

# Sampling From Finite Populations

Yves Tillé

University of Neuchâtel

Les Diablerets, 2021

Doctoral school

1 Introduction and Motivation

2 Examples

3 Basic Concepts and Notations

4 Frameworks and Estimation

5 Simple Designs and their Implementations

6 Stratification

7 Unequal Probability Sampling

8 Balanced Sampling

9 Sampling in Space and Spread Sampling

0 Back to the Examples

# Introduction and Motivation

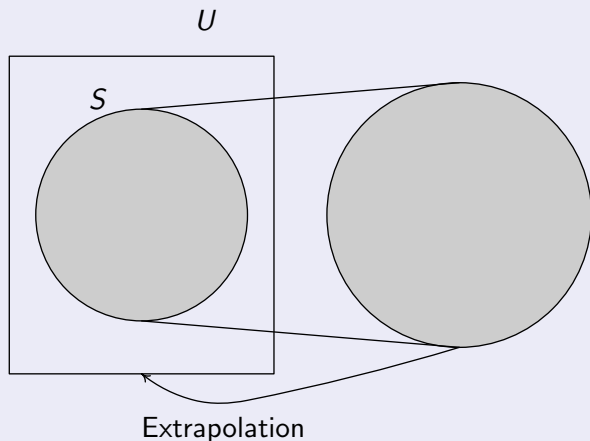
# Introduction

- Sample: subset of a population.
- Set of tools allowing to study a subset of the population for extrapolating the results to the whole population.
- Census: list of all the units of a population.
- Use: official statistics, epidemiology, biology, computer and social sciences, auditing.

# Introduction

## Extrapolation

- Sampling consists of selecting randomly a subset  $S$  of a population  $U$ .
- The aim is to extrapolate the subset  $S$  to population  $U$ .



# Examples

# Example 1.

New repayment system for the hospitals.

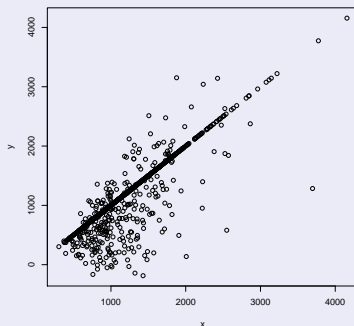
## Diagnostic Related Groups (RDG)

- Reimbursement based on “Major Diagnostic Categories”.
- Auditors to control the codification of the hospital (fraud, codification error).
- Selection of a sample of medical records.
- Two aims.
  1. Estimation of the amount of errors.
  2. Improvement of the codification.

# Example 1.

## New repayment system for the hospitals (RDG)

- Errors are rare.
- Optimization of the sampling design (Marazzi & Tillé, 2017).
- Oversampling when an error is suspected.
- Example with simulated data.

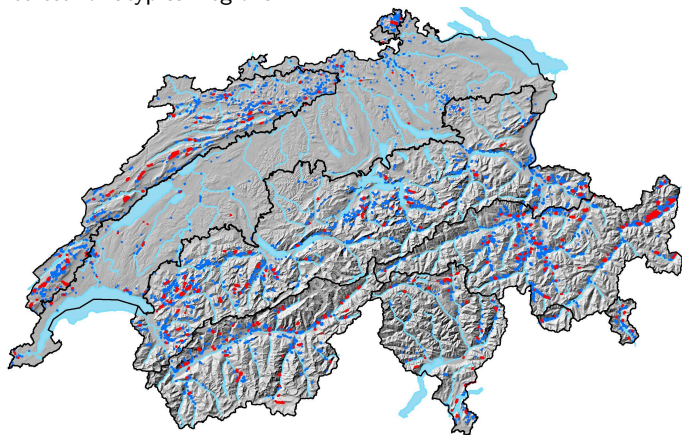




## Example 2.

Selection of a sample of 2100 circles of  $10\text{m}^2$  in the dry grasslands (Tillé & Ecker, 2013)

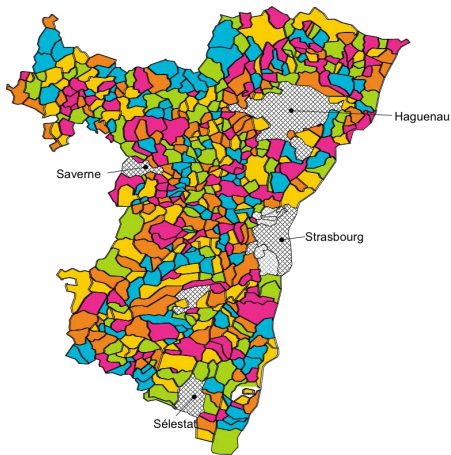
Analysis of the biodiversity of the vegetation.  
Particular interest for atypical regions.



## Example 3.

New French Rolling Census For the small municipalities (less than 10000 inhabitants), five rotations groups are created.

One group is surveyed each year.



(exemplo do Bas-Rhin)



"communes 2004"  
(menos de 10 000 habitantes)



"communes 2005"  
menos de 10 000 habitantes)



"communes 2006"  
(menos de 10 000 habitantes)



"communes 2007"  
(menos de 10 000 habitantes)



"communes 2008"  
(menos de 10 000 habitantes)



Communes de mais de 10 000 habitantes

From Durr & Clanché (2013).

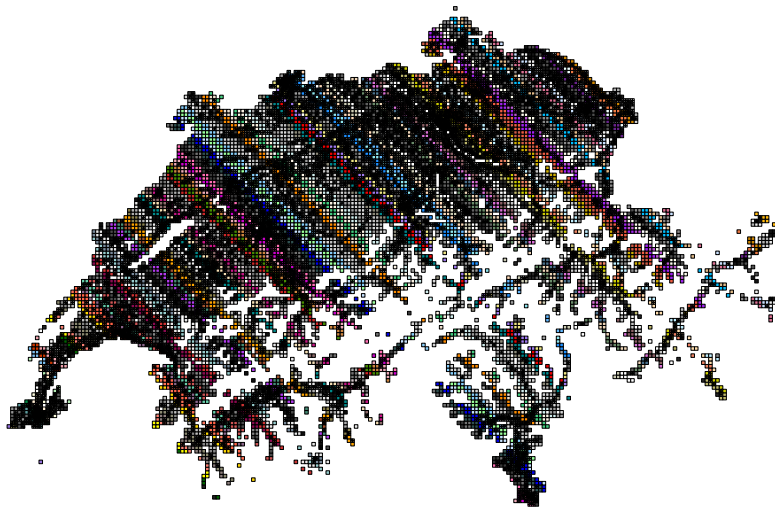
## Example 4.

### Swiss Census System

- In Switzerland, the old census method has been abolished.
- Data are collected from registers (Control of the inhabitants, buildings).
- Quality Survey of the National Census System.
- Important to evaluate the quality of the new method.

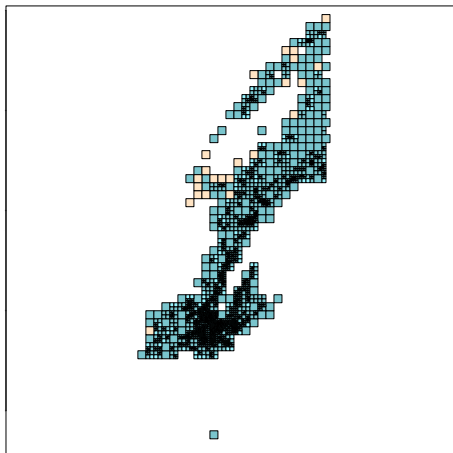
## Example 4.

Swiss Census System



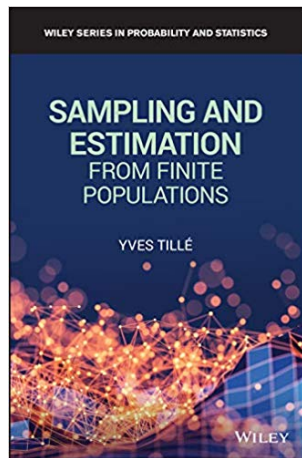
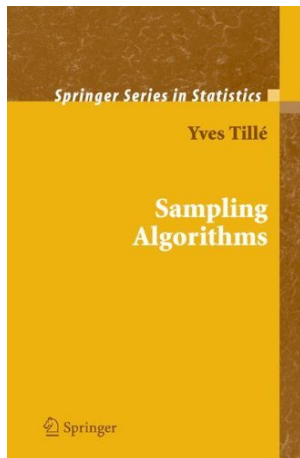
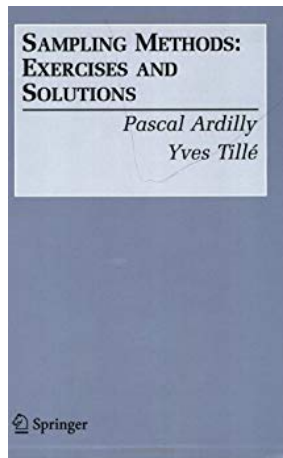
## Example 4.

Swiss Census System



The sampling design is optimized.

# References



See also Tillé & Wilhelm (2017).

# Basic Concepts and Notations

# Population

- Finite population  $U = \{1, \dots, N\}$  of size  $N$ .
- The aim is to study the variable of interest  $y$ , that takes the values  $y_k, k \in U$ .
- The  $y_k$ 's are not supposed to be random.
- The parameter of interest is usually a function of the  $y_k$ .

$$\theta = f(y_1, \dots, y_k, \dots, y_N).$$

- Usual functions of interest include the total,

$$Y = \sum_{k \in U} y_k$$

the mean

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k,$$

the variance

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2$$

- More difficult: Gini index, poverty index.



# Sampling Design

## Definition

A sample  $s$  without replacement is a subset of the population ( $s \subset U$ ).

## Definition

A sampling design  $p(\cdot)$  without replacement is a probability distribution on the set  $\mathcal{S}$  of all the subsets of  $U$  such that

$$\sum_{s \in \mathcal{S}} p(s) = 1.$$

The random sample  $S$  takes a value  $s$  with probability  $\Pr(S = s) = p(s)$ .

## Definition

Indicator variable of the presence of the unit in the random sample

$$a_k = \begin{cases} 1 & \text{if unit } k \in S \\ 0 & \text{otherwise.} \end{cases}$$

Vector  $\mathbf{a} = (a_1, \dots, a_k, \dots, a_N)^\top$  is the vector of random variables.

# Entropy

## Definition

Entropy of a design

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

Entropy is a measure of randomness. The basis principle is to select a sample as random as possible.

# Samples and sampling design

## Example

In a population  $U = \{1, 2, 3\}$ , there are 8 samples

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, U.$$

or

$$(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1).$$

The sampling design is

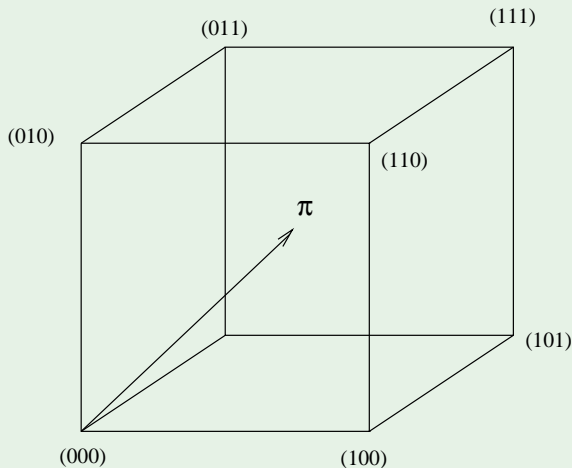
$$p(\{1, 2\}) = \frac{1}{2}, p(\{1, 3\}) = \frac{1}{4}, p(\{2, 3\}) = \frac{1}{4},$$

and the other samples have a null probability.

# Cube

## Example

The samples are the vertices of the  $N$ -cube:



# Inclusion probability

## Definition

The inclusion probability  $\pi_k$  (or *first-order inclusion probability*) of a unit  $k$ :

$$\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s) = E_a(a_k),$$

The joint inclusion probability (or *second-order inclusion probability*) of a couple of unit  $k, \ell \in U$ :

$$\pi_{k\ell} = \Pr(k \text{ and } \ell \in S) = \sum_{s \ni k, \ell} p(s) = E_a(a_k a_\ell).$$

Note that if  $k = \ell$ , we have  $\pi_{k\ell} = \pi_k$ .

The sample size (that can be random) is  $n = \sum_{k \in U} a_k$ . The sum of the inclusion probabilities is equal to the expectation of the sample size. Indeed,

$$\sum_{k \in U} \pi_k = E_a \left( \sum_{k \in U} a_k \right) = E(n).$$

## Example

### Example

If, on a population of size  $N = 3$ , the sampling design is

$$p(\{1, 2\}) = \frac{1}{2}, p(\{1, 3\}) = \frac{1}{4}, p(\{2, 3\}) = \frac{1}{4},$$

$$\pi_1 = p(\{1, 2\}) + p(\{1, 3\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

$$\pi_2 = p(\{1, 2\}) + p(\{2, 3\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

$$\pi_3 = p(\{1, 3\}) + p(\{2, 3\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Thus

$$\pi_1 + \pi_2 + \pi_3 = 2.$$

# Variance and covariance of $a_k$

## Definition

Let  $\Delta_{k\ell} = \text{cov}_a(\mathbf{a}_k, \mathbf{a}_\ell)$ . Hence,

$$\Delta_{k\ell} = \begin{cases} \text{cov}_a(\mathbf{a}_k, \mathbf{a}_\ell) & \text{if } k \neq \ell, \\ \text{var}_a(\mathbf{a}_k) & \text{if } k = \ell. \end{cases}$$

The variance

$$\text{var}_a(\mathbf{a}_k) = E_a(\mathbf{a}_k^2) - E_a^2(\mathbf{a}_k) = \pi_k(1 - \pi_k), \text{ for all } k \in U,$$

and

$$\text{cov}_a(\mathbf{a}_k, \mathbf{a}_\ell) = E_a(\mathbf{a}_k \mathbf{a}_\ell) - E_a(\mathbf{a}_k)E_a(\mathbf{a}_\ell) = \pi_{k\ell} - \pi_k \pi_\ell, \text{ for all } k, \ell \in U, k \neq \ell.$$

Thus

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell, \\ \pi_k(1 - \pi_k) & \text{if } k = \ell. \end{cases}$$

# Variance and covariances

## Result

*If the sample size is fixed, then*

$$\sum_{k \in U} \Delta_{kl} = 0.$$

Indeed,

$$\sum_{k \in U} \Delta_{kl} = \sum_{k \in U} \text{cov}_a(a_k, a_l) = \text{cov}_a\left(a_l, \sum_{k \in U} a_k\right) = \text{cov}_a(a_l, n) = 0.$$

Matrix  $\mathbf{\Delta} = (\Delta_{kl})$  has a null eigenvalue.



# Frameworks and Estimation

# Frameworks

We can consider two sources of randomness

- The sampling design  $p(s)$  and thus vector  $\mathbf{a}$ .
- A possible model, for instance,  $M : y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k$ .

The inference can be made either with respect to the design or with respect to the model or both

- Design-based framework (classic approach). The inference is made with respect to the sampling design. Horvitz-Thompson estimator (design-unbiased) (Horvitz & Thompson, 1952).

$$\hat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

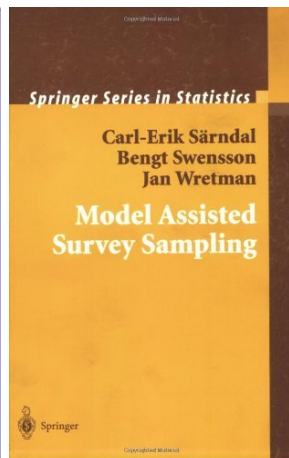
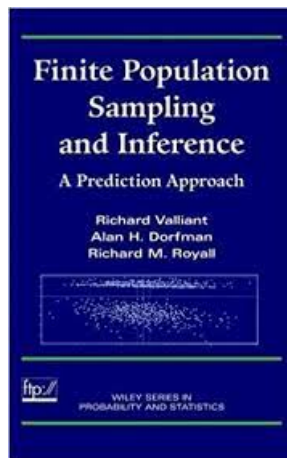
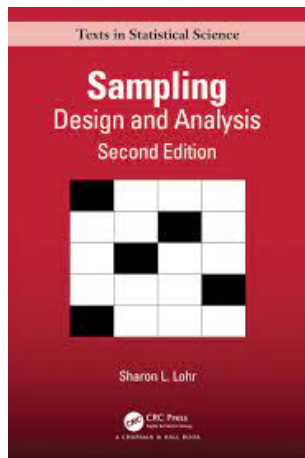
- Model-based framework. The inference is made with respect to the model  $\boldsymbol{\beta}$  is estimated by  $\hat{\boldsymbol{\beta}}$ . The unobserved values are predicted  $\hat{y}_k = \mathbf{x}_k^T \hat{\boldsymbol{\beta}}$  (model-unbiased) (Valliant, Dorfman & Royall, 2000).

$$\hat{Y} = \sum_{k \in S} y_k + \sum_{k \in U \setminus S} \hat{y}_k.$$

- Model-assisted (or mixed) framework. The two sources of randomness are considered but the estimator must remain approximately design-unbiased (Särndal, Swensson & Wretman, 1992).

$$\hat{Y} = \sum_{k \in U} \hat{y}_k + \sum_{k \in S} \frac{y_k - \hat{y}_k}{\pi_k}.$$

# References



# Design-Based inference

## Definition

Let  $y$  be a variable of interest defined on the population  $U$ . The total of the variable  $y$ , denoted by  $Y$ , is

$$Y = \sum_{k \in U} y_k.$$

## Definition

The Horvitz-Thompson estimator (or  $\pi$ -estimator or expanded estimator) of the total  $Y$  is given by

$$\hat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k a_k}{\pi_k}.$$

# Design-based estimation

## Theorem

If  $\pi_k > 0$  for all  $k \in U$ , then  $\hat{Y}$  is a design-unbiased estimator of  $Y$ .

## Proof.

$$E_a(\hat{Y}) = E_a\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) = E_a\left(\sum_{k \in U} \frac{a_k y_k}{\pi_k}\right) = \sum_{k \in U} \frac{E_a(a_k) y_k}{\pi_k} = \sum_{k \in U} y_k = Y.$$



# Notation

## Variance

- Variance of the estimator is equal to:

$$\begin{aligned}\text{var}_a(\hat{Y}) &= \text{var}\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) = \text{var}\left(\sum_{k \in U} \frac{y_k}{\pi_k} a_k\right) \\ &= \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \text{cov}(a_k a_\ell) \\ &= \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.\end{aligned}$$

- For fixed sample size  $n$ :

$$\text{var}_a(\hat{Y}) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell}.$$

# Design-based estimation: Variance and estimation

- $\text{var}_a(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{kl}$
- $\widehat{\text{var}}_a(\widehat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{kl}}{\pi_{kl}}$

## Anticipated variance

- Model  $y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$ , with  $\mathbb{E}_M(\varepsilon_k) = 0$ ,  $\text{var}_M(\varepsilon_k) = \sigma_k^2$ ,  $\text{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{kl}$
- $\mathbb{E}_a \mathbb{E}_M(\widehat{Y} - Y)^2 = \mathbb{E}_a \left[ \left( \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2 + \sum_{k \in U} \sum_{\ell \in U} \frac{\sigma_k \sigma_\ell \rho_{kl} \Delta_{kl}}{\pi_k \pi_\ell}$

# Designing with the Anticipated variance

- Write your intuition on the variable you want to estimate under the form of a model.
- Search an appropriate sampling design.
- Try to implement the design under the form of a sampling algorithm.



# Simple Designs and their Implementations

# Simple random sampling without replacement (WOR)

- All samples of size  $n$  have the same probability of being selected and the others have zero probabilities:

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{if } \#s = n \\ 0 & \text{otherwise} \end{cases}$$

- $\pi_k = \sum_{s \ni k} p(s) = \binom{N-1}{n-1} \binom{N}{n}^{-1} = \frac{n}{N}$

- $\pi_{k\ell} = \sum_{s \ni k, \ell} p(s) = \binom{N-2}{n-2} \binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}$

- $\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell = -\frac{n(N-n)}{N^2(N-1)} & \text{if } k \neq \ell \\ \pi_k(1 - \pi_k) = \frac{n(N-n)}{N^2} & \text{if } k = \ell. \end{cases}$

- $\hat{Y}_{\text{WOR}} = \sum_{k \in S} \frac{y_k}{\pi_k} = N \frac{1}{n} \sum_{k \in S} y_k.$

- $\text{var}_a(\hat{Y}_{\text{WOR}}) = N^2 \frac{N-n}{Nn} S_y^2$  where  $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2$

# Sampling design and sampling algorithms

## Sampling design and sampling algorithms

- A sampling design is a probability law on  $s \subset U$ .
- A sampling algorithm is a procedure to select a sample.

## Sampling design and sampling algorithms

Several very different sampling algorithms can implement the same sampling design. For instance, for simple random sampling:

- Loto method.
- Random sort (select a permutation).
- Reservoir method.
- Selection/rejection method.

# Simple random sampling without replacement (implementation)

There are a lot of implementations of each sampling design.

- Draw by draw implementation without replacement (or Lotto method) (very inefficient).



- Random sort (inefficient).
  - ▶ Assign realisations of i.i.d. continuous variables to each unit.
  - ▶ Sort in increasing order.
  - ▶ Take the first (or last)  $n$  units.

# Selection rejection method, simple random sampling

## Selection rejection method

### Selection-rejection method

Definition  $k, j$ : integer;  $u$ : real;

$k = 0$ ;

$j = 0$ ;

Repeat while  $j < n$   $\left\{ \begin{array}{l} u = \text{uniform random variable } [0, 1]; \\ \text{If } u < \frac{n-j}{N-k} \text{ then } \left\{ \begin{array}{l} \text{select unit } k+1; \\ j = j+1; \end{array} \right. \\ \text{otherwise do not select unit } k+1; \\ k = k+1. \end{array} \right.$

## SSRSWOR: Selection-Rejection (one pass method)

Fan et al. (1962), Bebbington (1975) and Devroye (1986, p. 620).  
This method is called Algorithm *S* by Knuth (1981, pp. 136-138)

```
rm(list=ls())  
n=5  
N=10  
sample(N,n)  
library(sampling)  
srswor(n,N)  
srswor1(n,N)  
srswor1
```

# Reservoir method, simple random sampling

Knuth (1981, p. 144), McLeod & Bellhouse (1983), Vitter (1985) and Devroye (1986, pp. 638–640)

Definition :  $k$  integer;  $u$  real;

The first  $n$  units are selected

Repeat for  $k = n + 1, \dots, N$

$u =$  uniform random number  $[0, 1]$ ;

select unit  $k$ ;

If  $u < \frac{n}{k}$  a unit is removed from the sample ;

the selected unit  $k$  takes the place of the removed unit ;

otherwise unit  $k$  is not selected.

# Bernoulli sampling

## Definition

The  $a_k$ 's are independent and have the same expectation  $\pi_k = \pi$ .

- Due to the independence,  $\Delta_{kl} = 0$  when  $k \neq l$ .
- The sample size is random  $n \sim \text{Bin}(N, \pi)$ .
- Implementation.

```
N=10
```

```
pi=0.2
```

```
as.integer(runif(N)<pi)
```

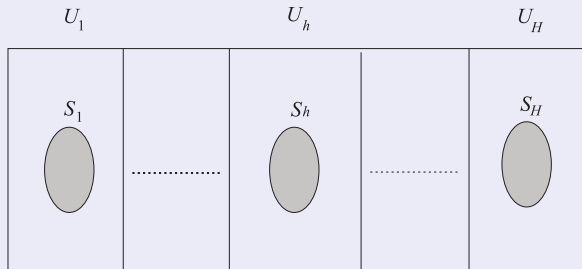


## A small simulation

```
library(sampling)
data(belgianmunicipalities)
attach(belgianmunicipalities)
plot(Tot04/1000,Totaltaxation/1000000)
N=length(Totaltaxation)
n=200
pi=n/N
TT=Totaltaxation/1000000
SIM=1000
ESTHT1=rep(0,SIM)
ESTHT2=rep(0,SIM)
for(i in 1:SIM)
{
ESTHT1[i]=N*mean(TT[srswor(n,N)==1])
ESTHT2[i]=sum( (TT/pi)[as.integer(runif(N)<pi)==1])
}
boxplot(ESTHT1,ESTHT2,names=c("SRSWOR","Bernoulli"))
```

# Stratification

# Stratification



- Stratification with  $H$  nonoverlapping strata  $U_1, \dots, U_H$ .

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

# Stratification

- Inclusion probabilities are  $\pi_k = n_h/N_h$ , if  $k \in U_h$ .
- All the samples with non null probability have the same probability.
- Two main allocations for the sample sizes:
  - ▶ In proportional allocation,  $n_h = \frac{nN_h}{N}$ .
  - ▶ In Neyman's optimal allocation (Neyman, 1934),  $n_h = \frac{nN_h S_h}{\sum_{\ell=1}^H N_\ell S_\ell}$ ,

where  $S_h$  is the standard deviation of the variable of interest in stratum  $U_h$ .

- Morality: Overrepresentation where the dispersion is larger.

## Jerzy Neyman

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



# Stratification

- In order to minimize the variance of the estimator for a global sample size  $n$  the sample size in each stratum must be

$$n_h = \frac{nN_h S_h}{\sum_{i=1}^H N_i S_i},$$

where  $S_h$  is the standard deviation of  $y$  in stratum  $h$  and  $N_h$  is the population size of stratum  $h$ .

- 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 units have null inclusion probabilities, there is a coverage problem.

## Representative sample



# 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.
- Sentences like “the sample is biased” means nothing. Bias is a property of an estimator.

# Unequal Probability Sampling

# Computation of the inclusion probabilities

Auxiliary variable  $x_k > 0, k \in U$  correlated with  $y_k$ . Compute

$$\pi_k = \min(C x_k, 1),$$

where  $C$  is defined such that

$$\sum_{k \in U} \min(C x_k, 1) = n \text{ (sample size).}$$

function `inclusionprobabilities(x, n)` of the R sampling package.

# Computation of the inclusion probabilities

```
N=12
n=4
library(sampling)
pik=inclusionprobabilities(1:N,n)
pik
sum(pik)
data(belgianmunicipalities)
attach(belgianmunicipalities)
pik=inclusionprobabilities(Tot04,200)
Commune[pik==1]
```

# The sample R function is false

Select a large number of samples for estimating  $\pi_k$  by  $\hat{\pi}_k = \frac{1}{\# \text{simulations}} \sum_{\text{simulations}} a_k$ .

Compute  $\frac{\hat{\pi}_k - \pi_k}{\sqrt{\pi_k(1 - \pi_k)/\# \text{ (simulation)}}$  that should remain in  $[-1.96, 1.96]$  in 95% of the cases.

```
N=12
n=4
pik=inclusionprobabilities(1:N,n)
sum(pik)
p=pik/n
p
sum(p)
s=sample(x=1:N, size=n, replace = FALSE, prob = p)
s
a=rep(0,N)
a[s]=1
a
pikest=rep(0,N)
SIM=10000
for(j in 1:SIM)
{
s=sample(x=1:N, size=n, replace = FALSE, prob = p)
a=rep(0,N)
a[s]=1
pikest=pikest+a
}
pikest=pikest/SIM
pik
pikest
(pikest-pik)/sqrt(pik*(1-pik)/SIM)
```

# The generalization of the loto procedure is false

- Compute the drawing probabilities

$$p_k = \frac{\pi_k}{n}, k \in U.$$

- At the first step, select a unit with unequal probability  $p_k, k \in U$ .
- The selected unit is denoted  $j$ .
- The selected unit is removed from  $U$ .
- Next we compute

$$p_k^j = \frac{p_k}{1 - p_j}, k \in U \setminus \{j\}.$$

- Select again a unit with unequal probabilities  $p_k^j, k \in U$ , amongst the  $N - 1$  remaining units, and so on.

This method is wrong.

We can see it by taking  $n = 2$ .

## The generalization of the loto procedure is false

If  $n = 2$

$$\begin{aligned}\Pr(k \in S) &= \Pr(k \text{ be selected at the first step}) \\ &\quad + \Pr(k \text{ be selected at the second step}) \\ &= p_k + \sum_{\substack{j \in U \\ j \neq k}} p_j p_k^j \\ &= p_k \left( 1 + \sum_{\substack{j \in U \\ j \neq k}} \frac{p_j}{1 - p_j} \right).\end{aligned}\tag{1}$$

We should have  $\pi_k = 2p_k, k \in U$ .

## The generalization of the loto procedure is false

We could use modified values  $p_k^*$  for the  $p_k$  in such a way that the inclusion probabilities is equal to  $\pi_k$ .

In the case where  $n = 2$ , we should have  $p_k^*$  such that

$$p_k^* \left( 1 + \sum_{\substack{j \in U \\ j \neq k}} \frac{p_j^*}{1 - p_j^*} \right) = \pi_k, k \in U.$$

This method is known as the Nairin procedure (see also Horvitz & Thompson, 1952; Yates & Grundy, 1953; Brewer & Hanif, 1983, p.25)



# Poisson sampling (generalization of Bernoulli sampling)

## Definition

The  $a_k$  are independent with expectation  $\pi_k$ .

Other definition:

## Definition

A sampling design is called a Poisson sampling design if it can be written

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k), \quad \text{for all } s \subset U.$$

# Poisson sampling (generalization of Bernoulli sampling)

Algorithm:

Definition $u$ : real ; $k$ : integer ;	
Repeat for $k = 1, \dots, N$	$u =$ uniform random variable $[0, 1]$ if $u < \pi_k$ select unit $k$ ; else pass unit $k$ .

Using the independence of each draw, the joint inclusion probabilities are

$$\pi_{k\ell} = \pi_k \pi_\ell,$$

Covariances of the indicator

$$\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell = 0, \text{ for all } k \neq \ell.$$

The distribution of the sample size is Poisson-Binomial (Hodges Jr. & Le Cam, 1960; Stein, 1990; Chen, 1993; Chen & Liu, 1997)

# Systematic sampling 1

## Systematic sampling

- Cumulated inclusion probabilities

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

- $u$  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

### Example

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

$k$	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	

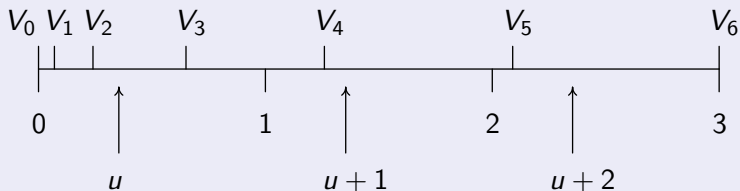
# Systematic sampling 3

## Systematic sampling

Suppose also that the value taken by the uniform random number is  $u = 0.354$ . The rules of selection 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  $\mathbf{a} = (0, 0, 1, 0, 1, 1)$ .



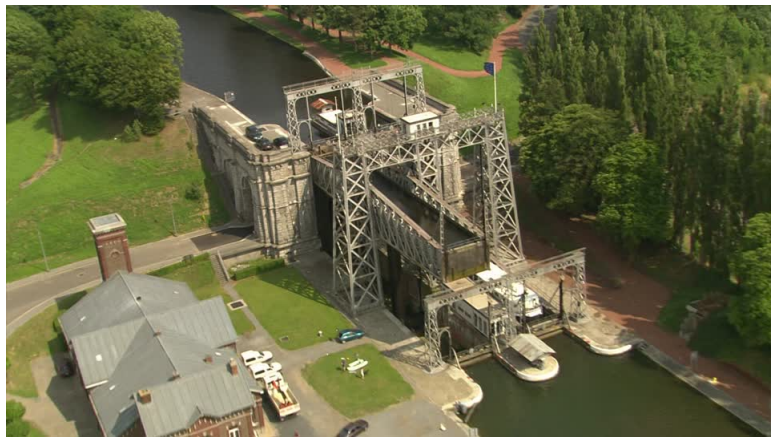
# Systematic sampling 4

## Systematic sampling

```
library(sampling)
pik=c(0.07,0.17,0.41,0.61,0.83,0.91)
UPsystematic(pik)
UPrandomsystematic(pik)
UPsystematicpi2(pik)
```

- Most of the joint inclusion probabilities are null.
- Random systematic sampling: the population is sorted at random before applying systematic sampling.

# Pivotal method 1



## Pivotal method 2





## Pivotal method 3

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

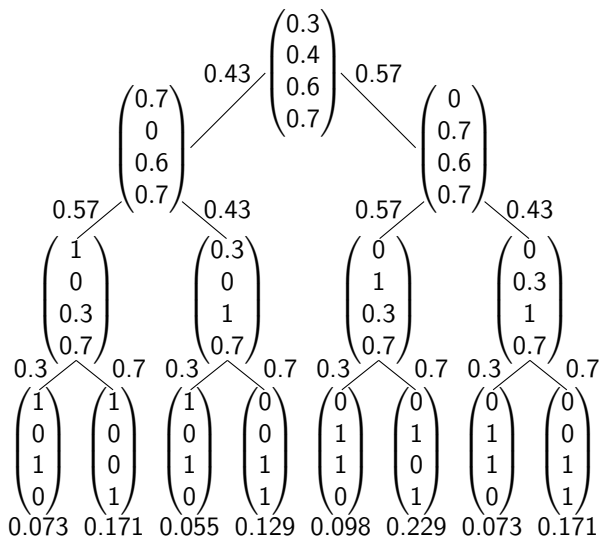
## Pivotal method 4

- Pivotal method (Deville & Tillé, 2000).
- At each step, two inclusion probabilities are modified randomly (called  $i$  and  $j$ ).
- 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 5



## Pivotal method 6

- 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

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

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

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

# Pivotal Method 7

If  $\pi_i + \pi_j < 1$ , then

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

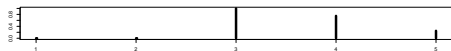
$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ \pi_i + \pi_j & k = i \\ 0 & k = j, \end{cases} \quad \text{and} \quad \pi_k^{(2)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ 0 & k = i \\ \pi_i + \pi_j & k = j. \end{cases}$$

# Different 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 method: Example



# Remarks on the entropy

## Designs that maximize the entropy

Design	Constraint(s)	Sample size
Bernoulli sampling	Equal inclusion probabilities $\pi$	random
Simple random sampling	Fixed sample size $n$	fixed
Poisson sampling	Unequal inclusion probabilities $\pi_k$	random
Maximum entropy designs	incl. prob. $\pi_k$ and fixed $n$	fixed



## Maximum entropy with fixed sample size

One of the main drawback of the Poisson sampling designs is that it is of random size. So, instead of maximizing the entropy only under constraints of prescribed inclusion probabilities, we may add a fixed size constraint. Let us define  $\mathcal{S}_n$ , the set of all the subset of  $U$  of size  $n$

$$\mathcal{S}_n = \{s \mid \#s = n\}.$$

The optimization problem can be written as

$$\begin{aligned} \text{maximize} \quad & I(p) = - \sum_{s \in \mathcal{S}_n} p(s) \log p(s) \\ \text{subject to} \quad & \sum_{\substack{s \ni k \\ s \in \mathcal{S}_n}} p(s) = \pi_k, \text{ and } \sum_{s \in \mathcal{S}_n} p(s) = 1. \end{aligned} \tag{2}$$

## Maximum entropy with fixed sample size (cont'd)

The solution satisfies

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

If  $\pi_k(n)$  is the probability the inclusion probability of a maximum entropy design of size  $n$ , it is possible to show that:

$$\pi_k(n) = \frac{\exp \lambda_k \{1 - \pi_k(n-1)\}}{\sum_{\ell \in U} \exp \lambda_\ell \{1 - \pi_\ell(n-1)\}}, k \in U.$$

It is thus possible to pass from  $\lambda_k$  to  $\pi_k$  without enumerating all the possible samples. It is also possible to pass from  $\pi_k$  to  $\lambda_k$  by means of the Newton method (method for finding the root of a function). Several implementations are possible:

- Conditional Poisson sampling or rejective method
- Sequential algorithm.

(see Chen, Dempster & Liu, 1994; Deville, 2000)

## Maximum entropy with fixed sample size (cont'd)

- Implemented but complicated (Tillé, 2006).

```
library(sampling)
pik=c(0.07,0.17,0.41,0.61,0.83,0.91)
UPmaxentropy(pik)
UPmaxentropypi2(pik)
```

# A larger simulation

```
library(sampling)
data(belgianmunicipalities)
attach(belgianmunicipalities)
plot(Tot04/1000,Totaltaxation/1000000)

N=length(Totaltaxation)
n=200
pi=n/N
pik=inclusionprobabilities(Tot04,n)
TT=Totaltaxation/1000000
SIM=100
ESTHT1=rep(0,SIM)
ESTHT2=rep(0,SIM)
ESTHT3=rep(0,SIM)
ESTHT4=rep(0,SIM)
ESTHT5=rep(0,SIM)
ESTHT6=rep(0,SIM)
ESTHT7=rep(0,SIM)
for(i in 1:SIM)
{
print(i)
ESTHT1[i]=N*mean(TT[srswor(n,N)==1])
ESTHT2[i]=sum( (TT/pi)*as.integer(runif(N)<pi))
ESTHT3[i]=sum( (TT/pik)*UPpoisson(pik) )
ESTHT4[i]=sum( (TT/pik)*UPrandomsystematic(pik) )
ESTHT5[i]=sum( (TT/pik)*UPrandompivotal(pik) )
ESTHT6[i]=sum( (TT/pik)*UPmaxentropy(pik) )
ESTHT7[i]=sum( (TT/pik)[sample(x=1:N,size=n,prob=pik/n) ] )
}
boxplot(ESTHT1,ESTHT2,ESTHT3,ESTHT4,ESTHT5,ESTHT6,ESTHT7,
names=c("SRSWOR","Bernoulli","Poisson",
"R. syst","R. piv.,""Max. entrop.,""False"))
abline(a=sum(TT),b=0)
```

# Balanced sampling

# Balanced sampling in survey sampling

- A sample is balanced if the Horvitz-Thompson estimators of totals calculated from the sample are equal or nearly equal to the population totals.
- Gini & Galvani (1929) had selected a sample of 29 districts (*circondari*) out of 214 in order to reconstitute some population averages. (Langel & Tillé, 2011; Tillé, 2016; Brewer, 2013). The method was criticized by Jerzy Neyman because the sample is not selected randomly.



# Balanced sampling in survey sampling

- Yates (1949) and Thionet (1953) proposed random methods of unit substitution to improve the balance.
- Hájek (1964, 1981) proposed rejective sampling (Choudhry & Singh, 1979; Dupačová, 1979; Fuller, 2009; Legg & Yu, 2010; Boistard, Lopuhaä & Ruiz-Gazen, 2012; Fuller, Legg & Li, 2017).
- Royall & Herson (1973) advocate for using balanced samples in the model-based framework.
- Ardilly (1991), Hedayat & Majumdar (1995) proposed enumerative methods.
- The cube method (Deville & Tillé, 2004, 2005): Random balanced sampling with equal or unequal inclusion probabilities (Tillé, 2006; Chauvet & Tillé, 2006; Chauvet, 2009; Chauvet, Deville & Haziza, 2010, 2011b; Tillé, 2011; Breidt & Chauvet, 2011; Chauvet, Bonnéry & Deville, 2011a; Breidt & Chauvet, 2012; Grafström & Tillé, 2013; Hasler & Tillé, 2014; Chauvet, Haziza & Lesage, 2015; Grafström & Lisic, 2019; Jauslin, Eustache & Tillé, 2021).

# Notation

- Auxiliary variables  $x_1, \dots, x_p$ , known for each unit of the population.
- $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^\top$ , is known for all  $k \in U$ .
- The vector of totals  $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$ .
- The Horvitz-Thompson estimator of the vector of totals

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

- The aim is always to estimate  $\hat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$ .



# Definition

## Definition

A sampling design  $p(s)$  is said to be balanced on the auxiliary variables  $x_1, \dots, x_p$ , if and only if it satisfies the balancing equations given by  $\widehat{\mathbf{X}} = \mathbf{X}$ , which can also be written

$$\sum_{k \in s} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for all  $s \in \mathcal{S}$  such that  $p(s) > 0$ , and for all  $j = 1, \dots, p$ , or in other words

$$\text{var}(\widehat{\mathbf{X}}) = 0.$$

## Particular case 1: Fixed sample size

- A sampling design of fixed sample size  $n$  is balanced on the variable  $x_k = \pi_k, k \in U$ . Indeed,

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k = n.$$

## Particular case 2: Stratification

- Stratification with strata  $U_h, h = 1, \dots, H, \#U_h = N_h$   
Simple random sample of size  $n_h$  in each stratum  
The design is balanced on variables  $\pi_k \delta_{kh}$  of values

$$\delta_{kh} = \begin{cases} 1 & \text{if } k \in U_h \\ 0 & \text{if } k \notin U_h. \end{cases}$$

Indeed 
$$\sum_{k \in S} \frac{\pi_k \delta_{kh}}{\pi_k} = \sum_{k \in S} \delta_{kh} = \sum_{k \in U} \pi_k \delta_{kh} = \sum_{k \in U_h} \pi_k = n_h,$$
 for  $h = 1, \dots, H$ .

## Particular case 3: Rounding problem

- $N = 10, n = 7, \pi_k = 7/10, k \in U,$   
 $x_k = k, k \in U.$

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

which gives that

$$\sum_{k \in S} k = 55 \times 7/10 = 38.5,$$

IMPOSSIBLE: Rounding problem.

- Aim: find a sample approximately balanced!

## Particular case 4: Balancing on the population size

- Balance on the variable  $x_k = 1, k \in U$ . The balancing equations become

$$\sum_{k \in S} \frac{1}{\pi_k} = \sum_{k \in U} 1 = N.$$

or

$$\hat{N} = N.$$

The population size is estimated without error.

- REMARK: there is always two free auxiliary variables

$$x_{k1} = \pi_k \text{ and } x_{k2} = 1, k \in U.$$

A sample should always be balanced on these variables.

## General Remark

- All the problems of sampling can theoretically be solved by using a linear program.
- Define a cost  $C(s)$  for each sample  $s$ . The cost is small if the sample is well balanced.
- Search the sampling design  $p(s)$  that minimizes the expected cost

$$\sum_{s \subset U} C(s)p(s)$$

subject to

$$\sum_{s \subset U} p(s) = 1 \text{ and } \sum_{s \subset U, s \ni k} p(s) = \pi_k, k \in U.$$

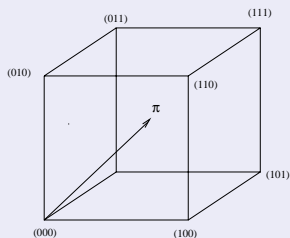
- Impossible in practice because of the combinatory explosion ( $2^N$  samples).
- The cube method is a shortcut that avoids the enumeration of the samples.

# Cube representation

- Geometric representation of a sampling design.

$$\mathbf{a}_s = (I[1 \in s] \dots I[k \in s] \dots I[N \in s])^\top,$$

where  $I[k \in s]$  takes the value 1 if  $k \in s$  and 0 if not.



Possible samples in a population of size  $N = 3$

# Cube representation

- Geometrically, each vector  $\mathbf{a}_s$  is a vertex of a  $N$ -cube.

$$E(\mathbf{a}_s) = \sum_{s \in \mathcal{S}} p(s) \mathbf{a}_s = \boldsymbol{\pi},$$

where  $\boldsymbol{\pi} = [\pi_k]$  is the vector of inclusion probabilities.



# Balancing equations

- The balancing equations

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

can also be written

$$\sum_{k \in U} \check{\mathbf{x}}_k a_k = \sum_{k \in U} \check{\mathbf{x}}_k \pi_k \text{ with } a_k \in \{0, 1\}, k \in U,$$

where  $\check{\mathbf{x}}_k = \mathbf{x}_k / \pi_k, k \in U$ .

- The balancing equations defines an affine subspace in  $\mathbb{R}^N$  of dimension  $N - p$  denoted  $Q$ .
- $Q = \pi + \text{Ker}(\mathbf{A})$  where  $\mathbf{A} = (\check{\mathbf{x}}_1, \dots, \check{\mathbf{x}}_k, \dots, \check{\mathbf{x}}_N)$ .  
**The problem:** Choose a vertex of the  $N$ -cube (a sample) that remains on the sub-space  $Q$ .

# System exactly verifiable

## Example

$$\pi_1 + \pi_2 + \pi_3 = 2.$$

$$x_k = \pi_k, k \in U \text{ and } \sum_{k \in U} s_k = 2.$$

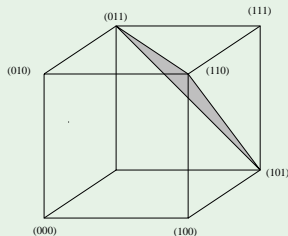


Figure: Fixed size constraint: all the vertices of  $K$  are vertices of the cube

# System approximately verifiable

## Example

- $6 \times \pi_2 + 4 \times \pi_3 = 5$ .
- $x_1 = 0, x_2 = 6 \times \pi_2$  and  $x_3 = 4 \times \pi_3$ .

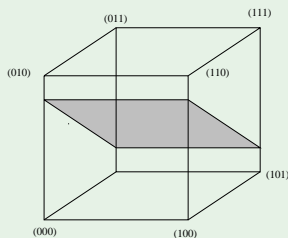


Figure: none of vertices of  $K$  are vertices of the cube

# System sometimes verifiable

## Example

$$\pi_1 + 3 \times \pi_2 + \pi_3 = 4.$$

$$x_1 = \pi_1, x_2 = 3 \times \pi_2 \text{ and } x_3 = \pi_3.$$

$$s_1 + 3s_2 + s_3 = 4.$$

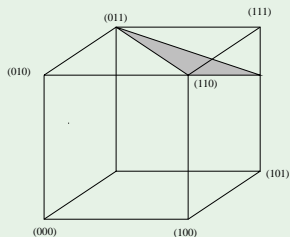


Figure: some vertices of  $K$  are vertices of the cube and others not

## Cube methods: phases

- **Cube method** (Deville & Tillé, 2004)
  - ▶ flight phase
  - ▶ landing phase (needed only if there exists a rounding problem)
- **The flight phase** is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace.  
This random walk stops at a vertex of the intersection of the cube and the constraint subspace.
- **The landing phase** At the end of the flight phase, if a sample is not obtained, a sample is selected as close as possible to the constraint subspace.

# Cube methods: examples

## Example

The only constraint is the fixed sample size.

The flight phase transforms a vector of inclusion probabilities into a vector of 0 and 1.

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6666 \\ 0.6666 \\ 0.6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{a}.$$

Maximum  $N - p$  steps.

## Cube methods: examples

### Example

If there exists a rounding problem, then some components cannot be put to zero.

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.25 \\ 1 \\ 0.25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \pi^*.$$

In this case, the flight phase lets one non-integer component.

## Idea of the algorithm

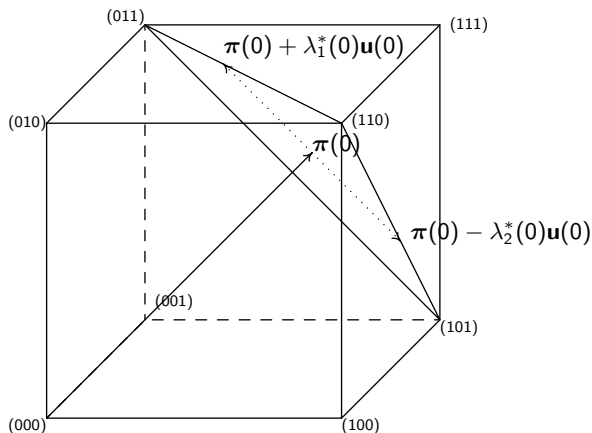


Figure: Flight phase in a population of size  $N = 3$  with a sample size constraint  $n = 2$



# The algorithm

First initialize with  $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ . Next, at time  $t = 0, \dots, T$ ,

- 1 Generate any vector  $\mathbf{u}(t) = [u_k(t)] \neq 0$  such that
  - (i)  $\mathbf{u}(t)$  is in the kernel of matrix  $\mathbf{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_k/\pi_k, \dots, \mathbf{x}_N/\pi_N)$
  - (ii)  $u_k(t) = 0$  if  $\pi_k(t)$  is integer.
- 2 Compute  $\lambda_1^*(t)$  and  $\lambda_2^*(t)$ , the largest values such that
$$0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$$
$$0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$$
- 3 Compute

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with a proba } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with a proba } q_2(t), \end{cases}$$

where  $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$  and  $q_2(t) = 1 - q_1(t)$ .

# Chauvet Tillé Implementation

- Chauvet & Tillé (2005a,b, 2006, 2007); Tillé & Matei (2021)
- Fast algorithm. Execution time  $O(N \times p^2)$ .
- Apply each step on the algorithm only on the first  $p + 1$  units with non integer  $\pi_k(t)$ .
- If the only constraint is the fixed sample size: pivotal method.
- The order of the file change the sampling design.
- Two solutions:
  - ▶ Random order of the sample. (Increasing of the randomness of the sample, of the entropy).
  - ▶ Decreasing order of size (reduce the rounding problem).

# Landing Phase 1

- Let  $\boldsymbol{\pi}^* = [\pi_k^*]$  the vector obtained at the last step of the flight phase.

	Inclusion	Flight	Landing
•	probabilities	Phase	phase
	$\boldsymbol{\pi}$	$\rightarrow \boldsymbol{\pi}^*$	$\rightarrow \mathcal{S}$

- It is possible to prove that

$$\text{card}U^* = \text{card} \{k \in U \mid 0 < \pi_k^* < 1\} = q \leq p.$$

- The aim of the landing phase is to find a sample  $\mathbf{a}$  such that  $E(\mathbf{a} \mid \boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$ , and that is almost balanced.

# Landing Phase 1

- Solution: linear program defined only on  $q \leq p$  units.
- Search the sampling design on  $U^*$  that minimize

$$\sum_{s^* \subset U^*} p(s^*) C(s^*)$$

subject to

$$\sum_{s^* \subset U^*, s^* \ni k} p(s^*) = \pi_k^* \text{ and } \sum_{s^* \subset U^*} p(s^*) = 1.$$

- $C(s^*)$  is the cost of sample  $s^*$  (for instance the distance between the sample and the subspace of constraints).

## Landing Phase 2

- If the number of auxiliary variables is too large for the linear program to be solved by a simplex algorithm,  $p > 20$  then, at the end of the flight phase, an auxiliary variable can be dropped.
- Next, one can return to the flight phase until it is no longer possible to 'move' within the constraint subspace. The constraints are thus relaxed successively.

# Model without autocorrelation

- Model  $y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$ , with  $E_M(\varepsilon_k) = 0$ ,  $\text{var}_M(\varepsilon_k) = \sigma_k^2$ ,  $\text{cov}_M(\varepsilon_k, \varepsilon_\ell) = 0$
- Anticipated variance

$$E_a E_M(\hat{Y} - Y)^2 = E_a \left[ \left( \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2 + \sum_{k \in U} \frac{\sigma_k^2 \pi_k (1 - \pi_k)}{\pi_k^2}$$

- In order to minimize the anticipated variance
  - ▶ Select a balanced sample:  $\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k$ .
  - ▶ Take inclusion probabilities  $\pi_k \propto \sigma_k$ .

## Example 1: The Annual County Population

- The "CO-EST2006-alldata": is the Annual County Population Estimates and Estimated Components of Change: April 1, 2000 to July 1, 2006.
- County level.
- Source : web site of the Census Bureau in csv format.

# Example 1: The Annual County Population

**Table:** Metropolitan and Micropolitan Statistical Area Population Estimates File for Internet Display: source Census Bureau

Variable	Description
CTYNAME	County name
STNAME	State name
CENSUS2000POP	4/1/2000 resident Census 2000 population
POPESTIMATE2005	7/1/2005 resident total population estimate
POPESTIMATE2006	7/1/2006 resident total population estimate
BIRTHS2000	Births 4/1/2000 to 7/1/2000
BIRTHS2006	Births 7/1/2005 to 7/1/2006
DEATHS2000	Deaths 4/1/2000 to 7/1/2000
DEATHS2006	Deaths 7/1/2005 to 7/1/2006
INTERNATIONALMIG2000	Net international migration 4/1/2000 to 7/1/2000
INTERNATIONALMIG2006	Net international migration 7/1/2005 to 7/1/2006



# Example 1: The Annual County Population

- A sample of size  $n = 400$  from the population of  $N = 3141$  counties.
- Inclusion probabilities proportional to the variable `popestimate2006`.
- All the auxiliary variables are used as balancing variables.
- The R sampling package (see Tillé & Matei, 2021) allows the sample to be selected directly.
- Function `samplecube(X, pik)`.

## Example 1: The Annual County Population

```
# loading the sampling package library(sampling)
# reading of the file
V=read.csv("C://CO-EST2006-ALLDATA.csv", header = TRUE,
           sep=";", dec=".", fill = TRUE,comment.char="")
attach(V)
# definition of the matrix of balancing variables
X=cbind(
  popestimate2006,
  births2000,births2006,
  deaths2000,deaths2006,
  internationalmig2000,internationalmig2006,
  one=rep(1,length(popestimate2006))
)
# selection of the counties X=X[sumlev==50,]
# definition of the vector of inclusion probabilities
pik=inclusionprobabilities(X[,1],400)
# selection of the sample
s=samplecube(X,pik)
```

# Example 1: The Annual County Population

## BEGINNING OF THE FLIGHT PHASE

The matrix of balanced variable has 8 variables and 3141 units

The size of the inclusion probability vector is 3141

The sum of the inclusion probability vector is 400

The inclusion probability vector has 3038 non-integer elements

Step 1

## BEGINNING OF THE LANDING PHASE

At the end of the flight phase, there remain 8 non integer probabilities

The sum of these probabilities is 4

This sum is integer

The linear program will consider 70 possible samples

The mean cost is 0.005863816

The smallest cost is 0.001072027

The largest cost is 0.00987782

The cost of the selected sample is 0.002267229

QUALITY OF BALANCING	TOTALS	HorvitzThompson_estimators	Relative_deviation
poestimate2006	299398484	2.993985e+08	-2.189894e-13
births2000	989020	9.897919e+05	7.804513e-02
births2006	4151889	4.154413e+06	6.079854e-02
deaths2000	560891	5.599909e+05	-1.604724e-01
deaths2006	2464633	2.462009e+06	-1.064653e-01
internationalmig2000	364221	3.658406e+05	4.446634e-01
internationalmig2006	1204167	1.209396e+06	4.342091e-01
one	3141	3.183646e+03	1.357711e+00

# Example 1: The Annual County Population

## The Selected Sample : Virginia, Washington, Wisconsin, Wyoming

	Name	State	Incl. prob.	Pop2006
371	Fairfax County	Virginia	1.000000000	1010443
372	Giles County	Virginia	0.030210273	17403
373	Greensville County	Virginia	0.019105572	11006
374	Hanover County	Virginia	0.171826896	98983
375	Loudoun County	Virginia	0.466645695	268817
376	Rockingham County	Virginia	0.125965539	72564
377	Stafford County	Virginia	0.208605903	120170
378	Alexandria city	Virginia	0.237776359	136974
379	Danville city	Virginia	0.079133800	45586
380	Hampton city	Virginia	0.251738390	145017
381	Norfolk city	Virginia	0.397720860	229112
382	Richmond city	Virginia	0.334882172	192913
383	Suffolk city	Virginia	0.140733038	81071
384	Virginia Beach city	Virginia	0.756201174	435619
385	King County	Washington	1.000000000	1826732
386	Pierce County	Washington	1.000000000	766878
387	Snohomish County	Washington	1.000000000	669887
388	Spokane County	Washington	0.775447356	446706
389	Yakima County	Washington	0.404652402	233105
390	Marshall County	West Virginia	0.058840856	33896
391	Dane County	Wisconsin	0.805166363	463826
392	Kenosha County	Wisconsin	0.281221311	162001
393	Langlade County	Wisconsin	0.035813834	20631
394	Lincoln County	Wisconsin	0.052339824	30151
395	Milwaukee County	Wisconsin	1.000000000	915097
396	Outagamie County	Wisconsin	0.299852976	172734
397	Shawano County	Wisconsin	0.071868961	41401
398	Wood County	Wisconsin	0.129801929	74774
399	Laramie County	Wyoming	0.148220075	85384
400	Sweetwater County	Wyoming	0.067289595	38763

## Example 2: 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 2: 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  $\hat{X}_{j\pi}$ ,
- the relative deviation in % defined by

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

## Example 2: 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

## Example 2: Results

```
rm(list=ls())
library(sampling)
data(swissmunicipalities)
attach(swissmunicipalities)
summary(swissmunicipalities)
#
ONE=rep(1,nrow(swissmunicipalities))
X=cbind(POPTOT,ONE,HAPoly,P00BMTOT,P00BWTOT,
        Pop020,Pop2040,Pop4065,Pop65P,H00PTOT)
X=X[CT==21,] # selection of Ticino
pik=inclusionprobabilities(POPTOT[CT==21],50)
Nom[CT==21][pik==1]
#
s=samplecube(X,pik)
```



## Application: France's New Census

- Selection of the rotation groups for the French census.
- **Small municipalities (<10,000 inhabitants)**  
Five non-overlapping rotation groups were selected using a balanced sampling design with equal inclusion probabilities ( $1/5$ ). Each year, a fifth of the municipalities are surveyed.
- **Big municipalities (>10,000 inhabitants)**  
Five non-overlapping balanced samples of addresses are selected with inclusion probabilities  $1/8$ . So, after 5 years, 40% of the addresses are visited.
- The balancing variables are socio-demographic variables of the last Census.

# French Master Sample

- The primary units are geographical areas that are selected using a balanced sampling design.
- Self-weighted multi-stage sampling.
- So the primary units are selected with unequal probabilities proportional to their sizes.
- The balancing variables are socio-demographic variables of the least census.

# An even larger simulation

```
library(sampling)
data(belgianmunicipalities)
attach(belgianmunicipalities)
plot(Tot04/1000,Totaltaxation/1000000)
summary(belgianmunicipalities)
N=length(Totaltaxation)
n=100
pi=n/N
pik=inclusionprobabilities(Tot04,n)
X=cbind(pik,rep(1,N), DiffTOT,Women04, TaxableIncome,disjunctive(Province)*pik)
TT=Totaltaxation/1000000
SIM=100
ESTHT4=rep(0,SIM)
ESTHT6=rep(0,SIM)
ESTHT8=rep(0,SIM)
for(i in 1:SIM)
{
print(i)
ESTHT4[i]=sum( (TT/pik)*UPrandomsystematic(pik) )
ESTHT6[i]=sum( (TT/pik)*UPmaxentropy(pik))
ESTHT6[i]=sum( (TT/pik)*UPrandompivotal(pik))
ESTHT8[i]=sum( (TT/pik)*samplecube(X,pik,comment=FALSE))
}
boxplot(ESTHT4,ESTHT5,ESTHT6,ESTHT8,
names=c("R. syst","Max. entrop.,""R. piv.,""Balanced"))
abline(a=sum(TT),b=0)
```

# Variance Approximation by a Residual Technique

- Deville & Tillé (2005) proposed an approximation:
- Idea: the balanced sampling design is a Poisson sampling design conditionally to the balanced constraints.
- Assuming that  $(\widehat{Y}, \widehat{\mathbf{X}}^\top)^\top$  has a multivariate normal distribution, a simple reasoning allows us to compute:

$$\text{var}_p(\widehat{Y}) = \text{var}_{\tilde{p}}(\widehat{Y} | \widehat{\mathbf{X}} = \mathbf{X}),$$

where  $\tilde{p}(\cdot)$  is the Poisson design and  $p(\cdot)$  is the balanced design.

# Variance Approximation by a Residual Technique

- Approximation of the variance

$$\text{var}_p(\hat{Y}) \cong \text{var}_{app}(\hat{Y}) = \sum_{k \in U} b_k \frac{(y_k - \mathbf{x}_k^\top \mathbf{b})^2}{\pi_k^2},$$

where

$$\mathbf{b} = \left( \sum_{k \in U} b_k \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k^2} \right)^{-1} \sum_{k \in U} b_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

and  $b_k = \frac{N}{N-1} \pi_k (1 - \pi_k)$ .

# Estimation of Variance

- Deville & Tillé (2005) proposed a family of variance estimators

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k \in S} c_k \frac{(y_k - \mathbf{x}_k^\top \widehat{\mathbf{b}})^2}{\pi_k^2},$$

where

$$\widehat{\mathbf{b}} = \left( \sum_{\ell \in S} c_\ell \frac{\mathbf{x}_\ell \mathbf{x}_\ell^\top}{\pi_\ell^2} \right)^{-1} \sum_{\ell \in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

and

$$c_k = \frac{n}{n - q} (1 - \pi_k).$$

# Sampling in Space and Spread Sampling

## Model for spatial sampling

- Model for spatial sampling

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad (3)$$

$E(\varepsilon_k) = 0$ ,  $\text{var}(\varepsilon_k) = \sigma_k^2$  and  $\text{cov}(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{kl}$  when  $k \neq \ell$ .

- The model thus admits heteroscedasticity and autocorrelation.
- 

$$\text{AVar}(\hat{Y}) = E_p \left( \sum_{k \in S} \frac{\mathbf{x}_k^\top \boldsymbol{\beta}}{\pi_k} - \sum_{k \in U} \mathbf{x}_k^\top \boldsymbol{\beta} \right)^2 + \sum_{k \in U} \sum_{\ell \in U} \Delta_{kl} \frac{\sigma_k \sigma_\ell \rho_{kl}}{\pi_k \pi_\ell}.$$

Optimal design:

- ▶ using inclusion probabilities proportional to  $\sigma_k$ ,
- ▶ using a balanced sampling design on the auxiliary variables  $x_1, \dots, x_p$ .
- ▶ avoiding the selection of neighboring units, that is, selecting a well-spread sample (or spatially balanced)

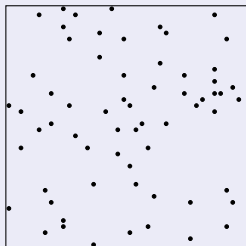


# Usual methods (Wang, Stein, Gao & Ge, 2012)

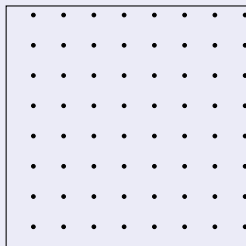
- Usual methods can be used: simple, stratified, cluster, two-stage sampling.
- Stratification can improve the spreading.
- Central role of systematic sampling (because spread).

# Usual methods

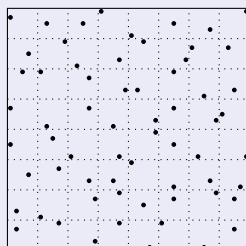
## Usual methods



simple design



systematic



stratification

# Biodiversity Monitoring

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



Table: WSL Swiss biodiversity Monitoring



# Biodiversity Monitoring



Table: Swiss biodiversity Monitoring: number of neophytes in the plots

# 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

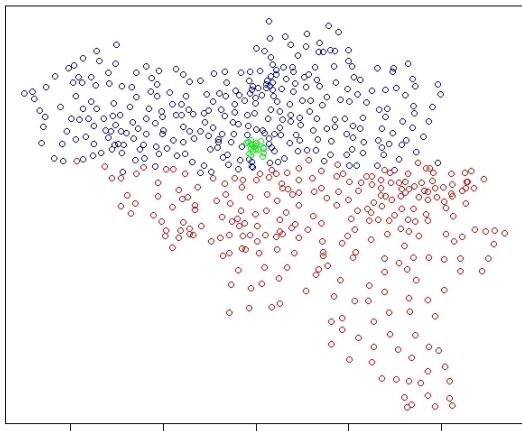


Figure: Centers of the Belgian Municipalities (Data IGN Belgium)

# Generalized Random Tessellation Sampling

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

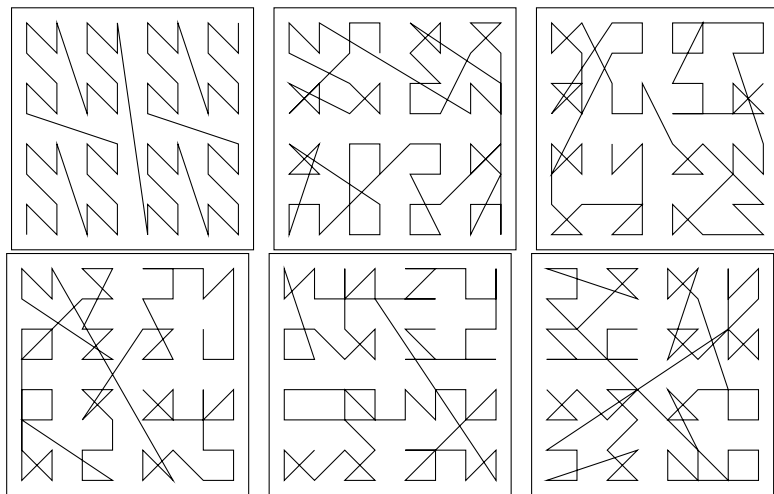
- 1 Create a hierarchical grid with addresses. Cut the squares till having maximum one unit per square.
- 2 Construct a sampling line using the addresses (quadrant-recursive).
- 3 Randomize the addresses.
- 4 Select a systematic sample on the line.

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





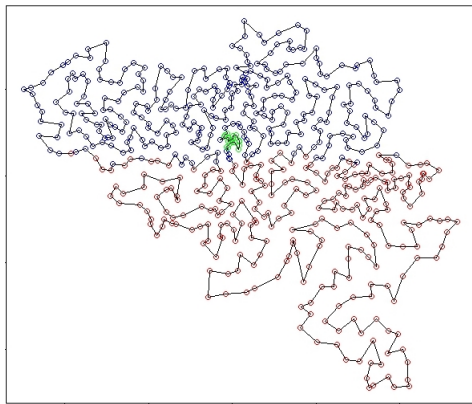
# Generalized Random Tessellation Sampling



# 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

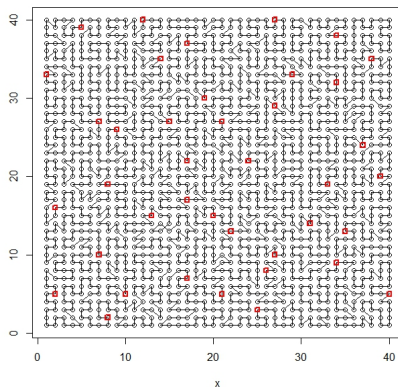
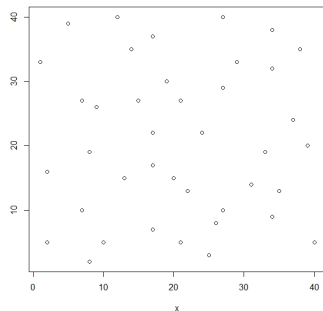


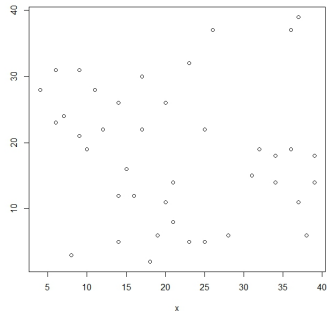
Table: Grid  $40 \times 40$ . Selection of 40 points.

# Travelling Salesman Problem

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.

- [Exempel\\_canvas1.html](#) Example 1
- [Exempel\\_canvas2.html](#) Example 2
- any sentence

# Local Cube Method (Grafström & Tillé, 2013)

- Cube method (Deville & Tillé, 2004) to obtain balanced sample
$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k.$$
- The cube method is composed of two phases
  - ▶ Flight phase
  - ▶ Landing phase
- In the flight phase, at each step the balancing equation are satisfied. Vector  $\pi$  is modified randomly. A component of  $\pi$  of the vector of inclusion probability is set to either 0 or 1.
- Idea: At each step apply the flight phase on a subset of  $p + 1$  neighboring units. ( $p$  is the number of balancing variables).
- The sample will be well spread and balanced.

# Algorithm for spread and balanced sampling (doubly balanced)

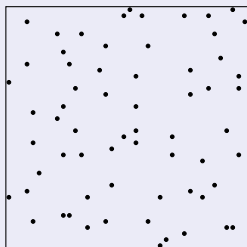
- Let  $p$  the number of auxiliary variables.
- For the cube method, the dimension of the subspace of constraints is  $N - p$ .
- In order to run a step of the flight phase of the cube method, the population size must have at least  $p + 1$  units.

**Algorithm** Repeat these steps:

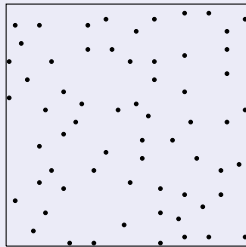
- (1) Select a set of  $p + 1$  neighboring units that have inclusion probabilities strictly between 0 and 1.
- (2) Run one step of the flight phase on these units.

# Complex methods

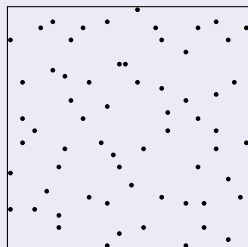
## Complex methods



GRTS



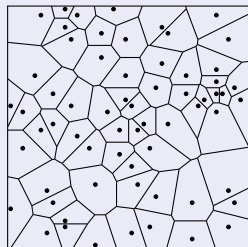
local pivotal



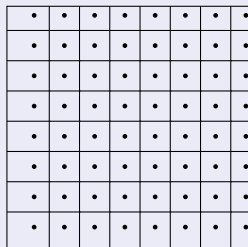
local cube



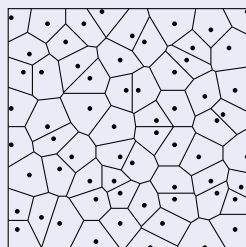
# Voronoi polygons



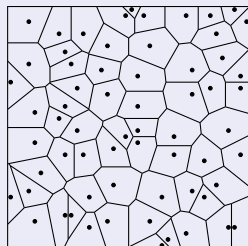
simple



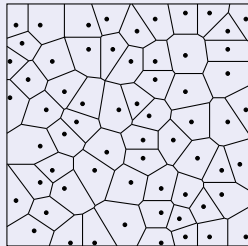
systematic



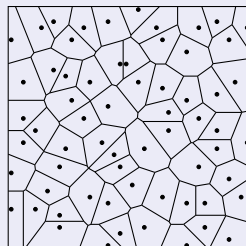
stratification



GRTS



pivot local



cube local

# Voronoi polygons

An index for spatial balance: (Stevens Jr. & Olsen, 2003, 2004; Theobald, Stevens Jr., White, Urquhart, Olsen & Norman, 2007)

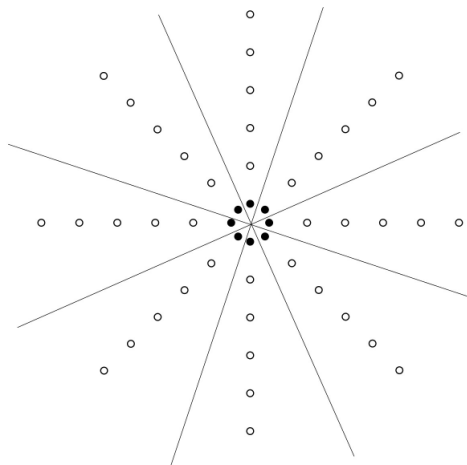
- Compute the Voronoi polygons for the unit selected in the samples.
- Compute  $v_k$  the sum of the inclusion probabilities of the population units that are in the Voronoi polygon of unit  $k \in S$ .

• Since  $\sum_{k \in S} v_k = \sum_{k \in U} \pi_k = n$ , we have  $\bar{v} = \frac{1}{n} \sum_{k \in S} v_k = 1$ .

• The index of spatial balance:  $\frac{1}{n} \sum_{k \in S} (v_k - 1)^2$ .

# Voronoi polygons

This index can have problems:



## 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

# 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  $\mathbf{a} = (0 \ 1 \ 0 \ 0 \ 1 \ \dots \ 0)$ .
  - ▶ 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

## An alternate measure of spreading based on the Moran index

- Use of a contiguity matrix  $\mathbf{W}$ . Compute the mean of the neighboring units  $\bar{\mathbf{a}}_w$ .

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

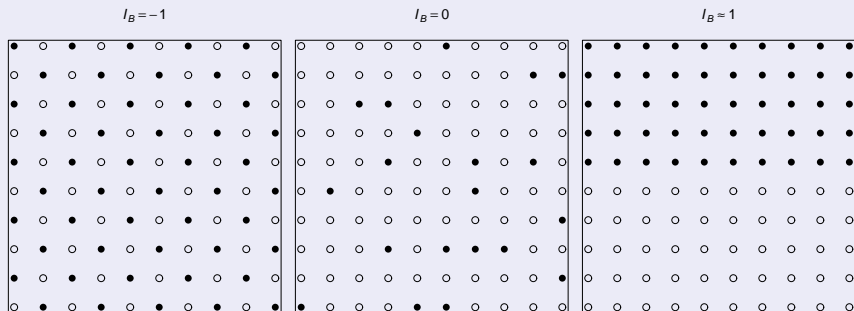
- where  $\mathbf{W} = (w_{ij})$ ,  $w_{ij}$  indicates how close is  $j$  to  $i$ ,  $w_{ii} = 0$ ,  $\mathbf{D}$  be the diagonal matrix containing  $\sum_{j \in U} w_{ij} = w_i$ .

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

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

# An alternate measure of spreading based on the Moran index

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



# A maybe better idea

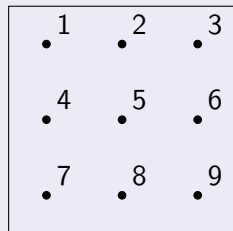
## A maybe better idea

- Grafström & Lundström (2013)  
“Why well spread probability samples are balanced?”.
- Spread samples are balanced everywhere.
- Spread samples are locally stratified.
- Why not to stratify everywhere? Wave method (Jauslin & Tillé, 2020; Jauslin & Tillé, 2019).



# The wave method 1: Definition of a distance (Jauslin & Tillé, 2020; Jauslin & Tillé, 2019)

Example of points in a space. One can use an Euclidian distance



# The wave method 2: Definition of a distance

## Tore distance

<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9	<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9	<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											
<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9	<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9	<table><tr><td>• 1</td><td>• 2</td><td>• 3</td></tr><tr><td>• 4</td><td>• 5</td><td>• 6</td></tr><tr><td>• 7</td><td>• 8</td><td>• 9</td></tr></table>	• 1	• 2	• 3	• 4	• 5	• 6	• 7	• 8	• 9
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											
• 1	• 2	• 3																											
• 4	• 5	• 6																											
• 7	• 8	• 9																											

## The wave method 3: Matrices of squares of the Euclidian and Tore distances

Let  $\{1, \dots, 9\}$  be on a regular grid of size  $3 \times 3$ .

$$\mathbf{M}_E = \begin{pmatrix} 0 & 1 & 4 & 1 & 2 & 5 & 4 & 5 & 8 \\ 1 & 0 & 1 & 2 & 1 & 2 & 5 & 4 & 5 \\ 4 & 1 & 0 & 5 & 2 & 1 & 8 & 5 & 4 \\ 1 & 2 & 5 & 0 & 1 & 4 & 1 & 2 & 5 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 5 & 2 & 1 & 4 & 1 & 0 & 5 & 2 & 1 \\ 4 & 5 & 8 & 1 & 2 & 5 & 0 & 1 & 4 \\ 5 & 4 & 5 & 2 & 1 & 2 & 1 & 0 & 1 \\ 8 & 5 & 4 & 5 & 2 & 1 & 4 & 1 & 0 \end{pmatrix}, \quad \mathbf{M}_T = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 & 1 & 2 & 2 & 1 & 2 \\ 1 & 1 & 0 & 2 & 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 & 1 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 2 & 2 & 1 & 1 & 1 & 0 & 2 & 2 & 1 \\ 1 & 2 & 2 & 1 & 2 & 2 & 0 & 1 & 1 \\ 2 & 1 & 2 & 2 & 1 & 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 2 & 2 & 1 & 1 & 1 & 0 \end{pmatrix}. \quad (4)$$

# The wave method 4: Definition of a distance

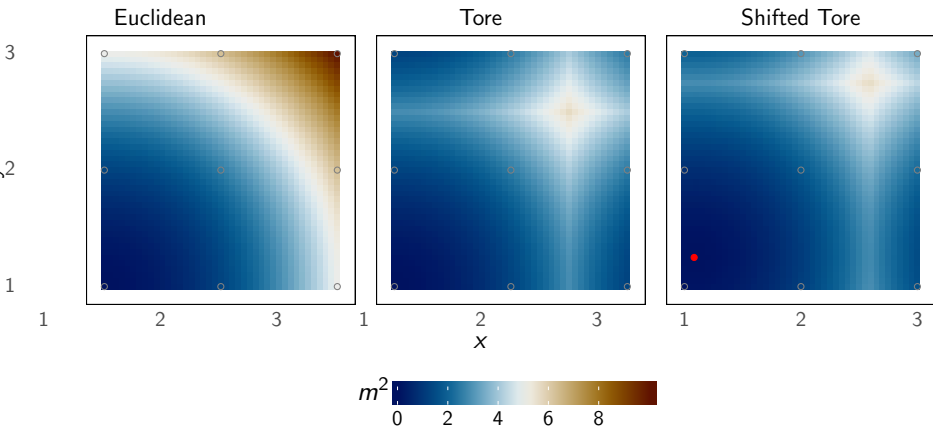
## Tore distance with a shift

1 •	2 •	3 •	1 •	2 •	3 •	1 •	2 •	3 •
4 •	5 •	6 •	4 •	5 •	6 •	4 •	5 •	6 •
7 •	8 •	9 •	7 •	8 •	9 •	7 •	8 •	9 •
1 •	2 •	3 •	1 •	2 •	3 •	1 •	2 •	3 •
4 •	5 •	6 •	4 •	5 •	6 •	4 •	5 •	6 •
7 •	8 •	9 •	7 •	8 •	9 •	7 •	8 •	9 •

## The wave method 5: Distance with a shift (no ties)

$$\mathbf{M}_S = \begin{pmatrix} 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 \\ 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 \\ 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 \\ 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 \\ 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 \\ 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 \\ 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 \\ 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 \\ 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 \end{pmatrix} .$$

# The wave method 6: Example of distance



# The wave method 7: Definition of a contiguity matrix

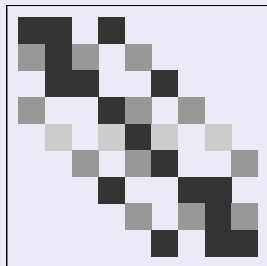
## Contiguity

$$w_{kl} = \begin{cases} \pi_l & \text{if unit } l \text{ is in } g_k - 1 \text{ nearest neighbour of } k, \\ \pi_l + 1 - \sum_{j \in G_k} \pi_k & \text{if unit } l \text{ is the } g_k \text{th nearest neighbour of } k, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

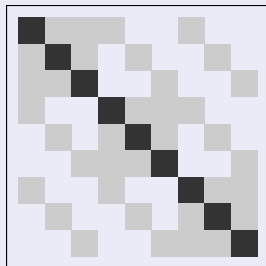
# The wave method 8: Contiguity matrices

## Contiguity

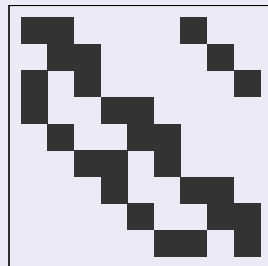
Euclidean



Tore



Shifted Tore



$w_{kl}$   1/6  2/9  1/3

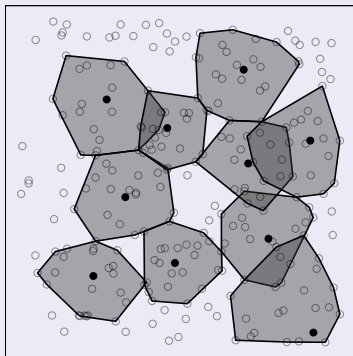
$\pi_k = 1/3$ . The three contiguity matrices.



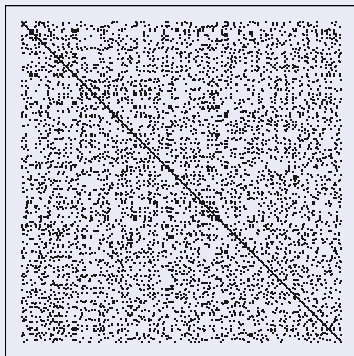
# Contiguity matrices

## Contiguity

Initial strata



Stratification matrix



The strata defined by a contiguity matrix.  
Ideally, one selects one unit per stratum.

## The wave method 9: selection of the sample by using weakly associated vectors

- Based of the cube method: to find a sample modify the vector  $\pi_k$  orthogonally to the constraints (the strata).
- Find a vector that is weakly associated with the contiguity matrix.
- Modify randomly the vector of inclusion probabilities by following the direction of this vector.

# Algorithm for WAVE sampling

Let  $\mathbf{W} \text{diag}(\boldsymbol{\pi})^{-1} = \mathbf{A} = \mathbf{A}_0$  and  $\boldsymbol{\pi}_0 = (\pi_1^{(0)}, \dots, \pi_N^{(0)}) = \boldsymbol{\pi}$  for the initialization step. For  $t = 0, 1, 2, \dots$

- 1 From  $\boldsymbol{\pi}_t$ , extract  $\tilde{\boldsymbol{\pi}}_t$  vector  $\boldsymbol{\pi}_t$  restricted to the  $k$  such that  $0 < \pi_k^{(t)} < 1$ . Let  $J$  be the length of  $\tilde{\boldsymbol{\pi}}_t$ .
- 2 Compute the  $J \times J$  matrix  $\mathbf{A}_t$  using inclusion probabilities  $\tilde{\boldsymbol{\pi}}_t$ .
- 3 Calculate the rank  $r$  of matrix  $\mathbf{A}_t$ .
  - 1 If matrix  $\mathbf{A}_t$  does not have full rank, choose  $\mathbf{v}_t = (v_1^{(t)}, \dots, v_J^{(t)}) \in \mathbb{R}^J$  a vector in the right null space of  $\mathbf{A}_t$ .
  - 2 If matrix  $\mathbf{A}_t$  has full rank, compute the singular value decomposition and seek for  $\mathbf{v}_t$  a right singular vector associated to the smallest singular value  $\sigma_t$ .
- 4 Next in order to ensure the fixed sample size, vector  $\mathbf{v}_t$  is centered:

$$\tilde{\mathbf{v}}_t = \mathbf{v}_t - \frac{1}{J} \sum_{i \in J} v_i^{(t)} \mathbf{1}_J,$$

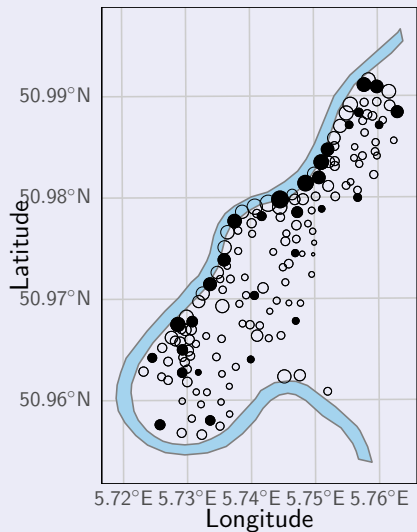
where  $\mathbf{1}_J$  is the  $J \times 1$  vector of one.

- 5 Find  $\lambda_1$  and  $\lambda_2$  the largest positive real numbers such that all the  $0 \leq \tilde{\pi}_k^{(t)} + \lambda_1 \tilde{v}_k^{(t)} \leq 1$  and  $0 \leq \tilde{\pi}_k^{(t)} - \lambda_2 \tilde{v}_k^{(t)} \leq 1$ ,  $k = 1, \dots, J$ .
- 6 Compute

$$\boldsymbol{\pi}_{t+1} = \begin{cases} \tilde{\boldsymbol{\pi}}_t + \lambda_1 \tilde{\mathbf{v}}_t & \text{with probability } \lambda_2 / (\lambda_1 + \lambda_2) \\ \tilde{\boldsymbol{\pi}}_t - \lambda_2 \tilde{\mathbf{v}}_t & \text{with probability } \lambda_1 / (\lambda_1 + \lambda_2). \end{cases}$$

- 7 Return at (a) with  $\boldsymbol{\pi}_{t+1}$  until no units  $k$  remains such that  $0 < \pi_k^{(t+1)} < 1$ .

# The wave method 10: Meuse data set



Copper • 25 ● 50 ● 75 ● 100 ● 125

# The wave method 11: Spreading measures results based on 10000 simulations on the Meuse dataset

	Sampling design									
	Equal probabilities					Unequal probabilities				
	wave	lpm1	scps	grts	srswor	wave	lpm1	scps	grts	maxent
$I_{B_1}$										
$n = 15$	-0.518	-0.337	-0.352	-0.224	-0.030	-0.340	-0.250	-0.246	-0.162	-0.004
$n = 30$	-0.663	-0.428	-0.429	-0.267	-0.018	-0.407	-0.297	-0.288	-0.172	0.024
$n = 75$	-0.926	-0.624	-0.508	-0.365	-0.007	-0.537	-0.372	-0.328	-0.226	0.033
$I_B$										
$n = 15$	-0.518	-0.337	-0.352	-0.224	-0.030	-0.355	-0.244	-0.248	-0.151	0.006
$n = 30$	-0.663	-0.428	-0.429	-0.267	-0.018	-0.427	-0.290	-0.284	-0.153	0.047
$n = 75$	-0.926	-0.624	-0.508	-0.365	-0.007	-0.532	-0.369	-0.315	-0.216	0.040
$B$										
$n = 15$	0.119	0.126	0.118	0.170	0.382	0.115	0.121	0.120	0.171	0.383
$n = 30$	0.119	0.123	0.125	0.163	0.357	0.120	0.120	0.120	0.162	0.345
$n = 75$	0.162	0.148	0.153	0.184	0.295	0.115	0.113	0.120	0.135	0.210

## The wave method 12: R package

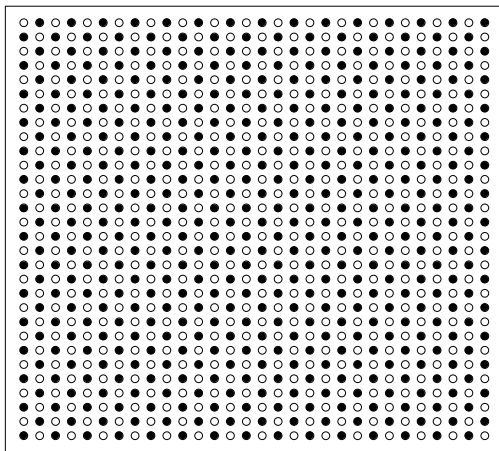
```
install.packages("WaveSampling")
library(WaveSampling)
N <- 200
n <- 80
X <- as.matrix(cbind(runif(N),runif(N)))
pik <- sampling::inclusionprobabilities(runif(N),n)
s <- wave(X,pik)
plot(X)
points(X[s==1,],pch=16)
```

## The wave method 13: Comment

- Possibility of obtaining the most spread sample.
- Computer intensive.
- Implementation in R Package: `WaveSampling`.

# The wave method 14: Examples of periodic sample in a grid

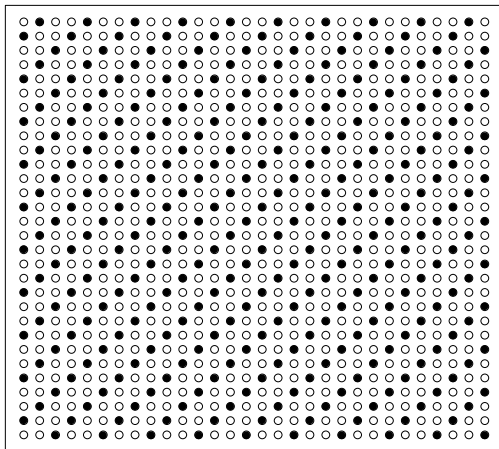
$$\pi_k = 1/2$$





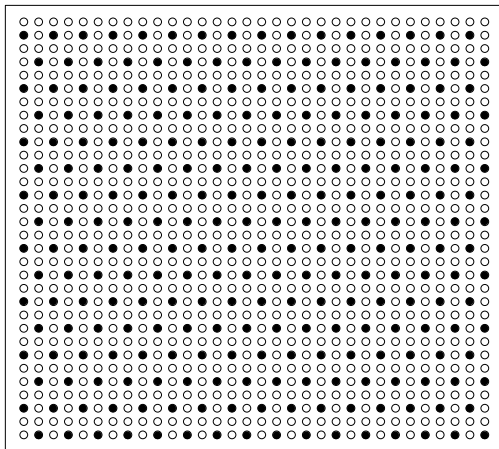
## Examples of periodic sample in a grid

$$\pi_k = 1/3$$



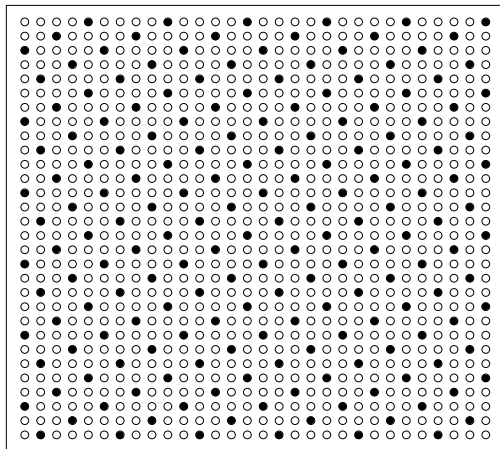
## Examples of periodic sample in a grid

$$\pi_k = 1/4$$



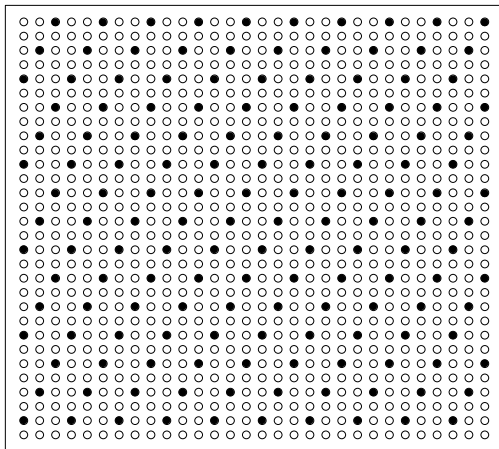
## Examples of periodic sample in a grid

$$\pi_k = 1/5$$



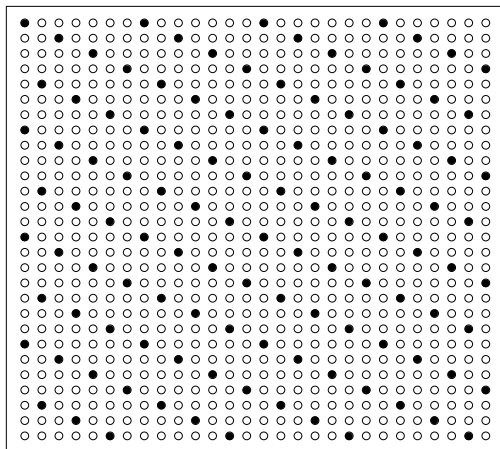
## Examples of periodic sample in a grid

$$\pi_k = 1/6$$



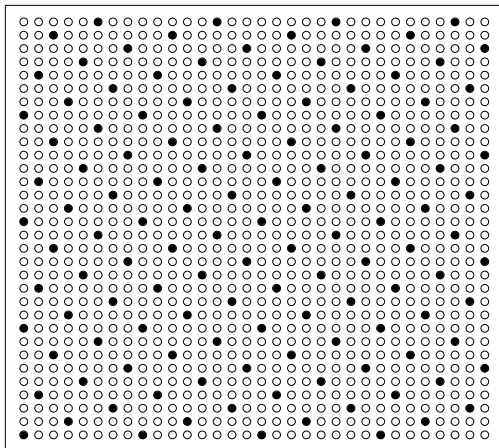
## Examples of periodic sample in a grid

$$\pi_k = 1/7$$



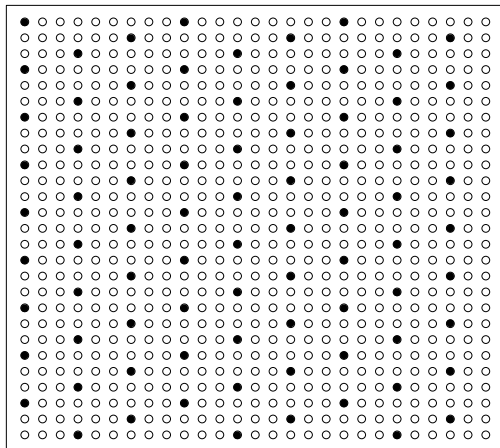
## Examples of periodic sample in a grid

$$\pi_k = 1/8$$



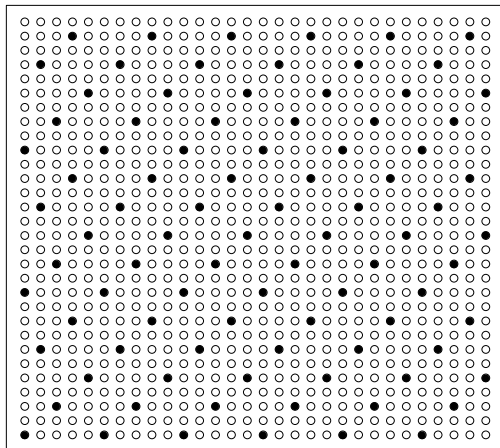
## Examples of periodic sample in a grid

$$\pi_k = 1/9$$



## Examples of periodic sample in a grid

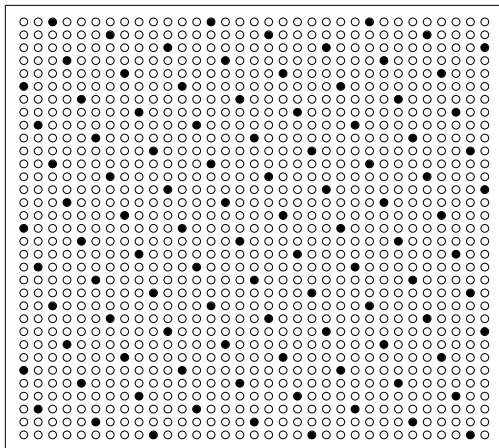
$$\pi_k = 1/10$$





## Examples of periodic sample in a grid

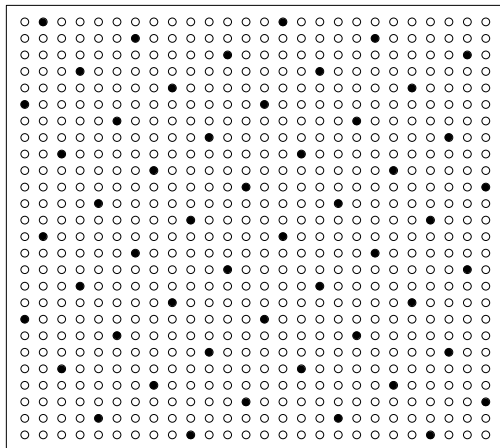
$$\pi_k = 1/11$$





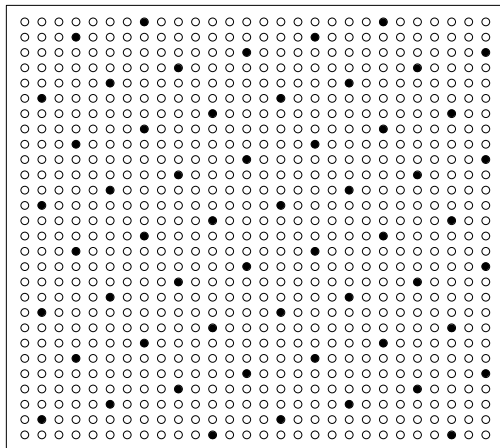
## Examples of periodic sample in a grid

$$\pi_k = 1/13$$



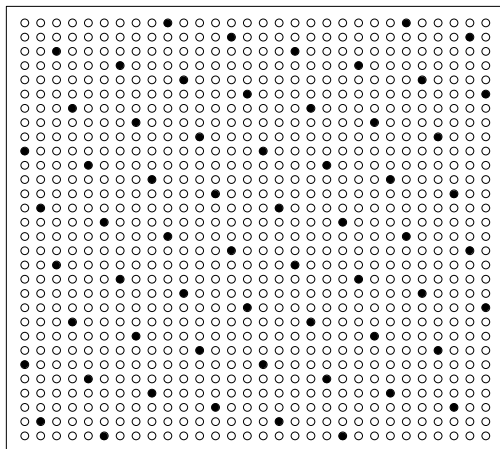
## Examples of periodic sample in a grid

$$\pi_k = 1/14$$



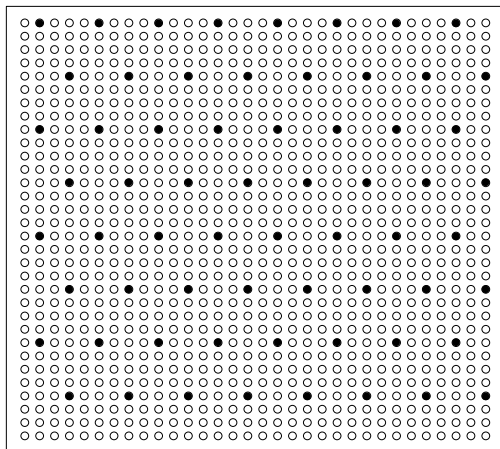
# Examples of periodic sample in a grid

$$\pi_k = 1/15$$



## Examples of periodic sample in a grid

$$\pi_k = 1/16$$



# Spatiotemporal spread sampling

- Rivest & Ebouele (2020). Bivariate sampling.
- Eustache, Jauslin & Tillé (2020). Spatiotemporal sampling with spreading and negative coordination.

## Example: spot method

```
install.packages("SpotSampling")
library(SpotSampling)
## Coordinates in two dimensions of 4 units ##
coord <- matrix(c(0.5,0.6,0.2,0.3,0.8,0.9,0.4,0.7), ncol=2)
## Temporal inclusion probabilities ##
## with 3 waves and 4 units ##
pik <- matrix(c(0.6,0.3,0.3,
0.2,0.4,0.9,
0.3,0.2,0.5,
0.9,0.1,0.3), ncol = 3, byrow = TRUE)
## SPOT method ##
Spot(pik, coord, EPS = 1e-6)
```



# Back to the Examples

# Back to the examples: Example 1.

## New repayment system for the hospitals.

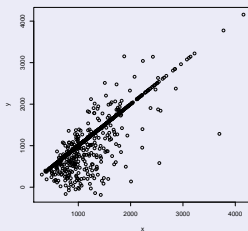
### Diagnostic Related Groups (RDG)

- Reimbursement based on “Major Diagnostic Categories” .
- Auditors to control the codification of the hospital (fraud, codification error).
- Selection of a sample of medical records.
- Two aims.
  1. Estimation of the amount of errors.
  2. Improvement of the codification.

# Back to the examples: Example 1.

## New repayment system for the hospitals (RDG)

- Errors are rare.
- Optimization of the sampling design (Marazzi & Tillé, 2017).
- Oversampling when an error is suspected.
- Example with simulated data.

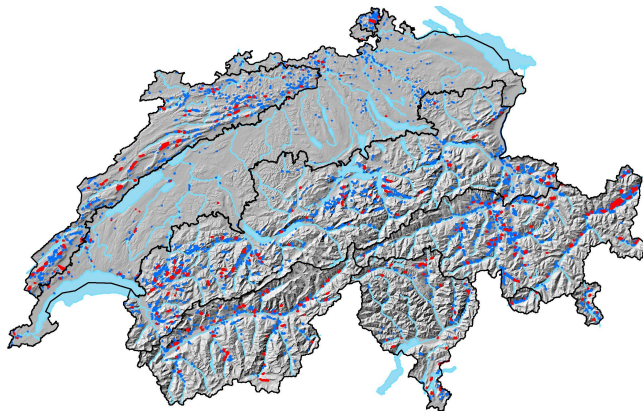


- Two-stage model. Balanced sampling with unequal inclusion probabilities.

## Back to the examples: Example 2.

Selection of a sample of 2100 circles of  $10\text{m}^2$  in the dry grasslands (Tillé & Ecker, 2013)

Analysis of the biodiversity with interest for atypical regions.

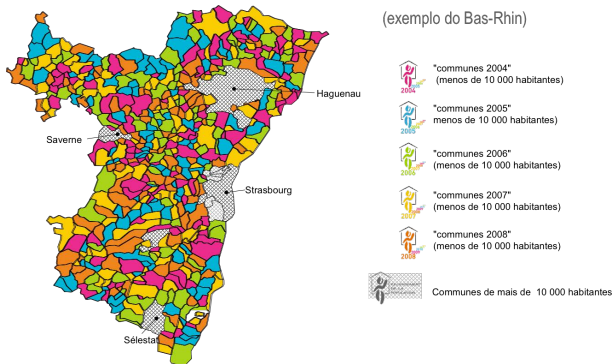


Balanced spread sampling with inclusion probabilities proportional to an index of biodiversity.

## Back to the examples: Example 3.

New French Rolling Census For the small municipalities (less than 10000 inhabitants), five rotations groups are created.

One group is surveyed each year.



From Durr & Clanché (2013).

Balanced sampling with equal inclusion probabilities.

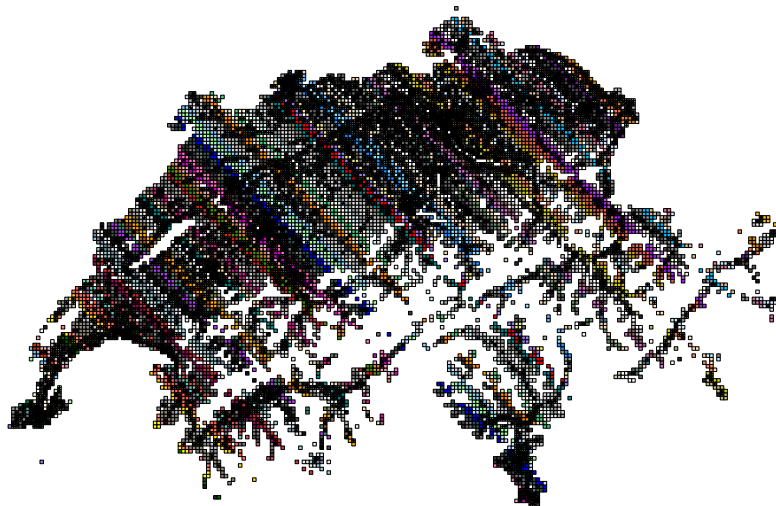
## Back to the examples: Example 4.

### Swiss Census System

- In Switzerland, the old census method has been abolished.
- Data are collected from registers (Control of the inhabitants, buildings).
- Quality Survey of the National Census System.
- Important to evaluate the quality of the new method.

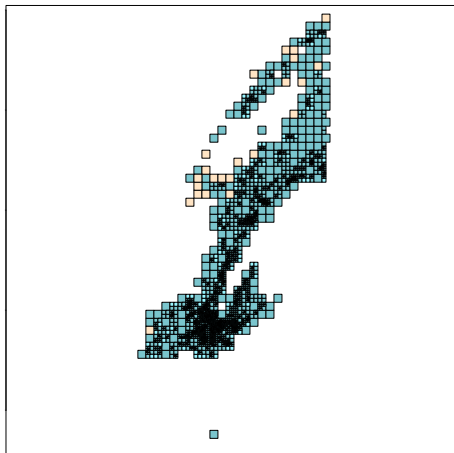
# Back to the examples: Example 4.

Swiss Census System



## Back to the examples: Example 4.

Swiss Census System




The sampling design is optimized: balanced spread sampling with inclusion probabilities proportional to the number of buildings.



# THANK YOU

# References

- ARDILLY, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique* **23**, 91–113.
- ARDILLY, P. & TILLÉ, Y. (2006). *Sampling Methods: Exercises and Solutions*. New York: Springer.
- BEBBINGTON, A. C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics* **24**, 136.
- BOISTARD, H., LOPUHAÄ, H. P. & RUIZ-GAZEN, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics* **6**, 1967–1983.
- 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.
- BREIDT, F. J. & CHAUVET, G. (2012). Penalized balanced sampling. *Biometrika* **99**, 945–958.
- BREWER, K. R. W. (2013). Three controversies in the history of survey sampling. *Survey Methodology* **39**, 249–262.
- BREWER, K. R. W. & HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer.
- CHAUVET, G. (2009). Stratified balanced sampling. *Survey Methodology* **35**, 115–119.
- CHAUVET, G., BONNÉRY, D. & DEVILLE, J.-C. (2011a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference* **141**, 984–994.
- CHAUVET, G., DEVILLE, J.-C. & HAZIZA, D. (2010). Adapting the cube algorithm for balanced random imputation in surveys. Rapport technique, ENSAI, Rennes.
- CHAUVET, G., DEVILLE, J.-C. & HAZIZA, D. (2011b). On balanced random imputation in surveys. *Biometrika* **98**, 459–471.
- CHAUVET, G., HAZIZA, D. & LESAGE, É. (2015). Examining some aspects of balanced sampling in surveys. *Statistica Sinica* **25**, 313–334.
- CHAUVET, G. & TILLÉ, Y. (2005a). De nouvelles macros SAS d'échantillonnage équilibré. In *Actes des Journées de Méthodologie Statistique, Insee*. Paris.
- CHAUVET, G. & TILLÉ, Y. (2005b). *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.
- CHAUVET, G. & TILLÉ, Y. (2007). Application of the fast SAS macros for balancing samples to the selection of addresses. *Case Studies in Business, Industry and Government Statistics* **1**, 173–182.

- CHEN, X. (1993). Poisson-binomial distribution, conditional Bernoulli distribution and maximum entropy. Rapport technique, Department of Statistics, Harvard University.
- CHEN, X.-H., DEMPSTER, A. P. & LIU, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457–469.
- CHEN, X.-H. & LIU, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.
- CHOUDHRY, G. H. & SINGH, M. P. (1979). Sampling with unequal probabilities and without replacement – a rejective method. *Survey Methodology* **5**, 162–177.
- DEVILLE, J.-C. (1998). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Rapport Technique 9804, Méthodologie Statistique, Insee, Paris.
- DEVILLE, J.-C. (2000). Note sur l’algorithme de Chen, Dempster et Liu. Rapport technique, 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.
- DEVROYE, L. (1986). *Non-uniform Random Variate Generation*. New York: Springer.
- DICKSON, M. M. & TILLÉ, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics* **31**, 1359–1372.
- DUPAČOVÁ, J. (1979). A note on rejective sampling. In *Contribution to Statistics (Jaroslav Hájek memorial volume)*. Academia Prague.
- DURR, J.-M. & CLANCHÉ, F. (2013). The french rolling census: a decade of experience. In *59th ISI World Statistics Congress*. Citeseer.
- EUSTACHE, E., JAUSLIN, R. & TILLÉ, Y. (2020). Spatiotemporal spread sampling with optimal rotation: the spot sampling method. Institut de Statistique, Université de Neuchâtel.
- FAN, C. T., MULLER, M. E. & REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association* **57**, 387–402. 

- FULLER, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika* **96**, 933–944.
- FULLER, W. A., LEGG, J. C. & LI, Y. (2017). Bootstrap variance estimation for rejective sampling. *Journal of the American Statistical Association* **112**, 1562–1570.
- GINI, C. & GALVANI, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica Series 6*, **4**, 1–107.
- GRAFSTRÖM, A. & LISIC, J. (2019). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.5.
- 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.
- HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523.
- HÁJEK, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.
- HASLER, C. & TILLÉ, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis* **74**, 81–94.
- HEDAYAT, A. S. & MAJUMDAR, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference* **44**, 237–247.
- HODGES JR., J. L. & LE CAM, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics* **31**, 737–740.
- HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.
- JAUSLIN, R., EUSTACHE, E. & TILLÉ, Y. (2021). Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science*, Accepted for publication, 1–20.
- JAUSLIN, R. & TILLÉ, Y. (2019). *WaveSampling: Weakly Associated Vectors (WAVE) Sampling*. R package version 0.1.0.
- JAUSLIN, R. & TILLÉ, Y. (2020). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics* **25**, 431–451.

- KNUTH, D. E. (1981). *The Art of Computer Programming (Volume II): Seminumerical Algorithms*. Reading, MA: Addison-Wesley.
- LANGEL, M. & TILLÉ, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. *Metron* **69**, 45–65.
- LEGG, J. C. & YU, C. L. (2010). Comparison of sample set restriction procedures. *Survey Methodology* **36**, 69–79.
- LOHR, S. L. (2009). *Sampling: Design and Analysis*. Boston: Brooks/Cole.
- MADOW, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20**, 333–354.
- MARAZZI, A. & TILLÉ, Y. (2017). Using past experience to optimize audit sampling design. *Review of Quantitative Finance and Accounting* **49**, 435–462.
- MCLEOD, A. I. & BELLHOUSE, D. R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics* **32**, 182–184.
- 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.
- RIVEST, L.-P. & EBOUELE, S. E. (2020). Sampling a two dimensional matrix. *Computational Statistics & Data Analysis* **149**, 106971.
- ROYALL, R. M. & HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association* **68**, 880–889.
- SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- STEIN, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Rapport Technique 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. 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.

- THIONET, P. (1953). *La théorie des sondages*. Paris: Institut National de la Statistique et des Études Économiques, Études théoriques vol. 5, Imprimerie nationale.
- 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. (2016). The legacy of Corrado Gini in survey sampling and inequality theory. *Metron* **74**, 167–174.
- TILLÉ, Y. (2020). *Sampling and Estimation From Finite Populations*. Hoboken: Wiley.
- 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. & 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. & MATEI, A. (2021). *sampling: Survey Sampling*. R package version 2.9.
- TILLÉ, Y. & WILHELM, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science* **32**, 176–189.
- VALLIANT, R., DORFMAN, A. H. & ROYALL, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.
- VITTER, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software* **11**, 37–57.
- WANG, J.-F., STEIN, A., GAO, B.-B. & GE, Y. (2012). A review of spatial sampling. *Spatial Statistics* **2**, 1–14.
- YATES, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin.
- YATES, F. & GRUNDY, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* **B15**, 235–261.