

# Échantillonnage spatial : l'État de l'art

Yves Tillé  
Université de Neuchâtel

2018  
Paris, JMS-Insee

# Table de contents

- 1 Introduction, notation
- 2 Échantillonnage avec autocorrélations
- 3 Méthodes en une dimension
- 4 Échantillonnage spatial: deux dimensions
- 5 Algorithme pour un échantillonnage étalé et équilibré
- 6 Conclusions

# Notation

## Notation

- Paper: Tillé & Wilhelm (2017).
- Population:  $U = \{1, \dots, k, \dots, N\}$ .
- Échantillon  $s \subset U$ .  
Exemple  $U = \{1, 2, 3, 4, 5\}$ , échantillon  $s = \{2, 3, 5\}$  autre notation  $\mathbf{s} = (0, 1, 1, 0, 1)^\top$ .
- 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 jointes  $\pi_{kl} = \Pr(\{k, \ell\} \in S) = \sum_{s \supset \{k, \ell\}} p(s)$ .
- Total  $Y = \sum_{k \in U} y_k$ .      Moyenne  $\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k$ .

# Notation

- L'estimateur de Narain-Horvitz-Thompson (NHT) :  $\widehat{Y} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$ .

- $\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l & \text{si } k \neq l \\ \pi_k(1 - \pi_k) & \text{si } k = l. \end{cases}$

- Variance de l'estimateur de Narain-Horvitz-Thompson :

$$\text{var}_p(\widehat{Y}) = \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{kl}.$$

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

- Estimateurs

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

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

# Modèle pour l'échantillonnage spatial

- Modèle pour l'échantillonnage spatial

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad (1)$$

$$E_M(\varepsilon_k) = 0, \text{ var}_M(\varepsilon_k) = \sigma_{\varepsilon_k}^2 \text{ et } \text{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_{\varepsilon_k} \sigma_{\varepsilon_\ell} \rho_{kl}.$$

- Le modèle admet de l'hétéroscédasticité et de l'autocorrélation.

$$\begin{aligned} \text{AVar}(\widehat{Y}) &= E_p E_M(\widehat{Y} - Y) \\ &= E_p \left( \sum_{k \in S} \frac{\mathbf{x}_k^\top \boldsymbol{\beta}}{\pi_k} - \sum_{k \in U} \mathbf{x}_k^\top \boldsymbol{\beta} \right)^2 + \sum_{k \in U} \sum_{\ell \in U} \Delta_{kl} \frac{\sigma_{\varepsilon_k} \sigma_{\varepsilon_\ell} \rho_{kl}}{\pi_k \pi_\ell}. \end{aligned}$$

Plan optimal :

- on utilise des probabilités d'inclusion inégales proportionnelles aux  $\sigma_{\varepsilon_k}$ ,
- on utilise un plan de sondage équilibré sur les variables auxiliaires  $\mathbf{x}_k$ ,
- on évite la sélection d'unités voisines, autrement dit, on sélectionne un échantillon bien étalé.

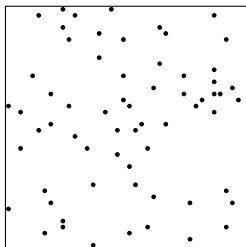
## Méthodes habituelles (Wang, Stein, Gao & Ge, 2012)

### Méthodes habituelles

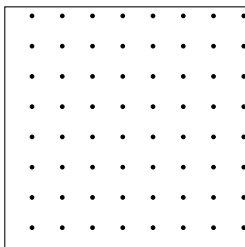
- Méthodes habituelles: plans simples, stratifiés, systématiques.
- La stratification améliore l'étalement.
- Rôle central du tirage systématique (parce que étalé).

## Méthodes habituelles

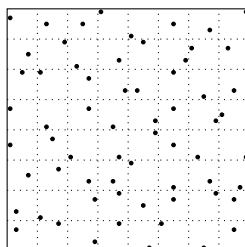
## Méthodes habituelles



plan simple



systématique



stratification

## Monitoring de Biodiversité

L'échantillonnage le plus étalé est le tirage systématique à deux dimensions.

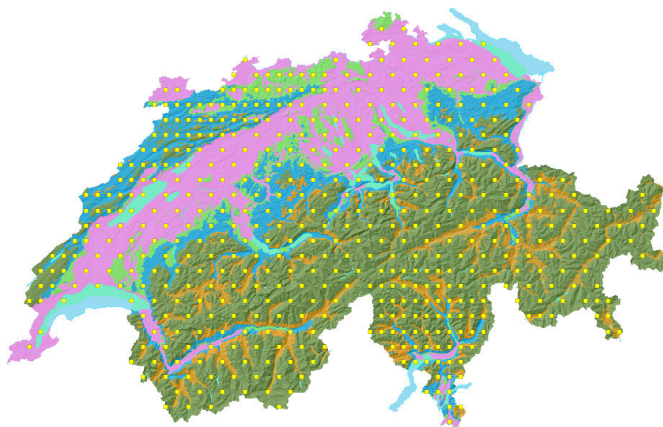


Table: WSL Monitoring de Biodiversité suisse



# Méthodes en une dimension: Tirage systématique 1

## Tirage systématique

- Probabilités d'inclusion cumulées

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

- $u$  une variable uniforme  $[0, 1]$ .
- Les unités telles que  $\lfloor V_k - u \rfloor \neq \lfloor V_{k-1} - u \rfloor$  sont sélectionnées dans l'échantillon. (Madow, 1949)
- Entropie minimale (Pea, Qualité & Tillé, 2007).

# Tirage systématique 2

## Tirage systématique

### Example

Suppose that  $N = 6$  et  $n = 3$ .

$k$	0	1	2	3	4	5	6	Total
$\pi_k$		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	

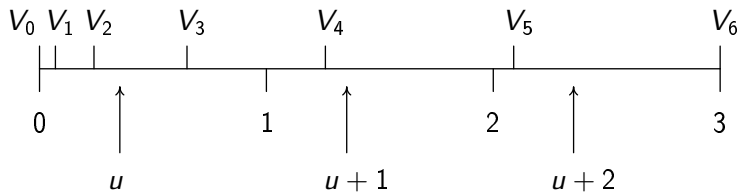
# Tirage systématique 3

## Tirage systématique

On suppose que  $u = 0.354$ .

- Comme  $V_2 \leq u < V_3$ , unité 3 est sélectionnée;
- Comme  $V_4 \leq u + 1 < V_5$ , unité 5 est sélectionnée;
- Comme  $V_5 \leq u + 2 < V_6$ , unité 6 est sélectionnée.

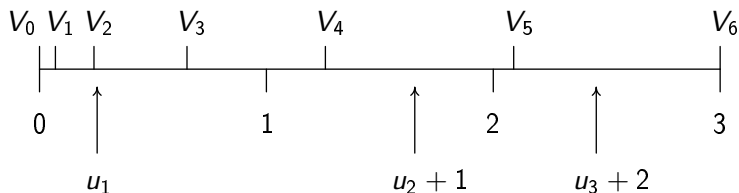
L'échantillon sélectionné est donc  $\mathbf{s} = (0, 0, 1, 0, 1, 1)$ .



# Deville Tirage systématique

## Deville (1998) Tirage systématique

Pour chaque intervalle de longueur 1, une variable uniforme est générée.



Une dépendance est introduite  $u_1, u_2,$  et  $u_3$  afin de ne pas sélectionner deux fois la même unité

# Deville Tirage systématique

## Deville Tirage systématique

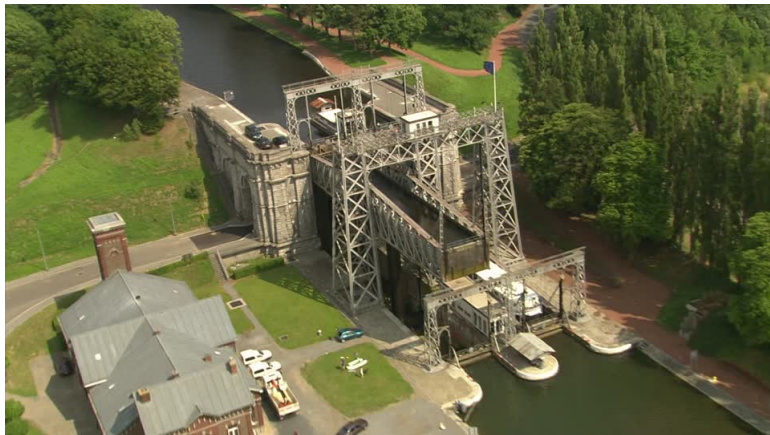
- Si l'unité frontière  $\ell$  est sélectionnée à l'étape  $i - 1$ ,  $u_i$  a la fonction de densité :

$$f_1(x) = \begin{cases} \frac{1}{i - V_\ell} & \text{si } x \geq V_\ell - (i - 1) \\ 0 & \text{si } x < V_\ell - (i - 1) \end{cases}, x \in [0, 1[.$$

- Si  $\ell$  n'est pas sélectionné à l'étape  $i - 1$ ,  $u_i$  a la fonction de densité :

$$f_2(x) = \begin{cases} 1 - \frac{(i - 1 - V_{\ell-1})(V_\ell - i + 1)}{[1 - (i - 1 - V_{\ell-1})][1 - (V_\ell - i + 1)]} & \text{si } x \geq V_\ell - i + 1 \\ \frac{1}{1 - (i - 1 - V_{\ell-1})} & \text{si } x < V_\ell - i + 1. \end{cases}$$

# Méthode du pivot



# Méthode du pivot



## Méthode du pivot

from Michel Maigre<sup>©</sup>, web site de Région Wallone: Direction des voies hydrauliques, canal du centre.



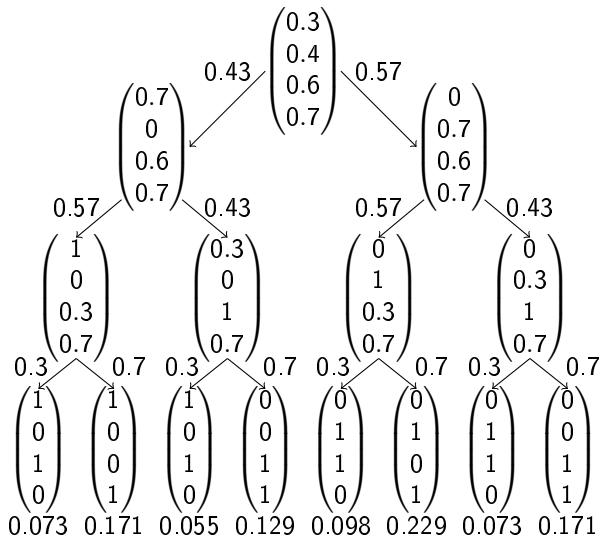
# Méthode du pivot

- Méthode du pivot (Deville & Tillé, 2000).
- À chaque étape, deux probabilités d'inclusion ( $i$  et  $j$ ) sont modifiées de manière aléatoire.
- Exemple

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

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

## Méthode du pivot



# Méthode du pivot

- Méthode du pivot (Deville & Tillé, 2000).
- On prend deux unités (notées  $i$  et  $j$ ) dans la population.
- Deux cas: Si  $\pi_i + \pi_j > 1$ , alors

$$\lambda = \frac{1 - \pi_j}{2 - \pi_i - \pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ 1 & k = i \\ \pi_i + \pi_j - 1 & k = j, \end{cases}$$

$$\pi_k^{(2)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ \pi_i + \pi_j - 1 & k = i \\ 1 & k = j. \end{cases}$$

## Méthode du pivot

Si  $\pi_i + \pi_j < 1$ , alors

$$\lambda = \frac{\pi_i}{\pi_i + \pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ \pi_i + \pi_j & k = i \\ 0 & k = j, \end{cases} \quad \text{et } \pi_k^{(2)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ 0 & k = i \\ \pi_i + \pi_j & k = j. \end{cases}$$

# Méthode du pivot

## Variantes

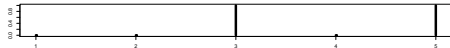
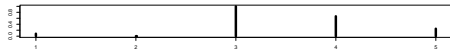
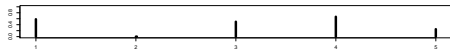
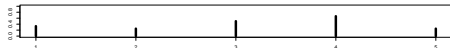
- Méthode du pivot ordonnée ou séquentielle ou tirage systématique de Deville  
(Deville, 1998),
- Méthode du pivot aléatorisée  
Deville & Tillé (1998),
- Méthode du pivot locale ou Méthode du pivot spatiale.  
(Grafström, Lundström & Schelin, 2012).

# Méthode du pivot

## Variantes

- Chauvet (2012) a montré que la méthode du pivot ordonné est la même que le tirage systématique de Deville.
- Fuller (1970) a proposé une méthode très similaire à la méthode du pivot.
- Tillé (2018) a proposé une implémentation simple de la méthode de Fuller au moyen d'une unité fantôme..

## Méthode du pivot: Exemple



## Problèmes

Le tirage systématique ne peut pas être utilisé quand

- 1 les probabilités d'inclusion sont inégales,
- 2 les unités statistiques sont disposées irrégulièrement dans l'espace.



## Centres des communes belges

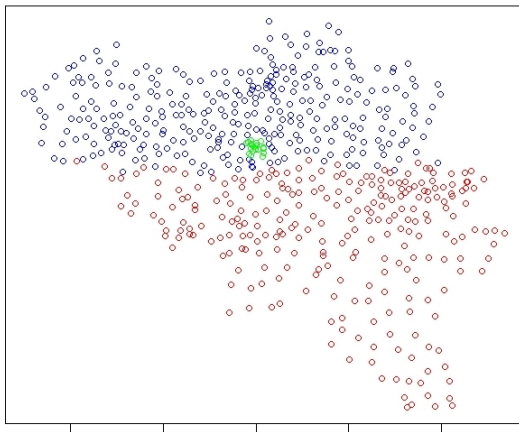


Table: Centres des communes belges (Données IGN Belgique)

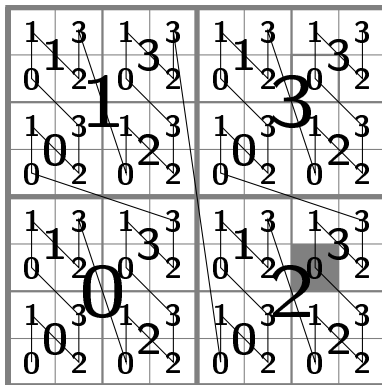
## Generalized Random Tessellation Sampling

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

- 1 Création d'une grille hiérarchisée avec des adresses.
- 2 On randomise les adresses.
- 3 On construit une courbe qui passe par toute les adresses.
- 4 On sélectionne un échantillon systématique le long de cette courbe.

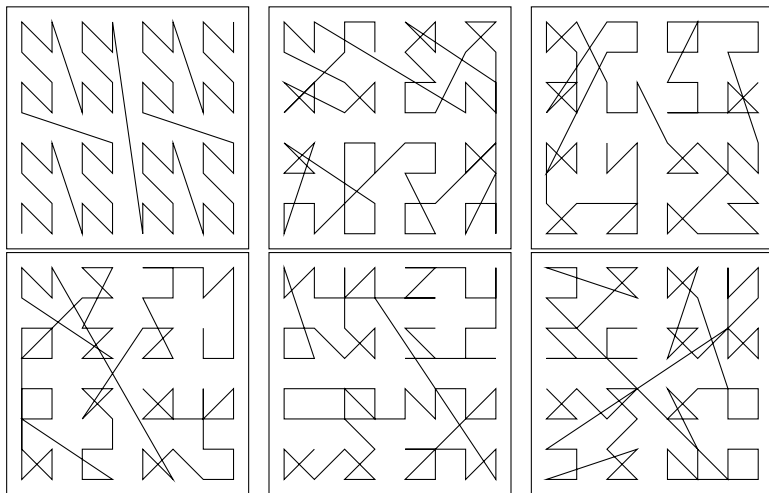
L'échantillon est bien étalé.

## Generalized Random Tessellation Sampling



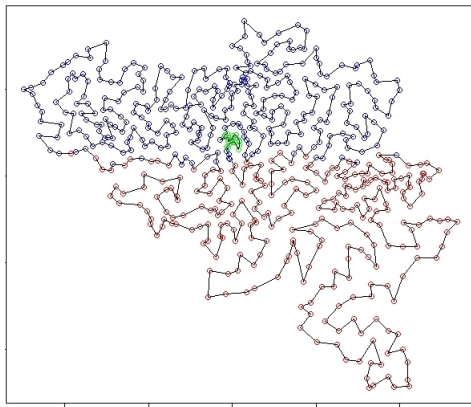
L'échantillon est bien étalé.

## Generalized Random Tessellation Sampling



## Problème du voyageur de commerce

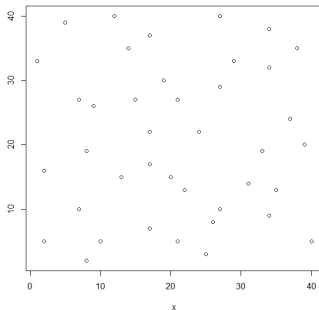
Autocorrélation le court du chemin pour le revenu moyen dans les communes : 0.4835873



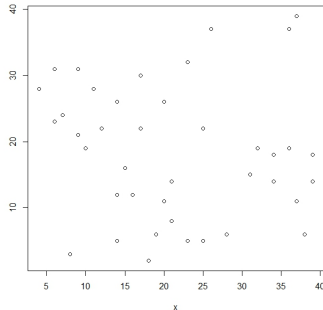
**Table:** Tirage systématique le long du plus court chemin. (Dickson & Tillé, 2016).

## Problème du voyageur de commerce

### Problème du voyageur de commerce et tirage systématique



### Plan simple



## La méthode du pivot locale

Algorithme de Grafström, Lundström & Schelin (2012)

- 1 On choisit deux unités  $i$  et  $j$  avec des  $\pi_k$  strictement entre 0 et 1 qui sont proches.
- 2 On applique une étape de la méthode du pivot  $i$  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 &amp; Tillé, 2013)

- Méthode du cube (Deville & Tillé, 2004) pour obtenir un échantillon équilibré  $\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k$ .
- The méthode du cube est composée de deux phases
  - La phase de vol,
  - La phase d'atterrissage.
- Durant la phase de vol, à chaque étape, les équations d'équilibrage sont satisfaites. Le vecteur  $\boldsymbol{\pi}$  est modifié de manière aléatoire. Une composante de  $\boldsymbol{\pi}$  est mise à 0 ou à 1.
- Idée: À chaque étape, on applique la phase de vol sur un sous-ensemble de  $p + 1$  unités voisines. ( $p$  est le nombre de variables d'équilibrage).
- L'échantillon sera bien étalé et équilibré.



## Algorithm for étalé et équilibré sampling (doubly équilibré)

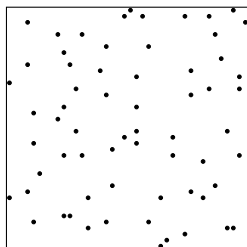
- Soit  $p$  le nombre de variables auxiliaires.
- Pour la méthode du cube, la dimension du sous-espace des contraintes est  $N - p$ .
- Afin de pouvoir appliquer la phase de vol de la méthode du cube, la taille de la population doit avoir au moins  $p + 1$  unités.

**Algorithme** On répète:

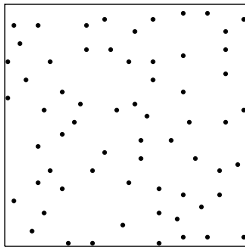
- (1) On sélectionne un ensemble de  $p + 1$  unités voisines qui ont des probabilités strictement entre 0 et 1.
- (2) On applique une étape de la phase de vol.

## Méthodes complexes

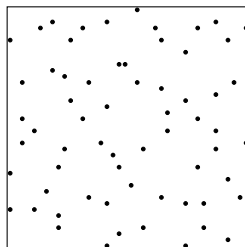
## Méthodes complexes



GRTS

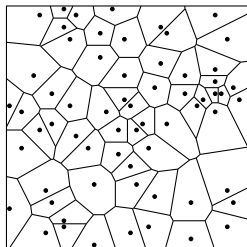


local pivotal

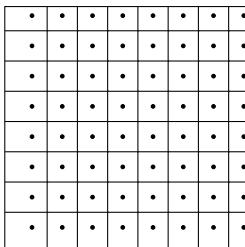


local cube

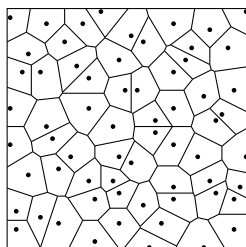
## Polygones Voronoï



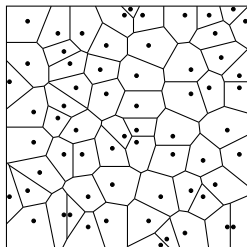
simple



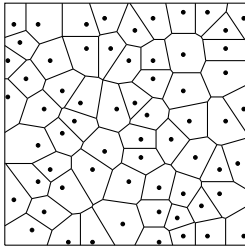
systématique



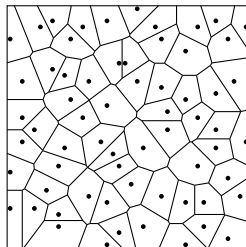
stratification



GRTS

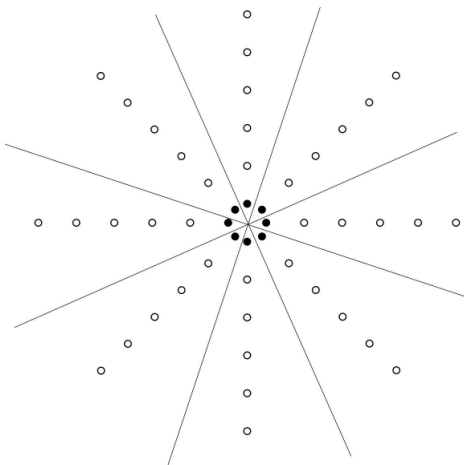


pivot local



cube locale

# Polygones Voronoï



## Qualité de l'étalement

**Table:** Indice d'étalement spatial pour les principaux plans de sondage (Variances de la somme des probabilités d'inclusion dans les polygones de Voronoï autour des unités)-

Plans	Indicateur d'étalement
Systematique	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 intermédiaires

### Conclusions intermédiaires

- Le plan le plus étalé est le tirage systématique.
- La méthode du pivot local est très bien étalée, mais n'est pas la solution optimale.
- Et-il possible de faire mieux? Un algorithme qui donne le tirage systématique quand c'est possible.

# Une autre mesure d'étalement basée sur l'indice de Moran

## Une autre mesure d'étalement basée sur l'indice de Moran

- Tillé, Dickson, Espa & Giuliani (2018).
- Corrélation entre:
  - le vecteur des indicatrices  $\mathbf{s} = (0 \ 1 \ 0 \ 0 \ 1 \ \dots \ 0)$ .
  - La moyenne locale de ce vecteur. La moyenne locale de  $k$  est la moyenne des  $\frac{1}{\pi_k} - 1$  plus proches voisins  $k$ .

# Une autre mesure d'étalement basée sur l'indice de Moran

Une autre mesure d'étalement basée sur l'indice de Moran



$$I_B = \frac{(\mathbf{s} - \bar{\mathbf{s}}_w)^\top \mathbf{W} (\mathbf{s} - \bar{\mathbf{s}}_w)}{\sqrt{(\mathbf{s} - \bar{\mathbf{s}}_w)^\top \mathbf{D} (\mathbf{s} - \bar{\mathbf{s}}_w)} (\mathbf{s} - \bar{\mathbf{s}}_w)^\top \mathbf{B} (\mathbf{s} - \bar{\mathbf{s}}_w)}.$$

- où  $\mathbf{W} = (w_{kl})$ ,  $w_{ij}$  indique si  $j$  est près de  $i$ ,  $w_{ii} = 0$ ,  $\mathbf{D}$  est la matrice diagonale des  $w_i$ .

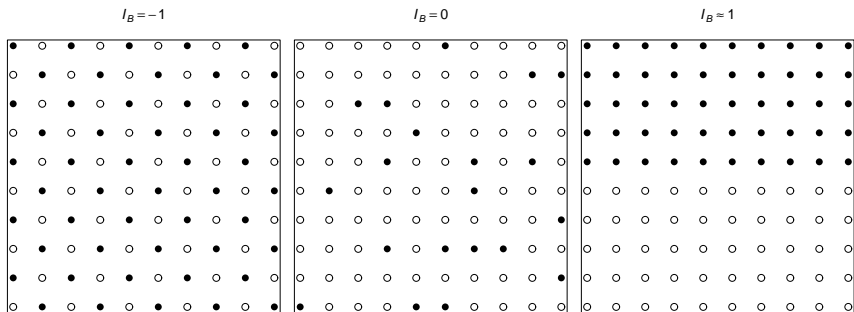
$$\mathbf{A} = \mathbf{D}^{-1} \mathbf{W} - \frac{\mathbf{1} \mathbf{1}^\top \mathbf{W}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}}$$

$$\mathbf{B} = \mathbf{A}^\top \mathbf{D} \mathbf{A} = \mathbf{W}^\top \mathbf{D}^{-1} \mathbf{W} - \frac{\mathbf{W}^\top \mathbf{1} \mathbf{1}^\top \mathbf{W}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}}.$$



# Une autre mesure d'étalement basée su l'indice de Moran

Exemples (Tillé, Dickson, Espa & Giuliani, 2018)



# Une autre idée

## Une autre idée

- Grafström & Lundström (2013)  
“Pourquoi les échantillons bien étalés sont équilibrés?”.
- Les échantillons étalés sont équilibrés partout.
- Les échantillons étalés sont localement stratifiés.
- Pourquoi ne pas stratifier partout?

# Définition d'une distance

Définition d'une distance : pour le tableau

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

## Définition d'une distance

Définition d'une distance : pour le tableau

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

$$d(1,2) = d(1,5) = d(1,13) = d(1,4) = 1,$$

$$d(1,6) = d(1,8) = d(1,16) = d(1,14) = \sqrt{2}.$$

## Définition d'une distance

Définition d'une distance : pour le tableau

- Utilisation de la méthode du cube.
- Construction d'une matrice de contrainte avec une strate par unité.
- $\mathbf{X} = x_{kl}$
- $x_{kl} = \begin{cases} \pi_l & \text{si } l \text{ est une unité voisine de } k \\ 0 & \text{sinon} \end{cases}$

$$\pi_k = 0.2$$

0	0	0.2	0	0
0	0.2	0.2	0.2	0
0	0	0.2	0	0
0	0	0	0	0
0	0	0	0	0

$$\pi_k = 0.2$$

0.2	0.2	0	0	0.2
0.2			0	0
0	0	0	0	0
0	0	0	0	0
0.2	0	0	0	0.2

- Idée: Dans chaque strate, une seule unité est sélectionnée.
- $N$  strates.

$D =$ 

0	1	4	4	1	1	2	5	5	2	4	5	8	8	5	4	5	8	8	5	1	2	5	5	2
1	0	1	4	4	2	1	2	5	5	5	4	5	8	8	5	4	5	8	8	2	1	2	5	5
4	1	0	1	4	5	2	1	2	5	8	5	4	5	8	8	5	4	5	8	5	2	1	2	5
4	4	1	0	1	5	5	2	1	2	8	8	5	4	5	8	8	5	4	5	5	5	2	1	2
1	4	4	1	0	2	5	5	2	1	5	8	8	5	4	5	8	8	5	4	2	5	5	2	1
1	2	5	5	2	0	1	4	4	1	1	2	5	5	2	4	5	8	8	5	4	5	8	8	5
2	1	2	5	5	1	0	1	4	4	2	1	2	5	5	5	4	5	8	8	5	4	5	8	8
5	2	1	2	5	4	1	0	1	4	5	2	1	2	5	8	5	4	5	8	8	5	4	5	8
5	5	2	1	2	4	4	1	0	1	5	5	2	1	2	8	8	5	4	5	8	8	5	4	5
2	5	5	2	1	1	4	4	1	0	2	5	5	2	1	5	8	8	5	4	5	8	8	5	4
4	5	8	8	5	1	2	5	5	2	0	1	4	4	1	1	2	5	5	2	4	5	8	8	5
5	4	5	8	8	2	1	2	5	5	1	0	1	4	4	2	1	2	5	5	5	4	5	8	8
8	5	4	5	8	5	2	1	2	5	4	1	0	1	4	5	2	1	2	5	8	5	4	5	8
8	8	5	4	5	5	5	2	1	2	4	4	1	0	1	5	5	2	1	2	8	8	5	4	5
5	8	8	5	4	2	5	5	2	1	1	4	4	1	0	2	5	5	2	1	5	8	8	5	4
4	5	8	8	5	4	5	8	8	5	1	2	5	5	2	0	1	4	4	1	1	2	5	5	2
5	4	5	8	8	5	4	5	8	8	2	1	2	5	5	1	0	1	4	4	2	1	2	5	5
8	5	4	5	8	8	5	4	5	8	5	2	1	2	5	4	1	0	1	4	5	2	1	2	5
8	8	5	4	5	8	8	5	4	5	5	5	2	1	2	4	4	1	0	1	5	5	2	1	2
5	8	8	5	4	5	8	8	5	4	2	5	5	2	1	1	4	4	1	0	2	5	5	2	1
1	2	5	5	2	4	5	8	8	5	4	5	8	8	5	1	2	5	5	2	0	1	4	4	1
2	1	2	5	5	5	4	5	8	8	5	4	5	8	8	2	1	2	5	5	1	0	1	4	4
5	2	1	2	5	8	5	4	5	8	8	5	4	5	8	5	2	1	2	5	4	1	0	1	4
5	5	2	1	2	8	8	5	4	5	8	8	5	4	5	5	5	2	1	2	4	4	1	0	1
2	5	5	2	1	5	8	8	5	4	5	8	8	5	4	2	5	5	2	1	1	4	4	1	0

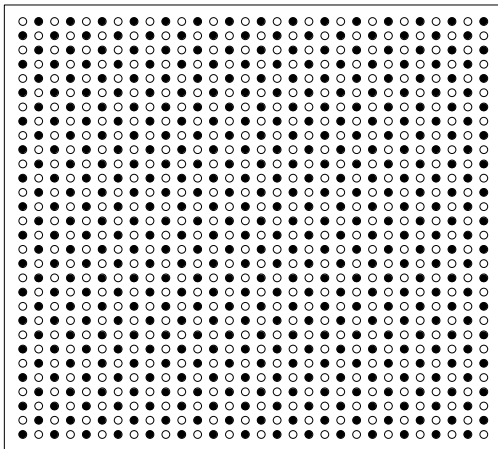




- La méthode du cube Deville & Tillé (2004) peut sélectionner un échantillon avec des strates qui se chevauchent.
- Problème: la matrice  $\mathbf{X}$  est carrée.
- La méthode consiste à chercher un vecteur dans le noyau de  $\mathbf{X}^\top \text{diag}(\mathbf{1}/\boldsymbol{\pi})$ .
- Heureusement, la matrice  $\mathbf{X}$  n'est pas de plein rang dans certains cas.
- Grille  $c_1 r \times c_2 r$ , où  $r = 1/\pi_k$  est constant et entier.

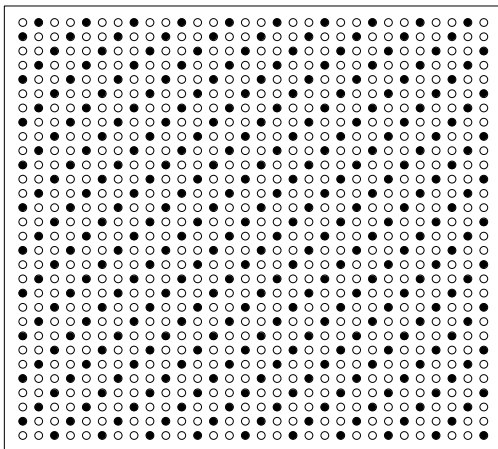
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/2$$



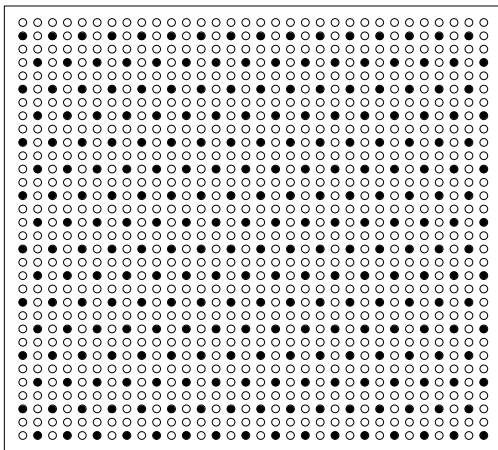
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/3$$



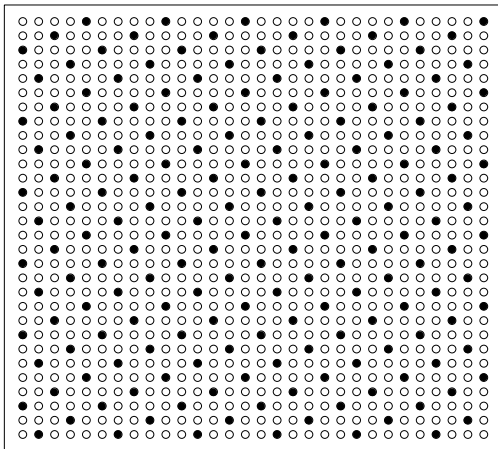
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/4$$



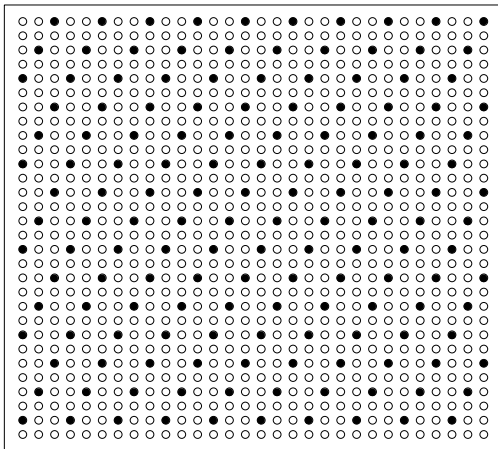
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/5$$



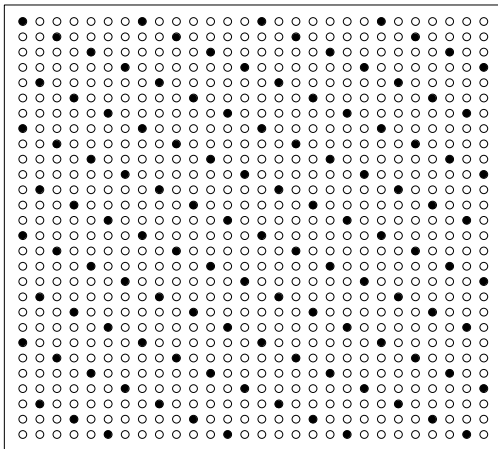
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/6$$



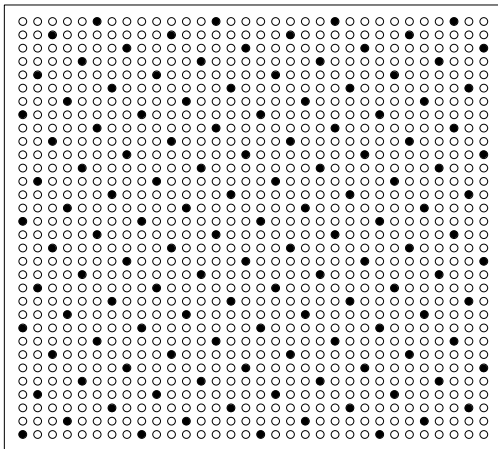
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/7$$



## Exemple d'échantillon périodique dans une grille

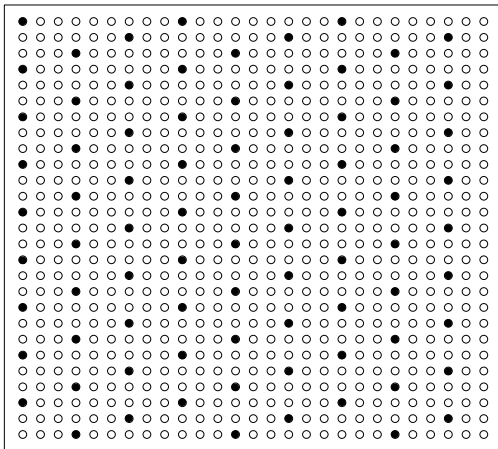
$$\pi_k = 1/8$$





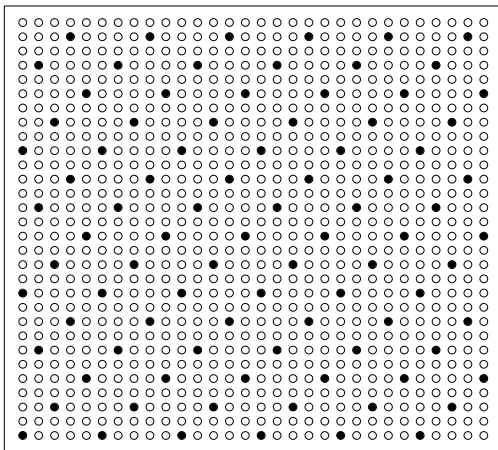
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/9$$



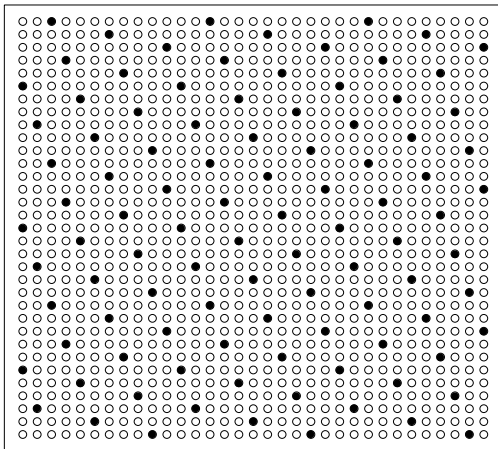
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/10$$



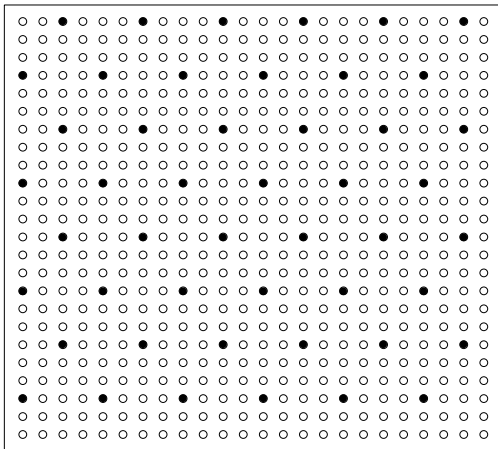
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/11$$



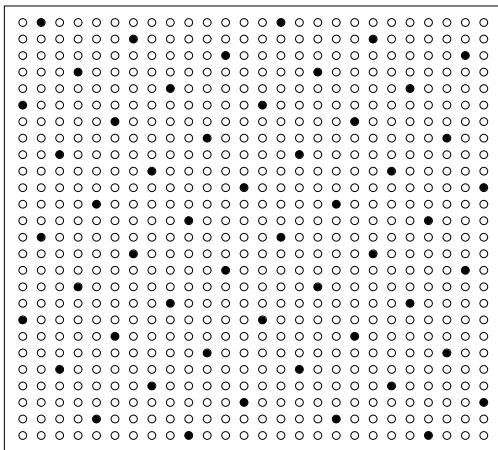
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/12$$



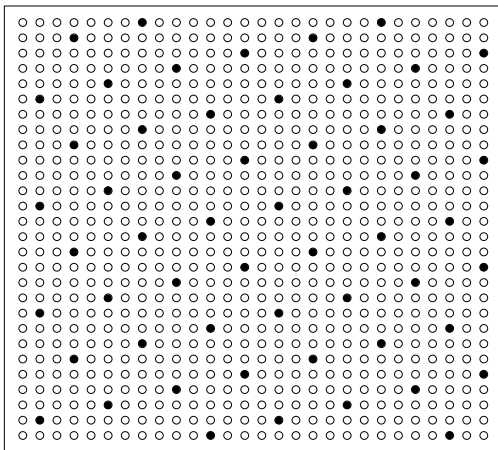
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/13$$



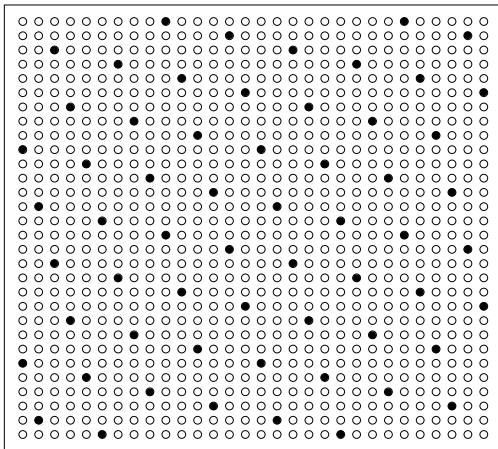
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/14$$



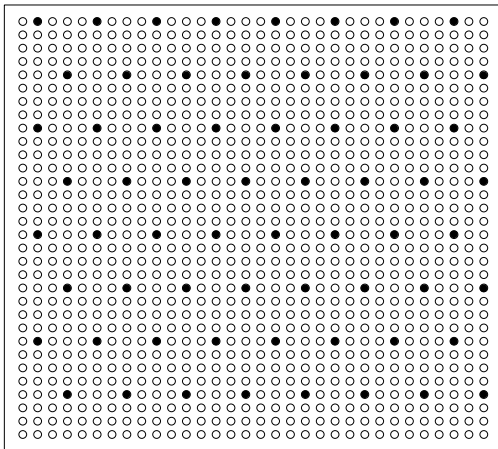
## Exemple d'échantillon périodique dans une grille

$$\pi_k = 1/15$$



## Exemple d'échantillon périodique dans une grille

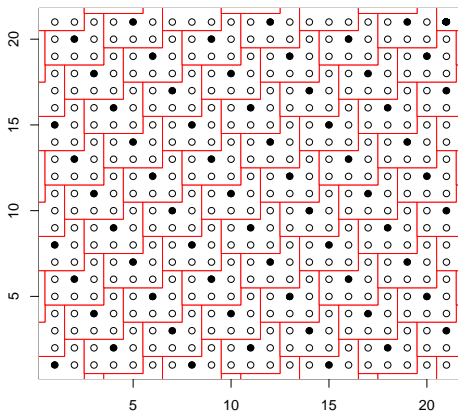
$$\pi_k = 1/16$$





# Tessellation

## Tessellation 1/7



## Etape de l'algorithme

- 1 Soit  $\tilde{\pi}$  le vecteur des probabilités d'inclusion restreints aux  $k$  tels que  $0 < \pi_k < 1$ . Soit  $\tilde{\mathbf{X}}$  la matrice  $\mathbf{X}$  avec seulement les lignes des  $k$  tels que  $0 < \pi_k < 1$ .
- 2 On construit la matrice  $\tilde{\mathbf{A}} = (\tilde{\mathbf{X}}^\top \text{diag}(\tilde{\pi})^{-1})$ .
  - 1 Si  $\tilde{\mathbf{A}}$  n'est pas de plein rang,  $\mathbf{u} = (u_1, \dots, u_k, \dots, u_N)^\top \in \mathbb{R}^N$  est un vecteur dans le noyau de  $\tilde{\mathbf{A}}^\top$ , i.e.  $\tilde{\mathbf{A}}^\top \mathbf{u} = \mathbf{0}$ .
  - 2 Si  $\tilde{\mathbf{A}}$  est de plein rang,  $\mathbf{u}$  est le vecteur propre droit associé à la plus petite valeur singulière de  $\mathbf{A}$ .
- 3 On identifie  $\lambda_1$  et  $\lambda_2$  les plus grandes valeurs telles que  $0 \leq \tilde{\pi}_k + \lambda_1 u_k \leq 1$  et  $0 \leq \tilde{\pi}_k - \lambda_2 u_k \leq 1$  pour tout  $k$  tels que  $0 < \pi_k < 1$ .
- 4 On calcule

$$\pi^* = \begin{cases} \tilde{\pi} + \lambda_1 \mathbf{u} & \text{avec la probabilité } \lambda_2 / (\lambda_1 + \lambda_2) \\ \tilde{\pi} - \lambda_2 \mathbf{u} & \text{avec la probabilité } \lambda_1 / (\lambda_1 + \lambda_2). \end{cases}$$

- 5 On remplace dans  $\pi$  les valeurs correspondantes de  $\pi^*$ .

# Conclusions

- Possibilité d'obtenir les échantillons les plus équilibrés.
- Computer intensive.
- Possibilité d'utiliser des matrices "sparse" dans R.

# Bibliography I

- Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli* 18, 1099–1471.
- Deville, J.-C. (1998). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Tech. Rep. 9804, Méthodologie Statistique, Insee.
- 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.
- Dickson, M. M. & Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics* 31, 1359–1372.
- Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society* B32, 209–226.
- 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.
- Grafström, A. & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 14, 120–131.
- Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* 20, 333–354.
- Pea, J., Qualité, L. & Tillé, Y. (2007). Systematic sampling is a minimal support design. *Computational Statistics & Data Analysis* 51, 5591–5602.
- Stevens Jr., D. L. & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14, 593–610.

# Bibliography II

- 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. E., 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. (2018). Fast implementation of Fuller's unequal probability sampling method. Tech. rep., University of Neuchâtel.
- Tillé, Y., Dickson, M. M., Espa, G. & Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on the Moran's  $I$  index. *Spatial Statistics* 23, 182–192.
- Tillé, Y. & Wilhelm, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science* 32, 176–189.
- Wang, J.-F., Stein, A., Gao, B.-B. & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics* 2, 1 – 14.