# Evaluation and development of strategies for sample coordination and statistical inference in finite population sampling

**Desislava Nikolova Nedyalkova**

Thesis submitted in fulfillment of the requirements for the degree of

PhD in Statistics

Under the supervision of

**Yves Tillé**, Université de Neuchâtel

Members of the Evaluation committee:

**Catalin Starica**, Université de Neuchâtel, Chairman
**Ray Chambers**, University of Wollongong, Australia
**Jean-Claude Deville**, CREST-ENSAI, France
**Wayne Fuller**, Iowa State University, USA

Defended on June, 23rd 2009

Faculté des Sciences Economique, Université de Neuchâtel
2009

Université de Neuchâtel

**IMPRIMATUR POUR LA THESE**

Evaluation and development of strategies for sample coordination and statistical inference in finite population sampling

**Desislava Nikolova NEDYALKOVA**

---

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES ECONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
M. Jean-Claude Deville (Ecole Nationale de la Statistique et de l'Analyse de l'Information, Rennes, France),
Prof. Wayne Fuller (Iowa State University, USA)
Prof. Ray Chambers (University of Wollongong, Australie)
Prof. Catalin Starica (Université de Neuchâtel)

Autorise l'impression de la présente thèse.

Neuchâtel, 2 juillet 2009

Le doyen

Kilian Stoffel

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

**To Petko and Maya**

# Acknowledgments

First, I would like to thank my Ph.D. supervisor Prof. Yves Tillé for his guidance, support and help during my Ph.D. studies. He showed me the vast researcher's world and taught me how to express and defend my ideas scientifically. I have benefited greatly from his knowledge and experience. Thank you, Yves.

In addition, I would like to thank the members of the thesis committee: the president of the jury, Prof. Catalin Starica, Jean-Claude Deville, Prof. Ray Chambers and Prof. Wayne Fuller for kindly accepting to be part of it.

I extend my sincere thanks to the Department of Statistics, University of Neuchâtel for providing me with an excellent working environment. I would like to acknowledge and thank as well the Swiss National Science Foundation which financially supported a big part of this research through the grant FN205121-105187/1, the Swiss Federal Statistical Office, and in particular, the Statistic Methodology group, for supporting a part of this research, and the Office of equal opportunities at the University of Neuchâtel which financially supported me at the final stage of my Ph.D. studies through the 'Subside Tremplin' grant.

I would also like to express my gratitude to my friends and colleagues, and in particular, Alina Matei, Lionel Qualité and Erika Antal, for always being ready to discuss with me and help me with the various problems I encountered, and for their encouragement.

Last, but not least, I would like to thank my husband and friend Petko and my sweet daughter Maya, for being the light of my life, and my family in Bulgaria, for their love and support.

# Abstract and key words

**Mots clés** Coordination d'échantillons dans le temps, Echantillonage répété, Nombres aléatoires permanents, Microstrates, Algorithme d'échantillonage, Echantillonage équilibré, Inférence basée sur le plan d'échantillonage, Inférence basée sur le modèle, Robustnesse au biais.

**Key words** Sample coordination, Repeated survey sampling, Permanent random numbers, Microstrata, Sampling algorithm, Balanced sampling, Design-based inference, Model-based inference, Optimal strategy, Bias-robustness.

**Evaluation et développement de stratégies pour la coordination d'échantillons et l'inférence statistique dans les enquêtes par sondages**

**Résumé** Cette thèse de doctorat se concentre sur deux sujets importants de la théorie des sondages. La première partie traite du problème du fondement de l'inférence statistique en populations finies. La seconde partie traite de la question de la coordination d'échantillons dans le temps. La thèse est basée sur quatre articles, dont trois ont été déjà publiés dans des revues internationales et le quatrième a été soumis pour publication.

Dans les premières chapitres de la thèse, on discute de l'optimalité de strategies composées d'un plan d'échantillonage et d'un estimateur. On démontre que la strategie qui consiste à utiliser l'échantillonage équilibré avec des probabilités proportionnelles aux erreurs du modèle linéaire, et l'estimateur de Horvitz-Thompson est optimale sous le plan et sous le modèle. En suite, on montre que cette strategie est toujours robuste et efficace dans le cas où le modèle s'avère faux en prenant un exemple sous le modèle polynomial.

Les dernières chapitres traitent un premier temps de la coordination d'échantillons stratifiés, des méthodes existante dont on compare la qualité de coordination et l'optimalité à l'aide d'une étude de simulation. On propose de nouvelles méthodes basées sur des microstrates et on teste, à nouveau par simulations, leur validité. Enfin, on a réalisé une étude plus fondamentale de l'échantillonage répété dans le temps. On y présente les plans longitudinaux les plus connus. On note qu'il y a un antagonisme entre une bonne coordination et le choix libre d'un plan transversal. On propose également une nouvelle méthode qui peut remédier à ce problème.

**Abstract** This Ph.D. thesis concentrates on two important subjects in survey sampling theory. One is the problem of the foundation for statistical inference in finite population sampling, and the other is the problem of coordination of samples over time. The thesis is based on four articles. Three of them are already published and the last one is submitted for publication.

First, we show that the model-based and design-based inferences can be reconciliated if we search for an optimal strategy rather than just an optimal estimator, a strategy being a pair composed of a sampling design and an estimator. If we accept the idea that balanced samples are randomly selected, e.g. by the cube method, then we show that, under the linear model, an optimal strategy consists of a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model and the Horvitz-Thompson estimator. Moreover, if the heteroscedasticity of the model is 'fully explainable' by the auxiliary variables, then the best linear unbiased estimator and the Horvitz-Thompson estimator are equal. We construct a single estimator for both the design and model variance. The inference can thus be valid under the sampling design and under the model. Finally, we show that this strategy is robust and efficient when the model is misspecified.

Coordination of probabilistic samples is a challenging theoretical problem faced by statistical institutes. One of their aims is to maximize or minimize the overlap between several samples drawn successively in a population that changes over time. In order to do that, a dependence between the samples must be introduced. Several methods for coordinating stratified samples have already been developed. Using simulations, we compare their optimality and quality of coordination. We present new methods based on Permanent Random Numbers (PRNs) and microstrata which have the advantage of allowing us to choose between positive or negative coordination with each of the previous samples. Simulations are run to test the validity of each of them. Another aim of sampling coordination is to obtain good estimates for each wave while spreading the response burden across the entire population. We review the existing solutions. We compute their corresponding longitudinal designs and discuss their properties. We note that there is an antagonism between a good rotation and control over the cross-sectional sampling design. In order to reach a compromise between the quality of coordination and the freedom of choice of the cross-sectional design, we propose an algorithm that uses a new method of longitudinal sampling.

# Contents

**1 Introduction** **1**

1.1 Finite population, Sample, and Sampling Design . . . . . 1

1.2 Statistical inference in finite population survey sampling . . . . . 2

1.3 The sample coordination problem . . . . . 4

1.4 Structure of the thesis . . . . . 7

**2 Optimality under the linear model** **9**

2.1 Introduction . . . . . 10

2.2 Notation and definitions . . . . . 11

2.3 Linear estimators . . . . . 13

2.4 Balanced sampling . . . . . 15

2.5 Model-assisted approach . . . . . 19

2.6 Model-based approach . . . . . 20

2.7 A combined model-based and model-assisted approach . . . . . 22

2.8 Estimation of variance . . . . . 23

2.9 Examples . . . . . 25

2.10 Discussion . . . . . 28

**3 Robustness and efficiency** **35**

3.1 Introduction . . . . . 35

3.2 Notation and definitions . . . . . 37

3.3 Model-based strategy and BLU estimator . . . . . 38

3.4 Balanced estimator under a linear model . . . . . 40

3.5 Bias robustness in submodels . . . . . 42

3.6 Application to the polynomial model . . . . . 44

3.6.1 Presentation of the model . . . 44
3.6.2 A first solution . . . 45
3.6.3 An alternative solution for the polynomial model . . . 47
3.6.4 A particular case: the ratio model . . . 48
3.7 Discussion . . . 49

**4 Coordination of stratified samples** **51**
4.1 Introduction . . . 52
4.2 Population, Sample, and Sampling Design . . . 53
4.3 Sample coordination, Overlap, and Burden . . . 55
4.4 The Kish & Scott Method . . . 57
4.5 The Cotton & Hesse Method . . . 57
4.6 Comparison of the Kish & Scott and Cotton & Hesse Methods . . . 59
4.7 The Rivière Method . . . 62
4.8 Other Methods Using Microstrata . . . 65
4.9 Simulation Study and Results . . . 69
4.10 Conclusions . . . 74

**5 Repeated survey sampling** **77**
5.1 Introduction . . . 78
5.2 Basic Concepts and Notation . . . 79
5.2.1 Sampling on one occasion . . . 79
5.2.2 Sampling on several occasions . . . 81
5.2.3 Average time out of the sample . . . 84
5.3 Classical sampling designs . . . 85
5.3.1 Poisson sampling design . . . 86
5.3.2 Simple random sampling . . . 87
5.3.3 Systematic sampling . . . 88
5.3.4 Deville's systematic sampling . . . 89
5.3.5 A new sampling algorithm for unequal probability sampling . . . 93
5.3.6 Remark on the variables $\phi_k^t$ . . . 94
5.4 Usual methods of coordination . . . 94
5.4.1 The systematic-Poisson (or Brewer) repeated design . . . 95
5.4.2 The systematic-simple repeated design . . . 96

5.4.3 Use of a measure of burden or the Deville's systematic-simple repeated design . . . . . 98
5.5 Other repeated sampling designs . . . . . 100
5.5.1 General method for the coordination of samples . . . . . 100
5.5.2 Application to a systematic longitudinal design . . . . . 101
5.6 Other solutions to the coordination problem . . . . . 102
5.6.1 The dilemma of sampling coordination . . . . . 102
5.6.2 The minimum time out of sample method . . . . . 103
5.7 Conclusions . . . . . 104

**6 Conclusion** **107**

**Bibliography** **109**

# List of Figures

4.1 First- and second-order inclusion probabilities for Algorithms 4, 5, 6. . . . 72
4.2 First- and second-order inclusion probabilities for Algorithms 7, 8, 9. . . . 73

# List of Tables

2.1 Optimal strategies in the model-assisted, model-based and combined model-based and model-assisted approaches . . . . . 29

4.1 Overlap between the possible samples at times 1 and 2. . . . . 60
4.2 Set of optimal solutions for $c \in [0, 1/4]$. . . . . 60
4.3 Negative Coordination with the Kish & Scott Method. . . . . 61
4.4 Negative Coordination with the Cotton & Hesse Method. . . . . 62
4.5 Definition of strata used for the simulations . . . . . 64
4.6 Wave number for each plot . . . . . 71
4.7 Expected overlaps. . . . . 74
4.8 Summary table. . . . . 75

# List of Algorithms


1 Positive Coordination using the Kish & Scott Method. . . . . 58
2 Negative Coordination using the Cotton & Hesse Method. . . . . 58
3 Negative coordination with The Rivière Method. . . . . 63
4 First Method of Chronological Permutations (Sim-false). . . . . 66
5 Second Method of Chronological Permutations (Sim-false). . . . . 67
6 Third Method of Chronological Permutations (Sim-false). . . . . 68
7 First Method of Retrospective Permutations (Sim-correct). . . . . 68
8 Second Method of Retrospective Permutations (Sim-correct). . . . . 69
9 Fourth Method of Chronological Permutations (Sim-correct). . . . . 69
10 General longitudinal sequential algorithm. . . . . 84
11 Poisson strictly sequential. . . . . 86
12 SRSWOR sequential. . . . . 87
13 Usual strictly sequential algorithm for systematic sampling. . . . . 89
14 Sequential systematic algorithm. . . . . 90
15 Deville's systematic sampling. . . . . 91
16 Deville's Systematic Sequential. . . . . 92
17 Minimum time out of sample. . . . . 93
18 Coordination of Poisson samples in the case of a static population. . . . . 95
19 Coordination of SRSWOR using random numbers in the case of a static population. . . . . 97
20 Coordination of SRSWOR using a measure of burden in the case of a static population. . . . . 99
21 Minimum time out of sample sequential algorithm, without conditions on the inclusion probabilities. . . . . 105

# Chapter 1

# Introduction

## 1.1 Finite population, Sample, and Sampling Design

The term finite population survey sampling consists of the three basic elements: survey, finite population, and sampling, which are genuinely interrelated to describe the statistical methods used for the collection of data from a finite population, selecting a part of this population, observing the selected part with respect to some characteristic of interest and then making inference about the whole population. Particularly, a finite population $U$ is a set of $N$ units. Each unit can be identified by a unique label. Let $\{1, \ldots, k, \ldots, N\}$ denote the set of these labels. The size of the population, $N$, is not necessarily known. A sample without replacement is a subset of $U$ and in vector notation is presented as

$$\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)' \in \{0, 1\}^N,$$

where

$$s_k = \begin{cases} 1 & \text{if unit } k \text{ is in the sample} \\ 0 & \text{if unit } k \text{ is not in the sample,} \end{cases}$$

for all $k \in U$. The sample size is $n(\mathbf{s}) = \sum_{k \in U} s_k$.

A support $\mathcal{Q}$ is a set of samples. In probability sampling (see, for instance, Särndal et al., 1992) the selection of the sample is based on a random procedure on $Q$. A sampling design $p(\mathbf{s})$ is a probability distribution on the samples of $U$. Let $\mathbf{S}$ be the random sample,

i.e. the random vector of $\mathbb{R}^N$, whose distribution is given by

$$\Pr(\mathbf{S} = \mathbf{s}) = p(\mathbf{s}), \mathbf{s} \in \mathcal{Q}.$$

The support of a sampling design $p(\cdot)$ is defined by:

$$p(\mathbf{s}) > 0, \text{ for all } \mathbf{s} \in \mathcal{Q},$$

and

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

Given a sampling design, the probability that unit $k$ is in the sample, or the first-order inclusion probability, is denoted by $\pi_k$ and $\boldsymbol{\pi} = (\pi_k)_{1 \leq k \leq N}$ is the inclusion probability vector which van be derived from the sampling design as follows:

$$\boldsymbol{\pi} = \sum_{\mathbf{s} \subset U} \mathbf{s} p(\mathbf{s}).$$

The second-order, or joint, inclusion probability $\pi_{k\ell}$ is the probability of selecting units $k$ and $\ell$ together in the sample, and $\pi_{kk} = \pi_k$. The matrix of joint inclusion probabilities is given by

$$\boldsymbol{\Pi} = \sum_{\mathbf{s} \subset U} \mathbf{s}\mathbf{s}' p(\mathbf{s}).$$

## 1.2 Statistical inference in finite population survey sampling

In survey sampling, the foundational aspects of inference have been a main topic of interest in the last 40 years. The statistical inference in finite population sampling can be design-based, model-assisted or model-based. In the first two approaches, the inference is based on the stochastic structure induced by the sampling design. In the model-based approach, however, the inference is based on the probability structure of an assumed statistical model, often called a superpopulation model.

Design-based inference is the standard mode of inference in finite population sampling and is described in many papers (see, for instance, Hansen et al., 1993a,b; Kish, 1965; Cochran, 1977). In design-based inference, the population units have fixed but unknown values $y_k, k \in U$. The variable of interest, $Y$, which is a function of $y_k$ can be a total, a mean

or a more complex function. The aim is to obtain an estimator of $Y$ and a variance estimator which are unbiased or approximately unbiased in expectation over the distribution of all possible samples that could be selected with the given sampling design. A well-known design-based estimator is the Horvitz-Thompson estimator, or the $\pi$-estimator, developed by Narain (1951), and Horvitz and Thompson (1952). The observations are weighted by the inverse of the inclusion probabilities, which are also called design weights. The Horvitz-Thompson estimator of the population total $Y = \sum_{k \in U} y_k$ is given by:

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

This estimator is design-unbiased, for any sampling design with $\pi_k > 0, k \in U$, since $\mathrm{E}_p(\widehat{Y}) = Y$, where $\mathrm{E}_p$ denotes the expectation taken with respect to the sampling design $p(\cdot)$. The design variance is given by $\mathrm{var}_p(\widehat{Y}) = \mathrm{E}_p \left\{ \widehat{Y} - \mathrm{E}_p(\widehat{Y}) \right\}^2$. Design-unbiased variance estimators have been derived by Horvitz and Thompson (1952), and Sen (1953) and Yates and Grundy (1953) for a fixed sample size design.

Auxiliary variables are often available in the sampling frame and can be used to increase the sampling efficiency. In some cases, the values of an auxiliary variable, denoted $x$, are known for all the units of the population, in other cases we know just the population total $X = \sum_U x_k$. An auxiliary variable can be used to create the sample design in order to increase the precision of the $\pi$-estimator, e.g. in probability proportional to size (pps) sampling where the inclusion probabilities satisfy $\pi_k \propto x_k$, where $x_k,\ k \in U$ are known, positive values. The available auxiliary information can also be used at the estimation stage, i.e. the auxiliary variables enter directly into the estimator formula, under the assumption that they covary with the study variable. This is the case in the model-based and model-assisted frameworks.

The model-based, or prediction, approach uses a superpopulation model to describe the finite population, i.e. the values $y_1, \ldots, y_N$ are assumed to be the realization of a superpopulation model $\xi$ defined as:

$$\begin{cases} y_k = x_k'\beta + \varepsilon_k \text{ with} \\ \mathrm{E}_\xi(\varepsilon_k) = 0, \\ \mathrm{var}_\xi(\varepsilon_k) = \nu_k^2 \sigma^2, \\ \mathrm{cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0, \end{cases}$$

where $k \neq \ell \in U$, the $x_k$ are not random and known, and the quantities $\nu_k$ are assumed known. We can write the population total $Y = \sum_{k \in U} y_k = \sum_{k \in S} y_k + \sum_{k \in \bar{S}} y_k$, where $\bar{S}$ denotes the set of units in the population which are not in the sample $S$. As the sample total is known, the problem of estimating $Y$ comes to predicting the sum of the non-sampled population units. The properties of the estimator are derived with respect to the model $\xi$. An estimator $\widehat{Y}$ is said to be model-unbiased if $\mathrm{E}_\xi(\widehat{Y} - Y) = 0$, where $\mathrm{E}_\xi$ denotes the expectation taken with respect to the model. Hence, the model mean squared error, also called the error variance, is $\mathrm{E}_\xi(\widehat{Y} - Y)^2 = \mathrm{var}_\xi(\widehat{Y} - Y) + \mathrm{E}^2_\xi(\widehat{Y} - Y)$. In the model-based approach, the challenge is how to specify the model correctly. The major weakness is that in case of misspecification of the model, the prediction could be unreliable. For a comprehensive review of the model-based approach, see Valliant et al. (2000).

Models are used also within the design-based framework, but in a model-assisted way according to the terminology of Särndal et al. (1992). The role of the model $\xi$ is however slightly different than in the model-based framework. The model describes the finite population, i.e. we suppose that the finite population looks as if it were generated by the model $\xi$, but no such assumption is made. Thus, the inference about the finite population parameters is independent of model assumptions, i.e. the basic properties of an estimator are independent of where the model holds or not.

The debate between the design-based and model based proponents goes back to Royall (1970). It continues in Smith (1976, 1994); Hansen et al. (1983); Kish (1995). Kalton (2002) presents a good review of the use of models in survey sampling. A good summary of the problem is given in Little (2004): 'Many survey statisticians adopt both design and model-based philosophies of statistical analysis, according to the context. For example, descriptive inference about finite population quantities based on large probability samples are carried out using design-based methods, but models are used for problems where this approach does not work, such as non-response or small area estimation. This pragmatic approach has increased in popularity since battles over the 'foundations of survey inference' in the 1980's subsided.'

## 1.3 The sample coordination problem

Another major issue in survey sampling is the problem of coordination of samples when dealing with repeated sampling designs. It is a commonly faced problem by the national

statistical agencies, i.e. in official statistics. Populations which can be households, businesses or other entities are sampled on two or more occasions. We distinguish two main types of coordination. In negative coordination, the aim is to minimize the overlap (the number of common units) between several samples drawn on consecutive occasions, while in positive coordination we want to maximize this number. This can be achieved by creating a dependance between the samples.

When sampling over time, we may be interested in two diametrically different aspects of the repeated sampling problem. One is the cross-sectional aspect, when we want to estimate some characteristic of the population on each occasion. The other is the longitudinal aspect, when we want to measure the changes being differences between, or ratios of, the corresponding estimates for the different time periods.

Some samples may be designed to retain the same sample units in all consecutive surveys. Subsequently, they may be subject to changes over time due to births (addition of new units to the population) or deaths (loss of units from the population). In this case we refer to the population as a dynamic population. Sometimes it is not possible to subject the same sample units to repeated surveys for a long time. In this case, we fix a proportion of the sample which will be replaced after a given number of surveys. This predetermined proportion is called the rotation rate, and the whole process, sample rotation.

The first papers on coordination were written by Patterson (1950) and Keyfitz (1951). These first works present methods which are in general restricted to two successive samples or small sample sizes. At a later stage, Kish and Scott (1971) generalized the coordination problem in the context of a larger sample size.

An important concept in coordination based on permanent random numbers (PRNs) was introduced by Brewer et al. (1972). Most of the national bureaus of statistics use variations of methods based on PRN sampling. Ohlsson (1995) gives an overview of PRN methods with implementation in different countries.

Let $U^t$ denote the population at time $t$, for $t = 1, 2, \ldots, T-1, T$. At time $t$, a sample without replacement is a subset of the population $U^t$. The sample is denoted by a vector

$$\mathbf{s}^t = (s_1^t, \ldots, s_k^t, \ldots, s_N^t)' \in \{0,1\}^N,$$

where

$$s_k^t = \begin{cases} 1 & \text{if, at time } t, \text{unit } k \text{ is in the sample} \\ 0 & \text{if, at time } t, \text{unit } k \text{ is not in the sample,} \end{cases}$$

for all $k \in U$.

At time $t$, the first-order inclusion probability and the joint inclusion probability are denoted, respectively, by $\pi_k^t$ and $\pi_{k\ell}^t$, where $k, \ell \in U^t, t = 1, \ldots, T$. The longitudinal inclusion probability, for times $t$ and $u$, is denoted by $\pi_k^{tu},\ k \in U^t \cap U^u, t, u = 1, \ldots, T$. Due to the Fréchet bounds, we have, for times $t$ and $u$:

$$\max(0, \pi_k^t + \pi_k^u - 1) \leq \pi_k^{tu} \leq \min(\pi_k^t, \pi_k^u).$$

If, at times 1 and 2, two samples are drawn independently without coordination, then, for all $k \in U, \pi_k^1 \pi_k^2 = \pi_k^{12}$. In positive coordination, for all $k \in U$, the longitudinal inclusion probability must satisfy the conditions

$$\pi_k^1 \pi_k^2 \leq \pi_k^{12} \leq \min(\pi_k^1, \pi_k^2).$$

In negative coordination, for all $k \in U$, the longitudinal inclusion probability must satisfy the conditions

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{12} \leq \pi_k^1 \pi_k^2.$$

In the last case, the longitudinal inclusion probability can be zero only if $\pi_k^1 + \pi_k^2 \leq 1$.

The response burden of a survey is usually quantified in terms of the time needed to complete the questionnaire. However, other aspects of response burden exist: for example, how difficult it is to provide the information or how sensitive the question sent to the respondent is. Therefore, the response burden can vary from one survey to another.

Consider a population $U$ split into $H$ parts $U_h$, called 'strata', such that

$$\cup_{h=1}^H U_h = U \quad \text{and} \quad U_h \cap U_i = \emptyset,$$

for all $(h, i)$ with $h \neq i$. A design is called stratified if a random sample $S_h$ of fixed size $n_h$ is selected in each stratum $U_h$, and if the sample selection in each stratum is taken independently of the selection done in all the other strata.

The coordination of stratified samples is a more complex problem. The main reason is that, over time, units usually change from one stratum to another. Several methods, the Kish & Scott method presented in Kish and Scott (1971), the Cotton & Hesse method presented in Cotton and Hesse (1992b), the Dutch method (EDS) described in De Ree (1983), Van Huis et al. (1994a,b), Koeijers and Willeboordse (1995), and the Rivière method pre-

sented in Rivière (1998, 1999, 2001a,b), have already been developed in order to obtain maximal or minimal coverage between samples drawn on different occasions.

The quality of a coordination procedure can be measured using four possible criteria:

1. the procedure provides a controllable degree of overlap;
2. the sampling design is respected in each selection;
3. for each unit, a fixed time out of sample is respected;
4. the procedure is computed easily.

## 1.4 Structure of the thesis

Each chapter of the thesis is self-contained[1]. Chapter 2 considers the model-based and model-assisted approaches. These two paradigms are shown to be similar if one searches for an optimal strategy rather than just an optimal estimator, a strategy being a pair composed of a sampling design and an estimator. It is shown that, under a linear model, the optimal model-assisted strategy consists of a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model and the Horvitz-Thompson estimator. If the heteroscedasticity of the model is 'fully explainable' by the auxiliary variables, then this strategy is also optimal in a model-based sense. This optimal strategy is a sufficient condition so that the best linear unbiased estimator and the Horvitz-Thompson estimator are equal. A single estimator for both the design and model variances is constructed.

Chapter 3 investigates the problem of bias robustness and efficiency in the model-based inference. If the idea that a balanced sample can be randomly selected is accepted, then a balanced sampling design with the Horvitz-Thompson estimator compose a strategy that is always robust. An extension to the polynomial model is given.

Chapter 4 presents several methods for coordinating stratified samples, such as the Kish & Scott method, the Cotton & Hesse method, and the Rivière method. Using simulations, the optimality of these methods and their quality of coordination are compared. Six new methods based on Permanent Random Numbers (PRNs) and microstrata are presented. These new methods have the advantage of allowing one to choose between positive or

[1]Each chapter is published, or is submitted for publication in a refereed international journal.

negative coordination with each of the previous samples. Simulations are run to test the validity of each of them.

In Chapter 5 are presented some classical sampling designs in the light of longitudinal sampling. For each design, a sequential or a strictly sequential algorithm is given. A new sampling algorithm for unequal probability sampling is also given. Next, some usual methods of negative coordination, i.e. the systematic-Poisson, the systematic-simple and the Deville's systematic-simple repeated designs are presented. A general method for the coordination of samples is given. It is also noted that there is an antagonism between a good rotation and control over the cross-sectional sampling design. In order to reach a compromise between the quality of the sample coordination, which appears to be optimal for a systematic longitudinal sampling design, and the freedom of choice of the cross-sectional design, an algorithm that uses a new method of longitudinal sampling is proposed.

# Chapter 2

# Optimal sampling and estimation strategies under the linear model

**Abstract:**

In some cases, model-based and model-assisted inferences can lead to very different estimators. These two paradigms are not so different if we search for an optimal strategy rather than just an optimal estimator, a strategy being a pair composed of a sampling design and an estimator. We show that, under a linear model, the optimal model-assisted strategy consists of a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model and the Horvitz-Thompson estimator. If the heteroscedasticity of the model is 'fully explainable' by the auxiliary variables, then this strategy is also optimal in a model-based sense. Moreover, under balanced sampling and with inclusion probabilities that are proportional to the standard deviation of the model, the best linear unbiased estimator and the Horvitz-Thompson estimator are equal. Finally, it is possible to construct a single estimator for both the design and model variance. The inference can thus be valid under the sampling design and under the model.

[1]This chapter is a reprint of the paper: D. Nedyalkova and Y. Tillé. Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95:521-537, 2008.

## 2.1 Introduction

In survey sampling theory there have long been contrasting views on which approach to use in order to obtain a valid inference in estimating population totals: a prediction theory based on a superpopulation model or a probability sampling theory based on a sampling design. Neither of these paradigms is false. Numerous articles compare the two approaches (Brewer, 1994, 1999b, 2002; Brewer et al., 1988; Hansen et al., 1983; Iachan, 1984; Royall, 1988; Smith, 1976, 1984, 1994). Valliant et al. (2000, p. 14), who favour the model-based theory, say that 'there is no doubt of the mathematical validity of either of the two theories'. Nevertheless, we believe that the choice between them depends on the point of view of the analyst.

In the model-based, or prediction, approach studied by Royall (1976, 1992); Royall and Cumberland (1981) and Chambers (1996), the optimality is conceived only with respect to the regression model without taking into account the sampling design. Royall (1976) proposed the use of the best linear unbiased predictor when the data are assumed to follow a linear model. Royall (1992) showed that under certain conditions there exists a lower bound for the error variance of the best linear unbiased predictor and that this bound is only achieved when the sample is balanced. More specifically, Royall and Herson (1973a,b) and Scott et al. (1978) discussed the importance of balanced sampling in order to protect the inference against a misspecified model. These authors have come to the conclusion that the sample must be balanced, but not necessarily random.

In the model-assisted approach advocated by Särndal et al. (1992), the estimator must be approximately design-unbiased under the sampling design. The generalized regression estimator uses auxiliary information coming from the linear model, but is approximately design-unbiased. Deville and Särndal (1992) proposed a purely design-based methodology which takes into account auxiliary information without considering a model. The main difference between the design-based and the model-based approaches arises because the statistical properties of an estimator are evaluated with respect to the sampling design and not with respect to the model.

Recently, Deville and Tillé (2004) developed the cube method, an algorithm that can select randomly balanced samples and satisfies exactly the given inclusion probabilities. In the model-based framework, balanced samples are essential for achieving the lower bound for the error variance proposed by Royall (1992). Moreover, it can be shown that balanced sampling is also optimal under model-assisted inference. Hájek (1981) defined a strategy

as a pair comprising a sampling design and an estimator. The purpose of this paper is to show that, if we search for an optimal strategy rather than just an optimal estimator, most of the differences between model-based and model-assisted inferences can be reconciled.

## 2.2 Notation and definitions

We consider a finite population $U$ of size $N$. Each unit of the population can be identified by a label $k = 1, \ldots, N$. Let $x_k = (x_{k1}, \ldots, x_{kq})'$ be the vector of the values of $q$ auxiliary variables for unit $k$, for $k = 1, \ldots, N$ and let

$$X = \sum_{k \in U} x_k$$

be the vector of totals which is also known. The values $y_1, \ldots, y_N$ of the variables of interest are unknown. The aim is to estimate the population total

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

A sample $s$ is a subset of the population $U$. Let $p(s)$ denote the probability of selecting the sample $s$, $S$ being the random sample such that $p(s) = \mathrm{pr}\,(S = s)$, and let $n(S)$ be the size of the sample $S$. The expected sample size is $n = \mathrm{E}_p\{n(S)\}$, where $\mathrm{E}_p$ denotes the expected value under the sampling design $p(\cdot)$. Let $\bar{S}$ denote the set of units of the population which are not in $S$. Let $\pi_k = \mathrm{pr}(k \in S)$ denote the inclusion probability of unit $k$, and let $\pi_{k\ell} = \mathrm{pr}(k \in S \text{ and } \ell \in S)$ denote, for $k \neq \ell$ the joint inclusion probability of units $k$ and $\ell$. The variable $y$ is observed on the sample only.

Under model-based inference, the values $y_1, \ldots, y_N$ are assumed to be the realization of a superpopulation model $\xi$. The model which we will study is the general linear model with uncorrelated errors given by

$$y_k = x_k'\beta + \varepsilon_k, \tag{2.1}$$

where the $x_k$'s are not random, $\beta = (\beta_1, \ldots, \beta_q)', \mathrm{E}_\xi(\varepsilon_k) = 0, \mathrm{var}_\xi(\varepsilon_k) = \nu_k^2\sigma^2$, for all $k \in U$, and $\mathrm{cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$, when $k \neq \ell \in U$. The quantities $\nu_k, k \in U$, are assumed

known. Moreover, we scale them so that

$$\sum_{k\in U} \nu_k = N.$$

The superpopulation model (2.1) includes the possibility of heteroscedasticity. Under homoscedasticity, $\nu_k = 1$ for all $k \in U$. An important and common hypothesis is that the random sample $S$ and the errors $\varepsilon_k$ of the model are independent. The symbols $\mathrm{E}_\xi, \mathrm{var}_\xi$ and $\mathrm{cov}_\xi$ denote, respectively, expected value, variance and covariance under the model.

In order to estimate the total $Y$, we will only use linear estimators which can be written as

$$\widehat{Y}_w = \sum_{k\in S} w_{kS} y_k = \sum_{k\in U} w_{kS} y_k I_k,$$

where the $w_{kS}, k \in S$, are weights that can depend on the sample, and where $I_k$ is equal to 1 if $k \in S$ and equal to 0 if $k \notin S$.

Definition 2.1. (Hájek, 1981, p. 153) *A strategy is a pair $(p(\cdot), \widehat{Y})$ comprising a sampling design and an estimator.*

Definition 2.2. *An estimator $\widehat{Y}$ is said to be model-unbiased if* $\mathrm{E}_\xi(\widehat{Y} - Y) = 0$.

Definition 2.3. *An estimator $\widehat{Y}$ is said to be design-unbiased if* $\mathrm{E}_p(\widehat{Y}) - Y = 0$.

Definition 2.4. *A linear estimator $\widehat{Y}_w$ is said to be calibrated on a set of auxiliary variables $x_k$ if and only if its weights satisfy*

$$\sum_{k\in S} w_{kS} x_k = \sum_{k\in U} x_k.$$

Definition 2.5. *The design variance of an estimator $\widehat{Y}$ is defined by*

$$\mathrm{var}_p(\widehat{Y}) = \mathrm{E}_p\left\{\widehat{Y} - \mathrm{E}_p(\widehat{Y})\right\}^2.$$

Definition 2.6. *The design mean-squared error of an estimator $\widehat{Y}$ is defined by*

$$\mathrm{MSE}_p(\widehat{Y}) = \mathrm{E}_p\left(\widehat{Y} - Y\right)^2.$$

DEFINITION 2.7. *The model variance of an estimator $\widehat{Y}$ is defined by*

$$\mathrm{var}_\xi(\widehat{Y}) = \mathrm{E}_\xi \left\{ \widehat{Y} - \mathrm{E}_\xi(\widehat{Y}) \right\}^2 .$$

DEFINITION 2.8. *The model mean-squared error of an estimator $\widehat{Y}$ is defined by*

$$\mathrm{E}_\xi \left( \widehat{Y} - Y \right)^2 .$$

The model mean-squared error is sometimes called the error variance. The model mean-squared error of an estimator $\widehat{Y}$ is generally smaller than its model variance due to the fact that $\widehat{Y}$ is closer to $Y$ than to $\mathrm{E}_\xi(\widehat{Y})$.

DEFINITION 2.9. *The anticipated mean square error of an estimator $\widehat{Y}$ is defined by*

$$\mathrm{MSE}_{p\xi}(\widehat{Y}) = \mathrm{E}_p \mathrm{E}_\xi \left( \widehat{Y} - Y \right)^2 = \mathrm{E}_\xi \mathrm{E}_p \left( \widehat{Y} - Y \right)^2 .$$

The anticipated mean-squared error is also called the anticipated variance, for example by Isaki and Fuller (1982).

## 2.3 Linear estimators

Consider the class of linear estimators

$$\widehat{Y}_w = \sum_{k \in S} w_{kS} y_k .$$

For all $k \in U$, define $C_k = \mathrm{E}_p(w_{kS} I_k) = \pi_k \mathrm{E}_p(w_{kS} | I_k = 1)$. Godambe (1955) showed that $\widehat{Y}_w$ is design-unbiased if and only if $C_k = 1$ or, equivalently, if $\mathrm{E}_p(w_{kS} | I_k = 1) = 1/\pi_k$. Moreover, its model bias is

$$\mathrm{E}_\xi(\widehat{Y}_w - Y) = \sum_{k \in S} w_{kS} x_k' \beta - \sum_{k \in U} x_k' \beta ,$$

for any value of $\beta \in \mathbb{R}^q$. Therefore, for the class of linear estimators under the linear model $\xi$, the definitions of a model-unbiased and a calibrated estimator are equivalent. For any linear estimator, a general expression of the anticipated mean-squared error can be given.

RESULT 2.1. *If* $\widehat{Y}_w$ *is a linear estimator, then*

$$\begin{aligned}
&\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 \\
&= \sigma^2\mathrm{E}_p\left\{\sum_{k\in S}(w_{kS}-1)^2\nu_k^2 + \sum_{k\in\bar{S}}\nu_k^2\right\} + \mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U}x_k'\beta\right)^2 \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{C_k^2\frac{1-\pi_k}{\pi_k} + \pi_k\mathrm{var}_p(w_{kS}|I_k=1) + (C_k-1)^2\right\} \\
&\quad +\mathrm{var}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + \left(\sum_{k\in U}C_kx_k'\beta - \sum_{k\in U}x_k'\beta\right)^2.
\end{aligned}$$

The proof is given in the Appendix.

The anticipated mean-squared error $\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2$ is the sum of five nonnegative terms,

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = A + B + C + D + E, \tag{2.2}$$

where

$$\begin{aligned}
A &= \sigma^2\sum_{k\in U}\nu_k^2C_k^2\frac{1-\pi_k}{\pi_k}, \quad B = \sigma^2\sum_{k\in U}\nu_k^2\pi_k\mathrm{var}_p(w_{kS}|I_k=1), \quad C = \sigma^2\sum_{k\in U}\nu_k^2(C_k-1)^2, \\
D &= \mathrm{var}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right), \quad \text{and } E = \left(\sum_{k\in U}C_kx_k'\beta - \sum_{k\in U}x_k'\beta\right)^2.
\end{aligned}$$

Term $A$ is a part of the anticipated mean-squared error; it depends on the inclusion probabilities and the variance of the errors. Term $B$ is only relevant if the weights $w_{kS}$ differ from sample to sample. Term $C$ depends on the design bias and the variance of the errors of the model; it is null if the estimator is design-unbiased. Term $D$ is the design variance of the model expectation of the estimator; it is null when the estimator is calibrated, or model-unbiased. Term $E$ is the square of the design bias of the model expectation of the estimator; it is also null when the estimator is calibrated, or model-unbiased, or when the estimator is design-unbiased.

Some particular cases of Result 2.1 are interesting.

COROLLARY 2.1. *If* $\widehat{Y}_w$ *is a model-unbiased linear estimator, or a calibrated estimator, then* $\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = A + B + C$.

COROLLARY 2.2. *If* $\widehat{Y}_w$ *is a design-unbiased linear estimator, then* $C_k = 1$ *for all* $k$ *in* $U$

*and* $\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = A + B + D$.

COROLLARY 2.3. *If* $\widehat{Y}_w$ *is a design-unbiased linear estimator with weights* $w_{ks}$ *that are constant from sample to sample, then* $C_k = 1$*, for all* $k$ *in* $U$*, and* $\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = A + D$.

COROLLARY 2.4. *If* $\widehat{Y}_w$ *is a design-unbiased and model-unbiased linear estimator, then* $\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = A + B$.

*Example* 2.1. The Horvitz-Thompson estimator, given by

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k},$$

is linear and design-unbiased when $\pi_k > 0$, for all $k \in U$, because

$$\mathrm{E}_p(\widehat{Y}_\pi) = \sum_{k \in U} \frac{y_k}{\pi_k} \mathrm{E}(I_k) = Y.$$

Under any sampling design, the design variance of this estimator is

$$\mathrm{var}_p(\widehat{Y}_\pi) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k}{\pi_k} \Delta_{k\ell} \frac{y_\ell}{\pi_\ell}, \tag{2.3}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell,\ k, \ell \in U$. The Horvitz-Thompson estimator is, however, model-biased and its bias is

$$\mathrm{E}_\xi(\widehat{Y}_\pi - Y) = \left( \sum_{k \in S} \frac{x'_k}{\pi_k} - \sum_{k \in U} x'_k \right) \beta. \tag{2.4}$$

Since the Horvitz-Thompson estimator is design-unbiased with weights $w_{ks} = 1/\pi_k$ that are constant from sample to sample, its anticipated mean-squared error can be deduced from Corollary 2.3:

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = A + D = \sigma^2 \sum_{k \in U} \nu_k^2 \frac{1 - \pi_k}{\pi_k} + \sum_{k \in U} \sum_{\ell \in U} \frac{x'_k \beta}{\pi_k} \Delta_{k\ell} \frac{x'_\ell \beta}{\pi_\ell}.$$

## 2.4 Balanced sampling

There exist several different definitions of the concept of balancing. A first definition of a balanced sample is that the sample mean is equal to the population mean. According to this definition, balancing is a property of a sample and a balanced sample can be constructed deliberately and deterministically without reference to a random procedure. A balanced

sample is then associated with the purposive selection and is thus in contradiction to the random selection of the sample (Brewer, 1999b).

A balanced sample can also be selected randomly by a procedure called a balanced sampling design. According to the definition of Deville and Tillé (2004), a sampling design $p(\cdot)$ is said to be balanced on the auxiliary variables $x_1, \ldots, x_q$ if the Horvitz-Thompson estimator satisfies the relationship

$$\widehat{X}_\pi = \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k = X. \tag{2.5}$$

Authors such as Cumberland and Royall (1981) and Kott (1986) would call this a '$\pi$-balanced sampling', opposed to a mean-balanced sampling defined by the equation

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

Below, we use the expression 'balanced sampling' to denote a sampling design that satisfies equation (2.5) for one or more auxiliary variables, a mean-balanced sampling being a particular case of this balanced sampling when the sample is selected with inclusion probabilities $n/N$.

The definition of balanced sampling includes the definition of sampling with fixed sample size. Suppose that one of the balancing variables is proportional to the inclusion probabilities or, more generally, that there exists a vector $\lambda$ such that $\lambda' x_k = \pi_k$, for all $k \in U$. In this case, the balancing equation

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

becomes for this variable, by multiplication by $\lambda'$,

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

or equivalently

$$\sum_{k \in S} 1 = \sum_{k \in U} \pi_k,$$

which means that the sample size must be fixed. In practice, it is always recommended to add the vector of inclusion probabilities in the balancing variables, because this allows one

to fix the sample size and thus the cost of the survey.

If a sampling design is balanced on the auxiliary variables, then $\widehat{X}_\pi$ is not a random variable. For a long time, balanced samples were considered difficult to construct, except for particular special cases such as sampling with fixed sample size or stratification. Partial procedures of balanced sampling have been proposed by Yates (1946), Thionet (1953), Deville et al. (1988), Ardilly (1991), Deville (1992), and Hedayat and Majumdar (1995), and a list of methods for constructing balanced samples is given in Valliant et al. (2000, pp. 65-78). Several of these methods are rejective: they consist of generating randomly a sequence of samples with an original sampling design until a sample is obtained that is sufficiently well balanced. Rejective methods are actually a way of constructing a conditional sampling design and have the important drawback that the inclusion probabilities of the balanced design are not necessarily the same as the inclusion probabilities of the original design. Moreover, if the number of balancing variables is large, the rejective methods can be very slow.

The cube method, proposed by Deville and Tillé (2004), is a non-rejective procedure that directly allows the random selection of balanced or nearly balanced samples and that satisfies exactly the given first-order inclusion probabilities. The cube method works with equal or unequal inclusion probabilities; see also Tillé (2006, pp. 147-76). If one of the balancing variables is proportional to the inclusion probabilities, then the cube method will produce samples of fixed size. However, it is not always possible for such a sample to be exactly balanced because of the rounding problem. For instance, in proportional stratification, which is a particular case of balanced sampling, it is generally impossible to select an exactly balanced sample because the sample sizes of the strata, $n_h = nN_h/N$, are seldom integers. Deville and Tillé (2004) also showed that the rounding problem, under reasonable hypotheses, is bounded by $O(q/n)$, where $q$ is the number of balancing variables and $n$ is the sample size. Thus, the rounding problem becomes negligible if the sample size is reasonably large with respect to the number of balancing variables.

Under model (2.1) and balanced sampling, the Horvitz-Thompson estimator is model-unbiased. Indeed, by equations (2.4) and (2.5), it follows that

$$\mathrm{E}_\xi(\widehat{Y}_\pi - Y) = \left(\sum_{k\in S}\frac{x_k}{\pi_k} - \sum_{k\in U} x_k\right)' \beta = 0.$$

Under model (2.1) and balanced sampling, we can compute the error variance and the

anticipated mean-squared error of the Horvitz-Thompson estimator.

RESULT 2.2. *Under model* (2.1)*, if the sample is balanced on* $x_k$ *and selected with inclusion probabilities* $\pi_k$*, then*

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k}{\pi_k}. \tag{2.6}$$

The proof is given in the Appendix.

If we fix the inclusion probabilities, then the expectation of the sample size is also fixed. The design mean-squared error of a balanced sampling design is, unfortunately, more difficult to determine. In their Method 4, Deville and Tillé (2005) have proposed the following approximation of the design variance given in (2.3):

$$\mathrm{var}_p(\widehat{Y}_\pi) \doteq \mathrm{var}_{\mathrm{app}}(\widehat{Y}_\pi) = \sum_{k\in U} d_k \frac{(y_k - x_k' b)^2}{\pi_k^2}, \tag{2.7}$$

where

$$b = \left(\sum_{k\in U} d_k \frac{x_k x_k'}{\pi_k^2}\right)^{-1} \sum_{k\in U} d_k \frac{x_k y_k}{\pi_k^2},$$

and the $d_k$ are the solution of the nonlinear system

$$\pi_k(1-\pi_k) = d_k - \frac{d_k x_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1} \frac{d_k x_k}{\pi_k}, k \in U. \tag{2.8}$$

This approximation, which uses only the first-order inclusion probabilities, was validated by Deville and Tillé (2005) under a variety of balanced samples regardless of how the $y$-values were generated. An additional argument in favour of using this approximation is that its model expectation is equal to its anticipated mean-squared error, as we see below.

RESULT 2.3. *Under model* (2.1)*, if the sample is balanced on* $x_k$*, then*

$$\mathrm{E}_\xi\left\{\mathrm{var}_{\mathrm{app}}(\widehat{Y}_\pi)\right\} = \mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2.$$

The proof is given in the Appendix.

## 2.5 Model-assisted approach

One approach to estimating $Y$ consists in finding the 'best' strategy that provides a valid inference under the sampling design. Godambe (1955) showed that there is no optimal estimator in the class of linear estimators for all $y_1, \ldots, y_N$ that minimises the design mean-squared error. It is, however, not possible to determine an optimal design-based strategy without formalizing the link between the auxiliary variables $x_k$ and the variables of interest $y_k$. A model must therefore be used to guide the choice of the estimator. Särndal et al. (1992) proposed the concept of 'model-assisted inference'. To be model-assisted, the estimator must be chosen so that it leads to a valid inference with respect to the sampling design, even if the model is misspecified. In order to make the inference, we need to estimate $\mathrm{E}_p(\widehat{Y}_w - Y)^2$, but in order to find the optimal strategy, we need to minimize $\mathrm{E}_\xi \mathrm{E}_p(\widehat{Y}_w - Y)^2$ under the constraint that the estimator is design-unbiased or that its design bias is small with respect to its design mean-squared error.

A bound for the model-assisted strategy was given by Godambe and Joshi (1965) for a set of fixed inclusion probabilities. The Godambe-Joshi bound can be derived directly from Corollary 2.2. If $\widehat{Y}_w$ is a design-unbiased linear estimator, then

$$\mathrm{E}_p \mathrm{E}_\xi (\widehat{Y}_w - Y)^2 \geq L_p = \sigma^2 \sum_{k \in U} \nu_k^2 \frac{1 - \pi_k}{\pi_k}. \tag{2.9}$$

If we suppose at least tentatively that the $\nu_k$ are known, a judicious choice of the inclusion probabilities allows a smaller anticipated mean-squared error to be determined. If we minimize $L_p$ in $\pi_k$ subject to

$$\sum_{k \in U} \pi_k = n, \quad 0 \leq \pi_k \leq 1, \tag{2.10}$$

for all $k$ in $U$, then we obtain the optimal inclusion probabilities given by

$$\pi_k^* = \min(1, \alpha \nu_k / N),$$

where $\alpha$ is such that

$$\sum_{k \in U} \min\left(1, \frac{\alpha \nu_k}{N}\right) = n.$$

The following general result gives a bound for any design-unbiased strategy with a sample

size $n$.

RESULT 2.4. *For any design-unbiased strategy,*

$$\begin{aligned}
\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 &\geq L_p = \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k}{\pi_k} \\
&\geq \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k^*}{\pi_k^*} = \sigma^2 \left( \frac{N}{\alpha} \sum_{\substack{k\in U \\ \pi_k^*<1}} \nu_k - \sum_{\substack{k\in U \\ \pi_k^*<1}} \nu_k^2 \right) \\
&\geq \sigma^2 \left( \frac{N^2}{n} - \sum_{k\in U} \nu_k^2 \right) = \sigma^2 N^2 \frac{N-n}{Nn} - \sigma^2 \sum_{k\in U} (\nu_k - 1)^2 .
\end{aligned}$$

The proof is given in the Appendix.

DEFINITION 2.10. *An optimal model-assisted strategy is one with a design-unbiased estimator that, subject to* (2.10)*, minimizes the anticipated mean-squared error of that estimator.*

From § 2.4 and Result 2.4, we obtain directly an optimal model-assisted strategy.

STRATEGY 2.1. *Under the superpopulation model* (2.1)*, an optimal model-assisted strategy consists of using inclusion probabilities that are proportional to $\nu_k$ subject to* (2.10)*, selecting the sample by means of a balanced sampling design on $x_k$, and using the Horvitz-Thompson estimator.*

## 2.6 Model-based approach

Under the model-based approach, the aim is to find a strategy that leads to a valid inference with respect to the model, i.e. a model-unbiased or approximately model-unbiased estimator and a sample that minimizes the error variance $\mathrm{E}_\xi(\widehat{Y} - Y)^2$.

DEFINITION 2.11. *An optimal model-based strategy is one with a linear model-unbiased estimator that, subject to a fixed sample size $n$, minimizes the error variance of that estimator.*

In the model-based approach, this strategy is strictly applied under ideal circumstances, which occur when the model is known to hold. In practice, the modeller must bear model failure in mind, and the model-based approach strongly emphasizes robustness to deviations from the working model. The strictly optimal strategies that are not robust in case of misspecification of the model are thus clearly rejected.

A well-known result (Royall, 1976) is that the model-unbiased linear estimator of $Y$ that minimizes the error variance turns out to be the best linear unbiased estimator given by

$$\widehat{Y}_{\mathrm{BLU}} = \sum_{k\in S} y_k + \sum_{k\in \bar{S}} x_k' \widehat{\beta}_{\mathrm{BLU}}, \tag{2.11}$$

where $\widehat{\beta}_{\mathrm{BLU}}$ is the weighted least-squares estimator of the regression coefficients vector $\beta$; defined by

$$\widehat{\beta}_{\mathrm{BLU}} = A^{-1} \sum_{k\in S} \frac{x_k y_k}{\nu_k^2},$$

where

$$A = \sum_{k\in S} \frac{x_k x_k'}{\nu_k^2}.$$

The error variance of the best linear unbiased estimator is

$$\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2 \left( \sum_{k\in \bar{S}} x_k' A^{-1} \sum_{\ell \in \bar{S}} x_\ell + \sum_{k\in \bar{S}} \nu_k^2 \right). \tag{2.12}$$

Consequently, to determine a model-based strategy, we look for a sample $s$ that minimizes (2.12), this sample being not necessarily unique.

STRATEGY 2.2. *Under the superpopulation model* (2.1)*, an optimal model-unbiased strategy consists of using the best linear unbiased estimator, and choosing a sample of size n that minimizes expression* (2.12)*.*

Again, this strategy must be put into perspective with respect to possible misspecification of the model. If the sample that minimizes (2.12) is very particular, then a more robust strategy should be considered.

With certain superpopulation models, expression (2.12) can be considerably simplified. Moreover, minimizing the anticipated mean-squared error given in (2.13) below in the class of linear model-unbiased estimators also leads to Strategy 2.2:

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2 \left\{ \mathrm{E}_p \left( \sum_{k\in \bar{S}} x_k' A^{-1} \sum_{\ell \in \bar{S}} x_\ell \right) + \sum_{k\in U} (1-\pi_k)\nu_k^2 \right\}. \tag{2.13}$$

Unfortunately, expression (2.13) cannot be much simplified.

DEFINITION 2.12. *Consider the following conditions:*

(i) *there exists a vector* $\lambda \in \mathbb{R}^q$ *such that* $\lambda' x_k = \nu_k^2$;

(ii) *there exists a vector* $\theta \in \mathbb{R}^q$ *such that* $\theta' x_k = \nu_k$.

*If both conditions are met, then model* (2.1) *is said to have fully explainable heteroscedasticity.*

RESULT 2.5. (Royall, 1992) *If the superpopulation model* (2.1) *is such that condition* (i) *of Definition* 2.12 *is met, then* $\widehat{Y}_{\mathrm{BLU}} = \sum_{k \in U} x_k' \widehat{\beta}_{\mathrm{BLU}}$, *and* $\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2 \left( X' A^{-1} X - \sum_{k \in U} \nu_k^2 \right)$.

RESULT 2.6. (Royall, 1992) *If the superpopulation model* (2.1) *has fully explainable heteroscedasticity, then*

$$\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 \geq \sigma^2 \left( \frac{N^2}{n} - \sum_{k \in U} \nu_k^2 \right),$$

*and, if the sample is such that*

$$\frac{1}{n} \sum_{k \in S} \frac{x_k}{\nu_k} = \frac{\sum_{k \in U} x_k}{N},$$

*then the bound for the error variance is achieved.*

Royall (1992) and later Valliant et al. (2000, pp. 98-100) in their Theorem 4.2.1 and consequent Remark 4 present results which from a design-based point of view can be used to prove the following result.

RESULT 2.7. *If the superpopulation model* (2.1) *has fully explainable heteroscedasticity and if the sample is balanced with inclusion probabilities proportional to* $\nu_k$, *then the best linear unbiased estimator* $\widehat{Y}_{\mathrm{BLU}}$ *equals the Horvitz-Thompson estimator* $\widehat{Y}_\pi$ *and the bound for the error variance is achieved.*

Note that, under the conditions of Result 2.7, $\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \mathrm{E}_p \mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2$.

## 2.7 A combined model-based and model-assisted approach

A third option for estimating $Y$ consists of finding a strategy that is simultaneously design-unbiased and model-unbiased. From Corollary 2.4, we know that such a strategy has an

anticipated mean-squared error equal to

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 \;=\; \sigma^2 \sum_{k\in U} \nu_k^2 \left\{ \pi_k \mathrm{var}_p(w_{kS}|I_k = 1) + \frac{1-\pi_k}{\pi_k} \right\}.$$

If the weights $w_{ks}$ are not random, then we obtain the Godambe-Joshi bound given by

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 \geq L_p = \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k}{\pi_k}. \tag{2.14}$$

Thus, an optimal strategy that is at the same time model-unbiased and design-unbiased consists simply of taking the Strategy 2.1. Indeed, in this case, the bound in expression (2.14) is achieved.

## 2.8 Estimation of variance

From the previous sections, it clearly appears that the Horvitz-Thompson estimator with a balanced sampling design is a valuable strategy that leads to valid inference under the model and under the sampling design. The estimation of the total should be complemented by a confidence interval. We will show that it is possible to construct a variance estimator that leads to a valid inference under the model and under the sampling design.

In order to estimate the variance, it is prudent to treat the $\nu_k$ as if they were unknown, even if the sample has been selected assuming known $\nu_k$. This will make the estimation of model variance in some sense robust to the failure of that assumption; see for example Cumberland and Royall (1981). In the model-assisted framework, Deville and Tillé (2005) have proposed a family of variance estimators for balanced sampling, of the form

$$\hat{\mathrm{var}}(\widehat{Y}_\pi) = \sum_{k\in S} c_k \frac{\left(y_k - x_k'\widehat{b}\right)^2}{\pi_k^2}, \tag{2.15}$$

where

$$\widehat{b} = \left( \sum_{\ell\in S} c_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2} \right)^{-1} \sum_{\ell\in S} c_\ell \frac{x_\ell y_\ell}{\pi_\ell^2}$$

and the $c_k$ are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k x_k'}{\pi_k} \left( \sum_{\ell \in S} c_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2} \right)^{-1} \frac{c_k x_k}{\pi_k}, \tag{2.16}$$

which can be solved by a fixed-point algorithm.

In Deville and Tillé (2005), simpler variants of $c_k$ are also proposed. These variants are based on the fact that

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

The estimator $\hat{\text{var}}(\widehat{Y}_\pi)$ is approximately design-unbiased because it is an estimator by substitution of the approximation given in expression (2.7), (on the estimators by substitution, see Deville, 1999), which is a reasonable approximation of the variance under the sampling design.

For the model-based framework, the question of estimating $\text{E}_\xi(\widehat{Y}_\pi - Y)^2$ is complicated because it depends on all the $\nu_k$ of the population and not just on the $\nu_k$ of the sample. Nevertheless, the following result shows that $\hat{\text{var}}(\widehat{Y}_\pi)$ is also a pertinent estimator of $\text{E}_\xi(\widehat{Y}_\pi - Y)^2$ and can be model-unbiased.

RESULT 2.8. *Under model* (2.1)*, if the sample is balanced on* $x_k$*, then*

$$\begin{aligned} \text{E}_\xi \left\{ \hat{\text{var}}(\widehat{Y}_\pi) \right\} &= \text{E}_\xi(\widehat{Y}_\pi - Y)^2 + \sigma^2 \left( \sum_{k \in S} \frac{\nu_k^2}{\pi_k} - \sum_{k \in U} \nu_k^2 \right), \\ \text{E}_p \text{E}_\xi \left\{ \hat{\text{var}}(\widehat{Y}_\pi) \right\} &= \text{E}_p \text{E}_\xi(\widehat{Y}_\pi - Y)^2. \end{aligned}$$

*If condition* (i) *of Definition* 2.12 *is met, then* $\hat{\text{var}}(\widehat{Y}_\pi)$ *is a model-unbiased estimator of* $\text{E}_\xi(\widehat{Y}_\pi - Y)^2$.

The proof is given in the Appendix.

If $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal variable, the confidence interval

$$\text{CI}(1-\alpha) = \left[ \widehat{Y}_\pi - z_{1-\alpha/2} \sqrt{\{\hat{\text{var}}(\widehat{Y}_\pi)\}}, \widehat{Y}_\pi + z_{1-\alpha/2} \sqrt{\{\hat{\text{var}}(\widehat{Y}_\pi)\}} \right]$$

leads to a reasonable design-based inference and a valid model-based inference provided that the $\nu_k^2$ can be expressed as linear combinations of the auxiliary variables. This inference does not depend on assumed values of the standard deviations of the errors of the model.

## 2.9 Examples

In the examples, we will use the notation

$$\bar{X} = \frac{1}{N}\sum_{k\in U} x_k,\ \bar{x} = \frac{1}{n}\sum_{k\in S} x_k,\ \bar{y} = \frac{1}{n}\sum_{k\in S} y_k,\ \bar{y}_h = \frac{1}{n_h}\sum_{k\in U_h\cap S} y_k,$$

where $U_1,\ldots,U_H$ are strata, i.e. the $U_h$, $h=1,\ldots,H$, are a partition of $U$. Moreover,

$$s_x^2 = \frac{1}{n-1}\sum_{k\in S}(x_k-\bar{x})^2,\ s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k-\bar{y})^2,$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{k\in S}(x_k-\bar{x})(y_k-\bar{y}),\ s_{yh}^2 = \frac{1}{n_h-1}\sum_{k\in U_h\cap S}(y_k-\bar{y}_h)^2.$$

*Example* 2.2. Suppose that the superpopulation model is the constant model $y_k = \beta+\varepsilon_k$, for all $k\in U$, with $\mathrm{var}_\xi(\varepsilon_k)=\sigma^2$. This simple model is homoscedastic and has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. The optimal model-based strategy consists of selecting any sample of fixed sample size $n$, deliberately or randomly. The optimal model-assisted strategy consists of selecting a sample that is balanced on the constant, which implies that it has a fixed sample size. This sample must be selected with equal inclusion probabilities $n/N$. In practice, a simple random sampling can be applied and the anticipated mean-squared error is

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2 N^2\frac{N-n}{Nn}.$$

In this case, $\widehat{Y}_\pi = N\bar{y}$,

$$c_k = \frac{(N-n)n}{N(n-1)},\quad \hat{\mathrm{var}}(\widehat{Y}_\pi) = N^2\frac{N-n}{Nn}s_y^2\ .$$

*Example* 2.3. Suppose that the superpopulation model consists of a constant and only one independent variable, i.e. $y_k = \beta_0 + x_k\beta_1+\varepsilon_k$, for all $k\in U$, with $\mathrm{var}_\xi(\varepsilon_k)=\sigma^2$. This model is homoscedastic and has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. For a particular sample $S$, balanced or not, and with fixed sample size, the error variance of the best linear

unbiased estimator is

$$\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2 N^2 \frac{N-n}{Nn} + \sigma^2 N^2 \frac{(\bar{x} - \bar{X})^2}{(n-1)Ns_x^2}.$$

The optimal model-based strategy consists of selecting a fixed-sample-size balanced sample in the sense that $\bar{x} = \bar{X}$. The optimal model-assisted strategy consists of selecting a sample that is balanced on $x_k$, of fixed sample size and with equal inclusion probabilities. This can be done by using the cube method. Next, one uses the Horvitz-Thompson estimator. The anticipated mean-squared error is then

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2 N^2 \frac{N-n}{Nn}.$$

Moreover,

$$c_k \doteq \frac{(N-n)n}{N(n-2)},$$

and, by using this approximation for the $c_k$, we obtain

$$\hat{\mathrm{var}}(\widehat{Y}_\pi) = N^2 \frac{N-n}{Nn} \frac{1}{n-2} \sum_{k \in S} (y_k - \widehat{\beta}_0 - \widehat{\beta}_1 x_k)^2,$$

where $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ and $\widehat{\beta}_1 = s_{xy}/s_x^2$.

*Example* 2.4. Suppose that the superpopulation model has only one independent variable, i.e. $y_k = x_k\beta + \varepsilon_k$, for all $k \in U$, with $\mathrm{var}_\xi(\varepsilon_k) = \nu_k^2\sigma^2$, where $\nu_k = Nx_k/X$, $x_k \geq 0$ and $X = \sum_{k \in U} x_k$. This model does not have fully explainable heteroscedasticity, which implies that the model-assisted and model-based optimal strategies are not the same. The optimal model-based strategy consists of using the best linear unbiased estimator. From expression (2.12), knowing that $A = X^2 n/N^2$, we obtain the anticipated mean-squared error

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2 \mathrm{E}_p \left\{ \frac{1}{n} \left( \sum_{k \in \bar{S}} \nu_k \right)^2 + \sum_{k \in \bar{S}} \nu_k^2 \right\}. \tag{2.17}$$

In this case, the best strictly model-based strategy consists of selecting a non-random sample containing the largest $n$ units. However, Valliant et al. (2000, p. 55) point out that, in this case, 'selecting this sample may be quite a risky procedure if the working model is wrong' because it fails to protect against model failure. By using an alternative more gen-

eral model, they conclude that a balanced sample will protect against model-bias resulting from a misspecification. From a design-based point of view, the strictly best model-based strategy leads to an incorrect design-based inference. The optimal model-assisted strategy consists of using a sampling design that is balanced on $x_k$ and has unequal inclusion probabilities proportional to $x_k$ with the Horvitz-Thompson estimator. The anticipated mean-squared error is then

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2 \left( \frac{N^2}{n} - \sum_{k \in U} \nu_k^2 \right).$$

This strategy has a larger anticipated mean-squared error than (2.17) but leads to correct model-assisted and model-based inferences. In this case, the estimator of the variance is

$$\hat{\mathrm{var}}(\widehat{Y}_\pi) = \sum_{k \in S} \frac{c_k}{\pi_k^2} \left( y_k - \pi_k \frac{\sum_{\ell \in S} c_\ell y_\ell / \pi_\ell}{\sum_{\ell \in S} c_\ell} \right)^2,$$

where $c_k$ are the solutions of the nonlinear system $1 - \pi_k = c_k - c_k^2 \left( \sum_{\ell \in S} c_\ell \right)^{-1}$ or more simply can be approximated by $c_k \doteq (1 - \pi_k)n/(n-1)$.

*Example* 2.5. We consider the superpopulation model presented in Kott (1986), given by $y_k = x_k\beta_1 + x_k^2\beta_2 + \varepsilon_k$, for all $k \in U$, with $\mathrm{var}_\xi(\varepsilon_k) = \nu_k^2\sigma^2$, where $\nu_k = Nx_k/X$ and $X = \sum_{k \in U} x_k$. This model has fully explainable heteroscedasticity, which implies that the model-assisted and the model-based optimal strategies are the same. Therefore, a strategy that is optimal for both the model-assisted and model-based frameworks consists of selecting a sample balanced on $x_k$ and $x_k^2$ with inclusion probabilities that are proportional to $x_k$, and using the Horvitz-Thompson estimator. The anticipated mean-squared error is then

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2 \left( \frac{N^2}{n} - \sum_{k \in U} \nu_k^2 \right).$$

This strategy leads to correct model-assisted and model-based inferences.

*Example* 2.6. Consider the stratified superpopulation model $y_{kh} = \alpha_h + \varepsilon_k$, for all $k \in U_h$, $h = 1, \ldots, H$. Moreover, suppose that $\mathrm{var}_\xi(\varepsilon_{kh}) = \nu_h^2\sigma^2$, with $\sum_{h=1}^H N_h\nu_h = N$. The stratified model has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. The optimal model-based strategy consists of defining the inclusion probabilities proportional to $\nu_h$, which gives

$\pi_{kh} = n\nu_h/N$, which is an optimal stratification. Next, a sample is selected with a fixed sample size $n_h = nN_h\nu_h/N$ in each stratum $U_h$. The Horvitz-Thompson estimator, $\widehat{Y}_\pi = \sum_{h=1}^{H} N_h\bar{y}_h$ has anticipated mean-squared error

$$\mathrm{E}_p\mathrm{E}_\xi(\widehat{Y}_\pi - Y)^2 = \sigma^2\left(\frac{N^2}{n} - \sum_{h=1}^{H} N_h\nu_h^2\right) = \sigma^2\frac{N^2}{n}\left(1 - \frac{1}{n}\sum_{h=1}^{H}\frac{n_h^2}{N_h}\right).$$

In this case,

$$c_k = \frac{(N_h - n_h)n_h}{N_h(n_h - 1)}, k \in U_h,$$

and thus

$$\hat{\mathrm{var}}(\widehat{Y}_\pi) = \sum_{h=1}^{H} N_h^2\frac{N_h - n_h}{N_h n_h}s_{yh}^2.$$

## 2.10 Discussion

The search for an optimal strategy rather than an optimal estimator allows the proponents of the model-based and the model-assisted approaches to resolve their differences because, when the superpopulation model has fully explainable heteroscedasticity, one chooses the same sampling design which is a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model. In this case, the best linear unbiased estimator is equal to the Horvitz-Thompson estimator. As a complement to this estimator, an estimator of the variance can be given, which in turn leads to valid model-based and design-based inferences. The controversy makes sense only if the sample is chosen inappropriately. If the superpopulation model has fully explainable heteroscedasticity, then Strategy 2.1 is the best strategy in the model-based, model-assisted and combined model-based and model-assisted frameworks, as presented in Table 1.

If the heteroscedasticity is not fully explainable, the optimal strategy is not the same in the model-assisted and model-based frameworks. In fact, Strategy 2.1 always leads to the selection of a balanced sample, while the strict application of Strategy 2.2 can lead either to the selection of a balanced sample or to a purposive selection of the sample as in Example 2.4 in § 9. In this second case, a robustness argument is usually used by the modeller in order to protect against misspecification of the model. The robustness is obtained by balancing the sample for the variables that are in the alternative model, which

gives the same strategy as in the model-assisted framework. Thus the two approaches are not far apart. In any case, it can also be wise to balance the sampling design with respect to additional variables in order to protect against failure of the model, such as the presence of curvature or an intercept. However, we suggest the use of models that have fully explainable heteroscedasticity, which can be easily achieved by systematically using $\nu_k$ and $\nu_k^2$ as independent variables in the model. This was the advantage of the model developed by Kott (1986) and summarized in Example 5 over the model given in Example 4, which does not have fully explainable heteroscedasticity.

The theory developed in this paper shows that the best approach is to select a sample that is balanced on the auxiliary variables. If exact balancing is not possible, a nearly balanced sample must first be selected. In this case, the rounding problem can be solved by a small calibration, by using either the calibration estimator (Deville and Särndal, 1992) or the best linear unbiased estimator, depending on the basis of the inference. An interesting particular case is the so-called cosmetic calibration proposed by Brewer (1999a). In a set of simulations, Deville and Tillé (2004) showed that the balanced sampling design with a calibration estimator strategy achieves the best results among the following four strategies: (i) non-balanced sampling with the Horvitz-Thompson estimator, (ii) balanced sampling with the Horvitz-Thompson estimator, (iii) non-balanced sampling with a calibration estimator, and (iv) balanced sampling with a calibration estimator. With strategy (iv), the weights $w_{ks}$ are less random than in the case of strategy (iii), and this leads to a more accurate estimator.

Table 2.1: Optimal strategies in the model-assisted, model-based and combined model-based and model-assisted approaches

| Approach | Fully explainable heteroscedasticity | Non fully explainable heteroscedasticity |
|---|---|---|
| MB | Strategy 2.1 | Strategy 2.2 |
| MA | Strategy 2.1 | Strategy 2.1 |
| CMBMA | Strategy 2.1 | Strategy 2.1 |

MB, model-based; MA, model-assisted; CMBMA, combined model-based and model-assisted

## Acknowledgement

The authors would like to thank Alina Matei, Phil Kott and the two reviewers for their helpful comments and suggestions. This work is in part supported by the Swiss National Science Foundation Grant 205121-105187.

## Appendix

### *Proofs*

*Proof of Result* 2.1. Since

$$\begin{aligned}\widehat{Y}_w - Y &= \sum_{k\in S} w_{kS} y_k - \sum_{k\in U} y_k \\ &= \sum_{k\in S} w_{kS} x_k'\beta + \sum_{k\in S} w_{kS}\varepsilon_k - \sum_{k\in U} x_k'\beta - \sum_{k\in U}\varepsilon_k \\ &= \sum_{k\in S}(w_{kS}-1)\varepsilon_k - \sum_{k\in \bar{S}}\varepsilon_k + \sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta,\end{aligned}$$

we have that

$$\mathrm{E}_\xi(\widehat{Y}_w - Y)^2 = \sigma^2\left\{\sum_{k\in S}(w_{kS}-1)^2\nu_k^2 + \sum_{k\in\bar{S}}\nu_k^2\right\} + \left(\sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2, \qquad \text{(A1)}$$

which leads to the first equality of Result 2.1. The second term of (A1) can be simplified. Indeed,

$$\begin{aligned}
&\mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2 = \\
&= \mathrm{E}_p\left\{\sum_{k\in S} w_{kS}x_k'\beta - \mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + \mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) - \sum_{k\in U} x_k'\beta\right\}^2 \\
&= \mathrm{E}_p\left\{\sum_{k\in S} w_{kS}x_k'\beta - \mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right)\right\}^2 + \mathrm{E}_p\left\{\sum_{k\in U} \mathrm{E}_p(w_{kS}I_k)x_k'\beta - \sum_{k\in U} x_k'\beta\right\}^2 \\
&\quad +2\mathrm{E}_p\left[\left\{\sum_{k\in S} w_{kS}x_k'\beta - \mathrm{E}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right)\right\}\left\{\sum_{k\in U} \mathrm{E}_p(w_{kS}I_k)x_k'\beta - \sum_{k\in U} x_k'\beta\right\}\right] \\
&= \mathrm{var}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + \left(\sum_{k\in U} C_k x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2. \qquad \text{(A2)}
\end{aligned}$$

The first term of (A1) gives

$$\begin{aligned}
&\sigma^2\mathrm{E}_p\left\{\sum_{k\in S}(w_{kS}-1)^2\nu_k^2 + \sum_{k\in\bar{S}}\nu_k^2\right\} \\
&= \sigma^2\left[\sum_{k\in U}\mathrm{E}_p\left\{(w_{kS}-1)^2I_k\right\}\nu_k^2 + \sum_{k\in U}(1-\pi_k)\nu_k^2\right] \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left[\mathrm{E}_p\left\{(w_{kS}-1)^2I_k\right\} + 1 - \pi_k\right] \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{\mathrm{E}_p(w_{kS}^2I_k) - 2\mathrm{E}_p(w_{kS}I_k) + \pi_k + 1 - \pi_k\right\} \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{\mathrm{E}_p(w_{kS}^2I_k) - \mathrm{E}_p^2(w_{kS}I_k) + \mathrm{E}_p^2(w_{kS}I_k) - 2\mathrm{E}_p(w_{kS}I_k) + 1\right\} \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{\mathrm{var}_p(w_{kS}I_k) + (C_k-1)^2\right\}. \qquad \text{(A3)}
\end{aligned}$$

By the law of total variance,

$$\begin{aligned}\mathrm{var}_p(w_{kS}I_k) &= \mathrm{var}_p\mathrm{E}_p(w_{kS}I_k|I_k) + \mathrm{E}_p\mathrm{var}_p(w_{kS}I_k|I_k)\\ &= \pi_k\{\mathrm{E}_p(w_{kS}|I_k=1)\}^2 - \{\mathrm{E}_p(w_{kS}I_k)\}^2 + \pi_k\mathrm{var}_p(w_{kS}|I_k=1)\\ &= \frac{1-\pi_k}{\pi_k}C_k^2 + \pi_k\mathrm{var}_p(w_{kS}|I_k=1). \qquad \text{(A4)}\end{aligned}$$

By inserting (A4) into (A3), and by adding (A2) and (A3), we finally obtain the second equality of Result 2.1. □

*Proof of Result* 2.2. Result 2.2 comes directly from equation (2.2). Term $B$ vanishes because the weights $1/\pi_k$ do not differ from sample to sample. Term $C$ vanishes because the estimator is design-unbiased. Terms $D$ and $E$ vanish because the estimator is model-unbiased under balanced sampling. All that remains is term A with $C_k = 1$ because the estimator is design unbiased. □

*Proof of Result* 2.3. Since $y_k = x_k'\beta + \varepsilon_k$,

$$\begin{aligned}\mathrm{var}_{\mathrm{app}}(\widehat{Y}_\pi) &= \sum_{k\in U} d_k\frac{(y_k - x_k'b)^2}{\pi_k^2}\\ &= \sum_{k\in U} d_k\left\{\frac{\varepsilon_k}{\pi_k} - \frac{x_k{}'}{\pi_k}\left(\sum_{\ell\in U} d_\ell\frac{x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1}\sum_{\ell\in U} d_\ell\frac{x_\ell\varepsilon_\ell}{\pi_\ell^2}\right\}^2\\ &= \sum_{k\in U} d_k\frac{\varepsilon_k^2}{\pi_k^2} - \sum_{k\in U}\frac{d_k x_k'\varepsilon_k}{\pi_k^2}\left(\sum_{\ell\in U}\frac{d_\ell x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1}\sum_{\ell\in U}\frac{d_\ell x_\ell\varepsilon_\ell}{\pi_\ell^2}.\end{aligned}$$

Thus,

$$\mathrm{E}_\xi\left\{\mathrm{var}_{\mathrm{app}}(\widehat{Y}_\pi)\right\} = \sigma^2\sum_{k\in U} d_k\frac{\nu_k^2}{\pi_k^2} - \sigma^2\sum_{k\in U}\frac{\nu_k^2}{\pi_k^2}\frac{d_k x_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell\frac{x_\ell x_\ell'}{\pi_k^2}\right)^{-1}\frac{d_k x_k}{\pi_k}.$$

By using the definition of $d_k$, given in expression (2.8), we obtain

$$\mathrm{E}_\xi \left\{ \mathrm{var}_{\mathrm{app}}(\widehat{Y}_\pi) \right\} = \sigma^2 \sum_{k \in U} \pi_k (1 - \pi_k) \frac{\nu_k^2}{\pi_k^2} = \mathrm{E}_p \mathrm{E}_\xi (\widehat{Y}_\pi - Y)^2.$$

Moreover, this holds even when the $\nu_k$ are unknown. □

*Proof of Result* 2.4. The optimal inclusion probabilities $\pi_k^*$ are obtained by minimizing (2.9) subject to

$$\sum_{k \in U} \pi_k = n, \quad 0 \leq \pi_k \leq 1,$$

which gives the second inequality. Now, if we minimize (2.9) subject to $\sum_{k \in U} \pi_k = n$, but without the constraint $\pi_k \leq 1$, then we obtain $\widetilde{\pi}_k = n\nu_k / N$, and we obtain a still lower bound in the third inequality. □

*Proof of Result* 2.8. By Result 2.3, following the same steps, we obtain

$$\begin{aligned}
\mathrm{E}_\xi \left\{ \hat{\mathrm{var}}(\widehat{Y}_\pi) \right\} &= \sigma^2 \sum_{k \in S} (1 - \pi_k) \frac{\nu_k^2}{\pi_k^2} \\
&= \sigma^2 \left\{ \sum_{k \in S} (1 - \pi_k)^2 \frac{\nu_k^2}{\pi_k^2} + \sum_{k \in \bar{S}} \nu_k^2 \right\} + \sigma^2 \left( \sum_{k \in S} \frac{\nu_k^2}{\pi_k} - \sum_{k \in U} \nu_k^2 \right) \\
&= \mathrm{E}_\xi (\widehat{Y}_\pi - Y)^2 + \sigma^2 \left( \sum_{k \in S} \frac{\nu_k^2}{\pi_k} - \sum_{k \in U} \nu_k^2 \right).
\end{aligned}$$

Obviously, if there exists a vector $\lambda$ such that $\lambda' x_k = \nu_k^2$, then

$$\sum_{k \in S} \frac{\nu_k^2}{\pi_k} - \sum_{k \in U} \nu_k^2 = 0.$$

□

# Chapter 3

# Bias Robustness and Efficiency in Model-Based Inference

**Abstract:**

In model-based inference, selecting balanced samples has been considered to give protection against misspesification of the model. Recent developments in finite population sampling proved that balanced samples can be selected randomly. If we accept the idea that balanced samples are randomly selected, then a balanced sampling design with the Horvitz-Thompson estimator compose a strategy that is robust and efficient. This strategy generalizes numerous results obtained in the model-based framework.

## 3.1 Introduction

The principal difference between the model-based and the classical design-based approach lies in the source of randomness they use (Särndal, 1978). In design-based sampling, the inference is based on the stochastic structure induced by the sampling design. In the model-based, or prediction, approach the inference depends on the validity of the model

[1]This chapter is a reprint of the paper: D. Nedyalkova and Y. Tillé. Bias-Robustness and Efficiency in Model-Based Inference, Submitted, 2009.

used to describe the data. In this case, the randomness is due to the population and not to the sampling design as it is the case in the design-based approach.

The model-based approach was developed, amongst others, by Royall (1976, 1992); Royall and Cumberland (1981) and Chambers (1996). When the data are assumed to follow a linear model, Royall (1976) proposed the use of the best linear unbiased predictor. The model-based approach has been criticized due to the fact that it may lead to severe bias if the model assumptions are violated. In contrast to model-based inference, design-based inference is considered to be robust by definition. Brewer and Särndal (1983) point out that since the inference is not based on a model there is no need to worry what will happen if the model is misspecified.

Much of the work in recent model-based research has been devoted to constructing robust strategies. More specifically, in order to protect the inference against a misspecified model, Royall and Herson (1973a,b) and Scott et al. (1978) point out the importance of *balanced* samples, where balance is achieved by equalizing the sample moments of the independent variables with those in the population. They have come to the conclusion that the sample must be *balanced*, but not necessarily random.

Another way to accomplish robustness in the model-based approach is to choose an appropriate sampling design. Since Deville and Tillé (2004)'s paper, it is now possible to select balanced samples randomly using a procedure called the cube method. Nedyalkova and Tillé (2008) have shown that under a balanced sampling design, with inclusion probabilities proportional to the standard deviations of the errors of the model, and under certain conditions defined as 'fully explainable heteroscedasticity' the best linear unbiased estimator equals the Horvitz-Thompson estimator. This represents an optimal strategy which reconciles both approaches.

The problem of comparing the two paradigms is solved by a result due to Isaki and Fuller (1982). They showed that if an estimator is both design- and model-unbiased, then the design expectation of its model variance is equal to the model expectation of its design mean squared error, a quantity that they named the anticipated variance.

The purpose of this paper is to investigate the different strategies leading to bias-robust strategies under the model-based framework. This paper is organized as follows: the notation and basic definitions are given in Section 2. The model-based framework is briefly introduced in Section 3. In Section 4, the focus is put on the properties of the d-balanced estimator under a linear model. The subject of bias robustness in submodels

is discussed in Section 5. An application to the polynomial model is given in Section 6. Finally, in Section 7, a few concluding remarks are given.

## 3.2 Notation and definitions

Consider a population $U$ of size $N$. Each unit of the population can be identified by a label $k = 1, \ldots, N$. Suppose that a register is available, and that the values of $p$ auxiliary variables are known for each unit of the population. Let $y_k$ be the value taken by the variable of interest $y$ on the $k$th unit of the population. The values $y_k$ are unknown. We are interested in estimating the population total $Y = \sum_{k \in U} y_k$. The total $Y$ is estimated by a sample $s$ of size $n$, where $s$ is a subset of $U$. A sample can be selected randomly or not. A sampling design is defined by assigning to each sample $s$ a probability $p(s)$ of being selected. Let $S$ denote the random sample such that $\Pr(S = s) = p(s)$. The inclusion probability $\pi_k$ is then the probability that unit $k$ is selected in the sample. We also denote by $\mathrm{E}_p(\cdot)$ and $\mathrm{var}_p(\cdot)$, respectively, the expectation and the variance under the sampling design $p(\cdot)$ and by $\bar{S}$ the set of units of the population which are not in $S$.

Definition 3.1. *An estimator $\widehat{Y}$ is said to be design-unbiased if* $\mathrm{E}_p(\widehat{Y}) - Y = 0$.

Definition 3.2. *A sample $s$ is said to be d-balanced on a set of variables $\mathbf{x}_k = (x_{k1} \cdots x_{kp})$ if and only if*

$$\sum_{k \in s} d_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k,$$

*where $d_1, \ldots, d_k, \ldots, d_N$ is a set of weights that do not depend on the sample $s$.*

A balanced sample can eventually (but not necessarily) be selected randomly. When $\pi_k > 0$, for all $k \in U$, the Horvitz-Thompson estimator of $Y$ given by

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}$$

is design-unbiased, i.e. $\mathrm{E}_p(\widehat{Y}_\pi) = Y$.

In order to randomly select a balanced sample, the set of inclusion probabilities is defined by $\pi_k = 1/d_k$. A procedure that randomly selects a balanced sample is called a balanced sampling design. According to the definition of Deville and Tillé (2004), a sampling design $p(\cdot)$ is said to be balanced on the auxiliary variables $x_1, \ldots, x_p$ if and only

if

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k. \tag{3.1}$$

Authors such as Cumberland and Royall (1981) and Kott (1986) would call this a '$\pi$-balanced sampling', opposed to a mean-balanced sampling defined by the equation

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

Below, we use the expression 'balanced sampling' to denote a sampling design that satisfies equation (3.1) for one or more auxiliary variables, a mean-balanced sampling being a particular case of this balanced sampling when the sample is selected with inclusion probabilities $n/N$. If the population size is small, a balanced sampling design can be implemented by a linear program. For larger population sizes, the cube method may be used (see Deville and Tillé, 2004; Tillé, 2006).

## 3.3 Model-based strategy and BLU estimator

We assume that the population follows a linear model $M$

$$y_k=\mathbf{x}_k'\boldsymbol{\beta}+\varepsilon_k, \tag{3.2}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of random variables $\varepsilon_k$ such that

$$\mathrm{E_M}(\varepsilon_k)=0, \mathrm{var_M}(\varepsilon_k)=\sigma^2\nu_k^2, \mathrm{cov_M}(\varepsilon_k,\varepsilon_\ell)=0 \text{ if } k\neq\ell,$$

The quantities $\nu_k, k\in U$, are assumed known. For simplicity, we scale them so that

$$\sum_{k\in U}\nu_k=N.$$

Model (3.2) includes the possibility of heteroscedasticity. Under homoscedasticity, $\nu_k=1$ for all $k\in U$. An important and common hypothesis is that the random sample $S$ and the errors $\varepsilon_k$ of the model are independent. The symbols $\mathrm{E_M}(.)$, $\mathrm{var_M}(.)$, $\mathrm{cov_M}(.)$ denote, respectively, the expected value, the variance and the covariance under model $M$.

Definition 3.3. *An estimator $\widehat{Y}$ is said to be model-unbiased if* $\mathrm{E_M}(\widehat{Y}-Y)=0$.

DEFINITION 3.4. *The model mean squared error of an estimator* $\widehat{Y}$ *is defined by* $\mathrm{E_M}\left(\widehat{Y}-Y\right)^2$.

The model mean squared error is also called the error variance. Royall (1976) showed that, in the framework of the model-based inference, the Best Linear Unbiased (BLU) estimator is given by

$$\widehat{Y}_{\mathrm{BLU}} = \sum_{k\in S} y_k + \sum_{k\in \bar{S}} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} = \sum_{k\in U} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S}(y_k - \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}}) = \sum_{k\in U} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S} e_k, \tag{3.3}$$

where

$$e_k = y_k - \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}}, \tag{3.4}$$

and

$$\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} = \mathbf{A}^{-1}\sum_{k\in S}\frac{\mathbf{x}_k y_k}{\nu_k^2},$$

where

$$\mathbf{A} = \sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k'}{\nu_k^2}.$$

The error variance of the best linear unbiased estimator is

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2\left(\sum_{k\in\bar{S}}\mathbf{x}_k'\mathbf{A}^{-1}\sum_{\ell\in\bar{S}}\mathbf{x}_\ell + \sum_{k\in\bar{S}}\nu_k^2\right).$$

DEFINITION 3.5. (Hájek, 1981, p. 153) *A strategy is a pair* $(p(\cdot), \widehat{Y})$ *comprising a sampling design and an estimator.*

STRATEGY 3.1. *An optimal, purely model-unbiased, strategy consists of using the best linear unbiased estimator and choosing a sample of size* $n$ *that minimizes* $\mathrm{E}_M(\widehat{Y}_{\mathrm{BLU}} - Y)^2$.

This strategy is sometimes not robust because, in some cases, it can lead to the choice of a very extreme sample. The classical example (see for instance Royall and Herson, 1973a) is the model without intercept and with only one regressor

$$y_k = x_k\beta + \varepsilon_k, \tag{3.5}$$

with $\mathrm{var}_{\mathrm{M}}(\varepsilon_k) = \sigma^2\nu_k^2$, with $\nu_k^2 \propto x_k$, that leads to the BLU estimator

$$\widehat{Y}_{\mathrm{BLU}} = \frac{\sum_{k\in U} x_k}{\sum_{k\in S} x_k} \sum_{k\in S} y_k,$$

which, in this case, is equal to the ordinary ratio estimator $\widehat{Y}_{\mathrm{R}}$. As $\nu_k^2 \propto x_k$ and $\sum_{k\in U} \nu_k = N$, we have that

$$\nu_k^2 = \frac{N^2 x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2}.$$

The error variance of $\widehat{Y}_{\mathrm{R}}$ is given by the expression

$$\mathrm{E}_{\mathrm{M}}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2} \left( \frac{\sum_{k\in \bar{S}} x_k}{\sum_{k\in S} x_k} \sum_{k\in U} x_k \right).$$

Thus, in this case, the optimal purely model-based strategy consists of choosing the units with the $n$ largest values of the variable $x$ (Royall, 1970). This strategy can be very dangerous if the model is wrong. It is thus reasonable to opt for a strategy that guarantees correct estimation when the model is not correct.

## 3.4 Balanced estimator under a linear model

Now consider the $d$-weighted estimator

$$\widehat{Y}_d = \sum_{k\in S} d_k y_k = \sum_{k\in S} d_k \mathbf{x}_k' \widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S} d_k e_k, \tag{3.6}$$

where $e_k$ is defined in (3.4). Under $d$-balanced sampling, $\widehat{Y}_d$ is model-unbiased and its error variance is

$$\mathrm{E}_{\mathrm{M}}(\widehat{Y}_d - Y)^2 = \sigma^2 \left[ \sum_{k\in S} (d_k - 1)^2 \nu_k^2 + \sum_{k\in \bar{S}} \nu_k^2 \right]. \tag{3.7}$$

By comparing (3.6) with (3.3), we directly obtain the following result:

Result 3.1. *A sufficient condition in order that $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$ is that*

- *the sampling design is $d$-balanced on $\mathbf{x}_k$,*
- $\sum_{k\in S} e_k(d_k - 1) = 0.$

A particular case of Result 3.1 is given below.

COROLLARY 3.1. *A sufficient condition in order that* $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$ *is that*

- *the sampling design is d-balanced on* $\mathbf{x}_k$,
- *there exists a vector* $\boldsymbol{\lambda}$ *such that* $\boldsymbol{\lambda}'\mathbf{x}_k = \nu_k^2(d_k - 1)$, *for all* $k \in U$.

**Proof**

If

$$\frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2(d_k-1)} = 1,$$

then

$$\sum_{k\in S} e_k(d_k-1) \;=\; \sum_{k\in S} \frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2(d_k-1)} e_k(d_k-1) = \sum_{k\in S} \frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2} e_k = 0.$$

□

There exist several particular strategies that can be used to meet the conditions of Result 3.1.

STRATEGY 3.2. *Consider the strategy that consists of*

- *Using a d-balanced sampling design on* $\mathbf{x}_k$, *where the* $x_k$ *are chosen so that there exist two vectors* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\gamma}$ *such that* $\boldsymbol{\alpha}'\mathbf{x}_k = \nu_k^2$ *and* $\boldsymbol{\gamma}'\mathbf{x}_k = d_k\nu_k^2$, *for all* $k \in U$,
- *and using the d-weighted estimator.*

With this strategy, we have that $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$. This can be obtained by adding $\nu_k^2$ and $d_k\nu_k^2$ to the list of balancing variables and by selecting a $d$-balanced sample. The value of $d_k$ can be chosen freely. In this case,

$$\sum_{k\in S} d_k\nu_k^2 = \sum_{k\in U} \nu_k^2,$$

and the error variance of the $d$-weighted estimator given in (3.7) simplifies to

$$\mathrm{E_M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in U} (d_k - 1)\nu_k^2.$$

STRATEGY 3.3. *Consider the strategy that consists of*

- *Using a d-balanced sampling design on* $\mathbf{x}_k$*, with* $d_k \propto \nu_k^{-1}$ *and* $x_k$ *chosen so that there exist two vectors* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\gamma}$ *such that* $\boldsymbol{\alpha}'\mathbf{x}_k = \nu_k^2$ *and* $\boldsymbol{\gamma}'\mathbf{x}_k = \nu_k$*, for all* $k \in U$*, which Nedyalkova and Tillé (2008) call a 'fully explainable heteroscedasticity',*

- *and using the d-weighted estimator.*

With this strategy, we also have that $\widehat{Y}_{\text{BLU}} = \widehat{Y}_d$. It can be obtained by adding $\nu_k$ and $\nu_k^2$ to the list of balancing variables and by taking $d_k$ proportional to $\nu_k$. Next, a $d$-balanced sample is selected. This strategy is recommended by Nedyalkova and Tillé (2008) and is in fact a particular case of Strategy3.2. If $d_k = 1/(\nu_k n)$, then the error variance of the $d$-weighted estimator, given in (3.7), simplifies to

$$\mathrm{E_M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in U} \left(\frac{\nu_k}{n} - \nu_k^2\right) = \sigma^2 \left(\frac{N}{n} - \sum_{k\in U} \nu_k^2\right).$$

Strategy 3.4. *Consider the strategy that consists of*

- *Using a d-balanced sampling design on* $\mathbf{x}_k$*, with* $d_k$ *chosen so that* $d_k = (\nu_k^2 + \boldsymbol{\lambda}'\mathbf{x}_k)/\nu_k^2$*, for all* $k \in U$*,*

- *and using the d-weighted estimator.*

With this strategy, we also have that $\widehat{Y}_{\text{BLU}} = \widehat{Y}_d$. This can be obtained by using $d_k = 1 + \boldsymbol{\lambda}'\mathbf{x}_k/\nu_k^2$, for all $k \in U$ and selecting a $d$-balanced sample. Thus, a judicious choice of the $d_k$'s can always equalize the $d$-weighted estimator and the BLU estimator. This is the procedure that is recommended by Scott et al. (1978) in the case of a polynomial model, but we will see that this is not necessarily the best strategy. After some algebra, it is possible to show that

$$\mathrm{E_M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in \bar{S}} d_k \nu_k^2 = \sigma^2 \left(\sum_{k\in \bar{S}} \nu_k^2 + \sum_{k\in \bar{S}} \boldsymbol{\lambda}'\mathbf{x}_k\right).$$

## 3.5 Bias robustness in submodels

A large part of the model-based inference is dedicated to the robustness of the BLU estimator in the case of misspecification of the model. We assume that a model $M$ was used to conceive the strategy, but that the true underlying model is $M^*$. We refer to the following definition:

DEFINITION 3.6. *A strategy is said to be bias-robust for a model $M^*$ if*

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y} - Y) = 0.$$

Consider a model

$$M : y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$$

and an alternative model

$$M^* : y_k = g_k + \eta_k,$$

where $\mathrm{E}_{\mathrm{M}^*}(\eta_k) = 0$, the values $g_k$ can depend on a function of a parameter $\boldsymbol{\gamma}$ and of a set of other variables $\mathbf{z}_k$, and could eventually be written $g_k = f(\boldsymbol{\gamma}'\mathbf{z}_k)$. Note that there is no assumption on the covariance matrix of the vector of $\eta_k$.

DEFINITION 3.7. *Model $M^*$ is said to be a submodel of $M$ if there exists a vector $\mathbf{a}$ such that $\mathbf{a}'\mathbf{x}_k = g_k$.*

Let us consider the following result, which can look trivial, but is the fundamental argument needed to show that a strategy is robust.

RESULT 3.2. *The strategy that consists of using a d-balanced sampling design on $\mathbf{x}_k$ (with any vector of $d_k$) and the d-weighted estimator is bias-robust for any submodel of $M$.*

**Proof**

Under model $M^*$ and using a $d$-balanced sample,

$$\widehat{Y}_d = \sum_{k\in S} d_k y_k = \sum_{k\in S} d_k(g_k + \eta_k) = \sum_{k\in S} d_k(\mathbf{a}'\mathbf{x}_k + \eta_k) = \mathbf{a}'\sum_{k\in U}\mathbf{x}_k + \sum_{k\in S} d_k\eta_k.$$

Thus,

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \mathrm{E}_{M^*}\left(\sum_{k\in S} d_k\eta_k - \sum_{k\in U}\eta_k\right) = 0.$$

□

We will see that a large part of the discussion about bias-robustness is actually a particular case of Result 3.2. This result encourages the statistician to over-specify the model, i.e. to introduce additional variables into model $M$ in order to ensure that $M^*$ is really a submodel of $M$.

Now, suppose that model $M^*$ is not a submodel of $M$. If the sampling design is d-

balanced on the independent variables of model $M$, then

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \sum_{k\in S} d_k g_k - \sum_{k\in U} g_k.$$

If

$$f_k = g_k - \mathbf{x}_k'\boldsymbol{\gamma}, \text{ are the residuals of a linear regression of } g_k \text{ on } x_k$$

then

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \sum_{k\in S} d_k f_k - \sum_{k\in U} f_k, \tag{3.8}$$

for any value of $\boldsymbol{\gamma}$, in particular when

$$\gamma = \left(\sum \frac{\mathbf{x}_k\mathbf{x}_k'}{\mathrm{var}(g_k)}\right)^{-1} \sum \frac{x_k g_k}{\mathrm{var}(g_k)}.$$

If we do not possess information about the $g_k$, the only way in which this could be done consists of selecting a random sample with inclusion probabilities $d_k = 1/\pi_k$, because in selecting randomly the sample, the expected value under the sampling design of (3.8) is equal to zero and, moreover,

$$\frac{E_{\mathrm{M}^*}(\widehat{Y}_d - Y)}{N} = \sum_{k\in S} \frac{f_k}{\pi_k} - \sum_{k\in U} f_k = O_p(1/\sqrt{n}),$$

where $O_p(1/\sqrt{n})$ is a quantity that remains bounded in probability when multiplied by $\sqrt{n}$. The random selection of a sample and the use of a model-unbiased estimator give an ultimate bias protection in the case where it is not possible to over-specify the model, thus guarantee a negligible bias under $M^*$ when $n$ is large.

## 3.6 Application to the polynomial model

### 3.6.1 Presentation of the model

The polynomial model was studied, amongst others, by Royall and Herson (1973a); Scott et al. (1978); Valliant et al. (2000). The model is defined by

$$y_k = \sum_{j=0}^{J} \delta_j \beta_j x_k^j + \varepsilon_k, \tag{3.9}$$

where $x_k$ is the only independent variable, $\beta_j$ is the $j$th regression coefficient, $\delta_j$ is equal to 1 or 0 when the term $\beta_j x_k^k$ appears or not in the regression, $\mathrm{E_M}(\varepsilon_k) = 0$, $\mathrm{var_M}(\varepsilon_k) = \sigma^2\nu_k^2$, and $\mathrm{cov_M}(\varepsilon_k, \varepsilon_\ell) = 0$, when $k \neq \ell$. We also assume that

$$\sum_{k\in U} \nu_k = N.$$

Now, from Result 3.2, for any set of vectors of $d_k$, the $d$-weighted estimator is bias-robust provided that

$$\sum_{k\in S} d_k x_k^j = \sum_{k\in U} x_k^j, \text{ for } j = 0, \ldots, J, \tag{3.10}$$

for any submodel of (3.9). Again, this conclusion can look trivial, but it implies several results on the polynomial model.

### 3.6.2 A first solution

Let $S^*(J)$ be a particular sample for which the following equation holds :

$$\frac{\sum_{k\in\bar{S}} x_k^j}{\sum_{k\in\bar{S}} x_k} = \frac{\sum_{k\in S} x_k^{j+1}/\nu_k^2}{\sum_{k\in S} x_k^2/\nu_k^2}, \text{ for } j = 0, \ldots, J. \tag{3.11}$$

With a sample satisfying equation (3.11), Scott et al. (1978) showed that the estimator

$$\widehat{Y}_0 = \sum_{k\in S} y_k + \sum_{k\in\bar{S}} x_k \frac{\sum_{k\in S} y_k x_k/\nu_k^2}{\sum_{k\in S} x_k^2/\nu_k^2},$$

is BLU, for any polynomial model (3.9) and any value of $\nu_k$. This simple condition on the sample implies that the estimator is bias-robust for a large class of polynomial models.

It can easily be shown that a sufficient condition for a sample to satisfy equation (3.11) is:

$$\sum_{k\in S} x_k^j \left(1 + \frac{\lambda x_k}{\nu_k^2}\right) = \sum_{k\in U} x_k^j, \text{ for } j = 0, \ldots, J, \tag{3.12}$$

where $\lambda$ is a scalar that does not depend on $j$. Equality (3.12) can be satisfied by using

the unequal inclusion probabilities

$$\pi_k = \frac{1}{1 + \lambda x_k/\nu_k^2}, k \in U,$$

and by selecting a balanced sample such that

$$\sum_{k\in S} \frac{x_k^j}{\pi_k} = \sum_{k\in U} x_k^j, j = 0, \ldots, J.$$

The constant $\lambda$ must be chosen in function of the desired sample size, by solving in $\lambda$ the equation

$$\sum_{k\in U} \pi_k = \sum_{k\in U} \frac{1}{1 + \lambda x_k/\nu_k^2} = n.$$

In fact, this is similar to the solution given in Scott et al. (1978) which is simply a particular case of Strategy 3.4, with $\pi_k = 1/d_k$ and $d_k = 1 + \lambda x_k/\nu_k^2$. The inclusion probabilities are chosen so that the BLU estimator is equal to the Horvitz-Thomposon estimator which is far from being the best strategy.

Two particular cases of equation (3.11) are:

a) When $\nu_k^2 \propto x_k$, under the condition $\sum_{k\in U} \nu_k = N$, then

$$\nu_k^2 = \frac{N^2 x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2}.$$

In this case, equation (3.11) reduces to

$$\frac{\sum_{k\in \bar{S}} x_k^j}{\sum_{k\in \bar{S}} x_k} = \frac{\sum_{k\in S} x_k^j}{\sum_{k\in S} x_k}, \text{ for } j = 0, \ldots, J.$$

Thus, the sample should satisfy the condition

$$\frac{1}{n}\sum_{k\in S} x_k^j = \frac{1}{N-n}\sum_{k\in \bar{S}} x_k^j = \frac{1}{N}\sum_{k\in U} x_k^j, \text{ for } j = 0, \ldots, J.$$

Royall and Herson (1973a) call samples satisfying this condition *balanced*. In this

case, $\widehat{Y}_0$ reduces to the ordinary ratio estimator

$$\widehat{Y}_{\mathrm{R}} = \sum_{k\in U} x_k \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k}.$$

b) When $\nu_k^2 \propto x_k^2$, it is easily shown that $\nu_k^2 = N^2 x_k^2 / \left(\sum_{k\in U} x_k\right)^2$. In this case, equation (3.11) reduces to

$$\sum_{k\in S} x_k^{j-1}/n = \sum_{k\in \bar{S}} x_k^j / \sum_{k\in \bar{S}} x_k, \text{ for } j = 0, \ldots, J.$$

The sample $S^*(J)$ is called *overbalanced* (Scott et al., 1978) and $\widehat{Y}_0$ reduces to

$$\widehat{Y}_{\mathrm{OB}} = \sum_{k\in S} y_k + \left(\frac{1}{n}\sum_{k\in S} y_k/x_k\right) \sum_{k\in \bar{S}} x_k.$$

### 3.6.3 An alternative solution for the polynomial model

STRATEGY 3.5. *Consider the strategy that consists of*

- *Using inclusion probabilities that are proportional to $\nu_k$ subject to :*

$$\sum_{k\in U} \pi_k = n, \ 0 \leq \pi_k \leq 1,$$

- *selecting a balanced sample according to the following balancing equations :*

$$\sum_{k\in S} \frac{x_k^j}{\pi_k} = \sum_{k\in U} x_k^j, j = 0, \ldots, J, \tag{3.13}$$

- *and using the Horvitz-Thompson estimator.*

Two particular cases of this strategy are :

a) When $\nu_k^2 \propto x_k$, with $\sum_{k\in U} \nu_k = N$, then

$$\nu_k^2 = \frac{N^2 x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2}.$$

As $\pi_k \propto \nu_k$, with $\sum_{k\in U} \pi_k = n$, it follows that $\pi_k = (n/N)\nu_k$.

b) When $\nu_k^2 \propto x_k^2$, with $\sum_{k\in U} \nu_k = N$, then

$$\nu_k^2 = \frac{N^2 x_k^2}{\left(\sum_{k\in U} x_k\right)^2}.$$

Here too, we have $\pi_k = (n/N)\nu_k$.

This strategy is better than the solution proposed by Scott et al. (1978), because the inclusion probabilities are optimal, while for the strategy of Scott et al. (1978), the inclusion probabilities are chosen so that the BLU estimator is equal to the Horvitz-Thompson estimator.

### 3.6.4 A particular case: the ratio model

Consider again the model without intercept and with only one regressor, model (3.5). We have seen in Section 3 that the BLU estimator under this model is the ordinary ratio estimator

$$\widehat{Y}_{\mathrm{R}} = \sum_{k\in U} x_k \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k},$$

with error variance

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2} \left( \frac{\sum_{k\in \bar{S}} x_k}{\sum_{k\in S} x_k} \sum_{k\in U} x_k \right).$$

With a mean-balanced sample, satisfying the condition

$$\frac{1}{n}\sum_{k\in S} x_k = \frac{1}{N-n}\sum_{k\in \bar{S}} x_k,$$

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2(N-n)}{n} \frac{\sum_{k\in U} x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2} = \mathrm{E}_p \mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2.$$

Strategy 3.6. *Consider the strategy that consists of selecting a mean-balanced sample of size $n$ that minimizes $\mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2$ under model (3.5).*

In order to show that Strategy (3.5) is better than Strategy (3.6), we will compare the anticipated variance, defined by $\mathrm{E}_p \mathrm{E_M}(\widehat{Y} - Y)^2$, of the ratio and Horvitz-Thompson estimators under model (3.5).

Nedyalkova and Tillé (2008) have shown that under balanced sampling, and under

model (3.5), the anticipated variance of the Horvitz-Thompson estimator is

$$\mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2 = \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k}{\pi_k},$$

which under this model, with $\pi_k = n/N\nu_k$, gives:

$$\mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2 = \sigma^2 \left( \frac{N}{n} \sum_{k\in U} \nu_k - \sum_{k\in U} \nu_k^2 \right).$$

Finally, after replacing $\nu_k$ with $N\sqrt{x_k}/\sum_{k\in U}\sqrt{x_k}$, we obtain

$$\mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2 = \sigma^2 \left[ \frac{N^2}{n} - \frac{N^2 \sum_{k\in U} x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2} \right].$$

Let denote $D = \mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\mathrm{R} - Y)^2 - \mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2$. After simplifying, we obtain

$$D = \frac{N^2}{n} \left[ \frac{N \sum_{k\in U} x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2} - 1 \right] \geq 0.$$

Thus, strategy (3.5) is better than strategy (3.6) under model (3.5).

## 3.7 Discussion

A d-balanced sampling design with the d-weighted estimator is a bias robust strategy that assures protection against misspecification of the model. The d-weighted estimator can be equivalent to the BLU estimator if some technical conditions are met. These conditions can be met by either choosing the *ad hoc* inclusion probabilities or adding $\nu_k$ and $\nu_k^2$ to the list of balancing variables.

For the polynomial model, Scott et al. (1978) used ad hoc inclusion probabilities, but this strategy is not admissible in the sense where it is always possible to have a smaller anticipated variance by selecting the units with inclusion probabilities proportional to the standard deviations of the errors of the model and using the Horvitz-Thompson estimator, which is the strategy that we advocated in Nedyalkova and Tillé (2008).

# Chapter 4

# Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata

**Abstract:**

The aim of sampling coordination is to maximize or minimize the overlap between several samples drawn successively in a population that changes over time. Therefore, the selection of a new sample will depend on the samples previously drawn. In order to obtain a larger (or smaller) overlap of the samples than the one obtained by independent selection of samples, a dependence between the samples must be introduced. This dependence will emphasize (or limit) the number of common units in the selected samples. Several methods for coordinating stratified samples, such as the Kish & Scott method, the Cotton & Hesse method, and the Rivière method, have already been developed. Using simulations, we compare the optimality of these methods and their quality of coordination. We present six new methods based on Permanent Random Numbers (PRNs) and microstrata. These new methods have the advantage of allowing us to choose between positive or negative coordination with each of the previous samples. Simulations are run to test the validity of each of them.

[1]This chapter is a reprint of the paper: D. Nedyalkova, J. Péa and Y. Tillé. Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata. *ISR*, 76:368-386, 2008.

## 4.1 Introduction

The coordination problem has been a main topic of interest for many years. We distinguish two main types of coordination: negative and positive. In negative coordination, we want to minimize the number of common units between several samples drawn successively in a population that changes over time, while, in positive coordination, we want to maximize this number. The first papers on coordination were written by Patterson (1950) and Keyfitz (1951). The first works on coordination present methods which are in general restricted to two successive samples or small sample sizes. At a later stage, Kish and Scott (1971) generalized the coordination problem in the context of a larger sample size.

The concept of coordination based on PRNs was introduced by Brewer et al. (1972). Most of the national bureaus of statistics use variations of methods based on PRN sampling. Ohlsson (1995) presented a summary of the methods used in different countries. Another approach that takes into account the concept of PRNs, called *order sampling*, was proposed by Rosén (1997a,b).

The coordination of stratified samples is a more complex problem. The main reason is that, over time, units usually change from one stratum to another. Several methods, the Kish & Scott method presented in Kish and Scott (1971), the Cotton & Hesse method presented in Cotton and Hesse (1992b), the Dutch method (EDS) described in De Ree (1983), Van Huis et al. (1994a,b), Koeijers and Willeboordse (1995), and the Rivière method presented in Rivière (1998, 1999, 2001a,b), have already been developed in order to obtain maximal or minimal coverage between samples drawn on different occasions. However, the Dutch method is not of much interest to us because it does not allow strata to be changed.

The methods that we will introduce are based on the use of PRNs and microstrata. They allow us to choose between negative and positive coordination with the previous waves, which is a major advantage. To do positive coordination, we should just coordinate negatively with the complement of the sample. In some of the methods, the PRNs are permuted in a chronological manner, according to what happened at the previous stages, while in others the PRNs are permuted in a retrospective manner. To illustrate the advantages and drawbacks of each one of these methods, simulations have been run.

This paper is structured as follows: some basic notions and definitions are given in Sections 2 and 3. Section 4 introduces the Kish and Scott method. Section 5 presents the Cotton and Hesse method. A comparison of the two methods is given in Section 6. Section 7 presents the Rivière method. Section 8 is devoted to the new methods that we introduce. Section 9 presents the simulation results. Finally, in section 10, a few concluding remarks are given.

## 4.2 Population, Sample, and Sampling Design

We define a finite population as a set of $N$ units $\{u_1, \ldots, u_k, \ldots, u_N\}$. Each unit can be identified without ambiguity by a label. Let $U = \{1, \ldots, k, \ldots, N\}$ be the set of these labels. The size $N$ of the population is not necessarily known. In the problems of sampling coordination, the population can change over time. Suppose that we are interested in studying a population at times $t = 1, 2, \ldots, T-1, T$. Let $U^t$ denote the population at time $t$. The set $U^t \backslash U^{t-1}$ contains the births at time $t$. The set $U^{t-1} \backslash U^t$ holds the deaths at time $t$. The population $U$ contains all the units from time 1 to T

$$U = \bigcup_{t=1}^{T} U^t.$$

DEFINITION 4.1. *At time $t$, a sample without replacement is a subset of the population $U^t$. Since $U^t \subset U$, a sample is also a subset of $U$. The sample is denoted by a vector*

$$\mathbf{s}^t = (s_1^t, \ldots, s_k^t, \ldots, s_N^t)' \in \{0, 1\}^N,$$

*where*

$$s_k^t = \begin{cases} 1 & \textit{if, at time } t, \textit{ unit } k \textit{ is in the sample} \\ 0 & \textit{if, at time } t, \textit{ unit } k \textit{ is not in the sample,} \end{cases}$$

*for all $k \in U$.*

The joint sampling design, $p(\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T)$, is a probability distribution for all the occasions. Let $\mathbf{S}^1, \ldots, \mathbf{S}^t, \ldots, \mathbf{S}^T$ denote the random samples as follows:

$$\Pr(\mathbf{S}^1 = \mathbf{s}^1, \ldots, \mathbf{S}^t = \mathbf{s}^t, \ldots, \mathbf{S}^T = \mathbf{s}^T) = p(\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T).$$

The size of $\mathbf{S}^t$ is denoted by $n^t$.

From the joint sampling design, one can derive the marginal design for the particular time $t$:

$$\sum_{\mathbf{s}^1,\ldots,\mathbf{s}^{t-1},\mathbf{s}^{t+1},\ldots,\mathbf{s}^T} p(\mathbf{s}^1,\ldots,\mathbf{s}^t,\ldots,\mathbf{s}^T) = p_t(\mathbf{s}^t).$$

At time $t$, the first-order inclusion probability and the joint inclusion probability are denoted, respectively, by

$$\pi_k^t = \mathrm{E}(S_k^t) \quad \text{and} \quad \pi_{k\ell}^t = \mathrm{E}(S_k^t S_\ell^t),$$

where $k, \ell \in U^t, t = 1, \ldots, T$. The longitudinal inclusion probability, for times $t$ and $u$, is given by

$$\pi_k^{tu} = \mathrm{E}(S_k^t S_k^u),\ k \in U^t \cap U^u, t, u = 1, \ldots, T.$$

Finally, the joint longitudinal probability has the form:

$$\pi_{k\ell}^{tu} = \mathrm{E}(S_k^t S_\ell^u),\ k, \ell \in S^t \cup S^u, t, u = 1, \ldots, T.$$

Note that this probability is not symmetrical. Indeed, $\pi_{k\ell}^{tu} \neq \pi_{\ell k}^{tu}$ and $\pi_{k\ell}^{tu} \neq \pi_{\ell k}^{ut}$.

The following basic result gives bounds for the longitudinal inclusion probabilities.

RESULT 4.1. *For times $t$ and $u$, we have*

$$\max(0, \pi_k^t + \pi_k^u - 1) \leq \pi_k^{tu} \leq \min(\pi_k^t, \pi_k^u).$$

*Proof.* By definition, we have

$$\pi_k^{tu} = \Pr(S_k^t = 1 \text{ and } S_k^u = 1) \leq \min\left[\Pr(S_k^t = 1), \Pr(S_k^u = 1)\right] = \min\left(\pi_k^t, \pi_k^u\right).$$

Moreover,

$$\begin{aligned} \pi_k^t - \pi_k^{tu} &= \Pr(S_k^t = 1) - \Pr(S_k^t = 1 \text{ and } S_k^u = 1) = \Pr(S_k^t = 1 \text{ and } S_k^u = 0) \\ &\leq \min\left[\Pr(S_k^t = 1), \Pr(S_k^u = 0)\right] = \min\left(\pi_k^t, 1 - \pi_k^u\right). \end{aligned}$$

Thus,

$$\pi_k^{tu} \geq \pi_k^t - \min\left(\pi_k^t, 1 - \pi_k^u\right) = \max\left(0, \pi_k^t + \pi_k^u - 1\right).$$

□

Consider a population $U$ split into $H$ parts $U_h$, called 'strata', such that

$$\cup_{h=1}^{H} U_h = U \quad \text{and} \quad U_h \cap U_i = \emptyset,$$

for all $(h, i)$ with $h \neq i$. A design is called stratified if a random sample $S_h$ of fixed size $n_h$ is selected in each stratum $U_h$, and if the sample selection in each stratum is taken independently of the selection done in all the other strata.

## 4.3 Sample coordination, Overlap, and Burden

DEFINITION 4.2. *The overlap is the number of common units at two different times t and u:*

$$n^{tu} = \sum_{k \in U} S_k^t S_k^u.$$

*The overlap can be random. The expected overlap is*

$$\mathrm{E}(n^{tu}) = \mathrm{E}\left(\sum_{k \in U} S_k^t S_k^u\right) = \sum_{k \in U} \mathrm{E}\left(S_k^t S_k^u\right) = \sum_{k \in U} \pi_k^{tu}.$$

DEFINITION 4.3. *The overlap rate is defined by:*

$$v^{tu} = \frac{2n^{tu}}{n^t + n^u}.$$

*If $n^t$ and $n^u$ are fixed, then the expected overlap rate is given by:*

$$\tau^{tu} = \frac{2\mathrm{E}(n^{tu})}{n^t + n^u}.$$

Let $\mathrm{ALB} = \sum_{k \in U} \max(0, \pi_k^t + \pi_k^u - 1)$ denote the absolute lower bound and $\mathrm{AUB} = \sum_{k \in U} \min(\pi_k^t, \pi_k^u)$ denote the absolute upper bound (see Matei and Tillé, 2005). Then, from result 4.1, we can directly derive bounds for the expected overlap:

$$\mathrm{ALB} \leq \mathrm{E}(n^{tu}) \leq \mathrm{AUB}.$$

Unfortunately, except for very particular cases like simple random sampling (SRS), the ALB and AUB cannot be reached.

If, at times 1 and 2, two samples are drawn independently without coordination, then, for all $k \in U$,

$$\pi_k^1 \pi_k^2 = \pi_k^{12}.$$

In positive coordination, for all $k \in U$, the longitudinal inclusion probability must satisfy the conditions

$$\pi_k^1 \pi_k^2 \leq \pi_k^{12} \leq \min(\pi_k^1, \pi_k^2).$$

In negative coordination, for all $k \in U$, the longitudinal inclusion probability must satisfy the conditions

$$\max(0, \pi_k^1 + \pi_k^2 - 1) \leq \pi_k^{12} \leq \pi_k^1 \pi_k^2.$$

Note that, in the last case, the longitudinal inclusion probability can be zero only if $\pi_k^1 + \pi_k^2 \leq 1$.

The response burden of a survey is usually quantified in terms of the time needed to complete the questionnaire. However, other aspects of response burden exist: for example, how difficult it is to provide the information or how sensitive the question sent to the respondent is. Therefore, the response burden can vary from one survey to another.

At time $t$, a survey has a burden denoted by $b^t$, which can be proportional to the time needed to complete the form or can be simply equal to one. After $T$ waves, the total burden of unit $k$ is defined as the sum of the burdens of the surveys in which unit $k$ has been included:

$$c_k^T = \sum_{t=1}^{T} b^t S_k^t.$$

We also define the cumulative burden from survey $m$ to survey $T$, named $(m, T)$-cumulated burden, as:

$$c_k^{m,T} = \sum_{t=m}^{T} b^t S_k^t.$$

The quality of a procedure concerning coordination can be measured using four possible criteria:

1. the procedure provides a controllable degree of overlap;

2. the sampling design is respected in each selection;

3. for each unit, a fixed time out of sample is respected;

4. the procedure is computed easily.

## 4.4 The Kish & Scott Method

Kish and Scott (1971) have proposed a method of substitution for coordinating stratified samples which allows changes in the definition of the strata. Although they had introduced this method for positive coordination, presented in Algorithm 1, it also allows us to do negative coordination. At times 1 and 2, i.e. waves 1 and 2, the definition of the strata can change. From this point forward, we will use the terms times and waves interchangeably.

In order to present a rigorous algorithm, it is necessary to formalize the notation. We also assume that there are no births and deaths in the population. Suppose that the population $U$ is stratified at time 1 into $H$ strata $U_1^1, \ldots, U_h^1, \ldots, U_H^1$, and at time 2 into $G$ strata $U_1^2, \ldots, U_g^2, \ldots, U_G^2$ as follows:

$$U = \cup_{h=1}^{H} U_h^1 = \cup_{g=1}^{G} U_g^2.$$

Let $N_h^1$ be the size of $U_h^1$, $N_g^2$ the size of $U_g^2$, and $N_{hg}^{12}$ the size of $U_{hg}^{12} = U_h^1 \cap U_g^2$. Suppose that two independent stratified samples $\mathbf{s}^1$ and $\mathbf{s}^2$ are drawn, at time 1 and time 2, respectively. Also consider the following notations:

- $s_g^i$ the set of units of stratum $U_g^2$ that are selected in $\mathbf{s}^i$, for $i = 1, 2$, with $n_g^i = \text{card}(\mathbf{s}_g^i)$,
- $s_{hg}^i$ the set of units of $U_h^1 \cap U_g^2$ that are selected in $\mathbf{s}^i$, for $i = 1, 2$, with $n_{hg}^i = \text{card}(\mathbf{s}_{hg}^i)$,
- $s_{hg}^{12} = s_{hg}^1 \cap s_{hg}^2$, with $n_{hg}^{12} = \text{card}(\mathbf{s}_{hg}^{12})$.

This method is correct because it provides two conditional stratified samples. Nevertheless, only two waves can be coordinated because the coordination of more than two samples becomes very complicated. Simulations show that the coordination is not very good. Generally, the Cotton & Hesse method performs better which we will show in a simulation example in Section 6 of our paper.

## 4.5 The Cotton & Hesse Method

The Cotton & Hesse method from the Institut National de la Statistique et des Études Économiques (INSEE) of France is fully described in Cotton and Hesse (1992b). This

**Algorithm 1** Positive Coordination using the Kish & Scott Method.

1: At wave 1, draw a stratified sample from $U$, denoted by $\mathbf{s}^1$.
2: At wave 2, draw a stratified sample from $U$, denoted by $\mathbf{s}^2$ and
3: **for** Each possible intersection of strata $U_{hg}^{12}$ **do**
4: **if** $n_{hg}^2 - n_{hg}^{12} \geq n_{hg}^1 - n_{hg}^{12}$ **then**
5: Replace $n_{hg}^1 - n_{hg}^{12}$ units from $s_{hg}^2 \backslash s_{hg}^{12}$ with units of $s_{hg}^1 \backslash s_{hg}^{12}$ by means of SRS.
6: **else**
7: Replace $n_{hg}^2 - n_{hg}^{12}$ units from $s_{hg}^2 \backslash s_{hg}^{12}$ with units of $s_{hg}^1 \backslash s_{hg}^{12}$ by means of SRS.
8: **end if**
9: **end for**

method works when the strata change over time and can be used to obtain negative coordination. The principle is as follows: Each unit of the population receives a PRN $\omega_k$ from a uniform distribution $U[0, 1]$. At the first wave, the sample is defined, in each stratum, as the set of units that have the smallest random numbers. After the sample has been selected, the PRNs are permuted in such a way that the units selected at the first wave receive the largest PRNs, and the non-selected units receive the smallest PRNs. Within the two subsets of selected and non-selected units, the order of the permuted PRNs must remain unchanged. Then the same procedure is applied for the subsequent waves. The procedure for negative coordination is presented in Algorithm 2. Note that the dead units lose their PRNs, while the new units receive a new PRN.

**Algorithm 2** Negative Coordination using the Cotton & Hesse Method.

1: Assign, independently, a PRN $\omega_k^1$ to each unit $k \in U$ and construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Select the units that have the $n_h^t$ smallest $\omega_k^t$ to obtain the sample $\mathbf{s}^t$.
4: Assign the largest $\omega_k^t$ to the units that belong to $\mathbf{s}^t$.
5: Assign the smallest $\omega_k^t$ to the units that belong to $U \setminus \mathbf{s}^t$.
6: Construct $\boldsymbol{\omega}^{t+1}$ as a permutation of $\boldsymbol{\omega}^t$ so that the rank of $\boldsymbol{\omega}^t$ in $\mathbf{s}^t$ and in $U \setminus \mathbf{s}^t$ is respected.
7: **end for**

The major advantage of this method is that the strata can change over time. The method is correct, because after the permutation, the PRNs remain independent uniform random numbers; the method is thus very simple to apply. Another advantage of this method is that only the permuted PRNs must be retained from one wave to another. The drawback of this method is that it allows only one kind of coordination. Once you have

decided to do negative coordination between two time periods, you cannot do positive coordination while drawing another sample. Moreover, the order of the surveys is fixed and cannot be changed.

## 4.6 Comparison of the Kish & Scott and Cotton & Hesse Methods

We will consider a very simple example in order to compare both methods. Since each problem of coordination can be also viewed as a problem of optimization, we can find the optimal solution and then compare this solution to the solutions of the Kish & Scott and Cotton & Hesse methods. We will see that neither method is optimal.

Suppose that the population $U = \{1, 2, 3, 4\}$. At time 1, the strata are $\{1, 2\}, \{3, 4\}$, and at time 2, $\{1, 3\}, \{2, 4\}$. At times 1 and 2, two stratified samples are selected with only one unit in each stratum. The aim is to obtain the best negative coordination. At time 1, the possible samples are given by:

$$\mathbf{s}_1^1 = \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}', \quad \mathbf{s}_2^1 = \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}', \quad \mathbf{s}_3^1 = \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}', \quad \mathbf{s}_4^1 = \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}'.$$

All samples are selected with probability $p(s_i^1) = 1/4$, $i = 1, \ldots, 4$ and $\pi_k^1 = 1/2$ for all $k \in U$.

At time 2, the possible samples are

$$\mathbf{s}_1^2 = \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix}', \quad \mathbf{s}_2^2 = \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}', \quad \mathbf{s}_3^2 = \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}', \quad \mathbf{s}_4^2 = \begin{pmatrix} 0 & 0 & 1 & 1 \end{pmatrix}'.$$

Here, again, all samples are selected with probability $p(s_i^2) = 1/4$, $i = 1, \ldots, 4$ and $\pi_k^2 = 1/2$ for all $k \in U$.

Since $n^1 = 2$ and $n^2 = 2$, then

$$\text{ALB} = \sum_{k \in U} \max(0, \pi_k^1 + \pi_k^2 - 1) = 0.$$

Nevertheless, we will see that the ALB cannot be reached due to the constraints of stratification. The overlap between the different samples is given in Table 4.1.

If $n_{ij}^{12} = \mathbf{s}_i^{1\prime} \mathbf{s}_j^2$ is the number of common units of samples $\mathbf{s}_i^1$ and $\mathbf{s}_j^2$, and $p_{ij} = \Pr(\mathbf{s}_i^1, \mathbf{s}_j^2)$,

Table 4.1: Overlap between the possible samples at times 1 and 2.

| | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ |
|---|---|---|---|---|
| $\mathbf{s}_1^1$ | 1 | 1 | 1 | 1 |
| $\mathbf{s}_2^1$ | 1 | 2 | 0 | 1 |
| $\mathbf{s}_3^1$ | 1 | 0 | 2 | 1 |
| $\mathbf{s}_4^1$ | 1 | 1 | 1 | 1 |

the optimal solution is obtained by solving

$$\arg\min_{p_{ij}} \sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij}^{12} p_{ij} \quad \text{subject to} \quad \begin{cases} p_{ij} > 0, \\ \sum_{i=1}^{4} p_{ij} = 1/4, \text{ for } j = 1, \ldots, 4, \\ \sum_{j=1}^{4} p_{ij} = 1/4, \text{ for } i = 1, \ldots, 4. \end{cases}$$

This is a linear program, that can be solved easily. The set of optimal solutions is given in Table 4.2.

Table 4.2: Set of optimal solutions for $c \in [0, 1/4]$.

| | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ |
|---|---|---|---|---|
| $\mathbf{s}_1^1$ | $1/4 - c$ | 0 | 0 | $c$ |
| $\mathbf{s}_2^1$ | 0 | 0 | $1/4$ | 0 |
| $\mathbf{s}_3^1$ | 0 | $1/4$ | 0 | 0 |
| $\mathbf{s}_4^1$ | $c$ | 0 | 0 | $1/4 - c$ |

If we take $c = 0$, we find one of the following optimal solutions:

- If, at time 1, $\mathbf{s}_1^1$ is selected, then select $\mathbf{s}_1^2$ at time 2.
- If, at time 1, $\mathbf{s}_2^1$ is selected, then select $\mathbf{s}_3^2$ at time 2.
- If, at time 1, $\mathbf{s}_3^1$ is selected, then select $\mathbf{s}_2^2$ at time 2.
- If, at time 1, $\mathbf{s}_4^1$ is selected, then select $\mathbf{s}_4^2$ at time 2.

The expected overlap is $\mathrm{E}(n^{12}) = 0.5$, and the expected overlap rate is

$$\tau^{12} = \frac{2\mathrm{E}(n^{12})}{n^1 + n^2} = \frac{2 \times 0.5}{2 + 2} = 0.25.$$

The *relative lower bound* (RLB) (see Matei and Tillé, 2005) is the value of the objective function when the linear problem is solved. The following relation holds:

$$\text{RLB} = \arg\min_{p_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{q} n_{ij}^{12} p_{ij} \geq \sum_{k \in U} \max(0, \pi_k^1 + \pi_k^2 - 1) = \text{ALB}.$$

Here, RLB = 0.5 is strictly larger than ALB = 0.

If we apply a negative coordination using the Kish & Scott method to the strata we have defined above, then we obtain the following results, given in Table 4.3. Note that,

Table 4.3: Negative Coordination with the Kish & Scott Method.

| Time 1 | $\mathbf{s}_1^1$ | $\mathbf{s}_1^1$ | $\mathbf{s}_1^1$ | $\mathbf{s}_1^1$ | $\mathbf{s}_2^1$ | $\mathbf{s}_2^1$ | $\mathbf{s}_2^1$ | $\mathbf{s}_2^1$ | $\mathbf{s}_3^1$ | $\mathbf{s}_3^1$ | $\mathbf{s}_3^1$ | $\mathbf{s}_3^1$ | $\mathbf{s}_4^1$ | $\mathbf{s}_4^1$ | $\mathbf{s}_4^1$ | $\mathbf{s}_4^1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time 2 | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ | $\mathbf{s}_1^2$ | $\mathbf{s}_2^2$ | $\mathbf{s}_3^2$ | $\mathbf{s}_4^2$ |
| Overlap | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 1 |

the expected overlap is 1 and the expected overlap rate is

$$\tau^{12} = \frac{2\text{E}(n^{12})}{n^1 + n^2} = \frac{2 \times 1}{2 + 2} = \frac{1}{2}.$$

In this case, the quality of coordination is not better than for independent stratified samples.

If we apply a negative coordination using the Cotton & Hesse method, then we obtain the following results, presented in Table 4.4.

Note that, in this case, the expected overlap is 2/3 and the expected overlap rate is

$$\tau_{12} = \frac{2\text{E}(n^{12})}{n^1 + n^2} = \frac{2 \times 2/3}{2 + 2} = \frac{1}{3}.$$

Thus, the optimal solution, which has an expected overlap of 1/2, is not reached. The Cotton & Hesse method does not provide the best solution to the problem of coordination.

This small example is very interesting because it shows that the optimality is not reached by either the Cotton & Hesse method or the Kish & Scott method. However, the Cotton & Hesse method gives a slightly better solution than the Kish & Scott method. This result was also confirmed by a set of simulations. Therefore, we advocate the use of the Cotton & Hesse method rather than the Kish & Scott method provided that the drawbacks of the Cotton & Hesse method are not an issue.

Table 4.4: Negative Coordination with the Cotton & Hesse Method.

| Ranks | Time 1 | Permuted Ranks | Time 2 | Overlap |
|---|---|---|---|---|
| (1 2 3 4) | (1 0 1 0) | (2 1 4 3) | (1 1 0 0) | 1 |
| (1 2 4 3) | (1 0 0 1) | (2 1 3 4) | (1 1 0 0) | 1 |
| (1 3 2 4) | (1 0 1 0) | (3 1 4 2) | (1 1 0 0) | 1 |
| (1 3 4 2) | (1 0 0 1) | (3 1 2 4) | (0 1 1 0) | 0 |
| (1 4 2 3) | (1 0 1 0) | (4 1 3 2) | (0 1 1 0) | 1 |
| (1 4 3 2) | (1 0 0 1) | (4 1 2 3) | (0 1 1 0) | 0 |
| (2 1 3 4) | (0 1 1 0) | (1 2 4 3) | (1 1 0 0) | 1 |
| (2 1 4 3) | (0 1 0 1) | (1 2 3 4) | (1 1 0 0) | 1 |
| (2 3 1 4) | (1 0 1 0) | (3 2 4 1) | (1 0 0 1) | 1 |
| (2 3 4 1) | (1 0 0 1) | (3 2 1 4) | (0 1 1 0) | 0 |
| (2 4 1 3) | (1 0 1 0) | (4 2 3 1) | (0 0 1 1) | 1 |
| (2 4 3 1) | (1 0 0 1) | (4 2 1 3) | (0 1 1 0) | 0 |
| (3 1 2 4) | (0 1 1 0) | (1 3 4 2) | (1 0 0 1) | 0 |
| (3 1 4 2) | (0 1 0 1) | (1 3 2 4) | (1 1 0 0) | 1 |
| (3 2 1 4) | (0 1 1 0) | (2 3 4 1) | (1 0 0 1) | 0 |
| (3 2 4 1) | (0 1 0 1) | (2 3 1 4) | (0 1 1 0) | 1 |
| (3 4 1 2) | (1 0 1 0) | (4 3 2 1) | (0 0 1 1) | 1 |
| (3 4 2 1) | (1 0 0 1) | (4 3 1 2) | (0 0 1 1) | 1 |
| (4 1 2 3) | (0 1 1 0) | (1 4 3 2) | (1 0 0 1) | 0 |
| (4 1 3 2) | (0 1 0 1) | (1 4 2 3) | (1 0 0 1) | 1 |
| (4 2 1 3) | (0 1 1 0) | (2 4 3 1) | (1 0 0 1) | 0 |
| (4 2 3 1) | (0 1 0 1) | (2 4 1 3) | (0 0 1 1) | 1 |
| (4 3 1 2) | (0 1 1 0) | (3 4 2 1) | (0 0 1 1) | 1 |
| (4 3 2 1) | (0 1 0 1) | (3 4 1 2) | (0 0 1 1) | 1 |

## 4.7 The Rivière Method

This method, based on the use of microstrata, was proposed in a large set of publications of Rivière (1998, 1999, 2001a,b) under the framework of the 1996 SUPCOM project of Eurostat. As a result, two software applications were developed: SALOMON in 1998 (see Mészáros, 1999) and MICROSTRAT in 2001. The method is based on four basic ideas:

- the use of PRNs that are allocated to each statistical unit,
- the use of a measure of burden, which can be the number of times that a unit has already been selected for all the waves that one wants to coordinate,

- the use of microstrata constructed at each wave by intersecting all the strata of the waves that one wants to coordinate,
- the permutation of the PRNs in proportion to the measure of burden within the microstrata so that the units with the smallest measures of burden obtain the smallest random numbers.

As a preliminary to the algorithm, each unit receives a PRN from the uniform distribution on [0,1] and a response burden equal to 0. Also note that for the Rivière method, if $t$ is the first wave that we want to coordinate, a microstratum, at wave $T$, is defined by the intersection of the strata of waves $t$ to $T-1$. The permutations are done within each microstratum according to the cumulative burden. Note that, within the subsets of equal burden, the order of the permuted PRNs must remain unchanged. Then, one can apply Algorithm 3 to do coordination using the Rivière method. A proof of the validity of

**Algorithm 3** Negative coordination with The Rivière Method.

1: Assign a PRN $\omega_k^1$ to each unit $k \in U$, i.e. construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: Assign a burden equal to 0 to each unit $k \in U$, i.e. $c_k^1 = 0$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4: Compute the burden $c_k^T = \sum_{t=1}^{T-1} b^t S_k^t$.
5: Construct the microstrata by crossing the strata of waves 1 to $T-1$.
6: Permute the $\omega_k^1$, in each microstratum, so that the units are sorted by increasing burden and the ranks remain unchanged in the subsets of equal burden.
7: Select the first $n_h^T$ units in each stratum.
8: **end for**

the method has been given by Bleuer (2002). Nevertheless, if the algorithm is carried out just once, the procedure has a main drawback: by crossing the strata of all the previous surveys, the microstrata become very small and thus the coordination is not very good. For this reason, and in order to have a good coordination with the last surveys, Rivière (1999, p.5) advocated the use of only three sorts.

We ran a large set of simulations, trying to invalidate the method. We were sceptical about the use of multiple sorting. Nevertheless, after 4 waves, the method seems to provide the right inclusion probabilities of order 1 and 2, as long as the permutations are done in a strictly sequential manner. For instance, at wave $T$, if we want to coordinate with respect to wave $t$, the microstrata must be obtained by crossing all the strata of all the waves between $t$ and $T-1$. It is not possible to skip a wave, otherwise the inclusion probabilities

are not satisfied, as shown in the following example.

*Example* 4.1. Suppose that we have 4 waves during the first year. The population is of size $N = 16$. At each wave, we have two strata of size $N_h^t = 8$. The burden is equal to 1 for each survey. The strata are defined in Table 4.5.

Table 4.5: Definition of strata used for the simulations

| Units | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Strata$^1$ | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Strata$^2$ | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| Strata$^3$ | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| Strata$^4$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

The sample strata sizes are $n_1^1 = 3, n_2^1 = 5, n_1^2 = 6, n_2^2 = 2, n_1^3 = 4, n_2^3 = 4, n^4 = 6$. If we apply the procedure of Salomon, the samples are coordinated as follows.

- For the initialization, uniform random numbers $\omega_k$ are generated for each unit.
- At wave 1, the units that have the smallest $\omega_k$ in each stratum are selected.
- At wave 2, the $\omega_k$ are permuted according to the burden in the strata of wave 1 for obtaining $\omega_k^2$. Next, the units that have the smallest $\omega_k^2$ in each stratum are selected.
- At wave 3, the $\omega_k^2$ are permuted according to the burden in the strata of wave 2 to obtain $\omega_k^{3a}$. Next, the $\omega_k^{3a}$ are permuted according to the burden in the crossing of the strata of wave 1 and 2, to obtain $\omega_k^{3b}$. Finally, the units that have the smallest $\omega_k^{3b}$ in each stratum are selected.
- At wave 4, the $\omega_k^{3b}$ are permuted in function of the burden in the strata of wave 3 for obtaining $\omega_k^{4a}$. Next, the $\omega_k^{4a}$ are permuted according to the burden in the crossing of the strata of wave 1 and 3, (**wave 2 is skipped**) to obtain $\omega_k^{4b}$. Finally, the units that have the smallest $\omega_k^{4b}$ in each stratum are selected.

However, this procedure is not recommended by Rivière. The permutation is done in relation to the previous survey and in relation to all the surveys since the beginning of the year.

After 10000 simulations, we estimated the inclusion probabilities by $\widetilde{\pi}_k^t$. Then we computed

$$z_k^4 = \frac{\widetilde{\pi}_k^4 - \pi_k^4}{\sqrt{\pi_k^4(1-\pi_k^4)/sim}},$$

where $\pi_k^4$ are the inclusion probabilities that we want to obtain, and $sim$ is the number of simulations. If the method provides good inclusion probabilities, then the $z_k$ should have a Normal distribution. We obtained the following vector of $z_k^4$ :

$$(1.033, -6.383, -1.611, 8.138, 1.012, -8.076, -2.520, 7.333,$$

$$2.520, -7.415, -1.384, 6.403, 1.632, -8.882, -0.909, 9.109).$$

Several $z_k^4$ are larger than 1.96 in absolute value. We must therefore reject the hypothesis that the inclusion probabilities are correct.

This example does not show that the method is false. It only shows that the permutation must be done in a strictly sequential manner and that a wave should never be skipped. We ran a set of simulations in the same population as in Example 4.1 in order to compare the quality of the coordination of the Rivière and Cotton & Hesse methods. The results which we do not present here clearly showed that both methods give almost equivalent results.

## 4.8 Other Methods Using Microstrata

In this section, we introduce several new methods based on the idea of microstrata. A method of coordination must be evaluated through several waves, and it is a complicated matter to theoretically prove that a method works or not when we have a large number of waves. Moreover, on a large number of waves and large population and sample sizes, the methods seem to give equivalent results. So, in order to invalidate our methods and point out the differences between them, we decided to run simulations on 4 waves. Obviously, if we cannot prove by simulation that a method is false, it does not imply that the method works. From this point forward, we will refer to a method as sim-false if it was invalidated by simulations, and sim-correct, if the simulations fail to invalidate it.

Before coming to the idea of applying the microstrata technique for coordination, we had a very simple idea. First, for each wave $T$, we generate a new random number. Next,

these random numbers are permuted within each of the strata of the previous waves. These permutations are done chronologically from wave 1 to wave $T-1$. The method is described precisely in Algorithm 4.

**Algorithm 4** First Method of Chronological Permutations (Sim-false).

1: At wave 1, assign a uniform random number, $\omega_k^1$, to each unit $k \in U$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4: Generate and assign a new uniform random number, $\omega_k^t$, to each unit $k \in U$.
5: **for** $t = 1, \ldots, T-1$ **do**
6: Permute the $\omega_k^t$, within the strata, so that the units that are selected receive the largest random numbers, the units that are not selected receive the smallest random numbers and the ranks remain unchanged in the subsets of selected and non-selected units.
7: **end for**
8: Select the units that have the $n_h^T$ smallest $\omega_k^T$, in each stratum, to obtain the sample $\mathbf{s}^T$.
9: **end for**

This method seems interesting but is sim-false in the sense that it does not provide the sim-correct inclusion probabilities. Simulations on at least 3 waves were needed to detect the problem. Our explanation is the following: at wave 3, the random numbers are permuted according to the first wave, and next according to the second wave. Nevertheless, the selection of the units of the second wave depends on the permuted random numbers from the first wave. This correlation implies that, after the permutation, the random numbers are not independent and uniform anymore. From this sim-false method, we concluded that, in order to coordinate a sample at time $T$, if a permutation of the random numbers is done in the strata of time $t$, the method will be false. The permutations must be done in the crossing of all the strata (the microstrata) from time $t$ to $T-1$.

One of the main differences between the methods we introduce is the order in which the permutations are done. We differentiate two types of order: *chronological* and *retrospective*. Chronological means always starting with the first wave and going on to the next ones. As an example of chronological permutations at wave 4, we have: first, the permutations are done in the crossing of the strata of waves $1, 2$ and 3, after that in the crossing of the strata of waves 2 and 3, and finally, in the crossing of the strata of wave 3. On the other side, a retrospective order means that the permutations are first done in the crossing of the strata of the latest wave and then going backwards to the first wave. As an example

of retrospective permutations at wave 4, we have: first, the permutations are done in the crossing of the strata of wave 3, after that in the crossing of the strata of waves 2 and 3, and finally, in the crossing of the strata of waves $1, 2$ and 3. However, the retrospective order has the small disadvantage that it takes more time to compute the permutations than if they are done chronologically.

In order to overcome the problem posed by Algorithm 4, the permutations could be done in the microstrata as described in Algorithm 5. Note, that the permutations are done in chronological order. We can modify the method described in Algorithm 5 by using PRNs instead of generating a new random number at each wave. This method is described in Algorithm 6. The simulations invalidated both methods, so they are both sim-false. We

**Algorithm 5** Second Method of Chronological Permutations (Sim-false).

1: At wave 1, assign a uniform random number to each unit $k \in U$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4: Assign a new uniform random number to each unit $k \in U$.
5: **for** $t = 1, \ldots, T-1$ **do**
6: Compute the $(t, T-1)$-cumulated burden, i.e. $c_k^{t,T-1} = \sum_{u=t}^{T-1} b^u S_k^u$.
7: Construct the microstrata by crossing the strata of waves $t$ to $T-1$.
8: Permute the $\omega_k^t$, in each microstratum, such that the units are sorted by increasing burden and the ranks remain unchanged in the subsets of equal burden.
9: **end for**
10: Select the units that have the $n_h^T$ smallest $\omega_k^T$ in each stratum to obtain the sample $\mathbf{s}^T$.
11: **end for**

thus concluded that the use of burden in microstrata does not work if the permutations are done in a chronological order.

These conclusions led us to create 6 other methods that could not be invalidated by simulation. We have abandoned the methods based on generating a new random number at each wave due to the simulation results in which these methods do not perform any better than the PRNs methods. Thus, we will present only the three methods based on PRNs. Method 7 is based on the idea of microstrata, cumulative burden and multiple permutations which are, this time, done in a retrospective way. This method can be considered as a modification of the Rivière method. The difference between our method and the Rivière method is that, at a given time $T$, it uses multiple permutations while the Rivière method uses only one permutation done in the microstrata, constructed by

**Algorithm 6** Third Method of Chronological Permutations (Sim-false).

1: Assign a PRN $\omega_k^1$ to each unit $k \in U$, i.e. construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4:  **for** $t = 1, \ldots, T-1$ **do**
5:  Compute the $(t, T-1)$-cumulated burden, i.e. $c_k^{t,T-1} = \sum_{u=t}^{T-1} b^u S_k^u$.
6:  Construct the microstrata by crossing the strata of waves $t$ to $T-1$.
7:  Permute the $\omega_k^t$, in each microstratum, so that the units are sorted by increasing burden and the ranks remain unchanged in the subsets of equal burden.
8:  **end for**
9:  Select the units that have the $n_h^T$ smallest $\omega_k^T$ in each stratum to obtain the sample $\mathbf{s}^T$.
10: **end for**

crossing the strata from waves 1 to $T-1$. Based on the simulation results, this method is sim-correct.

**Algorithm 7** First Method of Retrospective Permutations (Sim-correct).

1: Assign a PRN $\omega_k^1$ to each unit $k \in U$, i.e. construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4:  **for** $t = T-1, \ldots, 1$ **do**
5:  Compute the $(t, T-1)$-cumulated burden, i.e. $c_k^{t,T-1} = \sum_{u=t}^{T-1} b^u S_k^u$.
6:  Construct the microstrata by crossing the strata of waves $t$ to $T-1$.
7:  Permute the $\omega_k^t$, in each microstratum, so that the units are sorted by increasing burden and the ranks remain unchanged in the subsets of equal burden.
8:  **end for**
9:  Select the units that have the $n_h^T$ smallest $\omega_k^T$ in each stratum to obtain the sample $\mathbf{s}^T$.
10: **end for**

The last two methods are presented in Algorithms 8 and 9 and are based on the idea of multiple permutations done in microstrata which are the crossing of the strata of the previous waves and the subsets defined as $\mathbf{s}^t$. In these methods, the cumulative burden is not taken into account. The permutations are done according to the sample indicator variables and not according to the cumulative burden. In Algorithm 8, the permutations are done in a retrospective order, while in Algorithm 9 they are done in a chronological order. Based on the simulation results, these methods are sim-correct.

**Algorithm 8** Second Method of Retrospective Permutations (Sim-correct).

1: Assign a PRN $\omega_k^1$ to each unit $k \in U$, i.e. construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4: **for** $t = T-1, \ldots, 1$ **do**
5: Construct the microstrata as the intersection of the crossing of the strata $t, \ldots, T-1$ and the crossing of the subsets defined by $\mathbf{s}^{t+1}, ..., \mathbf{s}^{T-1}$.
6: Construct $\boldsymbol{\omega}^{t+1}$ by permuting $\boldsymbol{\omega}^t$, within each microstratum, so that the units that are selected receive the largest random numbers, the units that are not selected receive the smallest random numbers and the ranks remain unchanged in the subsets of selected and non-selected units.
7: **end for**
8: Select the units that have the $n_h^T$ smallest $\omega_k^T$ in each stratum to obtain the sample $\mathbf{s}^T$.
9: **end for**

**Algorithm 9** Fourth Method of Chronological Permutations (Sim-correct).

1: Assign a PRN $\omega_k^1$ to each unit $k \in U$, i.e. construct $\boldsymbol{\omega}^1 = \{\omega_1^1, \ldots, \omega_N^1\}$.
2: Select the units that have the $n_h^1$ smallest $\omega_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $T = 2, \ldots,$ Number of Waves **do**
4: **for** $t = 1, \ldots, T-1$ **do**
5: Construct the microstrata as the intersection of the crossing of the strata $t, \ldots, T-1$ and the crossing of the subsets defined by $\mathbf{s}^{t+1}, ..., \mathbf{s}^{T-1}$.
6: Construct $\boldsymbol{\omega}^{t+1}$ by permuting $\boldsymbol{\omega}^t$, within each microstratum, so that the units that are selected receive the largest random numbers, the units that are not selected receive the smallest random numbers and the ranks remain unchanged in the subsets of selected and non-selected units.
7: **end for**
8: Select the units that have the $n_h^T$ smallest $\omega_k^T$ in each stratum to obtain the sample $\mathbf{s}^T$.
9: **end for**

## 4.9 Simulation Study and Results

In this section, we will test the new methods presented in Section 8. We should note that simulations were done in larger sample (population) sizes but in this case all methods were performing well. Thus, in order to find which methods result in false inclusion probabilities, we decided to use small sample (population) sizes. We simulated 500 000 drawings of stratified simple random samples in a population of $N = 16$ units. Four waves were taken

into account. The strata are defined in Table 4.5. The sample strata sizes are

$$n_1^1 = 3, n_2^1 = 5, n_1^2 = 6, n_2^2 = 2, n_1^3 = 4, n_2^3 = 4, n^4 = 6.$$

To compare the results of the simulations, we decided to analyze three different simulation outputs:

1. The first-order inclusion probabilities.
2. The second-order inclusion probabilities.
3. The quality of the coordination.

To analyze the first- and second-order inclusion probabilities, which we denote, respectively, by $\pi_k^{sim}$ and $\pi_{k\ell}^{sim}$, we calculated a kind of $''z$-value$''$, which enables us to do a Normal test on the value obtained by simulation. This '$z$-value' was obtained using the following formula:

For the first-order inclusion probabilities:

$$z_{\pi_k} = \sqrt{sim} \cdot \frac{\pi_k^{sim} - \left(\frac{n_h}{N_h}\right)}{\sqrt{\frac{n_h}{N_h} \cdot \left(1 - \frac{n_h}{N_h}\right)}}$$

For the second-order inclusion probabilities:

$$z_{\pi_{k\ell}} = \sqrt{sim} \cdot \frac{\pi_{k\ell}^{sim} - \left(\frac{n_h^k \cdot n_h^\ell}{N_h^k \cdot N_h^\ell}\right)}{\sqrt{\frac{n_h^k \cdot n_h^\ell}{N_h^k \cdot N_h^\ell} \cdot \left(1 - \frac{n_h^k \cdot n_h^\ell}{N_h^k \cdot N_h^\ell}\right)}}$$

The obtained values are the '$z$-values' for the centered inclusion probabilities. An acceptable value of the centered $\pi_k$ and $\pi_{k\ell}$ should lie in the interval [-2,2] (95% confidence interval).

On each graph, there are four plots corresponding to each of the four waves, as shown in Table 4.6.

On each plot we can see:

- on the horizontal axis - the units of the population,

Table 4.6: Wave number for each plot

| | |
|---|---|
| 1 | 2 |
| 3 | 4 |

- on the vertical axis - the centered '$z$-value',
- the limit values of the confidence interval, given by dashed lines,
- the acceptable values - the circles between the dashed lines,
- the unacceptable values - the circles outside of the confidence interval.

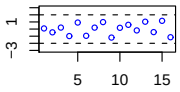
The third output used to compare the methods is the quality of the coordination which is simply given by the average number of common units in the samples.

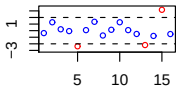
To interpret the results, we shall analyze each of the graphs. It is very important to note that, for all methods, the results are always correct for the first two waves. Thus, we will be considering only the last two waves.

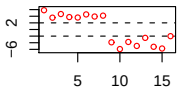
On Figure 1 and Figure 2, the first- and second-order inclusion probabilities, respectively, for Algorithms 4, 5, 6 and Algorithms 7, 8, 9 are plotted. We can see that:

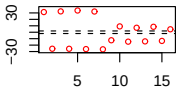
- Most or all of the first-order inclusion probabilities lie outside the confidence interval.
- Most of the second-order inclusion probabilities lie outside the confidence interval.

In conclusion, we can say that according to the graphs, the methods given by Algorithms 4, 5 and 6 are sim-false methods.

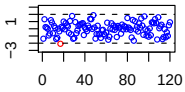


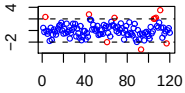



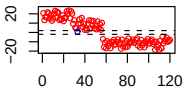



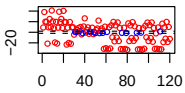




(a) First-order inclusion probabilities for Algorithm 4

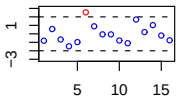


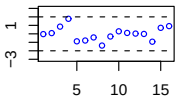



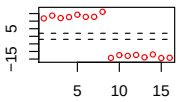



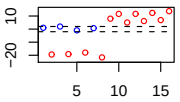




(b) Second-order inclusion probabilities for Algorithm 4

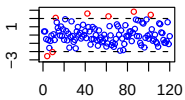


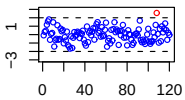



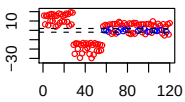



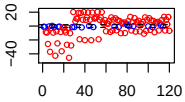




(c) First-order inclusion probabilities for Algorithm 5

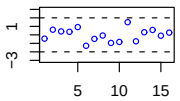


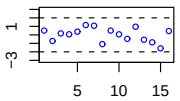



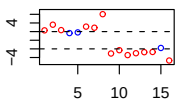



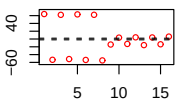




(d) Second-order inclusion probabilities for Algorithm 5

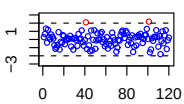


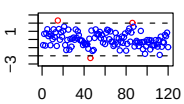



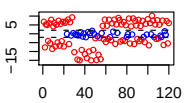



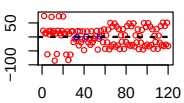




(e) First-order inclusion probabilities for Algorithm 6



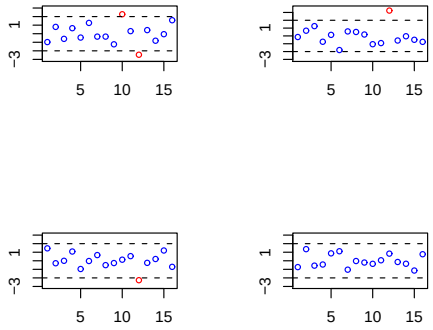



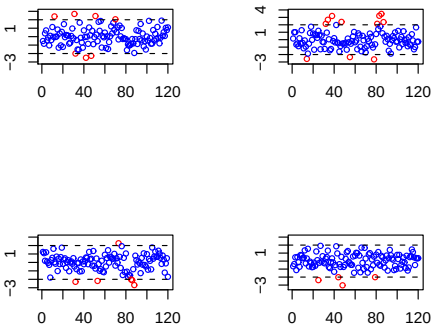




(f) Second-order inclusion probabilities for Algorithm 6

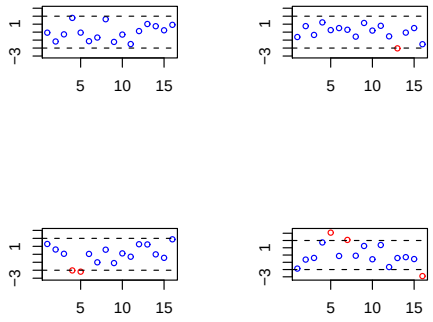
Figure 4.1: First- and second-order inclusion probabilities for Algorithms 4, 5, 6.



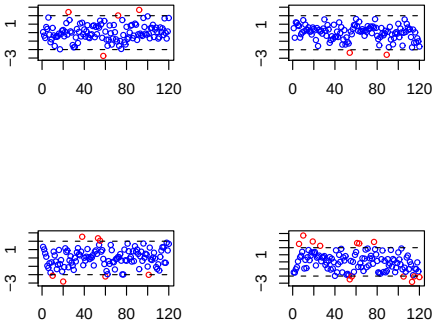

(a) First-order inclusion probabilities for Algorithm 7



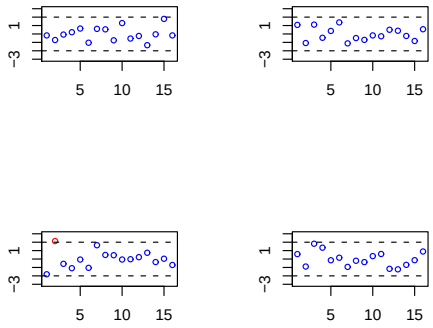

(b) Second-order inclusion probabilities for Algorithm 7



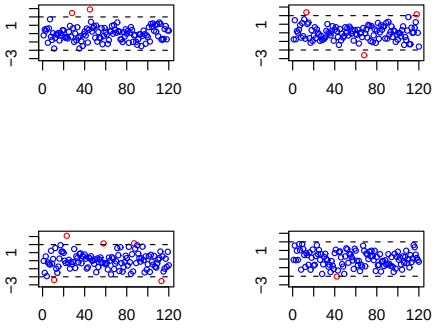

(c) First-order inclusion probabilities for Algorithm 8




(d) Second-order inclusion probabilities for Algorithm 8




(e) First-order inclusion probabilities for Algorithm 9




(f) Second-order inclusion probabilities for Algorithm 9

Figure 4.2: First- and second-order inclusion probabilities for Algorithms 7, 8, 9.

The quality of the coordination is given in Table 4.7. As we are coordinating negatively, the aim is to minimize the expected overlap, i.e. the number of common units in the samples. First, we compare the overlap between the samples of waves 3 and 4. If it is approximately the same, then we compare the overlap between the samples of waves 2 and 4. On this basis, we can conclude that the coordination works equally good for the Algorithms 7, 8 and 9.

Table 4.7: Expected overlaps.

| | Algorithm 7 | | | |
|---|---|---|---|---|
| | wave 1 | wave 2 | wave 3 | wave 4 |
| wave 1 | 8.000 | 2.159 | 4.027 | 3.085 |
| wave 2 | 2.159 | 8.000 | 2.073 | 3.941 |
| wave 3 | 4.027 | 2.073 | 8.000 | 0.126 |
| wave 4 | 3.085 | 3.941 | 0.126 | 6.000 |

| | Algorithm 8 | | | |
|---|---|---|---|---|
| | wave 1 | wave 2 | wave 3 | wave 4 |
| wave 1 | 8.000 | 2.161 | 4.028 | 3.086 |
| wave 2 | 2.161 | 8.000 | 2.071 | 3.945 |
| wave 3 | 4.028 | 2.071 | 8.000 | 0.125 |
| wave 4 | 3.086 | 3.945 | 0.125 | 6.000 |

| | Algorithm 9 | | | |
|---|---|---|---|---|
| | wave 1 | wave 2 | wave 3 | wave 4 |
| wave 1 | 8.000 | 2.163 | 4.024 | 3.088 |
| wave 2 | 2.163 | 8.000 | 2.073 | 3.943 |
| wave 3 | 4.024 | 2.073 | 8.000 | 0.125 |
| wave 4 | 3.088 | 3.943 | 0.125 | 6.000 |

## 4.10 Conclusions

The Kish & Scott, Cotton & Hesse, and Rivière methods allow the definition of the strata to be changed, which enables us to create a dynamic system of coordination. In the case of negative coordination, two bounds can be used as benchmarks for comparing the quality of the coordination: the absolute lower bound (ALB), which is rarely reached, and

Table 4.8: Summary table.

| Algorithm | Burden | Sample | Retrospective | Chronological | Sim-false |
|---|---|---|---|---|---|
| 3 | Yes | No | Yes | No | No |
| 4 | No | No | No | Yes | Yes |
| 5 | Yes | No | No | Yes | Yes |
| 6 | Yes | No | No | Yes | Yes |
| 7 | Yes | No | Yes | No | No |
| 8 | No | Yes | Yes | No | No |
| 9 | No | Yes | No | Yes | No |

the relative lower bound (RLB), defined as the solution of the linear program. A simple counter-example shows that neither the Kish & Scott nor the Cotton & Hesse method allow us to reach the RLB. Nevertheless, no other solution has been proposed to avoid the enumeration of all possible samples.

Based on the simulation results, we can see that the quality of coordination of the Kish & Scott method is worse than that of the Cotton & Hesse method. Moreover, the Kish & Scott method does not allow more than two samples to be coordinated. For these reasons, the Cotton & Hesse method should be preferred to the Kish & Scott method.

The Rivière method is slightly more complex to implement but, at present, the capacity of computation does not constitute a barrier to rapid implementation. We show that the condition for validity of the method is that the crossing of the strata is done for all the waves since the first survey that we want to coordinate. This limits our capacity to coordinate with really old surveys. The problem that can occur is that there are no more or very few units left in the microstrata. In a simulation based on four waves, the quality of the coordination of the Cotton & Hesse method is the same as that of the Rivière method. The concordance is so accurate, that we could conjecture that both methods provide the same joint sampling design.

However, the Rivière method is more flexible, the use of burden allows us to give more importance to chosen surveys. The burden can also change over time, for instance, if we want to have a positive coordination. Nevertheless, we do not understand why, in the implementation of MICROSTRAT and SALOMON software, only three sorts are done. We should rather advocate the use of one sort per wave, which seems to be possible with current computers, even with several decades of waves.

We have seen that it is easy to construct new methods because the ideas on which they are based are simple and intuitive. The method presented in Algorithm 4 is a non-PRN method where, at each wave, a new random number is generated for all the units of the population. The methods given in Algorithms 5 and 6 can be seen as modifications of the Rivière method. Because of their simplicity, it is difficult to understand why these methods do not work. After running simulations on only four waves, we showed that they are sim-false.

We have also seen that there is a way to construct modifications of the Cotton and Hesse method while permuting the random numbers in the crossing of all the strata. These methods have proven to be sim-correct.

Although it is difficult to give preference to one or another of the newly introduced methods, we believe that the method presented in Algorithm 8, which proved to be sim-correct, can be a good solution to the sample coordination problem. Like the Rivière method, it is based on the use of PRNs and is retrospective. Like the Cotton and Hesse method, the permutations are done while respecting the ranks in the vector of random numbers. Its innovation comes from the way in which the microstrata are constructed.

The general conclusion is that using methods based on microstrata makes coordination with old surveys very difficult because of the need to cross all the intermediate strata, which finally results in a very small sample size in the microstrata. Unfortunately, this seems to be a constraint of the coordination problem that cannot be ignored.

# Chapter 5

# General Framework for the Rotation of Units in Repeated Survey Sampling

**Abstract:**

Coordination of probabilistic samples is a challenging theoretical problem faced by statistical institutes. One of their aims is to obtain good estimates for each wave while spreading the response burden across the entire population. There is a collection of existing solutions that try to attend to these needs. These solutions, which were developed independently, are integrated in a general framework and their corresponding longitudinal designs are computed. The properties of these longitudinal designs are discussed. It is also noted that there is an antagonism between a good rotation and control over the cross-sectional sampling design. A compromise needs to be reached between the quality of the sample coordination, which appears to be optimal for a systematic longitudinal sampling design, and the freedom of choice of the cross-sectional design. In order to reach such a compromise, an algorithm that uses a new method of longitudinal sampling is proposed.

[1]This chapter is a reprint of the paper: D. Nedyalkova, L. Qualité and Y. Tillé. General Framework for the Rotation of Units in Repeated Survey Sampling. *Statistica Neerlandica*, 63:269-293, 2009.

## 5.1 Introduction

The negative coordination of samples is a challenging theoretical problem faced by statistical institutes. In business surveys, for example, several survey samplings are conducted each year on a relatively small population of large or medium-sized companies. The paperwork burden asked of these companies can lead to reduced response rates and lesser quality. It is thus important for statistical institutes to have some control over the response burden of the units in the population while maintaining a probabilistic sampling system. In business surveys, the aims of such a system can be diametrically opposed: the institutes may want to ensure that a company will not be selected too often in order to limit its burden, or on the contrary, they may want to have a large overlap between the samples of two consecutive waves in order to have accurate estimations for the evolutions. These opposite properties are respectively called negative and positive coordination of samples.

There is a collection of existing solutions that try to attend to these needs. For instance, the use of random numbers for coordinating Poisson designs (Brewer et al., 1972), collocated sampling (Brewer et al., 1984) and the use of a measure of burden (De Ree, 1983; Van Huis et al., 1994a,b). These methods give partial but important solutions to real-life problems. However, one drawback of these methods is that they do not allow the important advances made in the domain of one-sample selection over the last decades to be integrated. For example, none of these allow to use maximum fixed-size entropy sampling (see Chen et al., 1994), or balanced sampling (Deville and Tillé, 2004) as a cross-sectional sampling design.

Our aim is to provide the core of a general theory that includes the main existing sampling designs. However, in an attempt to keep this presentation simple, we will only describe negative coordination methods. The case of unit rotation (e.g. in partially renewed panels) is thus taken into account as it can be seen as a negative coordination problem. Indeed, sample rotation is usually achieved by splitting a sample into different parts and drawing for each new wave a non-overlapping sample that replaces one of these parts.

Another important issue where we made simplifications is the one of dynamic populations. In real-life problems, especially when we deal with business surveys, we need to allow for births and deaths of units in the population. This point is an important part of every rotation system and adds to its practical values. The algorithms present in this paper only require slight adaptations to work with a dynamic population.

The first part of our paper, in Section 2, is devoted to a review of the useful concepts and notations for sampling on one occasion and on several occasions. In Section 3, we

present some classical sampling designs in the context of longitudinal selection of a given unit over time. We also propose a new sampling algorithm that allows us to impose a minimum time between two selections of a unit.

After that, in Section 4, we review the main existing methods for negative coordination of samples: the Brewer method that selects Poisson samples, the method of permutation of random numbers (see Cotton and Hesse, 1992a,b), and the burden method. We show that it is possible to compute the cross-sectional and longitudinal sampling designs and in some cases even the joint sampling design. We also show that the longitudinal designs, which were never calculated before, are either systematic or Deville's systematic (Deville, 1998). These methods are not suitable if one wants to use a complex cross-sectional design (e.g. unequal inclusion probabilities and fixed size at the same time).

In Section 5, we give a general method that makes it possible to use any cross-sectional design along with a systematic longitudinal design. However, the cross-sectional design should be applied, at each step, on the conditional selection probabilities. That will result in a progressive loss of control over the cross-sectional designs. This difficulty sheds light on the antagonism between the requirements for the cross-sectional design and those for the longitudinal design. It seems that the quality of the coordination is contradictory with the control of the cross-sectional sampling design. Finally, in Section 6, we develop new sampling strategies that allow us to have a good coordination while leaving a relatively free choice of cross-sectional designs.

## 5.2 Basic Concepts and Notation

### 5.2.1 Sampling on one occasion

A finite population is a set of $N$ units. Each unit can be identified by a label. Let

$$U = \{1, \ldots, k, \ldots, N\}$$

be the set of these labels. The size $N$ of the population is not necessarily known. A sample without replacement is a subset of the population and in vector notation is presented as

$$\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)' \in \{0,1\}^N,$$

where

$$s_k = \begin{cases} 1 & \text{if unit } k \text{ is in the sample} \\ 0 & \text{if unit } k \text{ is not in the sample,} \end{cases}$$

for all $k \in U$. The sample size is

$$n(\mathbf{s}) = \sum_{k \in U} s_k.$$

A sampling design $p(\mathbf{s})$ is a probability distribution on the samples of $U$. Let $\mathbf{S}$ be the random sample, i.e. the random vector of $\mathbb{R}^N$, whose distribution is given by

$$\Pr(\mathbf{S} = \mathbf{s}) = p(\mathbf{s}).$$

The first-order inclusion probability $\pi_k$ is the probability of selecting unit $k$ in the sample, and $\boldsymbol{\pi} = (\pi_k)_{1 \leq k \leq N}$ is the inclusion probability vector. It can be derived from the sampling design as follows:

$$\boldsymbol{\pi} = \sum_{\mathbf{s} \subset U} \mathbf{s} p(\mathbf{s}).$$

When the design has a fixed sample size $n$, then

$$\sum_{k \in U} \pi_k = n.$$

The joint inclusion probability $\pi_{k\ell}$ is the probability of selecting units $k$ and $\ell$ together in the sample, and $\pi_{kk} = \pi_k$. The matrix of joint inclusion probabilities is given by

$$\boldsymbol{\Pi} = \sum_{\mathbf{s} \subset U} \mathbf{s}\mathbf{s}' p(\mathbf{s}).$$

A support $\mathcal{Q}$ is a set of samples. The support $\mathcal{Q}$ of a sampling design $p(\cdot)$ is defined by:

$$p(\mathbf{s}) > 0, \text{ for all } \mathbf{s} \in \mathcal{Q},$$

and

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

The full support $\mathcal{S}$ is the set of all the possible samples, i.e. $\mathcal{S} = \{0, 1\}^N$ and $\text{card}(\mathcal{S}) = 2^N$. The support corresponding to the samples of fixed sample size $n$ is defined by $\mathcal{S}_n =$

$\left\{\mathbf{s} \in \mathcal{S} | \sum_{k\in U} s_k = n\right\}$. Note that $\text{card}(\mathcal{S}_n) = \binom{N}{n}$. Some sampling designs have very small supports. They are called minimum support designs. We refer to the following definition:

DEFINITION 5.1. *A sampling design $p_0(\cdot)$ with inclusion probabilities $(\pi_k)_{1\leq k\leq N}$ is said to be defined on a minimum support $\mathcal{Q}_0$ if, for every $\mathcal{Q} \subset \mathcal{Q}_0$ with $\mathcal{Q} \neq \mathcal{Q}_0$, there is no design $p(\cdot)$ with support $\mathcal{Q}$ and with $\sum_{\mathbf{s}\in\mathcal{Q}} s_k p(\mathbf{s}) = \pi_k,\ k = 1, \ldots, N$.*

Péa et al. (2007) showed that the systematic design is a minimum support design. They also presented new methods to construct minimum support designs.

### 5.2.2 Sampling on several occasions

In coordination problems, we are interested in drawing samples from a population at times $t = 1, 2, \ldots, T$. At time $t$, a sample without replacement is a subset of the population.

DEFINITION 5.2. *The cross-sectional sample is denoted by a vector*

$$\mathbf{s}^t = (s_1^t, \ldots, s_k^t, \ldots, s_N^t)' \in \{0, 1\}^N,$$

*for all $t \in \{1, 2, \ldots, T\}$, and the longitudinal sample by a vector*

$$\mathbf{s}_k = (s_k^1, \ldots, s_k^t, \ldots, s_k^T)' \in \{0, 1\}^T,$$

*where*

$$s_k^t = \begin{cases} 1 & \text{if, at time } t, \text{ unit } k \text{ is in the sample } \mathbf{s}^t \\ 0 & \text{if, at time } t, \text{ unit } k \text{ is not in the sample } \mathbf{s}^t, \end{cases}$$

*for all $k \in U$.*

DEFINITION 5.3. *A sampling design $p(\mathbf{s}^t)$, $t = 1, 2, \ldots, T$, will be called a cross-sectional sampling design.*

DEFINITION 5.4. *A sampling design $p(\mathbf{s}_k)$, $k = 1, 2, \ldots, N$, will be called a longitudinal sampling design.*

The joint (or complete) sampling design $p(\mathbf{s})$ is given by

$$p(\mathbf{s}) = p(\mathbf{s}^1, \ldots, \mathbf{s}^t, \ldots, \mathbf{s}^T).$$

From this joint sampling design, we can derive the marginal cross-sectional design for a

time $t$

$$p(\mathbf{s}^t) = \sum_{\mathbf{s}^1,\ldots,\mathbf{s}^{t-1},\mathbf{s}^{t+1},\ldots,\mathbf{s}^T} p(\mathbf{s}^1,\ldots,\mathbf{s}^t,\ldots,\mathbf{s}^T),$$

and the marginal longitudinal design for a unit $k$,

$$p(\mathbf{s}_k) = \sum_{\mathbf{s}_1,\ldots,\mathbf{s}_{k-1},\mathbf{s}_{k+1},\ldots,\mathbf{s}_N} p(\mathbf{s}_1,\ldots,\mathbf{s}_k,\ldots,\mathbf{s}_N).$$

Let $S_k^t$ be the random variable that takes the value 1 if unit $k$ is selected at time $t$ and 0 otherwise. The first-order inclusion probabilities and the joint inclusion probabilities of the cross-sectional design at time $t$ are given respectively by:

$$\pi_k^t = \mathrm{E}(S_k^t) \quad \text{and} \quad \pi_{k\ell}^t = \mathrm{E}(S_k^t S_\ell^t),$$

where $\mathrm{E}(\cdot)$ is the expectation under the probability distribution $p(\cdot)$, $k,\ell \in U, t = 1,\ldots,T$. The longitudinal joint inclusion probabilities for times $t$ and $u$ are given by:

$$\pi_k^{tu} = \mathrm{E}(S_k^t S_k^u),\ k \in U, t, u = 1,\ldots,T.$$

Finally, we can define:

$$\pi_{k\ell}^{tu} = \mathrm{E}(S_k^t S_\ell^u),\ k,\ell \in U, t, u = 1,\ldots,T.$$

We have that $\pi_{k\ell}^{tu} = \pi_{\ell k}^{ut}$, where $k,\ell \in U$, but $\pi_{k\ell}^{tu}$ is not necessarily equal to $\pi_{\ell k}^{tu}$. These definitions can easily be adapted to a dynamic population denoted $U^t$, $t = 1,\ldots,T$, which holds $N^t$ units at time $t$. Naturally, if a unit $k$ does not belong to $U^t$, then $s_k^t = 0$ and the inclusion probabilities $\pi_k^t$ and $\pi_{k\ell}^{tu}$ are also null.

In a repeated sampling design, the objectives of the cross-sectional and longitudinal designs are completely different. The cross-sectional design must be organized so as to obtain a complete coverage of the population and optimize the accuracy of the estimators. The aim of the longitudinal design could be to organize an equitable rotation of the units in the samples in order to fairly share the response burden of the companies. In some studies, the aim could be to have the best possible longitudinal estimations. Fixed size of the longitudinal design is not necessarily required, but it could be if one wants to inform the units that they will be surveyed a certain number of times over a given period. Longitudinal

estimators do not necessarily need to be optimized.

Yet, up to now, no algorithm enables us to combine all these requirements. That is why relatively simple cross-sectional designs such as optimal stratified designs are generally used in repeated business surveys. It would be interesting to be able to have cross-sectional sampling designs with nice properties such as balanced sampling.

Another requirement is that we may want to be able to draw a sample at time $t$ without knowing how many other samples $\mathbf{s}^{t+1}, \ldots \mathbf{s}^{T}$ will have to be drawn, or the future inclusion probabilities $\pi_k^u$, $k \in U$, $u = t+1, \ldots, T$. In order to do so, we need to have an adequate longitudinal sampling algorithm. One should not confuse the definition of a sampling design with that of a sampling algorithm. The sampling design is given by the probability measure $p(.)$, while a sampling algorithm is a procedure that allows us to select a random sample. Usually, there are several algorithms that allow us to implement the same sampling design. For instance, in Tillé (2006, pp. 47-50), four sampling algorithms for simple random sampling design without replacement are proposed.

A sequential algorithm is a method that is applied to a list of units (or, in this case, occurrences) denoted $1, \ldots, t, \ldots, T$, which are sorted according to a particular order. Tillé (2006) gives two definitions of a sequential algorithm.

DEFINITION 5.5. *A longitudinal sampling algorithm, for a unit $k$, is said to be weakly sequential if at step $t = 1, \ldots, T$ of the procedure, the decision concerning whether the unit $k$ is in the sample $\mathbf{s}^t$ is definitively taken.*

DEFINITION 5.6. *A longitudinal sampling algorithm is said to be strictly sequential if it is weakly sequential and if the decision concerning the unit $k$ at time $t$ does not depend on the inclusion probabilities of the unit $k$ at times $t+1, \ldots, T$ and on the number $T$ of sampling occasions.*

A strictly sequential procedure may be necessary for the longitudinal design when we are sampling over time. This is the case when the inclusion probabilities for the future occasions are not known (e.g. they are proportional to a variable that is not available in advance), or when the total number of occasions is not known. Moreover, a strictly sequential algorithm allows for an indefinite number of sampling occasions, and can be used with dynamic populations. Indeed, with such an algorithm, the death of a unit has no influence on its previous selections and the unit can just stay in the population with null inclusion probabilities from then on. A newborn unit can always be added to the population and receive null inclusion probabilities for the previous sampling occasions.

A general schema for constructing a sequential algorithm consists in computing the conditional selection probabilities of a unit as described in Algorithm 10. First, a uniform random number is generated for each unit of the population. A unit is selected if its random number is at most equal to its inclusion probability. Next, for each consecutive wave, a uniform random number is generated for each unit $k \in U$. Then, a conditional probability of selection is computed. A unit is selected if its random number is at most equal to its conditional selection probability.

**Algorithm 10** General longitudinal sequential algorithm.

Generate $u$, a uniformly distributed random number in $[0, 1)$.
**if** $u < \pi_k^1$ **then**
  $s_k^1 = 1$
**else**
  $s_k^1 = 0$
**end if**
**for** $t = 2, \ldots, T$ **do**
  Generate $u$, a uniformly distributed random number in $[0, 1)$.
  Compute $p = \Pr(S_k^t = 1 | S_k^{t-1} = s_k^{t-1}, \ldots, S_k^1 = s_k^1)$.
  **if** $u < p$ **then**
    $s_k^t = 1$
  **else**
    $s_k^t = 0$
  **end if**
**end for**

The computation of the conditional selection probabilities can be intricate. In most cases, these probabilities depend on the inclusion probabilities at times $t+1, t+2, \ldots, T$, and thus, in those cases there is no strictly sequential algorithm to implement the design. When the sampling design is such that these conditional selection probabilities do not depend on the future, Algorithm 10 is strictly sequential. In Section 3, several strictly sequential algorithms, that are particular cases of Algorithm 10, along with the new algorithm that we propose, are presented.

### 5.2.3 Average time out of the sample

The distribution of the time between two selections of a given unit is an important characteristic for the coordination problem. Let $\psi_k^t$ be the random variable defined for $t = 1, \ldots, T$

by

$$\psi_k^t(\mathbf{s}_k) = \begin{cases} \min\left(T-t, \min\{r \geq 1 | s_k^{t+r} = 1\}\right) \text{ if } k \in \mathbf{s}^t, \\ 0 \text{ if } k \notin \mathbf{s}^t, \end{cases}$$

and $\psi_k^0(\mathbf{s}_k) = \min\left(T, \min\{r \geq 1 | s_k^r = 1\}\right)$. Let $\phi_k^t$, $t = 1, \ldots, T$, be a random variable with the same distribution as $\psi_k^t(\mathbf{s}_k)$ conditionally to $s_k^t = 1$, so that, if $1 \leq t < T$:

$$\phi_k^t = \begin{cases} 1 & \text{with probability } \Pr(S_k^{t+1} = 1 | S_k^t = 1) \text{ if } t+1 < T, \\ 2 & \text{with probability } \Pr(S_k^{t+2} = 1, S_k^{t+1} = 0 | S_k^t = 1) \text{ if } t+2 < T, \\ 3 & \text{with probability } \Pr(S_k^{t+3} = 1, S_k^{t+2} = 0, S_k^{t+1} = 0 | S_k^t = 1) \text{ if } t+3 < T, \\ \vdots & \\ T-t & \text{with probability } 1 - \sum_{r=1}^{T-t-1} \Pr(S_k^{t+r} = 1, S_k^{t+r-1} = 0, \ldots, S_k^{t+1} = 0 | S_k^t = 1). \end{cases}$$

We have the relation:

$$\sum_{t=0}^{T} \psi_k^t(\mathbf{s}_k) = T = \mathrm{E}\left(\psi_k^0\right) + \sum_{t=1}^{T} \pi_k^t \mathrm{E}\left(\phi_k^t | s_k^t = 1\right). \tag{5.1}$$

The quantity $\mathrm{E}\left(\phi_k^t | s_k^t = 1\right)$ can be seen approximately as the expected time out of the sample for a unit that has just been selected at time $t$.

In the subsequent sections, we will give the distribution of $\phi_k^t$ for several sampling designs. We will also show that the control of $\phi_k^t$ is the main issue in sampling coordination. We will consider particular sampling designs such as simple random sampling, Poisson sampling, systematic sampling, Deville's sampling, and give sequential algorithms for these designs.

## 5.3 Classical sampling designs

In this section, we will present a short summary of some of the classical sampling designs in the context of longitudinal sampling of a unit $k$ at times $t = 1, \ldots, T$ with inclusion probabilities $\pi_k^1, \ldots, \pi_k^T$. We will also give sequential or strictly sequential procedures to implement these designs.

### 5.3.1 Poisson sampling design

A longitudinal sampling design $p(\mathbf{s}_k)$ is said to be a Poisson sampling without replacement if it can be written

$$p(\mathbf{s}_k) = \prod_{t=1}^{T} (\pi_k^t)^{s_k^t} (1 - \pi_k^t)^{1 - s_k^t}.$$

The inclusion probabilities are equal to $\pi_k^t$ and the joint inclusion probabilities are equal to $\pi_k^{tu} = \pi_k^t \pi_k^u$ when $t \neq u$.

The random variables $S_k^1, S_k^2, \ldots, S_k^T$ are independent and thus the application of the general sequential Algorithm 10 to Poisson sampling gives the strictly sequential Algorithm 11:

**Algorithm 11** Poisson strictly sequential.

1: **for** $t = 1, \ldots, T$ **do**
2: Select unit $k$ at time $t$ with probability $\pi_k^t$.
3: **end for**

It is possible to compute the number of steps needed to select $k$ again given that it has been selected at time $t$:

$$\phi_k^t = \begin{cases} 1 & \text{with probability } \pi_k^{t+1} \text{ if } t+1 < T, \\ 2 & \text{with probability } \pi_k^{t+2}(1 - \pi_k^{t+1}) \text{ if } t+2 < T, \\ 3 & \text{with probability } \pi_k^{t+3}(1 - \pi_k^{t+2})(1 - \pi_k^{t+1}) \text{ if } t+3 < T, \\ 4 & \text{with probability } \pi_k^{t+4}(1 - \pi_k^{t+3})(1 - \pi_k^{t+2})(1 - \pi_k^{t+1}) \text{ if } t+4 < T, \\ \vdots & \end{cases}$$

With Poisson sampling, the sample size $n(\mathbf{S}_k)$ is random and has a Poisson-binomial distribution (see for example Hodges and LeCam, 1960). Its expected value and variance are, respectively, equal to:

$$\mathrm{E}\left[n(\mathbf{S}_k)\right] = \sum_{t=1}^{T} \pi_k^t \text{ and } \mathrm{var}\left[n(\mathbf{S}_k)\right] = \sum_{t=1}^{T} \pi_k^t (1 - \pi_k^t).$$

When all the inclusion probabilities are equal to $\pi_k$, the Poisson sampling design is called a Bernoulli design. In this case, $n(\mathbf{S}_k)$ has a binomial distribution: $n(\mathbf{S}_k) \sim \mathcal{B}(T, \pi_k)$ and

$$\Pr(\phi_k^t = j) = (1 - \pi_k)^{j-1} \pi_k, j = 1, \ldots, T - t - 1.$$

If $T$ is not finite, $\phi_k^t$ has a geometric distribution. In this case,

$$\mathrm{E}(\phi_k^t) = \frac{1}{\pi_k} \text{ and } \mathrm{var}(\phi_k^t) = \frac{1-\pi_k}{(\pi_k)^2}.$$

### 5.3.2 Simple random sampling

A longitudinal sampling design $p(\mathbf{s}_k)$ is said to be a simple random sampling without replacement (SRSWOR) (with fixed sample size $n_k$) if it can be written

$$p(\mathbf{s}_k) = \begin{cases} \binom{T}{n_k}^{-1} & \text{if } n(\mathbf{s}_k) = n_k, \\ 0 & \text{otherwise.} \end{cases}$$

The first-order inclusion probabilities are $\pi_k^t = n_k/T$, for all $t = 1, \ldots, T$, and the joint inclusion probabilities are $\pi_k^{tu} = n_k(n_k-1)/[T(T-1)]$, if $t \neq u$.

This design can be implemented using several sampling algorithms. An application of the general sequential Algorithm 10 was proposed by Fan et al. (1962) and is presented in Algorithm 12. First, a uniform random number $u$ is generated. Then, we calculate the probability of selection $p$. If the random number is less than the selection probability, then a unit is selected. The algorithm ends when exactly $n_k$ units are selected. This algorithm is sequential but not strictly sequential, as the inclusion probabilities depend on the number of sampling occasions.

**Algorithm 12** SRSWOR sequential.

1: Let $j = 0$.
2: Generate $u$, a uniformly distributed random number in $[0, 1)$.
3: **for** $t = 1, \ldots, T$ **do**
4: Calculate $p = \dfrac{n_k - j}{T - t + 1}$.
5: **if** $u < p$ **then**
6: Select unit $k$ in $\mathbf{s}^t$.
7: $j = j + 1$
8: **end if**
9: **end for**

It is possible to compute the number of steps needed to select unit $k$ again, given that

it has been selected at time $t$:

$$\Pr(\phi_k^t = j) = \frac{\binom{T-j-1}{n_k-2}}{\binom{T-1}{n_k-1}}, j = 1, \ldots, \min(T-t-1, T-n_k+1).$$

For the first $n_k - 2$ sampling occasions, $\phi_k^t$ has a negative (or inverse) hypergeometric distribution (see, for instance Johnson et al., 1992), and

$$\mathrm{E}(\phi_k^t) = \frac{T}{n_k} \text{ and } \mathrm{var}(\phi_k^t) = \frac{T(T-n_k)(n_k-1)}{(n_k+1)n_k^2}.$$

The moments of $\phi_k^t$ for larger values of $t$ are not as easy to obtain, due to the special treatment given to the last sampling occasion in the definition of $\psi_k^t$ and $\phi_k^t$.

### 5.3.3 Systematic sampling

Suppose that the longitudinal inclusion probabilities are such that $0 < \pi_k^t < 1$, $t = 1, \ldots, T$ with

$$\sum_{t=1}^{T} \pi_k^t = n_k.$$

Let $V_k^t$ be the cumulated inclusion probabilities defined by:

$$V_k^t = \sum_{i=1}^{t} \pi_k^i, \text{ for all } t = 1, \ldots, T, \tag{5.2}$$

with $V_k^0 = 0$ and $V_k^T = n_k$. The usual selection procedure for systematic sampling is given in Algorithm 13. This algorithm is sequential but is not a direct application of the general sequential Algorithm 10 to systematic sampling. The procedure is as follows. A uniform random number $u \in [0, 1)$ is generated. For all $t = 1, \ldots, T$, unit $k$ is selected in the sample $\mathbf{s}^t$ if there exists an integer $j$, $0 \leq j < n_k$, such that $u + j$ falls in the interval $[V_k^{t-1}, V_k^t)$.

The sampling design can be computed exactly:

$$p(\mathbf{s}_k) = \Lambda \left( \bigcap_{t|s_k^t=1} A_k^t \right),$$

**Algorithm 13** Usual strictly sequential algorithm for systematic sampling.

1: Generate $u$, a uniformly distributed random number in $[0, 1)$.
2: **for** $t = 1, \ldots, T$ **do**
3: **if** there is an integer $j > 0$ such that $V_k^{t-1} \leq u + j - 1 < V_k^t$ **then**
4: $s_k^t = 1$
5: **else**
6: $s_k^t = 0$
7: **end if**
8: **end for**

where $r_k^t = V_k^t \mod 1$,

$$A_k^t = \begin{cases} [r_k^{t-1}, r_k^t) & \text{if } r_k^{t-1} < r_k^t \\ [r_k^{t-1}, 1) \cup [0, r_k^t) & \text{otherwise,} \end{cases}$$

and $\Lambda\left(\bigcap_{t|s_k^t=1} A_k^t\right)$ is the sum of the length of the intervals in $\bigcap_{t|s_k^t=1} A_k^t$.

We propose an alternative to Algorithm 13 which is a direct application of the general sequential Algorithm 10 to systematic sampling. Algorithm 14 is strictly sequential, and as such it is practical for longitudinal sampling. It gives a simple procedure to compute the conditional probabilities of selecting the unit $k$ at time $t$ given the past. This algorithm can easily be adapted to the case of an indefinite number of sampling occasions.

If the sampling design is systematic with equal inclusion probabilities $\pi_k^t = \pi_k$, if $T$ is infinite, and if $c$ is the smallest integer such that $c\pi_k > 1$, then

$$\phi_k^t = \begin{cases} c - 1 & \text{with probability } (c\pi_k - 1)/\pi_k \\ c & \text{with probability } 1 - (c\pi_k - 1)/\pi_k. \end{cases}$$

$$\mathrm{E}(\phi_k^t) = \frac{1}{\pi_k} \text{ and } \mathrm{var}(\phi_k^t) = \frac{(c\pi_k - 1)(1 + \pi_k - c\pi_k)}{\pi_k^2}.$$

If $1/\pi_k$ is an integer, then $\mathrm{var}(\phi_k^t) = 0$.

### 5.3.4 Deville's systematic sampling

Deville (1998) presented a variant of the systematic algorithm that gives a new sampling design with unequal probabilities (see also Tillé, 2006, p.128). Deville's technique gives a fixed-size sampling design with a larger support than systematic sampling and is based

**Algorithm 14** Sequential systematic algorithm.

Define $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^T)'$, a vector of inclusion probabilities in $[0,1]^T$.
Define $\mathbf{s}_k = (s_k^1, \ldots, s_k^T)' = (0, \ldots, 0)'$, the empty sample.
Define $[a, b] = [0, 1]$.
Generate $u$, a uniformly distributed random number in $[0, 1)$.
**for** $j = 1, \ldots, T$ **do**
  Calculate $p = \dfrac{\max(\min(\pi_k^j, b) - \max(0, a), 0)}{b - a}$.
  **if** $p > 0$ **then**
    **if** $u < \Pr(S_k^j = 1 | S_k^{j-1} = s_k^{j-1}, \ldots, S_k^1 = s_k^1) = p$ **then**
      $s_k^j = 1$
      $b = \min(\pi_k^j, b)$
    **else**
      $a = \pi_k^j$
    **end if**
  **end if**
  $a = (a - \pi_k^j) \mod 1$
  $b = (b - \pi_k^j) \mod 1$
  **if** $b < a$ **then** $b = b + 1$ **end if**
**end for**

on a relatively simple algorithm. While only one random number is used for systematic sampling and its position relative to the cumulative inclusion probabilities $V_k^t$ defined by Expression (5.2) determines the whole sample, Deville's sampling uses several random numbers. The position of a random number has repercussions on a limited number of selection variables $s_k^t$. As a consequence, the conditional selection probabilities of a unit given its past are less constrained than in systematic sampling. A random number is selected uniformly in each interval between two integers. Its position relative to the $V_k^t$ that are also in this interval determines the values of the corresponding $s_k^t$. A slight adaptation has to be made so as to take into account the $t$ such that $[V_k^{t-1}, V_k^t)$ contains an integer. Deville's sampling can be implemented with Algorithm 15.

Deville's systematic sampling can also be implemented in the form of a strictly sequential algorithm. Algorithm 16 is a particular case of the general sequential Algorithm 10. At each step of the algorithm, a conditional selection probability $p$, is computed. This algorithm can easily be adapted to the case of an indefinite number of sampling occasions.

For this sampling design, in the particular case where $T$ is infinite, the $\pi_k^t$ are equal to $\pi_k$, and $K = 1/\pi_k$ is an integer, we can compute the distribution of the variables $\phi_k^t$. Let

**Algorithm 15** Deville's systematic sampling.

1: Generate $u_1$, a realization of a uniform random variable in $[0, 1)$.
2: **if** $V_k^{t-1} \leq u_1 < V_k^t$ **then** $s_k^t = 1$ **end if**
3: **for** $i = 2, \ldots, n_k$ **do**
4: **if** $\ell$ is such that $V_k^{\ell-1} \leq i - 1 < V_k^\ell$ **then**
5: **if** $s_k^\ell = 1$ **then**
6:
$$f(x) = \begin{cases} \dfrac{1}{i - V_k^\ell} & \text{if } x \geq V_k^\ell - (i-1) \\ 0 & \text{if } x < V_k^\ell - (i-1) \end{cases}, x \in [0, 1).$$
7: **else**
8:
$$f(x) = \begin{cases} 1 - \dfrac{(i - 1 - V_k^{\ell-1})(V_k^\ell - i + 1)}{[1 - (i - 1 - V_k^{\ell-1})][1 - (V_k^\ell - i + 1)]} & \text{if } x \geq V_k^\ell - (i-1) \\ \dfrac{1}{1 - (i - 1 - V_k^{\ell-1})} & \text{if } x < V_k^\ell - (i-1). \end{cases}$$
9: **end if**
10: **end if**
11: Generate $u_i$, a random variable with density $f(x)$.
12: **if** $V_k^{t-1} \leq u_i + i - 1 < V_k^t$ **then** $s_k^t = 1$ **end if**
13: **end for**

$r = K - (t \mod K)$, for all $t$. Then,

$$\phi_k^t = \begin{cases} r+1 & \text{with probability } 1/K \\ r+2 & \text{with probability } 1/K \\ \vdots & \\ r+K & \text{with probability } 1/K. \end{cases}$$

In this case,

$$\Pr(\phi_k^t = j) = \frac{1}{K}, j = r+1, \ldots, r+K,$$

$$\mathrm{E}(\phi_k^t) = r + \frac{K+1}{2} \text{ and } \mathrm{var}(\phi_k^t) = \frac{(K-1)(K+1)}{12}.$$

The variance of $\phi_k^t$ is larger than in the case of systematic sampling.

**Algorithm 16** Deville's Systematic Sequential.

Define $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^T)'$, a vector of inclusion probabilities of length $T$.
Define $\mathbf{s}_k = (s_k^1, \ldots, s_k^T)' = (0, \ldots, 0)'$, the empty sample.
Define $v = 0$, the cumulated inclusion probability.
Define $f = 0$.
Generate $u$, a uniformly distributed random number in $[0, 1)$.
**for** $j = 1, \ldots, T$ **do**
  $p = 0$
  **if** $v + \pi_k^j < 1$ **then**
    **if** $f = 0$ **then** $p = \pi_k^j/(1 - v)$ **end if**
  **else**
    **if** $f = 0$ **then**
      $p = 1$
    **else**
      $p = (v + \pi_k^j - 1)/v$
    **end if**
  **end if**
  $v = v + \pi_k^j$
  **if** $u < \Pr(S_k^j = 1 | S_k^{j-1} = s_k^{j-1}, \ldots, S_k^1 = s_k^1) = p$ **then** $s_k^j = 1$ **end if**
  **if** $v > 1$ **then**
    $v = v - 1$
    **if** $s_k^j = 0$ **then** $f = 0$ **end if**
  **else**
    **if** $s_k^j = 1$ **then** $f = 1$ **end if**
  **end if**
**end for**

Systematic sampling and Deville's systematic sampling share the property that if

$$j \leq V_k^t = \sum_{i=1}^{t} \pi_k^i,$$

for a given integer $j$, then

$$j \leq \sum_{i=1}^{t} s_k^i.$$

This property enables us to implement a design sequentially with a controlled size over an indefinite period of time.

### 5.3.5 A new sampling algorithm for unequal probability sampling

In this algorithm, we define a fixed number of steps during which a unit, once selected, is not selected anymore. At each step, a conditional probability is calculated. However, this is only possible if the sum of the inclusion probabilities over this number of consecutive steps does not exceed 1. If $r$ is the number of steps such that $s_k^t = 1$ implies $s_k^{t+1} = 0, \ldots, s_k^{t+r} = 0$, and if $\sum_{j=t}^{t+r-1} \pi_k^j < 1$, for $t = 1, \ldots, T - r + 1$, we consider Algorithm 17.

**Algorithm 17** Minimum time out of sample.

1: Define $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^T)'$, a vector of inclusion probabilities of length $T$.
2: Define $\mathbf{s}_k = (s_k^1, \ldots, s_k^T)' = (0, \ldots, 0)'$, the empty sample.
3: Generate $u^1$, a uniformly distributed random number in $[0, 1)$.
4: If $u^1 \leq \pi_k^1$ **then** $s_k^1 = 1$ **end if**
5: **for** $t = 2, \ldots, r$ **do**
6: Generate $u^t$, a uniformly distributed random number in $[0, 1)$.
7: If $s_k^1 = 0, \ldots, s_k^{t-1} = 0$ and $u^t \leq p = \pi_k^t/(1 - \sum_{i=1}^{t-1} \pi_k^i)$ **then** $s_k^t = 1$ **end if**
8: **end for**
9: **for** $t = r + 1, \ldots, T$ **do**
10: Generate $u^t$, a uniformly distributed random number in $[0, 1)$.
11: If $s_k^{t-r} = 0, \ldots, s_k^{t-1} = 0$ and $u^t \leq p = \pi_k^t/(1 - \sum_{i=t-r}^{t-1} \pi_k^i)$**then** $s_k^t = 1$ **end if**
12: **end for**

This algorithm is strictly sequential and can easily be adapted to an indefinite number of sampling occasions. In the Appendix, we give a modified version of this algorithm in which there is no condition on the inclusion probabilities. In that case, the fixed minimum time out of sample can not always be respected.

When $T$ is infinite, if all the $\pi_k^t$ are equal, i.e. $\pi_k^t = \pi_k$, and if $r\pi_k < 1$, then the distribution of the $\phi_k^t$ is as follows:

$$\Pr(\phi_k^t = j) = \begin{cases} 0 & , j = 1, \ldots, r \\ (1 - \nu_k)^{j-r-1}\nu_k & , j = r+1, r+2, r+3, \ldots, \end{cases}$$

where

$$\nu_k = \frac{\pi_k}{1 - r\pi_k}.$$

The variable $\phi_k^t$ has a shifted geometric distribution. We have:

$$\mathrm{E}(\phi_k^t) = r + \frac{1}{\nu_k} = \frac{1}{\pi_k} \text{ and } \mathrm{var}(\phi_k^t) = \frac{[(r+1)\pi_k - 1]\,[(r+1)\pi_k - \pi_k - 1]}{(\pi_k)^2}.$$

The minimum time out of sample design can be viewed as a compromise between Poisson sampling and systematic sampling. On the one hand, if all the $\pi_k^t$ are equal to $\pi_k$ and $\pi_k = (r+1)^{-1}$, then $\nu_k = 1$ and the sampling design is systematic. On the other hand, if $r = 0$, then we obtain a Poisson design. The Poisson design maximizes the entropy while the systematic design has a very small entropy because it is a minimum support design. Between these two extreme situations, the minimum time out of sample design provides a large range of intermediate solutions.

### 5.3.6 Remark on the variables $\phi_k^t$

For the sampling designs we just viewed, the expectation of the variables $\phi_k^t$ do not vary much. Expression (5.1), which is valid for any $T$, implies that in most cases this expectation will be close to $1/\pi_k^t$. For instance, if all the inclusion probabilities are equal to $\pi_k$, and $T$ is infinite, then systematic sampling, minimum time out of sample and Poisson sampling all have the same expectation for $\phi_k^t$. The variance of $\phi_k^t$, however, varies greatly from one sampling design to the other. When $\pi_k^t = \pi_k$ is constant and $1/\pi_k$ is integer, the only sampling design that gives a null variance for $\phi_k^t$ is the systematic sampling design.

The variable $\phi_k^t$ counts the number of waves a unit stays out of the sample after having been selected. The expectation of this variable does not depend much on the sampling design. Hence, a good method for negative coordination can not be a method that maximizes the number of times out of the sample after the selection of a unit. Instead, we can look for a method that organizes the rotation in a regular way, i.e. that minimizes the variance of the $\phi_k^t$. In this respect, systematic sampling is an interesting longitudinal design as it can give, in a very special case, perfect control over the frequency with which a unit is sampled.

In the next section we examine the most usual coordination methods and show that their longitudinal designs match the ones we have just described. In the simplest cases, we compute the cross-sectional, longitudinal and joint designs exactly.

## 5.4 Usual methods of coordination

There are several simple algorithms that allow us to draw coordinated samples with simple random or Poisson cross-sectional sampling designs. In this section, we describe three well-known methods. We give the corresponding longitudinal designs and compute the joint

sampling designs resulting from these algorithms.

### 5.4.1 The systematic-Poisson (or Brewer) repeated design

Brewer et al. (1972) suggested a very simple procedure to draw negatively coordinated Poisson samples. It gives a very convenient method to negatively coordinate samples with unequal probabilities. However, the cross-sectional samples do not have a fixed size. First, a uniform random number is generated for each unit of the population. A unit is selected if its random number is at most equal to its inclusion probability. Next, for each consecutive wave, we calculate a new uniform random number for all $k \in U$, which depends on the random number and on the inclusion probability at the previous wave. A unit is then selected if its new random number is at most equal to its new inclusion probability. The selection procedure is given in Algorithm 18.

**Algorithm 18** Coordination of Poisson samples in the case of a static population.

1: At time 1, assign a uniform random number $u_k^1$ to each unit $k \in U^1$.
2: **if** $u_k^1 \leq \pi_k^1$ **then** $s_k^1 = 1$ **end if**
3: **for** $t = 2, \ldots, T$ **do**
4: Compute $u_k^t = (u_k^{t-1} - \pi_k^{t-1}) \mod 1$.
5: **if** $u_k^t \leq \pi_k^t$ **then** $s_k^t = 1$ **end if**
6: **end for**

The cross-sectional design given by this algorithm is a Poisson design:

$$p(\mathbf{s}^t) = \prod_{k \in U} \left\{ (\pi_k^t)^{s_k^t} (1 - \pi_k^t)^{1 - s_k^t} \right\}.$$

The longitudinal design is a systematic sampling design with unequal probabilities:

$$p(\mathbf{s}_k) = \Lambda \left( \bigcap_{t | s_k^t = 1} A_k^t \right),$$

with the notations of Section 5.3.3. The selection of different units of the population being totally independent, the complete design is given by:

$$p(\mathbf{s}) = \prod_{k \in U} p(\mathbf{s}_k) = \prod_{k \in U} \Lambda \left( \bigcap_{t | s_k^t = 1} A_k^t \right).$$

From the complete design, we can derive all the properties of the sampling design. For example, we have

$$\pi_{k\ell}^t = \pi_k^t \pi_\ell^t, \text{ with } k \neq \ell, \text{ for all } t,$$

and

$$\pi_k^{t,t+j} = \Lambda\left(A_k^t \cap A_k^{t+j}\right).$$

This method can easily be adapted for dynamic populations. If a newborn unit enters the population at a given time $t > 1$, then it receives an inclusion probability $\pi_k^t$ and a uniform random number $u_k^t$. If its random number is not greater than its inclusion probability, then it is selected in the sample $\mathbf{s}^t$. At the following waves, its random number is subject to the same transformations as those of the other units of the population, described in line 4 of the algorithm. If a unit leaves the population at time $t$, then its inclusion probability becomes equal to zero for times $t, t+1, \ldots, T$. The adjustment of Algorithm 18 to dynamic populations is straightforward. We just need to replace $U$ by $U^t$ and add the following line between lines 4 and 5 of the algorithm:

| |
|---|
| 4b: Add newborn units to the sampling frame with their $u_k^t$ and $\pi_k^t$. |

As stated in the preceding section, a longitudinal systematic design can be desired in order to control the rotation of units in the sample. While Brewer's repeated design enjoys this property, it has a drawback: the cross-sectional design does not have a fixed sample size.

### 5.4.2 The systematic-simple repeated design

Suppose that the inclusion probabilities of the units in the population are constant at each wave, i.e. $\pi_k^t = \pi^t$, and that $\sum_{k\in U} \pi_k^t = n^t$ is integer. The following well-known procedure (see Cotton and Hesse, 1992a,b), given in Algorithm 19, can be used to negatively coordinate simple random samples without replacement. Its main drawback is that it can only be used in the case of simple random sampling or stratified sampling with fixed strata.

First, a uniform random number is generated for each unit of the population. In order to obtain the sample $\mathbf{s}^1$, the $n^1$ units having the smallest random numbers are selected. At the following waves, permute the uniform random numbers so that the selected units at the previous wave receive the largest random numbers and the non-selected receive the smallest. Select the $n^t$ units having the smallest random numbers to obtain the sample $\mathbf{s}^t$.

**Algorithm 19** Coordination of SRSWOR using random numbers in the case of a static population.

1: At time 1, assign a uniform random number, $u_k^1$, to each unit $k \in U$, i.e. construct the vector $\mathbf{u^1} = (\mathbf{u_1^1}, \ldots, \mathbf{u_N^1})$.
2: Select the units that have the $n^1$ smallest $u_k^1$ to obtain the sample $\mathbf{s}^1$.
3: **for** $t = 2, \ldots, T$ **do**
4: Construct $\mathbf{u^t}$ as a permutation of $\mathbf{u^{t-1}}$ so that the selected units at wave $t-1$ receive the largest $u_k^{t-1}$, the non-selected units receive the smallest $u_k^{t-1}$ and the ranks of the permuted random numbers remain unchanged within the subsets of selected and non-selected units.
5: Select the units that have the $n^t$ smallest $u_k^t$ to obtain the sample $\mathbf{s}^t$.
6: **end for**

This sampling algorithm results in a systematic longitudinal design. All the cross-sectional designs are simple and without replacement:

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

Moreover, for a static population, if

$$n(\mathbf{s}^1) = n^1, \ldots, n(\mathbf{s}^t) = n^t \text{ and } \sum_{j=1}^{t} n^j \leq N,$$

then

$$p(\mathbf{s}) = \left\{ \frac{N!}{n^1! n^2! \ldots n^t! (N - n^1 - \cdots - n^t)!} \right\}^{-1}.$$

This method can easily be adapted for dynamic populations. If a newborn unit $k$ enters the population at a given time $t > 1$, then a uniform random number $u_k^t$ is generated for this unit. The vector $\mathbf{u^t}$ is permuted as before and the new random number is added to this vector. Again we select the $n^t$ units having the smallest random numbers among the living $N^t$ units of the population at time $t$. If a unit leaves the population at time $t > 1$, then we simply remove its random number from the vector $\mathbf{u^t}$. The adjustment of Algorithm 19 is straightforward. We just need to replace $U$ by $U^t$ and add the following line between lines 4 and 5 of the algorithm:

4b: Add newborn units' random numbers $u_k^t$ to the vector $\mathbf{u^t}$ at their relative positions.

### 5.4.3 Use of a measure of burden or the Deville's systematic-simple repeated design

Another method of coordination with simple random cross-sectional designs, based on the use of permanent random numbers for the selection of the sample, is used by Statistics Netherlands for their business surveys (see De Ree, 1983; Koeijers and Willeboordse, 1995; Van Huis et al., 1994a,b). This method, called EDS, gives stratified cross-sectional samples with fixed size. The cumulative response burden of the units is factored in the selection process, and the surveys can have unequal response burdens. However, the choice of the stratification is not completely free. Indeed, the strata are constituted of response burden control groups, which are basic blocks of units that need to be defined once and for all. Rivière (2001a) proposed another method that takes into account the response burden, and that does not require predefined strata. However, it uses the intersection of all the previous stratifications. Consequently, it is not practical for a large number of sampling occasions. These methods, along with the Cotton and Hesse method, can be used with dynamic populations.

We describe another simple method of coordination that uses a measure of burden but does not use permanent random numbers. At each wave, every unit receives a measure of burden equal to the number of times it has previously been selected. At time $t$, the sample of size $n^t$ is selected among the units with the lowest burden measure. More precisely, the sample $\mathbf{s}^t$ at time $t$ is drawn with a simple random sampling of size $n^t$ in the set of units with the smallest burden, if this set is large enough. Or, if this set is too small, it is entirely selected and a sample is drawn with a simple random design in its supplement, in order to complete $\mathbf{s}^t$.

With this method, the burden measure, at any time $t$, can only take two values. It splits the population into a set of units with the lowest burden measure, denoted $\mathcal{M}^{t-1}$ and a set of units with the largest burden measure, denoted $U \backslash \mathcal{M}^{t-1}$. The procedure is given in Algorithm 20.

The cross-sectional design resulting from this algorithm is a SRSWOR, and the longitudinal design is the Deville's systematic sampling design, presented in Algorithm 16. Indeed, if the cumulated sum (over $t$), $V_k^\ell$, of the $\pi_k^t = n^t/N$ is such that there is an integer $i - 1 \geq 1$ between $V_k^{\ell-1}$ and $V_k^\ell$, then the population at time $\ell - 1$ is divided between units that have been selected $i - 2$ times and units that have been selected $i - 1$ times. Depending on its burden at time $\ell - 1$, unit $k$ is automatically selected in $\mathbf{s}^\ell$ (case $s_k^\ell = 1$

**Algorithm 20** Coordination of SRSWOR using a measure of burden in the case of a static population.

1: At time 1, assign a burden equal to 0 to each unit $k \in U$, i.e. $b_k^1 = 0$.
2: Select a SRSWOR of size $n^1$.
3: **if** $s_k^1 = 1$ **then** $b_k^1 = b_k^1 + 1$ **end if**
4: **for** $t = 2, \ldots, T$ **do**
5: Define $\mathcal{M}^{t-1}$, the set of units with the smallest burden.
6: **if** $\text{card}(\mathcal{M}^{t-1}) > n^t$ **then**
7: Select a SRSWOR of size $n^t$ from $\mathcal{M}^{t-1}$.
8: **else**
9: Select all the units from $\mathcal{M}^{t-1}$.
10: Complete the sample by a SRSWOR with the units from $U \backslash \mathcal{M}^{t-1}$.
11: **end if**
12: **if** $s_k^t = 1$ **then** $b_k^t = b_k^{t-1} + 1$ **end if**
13: **end for**

at line 5 in Algorithm 15) or it is drawn with equal probability among the units that have a burden equal to $i - 1$. At the following occasions, its conditional selection probability satisfies the equations in lines 6 and 8 of Algorithm 15.

This method can easily be adapted for dynamic populations. Following the idea of De Ree (1999), we randomly assign to the newborn units a measure of burden. For example, at the beginning of wave $t > 1$, there are $\text{card}(\mathcal{M}^{t-1})$ units with the smallest burden, denoted $b$, and $N^{t-1} - \text{card}(\mathcal{M}^{t-1})$ units with the largest burden, equal to $b + 1$. A newborn unit will be inserted into $\mathcal{M}^{t-1}$ and receive a burden equal to $b$ with probability $\text{card}(\mathcal{M}^{t-1})/N^{t-1}$, and, with probability $1 - \text{card}(\mathcal{M}^{t-1})/N^{t-1}$, it will receive a burden equal to $b+1$. The adjustment of Algorithm 20 is straightforward. We just need to replace $U$ by $U^t$, define $\mathcal{M}^{t-1}$ as the set of living units with the smallest burden, and add the following line between lines 5 and 6 of the algorithm:

5b: Add each newborn unit to $\mathcal{M}^{t-1}$ with probability $\text{card}(\mathcal{M}^{t-1})/N^{t-1}$ and to $U^t \backslash \mathcal{M}^{t-1}$ with probability $1 - \text{card}(\mathcal{M}^{t-1})/N^{t-1}$.

The coordination is not as good as with longitudinal systematic sampling since the measure of burden does not reflect the time spent out of the sample. For example, suppose that $N = 4$, all the inclusion probabilities are $\pi_k^t = 1/4$ and $t = 1, ..., T$. After four waves all the units have been selected once and have a burden $b_k^4 = 1$. Hence, at the fifth wave, any unit can be selected again with probability 1/4. The same unit can thus be

consecutively selected at $t = 4$ and $t = 5$. Contrariwise, in this case the systematic-simple design provides a strict rotation, in such a way that once a unit is selected it remains out of the sample during three waves. The EDS method by Van Huis et al. (1994a,b) is not affected by this problem.

## 5.5 Other repeated sampling designs

### 5.5.1 General method for the coordination of samples

The usual algorithms described in the previous section result either in a systematic or in a Deville's systematic longitudinal design. While systematic sampling seems to be a good choice for the longitudinal design, these algorithms do not allow for a wide selection of cross-sectional designs. We can wonder if there is a general way of obtaining a repeated sampling design with a given sequential longitudinal design and a given cross-sectional design. A weaker solution is possible if we are prepared to have unperfect control over the cross-sectional sampling designs.

- For each unit $k$ of the population at time $t$, compute the conditional inclusion probabilities
$$\pi_k^t(s_k^{t-1}, \ldots, s_k^1) = \Pr(S_k^t = 1 | S_k^{t-1} = s_k^{t-1}, \ldots, S_k^1 = s_k^1),$$
according to the chosen strictly sequential algorithm (systematic, Deville's systematic, Minimum time out of sample).

- When all the conditional probabilities are computed, apply to them any cross-sectional design in order to select $\mathbf{s}^t$. This design can be stratified, with unequal inclusion probabilities and fixed sample size, or even balanced (see Deville and Tillé, 2004).

With this method, conditionally to the past, we can choose any cross-sectional design provided that it is compatible with the conditional inclusion probabilities. This may in itself be a limiting factor, especially in the case of a systematic longitudinal design where these inclusion probabilities can rapidly become close or equal to 0 or 1. Moreover, the choice of the conditional cross-sectional design at time $t$ is not the same as the choice of the marginal (unconditional) design for $\mathbf{s}^t$. This method is perfectly applicable to dynamic populations.

### 5.5.2 Application to a systematic longitudinal design

We have seen that the systematic longitudinal design is well-suited for sampling coordination. It is thus of interest to know which cross-sectional designs can be implemented with a longitudinal systematic sampling design. We have seen in the previous section that it is the case of the Poisson design and the simple random sampling design.

The sampling design at the first wave can always be chosen at will. If we apply the sequential algorithm presented in Algorithm 14, then we can compute at wave 2 the conditional inclusion probabilities for each unit as follows:

$$\pi_k^2(s_k^1) = \Pr(S_k^2 = 1 | S_k^1 = s_k^1),$$

which, given that the longitudinal design is systematic, are such that

$$\pi_k^2(s_k^1) = \begin{cases} \pi_k^2/(1-\pi_k^1) & \text{if } s_k^1 = 0 \text{ and } \pi_k^1 + \pi_k^2 \leq 1 \\ 0 & \text{if } s_k^1 = 1 \text{ and } \pi_k^1 + \pi_k^2 \leq 1 \\ 1 & \text{if } s_k^1 = 0 \text{ and } \pi_k^1 + \pi_k^2 > 1 \\ (\pi_k^1 + \pi_k^2 - 1)/\pi_k^1 & \text{if } s_k^1 = 1 \text{ and } \pi_k^1 + \pi_k^2 > 1. \end{cases}$$

Then, any sampling design can be applied with the conditional inclusion probabilities $\pi_k^2(s_k^1)$. It must be noted that this free choice of conditional sampling design $p(\mathbf{s}^2|\mathbf{s}^1)$ does not mean that we know how to obtain a given marginal sampling design $p(\mathbf{s}^2)$ for the second wave. The identity $p(\mathbf{s}^2) = \sum_{\mathbf{s}^1} p(\mathbf{s}^2|\mathbf{s}^1)p(\mathbf{s}^1)$ is not readily reversible in a way that would enable us to select an adequate conditional design for a given marginal design.

Moreover, the conditional sampling design must respect the conditional inclusion probabilities. These constraints prevent us from using some conditional sampling designs. For instance, suppose that $p(\mathbf{s}^1)$ and $p(\mathbf{s}^2)$ have unequal inclusion probabilities $\pi_k^1$ and $\pi_k^2$ such that $\pi_k^1 + \pi_k^2 < 1$ for all $k \in U$. Then, even if $\mathbf{s}^1$ is selected with a fixed sample size $n^1$, there is no particular reason why

$$\sum_{k \in U} \pi_k^2(s_k^1) = \sum_{k | s_k^1 = 0} \pi_k^2/(1-\pi_k^1)$$

would be an integer, and there is even less reason for it to be equal to

$$\sum_{k \in U} \pi_k^2.$$

Hence, with a longitudinal systematic design, it is not possible to have complete control over the cross-sectional design of the second wave. The size of the conditional sampling design for the second wave may be random. This question is also discussed in Tillé and Favre (2004).

The method described in this section can be generalized for any number of sampling occasions. We compute the conditional inclusion probabilities at time $t$:

$$\pi_k^t(s_k^{t-1}, \ldots, s_k^1) = \Pr(S_k^t = 1 | S_k^{t-1} = s_k^{t-1}, \ldots, S_k^1 = s_k^1),$$

for a systematic longitudinal design. These conditional probabilities are computed at line 6 in Algorithm 14 for any time $t$. After several waves, more and more conditional inclusion probabilities are equal to 0 or 1, as the interval $[a, b]$ of Algorithm 14 becomes smaller. Hence, the conditional inclusion probabilities become more and more deterministic and the conditional sampling design can not be freely chosen. This problem sheds light on the antagonism between the desire for a good rotation system and the control over the cross-sectional sampling designs.

## 5.6 Other solutions to the coordination problem

### 5.6.1 The dilemma of sampling coordination

In the previous section we have seen that, while the systematic design is a good longitudinal sampling design, its use leads to a considerable loss of control over the cross-sectional sampling design. This is due to the fact that systematic sampling has a very small support. In order to avoid this problem, we must choose a longitudinal design that gives more freedom to the user, such as Deville's design or the minimum time out of sample design, and possibly cope with deteriorated coordination.

Until now, there were two main approaches to the coordination problem:

- To choose the cross-sectional design and try to get the best coordination. This is the aim of the Cotton and Hesse (1992b) method, and of the Rivière (1998, 1999,

2001a,b) method. However, a simple example shows that these methods do not always provide the best longitudinal design for a fixed cross-sectional design on a population with changing strata. Unfortunately the only way, that we know of, to identify this optimal design consists in applying a linear program on all the possible samples. In most situations, this method is not practical.

- To choose a longitudinal systematic design and accept the progressive loss of control over the cross-sectional design.

We propose a new solution that makes a compromise between the control of the longitudinal and of the cross-sectional sampling design. As in Deville's systematic-simple repeated design, we use a longitudinal design that has a larger support than systematic sampling.

### 5.6.2 The minimum time out of sample method

We have seen that any longitudinal design can be applied provided that there is a strictly sequential algorithm to implement it. Being able to inform a unit that it will not be sampled for a fixed minimum number of waves after it has been sampled is a nice feature, so we propose to use the minimum time out of sample method for the longitudinal design. Moreover, this method allows us to set the number of previous waves that can have an influence on the present. If $r$ is the fixed time out of sample, we have that

$$\Pr(S_k^{t+r}=1|S_k^1,\ldots,S_k^{t+r-1})=\Pr(S_k^{t+r}=1|S_k^t,\ldots,S_k^{t+r-1}).$$

At the first wave any cross-sectional sampling design can be applied on the inclusion probabilities $\pi_k^1$ for $k\in U$. Then, using Algorithm 21 or Algorithm 17, it is possible to compute the conditional inclusion probabilities $\Pr(S_k^2=1|S_k^1=s_k^1)$, $k\in U$. At the second wave, any compatible sampling design can be applied on these conditional inclusion probabilities.

At time $t$, the conditional inclusion probabilities $\Pr(S_k^t=1|S_k^1=s_k^1,\ldots,S_k^{t-1}=s_k^{t-1})$ can again be computed with Algorithm 17 or Algorithm 21. Any compatible cross-sectional sampling design can then be applied using these inclusion probabilities. If $t\geq r+1$, one just needs to know $\pi_k^{t-r},\ldots,\pi_k^t$ and $s_k^{t-r},\ldots,s_k^{t-1}$ in order to compute the conditional probabilities. The implementation of this algorithm is thus relatively simple and practical.

## 5.7 Conclusions

We made an attempt to develop a general theory for the problem of units rotation in repeated sampling. The main methods that are currently in use have well known cross-sectional designs and we derived their longitudinal designs. Longitudinal systematic sampling plays a fundamental role in sampling coordination because it provides a good coordination. However, it results in a rapid loss of control over the cross-sectional sampling design. This problem highlights the deep antagonism between control of the cross-sectional design and control of the coordination. Whatever the adopted solution may be, it is not possible to have at the same time the best coordination and a complete choice of cross-sectional design. We offer a compromise that allows us to have a relatively free choice of cross-sectional design while providing a good coordination between the samples.

## Acknowledgements

The authors wish to thank the editor and anonymous referees for their helpful suggestions and comments that helped improve this article. The authors are also grateful to the Swiss National Science Foundation (grant FN205121-105187/1) and the Swiss Federal Statistical Office for the financial support. The views expressed herein are those of the authors and not necessarily those of the Swiss Federal Statistical Office or of the Swiss National Science Foundation.

## Appendix: a new algorithm for unit rotation

The aim of Algorithm 17 was to impose a fixed number of steps during which a unit, once selected, is not selected anymore. However, this is only possible if the sum of the inclusion probabilities for any $r$ successive occasions does not exceed 1. If this condition is not verified, Algorithm 17 can not be applied. One solution that allows for any vector of inclusion probabilities is to use Algorithm 21. However, in this case the minimum time out of sample can not always be respected. Algorithm 21 gives the exact same results as Algorithm 17 when the sums of the inclusion probabilities do not exceed 1.

**Algorithm 21** Minimum time out of sample sequential algorithm, without conditions on the inclusion probabilities.

1: Define $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^T)'$, a vector of inclusion probabilities.
2: Define $\mathbf{s}_k = (s_k^1, \ldots, s_k^T)' = (0, \ldots, 0)'$, the empty sample.
3: Fix $r$, the number of times that a unit must stay out of the sample.
4: **for** $t = 1, \ldots, T$ **do**
5: Generate $u^t$, a uniformly distributed random number in $[0, 1)$.
6: $p = 0$
7: **if** $\pi_k^t \geq 1$ **then**
8: $j = t$
9: **else**
10: $j = \max(t - r - 1, 1)$
11: **while** $\sum_{i=j}^{t} \pi_k^i > 1$ **do**
12: $j = j + 1$
13: **end while**
14: **end if**
15: **if** $j = t$ **then**
16: $p = \pi_k^t$
17: **else**
18: **if** $\sum_{i=j}^{t-1} s_k^i = 0$ **then** $p = \pi_k^t / (1 - \sum_{i=j}^{t-1} \pi_k^i)$ **end if**
19: **end if**
20: **if** $u^t < p$ **then** $s_k^t = 1$ **end if**
21: **end for**

This method can easily be adapted for dynamic populations. A newborn unit, at time $t$, will receive a fictive past, i.e. $\pi_k^i = s_k^i = 0, i \leq t$. The adjustment of Algorithm 21 is straightforward. We just add the following line between lines 20 and 21 of the algorithm:

$20b$ : Add the newborn units to the population.

# Chapter 6

# Conclusion

This work is just a small fish in the vast ocean which is survey sampling. We do not pretend to have made some great invention, we have just made an attempt to bring some new ideas in. An important step has been done towards the reconciliation of the proponents of the model-based and design-based frameworks. If we search for an optimal strategy rather than just an optimal estimator, the strategy that consists of a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model and the Horvitz-Thompson estimator is optimal and robust. This is not a miracle solution to all inference problems in survey sampling, but could be a milestone to look at in case of a misspecification of the model.

We have also seen that the coordination of stratified samples has turned out to be a very interesting problem. We have compared the quality of coordination of the existing methods of coordination of stratified samples is almost equal. It is not difficult to construct new methods, because the technique is intuitive and simple. However, a problem which can not be ignored is, that if we use methods based on microstrata the coordination with very old surveys becomes very difficult due to small sample size.

In our study of the general theory of repeated sampling we have come to the conclusion that there is an antagonism between a good rotation and control over the cross-sectional sampling design. We have seen that longitudinal systematic sampling plays a crucial role in sampling coordination because it provides good coordination. His main drawback is however it that it leads to a rapid loss of control over the cross-sectional design. In order to reach a compromise between the quality of coordination and the freedom of choice of the cross-sectional design, we propose a sampling algorithm that uses a new method of

longitudinal sampling. This method which we called the 'minimum time out of sample method' imposes a fixed number of waves that a unit stays out of the sample once drawn.

# Bibliography


Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, 23:91–113.

Bleuer, S. R. (2002). Report on Rivière's random permutations method of sampling coordination. Report 2002-07-26, Statistics Canada.

Brewer, K. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10:15–30.

Brewer, K. (1999a). Cosmetic calibration for unequal probability sample. *Survey Methodology*, 25:205–12.

Brewer, K. (1999b). Design-based or prediction-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review*, 67:35–47.

Brewer, K. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. Arnold, London.

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.

Brewer, K., Hanif, M., and Tam, S. (1988). How nearly can model-based prediction and design-based estimation be reconciled. *Journal of the American Statistical Association*, 83:128–32.

Brewer, K. and Särndal, C.-E. (1983). Six approaches to enumerative survey sampling. In Madow, W. G., Olkin, I., and Rubin, D. B., editors, *Incomplete data in sample surveys*

*[Proceedings of the Symposium on incomplete data, Washington D.C.]*, volume 3, pages 363–405. Academic press, N.Y.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12:3–32.

Chen, S., Dempster, A., and Liu, J. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81:457–469.

Cochran, W. (1977). *Sampling Techniques*. Wiley, New York.

Cotton, F. and Hesse, C. (1992a). Co-ordinated selection of stratified samples. *Proceedings of Statistics Canada Symposium 92*, 92:47–54.

Cotton, F. and Hesse, C. (1992b). Tirages coordonnes d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Technical report, INSEE, Paris.

Cumberland, W. and Royall, R. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B 43:353–367.

De Ree, S. (1983). A system of co-ordinated sampling to spread response burden of enterprises. In *Contributed paper, 44th Session of the ISI Madrid*, pages 673–676.

De Ree, S. (1999). Co-ordination of business samples using measured response burden. In *Invited paper, 52nd Session of the ISI, Helsinki*.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proc. Workshop on the Uses of Auxiliary Information in Surveys*, pages 21–40, Örebro, Sweden. Statistics Sweden.

Deville, J.-C. (1998). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Technical Report 9804, Méthodologie Statistique, INSEE.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–204.

Deville, J.-C., Grosbras, J.-M., and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In Payne, R. and Green, P., editors, *COMPSTAT, Proceedings in Computational Statistics*. Physica Verlag, Ed. R. Payne and P. Green, pp. 255-66, Heidelberg.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–82.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:411–25.

Fan, C., Muller, M., and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association*, 57:387–402.

Godambe, V. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, B 17:269–78.

Godambe, V. and Joshi, V. (1965). Admissibility and bayes estimation in sampling finite populations I. *Annals of Mathematical Statistics*, 36:1707–22.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Hansen, M., Hurwitz, W., and Madow, W. (1953 reprint in 1993a). *Sample Survey Methods and Theory, I.* Wiley, New York.

Hansen, M., Hurwitz, W., and Madow, W. (1953 reprint in 1993b). *Sample Survey Methods and Theory, II.* Wiley, New York.

Hansen, M., Madow, W., and Tepping, B. (1983). An evaluation of model dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association*, 78:776–807.

Hedayat, A. and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44:237–47.

Hodges, J. and LeCam, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics*, 31:737–740.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Iachan, R. (1984). Sampling strategies, robustness and efficiency: the state of the art. *International Statistical Review*, 52:209–18.

Isaki, C. and Fuller, W. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77:89–96.

Johnson, N., Kotz, S., and Kemp, A. (1992). *Univariate Discrete Distributions*. Wiley, New York.

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18:129–154.

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

Kish, L. (1965). *Survey Sampling*. Wiley, New York.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2:813–830.

Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66:461–470.

Koeijers, E. and Willeboordse, A. (1995). Reference manual on design and implementation of business surveys. Technical report, Statistics Netherlands.

Kott, P. S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *American Statistician*, 40:202–4.

Little, R. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99:546–556.

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

Mészáros, P. (1999). A program for sample co-ordination: Salomon. In *Proceedings of the International Seminar on Exchange of Technology and Knowhow*, pages 125–130, Prague.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95:521–537.

Ohlsson, E. (1995). *Business Survey Methods*, volume 1, chapter 9, "Coordination of samples using permanent random numbers"., pages 153–169. Wiley. inc., New York, USA.

Patterson, H. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, B12:241–255.

Péa, J., Qualité, L., and Tillé, Y. (2007). Systematic sampling is a minimum support sesign. *Computational Statistics and Data Analysis*, 51:5591–5602.

Rivière, P. (1998). Description of the chosen method: Deliverable 2 of supcom 1996 project (part "co-ordination of samples"), pp. 11-33. Report, EUROSTAT.

Rivière, P. (1999). Coordination of samples : the microstrata methodology. In *13th International Roundtable on Business Survey Frames*. Paris: INSEE.

Rivière, P. (2001a). Coordinating samples using the microstrata methodology. *Proceedings of Statistics Canada Symposium 2001*.

Rivière, P. (2001b). Random permutations of random vectors as a way of co-ordinating samples. Report, University of Southampton, INSEE.

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62:135–158.

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62:159–191.

Royall, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377–387.

Royall, R. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–64.

Royall, R. (1988). The prediction approach to sampling theory. In *Handbook of Statistics*, volume 6, pages 399–413. Elsevier Science Publishers, Amsterdam, Holland.

Royall, R. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Method.*, 18:179–185.

Royall, R. and Cumberland, W. (1981). The finite population linear regression estimator and estimators of its variance. an empirical study. *Journal of the American Statistical Association*, 76:924–30.

Royall, R. and Herson, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68:880–9.

Royall, R. and Herson, J. (1973b). Robust estimation in finite populations II: Stratification on a size variable. *Journal of the American Statistical Association*, 68:891–3.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5:27–52.

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

Scott, A., Brewer, K., and Ho, E. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73:359–61.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.

Smith, T. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society*, A 139:183–204.

Smith, T. (1984). Sample surveys, present position and potential developments: Some personal views (with discussion). *Journal of the Royal Statistical Society*, A 147:208–21.

Smith, T. (1994). Sample surveys 1975-1990; An age of reconciliation (with discussion)? *International Statistical Review*, 62:5–34.

Thionet, P. (1953). *La théorie des Sondages*. INSEE, Imprimerie Nationale, Paris.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tillé, Y. and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91:913–927.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Van Huis, L., Koeijers, C., and de Ree, S. (1994a). EDS, sampling system for the central business register at statistics netherlands. Technical report, Statistics Netherlands.

Van Huis, L., Koeijers, C., and de Ree S.J.M. (1994b). Response burden and co-ordinated sampling for economic surveys. Technical report, Statistics Netherlands, Volume 9.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys (with discussion). *Journal of the Royal Statistical Society*, A 109:12–43.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15:235–261.