Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Coordination of spatially balanced samples

by Anton Grafström and Alina Matei

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 215-238
Statistics Canada, Catalogue No. 12-001-X

Coordination of spatially balanced samples

Anton Grafström and Alina Matei[1]

Abstract

Sample coordination seeks to create a probabilistic dependence between the selection of two or more samples drawn from the same population or from overlapping populations. Positive coordination increases the expected sample overlap, while negative coordination decreases it. There are numerous applications for sample coordination with varying objectives. A spatially balanced sample is a sample that is well-spread in some space. Forcing a spread within the selected samples is a general and very efficient variance reduction technique for the Horvitz-Thompson estimator. The local pivotal method and the spatially correlated Poisson sampling are two general schemes for achieving well-spread samples. We aim to introduce coordination for these sampling methods based on the concept of permanent random numbers. The goal is to coordinate such samples while preserving spatial balance. The proposed methods are motivated by examples from forestry, environmental studies, and official statistics.

Key Words: Coordination; Local pivotal method; Spatially correlated Poisson sampling; Permanent random numbers; Unequal probability sampling designs; Transformed spatially correlated Poisson sampling.

1 Introduction

In the classical survey sampling framework, a random sample is selected from a finite population with a probability provided by the sampling design. The sampling design can be extended to the case of several samples, defining a joint probability to select them. On the other hand, two or more samples can be drawn from the same population or from overlapping populations, independently or not. Sample coordination applies to the latter case and seeks to create a probabilistic dependence between samples' selections based on a joint sampling design. It is used in the case of repeated surveys or of several surveys. Two types of coordination are defined in the literature: positive and negative. In the former case, the goal is to maximize the overlap between different samples. In the latter, one wants to minimize it. Positive coordination can be used to reduce the survey costs or to induce a positive covariance between successive estimators of state in repeated surveys, and thus reduce the variance of an estimator of change. Negative coordination may be applied to reduce the response burden of units that have a risk of being selected for several surveys.

When updating a sample in repeated surveys over time (a panel), deaths, births or merge of the units can appear in the population. Thus, the population changes over time and the same sample can not be used at each time occasion. New samples are drawn at different time occasions, but a certain degree of overlap between samples can be required. This can be achieved using positive coordination. On the other hand, negative coordination is usually used to draw samples in several surveys, involving thus different but overlapping populations. Due to births, deaths, changes in activity or size, splits, mergers, etc. of units in the same population or due to the use of different overlapping populations, an important problem in sample coordination is the difficulty to manage the population changes over time or different overlapping populations. Usually, to overcome this problem, an overall population is constructed as a union of all units that ever existed, or as a union of different overlapping populations.

1. Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183, Umea, Sweden. E-mail: anton.grafstrom@slu.se; Alina Matei, Institute of Statistics, University of Neuchâtel, Rue Bellevaux 51, 2000, Neuchâtel and Institute of Pedagogical Research and Documentation (IRDP) Neuchâtel, Switzerland. E-mail: alina.matei@unine.ch.

Various methods to provide sample coordination have been introduced in the literature. A summary of such methods is given for instance in Grafström and Matei (2015). An easy method to provide sample coordination is based on the use of so-called permanent random numbers introduced by Brewer, Early and Joyce (1972) for Poisson samples: one associates to each unit in the overall population an $U(0,1)$ random number. Such a number is called a permanent random number (PRN); these numbers are independent and are used in all sample selections. The probabilistic dependence of the samples' selection is thus created based on the use of permanent random numbers. Versions of the PRN method of Brewer et al. (1972) have been introduced in the literature (see Kröger, Särndal and Teikari, 1999; Kröger, Särndal and Teikari, 2003, for instance) and are widely used in different contexts. A recent example of a PRN method is the new system to coordinate business surveys by Statistics Canada. A two-phase stratified sampling design is used. The first-phase is a stratified sampling by Geography × Industry type × Business size and a Bernoulli sample is selected in each stratum by the use of PRNs. The main goal of the first-phase is to select a large sample covering all industries. For two consecutive first-phase waves a positive coordination is employed. In the second-phase, a sample is selected from the first-phase sample. For two consecutive second-phase waves, a negative coordination is applied to control the response burden of the business units (Haziza, 2013).

Our interest is to provide solutions to coordinate spatially balanced samples (for an overview on spatially balanced samples see Benedetti, Piersimoni and Postiglione, 2017). Usually, spatial sampling uses a space discretization, leading to the use of the classical sampling definition for finite populations. Thus, a population is defined as a finite set of units or locations having associated geographical coordinates. In most of the cases data are spatially autocorrelated and nearby locations tend to provide similar information. Consequently, it is desirable to sample units spread across the whole area of interest and to obtain a *spatially balanced sample*. The intuitive idea behind this is to cover through sampling the entire area of interest in order to obtain some representativeness. The selected sample should thus provide a full spatial coverage. Spatially balanced samples are efficient if a spatial trend is present in the variable of interest, denoted by y. Benedetti et al. (2017, page 447) note that "The motivation for the choice of selecting spatial well-spread samples is surely realistic if it is considered to be acceptable that increasing the distance between two units k and ℓ increases the difference, observed at units k and ℓ, namely, $|y_k - y_\ell|$. In this situation, it is evident that the variance of the Horvitz-Thompson estimator will necessarily decrease if we set high joint inclusion probabilities to pairs that have very different y values." Two spatial schemes useful for these goals are the local pivotal method (Grafström, Lundström and Schelin, 2012) and the spatially correlated Poisson sampling (Grafström, 2012). It was empirically found that both sampling schemes provide a good degree of spatial spreading, measured using Voronoi polytopes (see for instance Grafström et al., 2012, for some results).

We focus on coordination of spatially balanced samples using PRN methods, where sample selection is ensured using the local pivotal method (LPM) and the spatially correlated Poisson sampling (SCPS). Spatial sampling is used in many applications in environmental studies, forestry, agricultural surveys, but also in official statistics. We motivate the introduction of the coordinated spatially balanced samples by giving the following examples:

- In ecological monitoring, it is important to preserve over time the same sampled spatial locations in order to measure the changes in species abundance. However, over time, these locations may

disappear. Positive coordination can be applied in this case to ensure a significant overlap of the selected locations.

- In national forest inventories, both current state and change of several parameters, such as growing stock volume for different tree species, are of interest. The methodology we present can be used to make sure the sample is continuously updated, e.g., yearly, to be well-spread geographically or in auxiliary variables available from remote sensing (to improve estimates of current state); a high positive coordination would guarantee good estimates of change as well.
- Different 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). The business units can be selected based on their geographical coordinates (Dickson, Benedetti, Giuliani and Espa, 2014). A negative coordination would be useful in this case to control the response burden of the units sampled by different surveys.

Note that methods to coordinate spatial samples have not yet been introduced in the literature. The novelty of the paper consists in introducing methods to coordinate spatially balanced samples. All the benefits of the sample coordination described above are provided for spatially balanced samples. In both types of coordination, the proposed methods preserve the spatial balancing property of the selected samples. Note that our goal is to control the overlap size between balanced samples, and not to improve sample coordination in general.

The paper is organized as follows. Section 2 introduces the notation. Sections 3.1 and 3.2 remind the local pivotal (LP) method and spatially correlated Poisson (SCP) sampling, respectively, while Section 3.3 a measure of spatial balance based on the Voronoi polytopes. We introduce methods to coordinate LP samples and SCP samples in Section 4. The same section introduces a new family of balanced sampling designs derived from SCP sampling, that provides good results for sample coordination. The coordination performances of the methods are presented in Section 5.1. Section 5.2 compares the new family of balanced sampling designs with Poisson sampling, while Section 5.3 provides simulation results for two typical estimators in repeated surveys. Section 6 shows an application of the proposed methods on real data. Discussion of the proposed methods and conclusions are provided in Section 7.

2 Notation

Let U_1 and U_2 be a population (subject to change over time) at time 1 and time 2, respectively, or consider that U_1 and U_2 are two overlapping populations. Consider samples s_1 and s_2 drawn from U_1 and U_2, using the sampling designs p_1 and p_2, respectively. No restriction about the sampling designs p_1 and p_2 is necessary to introduce the definitions in this section: they can be fixed or random size sampling designs, with or without replacement.

Let $U = U_1 \cup U_2$. We call U the "overall population". The set of labels of the units in U is $\{1, 2, \ldots, i, \ldots, N\}$. We define on U the joint sampling design p used to select a couple (s_1, s_2). The samples s_1 and s_2 are coordinated if $p(s_1, s_2) \neq p_1(s_1)\, p_2(s_2)$, that is the samples are not drawn independently (see Cotton and Hesse, 1992; Mach, Reiss and Şchiopu-Kratina, 2006). Let $\pi_{i1} = P(i \in s_1)$

and $\pi_{i2} = P(i \in s_2)$ be the first-order inclusion probabilities of unit $i \in U$ in the first and second sample, respectively. It follows that $\pi_{i1} = 0$ if $i \notin U_1$ and $\pi_{i2} = 0$ if $i \notin U_2$. Thus, it is not necessary to identify explicitly the subpopulation memberships.

Let $\pi_{i,12} = P(i \in s_1, i \in s_2)$ be the joint inclusion probability of unit $i \in U$ in both samples s_1 and s_2. If the samples s_1 and s_2 are selected independently, $\pi_{i,12} = \pi_{i1}\pi_{i2}$, for all $i \in U$.

Let c be the overlap between s_1 and s_2, which represents the number of common units of the two samples; it is in most of the cases a random variable. The coordination degree of s_1 and s_2 is measured by the expected overlap

$$E(c) = \sum_{i \in U} \pi_{i,12},$$

where $\pi_{i,12} = P(i \in s_1, i \in s_2)$. By using the Fréchet bounds of the joint probability $\pi_{i,12}$ it follows that

$$\sum_{i \in U} \max(0, \pi_{i1} + \pi_{i2} - 1) \leq E(c) = \sum_{i \in U} \pi_{i,12} \leq \sum_{i \in U} \min(\pi_{i1}, \pi_{i2}). \quad (2.1)$$

In negative coordination one wants to achieve the left bound in expression (2.1), that is $\sum_{i \in U} \max(0, \pi_{i1} + \pi_{i2} - 1) = E(c)$, while in positive coordination the right bound, that is $E(c) = \sum_{i \in U} \min(\pi_{i1}, \pi_{i2})$. Thus, to optimize the sample coordination process, the goal is to achieve these bounds, prior to coordination type, positive or negative. Using the terminology of Matei and Tillé (2005) the left side-part in (2.1) is called the Absolute Lower Bound (ALB) and the right side-part in (2.1) the Absolute Upper Bound (AUB).

The focus here is on sample coordination using PRNs. The PRN method was originally introduced by Brewer et al. (1972) to coordinate Poisson samples. Poisson sampling with PRNs reaches the Fréchet bounds given in equation (2.1). Yet, it results in a random sample size and does not provide spatially balanced samples. In order to achieve spatial balance, the local pivotal method (Grafström et al., 2012) and the spatially correlated Poisson sampling (Grafström, 2012) are used. Both sampling designs provide a good degree of spatial balance (see Grafström et al., 2012, for some empirical results). Moreover, since both are fixed size πps sampling designs (probability proportional to size sampling, see Särndal, Swensson and Wretman, 1992, page 90), the precision of the estimators is in general improved compared to Poisson sampling.

In what follows, we consider the sampling designs p_1 and p_2 to be without replacement, and the expected sample sizes of s_1 and s_2 are denoted by n_1 and n_2, respectively.

3 Spatial balanced sampling

The two spatial sampling designs we intend to introduce coordination for are briefly recalled below for a generic sample s of fixed size n.

3.1 Local pivotal method

The local pivotal method (Grafström et al., 2012) is a spatial application of the pivotal method (Deville and Tillé, 1998). Let $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_N)$ be a given vector of inclusion probabilities, with sum n,

$\pi_i = P(i \in s), i \in U$. The vector $\boldsymbol{\pi}$ is successively updated to become a vector with $N-n$ zeros and n ones, where the ones indicate the selected units. A unit that still has a (possibly updated) probability strictly between 0 and 1 is called *undecided*. In one step of the LPM, a pair of units $i, j \in U$ is chosen to compete. More precisely, we choose unit i randomly among the undecided units, and unit i's competitor j is the nearest neighbor of i among the undecided units. Thus we apply the pivotal method locally in space. The winner receives as much probability mass as possible from the loser, so the winner ends up with $\pi_w = \min(1, \pi_i + \pi_j)$ and the loser keeps what is possibly remaining $\pi_\ell = \pi_i + \pi_j - \pi_w$. The rules of the competition are

$$(\pi_i, \pi_j) := \begin{cases} (\pi_w, \pi_\ell) & \text{with probability } (\pi_w - \pi_j)/(\pi_w - \pi_\ell) \\ (\pi_\ell, \pi_w) & \text{with probability } (\pi_w - \pi_i)/(\pi_w - \pi_\ell) \end{cases}. \tag{3.1}$$

The final outcome is decided for at least one unit each update, so the procedure has at most N steps. Because neighboring units compete against each other for inclusion, they are unlikely to be simultaneously included in a sample.

3.2 Spatially correlated Poisson sampling

The spatially correlated Poisson sampling method (Grafström, 2012) is a spatial application of the correlated Poisson sampling method (Bondesson and Thorburn, 2008). Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ be a given vector of inclusion probabilities, with sum n, $\pi_i = P(i \in s), i \in U$. The vector $\boldsymbol{\pi}$ is sequentially updated to become a vector with $N-n$ zeros and n ones, where the ones indicate the selected units. First unit 1 is included with probability $\pi_1^{(0)} = \pi_1$. If unit 1 was included, we set $I_1 = 1$ and otherwise $I_1 = 0$. Generally at step j, when the values for $I_1, \ldots, I_{j-1}$ have been recorded, unit j is included with probability $\pi_j^{(j-1)}$. Then the inclusion probabilities are updated for the units $i = j+1, \ldots, N$, according to

$$\pi_i^{(j)} = \pi_i^{(j-1)} - \left(I_j - \pi_j^{(j-1)}\right) w_j^{(i)}, \tag{3.2}$$

where $w_j^{(i)}$ are weights given by unit j to the units $i = j+1, j+2, \ldots, N$ and $\pi_i^{(0)} = \pi_i$. The weight $w_j^{(i)}$, $j < i$, determine how the inclusion probability for unit i should be affected by the sampling outcome of unit j. More precisely, the weight $w_j^{(i)}$, $j < i$, may depend on the previous sampling outcome $I_1, I_2, \ldots, I_{j-1}$ but not on the future outcomes $I_j, I_{j+1}, \ldots, I_N$. The weights should also satisfy the following restrictions

$$-\min\left(\frac{1-\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \le w_j^{(i)} \le \min\left(\frac{\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{1-\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right)$$

in order for $0 \le \pi_i^{(j-1)} \le 1$, $i = j, j+1, \ldots, N$, to hold. The unconditional inclusion probabilities are not affected by the weights since the updating rule (3.2) gives

$$E\left(\pi_i^{(i-1)}\right) = E\left(E\left(\pi_i^{(i-1)} \middle| \pi_i^{(i-2)}\right)\right) = E\left(\pi_i^{(i-2)}\right) = \ldots = \pi_i.$$

Thus the method always gives the prescribed inclusion probabilities $\pi_i, i = 1, 2, \ldots, N$.

Bondesson and Thorburn (2008) showed that a fixed size sampling is obtained only if $\sum_{i=1}^{N} \pi_i = n$ and the weights are chosen such that $\sum_{i=j+1}^{N} w_j^{(i)} = 1$, $j \in U$.

To achieve spatial balance, the weights should be decided on the basis of the distance between units. The most common approach to choose weights in SCPS is that unit j first gives as much weight as possible to the closest unit (in distance) among the units $i = j+1, j+2, \ldots, N,$ then as much weight as possible to the second closest unit etc. with the restriction that the weights are non-negative and sum up to 1. This strategy is called the *maximal weight strategy*. If distances are equal, then the weight is distributed equally on those units that have equal distance if possible. The first priority is that weight is not put on a unit if it is possible to put the weight on a closer unit. The maximal weight strategy always produces samples of fixed size if the inclusion probabilities sum up to an integer. In what follows, when we refer to SCPS, the "maximal weight strategy" is used.

3.3 Voronoi polytopes

Voronoi polytopes are used to measure the level of spatial balance (or spread) with respect to the inclusion probabilities (Stevens and Olsen, 2004). 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 is (see Stevens and Olsen, 2004)

$$B = \frac{1}{n}\sum_{i\in s}(v_i - 1)^2, \tag{3.3}$$

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 the better the spread of the selected samples.

4 Coordination methods

We present below PRN methods based on the local pivotal (LP) method and the spatially correlated Poisson (SCP) sampling.

4.1 Coordination of LP samples

The positive coordination of LP samples with PRNs is implemented as follows:

1. independent permanent random numbers $v_{ij} \sim U(0,1)$ are associated to each pair $(i,j) \subseteq U \times U$;
2. s_1 is drawn using LPM as follows: if a pair of units (i,j) is chosen to compete, the number v_{ij} is used in the corresponding competition rule (3.1) and the pair (i,j) is saved into a list of pairs;
3. s_2 is drawn using LPM as follows: the pairs (i,j) are considered sequentially from the list of pairs constructed above for s_1, and the same numbers v_{ij} are used in the corresponding competition rule (3.1). If the sample size n_2 is achieved using pairs from this list, the algorithm stops; if not, the selection process continues with new pairs (i,j) (not included in this list) and selected as described in Section 3.1.

For negative coordination, the first two steps are the same, but the last step becomes:

3'. s_2 is drawn using LPM as follows: the pairs (i, j) are considered sequentially from the list of pairs constructed above for s_1, and the numbers $1 - v_{ij}$ are used in the corresponding competition rule (3.1). If the sample size n_2 is achieved using pairs from this list, the algorithm stops; if not, the selection process continues with new pairs (i, j) (not included in this list) and selected as described in Section 3.1.

4.2 Coordination of SCP samples

The coordination of SCP samples with PRNs is implemented as follows. Let u_i be the PRN associated to unit $i \in U$, with $u_1, u_2, \ldots, u_N$ iid $U(0, 1)$. Let $\pi_{it}^{(i-1)}$ be the (updated) selection probability for unit i in the selection of sample s_t, $t = 1, 2$. For positive coordination, the PRNs are introduced in the selection step similarly to Poisson sampling with PRNs: if $u_i < \pi_{it}^{(i-1)}$, unit i is selected in the sample s_t, $t = 1, 2$. For negative coordination, if $u_i < \pi_{i1}^{(i-1)}$, unit i is selected in s_1; if $1 - u_i < \pi_{i2}^{(i-1)}$, unit i is selected in s_2. This coordination method is general for spatially correlated Poisson sampling and can be used no matter what weights are applied within the method.

We utilize the maximal weight strategy advocated in Section 3.2 as the main alternative, but we also introduce two new alternative strategies to compute the weights $w_j^{(i)}$. The new strategies are intended to provide a good compromise between the degrees of spatial balance and coordination. By reducing the amount of spatial correlation in SCPS we can achieve any level of mixing between SCPS and Poisson sampling. Both of the new strategies are similar to the SCPS with maximal weights, but the weights $w_j^{(i)}$ given by the unit j to units $i = j + 1, \ldots, N$ do not sum up to 1 any more. Consequently, the result of Bondesson and Thorburn (2008) advocated in Section 3.2 does not apply and the new sampling designs do not any more provide fixed sample sizes. We denote the resulting family of designs Transformed Spatially Correlated Poisson Sampling (TSCPS).

The first mixing strategy is to modify SCPS by multiplying the maximal weight by a given scalar α, $0 \leq \alpha \leq 1$. Thus we no longer use maximal weight, but the proportion α of the maximal weight is the limit for the applied weight. This method is denoted TSCPS 1. With this method the positive weights will reach longer (more neighbors) than in SCPS. Each unit would distribute a total weight of maximum 1, starting with the nearest unit and then the second nearest etc. Say the maximal weights for the three nearest neighbors of a unit are 0.7, 0.5, 0.2. Then, in standard SCPS (with maximal weights) the unit would distribute the weights 0.7, 0.3, 0, and the new modified version would, with $\alpha = 0.5$, distribute the weights 0.35, 0.25, 0.1. The reach is longer but it is not guaranteed we can use all α. As a result, the total weight is not necessary 1, and the sample size becomes random.

The second mixing strategy is achieved by limiting the weights that a unit distributes to sum to a fixed scalar α, $0 \leq \alpha \leq 1$. This method is denoted TSCPS 2. In SCPS with maximal weight strategy, each unit is given a total weight 1 (the sum of the weights) to distribute on remaining units in the list. Instead, each unit is given a total weight α to distribute. Otherwise, this works as the maximal weight strategy, so that unit i first gives as much weight as possible to the nearest, then the second nearest etc. With this strategy the weights will reach a shorter distance (fewer neighbors). Say the maximal weights for the three nearest

neighbors of a unit are 0.7, 0.5, 0.2. Then standard SCPS (with maximal weights) would distribute the weights 0.7, 0.3, 0, and the new modified version would, with $\alpha = 0.5$, distribute 0.5, 0, 0. The reach is shorter and it is guaranteed we can use all α. However, if the total weight α is less than 1, there will be a random sample size.

Note that for both TSCPS 1 and 2 we have the following result. With $\alpha = 0$, we get Poisson sampling and with $\alpha = 1$ we get SCPS with maximal weights. We can scale with α between 0 and 1 to mix the two to any degree. 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. Sample size is likely to be more stable (given the same α) for TSCPS 1 than for TSCPS 2, as more weight is likely to be distributed with TSCPS 1. Since both TSCPS 1 and 2 use a given scalar α, $0 \leq \alpha \leq 1$, they provide a family of sampling designs. Each element in this family corresponds to a given α. Contrary to SCPS, for any value of $\alpha < 1$ both TSCPS 1 and 2 involve random sample sizes. The consequences of having random sample sizes on coordination is empirically studied in Section 5.1, on spatial balance degree in Section 5.2 and on variance estimation in Section 5.3.

5 Empirical results

5.1 Overlap performance

Monte Carlo simulation was used to study the overlap performance of the proposed methods. A number of $m = 10^4$ runs were considered for each of the four settings described below. In each run, samples were drawn using the proposed methods. The same permanent random numbers were employed for all methods. The Euclidean distance between units was used for all spatial sampling designs. In each run, for LPM with PRNs, a matrix of dimension $N \times N$ of PRNs was randomly generated; the diagonal elements of this matrix were used as PRNs for Poisson, SCPS and the transformed SCPS with PRNs. All sampling schemes were applied for positive and negative coordination, respectively, using in each run the same PRNs and the same matrix of distances. Samples s_1 and s_2 of following types were drawn in each run:

- two Poisson samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two LP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two SCP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two transformed SCP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs; the two strategies shown in Section 4.2 were employed using respectively $\alpha = 0.25$, 0.50 and 0.75.

Three measures were used to quantify the performance of the proposed methods, for positive and negative coordination, respectively:

- the Monte Carlo expected overlap

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

  $c_{\ell}^{1,2} = |s_{1\ell} \cap s_{2\ell}|$, and $s_{1\ell}$, $s_{2\ell}$, are the samples drawn in the ℓ^{th} run, where $|s_{1\ell} \cap s_{2\ell}|$ represents the number of common units of $s_{1\ell}$ and $s_{2\ell}$;
- the Monte Carlo variance of the overlap

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

- the Monte Carlo coefficient of variation of the overlap

$$\text{CV}_{\text{sim}}(c) = \frac{\sqrt{V_{\text{sim}}(c)}}{E_{\text{sim}}(c)}.$$

The correlation between $\boldsymbol{\pi}_1 = (\pi_{1i})_{i=1,\ldots,N}$ and $\boldsymbol{\pi}_2 = (\pi_{2i})_{i=1,\ldots,N}$ is an important factor of the sample coordination degree. This correlation varies and takes extreme values in the following four settings used to study the performance of the proposed methods:

- the static MU284 population: from the MU284 data set (see Appendix B in Särndal et al., 1992), the region 2 was selected. The population size is $N = 48$, and the expected sample sizes are $n_1 = 10$, $n_2 = 6$, respectively. The first-order inclusion probabilities π_{i1} are computed using the variable P75 (population in 1975 in thousands), and π_{i2} using the variable P85 (population in 1985 in thousands). The elements of the distance matrix were artificially generated using independent draws from the $N(0,1)$ distribution and taking their absolute values. The correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is 0.99.
- the Baltimore data set is about house sales prices and hedonics (see Dubin, 1992). The data set is available on-line at the GeoDa Center for Geospatial Analysis and Computation (2017). Information on $N = 211$ houses are provided by 17 variables. The geographical coordinates of the houses are available. We use $n_1 = n_2 = 25$. The first-order inclusion probabilities π_{i1} are computed using the variable AGE (the house age) and π_{i2} using AGE+5. The elements of the distance matrix are the Euclidean distances between the geographical coordinates on the Maryland grid of the houses included in this data set. The correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is 1.
- the MU284 dynamic population: from the MU284 data set, the regions 2 and 3 were used. A dynamic population was created using on the first occasion 50% of the units randomly selected from the region 2 using simple random sampling without replacement (these units are the "persistents" and the rest of the 50% of the units are "deaths"), and on the second occasion 50% of the units randomly selected from the region 3 using simple random sampling without replacement (these units are the "births"). The elements of the distance matrix were artificially

generated using independent draws from the $N(0,1)$ distribution and taking their absolute values. For a run, the correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ was 0.08.

- one artificial data set, with $N = 100$, $n_1 = 10$, $n_2 = 25$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ uncorrelated and randomly generated using independent draws from the $U(0,1)$ distribution and scaled to obtain the sum 10 and 25, respectively. The elements of the distance matrix were artificially generated using independent draws from the $N(0,1)$ distribution and taking their absolute values.

A number of 10^4 simulation runs was used to compute the Monte Carlo overlap measures using the nine methods in each setting. Tables 5.1, 5.2, 5.3, and 5.4 provide the results of the Monte Carlo studies based on the previous four settings. For TSCPS 1 and 2, the value of α is also specified in these tables.

Table 5.1
The static MU284 population, $N = 48$, expected sample sizes $n_1 = 10, n_2 = 6$, π_{i1} are computed using the variable P75 (population in 1975 in thousands), and π_{i2} using the variable P85 (population in 1985 in thousands). The distance matrix was artificially generated. The values of AUB and ALB are 6 and 1.96, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		3.04	1.89	0.45	6.03	4.06	0.33	1.96	1.13	0.54
LPM		3.03	1.22	0.36	5.10	0.71	0.17	2.64	1.20	0.41
SCPS		3.06	1.21	0.36	4.91	0.85	0.19	2.33	1.06	0.44
TSCPS 1	$\alpha = 0.25$	3.06	1.28	0.37	5.84	0.93	0.17	2.09	1.13	0.51
	$\alpha = 0.50$	3.04	1.27	0.37	5.54	0.79	0.16	2.21	1.10	0.47
	$\alpha = 0.75$	3.06	1.25	0.37	5.20	0.80	0.17	2.27	1.06	0.45
TSCPS 2	$\alpha = 0.25$	3.07	1.67	0.42	5.75	2.40	0.27	1.97	1.13	0.54
	$\alpha = 0.50$	3.06	1.45	0.39	5.40	1.57	0.23	2.05	1.10	0.51
	$\alpha = 0.75$	3.04	1.27	0.37	5.13	1.10	0.20	2.18	1.04	0.47

Table 5.2
Baltimore data, $N = 211$, expected sample sizes $n_1 = 25, n_2 = 25$, π_{i1} are computed using the variable AGE and π_{i2} using 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		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		4.08	3.93	0.49	24.20	20.63	0.19	0.10	0.09	3.00
LPM		4.09	3.15	0.43	21.50	2.86	0.08	1.76	1.51	0.70
SCPS		4.01	3.22	0.45	22.20	3.14	0.08	0.76	0.70	1.10
TSCPS 1	$\alpha = 0.25$	4.05	3.02	0.43	23.10	2.60	0.07	0.26	0.26	1.96
	$\alpha = 0.50$	4.06	3.06	0.43	22.50	2.93	0.08	0.45	0.43	1.46
	$\alpha = 0.75$	4.05	3.22	0.44	22.30	3.10	0.08	0.57	0.55	1.30
TSCPS 2	$\alpha = 0.25$	4.07	3.56	0.46	23.70	11.75	0.14	0.10	0.09	3.00
	$\alpha = 0.50$	4.07	3.37	0.45	23.20	6.35	0.11	0.29	0.27	1.79
	$\alpha = 0.75$	4.04	3.31	0.45	22.70	3.84	0.09	0.58	0.52	1.24

Table 5.3
The dynamic MU284 population – region 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"), $N = 72$, expected sample sizes $n_1 = 10, n_2 = 6$. The distance matrix was artificially generated. The values of AUB and ALB are 3.56 and 1.33, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		2.02	1.20	0.54	3.56	2.35	0.43	1.32	0.71	0.64
LPM		2.03	0.95	0.48	2.37	1.00	0.42	1.87	0.89	0.50
SCPS		2.02	1.02	0.50	3.01	1.19	0.36	1.54	0.79	0.58
TSCPS 1	$\alpha = 0.25$	2.02	0.94	0.48	3.42	1.31	0.33	1.39	0.70	0.60
	$\alpha = 0.50$	2.03	1.02	0.50	3.27	1.33	0.35	1.42	0.79	0.63
	$\alpha = 0.75$	2.02	1.02	0.50	3.16	1.26	0.36	1.47	0.80	0.61
TSCPS 2	$\alpha = 0.25$	2.02	1.04	0.50	3.36	1.67	0.38	1.33	0.64	0.60
	$\alpha = 0.50$	2.02	0.96	0.49	3.20	1.37	0.37	1.41	0.66	0.58
	$\alpha = 0.75$	2.02	0.94	0.48	3.10	1.24	0.36	1.50	0.71	0.56

Table 5.4
Artificial data, $N = 100$, expected sample sizes $n_1 = 10, n_2 = 25, \pi_{i1}$ and π_{i2} randomly generated, uncorrelated. The distance matrix was artificially generated. The values of AUB and ALB are 9.11 and 0, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		2.44	2.34	0.63	9.11	8.08	0.31	~ 0	~ 0	
LPM		2.45	1.82	0.55	5.42	2.35	0.28	1.03	0.91	0.93
SCPS		2.42	1.82	0.56	6.94	2.07	0.21	0.45	0.42	1.44
TSCPS 1	$\alpha = 0.25$	2.44	1.76	0.54	8.53	2.05	0.17	0.06	0.07	4.41
	$\alpha = 0.50$	2.46	1.79	0.54	7.95	1.90	0.17	0.21	0.22	2.23
	$\alpha = 0.75$	2.43	1.80	0.55	7.40	1.97	0.19	0.31	0.31	1.80
TSCPS 2	$\alpha = 0.25$	2.43	2.09	0.59	8.53	4.86	0.26	~ 0	~ 0	
	$\alpha = 0.50$	2.45	1.91	0.56	7.90	3.32	0.23	0.11	0.10	2.87
	$\alpha = 0.75$	2.44	1.83	0.55	7.34	2.51	0.22	0.28	0.26	1.82

Following the results given in Tables 5.1, 5.2, 5.3, and 5.4, SCPS shows in general better performance than LPM in terms of $E_{\text{sim}}(c), V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$ for both types of coordination; an exception is the case of the static MU284 population and positive coordination. In this setting, the pairs used for the selection of s_1 are also used for the selection of s_2, since deaths or births are not assumed. Without such changes in population, LPM may perform better than SCPS in terms of $E_{\text{sim}}(c)$, but also in terms of $V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$.

As expected, Poisson sampling achieves the AUB and ALB (minor differences are due to the sampling error) in all settings, but the overlap variance is very high in positive coordination. This is mainly due to the random sizes of s_1 and s_2. The large values of $V_{\text{sim}}(c)$ impact the values of $\text{CV}_{\text{sim}}(c)$. In all the examples shown, the latter is in general larger than the values of $\text{CV}_{\text{sim}}(c)$ provided by the other sampling schemes.

Results in Tables 5.1, 5.2, 5.3, and 5.4 confirm that the value of α in the transformed SCPS determines the coordination degree; a smaller value of α provides a better coordination degree, since one gets closer to Poisson sampling (we remind that $\alpha = 0$ in the TSCPS designs leads to Poisson sampling).

For a given α, the new strategies presented in Section 4.2 yield similar values of $E_{\text{sim}}(c)$ in positive coordination, but TSCPS 2 gives larger values of $V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$. For all α used, both TSCPS 1 and TSCPS 2 provides similar values of $\text{CV}_{\text{sim}}(c)$ in positive and negative coordination in our examples, excepting TSCPS 2 with $\alpha = 0.25$. The latter performs very close to Poisson sampling in negative coordination as the results in Tables 5.1, 5.2, 5.3, and 5.4 show.

An interesting result for Poisson sampling arises from Tables 5.1, 5.2, 5.3, and 5.4 in terms of $\text{CV}_{\text{sim}}(c)$. While the values of $V_{\text{sim}}(c)$ are large for positive coordination compared to LPM and SCPS, it is not the case for negative coordination. However, in the latter case, if $E_{\text{sim}}(c) \sim V_{\text{sim}}(c)$ and both are small as in Table 5.2, the corresponding value of $\text{CV}_{\text{sim}}(c)$ becomes very large. As we mentioned, that can also be the case for the TSCPS designs with small values of α. The improvement of introducing this new family of designs compared to Poisson sampling is measured for these situations in terms of spatial balance degree as shown in the next section.

5.2 Spatial balance and variance of sample size

The transformed SCPS is compared to the other sampling designs in terms of degree of spatial balance using Monte-Carlo simulation. The degree of spatial balance is measured using the B measure shown in expression (3.3). For the transformed SCPS the two strategies presented in Section 4.2 are used, and the four previous settings are employed. The B measure was computed on the same samples s_1 used to obtain the outcomes given in Tables 5.1, 5.2, 5.3, and 5.4, respectively. The following overall measure was used for each type of sample

$$E_{\text{sim}}(B) = \frac{1}{m}\sum_{\ell=1}^{m} B_\ell,$$

where B_ℓ represents the B measure computed on a realised sample in the ℓ^{th} run. For comparison, the average of the B measures computed over the Monte-Carlo runs for Poisson sampling and LPM were also reported.

TSCPS is also compared with Poisson sampling in terms of variance of sample size computed over the Monte-Carlo runs using:

$$V_{\text{sim}}(\text{size}) = \frac{1}{m-1}\sum_{\ell=1}^{m}\left(|s_\ell| - |\overline{s}|\right)^2,$$

where $|s_\ell|$ represents the sample size of a realised sample s_ℓ in the ℓ^{th} run and $|\overline{s}| = \frac{1}{m}\sum_{\ell=1}^{m}|s_\ell|$.

Tables 5.5, 5.6, 5.7 and 5.8 provide the results. Following these results, we note that the choice of α determines the performance of the transformed SCPS in terms of averaged B measure: a larger value of α results in a better spatial balance degree. However, in all settings, the resulting spatial balance degree is

worse than for LPM and SCPS, but better than for Poisson sampling as expected, since the latter is not a spatial balanced sampling.

For all four settings, the variance of sample size is much higher for Poisson sampling than for TSCPS 1 and TSCPS 2, for all values of α. While TSCPS 2 with $\alpha = 0.25$ performs very close to Poisson sampling in the examples shown in Section 5.1 for negative coordination, we note however that the corresponding values of $V_{\text{sim}}(\text{size})$ for the former method are much smaller than those provided by Poisson sampling.

As underlined in Section 4.2, TSCPS 1 shows smaller sample size variance than TSCPS 2 for the same α. The results in our settings confirm for both TSCPS 1 and TSCPS 2 that the variance of sample size decreases when α increases.

Table 5.5
The static MU284 population, $N = 48$, expected sample size 10, the inclusion prob. are computed using the variable P75 (population in 1975 in thousands). The distance matrix was artificially generated

Design		$E_{\text{sim}}(B)$	$V_{\text{sim}}(\text{size})$
Poisson		0.301	4.806
LPM		0.124	0
SCPS		0.131	0
TSCPS 1	$\alpha = 0.25$	0.209	0.727
	$\alpha = 0.50$	0.177	0.405
	$\alpha = 0.75$	0.146	0.187
TSCPS 2	$\alpha = 0.25$	0.215	2.692
	$\alpha = 0.50$	0.159	1.211
	$\alpha = 0.75$	0.134	0.399

Table 5.6
Baltimore data, $N = 211$, expected sample size 25, the inclusion prob. are computed using the variable AGE. The distance matrix uses real data

Design		$E_{\text{sim}}(B)$	$V_{\text{sim}}(\text{size})$
Poisson		0.416	21.107
LPM		0.137	0
SCPS		0.137	0
TSCPS 1	$\alpha = 0.25$	0.256	0.909
	$\alpha = 0.50$	0.198	0.449
	$\alpha = 0.75$	0.162	0.222
TSCPS 2	$\alpha = 0.25$	0.282	11.382
	$\alpha = 0.50$	0.195	4.811
	$\alpha = 0.75$	0.148	1.227

Table 5.7
The dynamic MU284 population, $N = 48$, expected sample size 10, the inclusion prob. are computed using the variable P75 (population in 1975 in thousands). The distance matrix was artificially generated

Design		$E_{sim}(B)$	$V_{sim}(\text{size})$
Poisson		0.422	5.683
LPM		0.202	0
SCPS		0.210	0
TSCPS 1	$\alpha = 0.25$	0.306	0.798
	$\alpha = 0.50$	0.255	0.427
	$\alpha = 0.75$	0.224	0.231
TSCPS 2	$\alpha = 0.25$	0.315	3.128
	$\alpha = 0.50$	0.252	1.370
	$\alpha = 0.75$	0.213	0.446

Table 5.8
Artificial data, $N = 100$, expected sample size 10, the inclusion prob. are randomly generated. The distance matrix was artificially generated

Design		$E_{sim}(B)$	$V_{sim}(\text{size})$
Poisson		0.485	8.892
LPM		0.134	0
SCPS		0.133	0
TSCPS 1	$\alpha = 0.25$	0.286	0.938
	$\alpha = 0.50$	0.213	0.446
	$\alpha = 0.75$	0.167	0.230
TSCPS 2	$\alpha = 0.25$	0.313	4.854
	$\alpha = 0.50$	0.204	2.121
	$\alpha = 0.75$	0.149	0.632

5.3 Variance estimation

In repeated surveys, estimates of net variation, period averages and gross change are of interest. Our proposed methods are suitable to estimate such parameters. Their variance estimation is, however, intractable for our methods and is not addressed here. We study only empirically the impact that each coordinated spatial balancing method has on the quality of the estimates of two of the above parameters. Note that there exist approximative variance estimators for state that can be used for LPM and SCPS (Grafström and Schelin, 2014), but further research is needed to derive an approximative estimator for the covariance between successive state estimators under coordination.

Consider a repeated survey over two time occasions. Let y be the variable of interest, measured in the first and second time occasion, respectively. We denote by y_{it} the value of this variable taken by the unit $i \in U$ on the time occasion t, with $t \in \{1, 2\}$. Let x_{it} be the value of an auxiliary variable taken by the unit $i \in U$ at occasion t; the variable x is well correlated with y, and available for all units $i \in U$ in both time occasions. It is assumed that x_{it} is known for all $i \in U$ from a previous census or that a two-phase sampling is used: in the first phase the value of x_{it} is obtained, while the coordination process is addressed in the

second phase of the sampling. The notation $E_M(.)$ and $\text{var}_M(.)$ indicate the expectation and variance under a model. We borrow from Grafström and Tillé (2013) the following cross-sectional superpopulation model with spatial correlation

$$y_{i,t-1} = \beta_0 + x_{i,t-1}\beta_1 + \varepsilon_{i,t-1}, \tag{5.1}$$

where β_0 and β_1 are parameters, where $\varepsilon_{i,t-1}$ are random variables, with $E_M(\varepsilon_{i,t-1}) = 0$, $\text{var}_M(\varepsilon_{i,t-1}) = \sigma_i^2$, $\text{cov}_M(\varepsilon_i, \varepsilon_j) = \sigma_i\sigma_j\rho^{d(i,j)}$, where $d(i,j)$ represents the distance between the units i and j, for $i, j \in U$. The particular form of $\text{cov}_M(\varepsilon_i, \varepsilon_j)$ in model (5.1) underlines a decreasing function of the distance between i and j, reflecting that the proximity of units implies a larger spatial correlation. The following autoregressive model is considered

$$y_{it} = \delta_0 + \delta_1 y_{i,t-1} + \gamma_{it}, \tag{5.2}$$

with δ_0 and δ_1 being parameters, and with γ_{it} being independent random variables, with $E_M(\gamma_{it}) = 0$, $\text{var}_M(\gamma_{it}) = u^2$. The following model is also assumed

$$x_{it} = \alpha_0 + \alpha_1 x_{i,t-1} + \tilde{\gamma}_{it}, \tag{5.3}$$

where α_0 and α_1 are parameters, where $\tilde{\gamma}_{it}$ are independent random variables, with $E_M(\tilde{\gamma}_{it}) = 0$, $\text{var}_M(\tilde{\gamma}_{it}) = \tilde{u}^2$. We obtain thus a spatial-temporal dependence of the data through models (5.1), (5.2) and (5.3).

We consider that π_{it} are constructed using the expression

$$\pi_{it} = \frac{n_t x_{it}}{\sum_{j\in U} x_{jt}},\ t \in \{1, 2\},$$

that leads to a correlation between $\boldsymbol{\pi}_{t-1}$ and $\boldsymbol{\pi}_t$ due to model (5.3).

The following parameters of interest are considered: the one period change $D = \sum_{i\in U_1} y_{i_1} - \sum_{i\in U_2} y_{i2}$ and the average over two periods $A = \frac{1}{2}\left(\sum_{i\in U_1} y_{i_1} + \sum_{i\in U_2} y_{i2}\right)$. The two parameters are estimated by

$$\hat{D} = \sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}} - \sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}},$$

and

$$\hat{A} = \frac{1}{2}\left(\sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}} + \sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}}\right),$$

respectively. We have

$$\text{var}\left(\hat{D}\right) = \text{var}\left(\sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}}\right) + \text{var}\left(\sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}}\right) - 2\text{cov}\left(\sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}}, \sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}}\right), \tag{5.4}$$

$$\text{var}\left(\hat{A}\right) = \frac{1}{4}\text{var}\left(\sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}}\right) + \frac{1}{4}\text{var}\left(\sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}}\right) + \frac{1}{2}\text{cov}\left(\sum_{i\in s_1} \frac{y_{i_1}}{\pi_{i1}}, \sum_{i\in s_2} \frac{y_{i2}}{\pi_{i2}}\right), \tag{5.5}$$

where $\text{var}(.)$ and $\text{cov}(.,.)$ represent the variance and the covariance operators, respectively.

Following expression (5.4), if s_1 and s_2 are positively coordinated, the variance of $\hat{D}$ is reduced in general through sample overlap, since a positive covariance between $\sum_{i \in s_1} y_{i_1} / \pi_{i1}$ and $\sum_{i \in s_2} y_{i2} / \pi_{i2}$ is achieved compared to independent samples' selection. Similarly, from expression (5.5), independent samples' selection reduces the variance of $\hat{A}$ compared to positively coordinated samples because this covariance is zero. Negative coordination of samples can lead to a negative covariance between $\sum_{i \in s_1} y_{i_1} / \pi_{i1}$ and $\sum_{i \in s_2} y_{i_2} / \pi_{i2}$, and the variance of $\hat{A}$ can diminish compared to independent samples' selection.

A population of size $N = 100$ was created using models (5.1), (5.2), and (5.3). No births or deaths were considered in the population. The distance matrix was artificially generated using absolute values of independent runs from the $N(0,1)$ distribution. We set $\beta_0 = 4$, $\beta_1 = 2$, $\rho = 0.9$, $\delta_0 = 0$, $\delta_1 = 1$, $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma}_i \sim N(0,1), i = 1, \ldots, N,$ iid and $\gamma_i = \beta_1 \tilde{\gamma}_i, i = 1, \ldots, N.$ We also generated artificially x_{i1} as independent random draws from the $N(4,1)$ distribution. The correlation between y_1 and y_2 was approximately 0.72, while between y_t and $x_t, t = 1, 2$ was approximately 0.9. Based on this population, two different settings were created, by varying n_1 and n_2: in the first setting $n_1 = 10, n_2 = 25,$ while in the second one $n_1 = n_2 = 50.$ The correlation between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ was approximately 0.7 in both settings.

Monte Carlo simulation was used to study empirically the impact that each proposed method has on $\text{var}\left(\hat{D}\right)$ and $\text{var}\left(\hat{A}\right)$. For each setting, $m = 10^4$ samples were drawn as described in the beginning of Section 5.1. Figures 5.1 and 5.2 show boxplots corresponding to the $\hat{D}$ values obtained through Monte Carlo simulation, for both settings. The white boxplots correspond to the $\hat{D}$ values obtained from independent samples s_1 and s_2, while the grey ones to positively coordinated samples s_1 and s_2. The sampling design is specified below each boxplot (for example, TSCPS1_indep_0.25 indicates TSCPS 1 with independent samples' selection and $\alpha =$ 0.25 for both selections, while TSCPS1_pos_0.25 indicates TSCPS 1 with positively coordinated samples and $\alpha =$ 0.25 for both selections).

Similarly, Figures 5.3 and 5.4 show boxplots corresponding to the $\hat{A}$ values obtained through Monte Carlo simulation, for both settings, respectively. The white boxplots correspond to the $\hat{A}$ values obtained from independent samples s_1 and s_2, while the grey ones to negatively coordinated samples s_1 and s_2. In all figures, LPM with PRNs as well SCPS with PRNs show smaller spread of the $\hat{D}$ values and $\hat{A}$ values compared to Poisson sampling designs since both provide fixed sample sizes and are able to manage the spatial correlation of the data.

Figures 5.1 and 5.2 show a similar pattern of the boxplots: a larger overlap between s_1 and s_2 leads to a smaller spread of the $\hat{D}$ values. As expected, the spread of the $\hat{D}$ values is reduced for each type of positively coordinated samples compared to independent samples' selection. For LPM and SCPS designs this reduction is, however, less important. This fact can be explained by the smaller overlap between positively coordinated samples in LPM and SCPS designs compared to the other ones, as the examples in Section 5.1 show it. The larger sample sizes in the second setting reduce the spread of the $\hat{D}$ values in the case of positively coordinated samples (grey boxplots) compared to the independent sample selection (white boxplots). In Figures 5.3 and 5.4, negative coordination reduces in general the spread of the $\hat{A}$ values

compared to independent sample selection. As in Figures 5.1 and 5.2, this reduction is less important for LPM and SCPS compared for example to Poisson sampling and TSCPS 2.

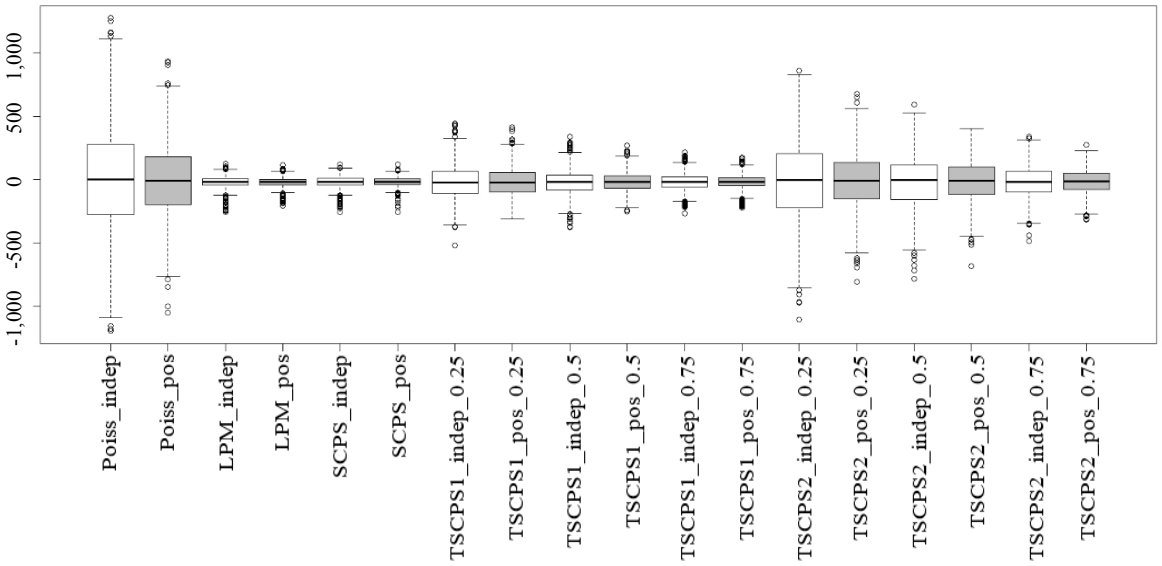

Figure 5.1 First setting: $N = 100$, $n_1 = 10$, $n_2 = 25$, boxplots of the $\hat{D}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to positively coordinated samples.

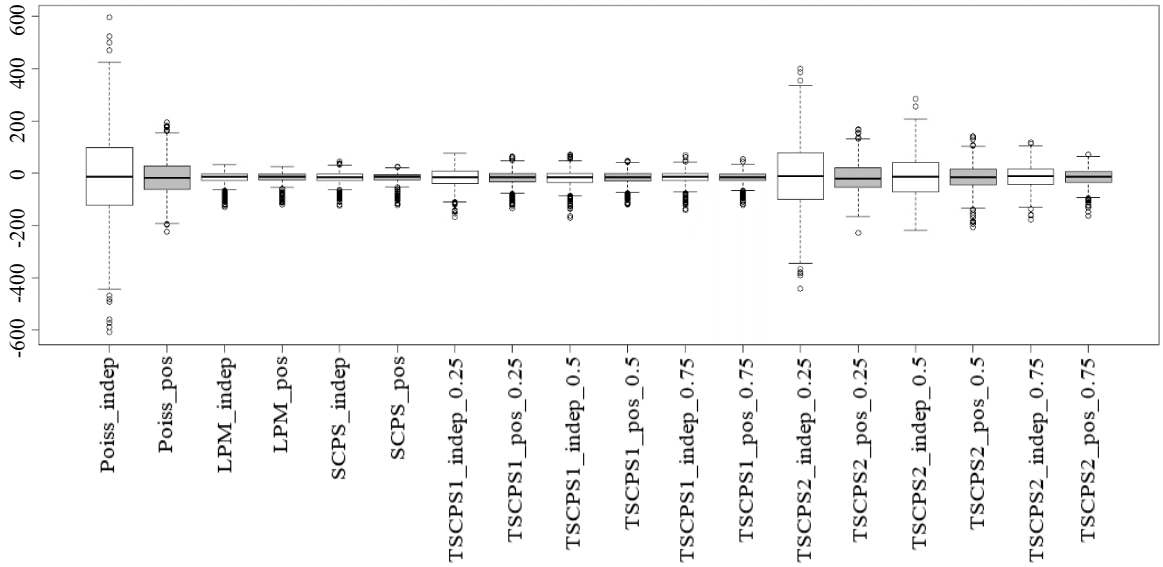

Figure 5.2 Second setting: $N = 100$, $n_1 = 50$, $n_2 = 50$, boxplots of the $\hat{D}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to positively coordinated samples.

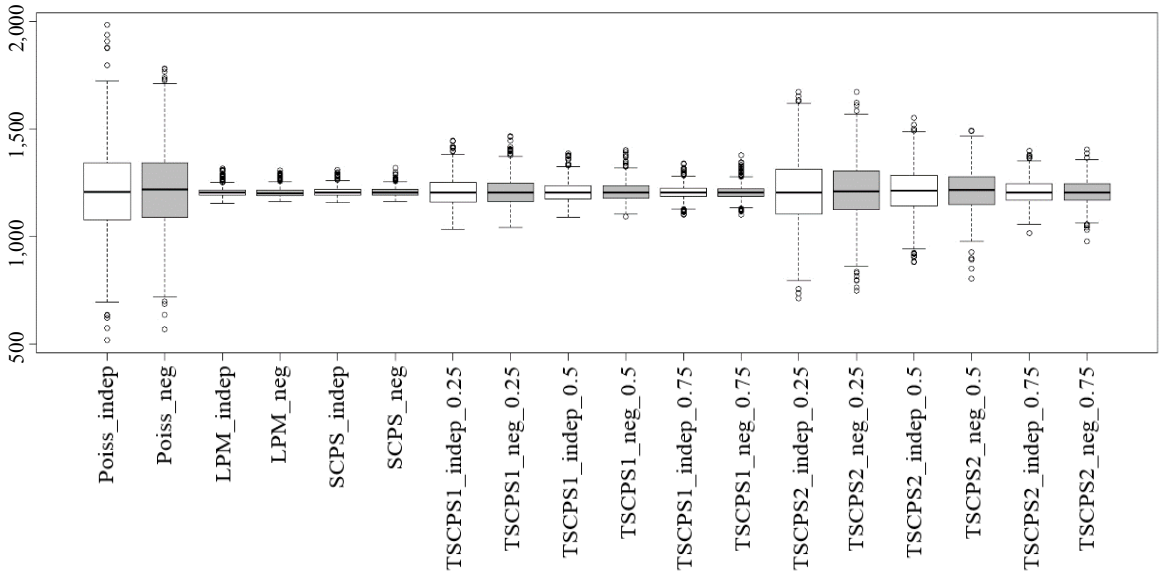

Figure 5.3 First setting: $N = 100$, $n_1 = 10$, $n_2 = 25$, boxplots of the $\hat{A}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to negatively coordinated samples.

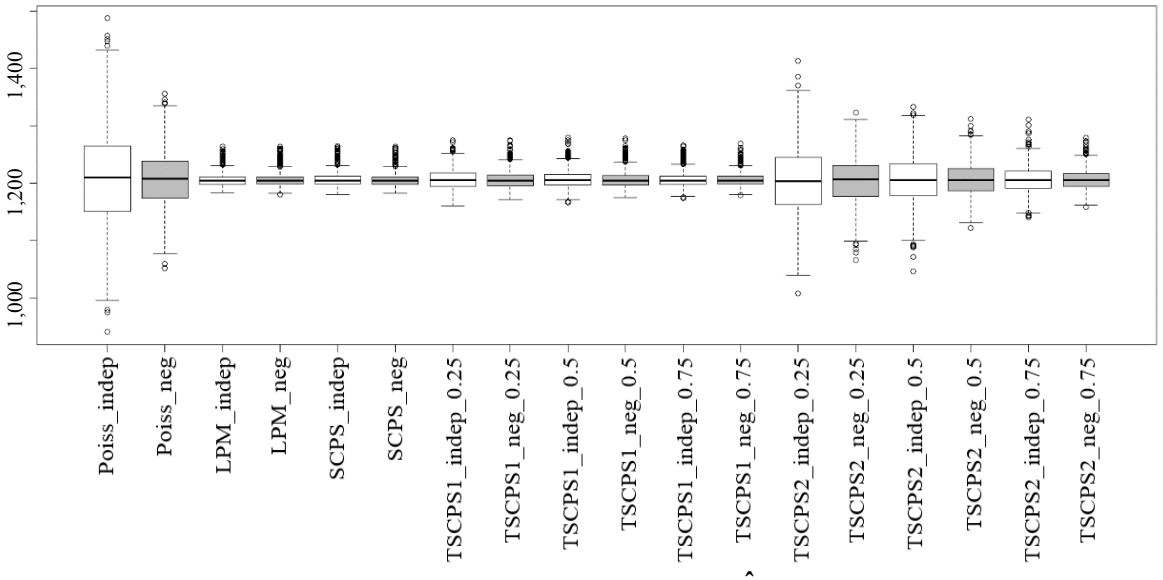

Figure 5.4 Second setting: $N = 100$, $n_1 = 50$, $n_2 = 50$, boxplots of the $\hat{A}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to negatively coordinated samples.

To quantify the performance of the proposed methods, for positive and negative coordination, respectively, the Monte Carlo variance was used

$$\mathrm{Var}_{\mathrm{MC}}(\theta) = \frac{1}{m-1}\sum_{\ell=1}^{m}(\theta_\ell - E_{\mathrm{sim}}(\theta))^2,$$

where θ_ℓ is the value of $\hat{D}$ or $\hat{A}$ obtained in the ℓ^{th} run and $E_{\mathrm{sim}}(\theta) = \frac{1}{m}\sum_{j=1}^{m}\theta_j$. The reduction in variance estimation through overlapped samples of $\hat{D}$ is summarized in Table 5.9. The table shows the values of the ratio between $\mathrm{var}_{\mathrm{MC}}\left(\hat{D}\right)$ obtained using positively coordinated samples and $\mathrm{var}_{\mathrm{MC}}\left(\hat{D}\right)$ using independent samples for both settings. We note that for all sampling designs this ratio is less than 1, indicating a variance reduction through sample overlap. Table 5.10 shows the values of the ratio between $\mathrm{var}_{\mathrm{MC}}\left(\hat{A}\right)$ obtained using negatively coordinated samples and $\mathrm{var}_{\mathrm{MC}}\left(\hat{A}\right)$ using independent samples for both settings. For the first setting, except for Poisson sampling, the ratio is close to 1, showing negligible improvement of the negatively coordinated samples compared to independent selections. Using larger sample sizes, the second setting shows an important improvement for TSCPS 2, but not for LPM and SCPS.

Table 5.9
Ratio between $\mathrm{var}_{\mathrm{MC}}\left(\hat{D}\right)$ obtained using positively coordinate samples and $\mathrm{var}_{\mathrm{MC}}\left(\hat{D}\right)$ using independent samples

Design		$n_1 = 10, n_2 = 25$ Ratio	$n_1 = 50, n_2 = 50$ Ratio
Poisson		0.481	0.178
LPM		0.759	0.679
SCPS		0.760	0.778
TSCPS 1	$\alpha = 0.25$	0.695	0.545
	$\alpha = 0.50$	0.739	0.700
	$\alpha = 0.75$	0.806	0.752
TSCPS 2	$\alpha = 0.25$	0.513	0.217
	$\alpha = 0.50$	0.571	0.319
	$\alpha = 0.75$	0.634	0.491

Table 5.10
Ratio between $\mathrm{var}_{\mathrm{MC}}\left(\hat{A}\right)$ obtained using negatively coordinate samples and $\mathrm{var}_{\mathrm{MC}}\left(\hat{A}\right)$ using independent samples

Design		$n_1 = 10, n_2 = 25$ Ratio	$n_1 = 50, n_2 = 50$ Ratio
Poisson		0.792	0.324
LPM		0.941	0.949
SCPS		0.921	0.901
TSCPS 1	$\alpha = 0.25$	0.932	0.679
	$\alpha = 0.50$	0.950	0.840
	$\alpha = 0.75$	0.953	0.876
TSCPS 2	$\alpha = 0.25$	0.828	0.387
	$\alpha = 0.50$	0.834	0.463
	$\alpha = 0.75$	0.919	0.597

In summary, LPM with PRNs, SCPS with PRNs and the TSCPS family reduce the Monte-Carlo variance of the differences through sample overlap compared to independent samples' selection in both settings. For the independent samples' selection, these methods are more precise than Poisson sampling because they are able to manage the spatial trend present in the variable of interest, and the sample sizes are fixed (for LPM and SCPS using the "maximal weight strategy") or less variable than for Poisson sampling. The Monte-Carlo variance of the averages is negligibly reduced by LPM and SCPS using negatively coordinated samples compared to independent samples in both settings. The transformed SCPS family shows a real improvement in the second setting, when n_1 and n_2 are relatively large, for all α.

6 Application to Swiss establishments

We illustrate the application of the proposed methods on real data. The data that we used was collected by the Swiss Federal Statistical Office and can be downloaded for free (https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/etablissements-emplois/statistique-structurel-entreprises-statent-depuis-2011.assetdetail.3303058.html). 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 an 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 and 2015 were considered the two time occasions. In 2013, a number of 7,057 units were available, while in 2015 this number was 7,104. The overall population was of size $N =$ 9,478. The difference in the sizes between the two time occasions was due to the 2,374 deaths and 2,421 births in 2015 compared to 2013. Figure 6.1 shows the geographical location of the units from the overall population. The parts inside of the figure with less locations correspond in majority to the Swiss Alps.

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 units.

We used the values of the expected sample sizes $n_1 =$ 1,000 and $n_2 =$ 800, while $\pi_{i,1}$ and $\pi_{i,2}$ were computed proportional to the same variable measured in 2013 and 2015, respectively. This variables was the total number of full-time equivalent employees of all establishments inside of a hectar. A matrix of size $N \times N$ of PRNs was generated for the LPM. For the other methods, the vector of PRNs was taken to be the

main diagonal of this matrix. In both time occasions respectively, we selected samples s_1 and s_2 using Poisson sampling with PRNs, LPM with PRNs, SCPS with PRNs, TSCPS 1 with PRNs (α = 0.25, 0.50, 0.75), and TSCPS2 with PRNs (α = 0.25, 0.50, 0.75). The Euclidean distance between locations was used in all methods, excepting Poisson sampling.

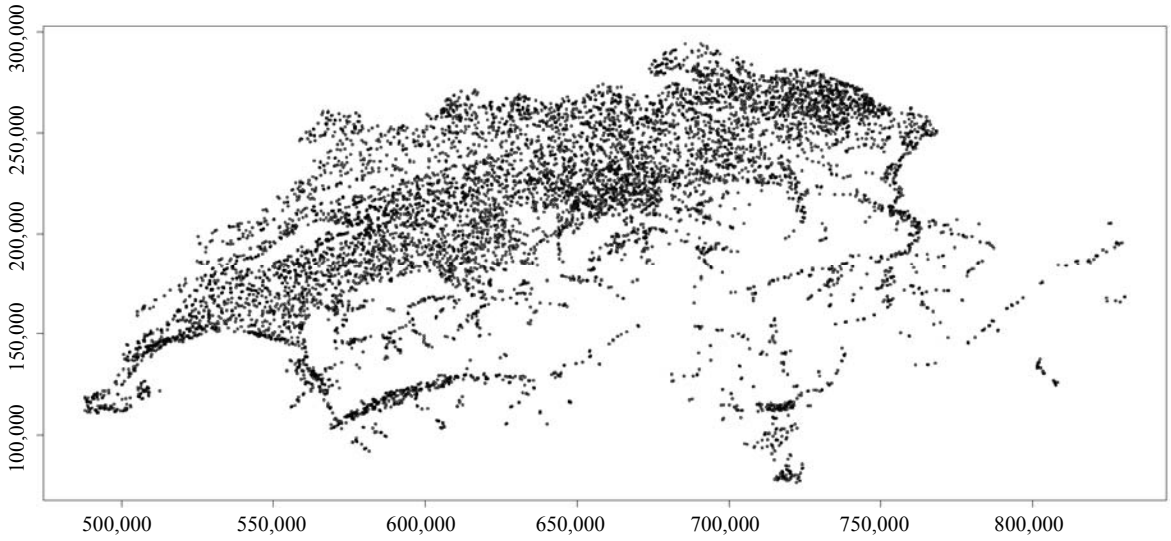

Figure 6.1 Swiss establishments aggregated data. Spatial distribution of the units in the overall population based on the census in 2013 and 2015.

We analyzed the selected samples in terms of realised overlap and B measure. To achieve this, positive and negative coordinations with PRNs were respectively applied. Table 6.1 shows the realised sample sizes as well as the overlap between different samples in both types of coordination. For the samples drawn in the first time occasion, the B measure given in expression (3.3) is also indicated. Poisson sampling presents the highest overlap in positive coordination (560, when AUB = 538.022), while LPM the smallest one. Due to the important changes in the population from 2013 to 2015, SCPS performs better than LPM, with an overlap of 329, but worse than Poisson sampling. All the members of the TSCPS family perform intermediately between Poisson sampling and SCPS, in function of the value of α. Negative coordination shows the same superiority of Poisson sampling, while the other designs exhibit smaller values of the realised overlap, with SCPS performing again better than LPM. Moving now to the spatial balancing feature, Poisson sampling yields the largest realised B measure, while LPM and SCPS as expected indicate the smallest ones. As in the results shown in Section 5.2, the members of the TSCPS family exhibit smaller realised B measure than Poisson sampling, but larger than SCPS. The application of the proposed methods on these real data indicates similar behavior of them with the simulation results shown in Sections 5.1 and 5.2.

Table 6.1
Swiss establishments aggregated data. N = 9,478, n_1 = 1,000, 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		1,010	840	560	779	46	0.387
LPM		1,000	800	270	800	93	0.161
SCPS		1,000	800	329	800	70	0.151
TSCPS 1	α = 0.25	999	799	459	800	64	0.178
	α = 0.50	1,000	799	420	800	66	0.217
	α = 0.75	1,000	800	366	800	67	0.178
TSCPS 2	α = 0.25	1,012	830	469	808	49	0.275
	α = 0.50	1,020	828	409	799	58	0.194
	α = 0.75	1,010	816	377	797	66	0.153

7 Conclusions

New methods are proposed to coordinate spatially balanced samples based on PRNs. The objective is two-fold: first, to achieve a good coordination degree between samples, and second to preserve a good spatial balance degree. With the coordination of LPM and SCPS a good degree of spatial balance is ensured. SCPS with PRNs is less memory consuming since only a PRN vector of size N is used, while for LPM one uses a matrix of dimension $N \times N$. Our examples concern moderate size populations, and a large N quickly introduces limits in the calculations. In practice, a large N leads to an oversized matrix to be employed in the coordination of LPM samples. In these cases, the method can be implemented using dynamic allocation of the computer memory. Despite this solution, limits of the proposed method are possible in practice.

In our simulations, SCPS tends to perform better than LPM in terms of overlap expectation and variance, for both positive and negative coordination. A good coordination of LPM samples is more difficult to achieve than of SCPS samples, because the same pairs of units should be considered in the sample selection process, instead of single units. If births or deaths appear in the population, the pairs used for the selection of s_1 may not be available any more for the selection of s_2. Thus, the sample coordination becomes poor. SCPS does not have this weakness, but instead the coordination level may drop as well if the weights are distributed very differently the second time compared to the first time. This is the reason why SCPS with PRNs performs worse than Poisson sampling with PRNs. LPM with PRNs may have better behavior in terms of overlap than SCPS with PRNs if changes in the population are not detected. This situation is exemplified in Table 5.1 when the static MU284 population is used.

As shown in our examples in Section 5.1 both methods show a weaker performance in terms of expected overlap than Poisson sampling. This is a normal feature of these methods since one imposes the fixed sample size constraint for LPM and SCPS. In order to overcome this weakness, we introduced a new family of designs, based on a transformation of SCPS and a choice of scalar α, $0 \leq \alpha \leq 1$. Each value of α leads to a member of this family. For $\alpha = 1$ one obtains SCPS, while for $\alpha = 0$ Poisson sampling. This family

of designs reminds us another family depending upon a scalar, the Pomix design (Kröger et al., 1999). The Pomix design is a mixture between Bernoulli and Poisson sampling, also used for coordination with PRNs.

For the transformed version of SCPS, the degree of coordination and spatial balance depend on the choice of α. Being a mixture of Poisson sampling and SCPS, it achieves a better coordination degree than SCPS. However, the improved degree of coordination comes at the cost of increased variance of sample size and reduced spatial balance as our examples in Section 5.1 and Section 4 showed. Based on our results, for the transformed SCPS, our recommendation is to use $\alpha = 0.5$ that represents a compromise between a good spatial balance degree and a good coordination degree. On the other hand, $\alpha = 0.5$ seems a good all-purpose suggestion since the results for variance estimation of differences and averages shown in Section 5.3 also indicate this value as a reasonable choice.

In our results shown in Section 5.3 LPM with PRNs, SCPS with PRNs and the TSCPS family reduce the Monte-Carlo variance of the differences when positively coordinated samples are used compared to independent samples' selection. In both used settings, it seems, however, that in the case of LPM with PRNs and SCPS with PRNs, variance reduction comes mainly from the combined effect of spatial balance and fixed sample size rather than from the effect of positive coordination. The Monte-Carlo variance of the averages is not always reduced in our examples when negatively coordinated samples are selected compared to independent samples; LPM with PRNs and SCPS with PRNs show in this case negligible improvement when negative coordination is used.

All the proposed methods can also be applied in the case where the spatial distance is replaced by a distance between auxiliary variables like the Mahalanobis distance. Thus, the sample coordination can be performed in the space spanned by these variables. The proposed methods allow thus not only a spatial sample coordination, but also the coordination of representative samples, in the terminology used by Grafström and Schelin (2014).

Aknowledgements

The authors wish to thank the Associate Editor and two referees for their valuable comments and suggestions that helped in improving the quality of the paper significantly.

References

Benedetti, R., Piersimoni, F. and Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85, 439-454.

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.

Cotton, F., and Hesse, C. (1992). Tirages coordonnés d'échantillons. Technical Report E9206, Direction des Statistiques Économiques, INSEE, Paris, France.

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.

Dubin, R.A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22, 3, 433-452.

GeoDa Center for Geospatial Analysis and Computation (2017). Sample data. http://spatial.uchicago.edu/sample-data. Accessed: 6-April-2017.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142, 139-147.

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

Grafström, A., and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 2, 277-290.

Grafström, A., and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14, 2, 120-131.

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

Haziza, D. (2013). Sampling and estimation procedures in business surveys: A discussion of some specific features. Seminar of the Royal Statistical Society, London, England.

Kröger, H., Särndal, C.-E. and Teikari, I. (1999). Poisson mixture sampling: A family of designs for coordinated selection using permanent random numbers. *Survey Methodology*, 25, 1, 3-11. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4707-eng.pdf.

Kröger, H., Särndal, C.-E. and Teikari, I. (2003). Poisson mixture sampling combined with order sampling. *Journal of Official Statistics*, 19, 59-70.

Mach, L., Reiss, P.T. and Şchiopu-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 Tillé, Y. (2005). Maximal and minimal sample co-ordination. *Sankhyā*, 67, part 3, 590-612.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

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