Coordinated sampling: the current state and the research frontier

Alina Matei[1] and Anton Grafström[2]

[1]University of Neuchâtel, Switzerland and [2]Swedish University of Agricultural Sciences, Umea, Sweden

ENBES workshop,
Statistics Netherlands (CBS), The Hague, The Netherlands

Université de Neuchâtel

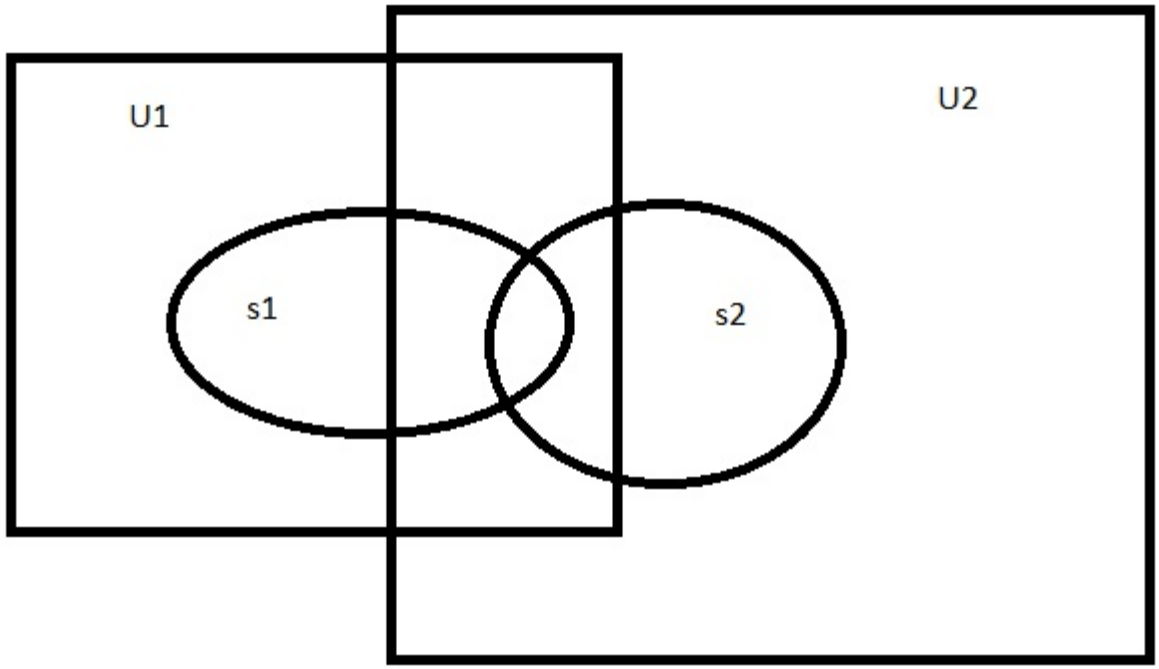
Overview

- Framework.
- Methods to coordinate samples.
- Poisson sampling with permanent random numbers.
- Methods to CP-samples and spatially balanced samples.
- Advantages/disadvantages of the methods.

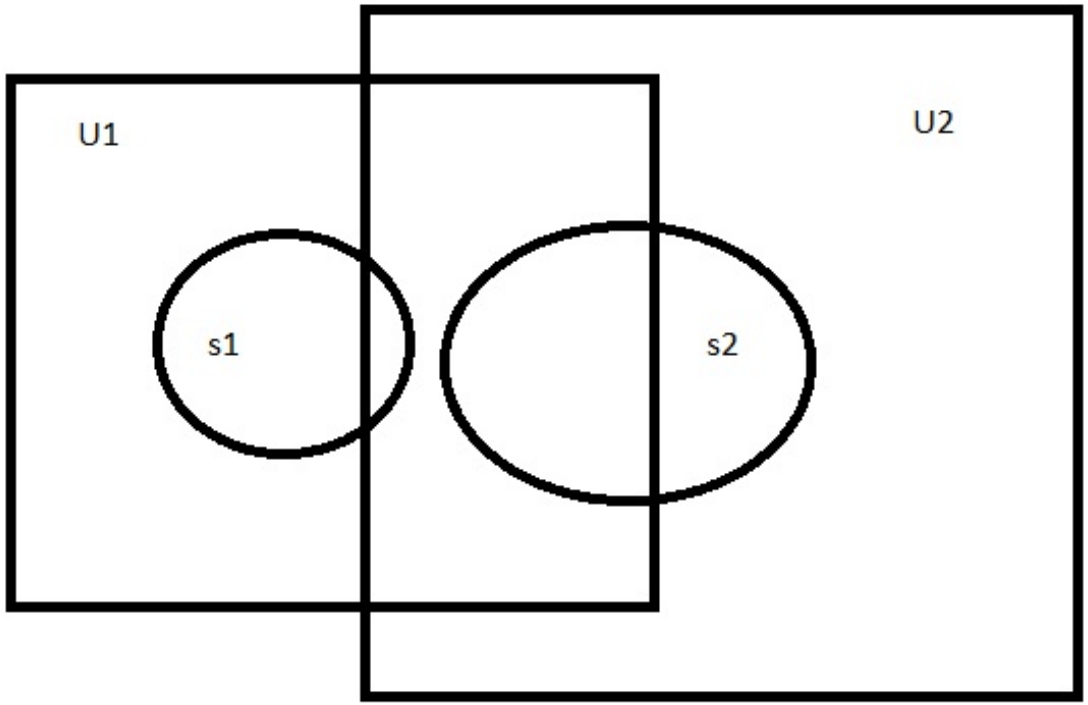
Sampling over time

- Consider a repeated survey over two time occasions 1 and 2.
- The finite population at time $t \in \{1, 2\}$ is U_t. U_1 and U_2 overlap.
- Samples s_t are selected from $U_t, t \in \{1, 2\}$.
- Let $\pi_{kt} = P(k \in s_t), k \in U_t, t \in \{1, 2\}$.

Overview - positive coordination



Overview - negative coordination



Sample coordination

- Sample coordination seeks to maximize/minimize the overlap between samples drawn in repeated surveys or several surveys (positive/negative coordination).
- Main goals:
  - positive coordination: reduce the variance of an estimator of change, reduce data collection costs;
  - negative coordination: diminish the response burden of the units that have a risk of being selected for several surveys.
- An important difficulty in sample coordination is due to changes in population definition. Thus, births, deaths, or splits of units frequently occur.

Coordination methods

- Sample coordination methods can be roughly divided into two categories: Permanent Random Numbers (PRN) methods and non-PRN methods
- PRN methods: assign to each unit in the overall population a uniform random number and use this number in all sample selections. The coordination between samples is created based on the use of the same permanent random number of a unit over different surveys. For an overview, see, for example Ernst (1999); Mach et al. (2006), and the references therein. More recent, Nedyalkova et al.(2008), Nedyalkova et al.(2009), Grafström and Matei (2015), Grafström and Matei (2018).
- non-PRN methods: Keyfitz (1951), Kish and Scott (1971), Matei and Tillé (2005), and methods based on mathematical programming (e.g. Raj, 1968; Arthnari and Dodge, 1981; Causey et al. (1985); Ernst and Ikeda, 1995; Ernst, 1996, 1998; Ernst and Paben, 2002; Mach et al., 2006; Matei and Skinner (2009); Schiopu-Kratina et al., 2014).

Framework

- We consider the overall population of units: $U = \{1, \ldots, k, \ldots, N\} = U_1 \cup U_2$. From this population we select samples.
- $s_{12} = (s_1, s_2)$ is a bi-sample having the selection probability $p_{12} = p(s_1, s_2)$.
- The marginal sampling designs for s_1 and s_2 are given by the probabilities $p_1(s_1)$ and $p_2(s_2)$, respectively.

The overall sampling design is said to be co-ordinated if

$$p(s_1, s_2) \neq p_1(s_1)p_2(s_2)$$

(the two samples are not selected independently).

see Cotton and Hesse, 1992; Mach et al., 2006.

Degree of coordination

- $\pi_{k1} = P(k \in s_1), \pi_{k2} = P(k \in s_2), \pi_k^{1,2} = P(k \in s_1, k \in s_2), k \in U.$

The expected overlap between s_1 and s_2 is defined as

$$E(c) = \sum_{k \in U} \pi_k^{1,2}.$$

Fréchet bounds

Any bivariate distribution function H with marginal distribution functions F and G satisfies

$$\max(0, F(x) + G(y) - 1) \leq H(x, y) \leq \min(F(x), G(y)).$$

Theoretical bounds on unit level

$$\sum_{k\in U} \max(0, \pi_{k1} + \pi_{k2} - 1) \leq E(c) = \sum_{k\in U} \pi_k^{1,2} \leq \sum_{k\in U} \min(\pi_{k1}, \pi_{k2}).$$

- $\sum_{k\in U} \min(\pi_{k1}, \pi_{k2})$ is the absolute upper bound - AUB;
- $\sum_{k\in U} \max(0, \pi_{k1} + \pi_{k2} - 1)$ is the absolute lower bound - ALB.
- These bounds are usually used to quantify the performance of different sample coordination methods.
- Yet, there are few methods in the literature capable to reach these bounds.

Poisson sampling design

- It is an unequal probability sampling design with random sample size.
- A Poisson sample s with $\pi_k = P(k \in s)$, $\sum_{k \in U} \pi_k = n$ is drawn as follows:

- generate $u_1, \ldots, u_k, \ldots, u_N \sim U(0,1)$ iid,
- select k in s if

$$u_k < \pi_k = P(k \in s).$$

Poisson sampling design with PRN

- Positive coordination:

- Generate generate $u_1, \ldots, u_k, \ldots, u_N \sim U(0,1)$ iid;
- Sample s_1 : select k in s_1 if $u_k < \pi_{k1}, \ \ \pi_{k1} = P(k \in s_1)$;
- Sample s_2 : select k in s_2 if $u_k < \pi_{k2}, \ \ \pi_{k2} = P(k \in s_2)$;

- $u_1, \ldots, u_k \ldots u_N$ are called permanent random numbers (PRN) (Brewer et al., 1972).

Poisson sampling design with PRN

- Negative coordination:

- Generate generate $u_1, \ldots, u_k, \ldots, u_N \sim U(0,1)$ iid;
- Sample s_1 : select k in s_1 if $u_k < \pi_{k1}, \;\; \pi_{k1} = P(k \in s_1)$;
- Sample s_2 : select k in s_2 if $1 - u_k < \pi_{k2}, \;\; \pi_{k2} = P(k \in s_2)$;

Poisson sampling with PRN

- Good point: the bounds AUB and ALB are achieved in positive and negative coordination, respectively
  - in positive coordination we achieve the AUB:

$$E(c) = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \min(\pi_{k1}, \pi_{k2}).$$

  - in negative coordination we achieve the ALB:

$$E(c) = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \max(0, \pi_{k1} + \pi_{k2} - 1).$$

- Bad point: the sample size is random and increases the variance of the estimates.

Our goals and constraints

- Maximize/minimize the expected overlap,
- Preserve (as much as possible) the sampling design in each time occasion,
- Use of unequal probability sampling designs (with fixed sample size).
- Presentation based on Grafström and Matei (2015) and Grafström and Matei (2018).

Conditional Poisson sampling design

Grafström and Matei (2015)

- Conditional Poisson (CP) sampling is a modification of the classical Poisson sampling that produces a fixed-size sample, and has the maximum entropy property subject to given inclusion probabilities.

Entropy of a generic sampling design $\tilde{p}$

$$I(\tilde{p}) = -\sum_{s \in \mathcal{S}} \tilde{p}(s) \log(\tilde{p}(s)),$$

where $\mathcal{S} = \{s | \tilde{p}(s) > 0\}$ is the support of $\tilde{p}$.

Entropy

1. The entropy is a measure of sample randomness: higher entropy of the sampling design implies more randomness in sample selection.
2. High entropy is important for variance estimation. Tillé and Haziza (2010) noted that: 'The concept of entropy is useful in the context of variance estimation. When a sampling design has a high entropy, it is possible to obtain approximation of the second-order inclusion probabilities in terms of the first-order inclusion probabilities, which simplifies considerably the problem of variance estimation in the context of unequal probability sampling.'
3. 'Higher entropy of a design results in a faster convergence to a normal distribution of the Horvitz-Thompson estimator' (Berger, 1998).

CP-sampling

- Consider the selection of a generic CP-sample of fixed size n.
- Different implementations exist.
- Let p_k be the parameters for Poisson sampling.

Rejective implementation

Draw Poisson samples (with parameters p_k) until we get a sample of size n.

- Usually it is assumed that $\sum_{k=1}^{N} p_k = n$ because it maximizes the probability of obtaining samples of size n.
- The assumption $\sum_{k=1}^{N} p_k = n$ is, however, not restrictive.

Inclusion probabilities of CP-sampling from p_k

- When implementing CP-sampling of size n with parameters p_k, $\sum_{k=1}^{N} p_k = n$, the true inclusion probabilities will only approximately equal the p_ks.
- Let $\pi_k^{CP(n)}$ denote the achieved inclusion probabilities for CP-sampling of size n. The formula is (see Chen et al., 1994, Deville, 2000)

$$\pi_k^{CP(n)} = n \frac{\frac{p_k}{(1-p_k)} \cdot \left(1 - \pi_k^{CP(n-1)}\right)}{\sum_{\ell=1}^{N} \frac{p_\ell}{(1-p_\ell)} \cdot \left(1 - \pi_\ell^{CP(n-1)}\right)}, \tag{1}$$

  and the start is given by $\pi_k^{CP(0)} = 0,\ k = 1, 2, \ldots, N$.
- Similarly, the second-order inclusion probabilities for CP-sampling can be calculated recursively.

p_k from the inclusion probabilities π_k of CPS

- It is also possible to adjust the p_ks to obtain desired inclusion probabilities (Dupacová, 1979; Chen et al., 1994; Deville, 2000; Aires, 2000; Tillé, 2006).
- Let π_k be the inclusion probabilities for CP-sampling.
- Let $\pi_k^{CP(n,t)}$ be the inclusion probabilities derived by Equation (1) with the parameters p_k^t, where t denotes the current iteration of the algorithm, and let $p_k^0 = \pi_k$. Then, practically, only a few iterations of

$$p_k^t = p_k^{t-1} + (\pi_k - \pi_k^{CP(n,t-1)}), \tag{2}$$

is enough to find parameters p_k^t that yield inclusion probabilities π_k.

Sequential implementation of CP-sampling for a generic s

see Chen and Liu (1997); Traat et al. (2004); Tillé (2006)

- Let p_k the parameters associated to Poisson sampling design, and $\sum_{k \in U} p_k = n$, where n is the sample size.
- Let $I_k \sim Bin(1, p_k)$, $k = 1, 2, \ldots, N$ be independent random variables.
- A unit k is selected in s with an updated probability $\pi_k^{(k-1)}$: include unit k in s if $u_k \leq \pi_k^{(k-1)}$.
- The updated probabilities is calculated as

$$\pi_k^{(k-1)} = P\left(I_k = 1 \,|S_k = n - n_{k-1}\right),$$

where $S_k = \sum_{\ell=k}^{N} I_\ell$ and $n_k = \sum_{\ell=1}^{k} I_\ell$, $n_0 = 0$.

Positive coordination

1. Given π_{k1} and π_{k2}, $k = 1, 2, ..., N$, use Equation (2) to find the corresponding Poisson parameters p_{1k} and p_{2k}, respectively.
2. To coordinate two CP-samples, apply the list-sequential method with the parameters p_{1k} and p_{2k} and the permanent random numbers u_k in each selection.

Negative coordination

- For negative coordination of 2 samples, antithetic random numbers $u_k^* = 1 - u_k$ can be used in the second selection.
- For $\beta > 2$ samples, new random numbers can be constructed by shifting the PRN an amount α to the right before the selection of each sample different from the first one: $u_k + \alpha$, or $(u_k + \alpha) \mod 1$, where a possible choice of $\alpha = 1/\beta$ (see Ohlsson, 2000).

Coordination performance

Five different sampling schemes:

a) two CP-samples are drawn independently (IND) (using the rejective method for both);

b) two Poisson samples are drawn using Poisson sampling (POI) with PRN;

c) two Pareto samples are drawn using Pareto sampling (PAR) with PRN;

d) two CP-samples are drawn using the list-sequential method (SEQ) with PRN;

e) the first sample is a CP one drawn using the rejective method; the second one is selected using the rejective method with updated parameters (it is an adaptive sampling design for the first one, in order to reach AUB or ALB, but the second sampling design is not exactly CPS). We call this method the mixed one (MIX).

Pareto sampling

see Rosén, 1997a,b; Saavedra, 1995

- To all units in the population independent random numbers $u_k \sim U(0,1)$ are permanently assigned.
- On the first occasion, the n_1 units having the smallest values of $H(u_k)/H(\pi_{k1})$ are selected as a sample of size n_1.
- On the second occasion, the n_2 units having the smallest values of $H(u_k)/H(\pi_{k2})$ are selected as a sample of size n_2.
- The sample coordination is assured by the use of the same u_k in both occasions.
- The shape function is $H(x) = x/(1-x)$.

MC measures

- the Monte Carlo expected overlap

$$E_{sim}(c) = \frac{1}{m}\sum_{\ell=1}^{m} c_{\ell}^{1,2},$$

where $m = 10^5$ is the number of runs, $c_{\ell}^{1,2} = |s_{1\ell} \cap s_{2\ell}|$, and $s_{1\ell}$, $s_{2\ell}$, are the samples drawn in the ℓ^{th} run of the simulation.

- The Monte Carlo variance of the overlap

$$V_{sim}(c) = \frac{1}{m-1}\sum_{\ell=1}^{m}(c_{\ell}^{1,2} - E_{sim}(c))^2.$$

Some results - positive coordination

Table: Monte Carlo expected overlap and variance based on 10^5 simulation runs, MU284 dynamic population – stratum 2 from the MU284 population, where 50% of the units are new in the second occasion (births), and 50% of the units change the stratum (deaths), $N = 72, n_1 = 10, n_2 = 6$.

Method	$E_{sim}(c)$	$V_{sim}(c)$
IND	1.55	0.78
POI	2.79	1.94
PAR	2.76	1.04
CP-SEQ	2.55	1.00
CP-MIX	2.79	0.99
AUB	2.79	

π_{k1} are computed using the variable P75 (population in 1975 in thousands), and π_{k2} using the variable P85 (population in 1985 in thousands).

Coordination of spatially balanced sampling
Grafström and Matei (2018)

Spatial units

- Geographical position of statistical units is important e.g. in agricultural and environmental surveys because the units themselves are defined using spatial criteria;
- Also many sampling frames contains information regarding the exact or estimated geographical position of each record;
- To simplify the problem, spatial units are artificially defined over a domain partitioned into a number of predetermined regularly, or irregularly, shaped sets of spatial objects, leading to the use of the traditional sampling definition for finite populations;
- Intuitively, contiguous units can provide similar data, and more information could be obtained if the random sample avoids pairs of contiguous units. Thus, the selected units should be spread in the space (spatially balanced sampling).
- Spatially balanced designs ensure there is spatial coverage of the entire survey area.

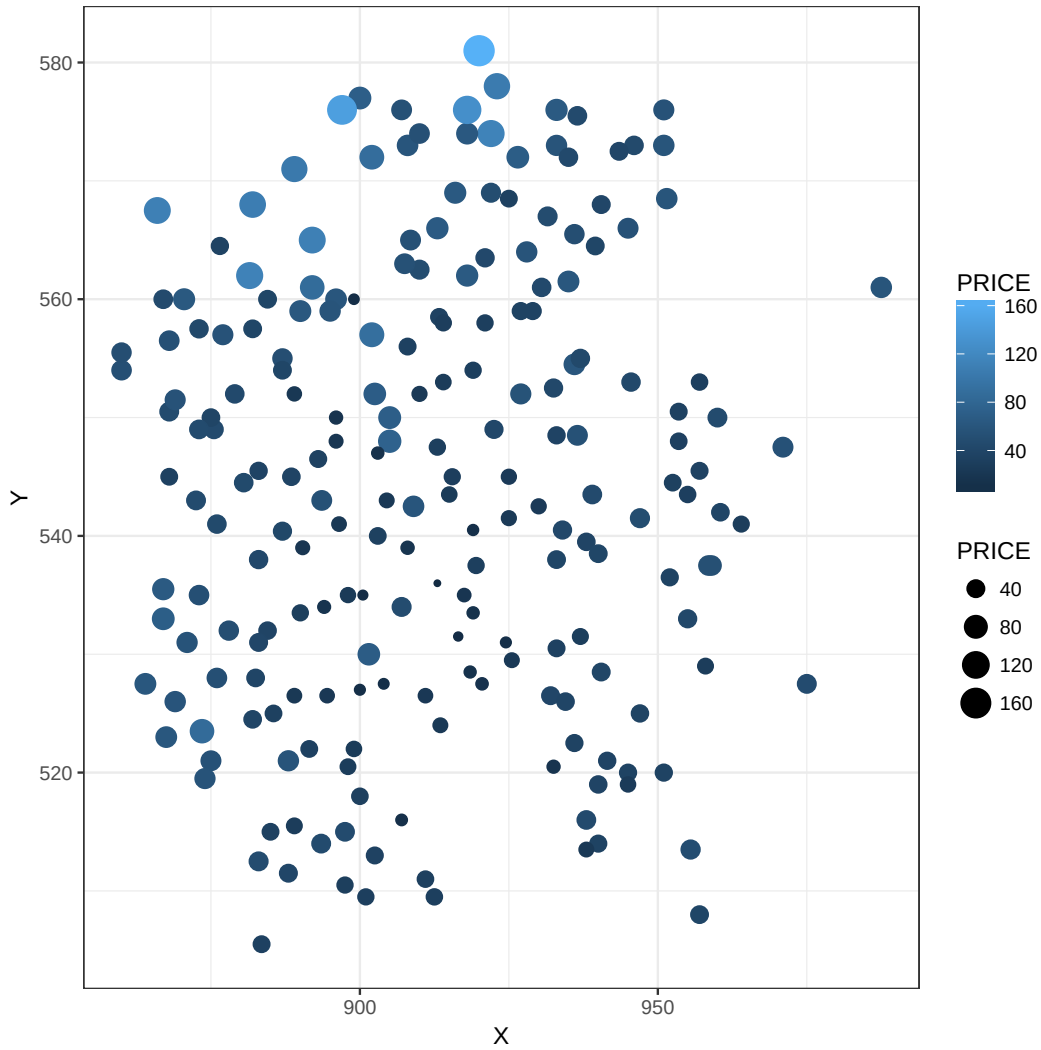


Figure: Baltimore data: house sales prices; available in R : 'spData' package (Bivand, Nowosad, Lovelace, and all., 2018)

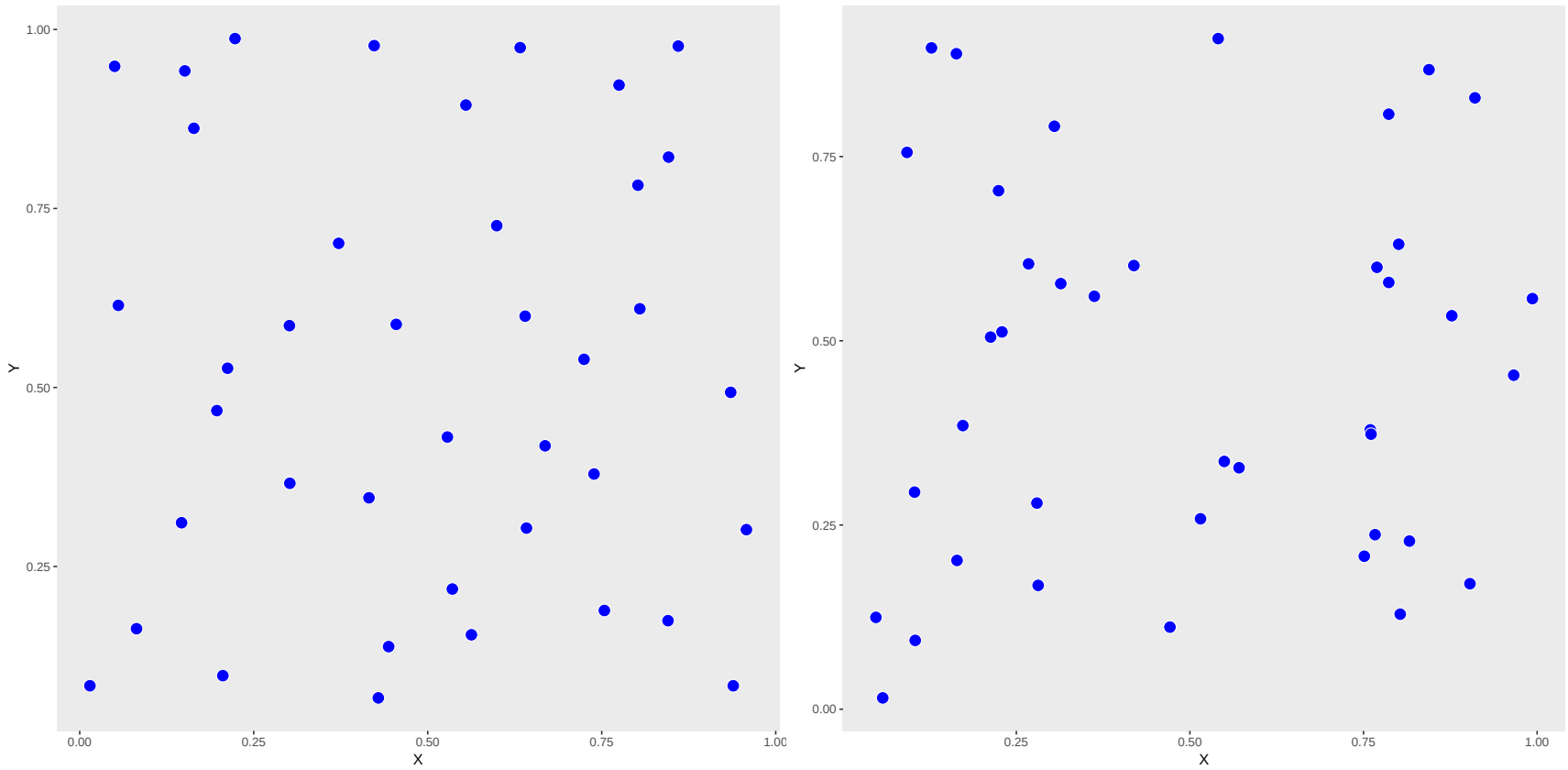


Figure: Left hand: a sample selected with a spatially balanced design; Right hand a sample selected with a non spatially balanced design.

Voronoi polytopes

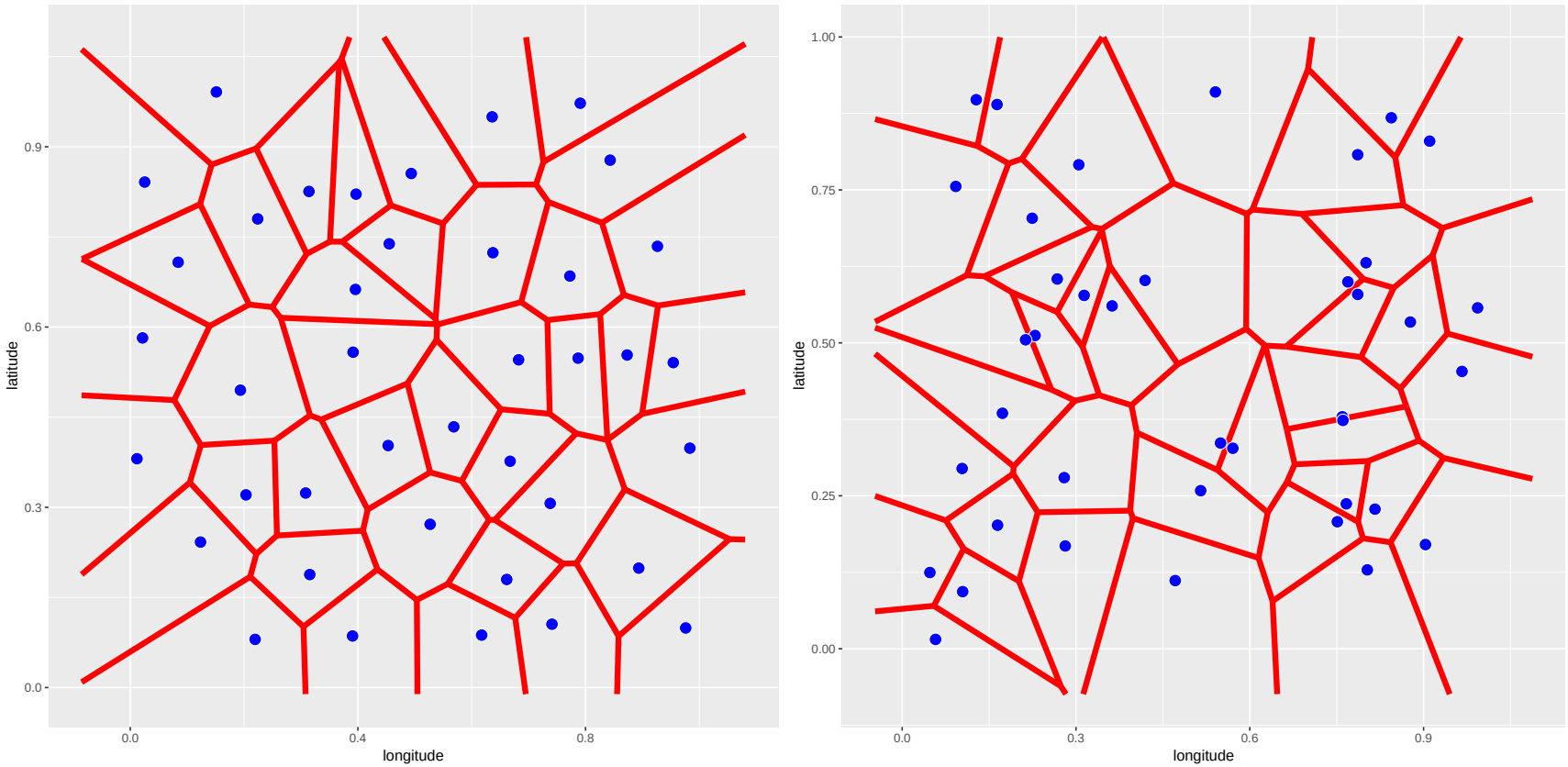


Figure: Left hand: a sample selected with a spatially balanced design; Right hand a sample selected with a non spatially balanced design.

Voronoi polytopes

- They are used to measure the level of spatial balance (or spread) with respect to the inclusion probabilities.
- A polytope P_i is constructed for each unit $i \in s$, and P_i includes all population units closer to unit i than to any other sample unit $j \in s, j \neq i$.
- Optimally, each polytope should have a probability mass that is equal to 1.
- A measure of spatial balance of a realised sample s of size n is (see Stevens and Olsen, 2004)

$$B = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2 ,$$

where v_i is the sum of the inclusion probabilities of the units in P_i.

- The expected value of B under repeated sampling is a measure of how well a design succeeds in selecting spatially balanced samples. The smaller the value of $E(B)$, the better the spread of the selected samples.

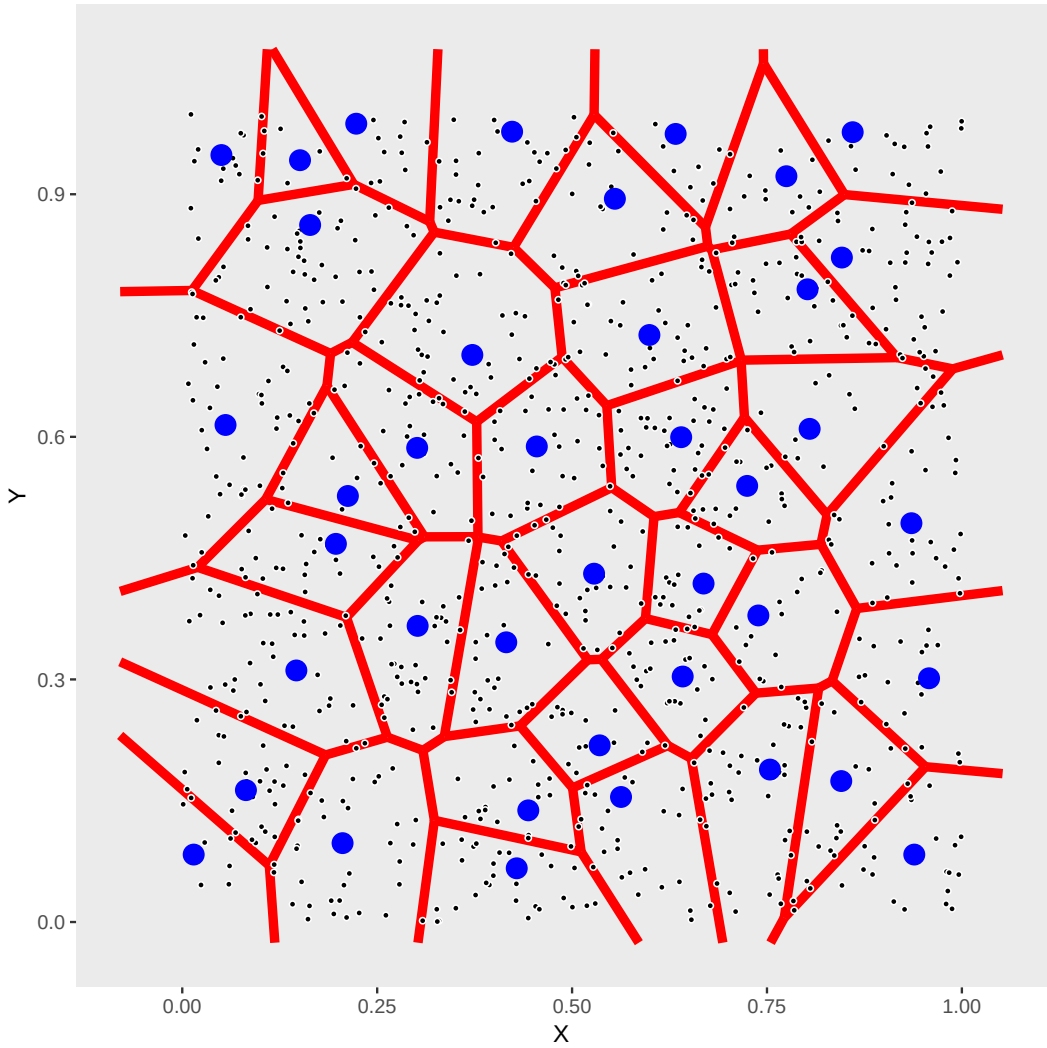


Figure: Voronoi polytopes for the sample shown in the previous left panel. The blue points are the selected ones.

Two spatial sampling schemes

1. local pivotal method (LPM) (Grafström et al., 2012), modification of the pivotal method (Deville and Tillé, 1998),
2. spatially correlated Poisson sampling (SCPS) (Grafström, 2012), modification of correlated Poisson sampling (Bondesson and Thorburn, 2008).

- Both sampling schemes are fixed-size πps designs and give spatially balanced samples.

Why is important to coordinate spatial samples?

- environmental monitoring: the sample is continuously updated, e.g. yearly, to match distributions of auxiliary variables from remote sensing (to give good estimates of current state), and where a high positive coordination would guarantee good estimates of change.
- official national business registers contain spatial coordinates of business units (e.g. US Census Bureau's Longitudinal Business Database, the Swiss GeoStat, the Italian Statistical Archive of Active Enterprisers) (Dickson et al., 2014).

Local pivotal method for a generic s

- Let $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_N)$ with $\sum_{i \in U} \pi_i = n$. The vector $\boldsymbol{\pi}$ is successively updated to become a vector with $N - n$ zeros and n ones, where the ones indicate selected units.

In each step, a pair of units (i, j) is chosen to compete. The unit i is chosen randomly, and the competitor j is the nearest neighbor to i.

- The winner takes as much probability mass as possible from the loser, so the winner ends up with $\pi_W = \min(1, \pi_i + \pi_j)$; the loser gets $\pi_L = \pi_i + \pi_j - \pi_W$.

The competition rule is

$$(\pi_i, \pi_j) := \begin{cases} (\pi_W, \pi_L) & \text{with probability } (\pi_W - \pi_j)/(\pi_W - \pi_L) \\ (\pi_L, \pi_W) & \text{with probability } (\pi_W - \pi_i)/(\pi_W - \pi_L) \end{cases}.$$

- The final outcome is decided for at least one unit at each updating, so the procedure has at most N steps. A unit that has received a probability that is 0 or 1 must not be chosen to compete again.
- The design succeeds in avoiding selection of nearby units, and hence forces the sample to be well spread (spatially balanced).

Competition rule

$$(\pi_i, \pi_j) := \begin{cases} (\pi_W, \pi_L) & \text{with probability } (\pi_W - \pi_j)/(\pi_W - \pi_L) \\ (\pi_L, \pi_W) & \text{with probability } (\pi_W - \pi_i)/(\pi_W - \pi_L) \end{cases}.$$

it means

- generate iid $u_{ij} \sim U(0,1)$,
- if $u_{ij} < (\pi_W - \pi_j)/(\pi_W - \pi_L)$ then $(\pi_i, \pi_j) := (\pi_W, \pi_L)$, else $(\pi_i, \pi_j) := (\pi_L, \pi_W)$.

First method - positive coordination

- It uses the concept of permanent random numbers and LPM.
- Samples s_1 and s_2 are drawn from U as follows:
  - a permanent random number $u_{ij} \sim U(0,1)$ is associated to each pair (i,j) of units in the overall population (and it will be used in all coordination process), all u_{ij} are iid.
  - s_1 is drawn using LPM with u_{ij} (see competition rule) and the order of pairs (i,j) is conserved,
  - s_2 is drawn using LPM with u_{ij} and the pairs (i,j) are considered in the same order as appear in the selection of s_1 (see selection). If the size of s_2 is not achieved, the method is applied as shown previously, using new pairs of units.

Remark: the negative coordination is similar using for s_2 the numbers $1 - u_{ij}$.

Correlated Poisson sampling for a generic s

- list sequential method, that is, it starts at unit 1, and once the sampling outcome is decided for that unit, it continues with unit 2 and so on. It does not revisit any unit, and once a sampling outcome has been decided, the selection probabilities for the remaining units are updated.
- Let $\pi_i, i = 1, \ldots, N$, be the prescribed inclusion probabilities and take $\pi_i^{(0)} = \pi_i, i = 1, \ldots, N$. Unit k is then selected is s with probability $\pi_k^{(k-1)}$, and we set $I_k = 1$ if it is included in s and 0 otherwise.
- The selection probabilities for the units $i = k+1, \ldots, N$ are then updated according to

$$\pi_i^{(k)} = \pi_i^{(k-1)} - (I_k - \pi_k^{(k-1)}) w_k^{(i)}.$$

- We have $E(\pi_k^{(k-1)}) = \pi_k$, for all $k \in U$.

Correlated Poisson sampling for a generic s

- In order for $0 \leq \pi_i^{(k-1)} \leq 1$, $i = k, k+1, ..., N$, to hold:

$$-\min\left(\frac{1-\pi_i^{(k-1)}}{1-\pi_k^{(k-1)}}, \frac{\pi_i^{(k-1)}}{\pi_k^{(k-1)}}\right) \leq w_k^{(i)} \leq \min\left(\frac{\pi_i^{(k-1)}}{1-\pi_k^{(k-1)}}, \frac{1-\pi_i^{(k-1)}}{\pi_k^{(k-1)}}\right)$$

- Remark: If $\sum_{k \in U} \pi_k = n$, to obtain a fixed sample size we should have for each $k \in U$ (Bondesson and Thorburn, 2008)

$$\sum_{i=k+1,...,N} w_k^{(i)} = 1.$$

Maximal weight strategy - spatial correlated Poisson sampling

- To avoid clustering of similar units and to obtain well-spread samples, the weights $w_k^{(i)}$ are chosen such that unit k gives maximal weight to the unit closest to k in distance, among the units $k+1, \ldots, N$ (Grafström, 2012).

$$\pi_i^{(k)} = \pi_i^{(k-1)} - (I_k - \pi_k^{(k-1)}) w_k^{(i)}, i = k+1, \ldots, N.$$

Selection

Unit k is then included in s with probability $\pi_k^{(k-1)}$ it means

- generate iid $u_k \sim U(0,1)$,
- if $u_k < \pi_k^{(k-1)}$ then $I_k = 1$ else $I_k = 0$.

Second method - positive coordination

- It uses the concept of permanent random numbers and SCPS.
- Samples s_1 and s_2 are drawn from U as follows:
  - a permanent random number $u_i \sim U(0,1)$ is associated to each unit $i \in U$, u_i are iid.
  - s_1 is drawn using SCPS with u_i (see selection),
  - s_2 is drawn using SCPS with u_i.

Remark: the negative coordination is similar using for s_2 the numbers $1 - u_i$.

Two new methods: mixing SCPS and Poisson sampling 1

- The new strategies are intended to provide a good compromise between the degrees of spatial balance and coordination.
- We denote the resulting family of designs Transformed Spatially Correlated Poisson Sampling (TSCPS).
- The coordination for them is similar to the coordination of SCPS.

Two new methods: mixing SCPS and Poisson sampling 2

- First strategy (TSCPS 1): modify SCPS by multiplying the maximal weight by a given scalar α, $0 \leq \alpha \leq 1$.
- Example:
  - Say the maximal weights for the three nearest neighbors of a unit k in SCPS (with maximal weights) are 0.7, 0.3, 0.
  - The new modified version would, with $\alpha = 0.5$, distribute the weights 0.35, 0.25, 0.1.
- Second strategy (TSCPS 2) is achieved by limiting the weights that a unit distributes to sum to a fixed scalar α, $0 \leq \alpha \leq 1$.
- Example:
  - Say the maximal weights for the three nearest neighbors of a unit k in SCPS (with maximal weights) are 0.7, 0.3, 0.
  - The new modified version would, with $\alpha = 0.5$, distribute 0.5, 0, 0.

Two new methods: mixing between SCPS and Poisson sampling 3

- For both TSCPS 1 and 2 with $\alpha = 0$, we get Poisson sampling and with $\alpha = 1$ we get SCPS with maximal weights.
- Maximum coordination, worst spatial balance and highest variance of sample size for $\alpha = 0$, and best spatial balance and guaranteed fixed sample size for $\alpha = 1$ while level of coordination will be to some extent worse.
- Both TSCPS 1 and 2 offer the possibility to make a trade-off between the Poisson and SCPS designs. Degree of spatial balance and coordination, as well as variance of achieved sample size depend on the parameter α.
- Drawback: for any value of $\alpha < 1$, the weights $w_k^{(i)}$ given by the unit k to units $i = k+1, \ldots, N$ do not sum up to 1 any more. Consequently, the new sampling designs do not any more provide fixed sample sizes.

Coordination performance

Different sampling schemes:

- two Poisson samples are drawn independently and with PRN, respectively;
- two LPM-samples are drawn independently;
- two samples are drawn using the LPM with PRN;
- two SCPS-samples are drawn independently;
- two samples are drawn using the SCPS with PRN;
- two samples are drawn using the TSCPS 1 with PRN and different values of α;
- two samples are drawn using the TSCPS 2 with PRN and different values of α.

MC measures

- the Monte Carlo expected overlap

$$E_{sim}(c) = \frac{1}{m}\sum_{\ell=1}^{m} c_{\ell}^{1,2},$$

where $m = 10^5$ is the number of runs, $c_{\ell}^{1,2} = |s_{1\ell} \cap s_{2\ell}|$, and $s_{1\ell}$, $s_{2\ell}$, are the samples drawn in the ℓ^{th} run of the simulation.

- The Monte Carlo variance of the overlap

$$V_{sim}(c) = \frac{1}{m-1}\sum_{\ell=1}^{m}(c_{\ell}^{1,2} - E_{sim}(c))^2.$$

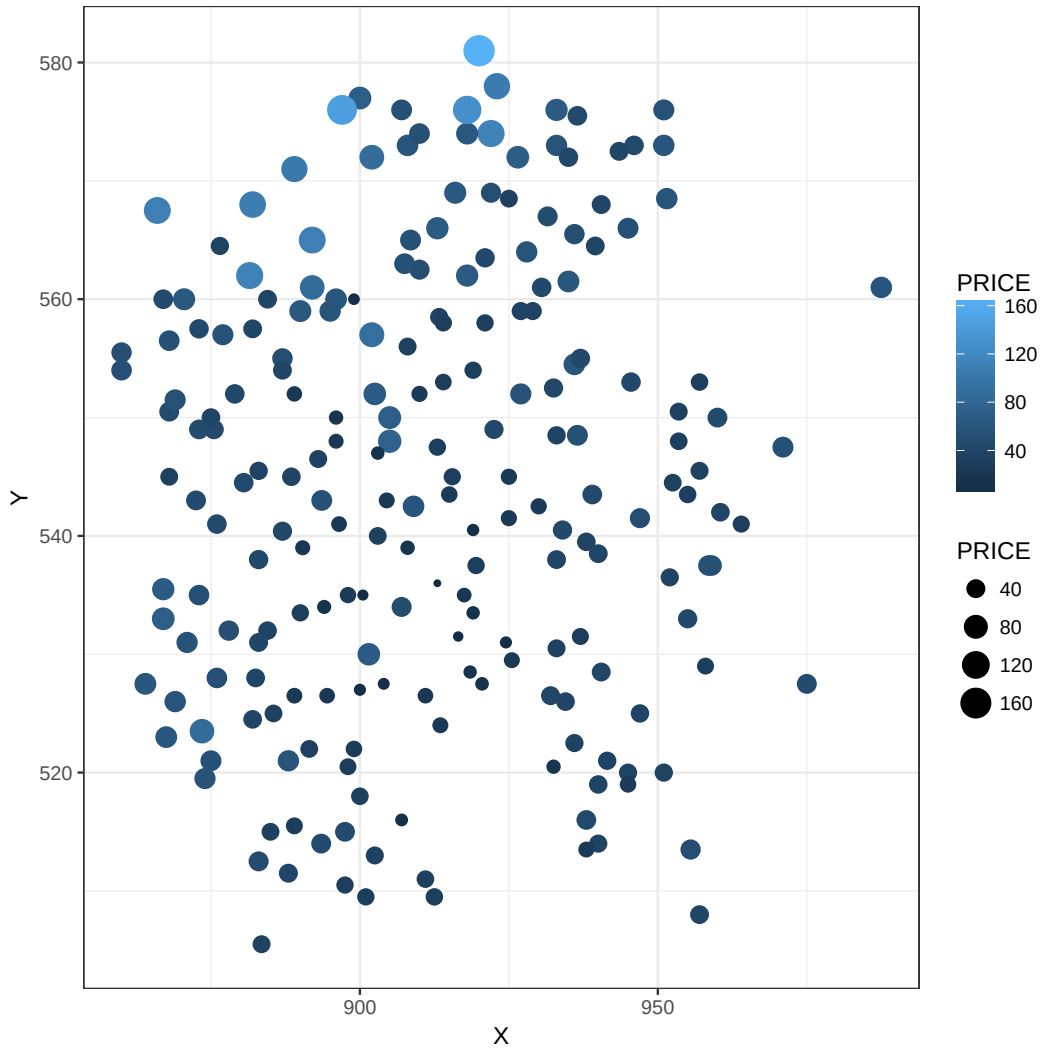


Figure: Baltimore data: house sales prices; available in R : 'spData' package (Bivand, Nowosad, Lovelace, and all., 2018)

Table: Baltimore data, $N = 211$, expected sample sizes $n_1 = 25, n_2 = 25$, π_{i1} are proportional to the variable AGE and π_{i2} to AGE+5. The distance matrix uses real data. The values of AUB and ALB are 24.20 and 0.10, respectively.

Method	independent		positive		negative	
	$E_{sim}(c)$	$V_{sim}(c)$	$E_{sim}(c)$	$V_{sim}(c)$	$E_{sim}(c)$	$V_{sim}(c)$
Poisson	4.08	3.93	24.20	20.63	0.10	0.09
LPM	4.09	3.15	21.50	2.86	1.76	1.51
SCPS	4.01	3.22	22.20	3.14	0.76	0.70
TSCPS 1						
$\alpha = 0.25$	4.05	3.02	23.10	2.60	0.26	0.26
$\alpha = 0.50$	4.06	3.06	22.50	2.93	0.08	0.43
$\alpha = 0.75$	4.05	3.22	22.30	3.10	0.08	0.55
TSCPS 2						
$\alpha = 0.25$	4.07	3.56	23.70	11.75	0.10	0.09
$\alpha = 0.50$	4.07	3.37	23.20	6.35	0.11	0.27
$\alpha = 0.75$	4.04	3.31	22.70	3.84	0.09	0.52

Application to Swiss establishments

- The data that we used was collected by the Swiss Federal Statistical Office.
- It contains census data from 2013 and 2015 on Swiss establishments.
- Data for all establishments are aggregated at the hectare level. The geographical coordinates are proper to each hectare, and not to establishments. Each hectare can contain several establishments. The statistical unit was in this application a hectare, and not an establishment.
- We considered only hectares containing establishments from the economic activity 1 (agriculture, hunting, forestry, fisheries and aquaculture), and having in total at least 3 full-time equivalent employees.
- The years 2013 (with 7057 units) and 2015 (7104 units) were considered the two time occasions. The overall population was of size $N = 9478$. The difference in the sizes between the two time occasions was due to the 2374 deaths and 2421 births in 2015 compared to 2013.

Coordination purposes

The data can be used with two main purposes:

- The location of each establishment in Switzerland has been geocoded since 1995. The register of establishments contains their geographical coordinates. Surveys are made to complete some missing information in this register. To achieve this, the Swiss Federal Statistical Office conducted such a survey in 2014. A positive coordination can be applied for example to check the quality of the the completed information from a time occasion to another one.
- Negative coordination can be applied to reduce the response burden of the establishments selected in several surveys. If the aggregated data are used, the hectares can be seen as primary selected units, while the establishments inside them as secondary selected units.

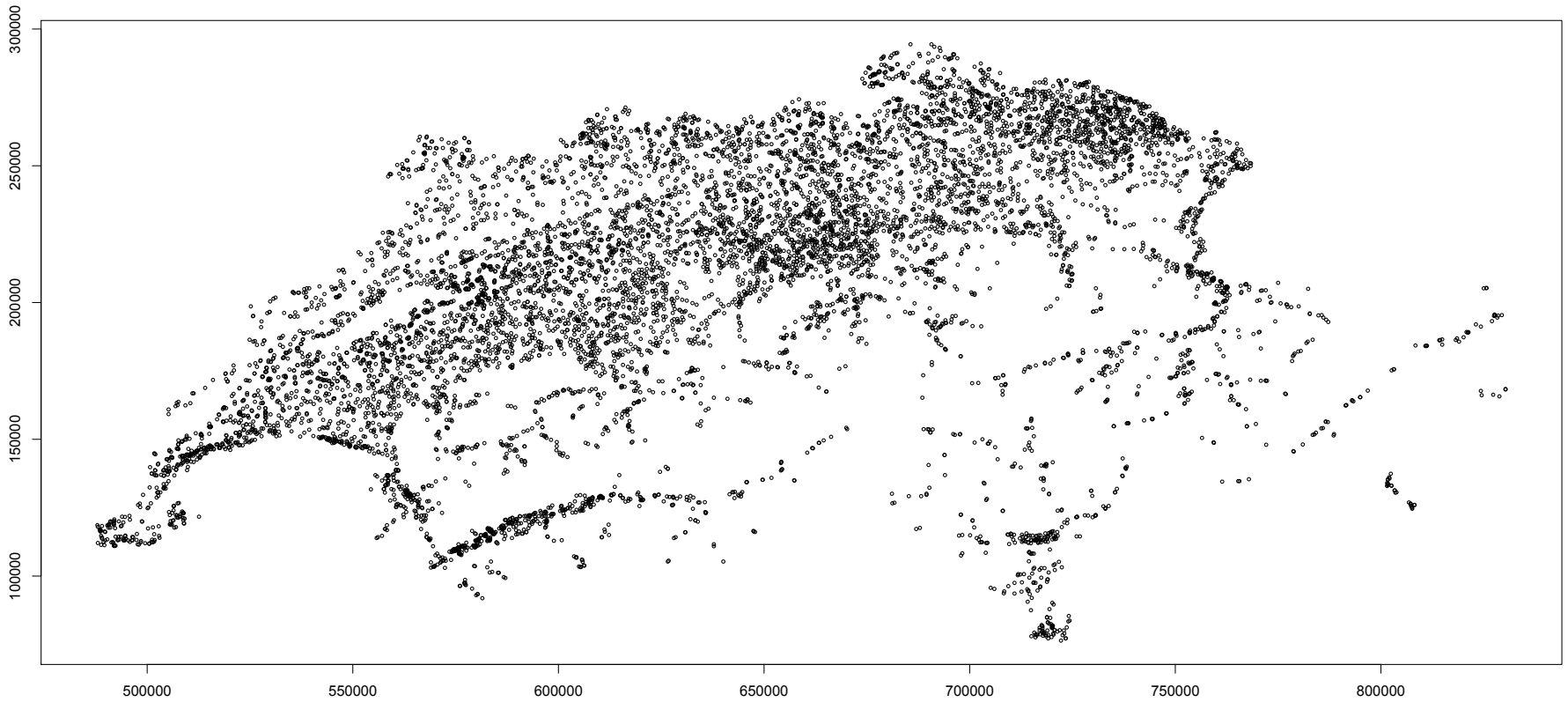


Figure: Swiss establishments aggregated data. Spatial distribution of the units in the overall population based on the census in 2013 and 2015.

Table: Swiss establishments aggregated data.
$N = 9478, n_1 = 1000, n_2 = 800, AUB = 538.022, ALB = 45.908$. Realised sample sizes, overlap between s_1 and s_2 in both types of coordination, and the B measure for s_1.

Design	size of s_1	Positive coord.		Negative coord.		B_{s_1}
		size of s_2	overlap	size of s_2	overlap	
Poisson	1010	840	560	779	46	0.387
LPM	1000	800	270	800	93	0.161
SCPS	1000	800	329	800	70	0.151
TSCPS 1						
$\alpha = 0.25$	999	799	459	800	64	0.178
$\alpha = 0.50$	1000	799	420	800	66	0.217
$\alpha = 0.75$	1000	800	366	800	67	0.178
TSCPS 2						
$\alpha = 0.25$	1012	830	469	808	49	0.275
$\alpha = 0.50$	1020	828	409	799	58	0.194
$\alpha = 0.75$	1010	816	377	797	66	0.153

Conclusions

- Methods based on permanent random numbers give partial but important solutions to real-life problems. Yet, there is no perfect method that can be applied in all circumstances.
- It is important to control the sampling design at each occasion in repeated surveys. If the goal is to maximize/minimize the expected overlap, near-optimal designs can be accepted (see MIX for CP-samples and TSCPS1 and TSCPS2 for spatial balanced samples).
- From the research point of view, it is necessary to develop coordinated sampling methods to include advances made in the domain of one-sample selection (e.g. balanced sampling).

Some references

Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35:466–483.

Brewer, K., Early, L., and Joyce, S. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 3:231–239.

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Application of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80:903–909.

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

Dickson, M. M., Benedetti, R., Giuliani, D., and Espa, G. (2014). The use of spatial sampling designs in business surveys. *Open Journal of Statistics*, 04:345–354.

Ernst, L. R. (1999). The maximization and minimization of sample overlap problems: a half century of results. In *Proceedings of the International Statistical Institute, 52nd Session*, pages 168–182, Helsinki, Finland.

Grafström, A. (2012). Spatially correlated Poisson sampling. *J. Statist. Plann. Inference*, 142:139–147.

Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

Grafström, A. and Matei, A. (2015). Coordination of Conditional Poisson samples. *Journal of Official Statistics*, 31(4):649–672.

Grafström, A. and Matei, A. (2018). Coordination of spatial balanced samples. *Survey Methodology*.

Mach, L., Reiss, P. T., and Șchiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the northwest corner rule. *Journal of the American Statistical Association*, 101(476):1671–1679.

Matei, A. and Skinner, C. (2009). Optimal sample coordination using controlled selection. *Journal of Statistical Planning and Inference*, 139(9):3112–3121.

Matei, A. and Tillé, Y. (2005). Maximal and minimal sample co-ordination. *Sankhyā*, 67, part 3:590–612.

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