Maximal, minimal sample co-ordination and associated bounds

Alina Matei*

Abstract

Sample co-ordination maximizes or minimizes the overlap of two or more samples selected from overlapping populations. It can be applied to designs with simultaneous or sequential sample selection. The degree of co-ordination is measured by the expected sample overlap, which is limited by theoretical bounds (absolute upper and lower bounds). Two types of bounds can be defined: on unit level and on marginal sampling designs' level. We consider the bounds on unit level, which depend on unit inclusion probabilities. If the expected overlap equals the absolute upper bound, the sample co-ordination is maximal; if the expected overlap equals the absolute lower bound, the sample co-ordination is minimal. It is possible to construct optimal sampling designs for given unit inclusion probabilities to realize maximal or minimal co-ordination. This approach was developed by Matei and Skinner (2009) and uses a combination of the iterative proportional fitting algorithm and the linear programming implementation to controlled sampling method. We study here the performance of this method using a real scenario survey based on Swiss municipality data set. Despite the computing facilities available nowadays, the problem can be prohibitively large even for moderately large population and sample size. The method is useful to solve moderate-sized sample co-ordination problems.

Key Words: Fréchet bounds, joint selection probability of two samples, linear programming.

1. Introduction

The sample co-ordination problem is typical for surveys where the goal is to create a dependency between the selected samples in repeated surveys given their selection probabilities; a similar problem can be defined when two surveys are conducted simultaneously. The co-ordination between two or more samples is positive if the sample overlap is maximized, and negative if the sample overlap is minimized. The advantage of having overlapping samples is for example to improve the precision in the estimation of change or to reduce data collection costs, while for non-overlapping samples is to diminish the response burden of the selected units. In some applications the units in the population might represent primary sampling units. In other applications each population might represent a stratum, with sample co-ordination taking place stratum by stratum.

Different approaches have been proposed in the literature to do sample co-ordination such as: permanent random numbers (PRNs) or mathematical programming; see, for example, Ernst (1999); Ohlsson (1995); Mach et al. (2006) and the references therein. The degree of co-ordination is measured by the expected sample overlap, which is limited by theoretical bounds. These bounds depend on unit inclusion probabilities or on marginal probability sampling designs. We are interested here in studying methods that achieve the theoretical bounds on unit level. The paper is organized as follows. In Section 2 the general framework is shown. In Section 3 two sample co-ordination criteria are presented: one on unit level and other one on probability sampling designs' level. We recall general conditions to achieve the theoretical bounds on unit level, and show some necessary conditions to achieve the equivalence of the two criteria. Section 4 gives a practical evaluation of a

*Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000, Neuchâtel, Switzerland and Institute of Pedagogical Research and Documentation (IRDP), Fbg. de l'Hôpital 43, Neuchâtel, Switzerland, alina.matei@unine.ch

method constructed to achieve the theoretical bounds on unit level. This method combines the iterative proportional fitting algorithm and the linear programming implementation of the controlled sampling method. Conclusions are drawn in Section 5.

2. Framework

Consider the selection of samples s^1 and s^2 from populations U^1, U^2, respectively, and let $U = \{1, \ldots, N\} = U^1 \cup U^2$. The sets of possible samples s^1 and s^2 are denoted $\mathcal{S}^1 = \{s_1^1, ..., s_m^1\}$ and $\mathcal{S}^2 = \{s_1^2, ..., s_q^2\}$, respectively. Let $s_{ij} = (s_i^1, s_j^2)$ referred to as the bi-sample. The set of all possible bi-samples is denoted $\mathcal{S} = \{s_{ij} | s_{ij} = (s_i^1, s_j^2), s_i^1 \subseteq \mathcal{S}^1, s_j^2 \subseteq \mathcal{S}^2, i = 1, \ldots, m, j = 1, \ldots q\}$. The overall sampling design is represented by the probability that bi-sample s_{ij} is selected, denoted $p_{ij} = p(s_i^1, s_j^2) = p(s_{ij})$ for $s_{ij} \subseteq \mathcal{S}$. The marginal sampling designs for s^1 and s^2 are given by the probabilities $p^1(s_i^1)$ and $p^2(s_j^2)$, respectively. We have $\sum_{s_i^1 \subseteq \mathcal{S}^1} p^1(s_i^1) = \sum_{s^2 \subseteq \mathcal{S}^2} p^2(s_j^2) = 1$, $\sum_{j=1}^{q} p_{ij} = p^1(s_i^1)$, and $\sum_{i=1}^{m} p_{ij} = p^2(s_j^2)$. The overall sampling design is said to be co-ordinated if $p(s_i^1, s_j^2) \neq p^1(s_i^1) p^2(s_j^2)$ (see Cotton and Hesse, 1992; Mach et al., 2006), i.e. if the two samples are not selected independently.

The size of the overlap between two samples s_i^1 and s_j^2 is denoted $c_{ij} = |s_i^1 \cap s_j^2|$ and, in general, is random. Let c denote the random overlap between two samples. The degree of co-ordination is measured by the expected sample overlap, given by:

$$E(c) = \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} = \sum_{k \in U} \pi_k^{1,2}, \tag{1}$$

where

$$\pi_k^{1,2} = \sum_{\substack{k \ni s_i^1, k \ni s_j^2 \\ s_{ij} = (s_i^1, s_j^2) \subseteq \mathcal{S}}} p(s_i^1, s_j^2)$$

is the probability that unit k is included in both samples.

3. Two sample co-ordination criteria

Two different criteria (among others) can be used to measure the quality of sample co-ordination schemes. Using the first criterium the degree of sample co-ordination is maximized/minimized on unit level; using the second one, the degree of sample co-ordination is maximized/minimized on marginal probability sampling designs' level.

The first one measures this quality by maximizing the expected sample overlap in (1) on unit level for a positive co-ordination and minimizing it in the negative case. To obtain bounds on the expected overlap on unit level, let

$$\pi_k^1 = \sum_{\substack{s_i^1 \ni k \\ s_i^1 \subseteq \mathcal{S}^1}} p^1(s_i^1)$$

be the first-order inclusion probability of unit $k \in U$ for the first design and let

$$\pi_k^2 = \sum_{\substack{s_j^2 \ni k \\ s_j^2 \subseteq \mathcal{S}^2}} p^2(s_j^2)$$

be the first-order inclusion probability of unit $k \in U$ for the second design. If $k \in U^1 \setminus U^2, \pi_k^2 = 0$ and if $k \in U^2 \setminus U^1, \pi_k^1 = 0$. Using the Fréchet bounds of $\pi_k^{1,2}$ we have

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{1,2} \leq \min(\pi_k^1, \pi_k^2),$$

and the expected sample overlap is limited by the following bounds

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

The first criterium applied to a sample co-ordination scheme implies that $E(c)$ equals the right-hand side in (2) for a positive co-ordination and the left-hand side in (2) for a negative co-ordination.

The second criterium uses the joint probability sampling design. Based on the Fréchet bounds of $p(s_i^1, s_j^2)$ we have in (1)

$$\sum_{k \in U} \sum_{k \ni s_i^1 \cap s_j^2} \max(0, p_i^1 + p_j^2 - 1) \leq E(c) \leq \sum_{k \in U} \sum_{k \ni s_i^1 \cap s_j^2} \min(p_i^1, p_j^2). \tag{3}$$

The second criterium applied to a sample co-ordination scheme implies that $E(c)$ equals the right-hand side in (3) for a positive co-ordination and the left-hand side in (3) for a negative co-ordination. To our knowledge, the second criterium has not been yet studied in the literature.

The right-hand side in Expression (3) depends on the number of bi-samples containing a unit $k \in U$, and it can be very large in practice (and thus never reached). To our knowledge, it is not possible, however, to give a monotonic relationship between the two upper bounds: the right-side in (2) and the right-side in (3).

Note that the methods proposed in the literature do not always achieve the proposed criteria. The first criterium is achieved using the Keyfitz's method (Keyfitz, 1951), Poisson sampling with permanent random numbers for each marginal design (Brewer et al., 1972, 1984), and the sequential simple random sample without replacement (srswor) with permanent random numbers (Ohlsson, 1995) in the case of stratified designs. The second criterium can be achieved in very particular cases given in Section 3.2.

In what follows, we recall general properties to reach the first criterium, and we develop conditions upon which the two criteria are equivalent.

3.1 General conditions for maximal/minimal co-ordination on unit level

Matei and Tillé (2005) call $\sum_{k \in U} \min(\pi_k^1, \pi_k^2)$ the Absolute Upper Bound (AUB) and say that *maximal sample co-ordination* occurs when equality holds in the right part of (2). Similar, $\sum_{k \in U} \max(0, \pi_k^1 + \pi_k^2 - 1)$ is called the Absolute Lower Bound (ALB) and the *minimal sample co-ordination* occurs when equality holds in the left part of (2). The following result summarizes theoretical conditions to reach the AUB and ALB, respectively, for two fixed marginal designs, but arbitrary (for a proof, see Matei and Tillé, 2005).

Proposition 1 *Let p_{ij} be the joint selection probability of two samples s_i^1 and s_j^2, with given marginal designs $\mathbf{p}^1$ and $\mathbf{p}^2$.*

1. *Let $I = \{k \in U | \pi_k^1 \leq \pi_k^2\}$ be the set of 'increasing' units, and let $D = \{k \in U | \pi_k^1 > \pi_k^2\}$ be the set of 'decreasing' units, with $U = I \cup D$, and $I \cap D = \emptyset$. The AUB is achieved iff the following two relations are fulfilled for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, q\}$:*

a) $p_{ij} = 0,$ *for all* s_i^1 *and* s_j^2 *for which* $(s_i^1 \backslash s_j^2) \cap I \neq \varnothing$;

b) $p_{ij} = 0,$ *for all* s_j^2 *and* s_i^1 *for which* $(s_j^2 \backslash s_i^1) \cap D \neq \varnothing$.

2. *Let* $\widetilde{I} = \{k \in U | \pi_k^1 + \pi_k^2 - 1 \leq 0\}$, *and let* $\widetilde{D} = \{k \in U | \pi_k^1 + \pi_k^2 - 1 > 0\}$, *with* $U = \widetilde{I} \cup \widetilde{D}$, *and* $\widetilde{I} \cap \widetilde{D} = \emptyset$. *The ALB is achieved iff the following two relations are fulfilled for all* $i \in \{1, \ldots, m\}$ *and* $j \in \{1, \ldots, q\}$:

   c) $p_{ij} = 0,$ *for all* s_i^1 *and* s_j^2 *for which* $(s_i^1 \cap s_j^2) \cap \widetilde{I} \neq \varnothing$;

   d) $p_{ij} = 0,$ *for all* s_i^1 *and* s_j^2 *for which* $U \setminus \left((s_i^1 \cup s_j^2) \cap \widetilde{D}\right) \neq \varnothing$.

Remark 1 *If the following conditions are fulfilled:* $(s_i^1 \backslash s_j^2) \cap I = \varnothing$ *and* $(s_j^2 \backslash s_i^1) \cap D = \varnothing$, *for all* s_i^1 *and* $s_j^2, i = 1, \ldots, m, j = 1, \ldots, q$ *it is possible to show that the AUB is also reached (for a proof see Matei and Skinner, 2009).*

3.2 Equivalence of the two criteria

The problem of interest is to see in which conditions the two criteria are equivalent. This problem is considered here because, in general, one measures the performance of a co-ordination method using only the theoretical bounds defined on unit level. Consider the case of positive co-ordination.

Proposition 2 *If the the right-hand in (2) is reached by a bi-design* $\mathbf{P}$*, the AUB is smaller or equal to the right-hand in (3).*

Proof 1 *Let* $k \in I$. *Following Proposition 1, point a), the AUB is reached if and only if* $p_{ij} = 0,$ *for all* s_i^1 *and* s_j^2 *for which* $(s_i^1 \backslash s_j^2) \cap I \neq \varnothing$. *We have*

$$\pi_k^1 = \sum_{k \ni s_i^1} p_i^1 = \sum_{k \ni s_i^1} \sum_{j=1}^{q} p_{ij} = \sum_{k \ni s_i^1} (\sum_{k \ni s_j^2} p_{ij} + \sum_{k \not\ni s_j^2} p_{ij}) = \sum_{k \ni s_i^1 \cap s_j^2} p_{ij} + \sum_{k \ni s_i^1 \backslash s_j^2} p_{ij} =$$

$$\sum_{k \ni s_i^1 \cap s_j^2} p_{ij} \leq \sum_{k \ni s_i^1 \cap s_j^2} \min(p_i^1, p_j^2). \tag{4}$$

Similarly, using Proposition 1, point b), for all $k \in D$, *we have*

$$\pi_k^2 \leq \sum_{k \ni s_j^2 \cap s_i^1} \min(p_i^1, p_j^2). \tag{5}$$

Finally

$$\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = \sum_{k \in I} \pi_k^1 + \sum_{k \in D} \pi_k^2 \leq$$

$$\sum_{k \in I} \sum_{k \ni s_i^1 \cap s_j^2} \min(p_i^1, p_j^2) + \sum_{k \in D} \sum_{k \ni s_j^2 \cap s_i^1} \min(p_i^1, p_j^2) = \sum_{k \in U} \sum_{k \ni s_i^1 \cap s_j^2} \min(p_i^1, p_j^2).$$

Remark 2 *Obviously, the right-hand side in (3) is reached by a bi-design* $\mathbf{P}$ *if* $p_{ij} = \min(p_i^1, p_j^2)$ *for all* s_i^1 *and* s_j^2 *such that* $s_i^1 \cap s_j^2 \neq \varnothing$.

Remark 3 *If the* AUB *is reached by a bi-design* $\mathbf{P}$, *the* AUB *equals the right-hand side in (3) if this bound is also reached (it follows from Expressions (4) and (5)).*

Proposition 3 *If the the right-hand side in (3) is reached by a bi-design* $\mathbf{P}$, *the AUB is larger or equal to the right-hand side in (3) if the following condition is fulfilled*

$$\sum_{k\in I}\sum_{k\ni s_i^1\backslash s_j^2} p_{ij} + \sum_{k\in D}\sum_{k\ni s_j^2\backslash s_i^1} p_{ij} \geq 0.$$

Proof 2 *Let* $k \in I$. *We have*

$$\pi_k^1 = \sum_{k\ni s_i^1} p_i^1 = \sum_{k\ni s_i^1}\sum_{j=1}^{q} p_{ij} = \sum_{k\ni s_i^1}(\sum_{k\ni s_j^2} p_{ij} + \sum_{k\not\ni s_j^2} p_{ij}).$$

Let $k \in D$. *We have*

$$\pi_k^2 = \sum_{k\ni s_j^2} p_j^2 = \sum_{k\ni s_j^2}\sum_{i=1}^{m} p_{ij} = \sum_{k\ni s_j^2}(\sum_{k\ni s_i^1} p_{ij} + \sum_{k\not\ni s_i^1} p_{ij}).$$

If the the right-hand side in (3) is reached, it follows that

$$\sum_{k\in U}\min(\pi_k^1, \pi_k^2) = \sum_{k\in I}\pi_k^1 + \sum_{k\in D}\pi_k^2 =$$

$$\sum_{k\in I}\sum_{k\ni s_i^1\cap s_j^2}\min(p_i^1, p_j^2) + \sum_{k\in D}\sum_{k\ni s_j^2\cap s_i^1}\min(p_i^1, p_j^2) + \sum_{k\in I}\sum_{k\ni s_i^1\backslash s_j^2} p_{ij} + \sum_{k\in D}\sum_{k\ni s_j^2\backslash s_i^1} p_{ij} =$$

$$\sum_{k\in U}\sum_{k\ni s_i^1\cap s_j^2}\min(p_i^1, p_j^2) + \sum_{k\in I}\sum_{k\ni s_i^1\backslash s_j^2} p_{ij} + \sum_{k\in D}\sum_{k\ni s_j^2\backslash s_i^1} p_{ij} \geq \sum_{k\in U}\sum_{k\ni s_i^1\cap s_j^2}\min(p_i^1, p_j^2).$$

If

$$\sum_{k\in I}\sum_{k\ni s_i^1\backslash s_j^2} p_{ij} + \sum_{k\in D}\sum_{k\ni s_j^2\backslash s_i^1} p_{ij} > 0$$

the AUB is larger then the right-hand side in (3).

4. Practical evaluation of a method to reach the maximal/minimal sample co-ordination on unit level

In this section we study the performance of the method developed by Matei and Skinner (2009) to reach the maximal/minimal sample co-ordination on a real data set. This method can satisfy the first criterium given in Section 3. The example of this study uses Swiss municipalities data set which is available in the R 'sampling' package (Tillé and Matei, 2012). It is a data set having 2896 observations and 22 variables. This data set is the official register of Swiss municipalities. Swiss Federal Statistical Office uses this register for its Environmental quality and behavior survey.

Before giving the application, we briefly describe here this method for the positive co-ordination (the negative coordination is similar). Matei and Tillé (2005) proposed to use the Iterative Proportional Fitting (IPF) procedure (Deming and Stephan, 1940) to obtain a joint probability design p_{ij} which fulfill the conditions a) and b) given in Proposition 1. Let $\mathbf{P} = (p_{ij})_{m\times q}$ be initially any matrix of $m \times q$ dimension. Using Proposition 1, matrix $\mathbf{P}$ is modified by assigning zero values to some p_{ij}, for example by applying p^1 and p^2 independently, i.e. $p_{ij} = p^1(s_i^1)p^2(s_j^2)$. Now, the total of row i and the total of column j of $\mathbf{P}$ are different from the given values $p^1(s_i^1)$ and $p^2(s_j^2)$ and the constraints of the joint probabilities p_{ij} are not respected, i.e. $\sum_{j=1}^{q} p_{ij} \neq p^1(s_i^1)$, $\sum_{i=1}^{m} p_{ij} \neq p^2(s_j^2)$. To respect these constraints, the non-zero values of $\mathbf{P}$ are modified using the IPF procedure. The IPF procedure iteratively modifies the matrix $\mathbf{P}$ and is applied until convergence is reached. The final matrix $\mathbf{P}$ has the property that $\sum_{i=1}^{m}\sum_{j=1}^{q} c_{ij}p_{ij}$ equals the AUB, if the latter can be achieved. The procedure is not constructive in the case where AUB cannot be achieved. The AUB cannot be achieved if Proposition 1 gives only zero values to the $\mathbf{P}$ elements for some row(s) i' and/or column(s) j' since the corresponding $p^1(s_{i'}^1)$ and $p^2(s_{j'}^2)$ are strictly positive. In terms of controlled selection method, these are 'nonpreferred' samples, and their influence should be reduced. Matei and Skinner (2009) extended the method given by Matei and Tillé (2005) and develop a method based on the controlled selection method to reduce the influence of these 'nonpreferred' samples in the case where the AUB is not achieved using the previous method. The extended method uses the linear programming implementation of the controlled selection method (Rao and Nigam, 1990). Contrary to the transportation method and the method of Matei and Tillé (2005), this approach does not necessarily maintain the constraints that the marginal sample designs are fixed. The extended method computes new selection probabilities p_*^1 and p_*^2, such that the marginal constraints are satisfied, i.e.

$$\sum_{j=1}^{q} p_{i'j} = p_*^1(\,s_{i'}^1) = 0 \text{ and } \sum_{i=1}^{m} p_{ij'} = p_*^2(\,s_{j'}^2) = 0,$$

and the inclusion probabilities π_k^1, π_k^2 are preserved.

Let $\mathcal{S}_*^1 = \{s_i^1 \subseteq \mathcal{S}^1 \,|\, p_{ij} = 0, \text{ for all } j = 1,\ldots,q\}$ and $\mathcal{S}_*^2 = \{s_j^2 \subseteq \mathcal{S}^2 \,|\, p_{ij} = 0, \text{ for all } i = 1,\ldots,m\}$. Consider the following problem

$$\min_{p_*^t} \sum_{s_\ell^t \subseteq \mathcal{S}_*^t} p_*^t(s_\ell^t), \tag{6}$$

subject to

$$\left|\begin{array}{l} \sum_{\substack{k\ni s_\ell^t \\ s_\ell^t\subseteq \mathcal{S}_*^t}} p_*^t(s_\ell^t) = \pi_k^t,\ k\in U, \\ \sum_{\ell=1}^{r} p_*^t(s_\ell^t) = 1, \\ p_*^t(s_\ell^t) \geq 0,\ \ell = 1,\ldots,r, \end{array}\right.$$

where $t = 1, 2$ and $r = m$ if $t = 1$ and $r = q$ if $t = 2$.

For $t = 1$, Problem (6) is used to reduce the probability of selecting samples $s_{i'}^1$ for which $p_{ij} = 0$, for all $j = 1,\ldots,q$. Similarly, for $t = 2$, Problem (6) is used to reduce the probability of selecting samples $s_{j'}^2$ for which $p_{ij} = 0$, for all $i = 1,\ldots,m$. A solution to the linear programming problem (6) always exists. If the value of the objective function equals zero, the AUB is achieved. In the case of simple random sample without replacement, Lahiri and Mukerjee (2000) developed a method to reduce the size of Problem (6) based on equivalence classes. Another approach useful for proportional to size sampling designs is to use quadratic programming with the same constraints as in Problem (6) (see Tiwari et al., 2007); the result is a nearest proportional to size sampling design.

Several examples using the Swiss municipalities data set are given below. Each population represents a stratum, and sample co-ordination takes place stratum by stratum. Samples in the two surveys are drawn simultaneously. Due the problem complexity (enumeration of all possible samples and linear programming), the examples below are restricted to small strata and small sample sizes. Two cases are taken into account: no changes in strata and different stratification in the two surveys.

1. No changes in strata: both strata are defined by the canton CT variable, giving 26 strata of sizes between 1 and 150. Consider the stratum 6 (canton of Obwald) for

both surveys, where the stratum size is equal to 7. We consider simple random sampling without replacement in both surveys, using the sample sizes $n_1 = 4, n_2 = 3$, respectively. The total number of samples in the first survey is 35. The same number of samples is available for the second survey. The vectors of inclusion probabilities are $\boldsymbol{\pi}^1 = (\pi_k^1)'_7 = (0.57)'_7$ and $\boldsymbol{\pi}^2 = (\pi_k^2)'_7 = (0.43)'_7$. The set D contains all the seven units, while the set I is empty. Proposition 1 allows the construction of the matrix $\mathbf{P}_{35\times 35}$ which does not have any zero row and zero column. The IPF procedure is successfully applied on the matrix $\mathbf{P}$ resulting in the matrix $\tilde{\mathbf{P}} = (\tilde{p}_{ij})_{35\times 35}$ which fulfills the given margins $\mathbf{p}_1 = (p^1(s_i^1))'_{35} = (0.028)'_{35}$ and $\mathbf{p}_2 = (p^2(s_j^2))'_{35} = (0.028)'_{35}$ and the expected overlap $\sum_i \sum_j c_{ij}\tilde{p}_{ij}$ equals $\sum_{k\in U} \min(\pi_k^1, \pi_k^2) = 3$.

2. Different stratification: The units in the first survey are stratified using the variable canton; the units in the second survey are stratified according to canton and municipality size, altogether 78 strata with 2896 units. The last variable is a categorical one (three categories, small, medium and large) and is derived from the number of inhabitants of each municipality.

   (a) We have considered for the first survey the stratum 15 (canton of Appenzell Ausserrhoden) containing 20 units. For the second survey, the canton 15 and small size municipalities are cross-stratified; the cross-stratum population size is 18. Thus, two units change stratum. We consider simple random sampling without replacement for both surveys, using the following sample sizes $n_1 = 4, n_2 = 2$, respectively. The total number of samples in the first survey is 4845 versus 153 in the second one. The vectors of inclusion probabilities are $\boldsymbol{\pi}_k^1 = (0.2)'_{20}$ and $\boldsymbol{\pi}_k^2 = (0.11)'_{18}$. The population $U = U^1 \cup U^2$ consists in 20 units. The set I is empty for this example, while the set D contains all 20 units. By applying the procedure given before the matrix $\mathbf{P}_{4845\times 153}$ is computed; it does not have any zero row and zero column. The IPF procedure is successfully applied on the matrix $\mathbf{P}$ resulting in the matrix $\tilde{\mathbf{P}} = (\tilde{p}_{ij})_{4845\times 153}$ and the expected overlap $\sum_i \sum_j c_{ij}\tilde{p}_{ij}$ equals $\sum_{k\in U} \min(\pi_k^1, \pi_k^2) = 2$.

   (b) Stratum 4 (canton of Uri) with 20 units was considered for the first survey. For the second survey, the cell cross stratum was determined by canton 4 (canton of Uri) and small size municipalities; the cross-strata population size is 10. Thus, 50% of the units change stratum. As before, simple random sampling without replacement was used in both cases. The number of total samples of size 4 in the first survey was 4845, while for the second survey a number of 120 samples of size 3 was computed. The vectors of inclusion probabilities $\boldsymbol{\pi}_k^1 = (0.2)'_{20}$ and $\boldsymbol{\pi}_k^2 = (0.3)'_{10}$ determine the sets I and D, each one of size 10. Using Proposition 1, the matrix $\mathbf{P}_{4845\times 120}$ was computed. The matrix contains 210 zero rows, but no zero columns. Problem (6) was applied to diminish the importance of the 210 non-preferred samples in the first survey. The linear programming gives a solution with 20 nonzero selection probabilities $p_*^1(s_i^1)$. The corresponding matrix is now $\mathbf{P}_{20\times 120}$. The IPF procedure was successfully applied on this matrix, resulting in the matrix $\tilde{\mathbf{P}} = (\tilde{p}_{ij})_{20\times 120}$. Finally, the expected overlap $\sum_i \sum_j c_{ij}\tilde{p}_{ij} = \sum_{k\in U} \min(\pi_k^1, \pi_k^2) = 2$.

   (c) A more complicated example is given bellow. We consider the same setting as in Example (a), but now we take $n_2 = 3$. The total number of possible samples in the second survey is 816. The corresponding vector of inclusion probabilities is $\boldsymbol{\pi}_k^2 = (0.16)'_{18}$. The set I is empty, while the set D contains all the

20 units. The matrix $\mathbf{P}_{4845\times 816}$ is computed using Proposition 1; it contains 153 zero rows and there is no single column in $\mathbf{P}$ consisting of zeros. Problem (6) was applied to reduce the importance of the 153 non-preferred samples. The linear programming reduced very much the sample support of the first survey, determining only 20 samples s_i^1 with non-zero probabilities. The corresponding matrix $\mathbf{P}_{20\times 816}$ contains 769 zero columns. Problem (6) was again applied to reduce now the importance of the 769 non-preferred samples s_j^2 in the second survey. The resulting probability sampling p_*^2 contains 17 non-zero values. Now, the current matrix $\mathbf{P}_{20\times 17}$ obtained using Proposition 1 contains five zero rows and four zero columns. The third application of the linear programming for the five non-preferred samples (on rows) does not eliminate any required sample. A similar result is obtained in the fourth application of the linear programming to eliminate the four non-preferred samples (on columns). The algorithm stopped here since the current matrix $\mathbf{P}_{20\times 17}$ with five zero rows and four zero columns cannot be any more modified; consequently the IPF procedure cannot be applied. The final marginal probabilities p_*^1 and p_*^2 are given in Table 1. For this example, the AUB cannot be reached. To obtain, however, a solution the elements of the zero rows and zero columns were replaced by a very small value (here 2.2×10^{-16}), and the IPF procedure was successfully applied, resulting in the matrix $\tilde{\mathbf{P}} = (\tilde{p}_{ij})_{20\times 17}$ which verify the marginal probabilities given in Table 1. Finally, the expected overlap $\sum_i \sum_j c_{ij}\tilde{p}_{ij} = 2.204$. Yet, $\sum_{k\in U} \min(\pi_k^1, \pi_k^2) = 3$.

If we come back to the initial matrix $\mathbf{P}_{4845\times 816}$ and replace the 153 null rows by 2.2×10^{-16}, we obtain (after the IPF application) the value $\sum_i \sum_j c_{ij}\widetilde{p'}_{ij} = 2.916$, which is much closer to the AUB. In this case, the application of the linear programming is not desirable, since a simpler application of the IPF procedure gives a better result. More research should be done to develop a controlled selection method which selects only the nonpreferred samples.

Table 1: Final marginal probabilities p_*^1 and p_*^2 for Example (c)

p_*^1	0.03415	0.00488	0.00488	0.05366	0.09919	0.08293
	0.00325	0.08455	0.03740	0.01789	0.02114	0.07642
	0.02927	0.09431	0.10244	0.05366	0.05366	0.04878
	0.04878	0.04878				
p_*^2	0.02500	0.02500	0.11458	0.05208	0.05208	0.06250
	0.06250	0.04167	0.04167	0.08333	0.08333	0.11667
	0.00208	0.07083	0.07083	0.04792	0.04792	

5. Conclusions

For the problem of two sample co-ordination, we have considered theoretical bounds of the expected overlap defined on unit level and on marginal probability sampling designs' level. We have also discussed conditions to reach these bounds in the case of positive co-ordination. The bounds constructed on unit level seem to be easier to reach in practice than the marginal bounds, which can be very large. In general, one measures the performance of the methods proposed in the literature using the theoretical bounds defined on unit level. Yet, there are particular cases where the bounds constructed on marginal sampling level are smaller than those on unit level, and it seems necessary to be also considered.

A method constructed to reach the bounds on unit level was developed by Matei and Skinner (2009). In the second part of the paper, the performance of this method has been investigated using a real scenario survey. In many cases, the method reach the theoretical bounds defined on unit level. Sometimes, however, it is better to search for sub-optimal solutions (see the last example), where these bounds are not reached. The method becomes impractical when the number of all possible samples is very large. For this case, the process of enumeration of all possible samples and definition of the constraints in the linear programming become infeasible in practice. Consequently, the method is useful to solve moderate-sized sample co-ordination problems.

REFERENCES

Brewer, K., Early, L., and Hanif, M (1984), "Poisson, modified Poisson and Collocated sampling", *Journal of Statistical Planning and Inference*, 10, 15–30.

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.

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

Deming, W. and Stephan, F. (1940), "On a least square adjustment of sampled frequency table when the expected marginal totals are known", *Annals of Mathematical Statistics*, 11, 427–444.

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, pp. 168–182, Helsinki, Finland.

Keyfitz, N. (1951), "Sampling with probabilities proportional to size: adjustment for changes in the probabilities", *Journal of American Statistics Association*, 46, 105–109.

Lahiri, P. and Mukerjee, R. (2000), "On a simplification of the linear programming approach to controlled sampling", *Statistica Sinica*, 10, 1171–1178.

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

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

Ohlsson, E. (1995), "Coordination of samples using permanent random numbers", in *Business Survey Methods*, eds. Cox, B. G., Binder, D. A., Chinnapa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., chapter 9, Wiley, pp. 153–169.

Pathak, P. K. and Fahimi, M. (1992), "Optimal integration of surveys", in *Current issues in statistical inference: Essays in honor of D. Basu*, eds. Ghosh, M. and Pathak, P. K., Hayward, CA: Institute of Mathematical Statistics, pp. 208–224.

Rao, J. N. K. and Nigam, A. K. (1990), "Optimal controlled sampling designs", *Biometrika*, 77, 807–814.

Tillé, Y. and Matei, A. (2012). sampling: Survey Sampling. R package version 2.5.

Tiwari, N., Nigam, A. K., and Pant, I. (2007), "On an optimal controlled nearest proportional to size sampling scheme", *Survey Mathodology*, 33(1), 87–94.