UNIVERSITÉ DE NEUCHÂTEL

Spatially Balanced Sampling: State of the Art

Yves Tillé
University of Neuchâtel

Seminar 2018
II Congress of Polish Statistics
100th anniversary of Polish Statistics

Table of contents

Notation

Notation

- Paper: Tillé & Wilhelm (2017).
- Population: $U = \{1, \ldots, k, \ldots, N\}$.
- Sample $s \subset U$.
  Example $U = \{1, 2, 3, 4, 5\}$, sample $s = \{2, 3, 5\}$ other notation $\mathbf{s} = (0, 1, 1, 0, 1)^\top$.
- Sampling design $p(s) \geq 0$ and $\sum_{s \subset U} p(s) = 1$.
- Random sample S, $\Pr(S = s) = p(s)$, for all $s \subset U$.
- Inclusion probabilities $\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s)$.
- Joint inclusion probabilities $\pi_{k\ell} = \Pr(\{k, \ell\} \in S) = \sum_{s \supset \{k,\ell\}} p(s)$.
- Total $Y = \sum_{k \in U} y_k$. Mean $\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k$.

Notation

- Narain-Horvitz-Thompson (NHT) estimator: $\widehat{\overline{Y}} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$.
- $\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell \\ \pi_k(1 - \pi_k) & \text{if } k = \ell. \end{cases}$
- Variance of the SHT-estimator is equal to:

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

$$\text{var}_p\left(\widehat{\overline{Y}}\right) = -\frac{1}{2N^2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell} \text{ (for fixed sample size).}$$

Survey sampling theory is not witchcraft

Survey sampling theory is not witchcraft

- If you do not like your boss you can make a small doll with his effigy.
- You push needles.
- It works because the doll looks like your boss.

Survey sampling theory is not witchcraft

- One can select units with unequal inclusion probabilities.
- Representativeness means that the sample is a reduced model of the population.
- Representativeness is not a scientific argument to justify estimation.
- Representativeness is only an argument to justify witchcraft.

Example

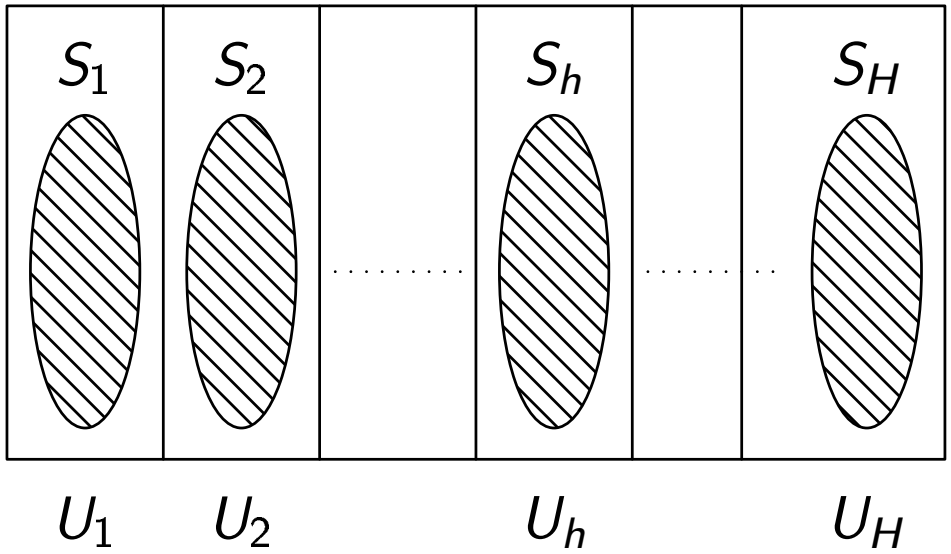


Figure: Stratified design: the samples are independent

Example

- Jerzy Neyman (1934) (1894-1981) defined the optimal stratification.

Example

- The more dispersed strata must be oversampled to reduce the variance.
- In all the business surveys, the big companies are selected with larger inclusion probabilities.
- Weighting by the inverse of the inclusion probabilities enables to have an unbiased estimation.
- Generalization : unequal probability sampling.
- Never use the word “representative”.
- Use the word coverage! If some π_k are null, there is a coverage problem.

Balanced sampling

- With equal inclusion probabilities, a sample is said to be balanced on p auxiliary variables $x_1, \ldots, x_p$ known on the population if

$$\frac{1}{n}\sum_{k\in S} x_{kj} = \frac{1}{N}\sum_{k\in U} x_{kj}, j = 1, \ldots, p.$$

- Generalisation with unequal inclusion probabilities.

$$\sum_{k\in S} \frac{x_{kj}}{\pi_k} = \sum_{k\in U} x_{kj}, j = 1, \ldots, p.$$

- Deville & Tillé (2004) proposed the cube method to select samples that are almost balanced.

Example: 245 municipalities of the Swiss Ticino canton

Table: Balancing variables of the population of municipalities of Ticino

POP	number of men and women
ONE	constant variable that takes always the value 1
ARE	area of the municipality in hectares
POM	number of men
POW	number of women
P00	number of men and women aged between 0 and 20
P20	number of men and women aged between 20 and 40
P40	number of men and women aged between 40 and 65
P65	number of men and women aged between 65 and over
HOU	number of households

Example: sampling design

- Inclusion probabilities proportional to size.
- Big municipalities are always in the sample Lugano, Bellinzona, Locarno, Chiasso, Pregassona, Giubiasco, Minusio, Losone, Viganello, Biasca, Mendrisio, Massagno.
- Sample size = 50.
- the population totals for each variable X_j,
- the estimated total by the Horvitz-Thompson estimator $\widehat{X}_{j\pi}$,
- the relative deviation in % defined by

$$\text{RD} = 100 \times \frac{\widehat{X}_{j\pi} - X_j}{X_j}.$$

Example: Results

Table: Quality of balancing

Variable	Population total	HT-Estimator	Relative deviation in %
POP	306846	306846.0	0.00
ONE	245	248.6	1.49
HA	273758	276603.1	1.04
POM	146216	146218.9	0.00
POW	160630	160627.1	-0.00
P00	60886	60653.1	-0.38
P20	86908	87075.3	0.19
P40	104292	104084.9	-0.20
P65	54760	55032.6	0.50
HOU	134916	135396.6	0.36

Model for spatial sampling

- Model for spatial sampling

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

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

- The model thus admits heteroscedasticity and autocorrelation.

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

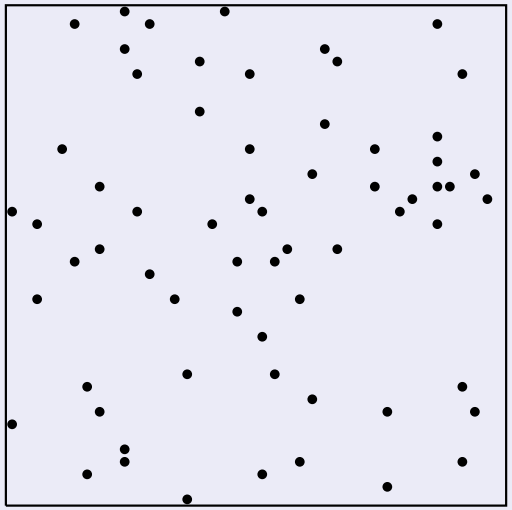
Optimal design:

- using inclusion probabilities proportional to $\sigma_{\varepsilon k}$,
- using a balanced sampling design on the auxiliary variables $\mathbf{x}_k$.
- avoiding the selection of neighboring units, that is, selecting a well-spread sample (or spatially balanced)

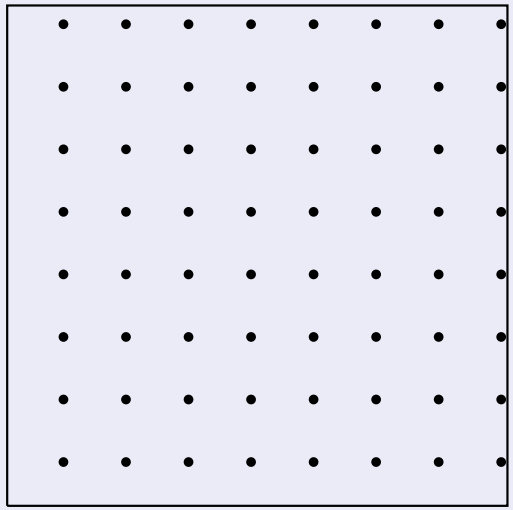
Usual methods

- Usual methods can be used: simple, stratified, cluster, two-stage sampling.
- Stratification can improve the spreading.
- Central role of systematic sampling (because spread).

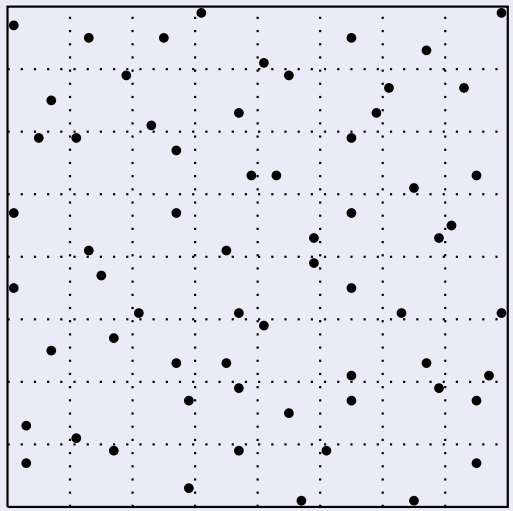
Usual methods



simple design



systematic



stratification

Biodiversity Monitoring

The most spread sampling design is the two-dimensional systematic sampling.

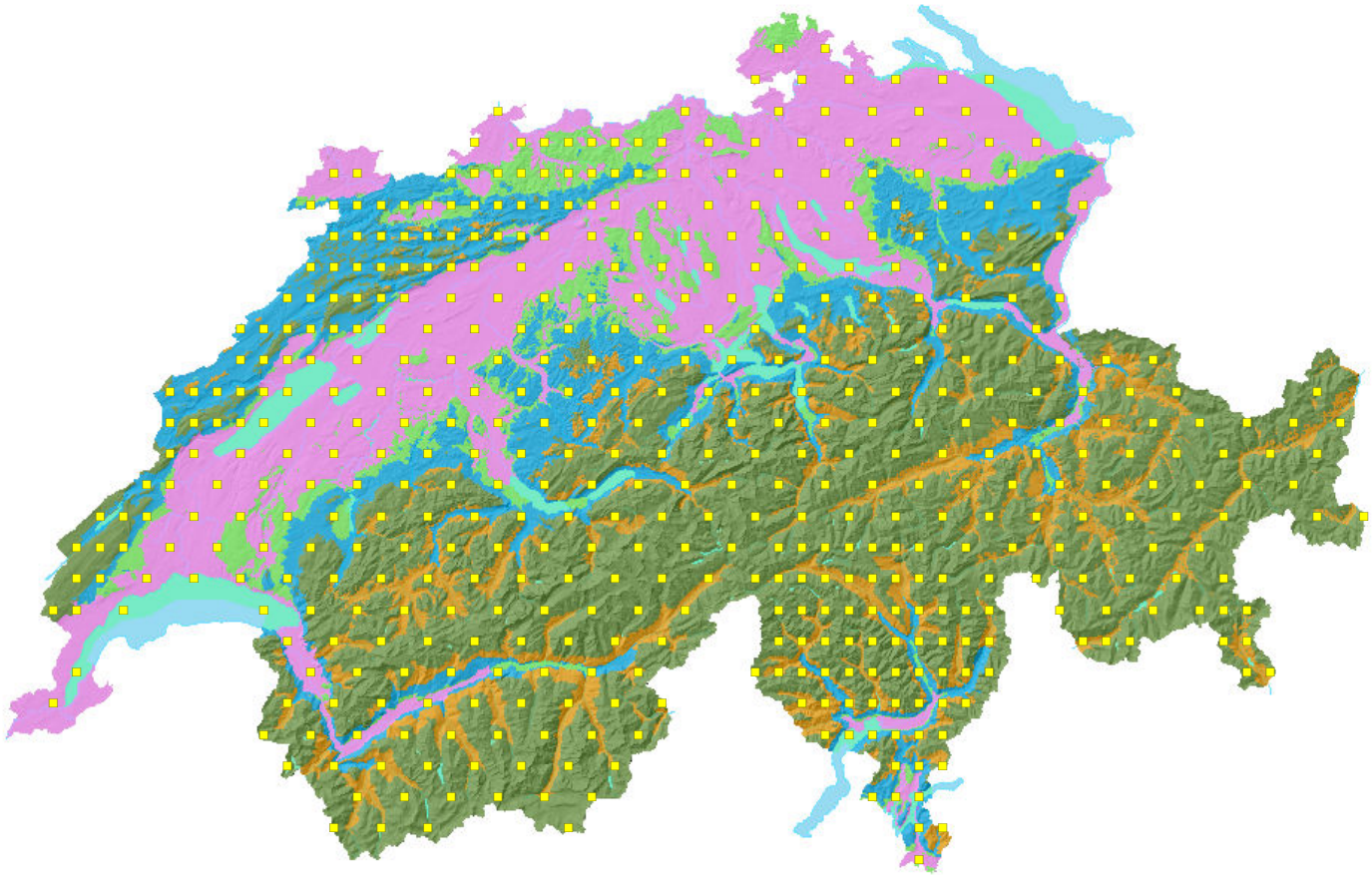


Table: WSL Swiss biodiversity Monitoring

Systematic sampling 1

Systematic sampling

- Cumulated inclusion probabilities

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

- u a uniform random number in $[0, 1]$.
- Units such that $\lfloor V_k - u \rfloor \neq \lfloor V_{k-1} - u \rfloor$ are selected in the sample. (Madow, 1949)
- Minimum entropy (Pea, Qualité & Tillé, 2007).

Systematic sampling 2

Systematic sampling

Example

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

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	

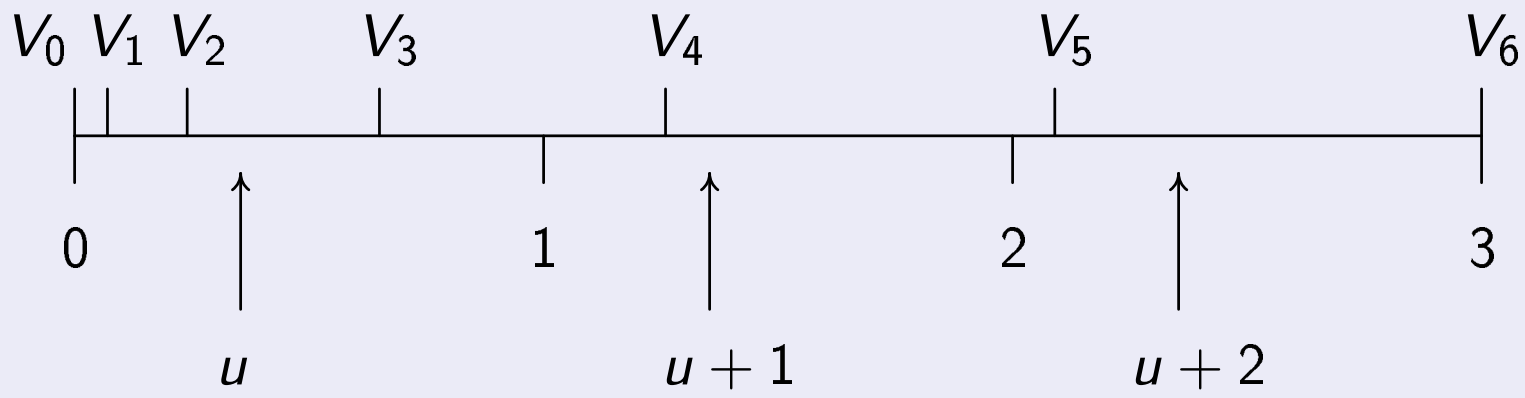
Systematic sampling 3

Systematic sampling

Suppose also that the value taken by the uniform random number is $u = 0.354$. The rules of selections are:

- Because $V_2 \leq u < V_3$, unit 3 is selected;
- Because $V_4 \leq u < V_5$, unit 5 is selected;
- Because $V_5 \leq u < V_6$, unit 6 is selected.

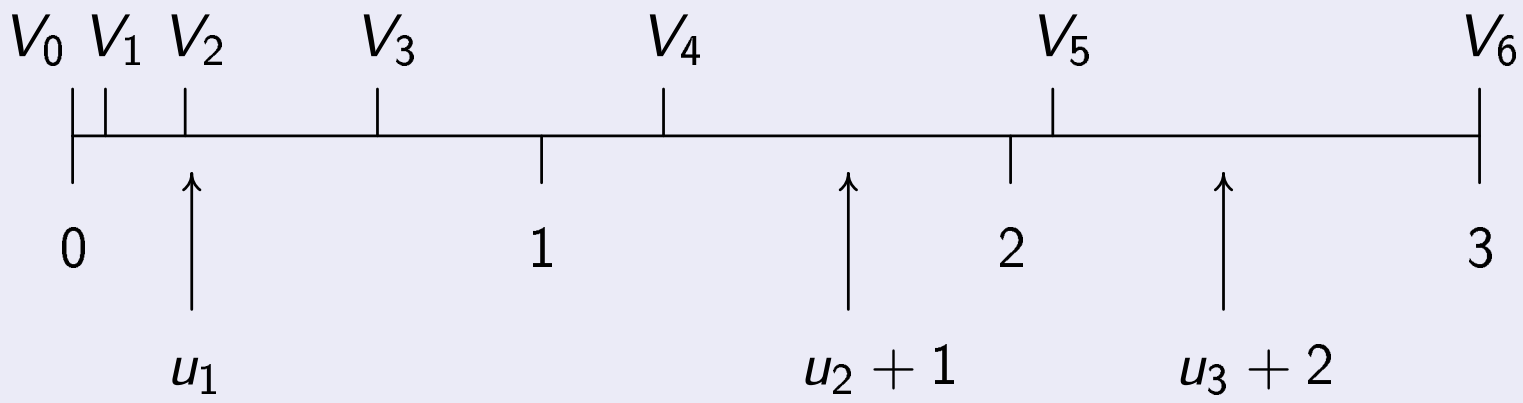
The sample selected is thus $\mathbf{s} = (0, 0, 1, 0, 1, 1)$.



Deville Systematic sampling

Deville (1998) Systematic sampling

For each interval of length 1, a uniform random variable is generated.



A dependency is introduces between u_1, u_2, and u_3 in order to not select twice the same unit.

Deville Systematic sampling

Deville Systematic sampling

- If frontier unit ℓ is selected at step $i-1$, u_i has the density function:

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

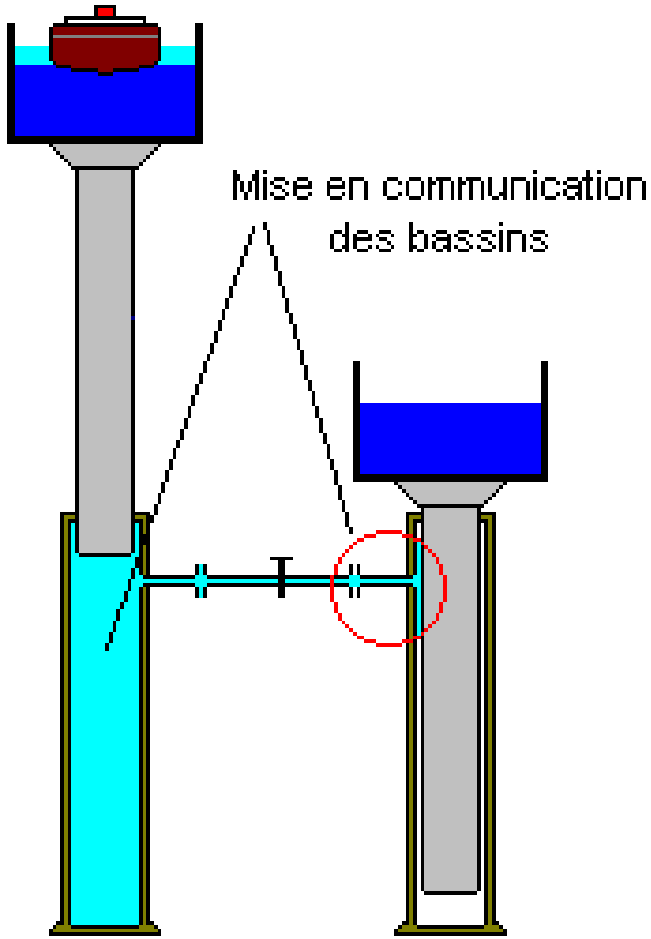
- If ℓ is not selected at step $i-1$, u_i has the density function:

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

Pivotal method

Pivotal method

from Michel Maigre©, web site of Région Wallone: Direction des voies hydrauliques, canal du centre.



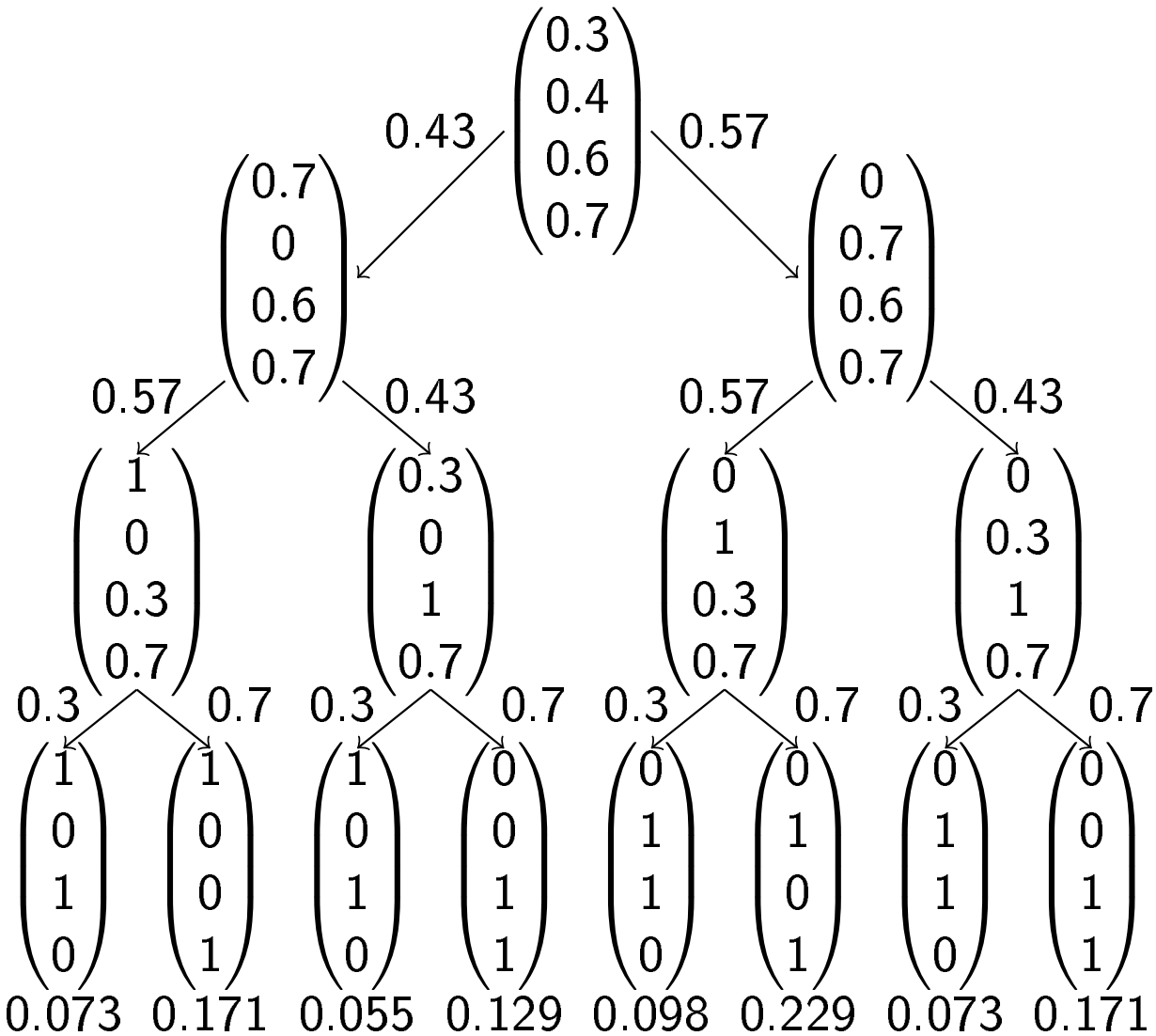
Pivotal method

- Pivotal method (Deville & Tillé, 2000).
- At each step, two inclusion probabilities (i are j) are randomly modified.
- Example

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

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

Pivotal method



Pivotal method

- Pivotal method (Deville & Tillé, 2000).
- Pick at each step two units (denoted by i and j) in the population.
- Two cases: If $\pi_i + \pi_j > 1$, then

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

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

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

Pivotal methods

Variants

- Ordered pivotal method or sequential pivotal method or Deville systematic sampling
  (Deville, 1998),
- Random pivotal method
  Deville & Tillé (1998),
- Local pivotal method or spatial pivotal method.
  (Grafström, Lundström & Schelin, 2012).

Pivotal methods

Variants

- Chauvet (2012) showed that ordered pivotal method is the same as Deville Systematic sampling.
- Fuller (1970) has proposed a method that is very similar to the ordered pivotal method.
- Tillé (2018) has proposed a simple implementation with a phantom unit and has generalised Fuller's method.

Problems

Systematic sampling cannot be used

1. when the inclusion probabilities are unequal,
2. when the statistical units are irregularly arranged on the territory, (ex. building, municipalities).

Centers of the Belgian Municipalities

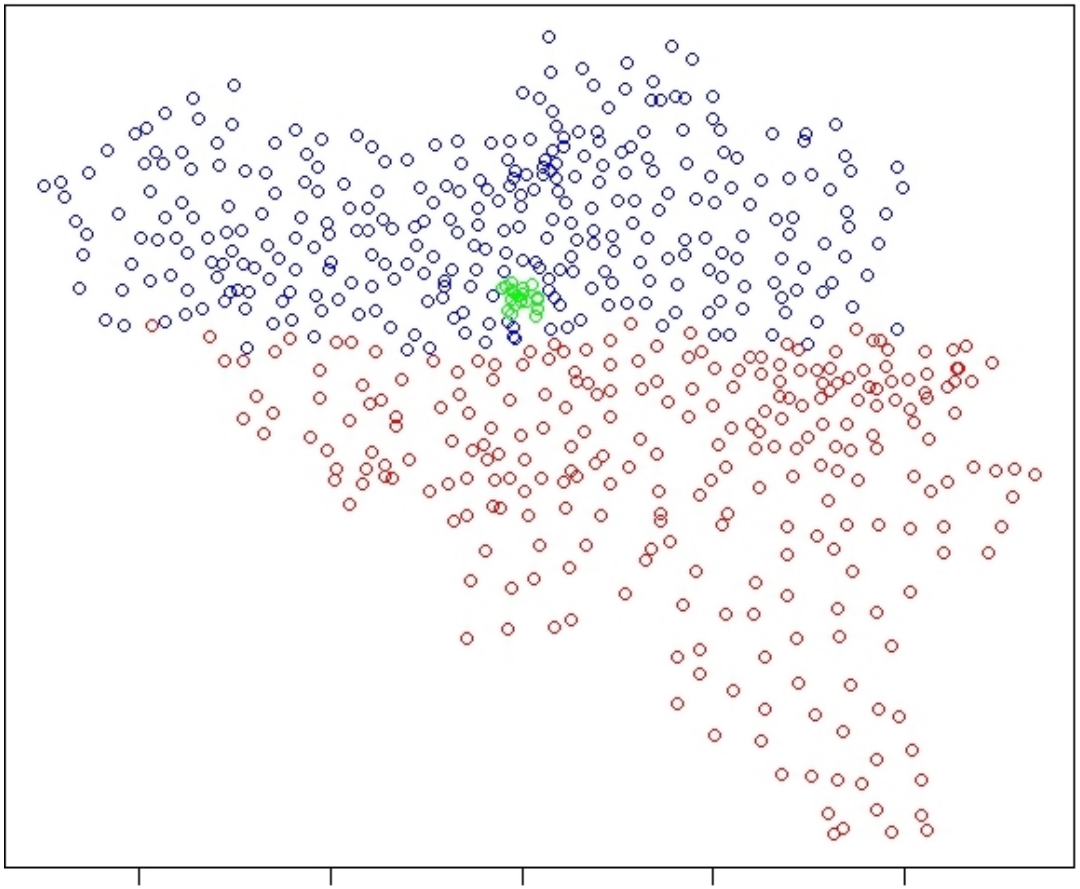


Table: Centers of the Belgian Municipalities (Data IGN Belgium)

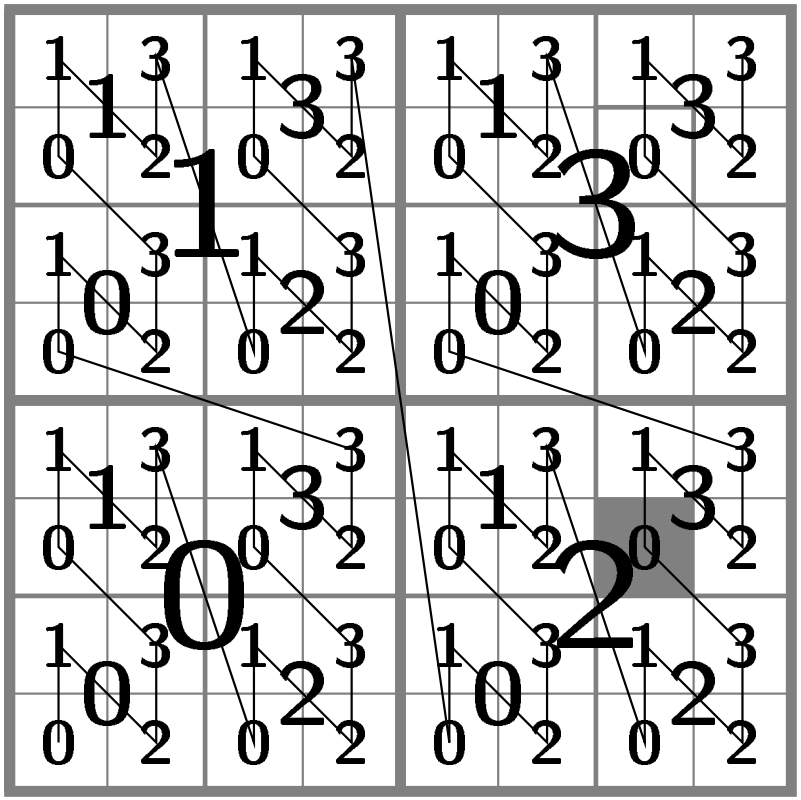
Generalized Random Tessellation Sampling

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

1. Create a hierarchical grid with addresses.
2. Randomize the addresses.
3. Construct a sampling line using the addresses
4. Select a systematic sample on the line.

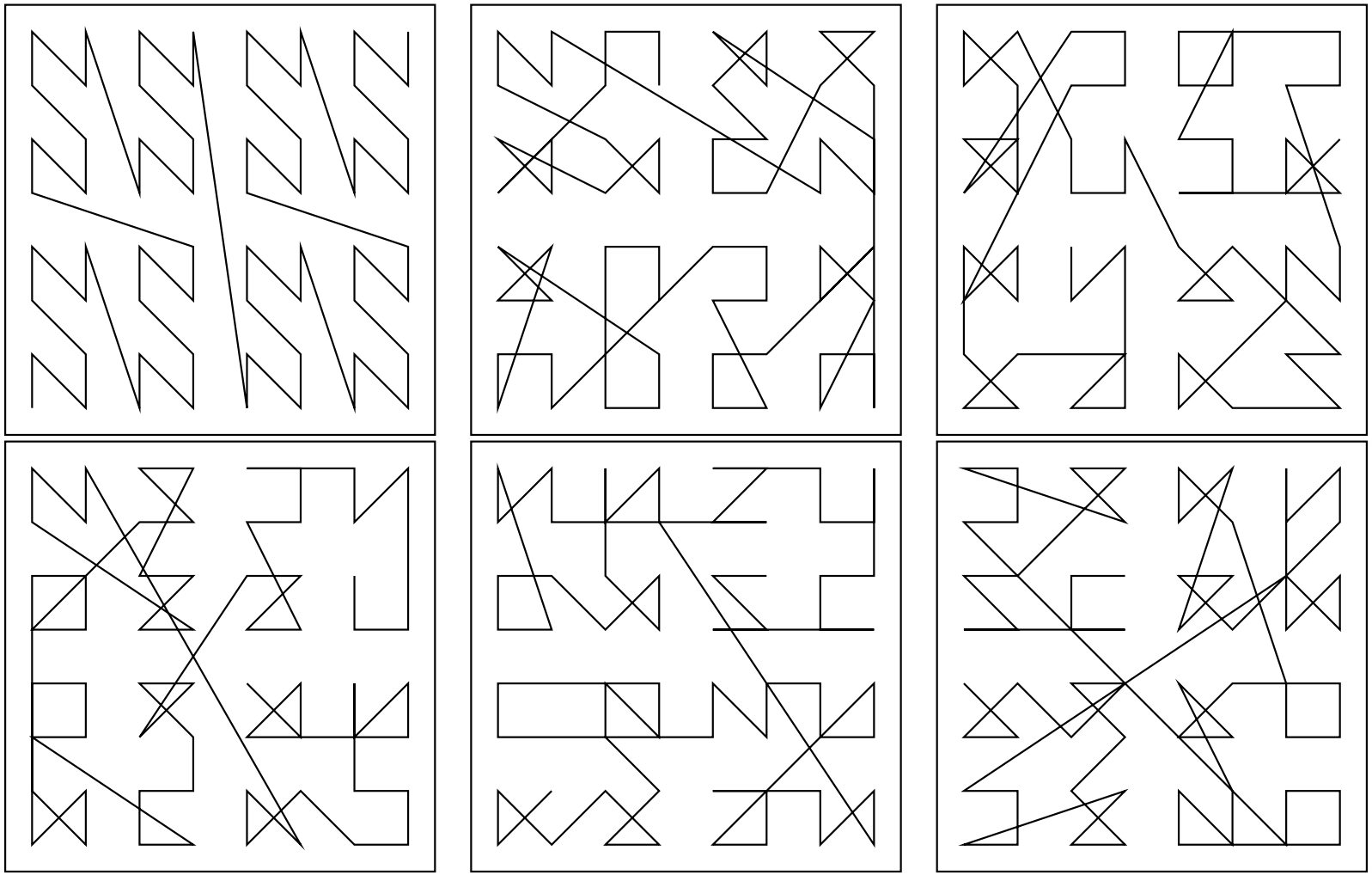
The sample is well spread, but the totals are not balanced.

Generalized Random Tessellation Sampling



The sample is well spread, but the totals are not balanced.

Generalized Random Tessellation Sampling



Travelling Salesman Problem

Autocorrelation along the path for the mean income in the municipalities: 0.4835873

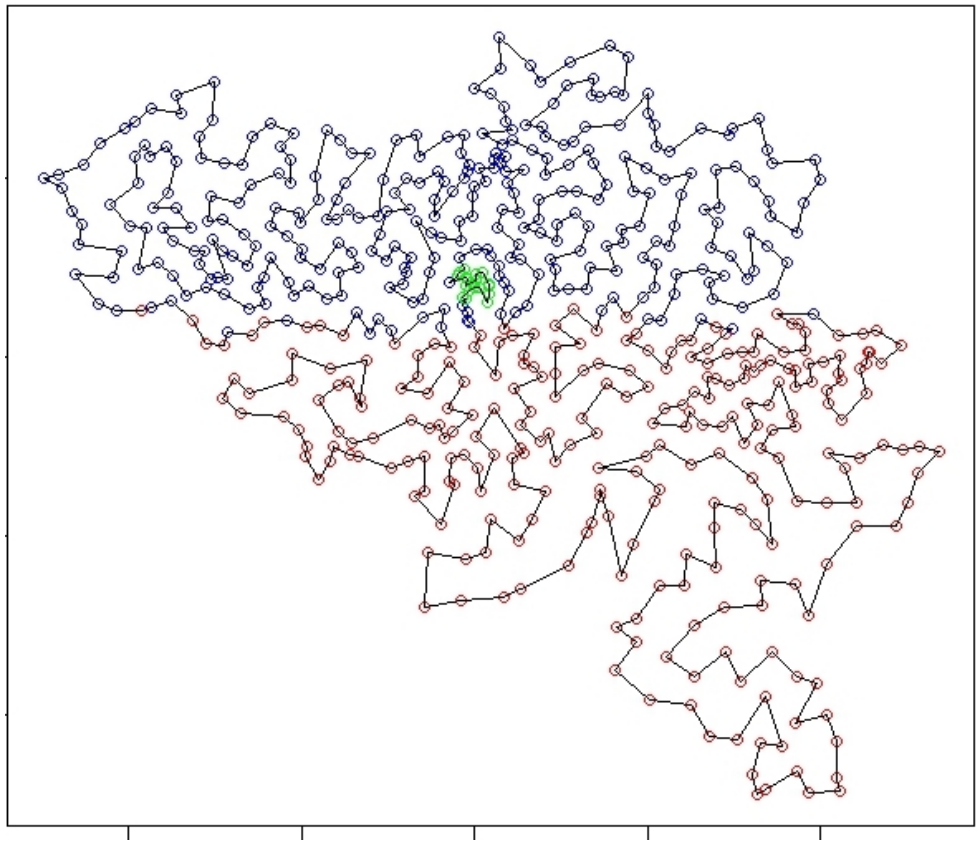
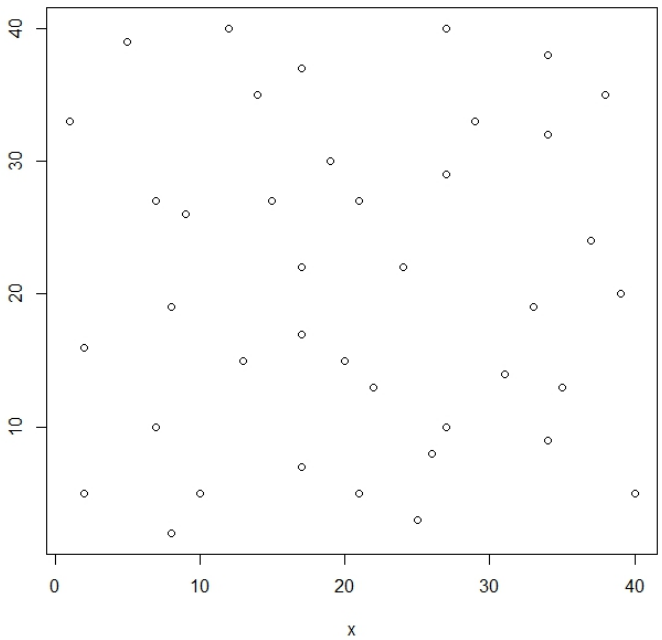
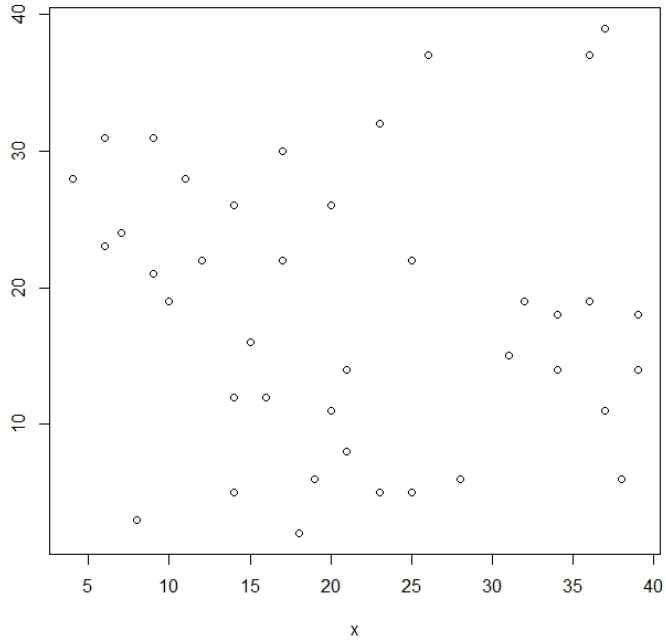


Table: Smallest path between the points. Next systematic sampling (Dickson & Tillé, 2016).

Travelling Salesman Problem and systematic sampling



Simple random sampling



The local pivotal method

Algorithm of Grafström, Lundström & Schelin (2012)

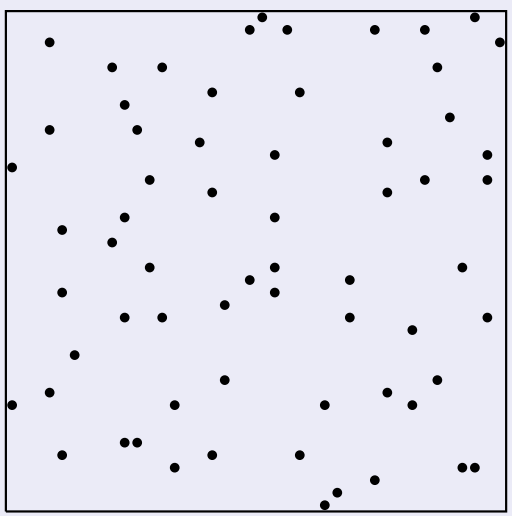
1. Choose randomly two units i and j with probabilities strictly between 0 and 1 that are spatially close.
2. Run one step of the pivotal method only on i and j.
3. Repeat these two steps.

The sample is well spread, but the totals are not balanced.

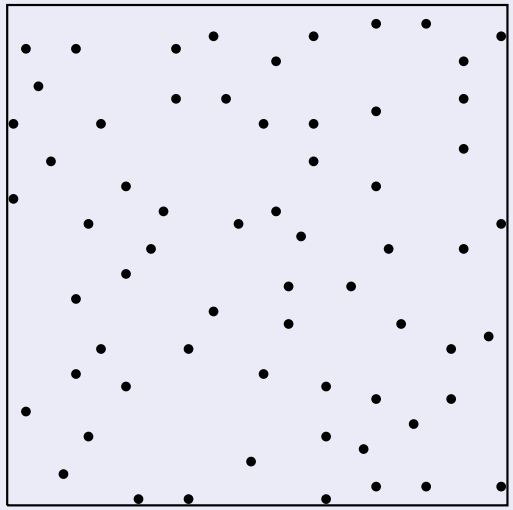
Local Cube Method (Grafström & Tillé, 2013)

- Generalization of the local pivotal.
- The sample is spread and is balanced on the auxiliary variables.

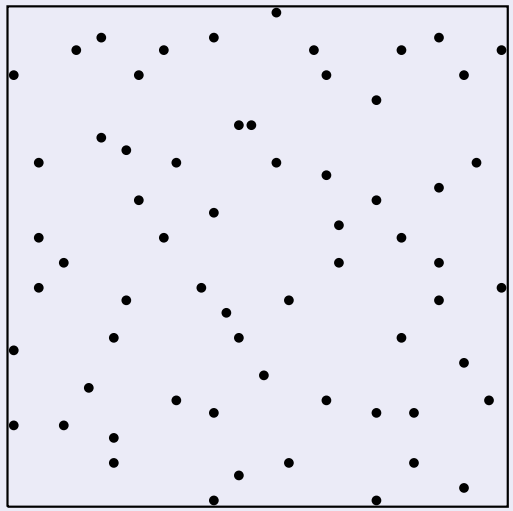
Complex methods



GRTS

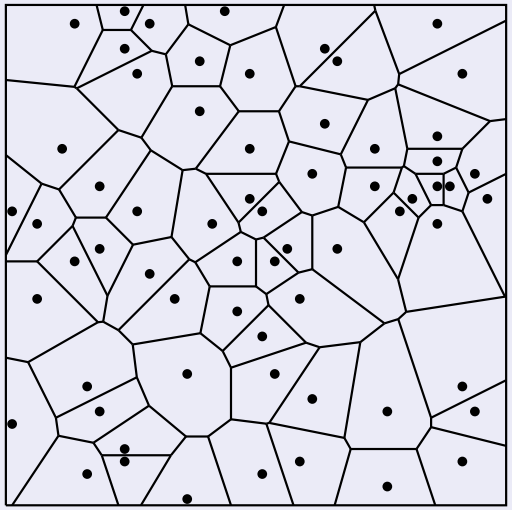


local pivotal

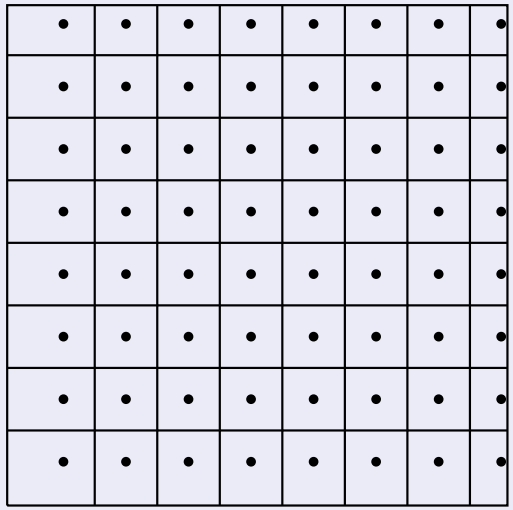


local cube

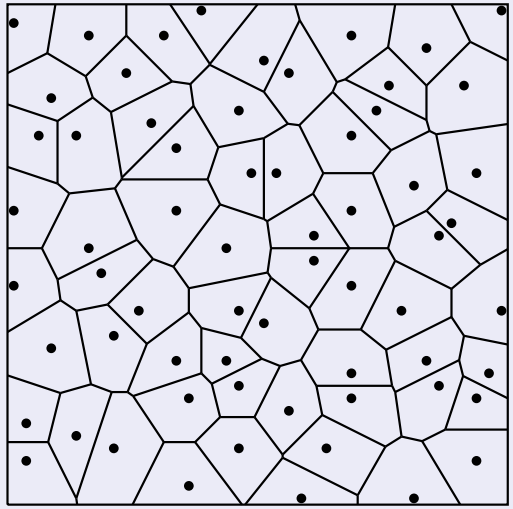
Voronoï polygons



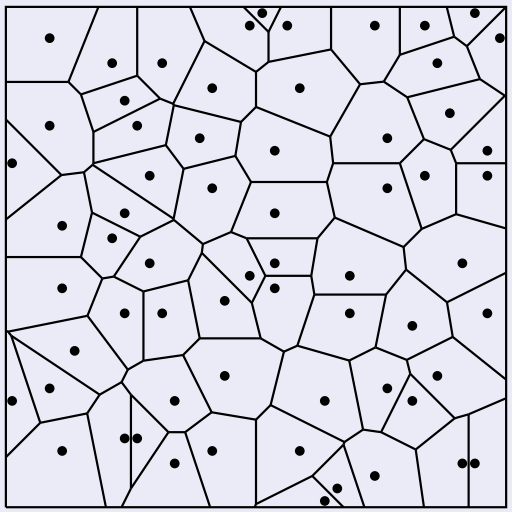
simple



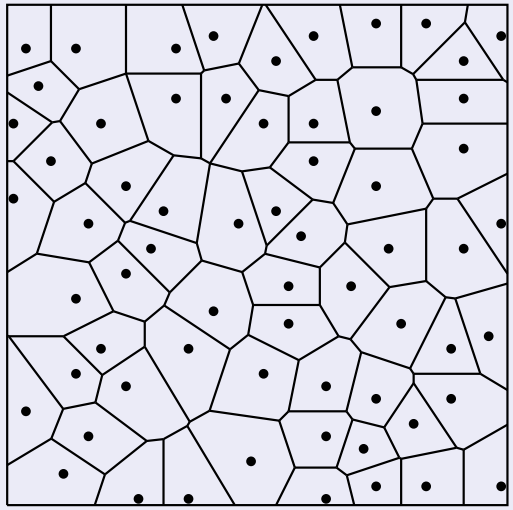
systematic



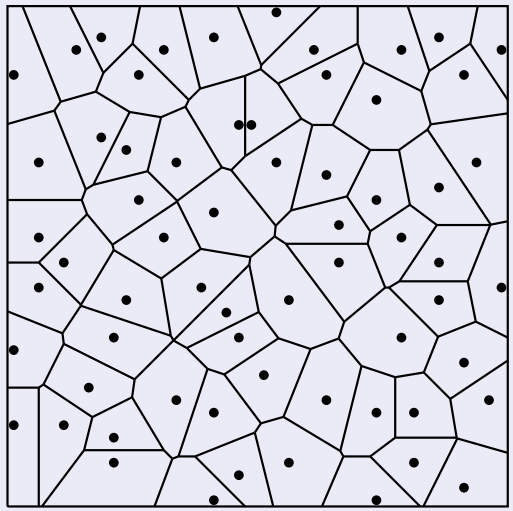
stratification



GRTS



local pivotal



local cube

Quality of balancing

Table: Indices of spatial balance for the main sampling designs (Variance of the sum of the inclusion probabilities of units in the Voronoï polygons around the selected units)-

Design	Balance indicator
Systematic	0.05
Simple random sampling	0.31
Stratification with H=25	0.11
Local pivotal	0.06
Cube method	0.21
Local Cube method	0.06
GRTS	0.09

Intermediate conclusions

Intermediate conclusions

- The most spread method is systematic sampling.
- The local pivotal method does not give the most spread method.
- Is it possible to do better? A general algorithm that gives systematic sampling in a grid with equal inclusion probabilities.

An alternate measure of spreading based on the Moran index

An alternate measure of spreading based on the Moran index

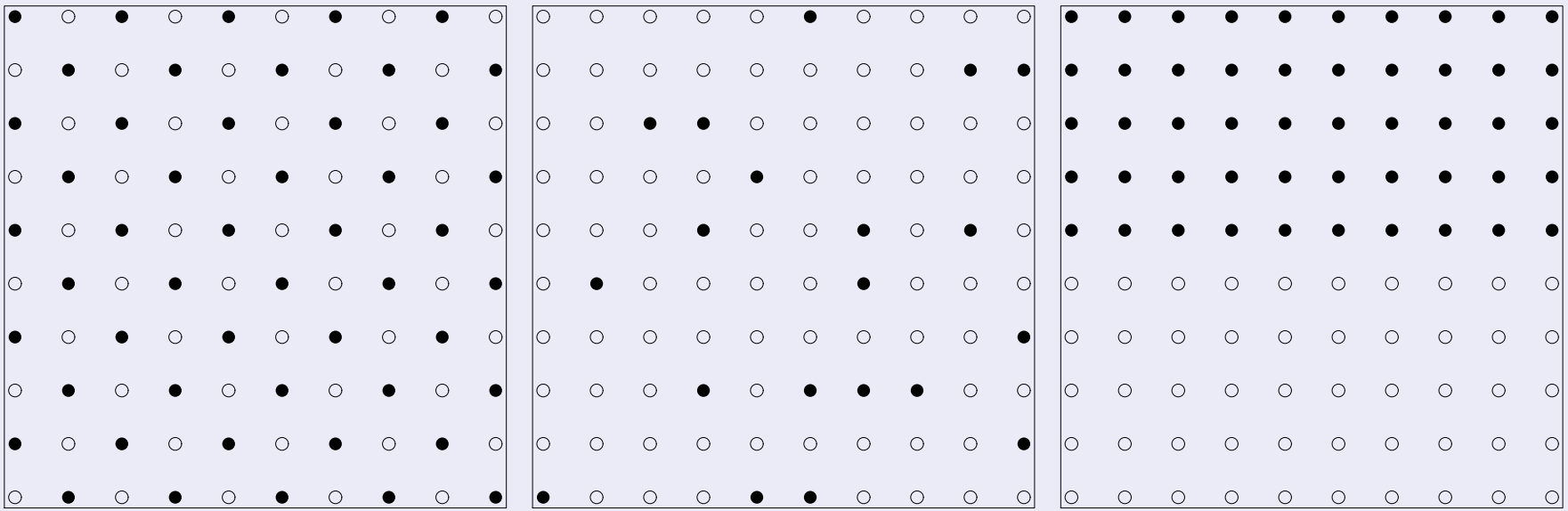
- Tillé, Dickson, Espa & Giuliani (2018).
- Correlation between:
  - the vector of indicator $\mathbf{s} = (0\ 1\ 0\ 0\ 1\ \cdots\ 0)$.
  - The local mean of this vector. The local mean of k is the mean of the $\frac{1}{\pi_k} - 1$ nearest values of k.

An alternate measure of spreading based on the Moran index

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

Correlation between the indicators of the presence of the unit in the sample and the local mean (Tillé, Dickson, Espa & Giuliani, 2018).

$I_B = -1$ $I_B = 0$ $I_B \approx 1$

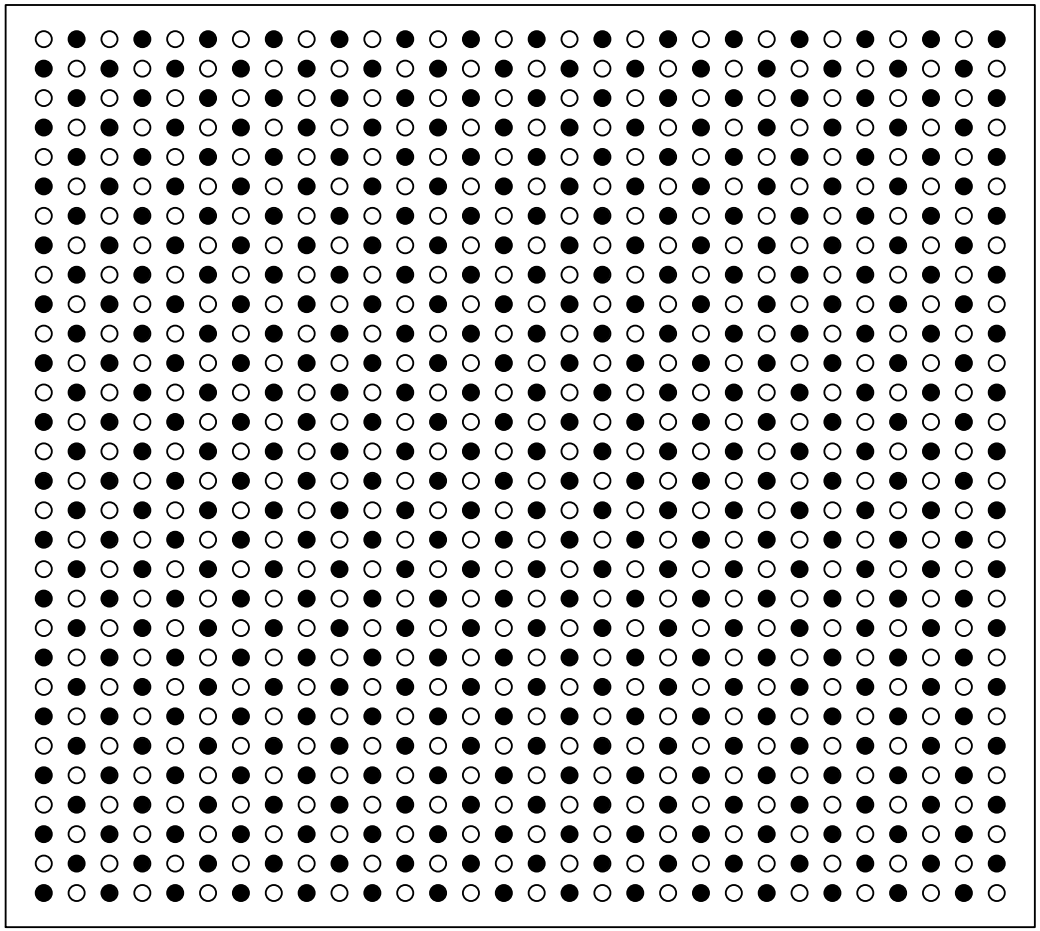


Other idea

- The cube method Deville & Tillé (2004) can select a sample with overlapping strata.
- Define one stratum for each unit.
- The stratum is the neighborhood of the units.
- Select a sample with the cube method with overlapping strata.

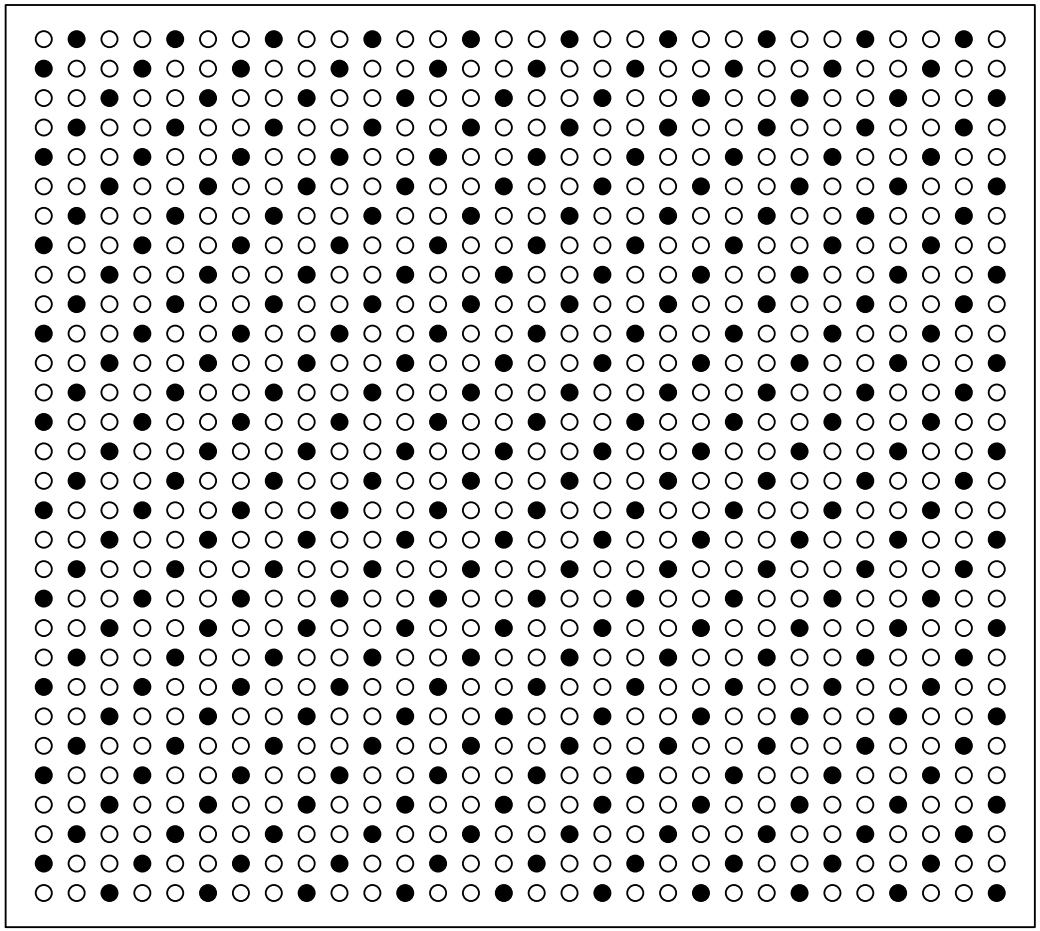
Examples of periodic sample in a grid

$$\pi_k = 1/2$$



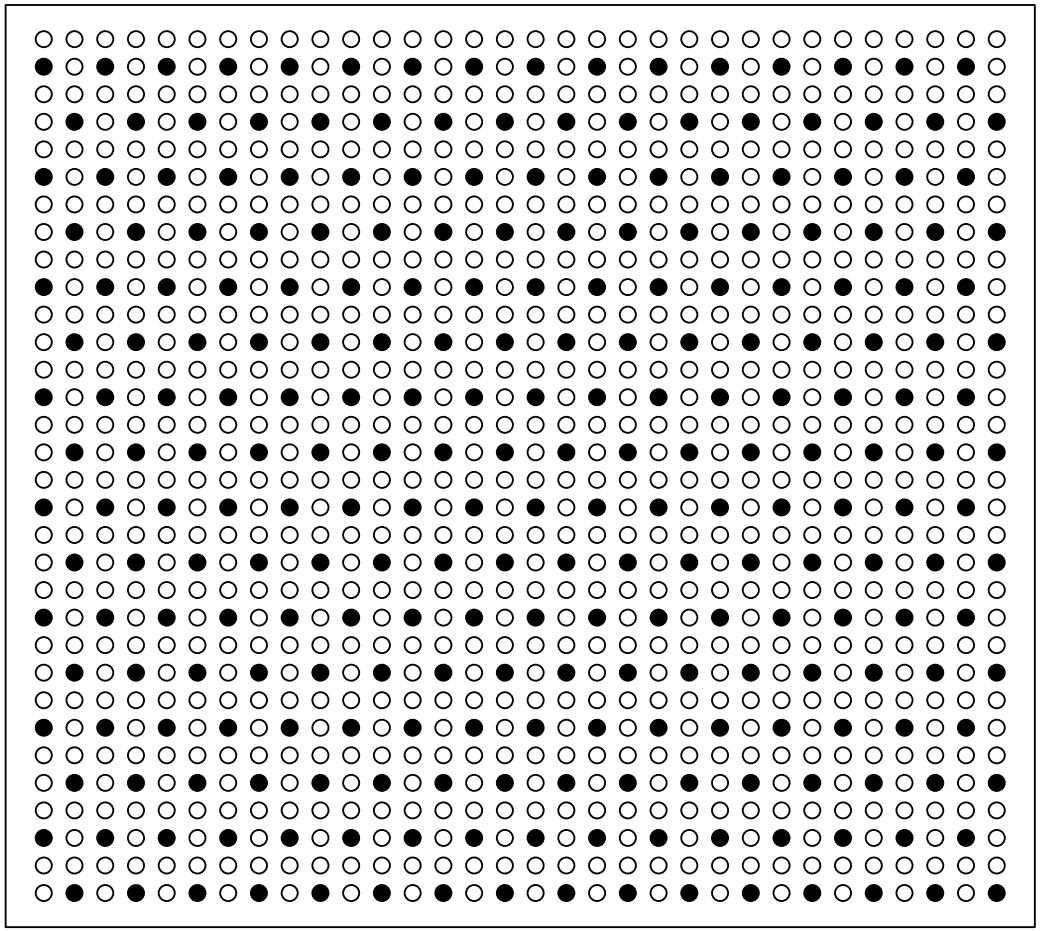
Examples of periodic sample in a grid

$$\pi_k = 1/3$$



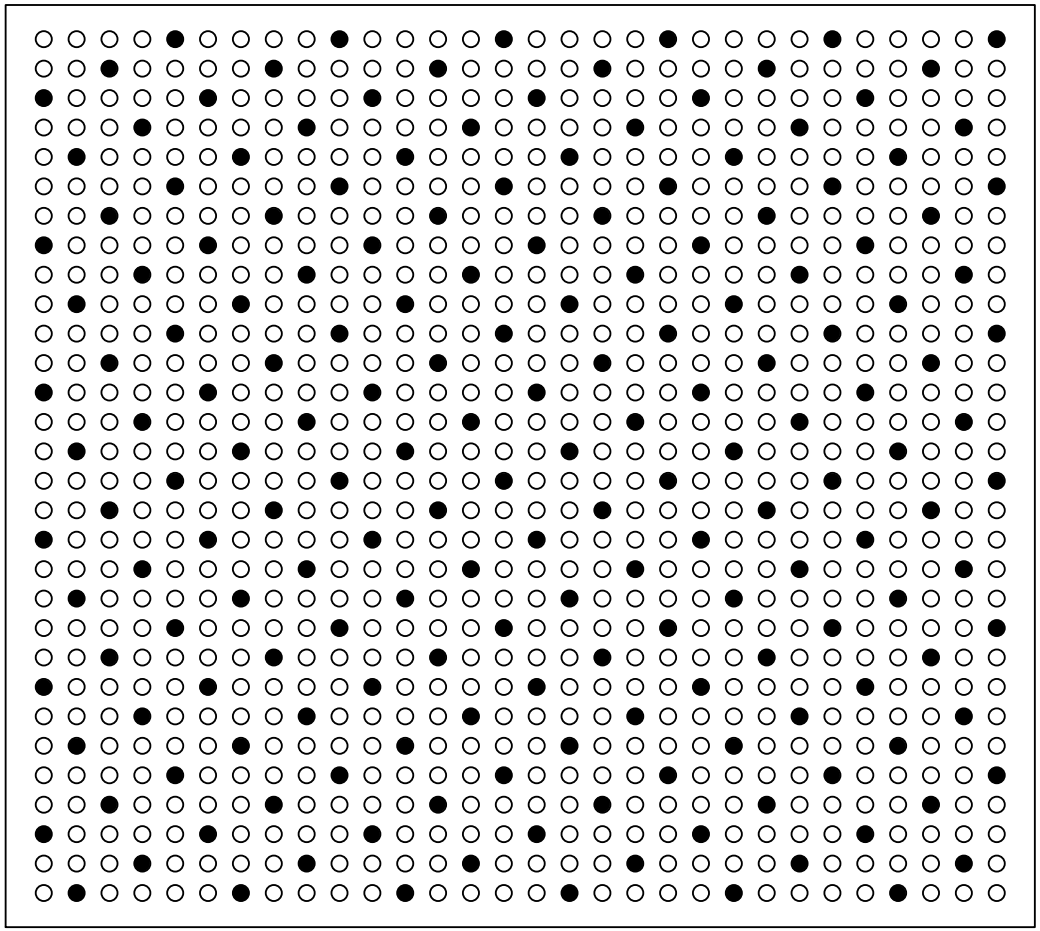
Examples of periodic sample in a grid

$$\pi_k = 1/4$$



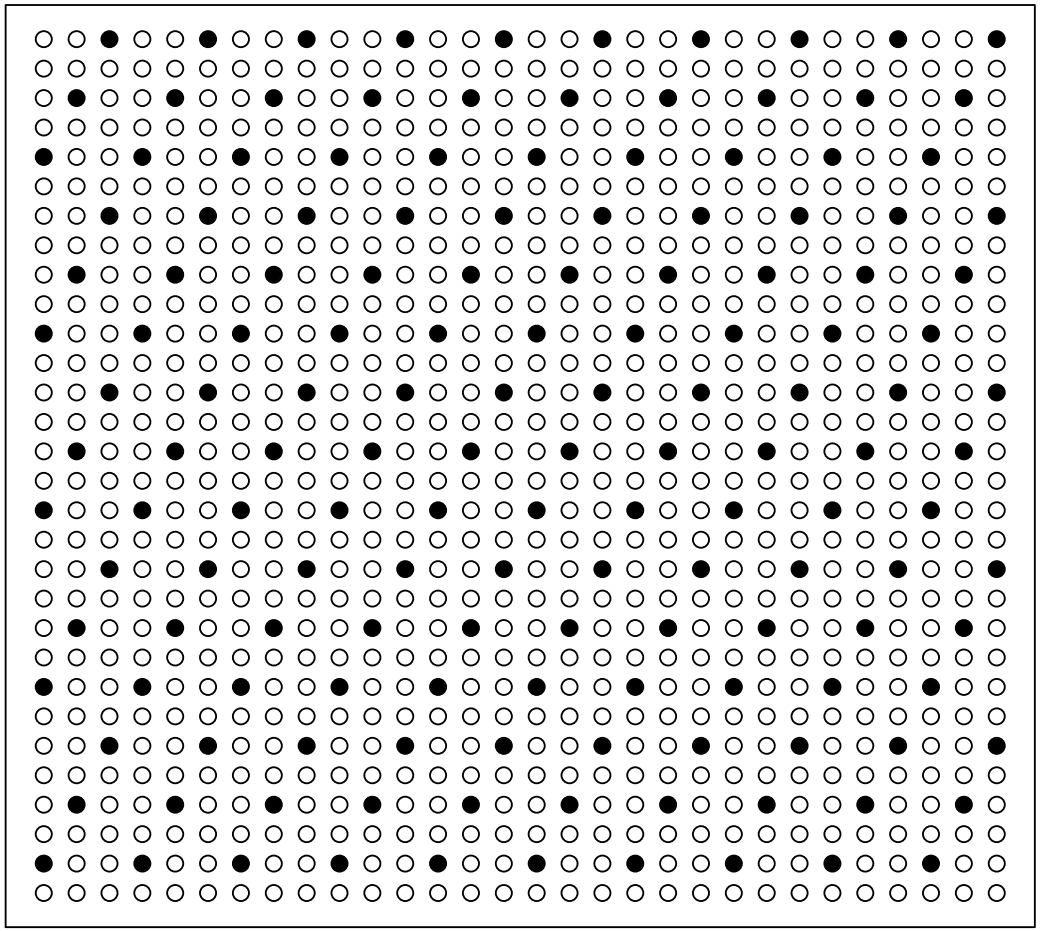
Examples of periodic sample in a grid

$$\pi_k = 1/5$$



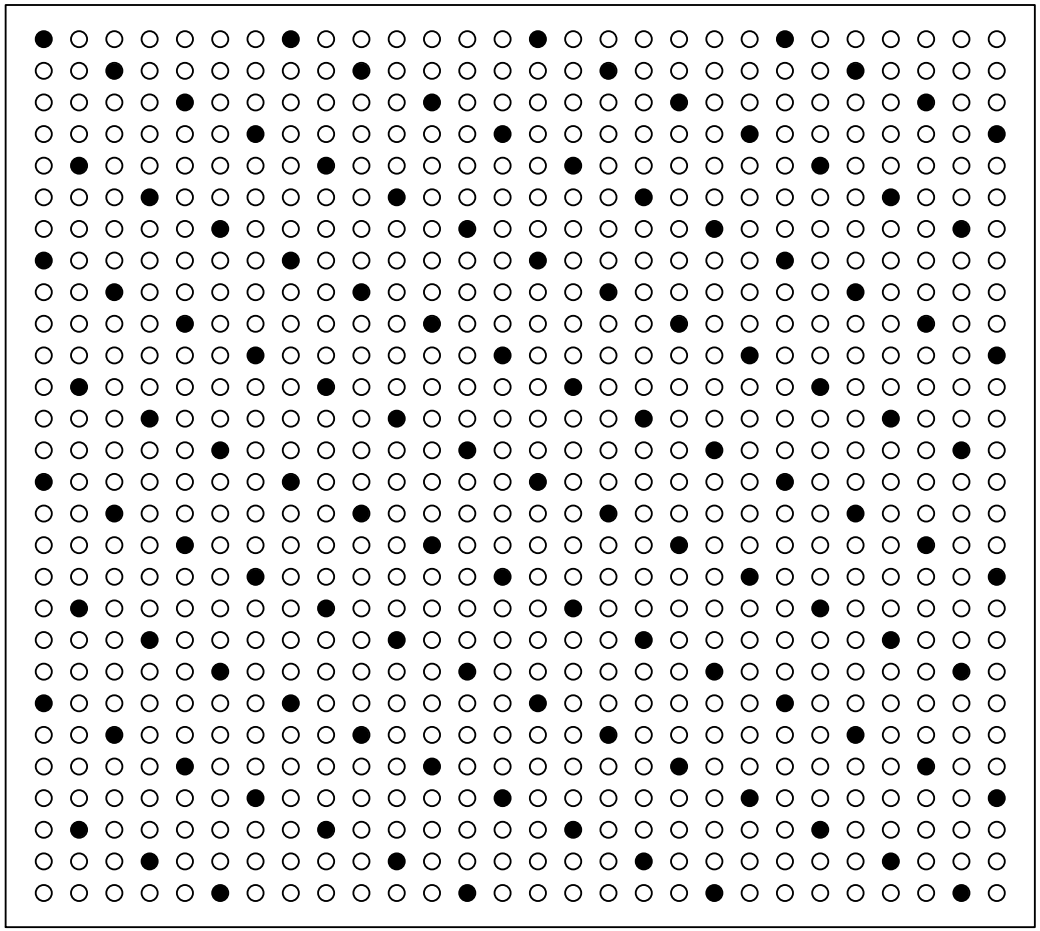
Examples of periodic sample in a grid

$$\pi_k = 1/6$$



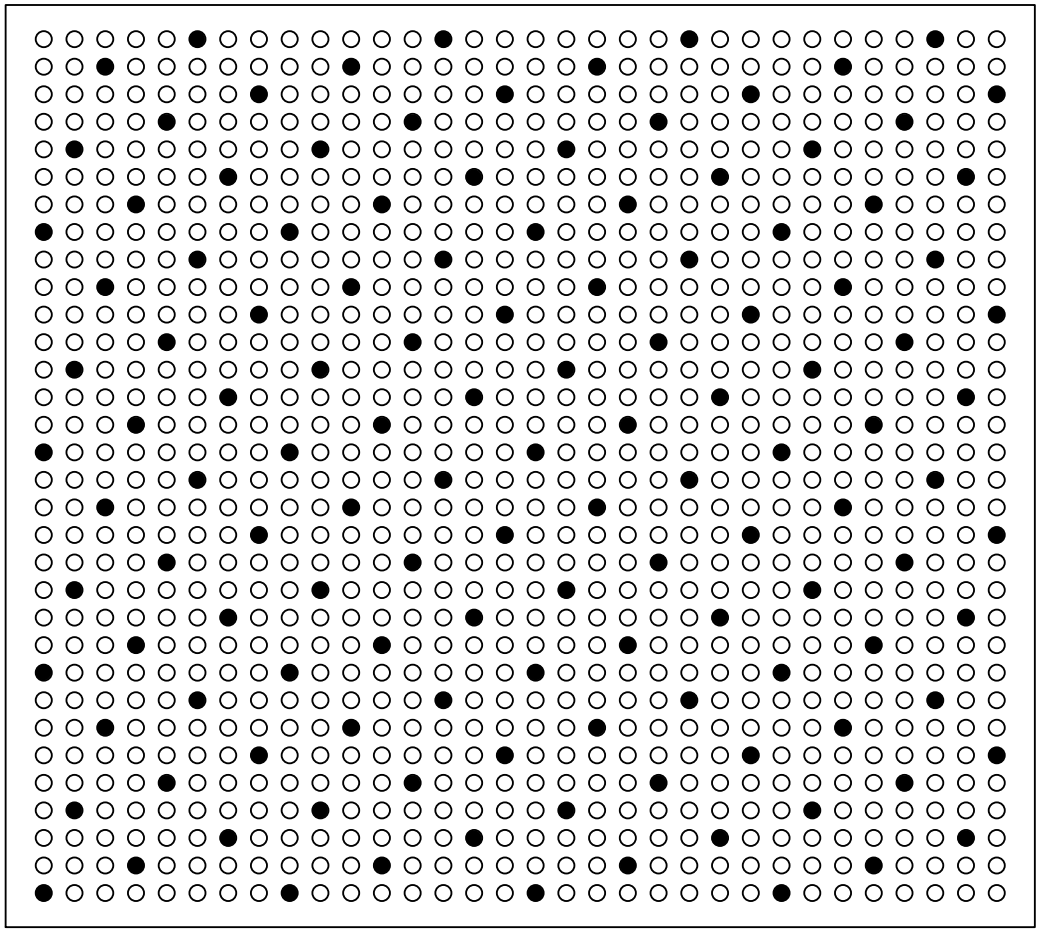
Examples of periodic sample in a grid

$$\pi_k = 1/7$$



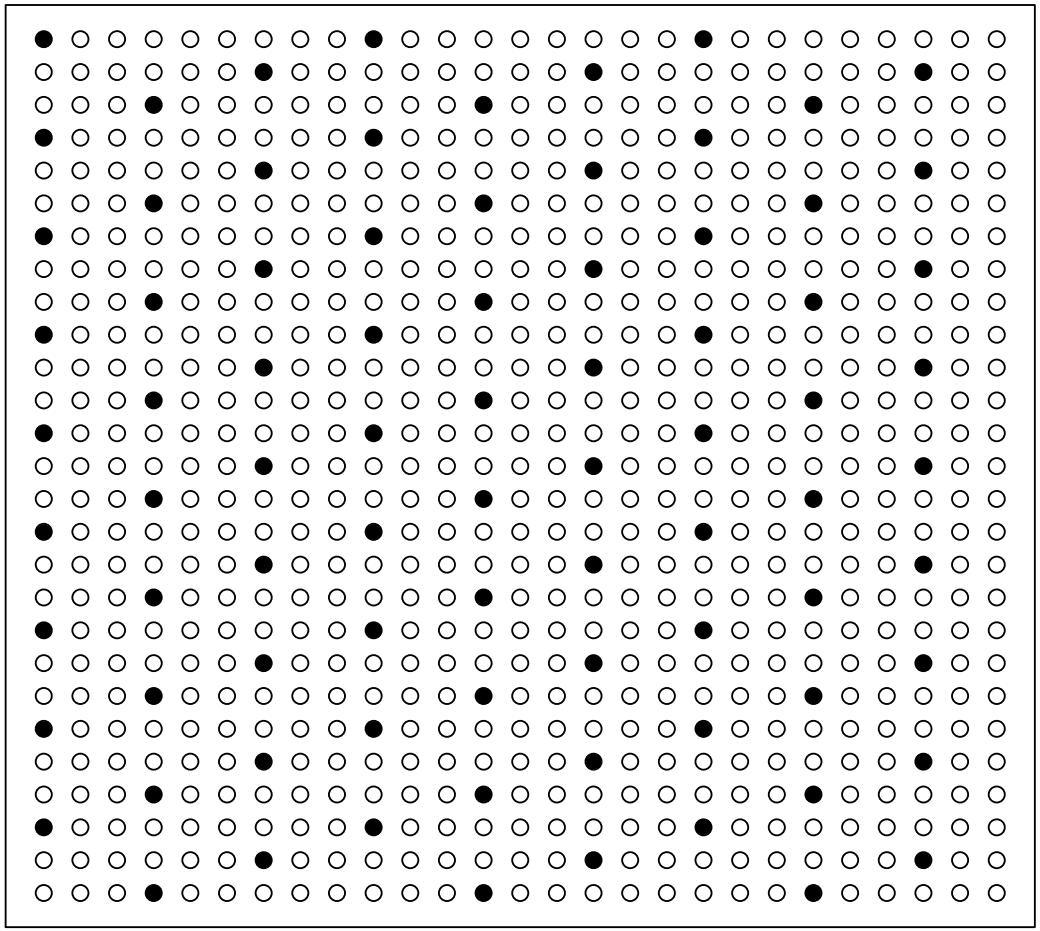
Examples of periodic sample in a grid

$$\pi_k = 1/8$$



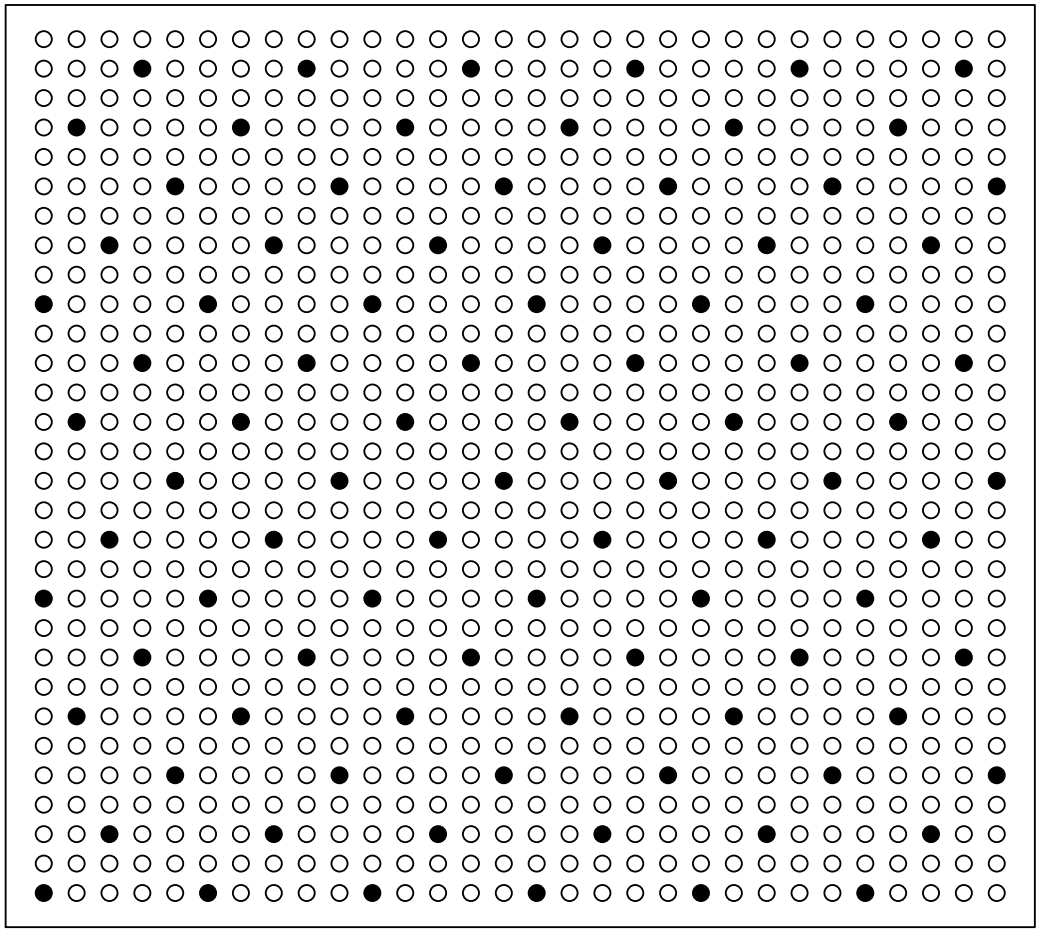
Examples of periodic sample in a grid

$$\pi_k = 1/9$$



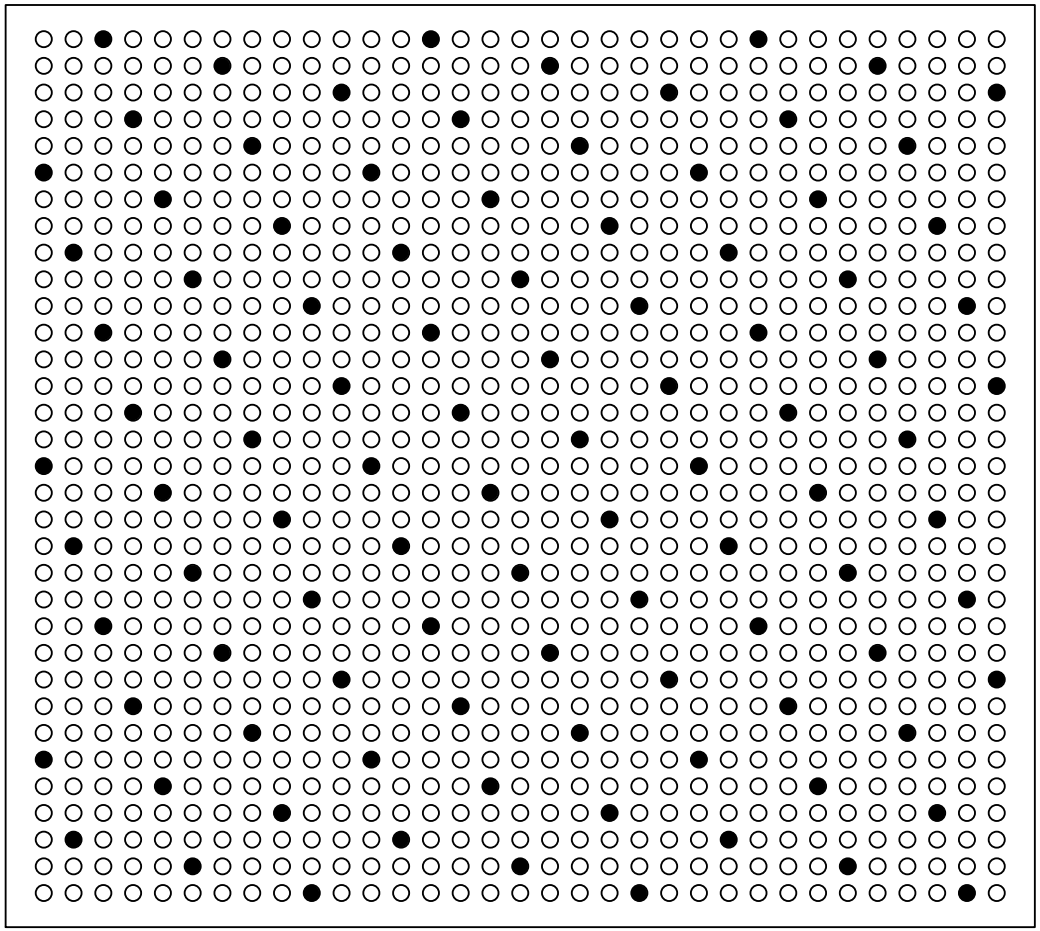
Examples of periodic sample in a grid

$$\pi_k = 1/10$$



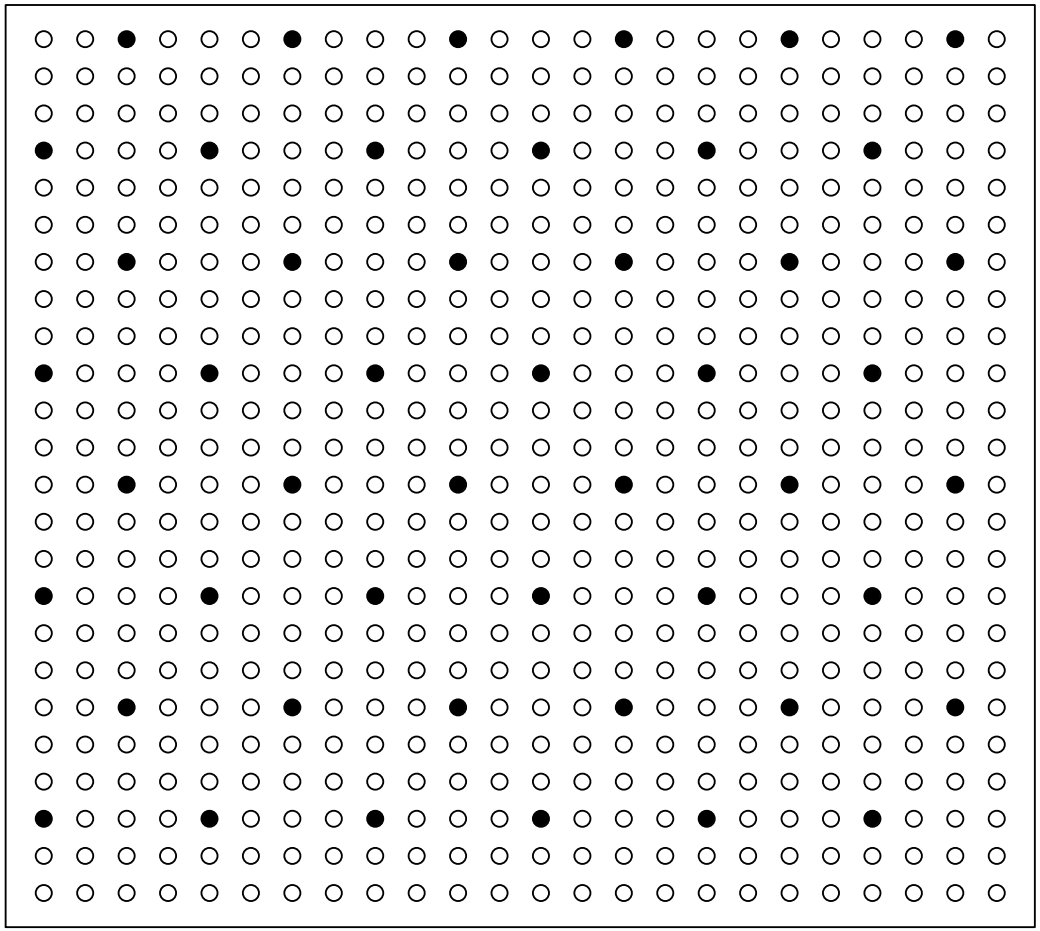
Examples of periodic sample in a grid

$$\pi_k = 1/11$$



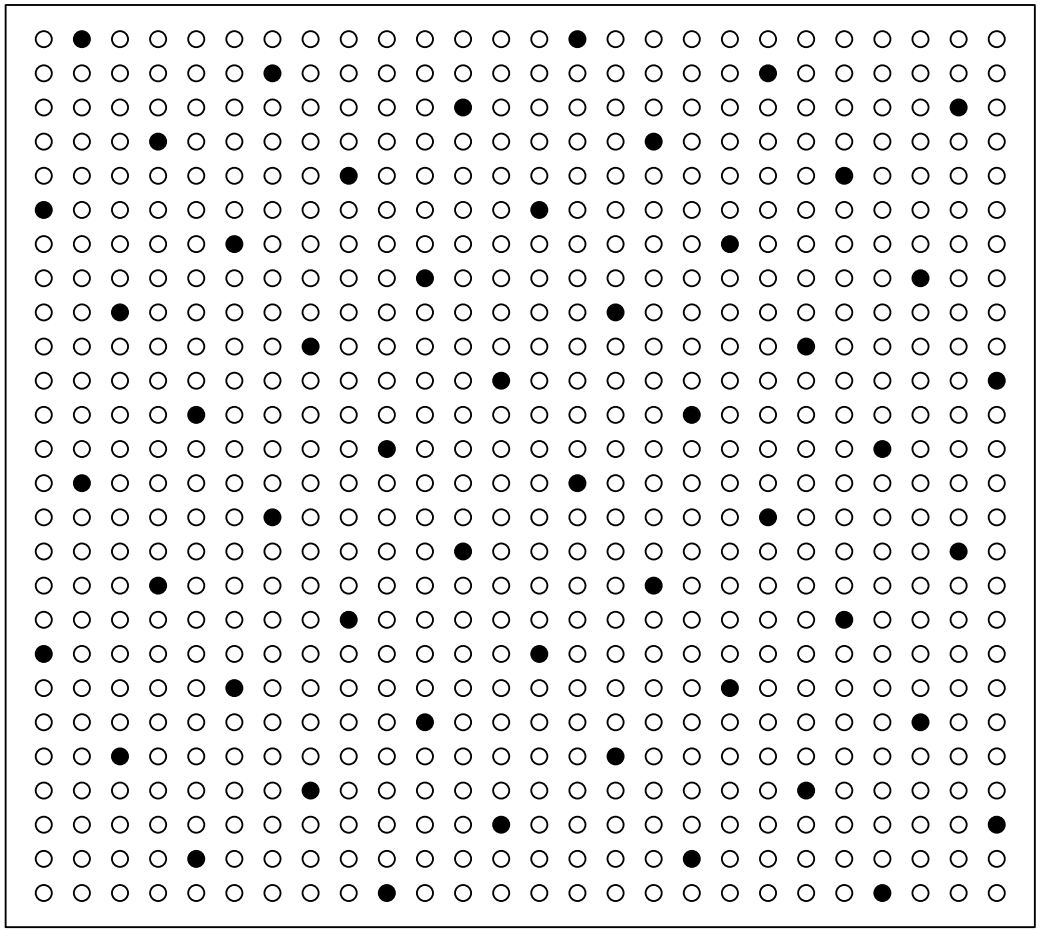
Examples of periodic sample in a grid

$$\pi_k = 1/12$$



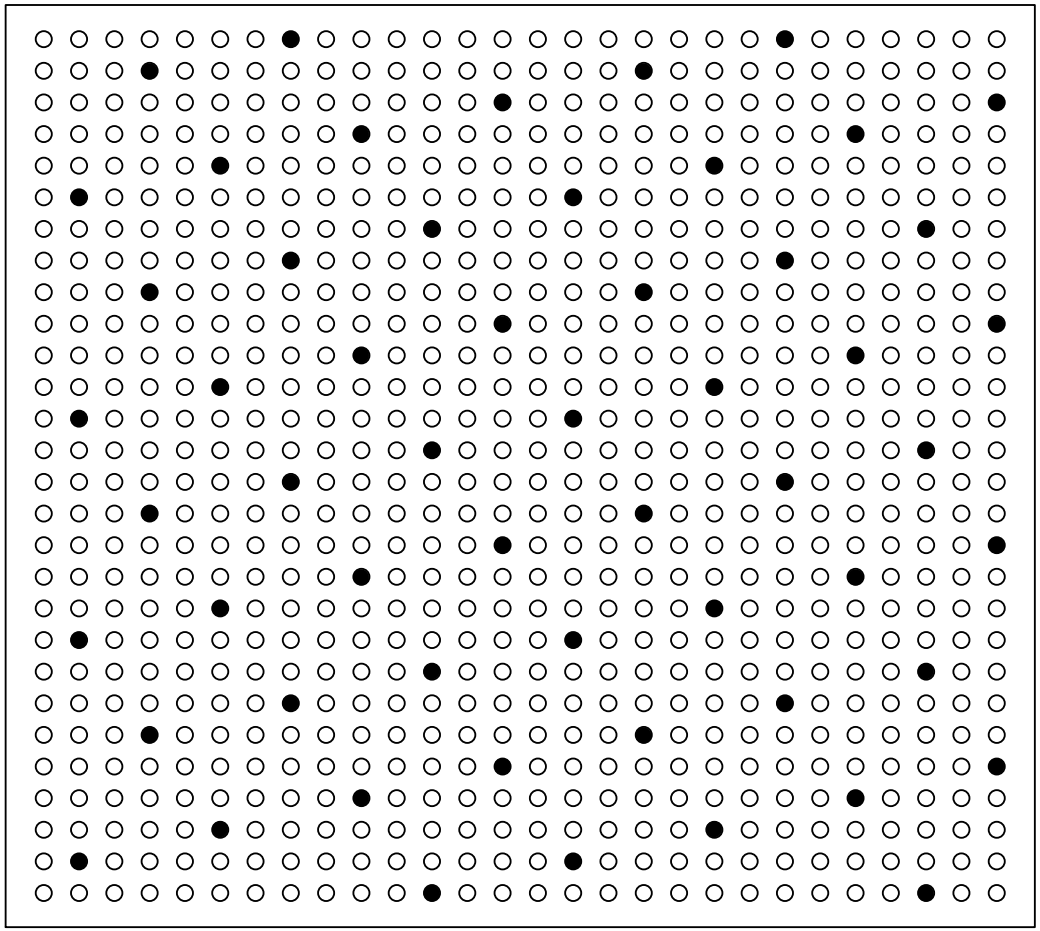
Examples of periodic sample in a grid

$$\pi_k = 1/13$$



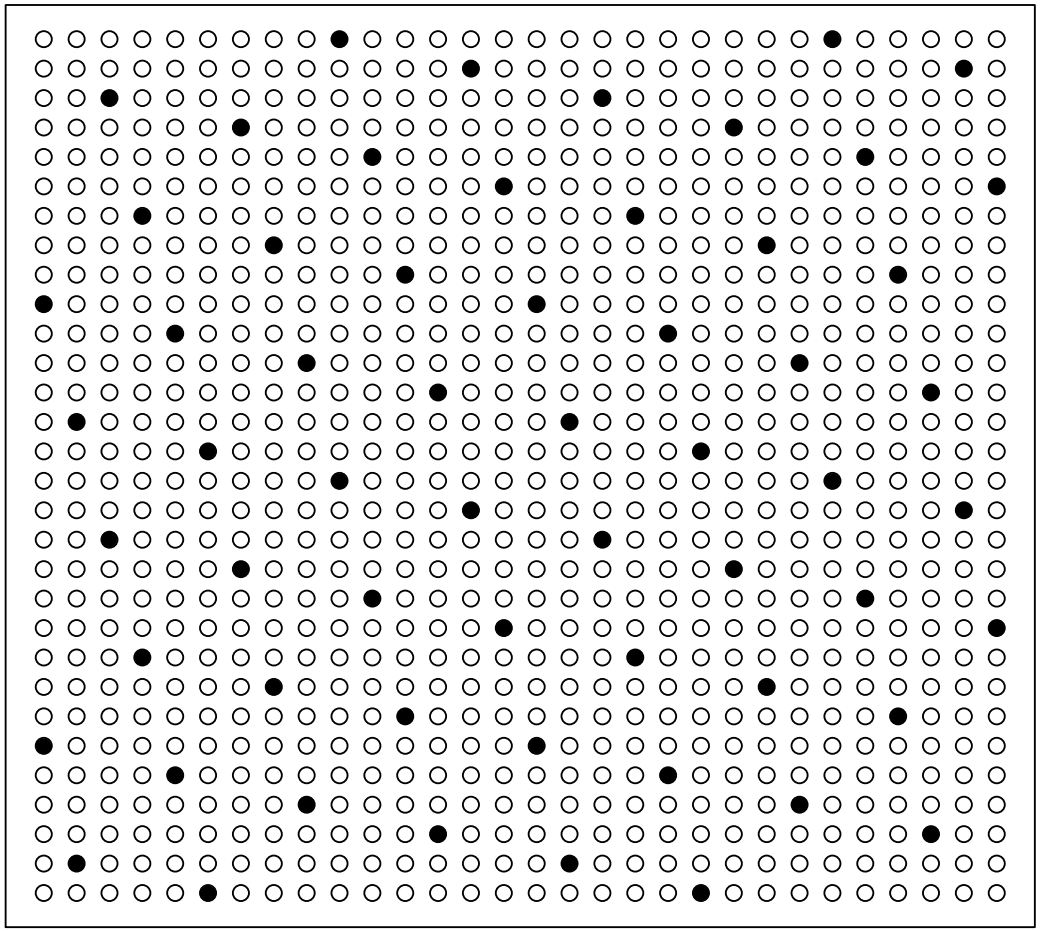
Examples of periodic sample in a grid

$$\pi_k = 1/14$$



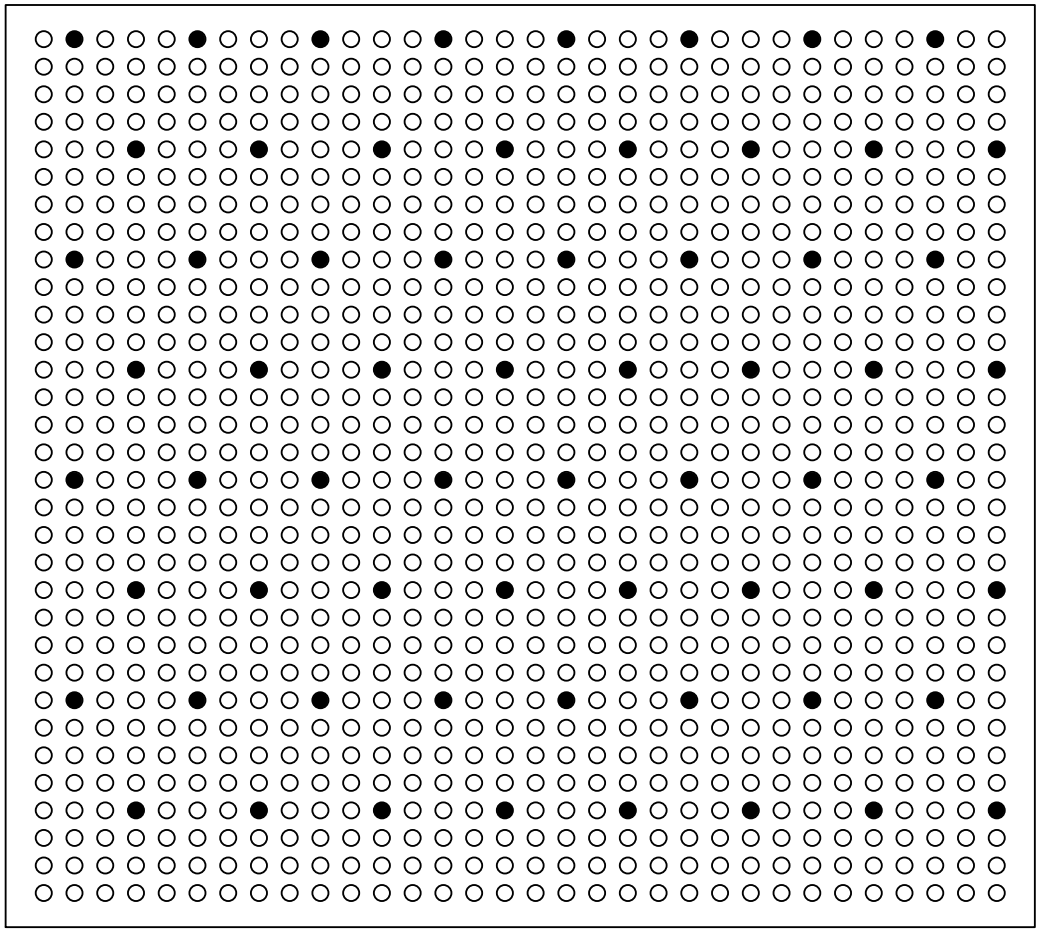
Examples of periodic sample in a grid

$$\pi_k = 1/15$$



Examples of periodic sample in a grid

$$\pi_k = 1/16$$



Algorithm fundamental step

1. Let $\widetilde{\pi}$ be the vector of inclusion probabilities restricted to the k such that $0 < \pi_k < 1$. Let also $\widetilde{\mathbf{X}}$ be the matrix $\mathbf{X}$ with the rows restricted to the k such that $0 < \pi_k < 1$.
2. Construct matrix $\widetilde{\mathbf{A}} = (\widetilde{\mathbf{X}}^\top \mathrm{diag}(\widetilde{\pi})^{-1})$.
   1. If $\widetilde{\mathbf{A}}$ does not have full rank, $\mathbf{u} = (u_1, \ldots, u_k, \ldots, u_N)^\top \in \mathbb{R}^N$ is vector in the kernel of $\widetilde{\mathbf{A}}^\top$, i.e. $\widetilde{\mathbf{A}}^\top \mathbf{u} = \mathbf{0}$.
   2. If $\widetilde{\mathbf{A}}$ is full rank, $\mathbf{u}$ is the right eigenvector associated to the smallest singular value of the SVD of $\mathbf{A}$.
3. Identify λ_1 and λ_2 the largest values such that all the $0 \leq \widetilde{\pi}_k + \lambda_1 u_k \leq 1$ and $0 \leq \widetilde{\pi}_k - \lambda_2 u_k \leq 1$ for all k such that $0 < \pi_k < 1$.
4. Compute

$$\boldsymbol{\pi}^* = \begin{cases} \widetilde{\boldsymbol{\pi}} + \lambda_1 \mathbf{u} & \text{with probability } \lambda_2/(\lambda_1 + \lambda_2) \\ \widetilde{\boldsymbol{\pi}} - \lambda_2 \mathbf{u} & \text{with probability } \lambda_1/(\lambda_1 + \lambda_2). \end{cases}$$

5. Replace in $\boldsymbol{\pi}$ the corresponding values by the values of $\boldsymbol{\pi}^*$.

Conclusion

- Possibility of obtaining the most spread sample.
- Computer intensive.
- Possibility of sparse implementation in 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. & 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.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, 558–606.

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.

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

Bibliography II

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.