\Delta} = \pi(1-\pi)\mathbf{I}_N$, valeurs propres $\lambda_i = \pi(1-\pi), i = 1, ..., N$
- $\text{var}_{BERN}(\widehat{\overline{Y}}) = \dfrac{1-\pi}{\pi N^2} \sum_{k \in U} y_k^2$

Plan simple

- $N-1$ valeurs propres égales à $\frac{N-n}{N-1}$ et une égale à 0 associée au vecteur propre $(1, \ldots, 1)^\top$.
- $\text{var}_{SRS}(\widehat{\overline{Y}}) = \dfrac{N-n}{nN} \dfrac{1}{N-1} \sum_{k \in U} (y_k - \overline{Y})^2 = \dfrac{1-\pi}{\pi N(N-1)} \sum_{k \in U} (y_k - \overline{Y})^2$
- Le plan simple a une variance nulle pour les variables constantes $y_k = C$. Le plan de Bernoulli est plus précis pour toutes les variables centrées.

Modélisation de la variable d'intérêt

- Modèle sur la population $y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$,
  $\mathbf{x}_k = (x_{k1}, \ldots, \ldots, x_{kp})^\top$, $\mathrm{var}_M(\varepsilon_k) = \sigma^2_{\varepsilon k}$. Pas d'autocorrélation.
- La variance anticipée de l'estimateur HT

$$\mathrm{AVar}(\widehat{Y}) = \mathrm{E}_p\mathrm{E}_M(\widehat{Y}-Y)^2 = \mathrm{E}_p\left(\sum_{k\in S}\frac{\mathbf{x}_k^\top\boldsymbol{\beta}}{\pi_k} - \sum_{k\in U}\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 + \sum_{k\in U}\pi_k(1-\pi_k)\frac{\sigma_k^2}{\pi_k^2}.$$

  Le plan de sondage qui minimise la variance anticipée consiste à
  - utiliser des probabilités d'inclusion proportionnelle aux $\sigma_{\varepsilon k}$,
  - équilibrer sur les auxiliaires $\mathbf{x}_k$.
- Sous ces contraintes, on peut maximiser l'entropie.

Modélisation de la variable d'intérêt

Plans optimaux qui maximisent l'entropie :

Plan	Modèle	Variance du modèle	π_k
Plan simple	$y_k = \beta + \varepsilon_k$	σ^2	n/N
Plan de Bernoulli	$y_k = \varepsilon_k$	σ^2	$\pi = \mathrm{E}(n_S)/N$
CPS	$y_k = x_k\beta + \varepsilon_k$	$x_k^2\sigma^2$	$\pi_k \propto x_k$
Plan de Poisson	$y_k = \varepsilon_k$	$x_k^2\sigma^2$	$\pi_k \propto x_k$
Stratification proportionnelle	$y_k = \beta_h + \varepsilon_k, k \in U_h,$	σ^2	n/N
Stratification optimale	$y_k = \beta_h + \varepsilon_k, k \in U_h,$	σ_h^2	$\pi_k \propto \sigma_h$

Modélisation de la variable d'intérêt

Cas général

- Cas général $y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$,
- Il n'existe pas de plans optimaux qui maximisent l'entropie en toute généralité.
- On ne peut pas maximiser l'entropie pour un plan équilibré sauf dans des cas particuliers.
- Deville "Conditions nécessaire et suffisante d'équilibrage exact : Toutes les $p \times p$ matrices de plein rang extraites de $\mathbf{A} = (x_{kj}/\pi_k)$ ont le même déterminant en valeur absolue.

Tirage systématique à probabilités inégales

Exemple

On suppose $N = 6$ et $n = 3$.

k	0	1	2	3	4	5	6	Total
π_k		0.07	0.17	0.41	0.61	0.83	0.91	3
V_k	0	0.07	0.24	0.65	1.26	2.09	3	

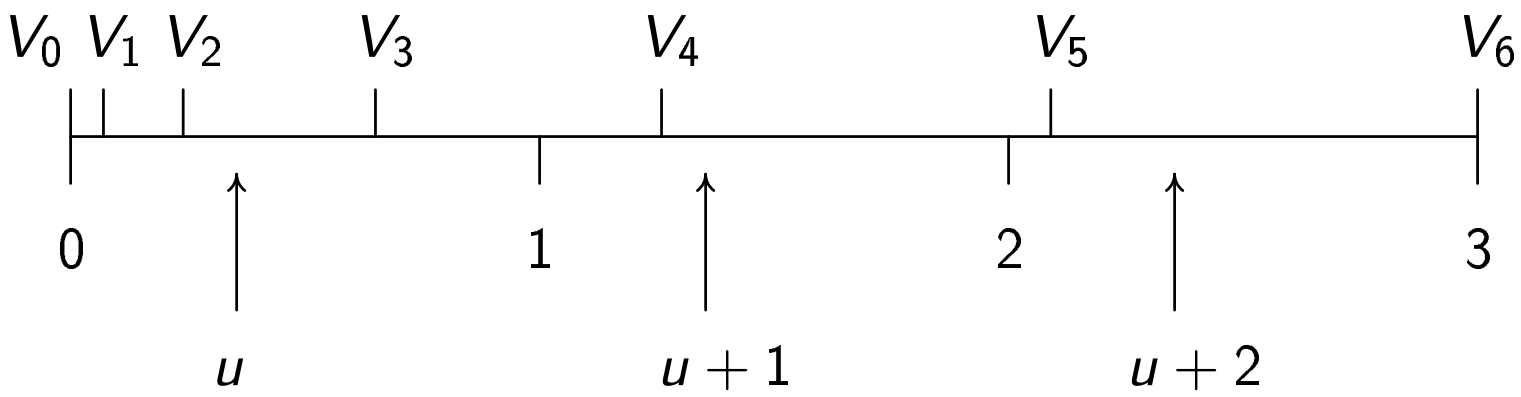
$$V_k = \sum_{j=1}^{k} \pi_k, \text{ avec } V_0 = 0 \text{ et } v_N = n.$$

Tirage systématique à probabilités inégales

Suppose que le nombre aléatoire $u = 0.354$. On sélectionne

- l'unité 3 est sélectionnée car $V_2 \leq u < V_3$,
- l'unité 3 est sélectionnée car $V_4 \leq u < V_5$,
- l'unité 3 est sélectionnée car $V_5 \leq u < V_6$,

Échantillon $s = \{3, 5, 6\}$.

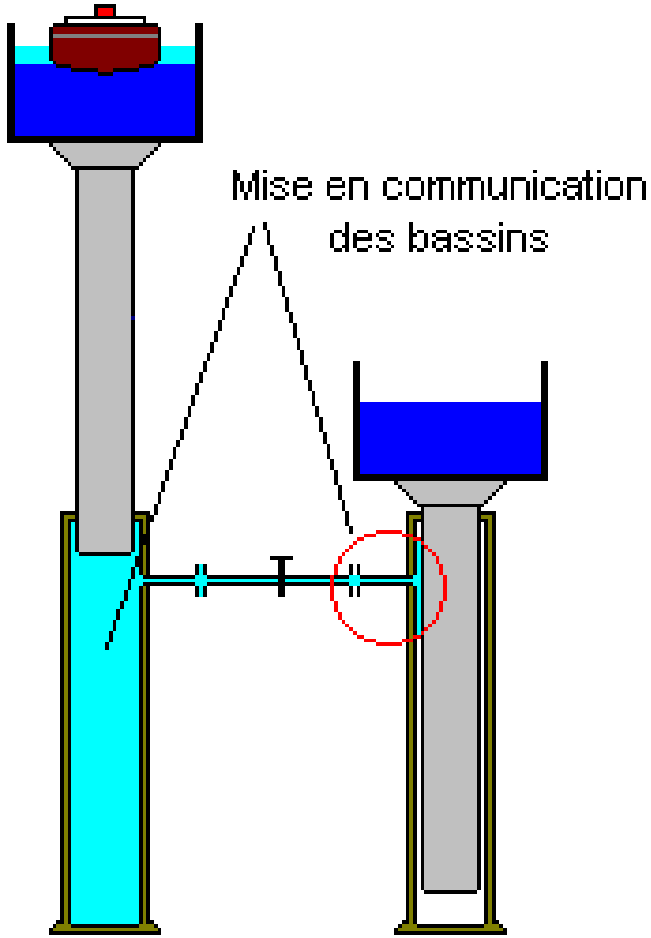


Méthode du pivot

Méthode du pivot

Méthode du pivot

de Michel Maigre©, site web de la Région wallonne : Direction des voies hydrauliques, canal du centre.



Méthode du pivot

- Méthode du pivot (Deville & Tillé, 2000).
- À chaque étape, on modifie les probabilités d'inclusion de deux unités (appelée i et j).
- Exemple

$$(0.07\ 0.17\ 0.41\ 0.61\ 0.83\ 0.91) \rightarrow \begin{cases} (0 \quad 0.24\ 0.41\ 0.61\ 0.83\ 0.91) & \text{proba} \quad 0.709 \\ (0.24\ 0 \quad 0.41\ 0.61\ 0.83\ 0.91) & \text{proba} \quad 0.291 \end{cases}$$

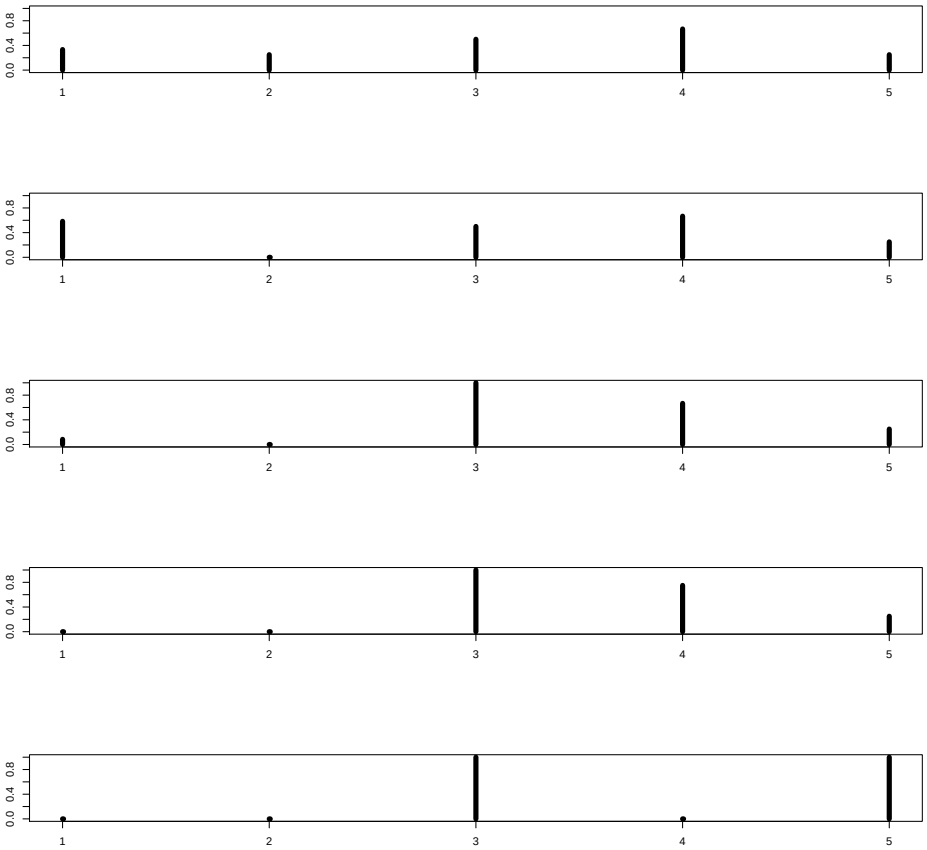
$$(0.07\ 0.17\ 0.41\ 0.61\ 0.83\ 0.91) \rightarrow \begin{cases} (0.07\ 0.17\ 0.41\ 0.61\ 1 \quad 0.74) & \text{proba} \quad 0.346 \\ (0.07\ 0.17\ 0.41\ 0.61\ 0.74\ 1 \quad) & \text{proba} \quad 0.654 \end{cases}$$

Méthode du pivot

Variantes de la méthode du pivot

- Méthode du pivot aléatoire
  Deville & Tillé (1998),
- Méthode du pivot ordonné (ou séquentiel) ou tirage systématique de Deville
  (Deville, 1998; Chauvet, 2012; Fuller, 1970),
- Méthode du pivot local (ou spatial)
  (Grafström, Lundström & Schelin, 2012).

Exemple, méthode du pivot ordonné



Échantillonnage spatial : Motivation

Exemples

- Sélection de cercles dans des prairies pour un monitoring d'observation de plantes.
- Sélection de communes dans un pays.
- Sélection d'arbres dans une forêt.
- On définit une distance entre des unités statistiques (par exemple des entreprises) basée sur des variables auxiliaires.

Problème : deux unités voisines sont en général similaires (autocorrélation).

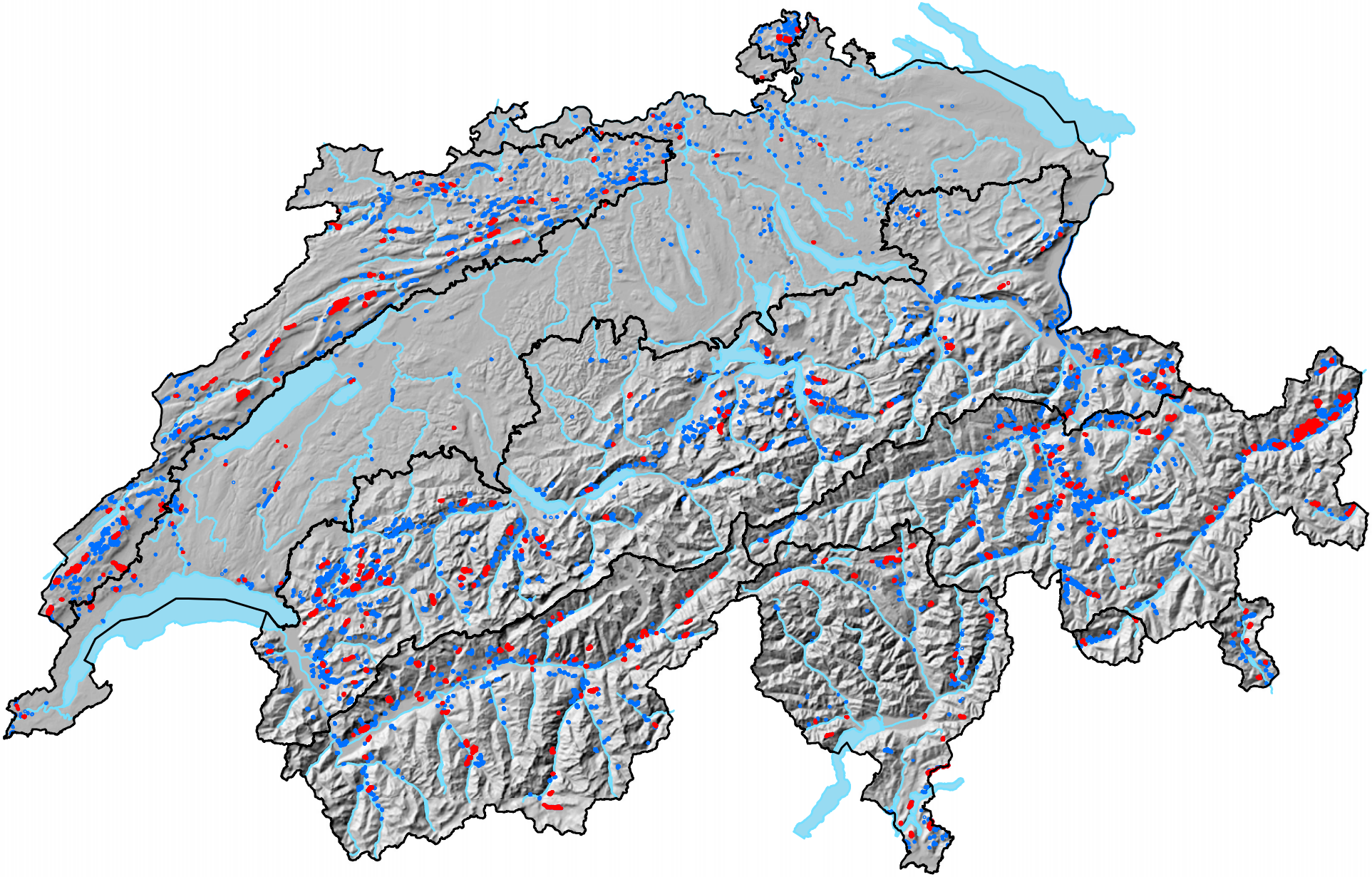
Population

- La population U peut être finie ou infinie.
- Une population infinie peut être discrétisée.
- Une information auxiliaire peut être disponible $\mathbf{x}_k$ (satellite).
- But : estimer le total des y_k.

Exemple : Monitoring des prairies sèches protégées en Suisse

(Tillé & Ecker, 2013)

Table – Exemple de prairies sèches dans le Jura suisse avec 9 polygones (limites noires). Six points (croix blanches) et deux points en réserve sont sélectionnés pour observation. Source orthophoto : swissimage© 2012 swisstopo (DV 033594).



Modèle pour l'échantillonnage spatial

- Modèle pour l'échantillonnage spatial

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{2}$$

$\mathrm{E}(\varepsilon_k) = 0$, $\mathrm{var}(\varepsilon_k) = \sigma_k^2$ et $\mathrm{cov}(\varepsilon_k, \varepsilon_\ell) = \sigma_{\varepsilon k}\sigma_{\varepsilon \ell}\rho_{k\ell}$

- Le modèle est hétéroscédastique avec autocorrélation.
-

$$\mathrm{AVar}(\widehat{Y}) = \mathrm{E}_p\left(\sum_{k\in S}\frac{\mathbf{x}_k^\top\boldsymbol{\beta}}{\pi_k} - \sum_{k\in U}\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 + \sum_{k\in U}\sum_{k\in U}\Delta_{k\ell}\frac{\sigma_{\varepsilon k}\sigma_{\varepsilon\ell}\rho_{k\ell}}{\pi_k\pi_\ell}.$$

Plan optimal :

- utilisation de probabilités d'inclusion proportionnelles aux $\sigma_{\varepsilon k}$,
- utilisation d'un plan équilibré sur les variables $\mathbf{x}_k$.
- évitement de la sélection d'unités voisines, donc sélection d'un échantillon bien étalé (ou équilibré spatialement)

Les méthodes habituelles (Wang, Stein, Gao & Ge, 2012)

- Les méthodes habituelles peuvent être utilisées : plans simples, stratifiés, par grappes, à deux degrés.
- Une stratification peut améliorer l'étalement.
- Rôle central de tirage systématique (car étalé).

Monitoring de biodiversité

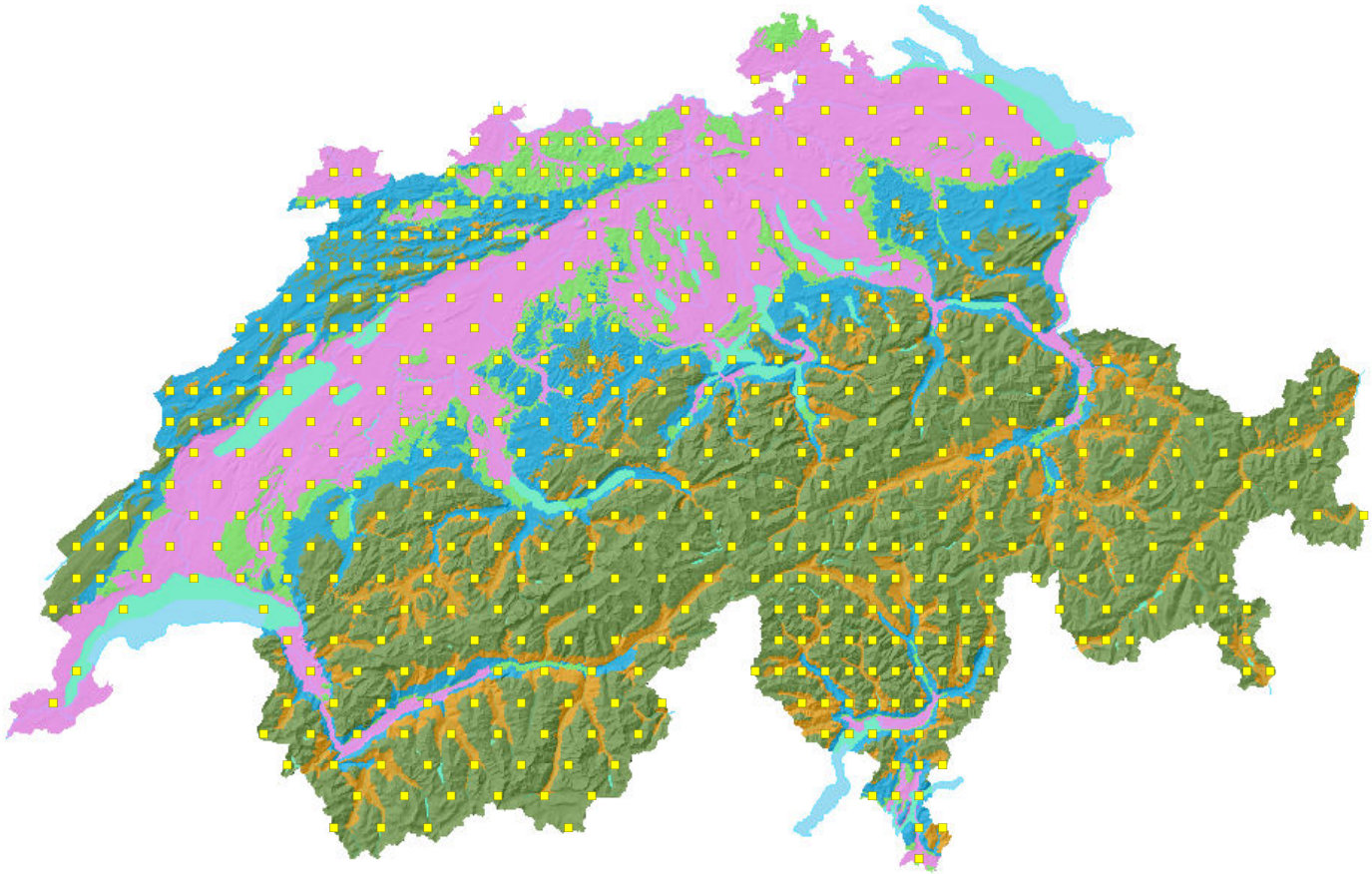


Table – Source : Monitoring suisse de la biodiversité : nombre de néophytes végétaux dans les placettes

Monitoring de biodiversité

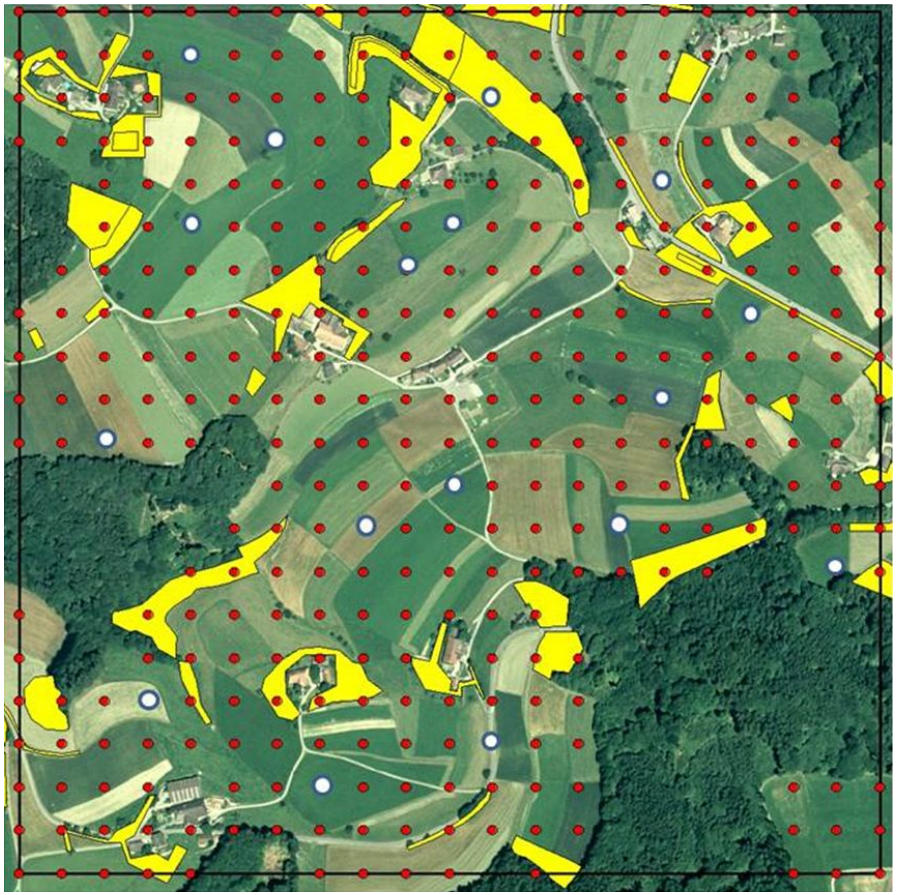


Table – Monitoring suisse de la biodiversité : nombre de néophytes végétaux dans les placettes

Monitoring forestier

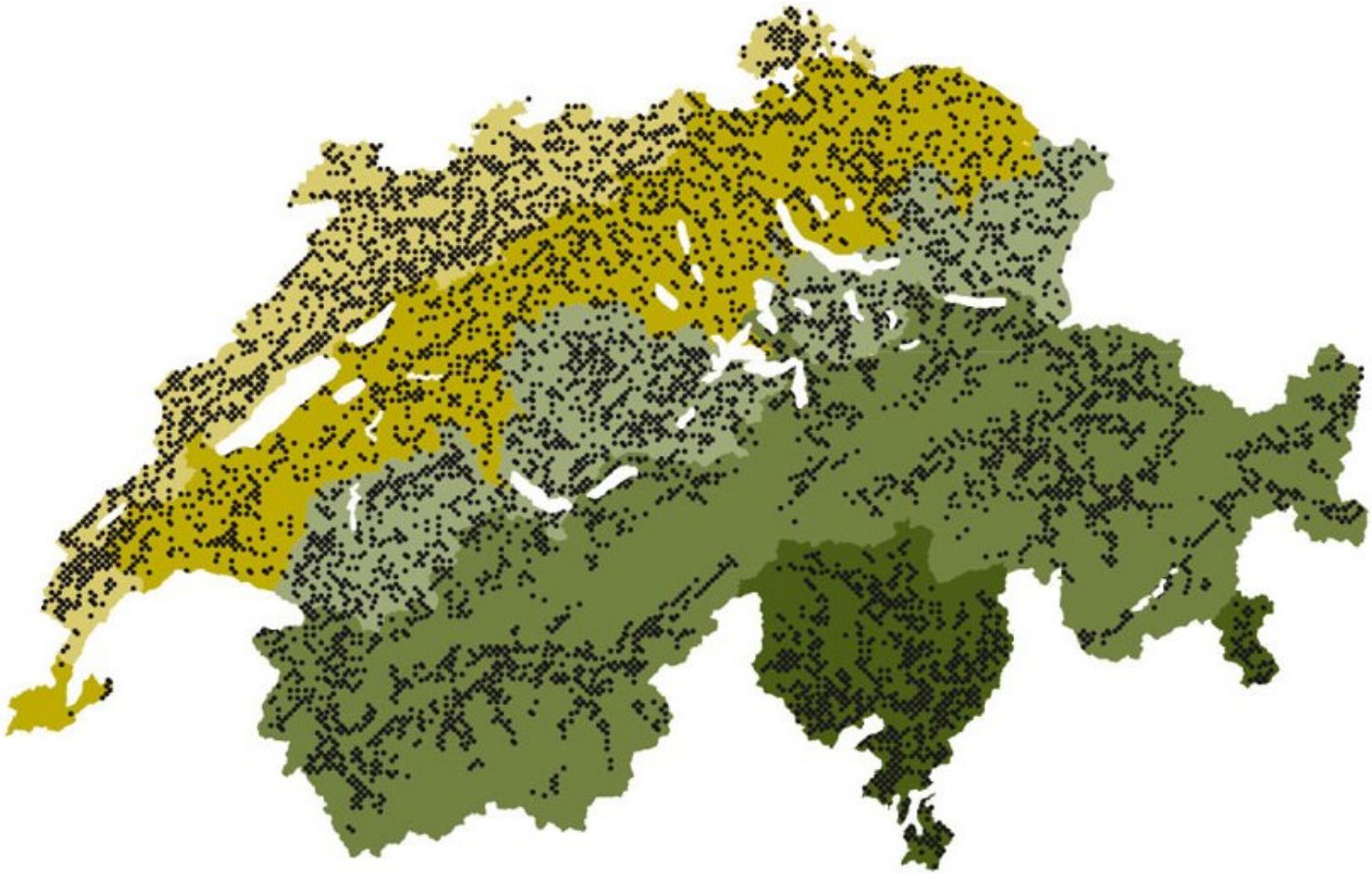


Table – Source : WSL (Wald, Schnee und Landschaft WSL) échantillon de placettes de forêt

Problèmes

L'échantillonnage systématique ne peut pas être utilisé :

1. quand les unités sont sélectionnées à probabilités inégales,
2. quand les unités sont sélectionnées de manière irrégulières sur le territoire (ex. constructions, communes),
3. quand le nombre d'unités est fini et qu'elles ne sont pas placées régulièrement sur le territoire.

Centres des communes belges

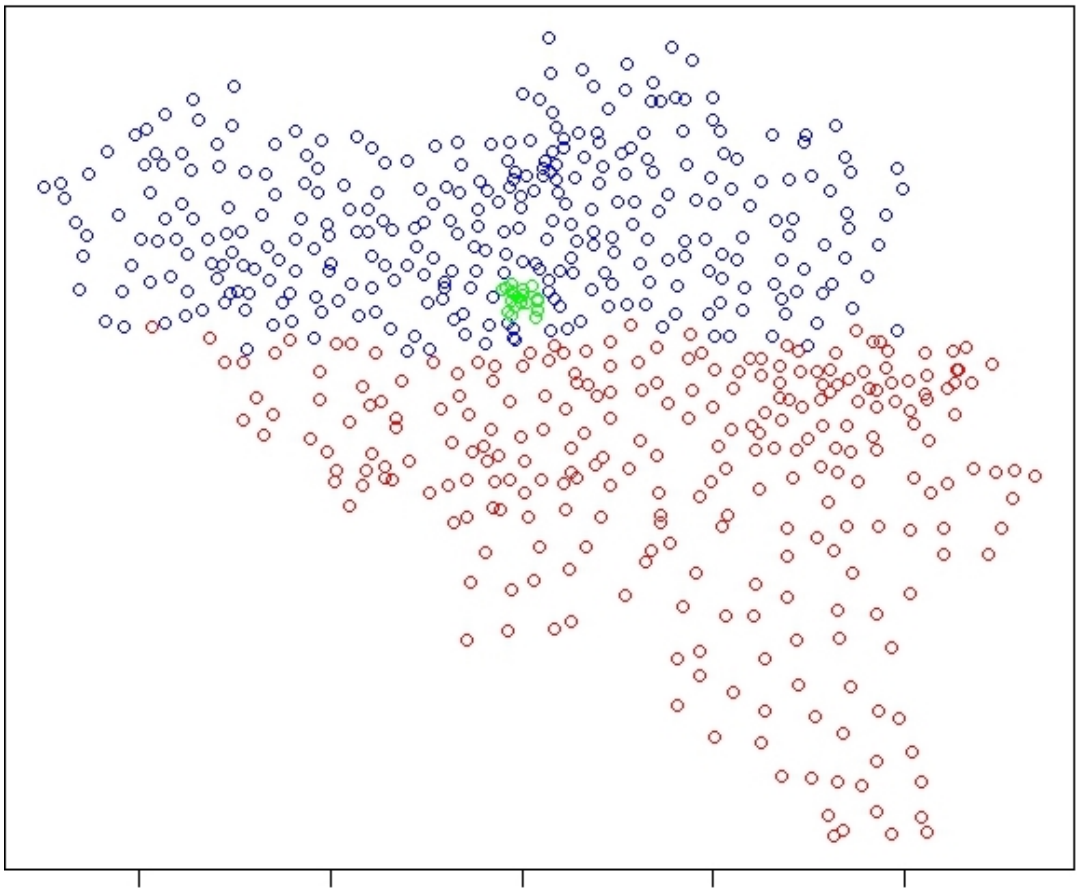


Table – Centres of des communes belges (Données IGN Belge)

Monitoring de prairies sèches (Tillé & Ecker, 2013)

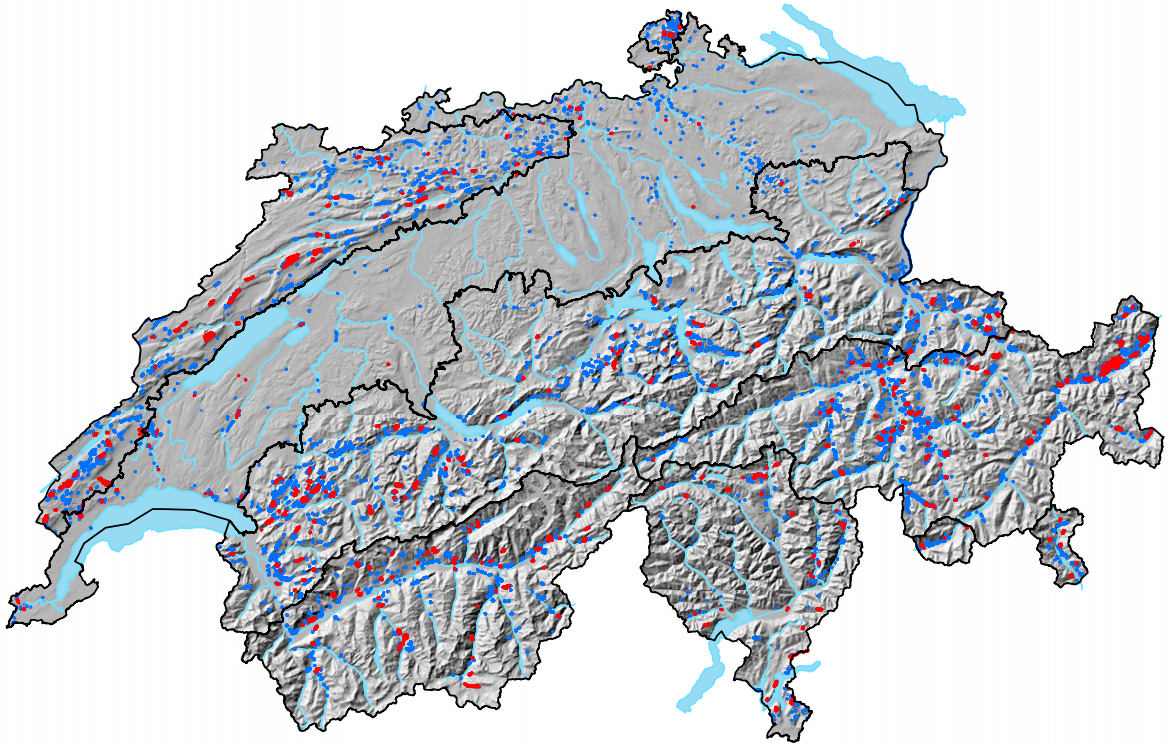


Table – Distribution des prairies sèches dans les six régions biogéographiques suisses. L'échantillon est en rouges sur la population en bleu.

Hilbert

Utilisation de courbes fractales (Quinn, Langbein, Martin & Elber, 2006; Lister & Scott, 2009).

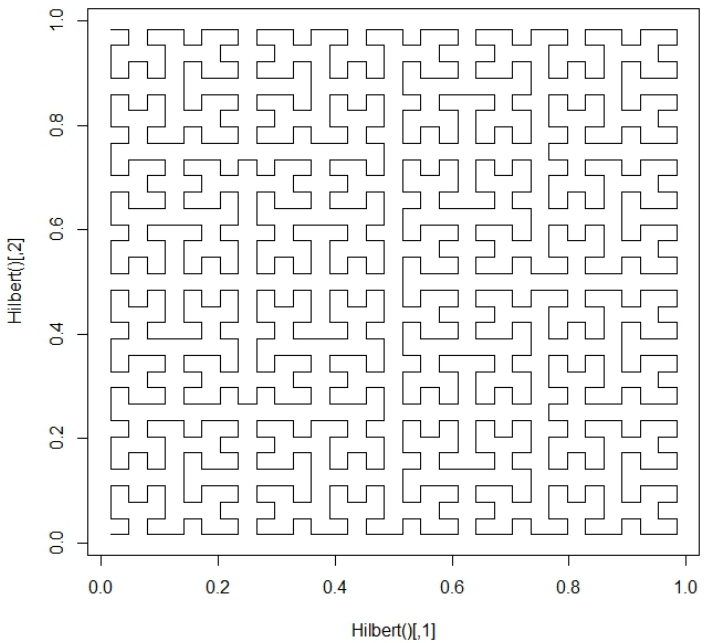


Table – Courbe de Hilbert

Generalized Random Tessellation Sampling GRTS,

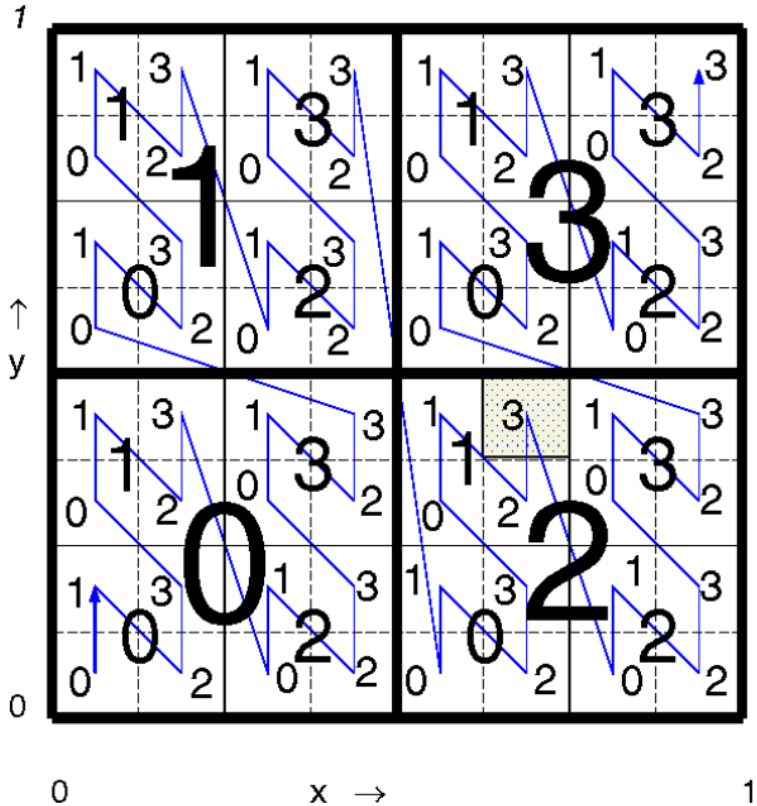
Échantillonnage aléatoire généralisé par tessellation

Algorithme de Stevens Jr. & Olsen (2003, 2004); Theobald, Stevens Jr., White, Urquhart, Olsen & Norman (2007)

1. Création d'une grille hiérarchique adressée.
2. Randomisation des adresses.
3. Ordonnancement des adresses en 1 dimension.
4. Sélection avec un tirage systématique.

L'échantillon est bien étalé mais il n'est pas équilibré.

Generalized Random Tessellation Sampling GRTS, Échantillonnage aléatoire généralisé par tessellation



(Source Stevens et Olsen) Courbe de Lebesgue. L'échantillon est bien étalé mais il n'est pas équilibré.

Problème du voyageur de commerce

Dickson & Tillé (2015)

- On détermine le plus court chemin passant par toutes les unités statistiques au moyen de la méthode du voyageur de commerce,
- On utilise le tirage systématique ou la méthode du pivot ordonnée le long du chemin.

Problème du voyageur de commerce

Autocorrélation le long du chemin pour la variable revenu moyen dans la commune : 0.4835873

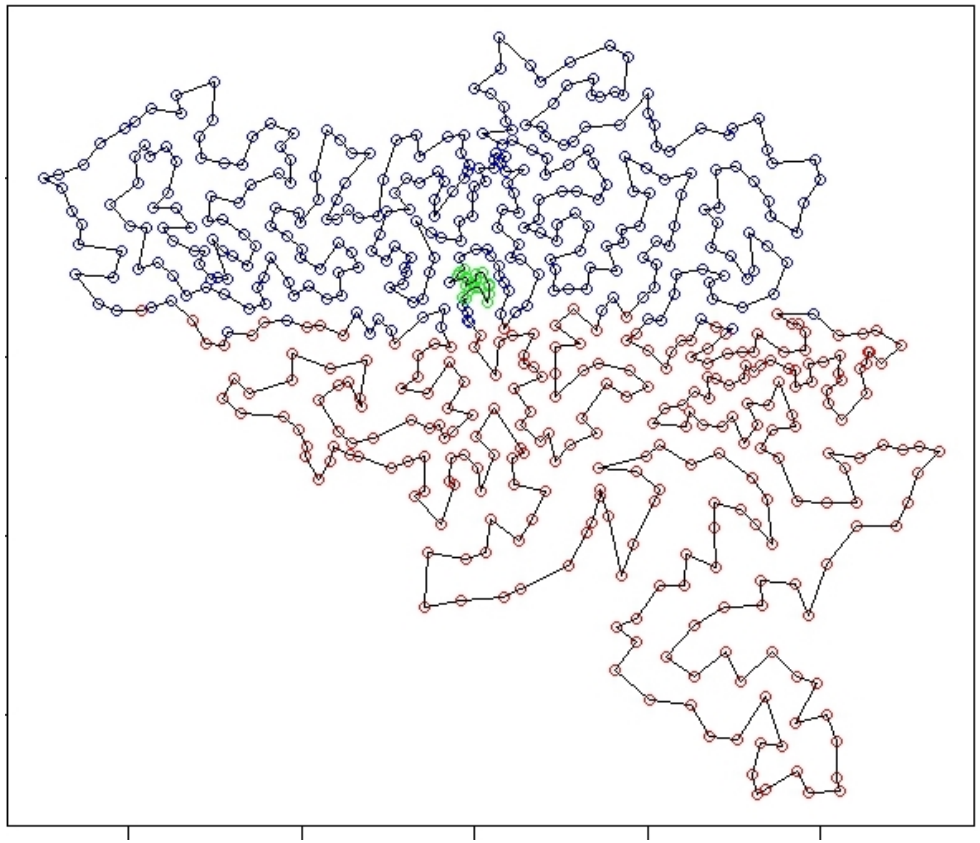


Table – Plus court chemin entre les centres des communes belges. Ensuite, on utilise le tirage systématique ou la méthode du pivot ordonnée.

Problème du voyageur de commerce

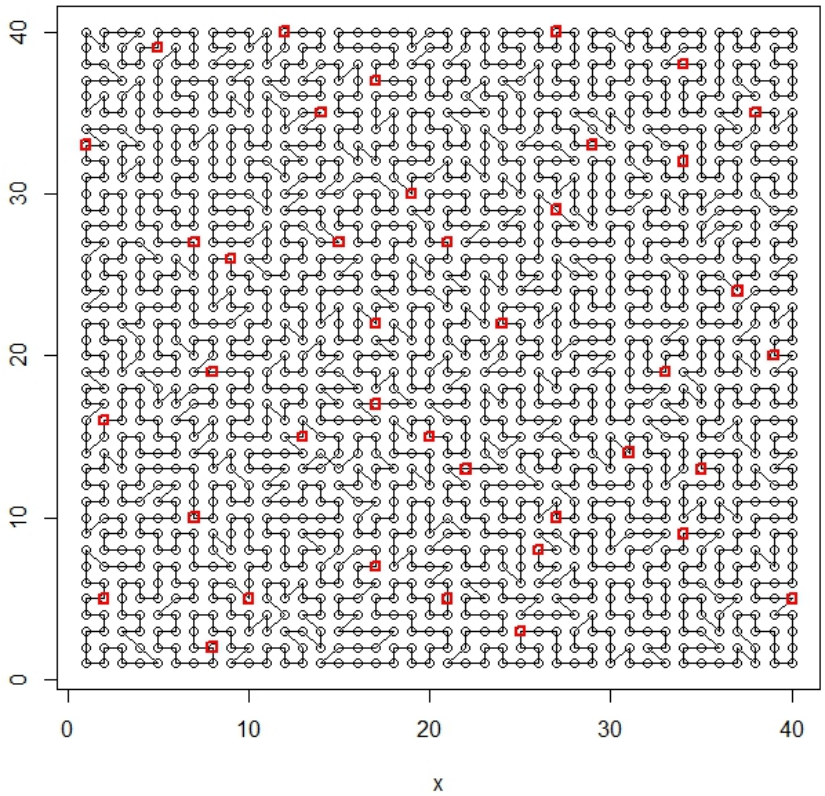
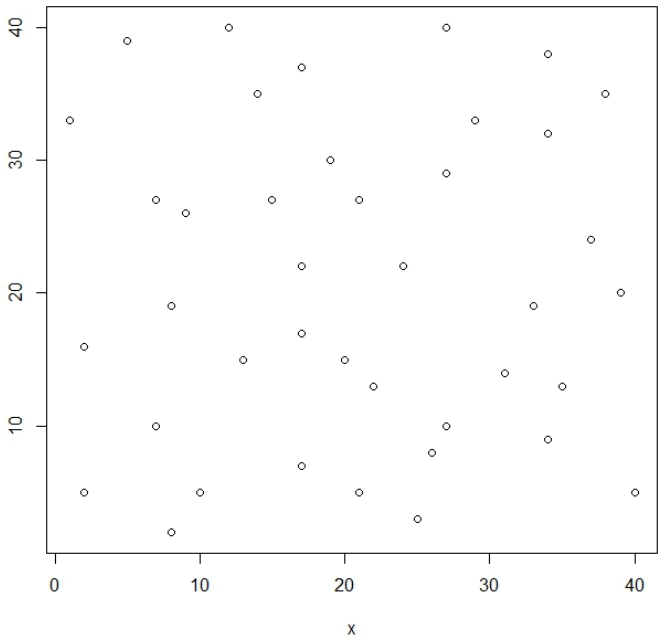


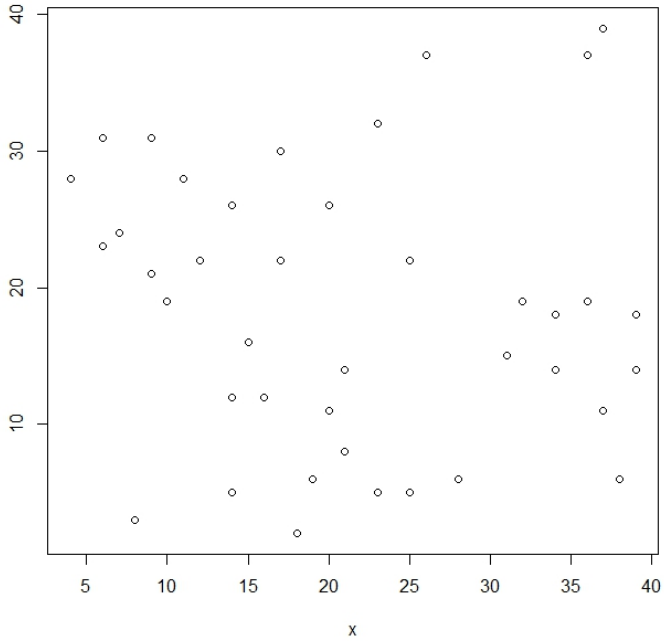
Table – Grille 40×40. Sélection de 40 points.

Problème du voyageur de commerce

Voyageur de commerce et tirage systématique



Plan simple



Méthode du pivot local

Algorithme de Grafström, Lundström & Schelin (2012)

Algorithme du pivot local

1. Sélection de deux unités voisines i et j avec des probabilités comprises strictement entre 0 et 1.
2. Application d'une étape de la méthode du pivot sur i et j.
3. On répète ces deux étapes.

L'échantillon est bien étalé mais les totaux ne sont pas équilibrés.

Méthode du cube locale(Grafström & Tillé, 2013)

- La méthode du cube (Deville & Tillé, 2004) permet d'obtenir des échantillons équilibrés $\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k$.
- La méthode du cube est composée de deux phases
  - Phase de vol
  - Phase d'atterrissage
- Durant la phase de vol, à chaque étape une composante de π passe à 0 ou à 1.
- Idée : Chaque étape de la phase de vol est appliquée seulement sur $p+1$ unités voisines. (p est le nombre de variable d'équilibrage).
- L'échantillon est à la fois équilibré et étalé.

Algorithmes pour un échantillonnage équilibré et étalé (doublement équilibré)

Idée

- Soit p le nombre de variables auxiliaires.
- Dans la méthode du cube, la dimension du sous-espace des contraintes est $N - p$.
- Pour faire tourner une étape de la phase de vol avec la méthode du cube, la population doit avoir une taille d'au moins $p + 1$ unités.

Algorithme

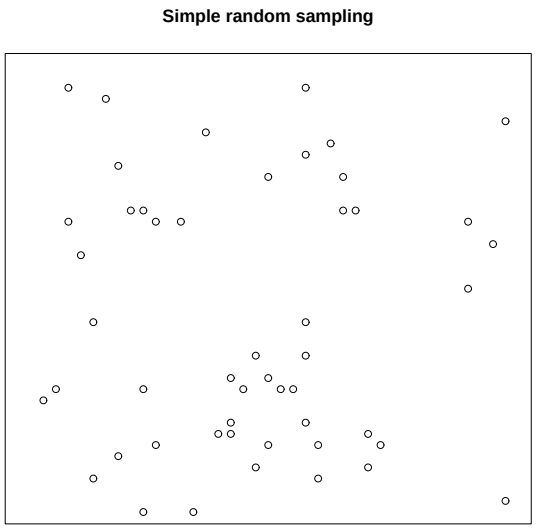
On répète ces étapes :

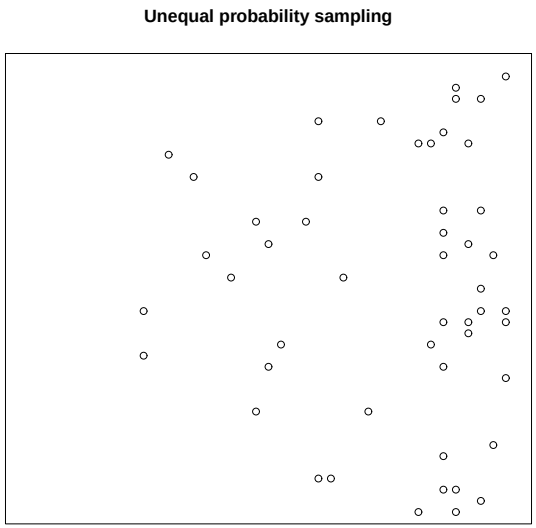
(1) Sélection d'un ensemble de of $p + 1$ unités voisines ayant des probabilités d'inclusion strictement entre 0 et 1.

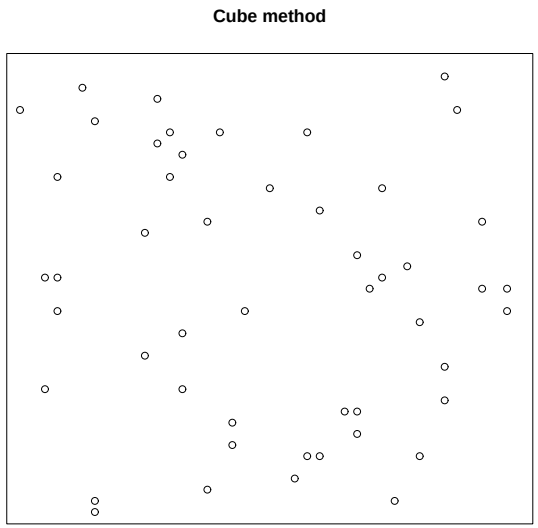
(2) On applique une étape de la phase de vol.

Méthodes non étalées

Méthodes non étalées : plans simple, à probabilités inégales et plan équilibré





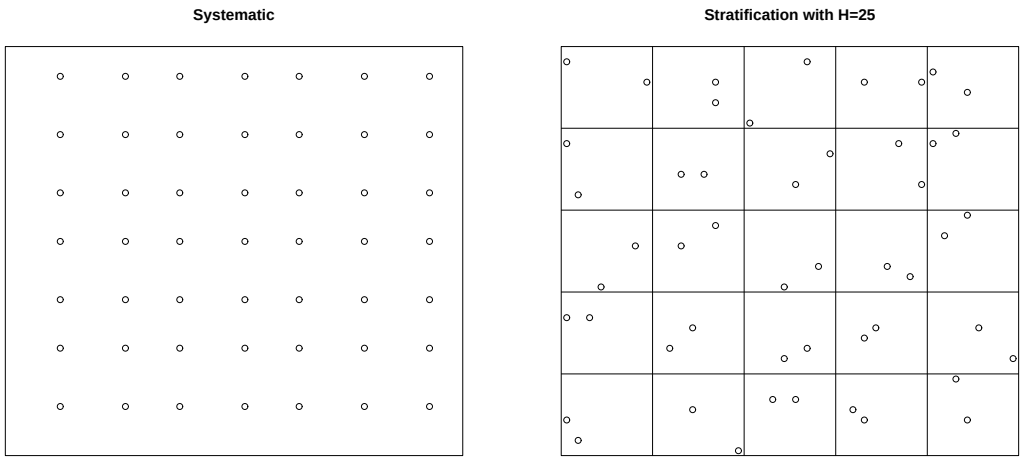


Méthode d'étalement habituelle

Méthodes de base

- Les méthodes les plus habituelles sont le tirage systématique et la stratification.
- Pas généralisable aux probabilités inégales.

Plans systématiques et stratifiés

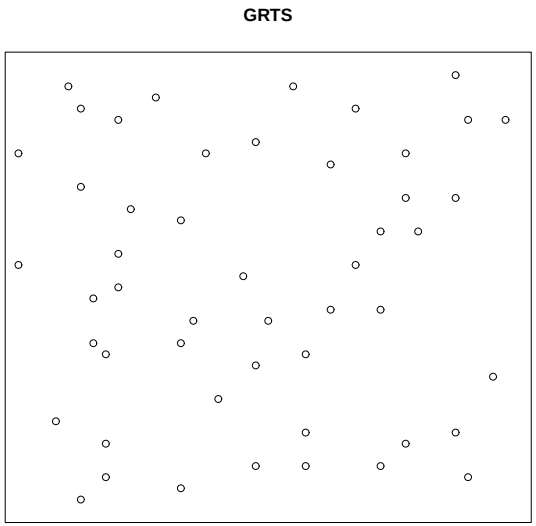


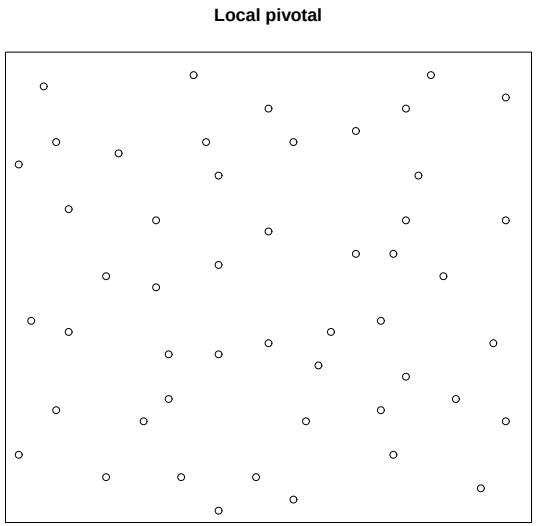
Échantillonnage systématique et stratification avec deux points par strate.

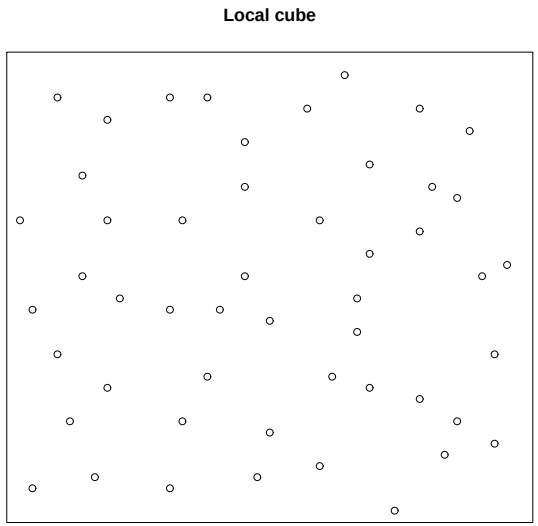
Méthode d'étalement

Méthode d'étalement

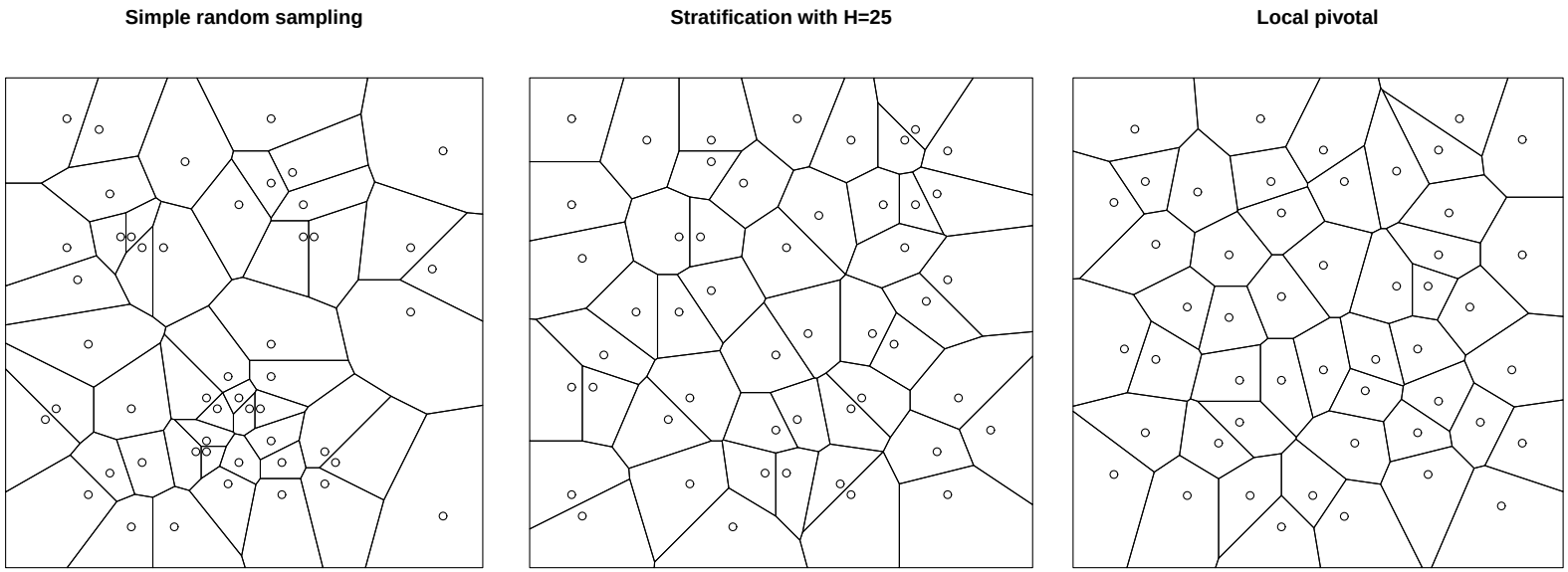
- Méthodes GRTS, méthode du pivot et cube local.
- Les échantillons sont bien étalés.
- À l'oeil, il est difficile de voir quel est l'échantillon le plus étalé.







Polygones de Voronoï



Qualité de l'équilibrage spatial

Table – Indices d'étalement pour les plans (Variance des polygones de Voronoï)

Plan	Indicateur d'équilibrage
Systématique	0.05
Plan simple	0.31
Stratification avec H=25	0.11
Pivot local	0.06
Méthode du cube	0.21
Méthode du cube locale	0.06
GRTS	0.09

Conclusions

Conclusions

- Il est important de modéliser la population avant de déterminer le plan de sondage.
- Le plan doit être déduit du modèle en appliquant les principes : restriction, randomisation et surreprésentation.
- Si autocorrélation, l'échantillon doit être étalé.
- L'échantillon représentatif selon Grafström & Lundström (2013); Grafström & Schelin (2014). Définition : similaire à la population pour tous les aspects important.
- L'échantillonnage spatial peut être appliqué à des données non-spatiales.
- Par exemple : distance entre entreprises (distance de Mahalanobis).
- L'étalement est comme une stratification dans chaque sous-ensemble convexe d'individus.

MERCI

MERCI

Bibliographie I

Berger, Y. G. (1996). Asymptotic variance for sequential sampling without replacement with unequal probabilities. *Survey Methodology* 22, 167–173.

Berger, Y. G. (1998a). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* 74, 149–168.

Berger, Y. G. (1998b). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* 67, 209–226.

Berger, Y. G. (1998c). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics* 14, 315–323.

Breidt, F. J. & Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the Cube method. *Journal of Statistical Planning and Inference* 141, 479–487.

Brewer, K. R. W. (1963). Ratio estimation in finite populations : Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5, 93–105.

Brewer, K. R. W. & Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York : Springer.

Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli* 18, 1099–1471.

Chauvet, G., Bonnéry, D. & Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference* 141, 984–994.

Chauvet, G., Haziza, D. & Lesage, É. (2015). Examining some aspects of balanced sampling in surveys. *Statistica Sinica* .

Chauvet, G. & Tillé, Y. (2005). *Fast SAS macros for balancing samples : user's guide*. Software Manual, University of Neuchâtel.

Chauvet, G. & Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics* 21, 9–31.

Chen, S. X., Dempster, A. P. & Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* 81, 457–469.

Bibliographie II

Chen, X. (1993). Poisson-binomial distribution, conditional bernoulli distribution and maximum entropy. Tech. rep., Department of Statistics, Harvard University.

Deville, J.-C. (1998). Une nouvelle (encore une !) méthode de tirage à probabilités inégales. Tech. Rep. 9804, Méthodologie Statistique, Insee.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Tech. rep., CREST-ENSAI, Rennes.

Deville, J.-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* 85, 89–101.

Deville, J.-C. & Tillé, Y. (2000). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference* 86, 215–227.

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. *Journal of Statistical Planning and Inference* 128, 569–591.

Dickson, M. M. & Tillé, Y. (2015). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics* , 1–14.

Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society* B32, 209–226.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika* , asp042.

Grafström, A. & Lisic, J. (2016). *BalancedSampling : Balanced and spatially balanced sampling*. R package version 1.5.2.

Grafström, A. & Lundström, N. L. P. (2013). Why well spread probability samples are balanced ? *Open Journal of Statistics* 3, 36–41.

Grafström, A., Lundström, N. L. P. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics* 68, 514–520.

Bibliographie III

Grafström, A. & Schelin, L. (2014). How to select representative samples ? *Scandinavian Journal of Statistics* 41, 277–290.

Grafström, A. & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 14, 120–131.

Hájek, J. (1981). *Sampling from a Finite Population*. New York : Marcel Dekker.

Hodges Jr., J. L. & Le Cam, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics* 31, 737–740.

Legg, J. C. & Yu, C. L. (2010). Comparison of sample set restriction procedures. *Survey Methodology* 36, 69–79.

Lister, A. J. & Scott, C. T. (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment* 149, 71–80.

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

Quinn, J., Langbein, F., Martin, R. & Elber, G. (2006). Density-controlled sampling of parametric surfaces using adaptive space-filling curves. In *Geometric Modeling and Processing - GMP 2006*, M.-S. Kim & K. Shimada, eds., vol. 4077 of *Lecture Notes in Computer Science*. New York : Springer, pp. 465–484.

Rousseau, S. & Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Tech. rep., Insee, Paris.

Royall, R. M. (1970a). Finite population sampling - On labels in estimation. *Annals of Mathematical Statistics* 41, 1774–1779.

Royall, R. M. (1970b). On finite population sampling theory under certain linear regression models. *Biometrika* 57, 377–387.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference*, V. P. Godambe & D. A. Sprott, eds. Toronto, Montréal : Holt, Rinehart et Winston.

Bibliographie IV

Royall, R. M. (1992). The model based (prediction) approach to finite population sampling theory. In *Current issues in statistical inference : Essays in honor of D. Basu*, M. Ghosh & P. K. Pathak, eds., vol. 17 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, pp. 225–240.

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Tech. Rep. TC 354, Department of Statistics, Stanford University.

Stevens Jr., D. L. & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14, 593–610.

Stevens Jr., D. L. & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99, 262–278.

Theobald, D. M., Stevens Jr., D. L., White, D., Urquhart, N. S., Olsen, A. R. & Norman, J. B. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management* 40, 134–146.

Tillé, Y. (2004). Estimation de la précision dans les enquêtes longitudinales. Tech. rep., Université de Neuchâtel, Neuchâtel.

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method : an appraisal. *Survey Methodology* 37, 215–226.

Tillé, Y. & Ecker, K. (2013). Complex national sampling design for long-term monitoring of protected dry grasslands in Switzerland. *Environmental and Ecological Statistics* 21, 1–24.

Tillé, Y. & Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika* 91, 913–927.

Tillé, Y. & Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters* 74, 31–37.

Tillé, Y. & Matei, A. (2015). *sampling : Survey Sampling*. R package version 2.7.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite Population Sampling and Inference : A Prediction Approach*. New York : Wiley.

Wang, J.-F., Stein, A., Gao, B.-B. & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics* 2, 1 – 14.