https://doi.org/10.35662/unine-thesis-3020

Faculté des Sciences
Institut de Statistique

Spatially balanced sampling, stratification and statistical matching

by

Raphaël Jauslin

Thesis submitted in fulfillment of the requirements
for the degree of
Doctorat ès Sciences

Accepted by the examination committee:

Prof.	Pascal	Felber	Université de Neuchâtel	Jury president
Prof.	Yves	Tillé	Université de Neuchâtel	Thesis director
Prof.	Maria	José Lombardia	Universidade da Coruña	Jury member
Prof.	Lorenzo	Fattorini	Università di Siena	Jury member
Prof.	Beat	Hulliger	Fachochschule Nordwestschweiz	Jury member

Thesis defended on 10 March 2023.

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THÈSE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Raphaël JAUSLIN

Titre :

"Spatially balanced, stratification and statistical matching"

sur le rapport des membres du jury composé comme suit :

- Prof. Yves Tillé, directeur de thèse, Université de Neuchâtel, Suisse
- Prof. Pascal Felber, Université de Neuchâtel, Suisse
- Prof. Lorenzo Fattorini, Università di Siena, Italie
- Prof. Marìa-José Lombardia, Universidade da Coruña, Espagne
- Prof. Beat Hulliger, Fachhochschule Nordwestschweiz FHNW, Zürich, Suisse

Neuchâtel, le 30 mars 2023

Le Doyen, Prof. R. Bshary

Résumé

Dans cette thèse, nous nous intéressons à trois champs de la théorie de l'échantillonnage. Ces trois champs sont l'échantillonnage spatial, la stratification et finalement l'appariement statistique. Après un premier chapitre qui rappelle les notions principales de la théorie de l'échantillonnage, la thèse est constituée de deux parties qui contiennent chacune deux chapitres.
La première partie concerne l'échantillonnage spatial. Dans le secteur de l'environnement en particulier, il est important de sélectionner un échantillon bien étalé. Les populations que nous étudions sont souvent auto-corrélées, c'est-à-dire que deux unités proches l'une de l'autre partagent les mêmes caractéristiques et ne devraient pas être sélectionnées dans le même échantillon. Dans le second chapitre, nous proposons une méthode qui permet de sélectionner un échantillon très bien étalé. Le troisième chapitre propose une méthode pour sélectionner un échantillon à la fois étalé sur des coordonnées géographiques et équilibré sur des variables auxiliaires. Cette méthode possède la particularité d'être séquentielle, ce qui offre un champ d'application plus large, notamment dans les très grands ensembles de données.
La deuxième partie de la thèse aborde la stratification et l'appariement statistique. Dans une enquête, on améliore presque toujours l'estimateur si on sépare la population en sous-groupes lorsque cette information est disponible. Ces sous-groupes peuvent être grands ou petits selon les caractéristiques des variables qui les conditionnent. Le quatrième chapitre propose un algorithme pour tirer un échantillon équilibré dans des populations fortement stratifiées. Finalement, le cinquième chapitre parle de l'appariement statistique qui consiste à fusionner deux enquêtes. Nous utilisons le problème du transport optimal pour combiner les deux enquêtes en une pseudo-population qui permet de tirer des conclusions sur des variables mesurées uniquement dans chacune des enquêtes respectives.

Mots-clés Théorie des sondages; échantillonnage étalé; méthode du cube; population fortement stratifiée; transport optimal

Abstract

In this thesis, we are interested in three fields of sampling theory. These three fields are spatial sampling, stratification and finally statistical matching. After the first chapter recapitulating the main notions of sampling theory, the thesis comprises two parts, each containing two chapters. The first part deals with spatial sampling. Particularly in the environmental sector, it is important to select a well-spread sample. The populations we study are often auto-correlated, i.e. two units close to each other share the same characteristics and should not be selected in the same sample. In the second chapter, we propose a method to select a very well-spread sample. The third chapter proposes a method to select a sample that is both spread on geographical coordinates and balanced on auxiliary variables. This method has the particularity of being sequential, which offers a wider scope of application, especially in very large datasets.
The second part of the thesis discusses stratification and statistical matching. In a survey, the estimator is almost always improved by separating the population into subgroups when this information is available. These subgroups can be large or small depending on the characteristics of the variables that condition them. The fourth chapter proposes an algorithm for drawing a balanced sample in highly stratified populations. Finally, the fifth chapter discusses statistical matching, which consists in merging two surveys. We use the optimal transport problem to combine the two surveys into a pseudo-population that allows conclusions to be drawn on variables measured only in each of the respective surveys.

Keywords Survey sampling; spread sampling; cube method; highly stratified population; optimal transport

Remerciements

Je remercie en premier lieu mon directeur de thèse le Professseur Yves Tillé pour l'accompagnement dont il a fait preuve pour moi. Son intuition statistique et sa passion pour le sondage et l'échantillonnage ont été une grande inspiration. Je voudrais tout particulièrement le remercier pour la proximité dont il fait preuve avec ses doctorants. Ceci a assurément participé à l'accomplissement de ma thèse. Yves, ce fut un immense plaisir de travailler avec toi ces quatres dernières années.

Je remercie les membres du Jury, la Professeure Maria José Lombardia, le Professeur Pascal Felber, le Professeur Lorenzo Fattorini et le Professeur Beat Hulliger pour leur minutieuse relecture et leurs commentaires qui vont assurément améliorer la qualité de ma thèse.

Je remercie aussi tous mes collègues de l'Institut de Statistique. Un remerciement tout particulier au Professeur Bardia Panahbehagh pour notre collabortation constructive lors de son séjour en Suisse en 2021. Merci aussi à Audrey-Anne, Esther, Arnaud, Ejub et Ziqing, pour tous les bons moments que nous avons passés ensemble autour d'un café ou sur une planche d'équilibre. Je remercie aussi Alina, Caren, Clément, Lionel, Michael et Pierre-Yves pour leur bienveillance et leur soutien. Finalement, je remercie Corine pour toute l'aide qu'elle m'a apporté au sein de l'Institut de Statistique.

Je remercie ma famille et mes amis, plus particulièrement Deirdre, Cliona, Moran et mon cher Papa, les premiers curieux relecteurs, doté toujours de commentaires constructifs malgré un contenu parfois un peu abstrait pour eux.

Finalement je remercie mon épouse Alice, pour son soutien indéfectible, ses encouragments et son réconfort au quotidien. Du premier jour de cours à l'EPFL au dernier jour de ma thèse, elle m'a toujours soutenu et je la remercie pour tous ce qu'elle m'a apporté et ce qu'elle m'apportera encore dans le futur.

Contents

Abstract iv

Remerciements ix

Contents xi

List of Figures xv

List of Tables xvii

List of Algorithms xix

Introduction 1

1 Survey sampling 5
- 1.1 Probability sampling . . . 5
- 1.2 Auxiliary information . . . 8
- 1.3 Spatially balanced sampling . . . 9

I Spatially balanced sampling 13

2 Spatial Spread Sampling Using Weakly Associated Vectors 15
- 2.1 Introduction . . . 15
- 2.2 Notation . . . 17
  - 2.2.1 Basic setup . . . 17
  - 2.2.2 Well-spread sample . . . 18
- 2.3 Weakly Associated Vector Sampling . . . 18
  - 2.3.1 General idea . . . 18
  - 2.3.2 Distance . . . 19
  - 2.3.3 The stratification matrix . . . 20
  - 2.3.4 Implementation . . . 22
- 2.4 Spatial balance . . . 23
  - 2.4.1 Voronoï polygons . . . 23
  - 2.4.2 Moran's I index . . . 25
- 2.5 Variance estimation . . . 26
- 2.6 Simulations on artificial spatial configurations . . . 28
- 2.7 Application to the Meuse dataset . . . 29
- 2.8 Discussion . . . 33
- 2.9 Appendix . . . 33

3 Sequential Spatially Balanced Sampling 37
3.1 Introduction . . . 37
3.2 Notation . . . 38
3.3 Balanced Sampling . . . 40
3.4 Spreading Measures . . . 41
3.5 Outline of the Proposed Method . . . 42
3.6 Variance Estimation . . . 45
3.7 Simulations . . . 48
3.7.1 Motivation on an artificial dataset . . . 48
3.7.2 Real example on amphibians dataset . . . 49
3.8 Conclusion . . . 52
3.9 Acknowledgements . . . 55

II Highly stratified sampling and statistical matching 57

4 Enhanced Cube Implementation For Highly Stratified Population 59
4.1 Introduction . . . 59
4.2 Basic sampling notations . . . 60
4.3 Stratified balanced sampling . . . 61
4.4 Cube Method . . . 62
4.5 Highly stratified population . . . 62
4.6 Proposed method . . . 63
4.7 Variance estimation . . . 65
4.8 Simulations . . . 66
4.9 Conclusion . . . 68

5 An Efficient Approach for Statistical Matching of Survey Data Through Calibration, Optimal Transport and Balanced Sampling 71
5.1 Introduction . . . 71
5.2 Problem and notation . . . 72
5.3 Harmonization by calibration . . . 74
5.4 Renssen's methods . . . 75
5.5 Matching by optimal transport . . . 76
5.5.1 Matching by using prediction . . . 77
5.5.2 Matching by using stratified balanced sampling . . . 78
5.6 Analysis of the data . . . 79
5.7 Simulations . . . 80
5.7.1 Gaussian example . . . 80
5.7.2 EU-SILC example . . . 82
5.8 Conclusion . . . 85

6 Conclusion 87

Bibliography 89

List of Figures

1.1 Example of coherent subset and incoherent subset. We suppose only spatial coordinates with Euclidean distance. 10

2.1 Simple example of a 3×3 regular grid set up on three different distances with a gradient calculated from the points $(1, 1)$. The left one is the classical Euclidean distance (2.4), the right one is the tore distance given in (2.5) and the central graph is the shifted tore distance with a shift equal to $(1/12, 1/4)$ (the black point on the graph). It illustrates the two different patterns and the values of the grid points corresponding to the entries of the first row of the three previous matrices (2.6). . 21

2.2 Sparsity pattern of three stratification matrices. Spatial coordinates are 3×3 regular grid and the inclusion probabilities are equal to $\pi = (1/3, \ldots, 1/3)$. Depending on the way of defining the nearest neighbours in Equation (2.8), different weight values are obtained. The left stratification matrix uses the classical Euclidean distance (2.4), the central one the tore distance (2.5) and the right one uses a shifted tore distance with a shift randomly generated from a random variable $\mathcal{N}(0, 1/100\mathbf{I})$. 22

2.3 Representation of the strata defined by the spatial weights Equation (2.8). Spatial coordinates of the units are generated randomly from a uniform distribution on the square unit $[0, 1] \times [0, 1]$. The overall population size is equal to $N = 250$ and the inclusion probabilities are identical and equal to $\pi_k = 1/25 = 0.04$. Meaning that the sample size is equal to $n = 10$. With these parameters the expected number of units in each stratum is equal to $1/0.04 = 25$. The left graph shows the population and the selected units with its initial strata. On the right, it shows the sparsity pattern of the matrix (2.9). All entries of the matrix are equal to 1. 23

2.4 Illustrated example of how the spatial balance measure based on the Voronoï is performed. The population and sample sizes are respectively equal to $N = 50$ and $n = 20$, the inclusion probabilities are identical and equal to $\pi_k = 0.4$. The spatial coordinates are generated from two random uniform $\mathcal{U}(0, 1)$. Two sampling design are compared. The left one is the simple random sampling without replacement and the right one is the weakly associated vector sampling. 25

2.5 Example of a sample selection by the WAVE sampling on the three different spatial configurations, Complete Spatial Randomness, Neyman-Scott and regular grid. For each of them, the inclusion probabilities are equal to $\pi_k = n/N$ for all $k \in U$. 28

2.6 Example of WAVE sampling on the Meuse dataset. The overall population size is equal to 155. The inclusion probabilities are proportional to copper level variable and the sample size is equal to 30. Plotted sizes of the units are proportional to the copper concentration. The Meuse River is filled in light blue. 30

2.7 Let $U = \{1, \ldots, 36\}$ be on a regular grid 6×6, inclusion probabilities are all equal to $\pi_k = \pi = 1/6$. Measure B given in Equation (10) is equal to 0 for the clustered sample while it should be a great value. $B = 0.106$ for the spread sample. On the other hand the Moran index shows a correct behaviour with the two samples. For the clustered sample $I_B =$ 0.71 while $I_B =$ -0.718 for the spread sample. 35
2.8 Let $U = \{1, \ldots, 36\}$. On the left figure, population is on a regular grid 6×6. On the right figure, population is clustered around four units. Inclusion probabilities are all equal to $\pi_k = \pi = 1/9$ (respectively $\pi_k = \pi = 1/12$). Measure B for both examples is equal to 0 while it should be a higher value. On the other hand, the Moran index shows correct behaviour with the two samples. For the Example 1 $I_B = 0.732$ while $I_B = 0.199$ for Example 2. 35

3.1 For six different inclusion probabilities π_i, we let π_k vary from 0 to 1. On the y-axis, the bounds of v_k in Equations (3.6) are calculated. The colored area represents eligible values of v_k, the shaded line is the upper bound, while the bottom line is the lower bound. 45
3.2 Simulated dataset used for the analysis of Section 3.7.1. The two datasets each contain 300 units. 52
3.3 The upper plot displays the biogeographical region of Switzerland, while the lower plot gives the different sites of the amphibians dataset 53

4.1 Illustration of how the **B** matrix is calculated within the flight phase in the discussed methods. The light grey area represent the auxiliary variable matrix **A** while the dark grey area is the **H** matrix. The **B** matrix is illustrated by the hatched area. The medium grey area is the one inside of the matrix **B** where all values are equal to zero. On the left, we see how the matrix **B** is computed in Chauvet and Hasler methods, while on the right, it is the matrix proposed by Algorithm 3. We see that the matrix **B** on the left contains some columns that are only equal to zero. 65
4.2 Data extracted from the Swiss establishments data base of the Swiss Federal Statistical Office (2020). The data are restricted to the NUTS region 2. The upper plot is showing the separation by Cantons $H_c = 5$, the lower one the separation by Municipalities $H_m = 675$. The grey gradient scale gives the number of units considered in each Canton. The data are selected such that each municipality contains 3 units. . . . 67

5.1 Representation of the statistical matching with intersection. Hatched areas are unknown quantities in each sample. 73
5.2 Outline of the statistical matching using optimal transport. 78
5.3 Boxplot of the household income by economic status. 84

List of Tables

2.1 Spreading measures results based on 10000 simulations on the Complete spatial randomness dataset. The population size is equal to 144. . 29
2.2 Spreading measures results based on 10000 simulations on the Meuse dataset. The population size is equal to 155. 31
2.3 Results of 10000 simulations on Meuse dataset. The population size is equal to 155. v_{SIM} is equal to the variance approximated by the simulations (2.16). $\widehat{\text{var}}$ depends on the sampling design. For the srswor and maxent methods, we used the estimator $\widehat{\text{var}}_{HAJ}$ (2.13) while for the other sampling designs, we use $\widehat{\text{var}}_{SB}$ (2.14). Coverage rate of the 95% confidence intervals are computed as well as the ratio between averages of $\widehat{\text{var}}$ and v_{SIM}. 32
2.4 Results of 10000 simulations on Meuse dataset. The population size is equal to 155. v_{SIM} (2.16) is equal to the variance approximated by the simulations. $\widehat{\text{var}}_{SB}$ (2.14) is the variance estimator based on the nearest neighbours in the sample. $\widehat{\text{var}}_{LM_j}$ is equal to the estimator (2.15) where the number of neighbouring units used is set to $j = 2, 3, 4$. v_{HAJ} (2.13) is the Hajek-Rosen estimator. 32
2.5 Results of 10000 simualtions of Moran's I_{B_1} spatial measure. The three spatial configurations of the Section 6 are taken into account. 33
2.6 Results of 10000 simulations of Moran's I_B spatial measure. The three spatial configurations of the Section 6 are taken into account. 34
2.7 Results of 10000 simulations of the spatial balance measure B. The three spatial configurations of the Section 6 are taken into account. . . 34

3.1 Results of 10,000 simulations on the variables of interest (3.12). The first column represents relative variance reduction (3.15). The second columns contains the relative variance estimator efficiency (3.16). The third and fourth columns correspond to the two spatial measures (3.3) and (3.4), respectively. 50
3.2 Results of 10,000 simulations of the relative deviation (3.13) on the artificial dataset presented in Section 3.7.1. 51
3.3 Results of 10,000 simulations on the diversity score of the amphibian dataset. The first column represents the relative variance reduction (3.15). The second columns contains the relative variance estimator efficiency (3.16). The third and fourth columns correspond to the two spatial measures (3.3) and (3.4) respectively. 54
3.4 Results of 10,000 simulations of the relative deviation (3.13) for the amphibians dataset presented in Section 3.7.2. 54

4.1 Results of 1000 simulations on the Swiss establishments dataset. The population size is equal to 2025. We compute the mean time execution in seconds of each sampling procedure. We vary the number of strata H and the number of units selected within each stratum n_h. 70

4.2 Results of 1000 simulations on a population of size 2025. The number of strata is equal to 5 for Cantons and 675 for Municipalities. For each variable of interest y_j, $j = 1, 2, 3$ and for each sampling methods, we compute the ratio between the different estimators (i.e. approximated variance (4.5) as well as the variance estimator (4.6)) and the variance approximated by the simulations (4.7). 70

5.1 Mean squared errors and their bias-variance decomposition on 10 000 simulations of the estimation of the variance-covariance matrix $\mathbf{\Sigma}_{yz}$. Matrix notation $\mathbf{B}$ stands for the squared bias while $\mathbf{V}$ denote the variance of the estimator $\widehat{\mathbf{\Sigma}}_{yz}$. 81

5.2 Selected variables of the `eusilc` dataset of the `R` package developed by Alfons and Templ (2013). The first five variables are the ones used for the matching while the two last ones are the variables of interest. . 83

5.3 Results of 10 000 simulations for the estimation of average income per category. Relative root mean squared errors as well as the relative bias are calculated for each economic status. The overall mean squared errors are equal to 47.443 for the optimal transport, 61.154 for the balanced imputation and 53.065 for the method of Renssen. 84

List of Algorithms

1 Algorithm for WAVE sampling 24

2 Algorithm of sequential balanced sampling 46

3 Find the submatrix **B** of (**H A**) 64

4 Complete proposed algorithm for highly stratified population 69

Introduction

Survey sampling is a field of statistics that studies finite populations. There are different steps for statisticians working in the field of survey sampling. The first step is data collection. All information that might be useful about the target population is collected, and a database is created containing all the units on which the sample will be selected. At this stage, some auxiliary variables can be used to establish unequal inclusion probabilities and increase the precision of future estimators. Next, a random sampling design is used to draw a sample. Statisticians working in survey sampling always strive to find appropriate sampling designs to obtain better estimates. Next, the variable of interest is recorded on the sampled units, and various treatments are applied to deal with non-response and measurement errors. Finally, an estimation of the target variable is performed, and variance estimation is generally achieved. All of these steps are of concern to survey statisticians and have room for improvement and constitute a vast area of research. In this research, we mainly focus on the process of a random selection of the sample. This thesis is an assortment of published articles. It investigates three main fields of survey sampling theory, spatial statistics, stratification and statistical matching.

Spatial statistics has been a productive field in recent decades. Especially in environmental surveys, data are auto-correlated meaning that two units close to each other share the same characteristics. In this context, the survey statisticians need to use a sampling method that selects a well-spread sample to ensure that units close to each other are not selected in the sample. There are simple methods such as systematic sampling or cluster sampling to select a well-spread sample in space. Systematic sampling has some disadvantages. The method is not generalizable to more than one dimension with prescribed unequal inclusion probabilities, thus unbiased estimation of sampling variance is not possible. Clustered sampling divides the space into strata and selects a few units in each stratum. Depending on the spatial coordinates, it may be difficult to construct strata that have the same sizes. In this thesis, we propose two methods that can select a well-spread sample. They handle equal and unequal inclusion probabilities and every spatial configuration. They increase the quality of the estimators.

Stratification is a sampling strategy that consists of partitioning the population into subgroups and then selecting a small sample from each subgroup. Stratification is a useful technique for incorporating auxiliary information into the sample design and reducing sampling error. In official statistics, it occurs naturally with auxiliary variables that could define strata, for example, gender, age or any categorical variables. Stratification can also be defined by statisticians to use the appropriate structure to reduce errors. If the information is available, the use of stratified sampling techniques almost always reduces the errors of the estimators. This thesis presents a new algorithm to select a sample in a highly stratified population.

Statistical matching occurs when two surveys have been performed on the same population, but the quantities recorded are different. There are two different approaches to statistical matching. The micro approach aims to complement one file with information from the other. The macro approach focuses on the characteristics between these two variables of interest, for example, contingency tables if the variables are categorical or correlation if the variables are continuous. In addition, sur-

veys usually come with their own weighting scheme. This weighting scheme may have been used to account for unequal inclusion probabilities or non-response treatment. Weights are generally calculated to allow the files to be extrapolated to the full population. These weights must be taken into account in the merging process. In this thesis, we present an elegant method of merging two surveys. The first step consists of harmonizing the weight schemes. Then, the method uses the optimal transport problem to create a pseudo-population to analyse correlations or contingency tables.

After the first introductory chapter on sampling theory, the thesis is separated into two parts, each containing two chapters. Spatial sampling is discussed in the first part while stratification and statistical matching are discussed in the second part.

Chapter 1 is a brief introduction to the sample survey framework. It introduces the concept of probability sampling design and the Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Narain, 1951). It presents the concept of balanced sampling that will be used throughout this thesis.

In Chapter 2, which is a reprint of Jauslin and Tillé (2020), we discuss a spatial sampling method that defines an elegant way to create strata with the aim of selecting a well-spread sample. Spatial sampling is a broad field. Many articles have been published on spatial sampling designs, Stevens Jr. and Olsen (2004); Grafström and Tillé (2013); Robertson et al. (2018) to mention just a few. Here we propose a new method that can treat equal or unequal inclusion probabilities. We also show that in some cases where the distance is carefully chosen, we obtain a regularity suggestive of systematic sampling in one dimension. It was also of interest to think about what systematic sampling in more than one dimension with unequal inclusion probabilities might look like. Another novelty is a modification to measure the spread of the sample. Tillé et al. (2018) have normalized Moran's I index to be used in a more general context than just on a grid. In this paper, we propose another modification of Moran's I index, that further increases the precision in the context of unequal inclusion probabilities.

In Chapter 3, which is a reprint of Jauslin et al. (2022b), we develop a method to select a sample that is both well-spread on geographical coordinates and balanced on a set of auxiliary variables. Grafström and Tillé (2013) have already proposed a similar method. Nevertheless, the method we propose has the particularity of being sequential. It means that the method does not need the entire dataset and can be managed one unit after the other. This feature can be very interesting for stream populations or for a sizeable dataset.

In Chapter 4, which is a reprint of Jauslin et al. (2021), we consider the problem of a highly stratified population. A population is highly stratified when the number of different strata is large. In this case, the usual technique, e.g. (Hasler and Tillé, 2014; Chauvet, 2009), becomes very time-consuming, especially if the sampling design has unequal inclusion probabilities. The proposed technique is an improved algorithm that reduces the computational problem. The method appears to be valuable for other sampling methods. It is already used in another approach developed to do statistical matching which is presented in Chapter 3 and a method to select a spatio-temporal sample with optimal rotation of unit in time (Eustache et al., 2022a).

In Chapter 5, which is a reprint of Jauslin and Tillé (2023), we present a new method to do statistical matching. It is common that two survey samplings have been done on the same population but where the variable of interest recorded is not the same. In that case, we are interested in merging the two analyses so that we can have some results on the contingency table or correlation between variables of interest. The pro-

posed method uses the optimal transport problem and the stratified balanced sampling algorithm to perform statistical matching.

Chapter 1

Survey sampling

Survey sampling is a field of statistics that studies the characteristics of a finite population. Throughout this manuscript, the population will be denoted U and supposed to be discrete and with cardinality N. The population is labelled from 1 to N

$$U = \{1, \dots, N\}.$$

Some characteristics are of interest on this population and are denoted as variable $y_k, k \in U$. This quantity is called the variable of interest. The survey statistician is interested in the characteristics (or functions) of the variable of interest. For example, the population total or the mean

$$Y = \sum_{k \in U} y_k, \overline{Y} = \frac{1}{N} \sum_{k \in U} y_k \tag{1.1}$$

Sometimes the population total is unknown and is also subject to estimation through the relation

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

To estimate these quantities, one possibility is to gather all the values of the population. This is called a census. But due to cost or time constraints, collecting all the values of the variable of interest is generally impracticable. The survey statistician uses a sample to estimate the functions of interest such as the population total. It is important to emphasize here that the variable of interest has no source of randomness. In fact, in this thesis, we restrict ourselves to the design-based approach. This approach supposes that the population is a complete reflection of reality and that the only source of randomness comes from the sample selection process. Another approach exists and is called model-based. It supposes that the population U and the variables $y_k, k \in U$ are realizations of a super population model. These two approaches are completely different. If the reader is interested in a model-based approach one could refer to Valliant et al. (2013).

1.1 Probability sampling

To estimate functions of variables of interest such as (1.1), the survey statistician uses sampling design to then conduct a survey on this subset of the population.

Definition 1.1.1 *Let $\mathcal{S} = \{s \mid s \subset U\}$ define the power set of U. A sampling design is a probability function $p : \mathcal{S} \to [0, 1]$ such that*

$$p(s) \geqslant 0 \text{ for all } s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1.$$

Definition 1.1.2 *A sample S is a random variable that maps the elements of $\mathcal{S}$ to an N vector of 0 or 1 such that*

$$P(S = s) = p(s) \text{ for all } s \in \mathcal{S}.$$

Define $\delta_k(s)$, $k = 1, \ldots, N$, the indicator variable

$$\delta_k(S) = \begin{cases} 1 & \text{if } k \in S \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we write $\delta_k(S) = \delta_k$. There exist then a complete bijection between the random variable S and the random vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$. A sample s is a realization of the random vector $\boldsymbol{\delta}$.

Definition 1.1.3 *Inclusion probability π_k, for all $k = 1, \ldots, N$, is the probability that a particular unit k is being collected in the random sample S. This inclusion probability can be deduced from the sampling design*

$$\pi_k = P(k \in S) = \mathrm{E}_p(\delta_k) = \sum_{s \in \mathcal{S} | s \ni k} p(s), \text{ for all } k \in U.$$

where $\mathrm{E}_p(.)$ is a notation to stress that the expectation is taken on the sampling design p. Second-order inclusion probability $\pi_{k\ell}$ is the probability that both unit k and ℓ are selected in the random sample S:

$$\pi_{k\ell} = P(k \in S \text{ and } \ell \in S) = \mathrm{E}_p(\delta_k \delta_\ell) = \sum_{s \in \mathcal{S} | s \ni \{k\ell\}} p(s).$$

In without replacement sampling design, δ_k, $k = 1, \ldots, N$ are Bernoulli distributed with parameters π_k. In general, the survey statistician does not deduce the inclusion probabilities from the sampling design but sets them up in advance to fulfil the right requirements. Some sampling design does not necessarily respect fixed sample size. This means that the sum of the random vector $\boldsymbol{\delta}$ is not necessarily equal for all realizations. The expectation of the sample size can be calculated from inclusion probabilities through the relation

$$\mathrm{E}_p(n_s) = \mathrm{E}_p\left(\sum_{k \in U} \delta_k\right) = \sum_{k \in U} \mathrm{E}_p(\delta_k) = \sum_{k \in U} \pi_k.$$

where n_s is a notation to emphasize that this is a random variable and not a scalar. To estimate the total (1.1), Narain (1951); Horvitz and Thompson (1952) have proposed the following estimator.

Definition 1.1.4 *The Horvitz-Thompson estimator (also called the expansion estimator) of the total Y is defined as:*

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k \delta_k}{\pi_k}. \tag{1.2}$$

Note that the last notation is only valid if $\pi_k > 0$ for all $k \in U$. Throughout this thesis, all inclusion probabilities are supposed to be greater than zero. Indeed, a well-known result from survey theory is the following:

Result 1.1.1 *The Horvitz-Thompson estimator is unbiased if and only if we have that $\pi_k > 0$ for all $k \in U$.*

Proof. Directly we have,

$$\begin{aligned} \mathrm{E}_p(\widehat{Y}) &= \mathrm{E}_p\left(\sum_{k\in U|\pi_k>0}\frac{y_k\delta_k}{\pi_k}\right) = \sum_{k\in U|\pi_k>0}\frac{y_k\mathrm{E}_p(\delta_k)}{\pi_k} \\ &= \sum_{k\in U|\pi_k>0}\frac{y_k\pi_k}{\pi_k} = \sum_{k\in U|\pi_k>0} y_k \\ &= \sum_{k\in U} y_k - \sum_{k\in U|\pi_k=0} y_k. \end{aligned}$$

The bias is then equal to

$$\mathrm{E}_p(\widehat{Y}) - Y = -\sum_{k\in U|\pi_k=0} y_k,$$

which is null if and only if $\pi_k > 0$ for all $k \in U$.

□

Result 1.1.2 *The variance of the Horvitz-Thompson estimator is equal to*

$$var_p(\widehat{Y}) = \sum_{k\in U}\sum_{\ell\in U}\frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell},$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$. An estimator of this variance is given by

$$\widehat{var}_p(\widehat{Y}) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

If the sampling design is of fixed sample size, the variance could be rewritten

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

Moreover, Sen (1953); Yates and Grundy (1953) have proposed the estimator

$$\widehat{var}_{SYG}(\widehat{Y}) = -\frac{1}{2}\sum_{k\in S}\sum_{\ell\in S}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2\frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

This estimator can take negative values, but if $\Delta_{k\ell} \leqslant 0$ for all $k \neq \ell \in U$, then the estimator is always positive. The proof of these results can be found in Tillé (2020, Chapter 2).

Definition 1.1.5 *A sampling design that assigns the same probability for each sample of a fixed sample size is called simple without replacement. The probability of each sample $s \in \mathcal{S}$ is then equal to*

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

where $n \in \{1, \ldots, N\}$. The first order inclusion probabilities of this sampling design are $\pi_k = n/N, k \in U$, and the Horvitz-Thompson estimator simplifies then to

$$\widehat{Y} = \frac{N}{n}\sum_{k\in S} y_k.$$

Note that simple random sampling without replacement does not define a way of selecting the sample. It just says that if the probability of selecting a particular sample is equal to $p(s) = \binom{N}{n}^{-1}$, then it is a without replacement sampling design. There are many ways of selecting the samples and different algorithms can lead to the same sampling design. Tillé (2020, Chapter 3) gives many algorithms that lead to simple random sampling without replacement. The survey statistician must then also be aware of the algorithm that he used to select the sample. Depending on the population some algorithms could be time-consuming and not efficient.

1.2 Auxiliary information

In most cases, the survey statistician also has auxiliary information. Throughout this thesis, auxiliary information will be denoted as vector $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \ldots, x_{kq}) \in \mathbb{R}^q, k \in U$. As for the variable of interest, we can define the total of each auxiliary variable as

$$X_i = \sum_{k \in U} x_{ki}, i = 1, \ldots, q.$$

Using the vector notation, we obtain

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$$

Note that using this notation, $\mathbf{X} = (X_1, \ldots, X_q) \in \mathbb{R}^q$ while generally the bold capital letter is used for matrix notation. The notation $\mathbf{X} \in \mathbb{R}^q$ and $\mathbf{Y} \in \mathbb{R}^r$ are the only exceptions to this thesis. If the sample is known, we can, as for the variable of interest, calculate the estimation of the totals,

$$\widehat{X}_i = \sum_{k \in S} \frac{x_{ki}}{\pi_k}, i = 1, \ldots, q.$$

and in vector notation,

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

Definition 1.2.1 *A sampling design $p(s)$ is said to be balanced on the q auxiliary variables if and only if it satisfies the following balancing equation:*

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

Deville and Tillé (2004) developed the cube method that selects a balanced sample. This method will be presented in more details in the next chapters. As well as the auxiliary variable, the population in environmental surveys contains spatial coordinates $\mathbf{z}^\top = (z_{k1}, z_{k2}, \ldots, z_{kp}) \in \mathbb{R}^p$, where p is the dimension of the considered space. Spatial coordinates can be seen as auxiliary information. Grafström and Lundström (2013) have shown that well-spread samples are also balanced. We discuss this statement in the second and third chapters of this thesis. In an ideal case, the more the sample is well-spread the better the estimates. Different measures exist to assess if a sample is well-spread or not.

Another way of incorporating auxiliary information in sampling design is by using stratification. Due to limitations, e.g. data collection is complex because of high distance or terrain, we stratify the population into several subgroups named strata.

Definition 1.2.2 *Suppose that the population is divided into H strata $U = \{U_1, U_2, \ldots, U_H\}$ with respective sizes of $N_1, N_2, \ldots, N_H$. We say that the population is stratified if the strata form a partition with the following properties:*

$$U = \bigcup_{h=1}^{H} U_h, N_h > 0, U_h \cap U_j = \emptyset, \text{ for all } h, j \in 1, \ldots, H.$$

The sizes of the strata sum to the population size,

$$\sum_{h=1}^{H} N_h = N.$$

Let define $\mathbf{h}^\top = (h_1, \ldots, h_N) \in \mathbb{N}^N$ be a categorical vector that specifies the stratum to which each unit belongs i.e. $h_k = j$ means that unit k belongs to strata U_j, $k \in U$ and $j \in \{1, \ldots, H\}$. Let $\mathbf{1}(U_j) \in \{0, 1\}^N$ the vector where the kth element is equal to 1 if the unit k belongs to the stratum U_j and 0 otherwise. Let $\mathbf{H}$ be the $N \times N$ disjunctive matrix of the corresponding vector $\mathbf{h}$ such that

$$\mathbf{H} = (\mathbf{1}(U_1), \mathbf{1}(U_2), \ldots, \mathbf{1}(U_H)).$$

The aim is still to estimate the total of a variable of interest y. The total Y can be then rewritten

$$Y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} Y_h,$$

where Y_h is the total of the variable y for the stratum U_h.

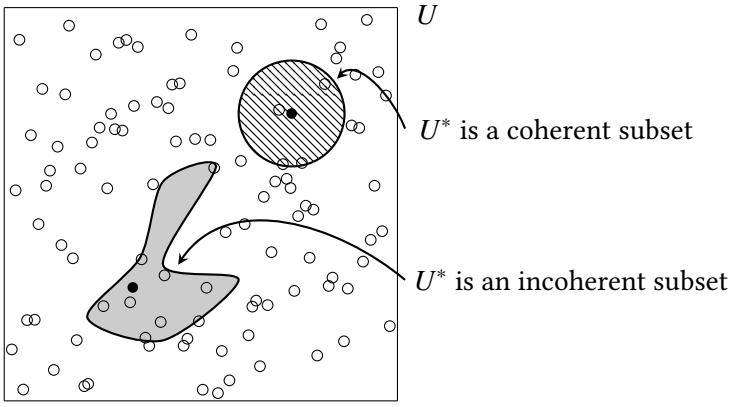
1.3 Spatially balanced sampling

When estimating the population total (1.2) the aim is to select a sample so that characteristics of the population can be estimated with a relatively good precision. When the population of interest came with spatial coordinates and inclusion probabilities $\pi_k,\ k = 1, \ldots, N$ are equal, a spatially balanced sample can be intuitively defined as a sample such that

1. The distances between selected units are equal.
2. In average, the selected samples cover the entire spatial configuration.

This informal definition is related to the intuitive thinking, sometimes incorrect, that a sample should be representative of the population, i.e. a scaled-down version of the population. Indeed, representativeness of a sample is generally a wrong intuitive thinking in sampling theory. We generally over sample some part of the population where the variance is larger, this at the end reduce the errors due to the sampling design. That said, in environmental context representative sample plays an important role. In particular, Grafström and Schelin (2014) have proposed a formal definition of the representativeness of a sample which give an appropriate definition of what is a spatially balanced sample.

Figure 1.1 – *Example of coherent subset and incoherent subset. We suppose only spatial coordinates with Euclidean distance.*



Definition 1.3.1 *Let $i \in U$ and $r \geqslant 0$, we say that U^* is a coherent subset of U if for all $j \in U$, j is included in U^* if and only if $d(i, j) \leqslant r$, where $d(i, j)$ is the distance between the units i and j.*

This definition is of course dependent of the distance. Depending on the presence of spatial coordinates or auxiliary variables, we could use Mahalanobis distance, Euclidean distance or Hamming distance.

Definition 1.3.2 *A sample $s \subset U$ is said to be representative if for all coherent subset $U^* \subset U$, we have*

$$n^* \approx \frac{n}{N} N^*$$

where n^* is the number of sampled units from U^* and N^* the size of U^*.

In other words, a sample is representative if for every coherent subset $U^* \subset U$, the proportion of sampled units in the coherent subset is equal to the one at the population level.

Definition 1.3.3 *A sample s is said to be spatially balanced or well-spread with respect to the inclusion probabilities if for all coherent subset $U^* \subset U$, we have*

$$n^* \approx \sum_{k \in U^*} \pi_k$$

where n^* is the number of sampled units from U^*.

Note that in presence of equal inclusion probabilities $\pi_k = n/N$, this is exactly equal to the definition of representative sample. Indeed, we have

$$\sum_{k \in U^*} \pi_k = \sum_{k \in U^*} \frac{n}{N} = \frac{n}{N} N^*.$$

This definition is completely in agreement of the intuition that a well-spread sample with equal inclusion probabilities is representative. But in presence of unequal inclusion probabilities, a sample can be spatially balanced but far from being representative. Stevens Jr. and Olsen (2004) have also proposed another way to defining a well-spread sample. They proposed to use the Voronoï polygons to measure the spread of a sample.

Definition 1.3.4 *Let define $V_k \subset U$, the kth Voronoï polygon, as the subset of population units $j \in U$ that satisfy*

$$d(k, j) \leqslant d(i, j). \text{ for all } i \in S.$$

Less formally, V_k is the subset of all population units that are closer to the kth unit than any other units selected in the sample. Let v_k denote the sum of the inclusion probabilities inside $V_k, k \in S$. By observing that

$$\sum_{k \in S} v_k = \sum_{j \in U} \pi_j = n,$$

and from the following equality:

$$\frac{1}{n} \mathrm{E}\left(\sum_{k \in S} v_k \right) = \frac{1}{n} \sum_{k \in S} \mathrm{E}(v_k) = 1,$$

we can deduce that $\mathrm{E}(v_k) = 1$ for all $k \in S$. Stevens Jr. and Olsen (2004) have defined then a measure of the spread B based on the estimated variance of this expectation:

$$B = \frac{1}{n} \sum_{k \in S} (v_k - 1)^2.$$

The closer this value is to zero, the better the sample is well-spread. Hence, this gives us another definition of a well-spread sample.

Definition 1.3.5 *A sample is said to be well-spread or spatially balanced with respect to the inclusion probabilities if for all Voronoï polygons, v_k is equal or close to* 1.

In the first part of this thesis we also introduce another way of defining a well-spread sample. This quantity is called the Moran index (Moran, 1950) and measure the spatial auto-correlation of the sample. A complete Section is dedicated on this measure in Chapter 2.

Part I

Spatially balanced sampling

Chapter 2

Spatial Spread Sampling Using Weakly Associated Vectors

Abstract

Geographical data are generally auto-correlated. In this case, it is preferable to select spread units. In this paper, we propose a new method for selecting well-spread samples from a finite spatial population with equal or unequal inclusion probabilities. The proposed method is based on the definition of a spatial structure by using a stratification matrix. Our method exactly satisfies given inclusion probabilities and provides samples that are very well-spread. A set of simulations shows that our method outperforms other existing methods such as the Generalized Random Tessellation Stratified (GRTS) or the Local Pivotal Method (LPM). Analysis of the variance on a real dataset shows that our method is more accurate than these two. Furthermore, a variance estimator is proposed. [1]

Keywords: GRTS, local pivotal method, cube method, stratification

2.1 Introduction

In most natural resources surveys, data contains spatial coordinates. The process of selecting units from a population defined over a region of space is called spatial sampling. These kinds of data are usually auto-correlated, meaning that two close measurements are generally similar. In general, to estimate a total of a target variable, selecting the units spatially best spread increases information colection and provides better estimation. An important problem of spatial sampling is thus to optimize the spread of the sampled units in space. A well-spread sample is called spatially balanced. Grafström and Lundström (2013) and Grafström and Schelin (2014) give the formal definition of a representative sample and discuss the theoretical justification of taking a well-spread sample with unequal probabilities. Marker and Stevens Jr. (2009) and Hankin et al. (2019) present some examples of studies where the population considered is in an environmental context such as lakes, wetlands, rangelands, and forests. Vallée et al. (2015) discuss forest ecosystem evolution using a well-spread spatial sampling design. Tillé (2020, Chapter 8), Tillé and Wilhelm (2017), Benedetti et al. (2017) and Wang et al. (2012) give a review of the main spatial sampling methods. Quenouille (1949) and Bellhouse (1977) showed that systematic sampling is the optimal design for auto-correlated data.

Generalized Random Tessellation Stratified (GRTS) sampling is a spatial sampling method proposed by Stevens Jr. and Olsen (1999, 2003, 2004). They use a mapping by means of a quadrant-recursive function to map a finite subset of a multi-dimensional

[1] This chapter is essentially a reprint of: Jauslin, R., & Tillé, Y., (2020). Spatial Spread Sampling Using Weakly Associated Vectors, *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):431–451

space into the real line. A one-dimension systematic sampling is then applied, possibly with unequal probabilities (see also Theobald et al., 2007; Brown et al., 2015; Kincaid et al., 2019). Robertson et al. (2018) have proposed a similar method called Halton iterative partitioning (HIP). It uses structural properties of the Halton sequence to draw a well-spread sample. Dickson and Tillé (2016) have simply used the Traveling Salesman Problem (TSP) in order to map the population points in one dimension. Systematic sampling is then applied. Grafström (2011) has proposed spatially correlated Poisson sampling (SCPS). This method uses weights to create strong negative correlations between the inclusion probabilities of nearby units. Grafström et al. (2012) proposed the Local Pivotal Method (LPM). It is a particular case of the splitting methods proposed by Deville and Tillé (1998). It consists of randomly choosing between two nearby units at each step and produces an automatic repulsion in the selection of the neighbour units. Grafström and Tillé (2013) have generalized the LPM to obtain spread samples that are also balanced on totals of auxiliary variables. All these methods are implemented in the BalancedSampling R package (Grafström and Lisic, 2019).

Stevens Jr. and Olsen (2004) have proposed computing the Voronoï polygons around the sampled units, after which they sum the inclusion probabilities of the population units belonging to each Voronoï polygon. The variance of these sums, called "spatial balance", is an indicator of the quality of spreading. Tillé et al. (2018) have modified the index proposed by Moran (1950) so that it can be interpreted as a coefficient of correlation between the units and their neighbourhood. The index provides another measure of the quality of spreading.

Diggle et al. (2010) defined preferential sampling as a sample selection where the sampling method is not independent of the spatial process, and where unequal inclusion probabilities cannot be explained by auxiliary variables. It is important to emphasize that, in this manuscript, the inclusion probabilities are supposed to be established in advance. The sample selection is a random realisation of the sampling model and is independent of all of the variables.

In this paper, we propose a new spatial sampling method. We start with the vector of inclusion probabilities. Like in the cube method (Deville and Tillé, 2004; Tillé, 2006) inclusion probabilities are randomly modified at each step. It can be seen as random walk that from the vector of inclusion probabilities ends up with a sample. By choosing well, the modification direction at each step the sample selected is very well-spread.

The paper is organized as follows. Section 2.2 gives the notation and a basic setup of the problem as well as the insight that a well-spread sample results in a Horvitz-Thompson estimator with a smaller variance. In Section 2.3, we introduce the new method that we propose and the process of sample selection. In Section 2.4, we describe the indices that allow evaluation of the quality of the spreading: the spatial balance index and the measure based on Moran's I index. In Section 2.5, we present a variance estimator for our method. In Section 2.6, we give simulation results of the algorithm on artificial spatial configurations while Section 2.7 is dedicated to simulations on real data. We used the geo-referenced "Meuse" dataset available in the R package "sp" of Pebesma and Bivand (2005) with inclusion probabilities proportional to the "cadmium" variable. Simulations show that the proposed method surpasses LPM, GRTS and SCPS for the quality of the spreading, and the estimation accuracy.

2.2 Notation

2.2.1 Basic setup

Consider a finite population U of size N whose units can be defined by labels $k \in \{1, 2, \ldots, N\}$. Let $\mathcal{S} = \{s | s \subset U\}$ be the power set of U. These units are geo-referenced in a space that can have more than two dimensions. A sampling design is defined by a probability distribution $p(.)$ on $\mathcal{S}$ such that

$$p(s) \geqslant 0 \text{ for all } s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1.$$

A random sample S is a random vector that maps elements of $\mathcal{S}$ to an N vector of 0 or 1 such that $\mathrm{P}(S = s) = p(s)$. Define $a_k(S)$, for $k = 1, \ldots, N$:

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

Then a sample can be denoted by means of a vector notation: $\boldsymbol{\delta}^\top = (\delta_1, \delta_2, \ldots, \delta_N)$. For each unit of the population, the inclusion probability $0 \leqslant \pi_k \leqslant 1$ is defined as the probability that unit k is selected into sample S:

$$\pi_k = \mathrm{P}(k \in S) = \mathrm{E}(\delta_k) = \sum_{s \in S | k \in s} p(s), \text{ for all } k \in U.$$

Let $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_N)$ be the vector of inclusion probabilities. Then, $\mathrm{E}(\boldsymbol{\delta}) = \boldsymbol{\pi}$. In many applications, inclusion probabilities are such that samples have a fixed size n. Let the set of all samples that have fixed size equal to n be defined by

$$\mathcal{S}_n = \left\{ \boldsymbol{\delta} \in \{0, 1\}^N \;\middle|\; \sum_{k=1}^{N} \delta_k = n \right\}.$$

The sample is generally selected with the aim of estimating some population parameters. Let y_k denote a real number associated with unit $k \in U$, usually called the variable of interest. For example, the total

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

can be estimated by using the classic Horvitz-Thompson estimator of the total defined by

$$\widehat{Y}_{HT} = \sum_{k \in U} \frac{y_k \delta_k}{\pi_k}. \tag{2.1}$$

Usually, some auxiliary information $\mathbf{x}_k^\top = (x_{k1}, x_{k2}, \ldots, x_{kq}) \in \mathbb{R}^q$ regarding the population units is available. In the particular case of spatial sampling, a set of spatial coordinates $\mathbf{z}_k^\top = (z_{k1}, z_{k2}, \ldots, z_{kp}) \in \mathbb{R}^p$ is supposed to be available, where p is the dimension of the considered space. A sampling design is said to be balanced on the auxiliary variables x_k if and only if it satisfies the balancing equations

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

2.2.2 Well-spread sample

A sample is well-spread "if the number of selected units is close to what is expected on average in any part of the space" (Grafström and Lundström, 2013). In this section we provide some insight as to why selection of a well-spread sample minimizes the variance of the Horvitz-Thompson estimator. Let suppose we are in general linear superpopulation model:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad \text{for all } k \in U,$$

where $\mathbf{x}_k$ is a column vector of values taken by q auxiliary variables on unit k, $\boldsymbol{\beta} \in \mathbb{R}^q$ are q regression coefficients and ε_k is a random variable that satisfies $\mathrm{E}_M(\varepsilon_k) = 0$ and $\mathrm{var}_M(\varepsilon_k) = \varphi^2(\mathbf{x}_k) = \varphi_k^2$, with $\varphi^2(\cdot)$ a Lipschitz continuous function. Note that $E_M(\cdot)$ and $\mathrm{var}_M(\cdot)$ are the expectation and the variance under the model. Let also

$$\mathrm{cov}_M(\varepsilon_k, \varepsilon_\ell) = \varphi_k \varphi_\ell \rho_{k\ell}, \quad \text{with } k \neq \ell \in U,$$

where $\rho_{k\ell}$ is a function that decreases when the distance between two units increase. This notation shows that two close units are autocorrelated. Grafström and Tillé (2013) showed that

$$\mathrm{E}_p \mathrm{E}_M (\widehat{Y}_{HT} - Y)^2 = \mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \beta \right]^2 + \sum_{k \in U} \sum_{\ell \in U} \varphi_k \varphi_\ell \rho_{k\ell} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell}, \quad (2.2)$$

where E_p is the expectation of the design and $\pi_{k\ell} = \mathrm{E}_p(a_k a_\ell)$ are the joint inclusion probabilities. From equation (2.2) we can see that the first term of the right-hand side is minimized if the sample is balanced on the auxiliary variables $\mathbf{X}$. The second term is minimized if $\pi_{k\ell}$ is small whenever $\rho_{k\ell}$ is large. Meaning that choosing a well-spread sample (i.e. a sample where the $\pi_{k\ell}$ are small) minimizing the equation (2.2). Grafström and Lundström (2013) showed that if the inclusion probabilities are set up proportional to the φ_k then (2.2) is even more minimized. As result, selection of a well-spread sample jointly used with the Horvitz-Thompson estimator is a very efficient procedure in terms of variance reduction.

2.3 Weakly Associated Vector Sampling

2.3.1 General idea

Our sampling algorithm, Weakly Associated VEctor (WAVE) sampling starts with the inclusion probability vector. At each step, this vector is randomly modified so that at least one of the components of the vector is replaced by a 0 or a 1. So, in at most N steps, a sample is randomly selected. This idea is also used in the cube method proposed by Deville and Tillé (2004) to select balanced samples. The proposed method is different from the cube method by selecting in a completely different way the vector of modification of inclusion probabilities. By carefully choosing the direction of the modification of the working vector, we can ensure that the selection of the sample will be well-spread. This choice is described in Section 2.3.4.

2.3.2 Distance

In order to describe the spatial structure of the population, a distance is defined as a function m defined on the product set $U \times U$ such that

$$d : U \times U \to \mathbb{R}^+, \tag{2.3}$$

and satisfies the property of non-negativity, symmetry, and triangular inequality. More specifically, for all $k, \ell, j \in U$ the following properties hold:

$$\begin{aligned} &d(k,\ell) \geqslant 0, \quad d(k,\ell) = 0 \iff k = \ell, \\ &d(k,\ell) = d(\ell,k), \\ &d(k,j) \leqslant d(k,\ell) + d(\ell,j). \end{aligned}$$

In most of applications, the usual Euclidean squared distance is used. It is defined by,

$$d_E^2(k,\ell) = (\mathbf{z}_k - \mathbf{z}_\ell)^\top (\mathbf{z}_k - \mathbf{z}_\ell), \tag{2.4}$$

where $\mathbf{z}_k$ and $\mathbf{z}_\ell$ are the spatial coordinates of units $k, \ell \in U$. Sometimes it could be interesting to compute the distance on auxiliary variables. In this case, the Mahalanobis distance can be more appropriate,

$$d_M^2(k,\ell) = (\mathbf{x}_k - \mathbf{x}_\ell)^\top \mathbf{S}^{-1} (\mathbf{x}_k - \mathbf{x}_\ell),$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{k \in U} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top, \quad \bar{\mathbf{x}} = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k.$$

When the population is distributed on a $N_1 \times N_2$ regular grid of $\mathbb{R}^2$, a tore distance can be defined. We define a tore distance as the Euclidean metric calculated on a regular tore. An advantage of using this distance is that the surface which we are working on, no longer has corners and edges. With this tore distance, two units on the same column (respectively row) that are on the opposite side have a small distance. More precisely, a unit that is positioned at the right top corner of the grid will be equally distant to the left top corner and the right bottom corner. It is like seeing the grid curved such that it looks like a regular tore. The distance is then defined by:

$$\begin{aligned} d_T^2(k,\ell) \quad = \quad & \min[(z_{k1} - z_{\ell 1})^2, (z_{k1} + N_1 - z_{\ell 1})^2, (z_{k1} - N_1 - z_{\ell 1})^2] \\ & + \min[(z_{k2} - z_{\ell 2})^2, (z_{k2} + N_2 - z_{\ell 2})^2, (z_{k2} - N_2 - z_{\ell 2})^2] \end{aligned} \tag{2.5}$$

Example 2.3.1 *Let* $\{1, \ldots, 9\}$ *be on a regular grid of size* 3×3*, then the squared distance matrices defined by Equations* (2.4) *and* (2.5) *are equal to*

$$\mathbf{D}_E = \begin{pmatrix} 0 & 1 & 4 & 1 & 2 & 5 & 4 & 5 & 8 \\ 1 & 0 & 1 & 2 & 1 & 2 & 5 & 4 & 5 \\ 4 & 1 & 0 & 5 & 2 & 1 & 8 & 5 & 4 \\ 1 & 2 & 5 & 0 & 1 & 4 & 1 & 2 & 5 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 5 & 2 & 1 & 4 & 1 & 0 & 5 & 2 & 1 \\ 4 & 5 & 8 & 1 & 2 & 5 & 0 & 1 & 4 \\ 5 & 4 & 5 & 2 & 1 & 2 & 1 & 0 & 1 \\ 8 & 5 & 4 & 5 & 2 & 1 & 4 & 1 & 0 \end{pmatrix}, \quad \mathbf{D}_T = \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 & 1 & 2 & 2 & 1 & 2 \\ 1 & 1 & 0 & 2 & 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 & 1 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 2 & 2 & 1 & 1 & 1 & 0 & 2 & 2 & 1 \\ 1 & 2 & 2 & 1 & 2 & 2 & 0 & 1 & 1 \\ 2 & 1 & 2 & 2 & 1 & 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 2 & 2 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{2.6}$$

In the spatial configuration of a regular grid, some distances between points are equal. The rank of the nearest neighbours is then assigned and duplicated values appear. In order to obtain a different rank distance for each unit, a small random quantity is added to the coordinates so that it disturbs the given units and the distances are a bit different from each other. Let $\boldsymbol{\varepsilon} \in \mathbb{R}^2$ and $\tilde{\mathbf{z}}_k = \mathbf{z}_k + \boldsymbol{\varepsilon}$ the shifted coordinates, Equation (2.5) is then replaced by,

$$\begin{aligned} d_S^2(k,\ell) &= \min[(\tilde{z}_{k1} - z_{\ell 1})^2, (\tilde{z}_{k1} + N_1 - z_{\ell 1})^2, (\tilde{z}_{k1} - N_1 - z_{\ell 1})^2] \\ &\quad + \min[(\tilde{z}_{k2} - z_{\ell 2})^2, (\tilde{z}_{k2} + N_2 - z_{\ell 2})^2, (\tilde{z}_{k2} - N_2 - z_{\ell 2})^2]. \end{aligned}$$

$\boldsymbol{\varepsilon}$ is called a "shift" and m_S the shifted version of m_T, for example if $\boldsymbol{\varepsilon} = (1/12, 1/4)$, the distance matrix $\mathbf{D}_S$ becomes,

$$\mathbf{D}_S = \begin{pmatrix} 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 \\ 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 \\ 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 \\ 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 & 0.57 & 1.40 & 1.74 \\ 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 & 1.74 & 0.57 & 1.40 \\ 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 & 1.40 & 1.74 & 0.57 \\ 0.57 & 1.40 & 1.74 & 1.57 & 2.40 & 2.74 & 0 & 0.90 & 1.24 \\ 1.74 & 0.57 & 1.40 & 2.74 & 1.57 & 2.40 & 1.24 & 0 & 0.90 \\ 1.40 & 1.74 & 0.57 & 2.40 & 2.74 & 1.57 & 0.90 & 1.24 & 0 \end{pmatrix}. \tag{2.7}$$

The matrix is no longer a distance matrix since the symmetric axiom has been dropped. A distance that has an unsatisfied symmetry axiom is called a quasi-metric. Nevertheless, if an epsilon value is added instead of $(1/12, 1/4)$, then the values are almost the same and the order is preserved in each row. In Figure 2.1, three simple configurations are presented: Euclidean, tore and shifted tore distance on a 3×3 regular grid. In the shifted distance graph, all the distances from point $(1, 1)$ to the other grid points are different.

2.3.3 The stratification matrix

Let $k \in U$ a unit in the population. The idea is to construct a strata G_k under some distance metric such that the elements in G_k are ranked in increasing order. Define G_k the set of the nearest neighbours of unit k, including k, such that their inclusion probabilities are greater or equal than one by only one unit. Denote g_k the number of elements inside G_k, the spatial weights are then defined as follows

$$w_{k\ell} = \begin{cases} \pi_\ell & \text{if unit } \ell \text{ is in the set of the } g_k - 1 \text{ nearest neighbour of } k, \\ \pi_\ell + 1 - \sum_{j \in G_k} \pi_k & \text{if unit } \ell \text{ is the } g_k\text{th nearest neighbour of } k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

$\mathbf{W}$ denote an $N \times N$ stratification matrix and each row of matrix $\mathbf{W}$ represents a stratum. Each stratum is defined by a particular unit and its neighbouring units. Nearest neighbours are defined with a metric function (2.3). If the metric is such that ties values exist, then we can divide the quantity w_{kl} into the different g_k nearest neighbours of the unit k that have the same distance. Or a shifted metric can be used (exemplified in matrix (2.7)) such that all the distances are different. Each row of matrix

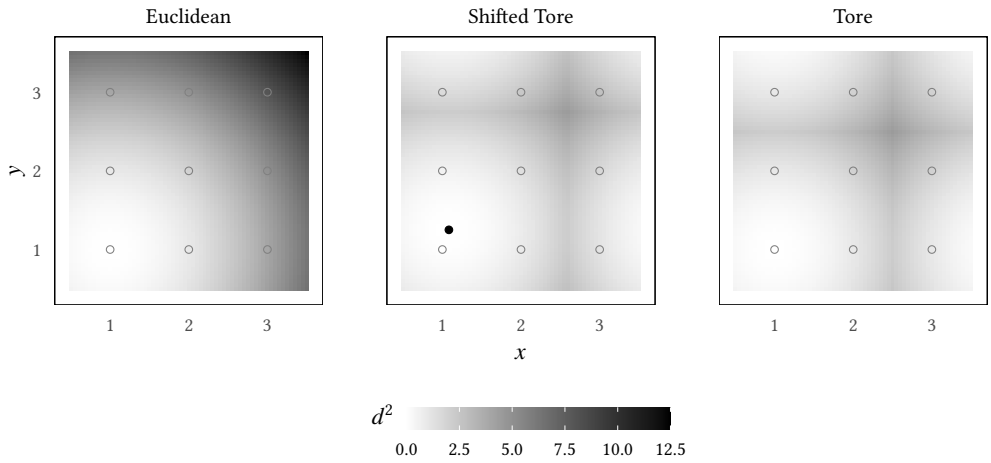


Figure 2.1 – *Simple example of a* 3×3 *regular grid set up on three different distances with a gradient calculated from the points* $(1, 1)$*. The left one is the classical Euclidean distance* (2.4)*, the right one is the tore distance given in* (2.5) *and the central graph is the shifted tore distance with a shift equal to* $(1/12, 1/4)$ *(the black point on the graph). It illustrates the two different patterns and the values of the grid points corresponding to the entries of the first row of the three previous matrices* (2.6)*.*

$\mathbf{W}$ sums to 1. Thus matrix $\mathbf{W}$ is a right stochastic matrix. Most of the components of matrix $\mathbf{W}$ are null. $\mathbf{W}$ can thus be encoded as a sparse matrix.

Example 2.3.2 *Let* $U = \{1, 2, 3, 4, 5\}$ *a population of 5 units. Suppose that the inclusion probabilities are equal to* $\boldsymbol{\pi} = (1/2, 1/3, 1/4, 1/5, 1/6)$ *and that the order in terms of distance metric from the unit 1 is exactly equal to* $1, 2, 3, 4, 5$*. Meaning that the 5th unit is the farthest from the first. Then* $G_k = \{1, 2, 3\}$ *because* $1/2 + 1/3 + 1/4 \cong 1.084 > 1$ *and* $w_{13} = 1/4 + 1 - (1/2 + 1/3 + 1/4) = 1/6$*.*

Example 2.3.3 *Let* $\{1, \ldots, 9\}$ *be on a regular grid of size* 3×3 *with inclusion probabilities equal* $\pi_k = 1/3$*, for all* $k \in U$*. Figure 2.2 shows different stratification matrices corresponding to* $\mathbf{D}_E$*,* $\mathbf{D}_T$ *and* $\mathbf{D}_S$ *with a shift randomly generated from a random variable* $\mathcal{N}(0, 1/100\mathbf{I})$ *where* $\mathbf{I}$ *is the identity matrix.*

Let $\mathbf{A}$ be defined by

$$\mathbf{A} = \begin{pmatrix} w_{11}/\pi_1 & w_{12}/\pi_2 & \cdots & w_{1N}/\pi_N \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1}/\pi_1 & w_{N2}/\pi_2 & \cdots & w_{NN}/\pi_N \end{pmatrix}. \tag{2.9}$$

Matrices $\mathbf{W}$ and $\mathbf{A}$ are square but not necessarily full rank. The sum of the rows of $\mathbf{A}$ is equal or approximately equal to the number of elements in each stratum. The strata are represented by the rows and the contribution of a unit i in each stratum is represented by the ith column. Figure 2.3 shows the sparsity pattern of the two stratification matrices.

Example 2.3.4 *Let* U *be a population of size* $N = 250$ *and inclusion probabilities equal to* $\pi_k = 1/25$*, for all* $k \in U$*. Suppose that spatial coordinates are generated independently from a uniform*

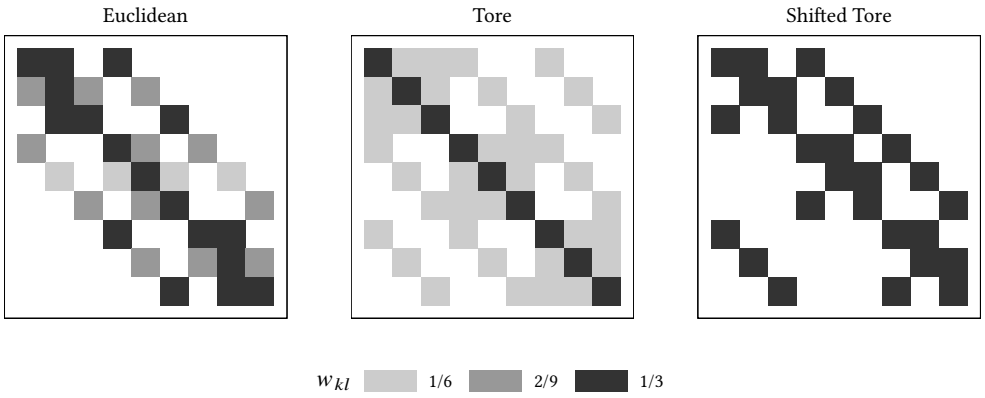


Figure 2.2 – *Sparsity pattern of three stratification matrices. Spatial coordinates are* 3×3 *regular grid and the inclusion probabilities are equal to* $\pi = (1/3, \ldots, 1/3)$*. Depending on the way of defining the nearest neighbours in Equation* (2.8)*, different weight values are obtained. The left stratification matrix uses the classical Euclidean distance* (2.4)*, the central one the tore distance* (2.5) *and the right one uses a shifted tore distance with a shift randomly generated from a random variable* $\mathcal{N}(0, 1/100\mathbf{I})$*.*

distribution on the square unit, so that with probability one there are no tied distance values. Since all $1/\pi_k = 25$ *the non-zero entries of* $\mathbf{A}$ *are all equal to 1. Based on the definition* (2.8)*, the weights are all equal to the inclusion probabilities or zero. Figure 2.3 shows the sparsity pattern of the stratification matrices and exemplifies some initial strata.*

2.3.4 Implementation

The method is described in detail in Algorithm 1. The main idea is derived from the cube method (Deville and Tillé, 2004). At each step, vector $\boldsymbol{\pi}$ is randomly modified. To modify $\boldsymbol{\pi}$, we choose a vector that spreads at best. Ideally, the aim consists of obtaining a sample $\boldsymbol{\delta}$ such that the following equality is satisfied:

$$\mathbf{A}\boldsymbol{\delta} = \mathbf{A}\boldsymbol{\pi} = \mathbf{1}.$$

This linear system define an affine subspace of $\mathbb{R}^N$:

$$\mathcal{A} = \{\boldsymbol{\delta} \in \mathbb{R}^N \mid \mathbf{A}\boldsymbol{\delta} = \mathbf{A}\boldsymbol{\pi}\}$$

which could also be rewritten:

$$\mathcal{A} = \boldsymbol{\pi} + \text{Null}(\mathbf{A})$$

where

$$\text{Null}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^N \mid \mathbf{A}\mathbf{v} = \mathbf{0}\}.$$

Depending on whether matrix $\mathbf{A}$ is full rank or not, the vector giving the direction is not selected in the same way. If matrix $\mathbf{A}$ is not full rank, a vector that is contained in

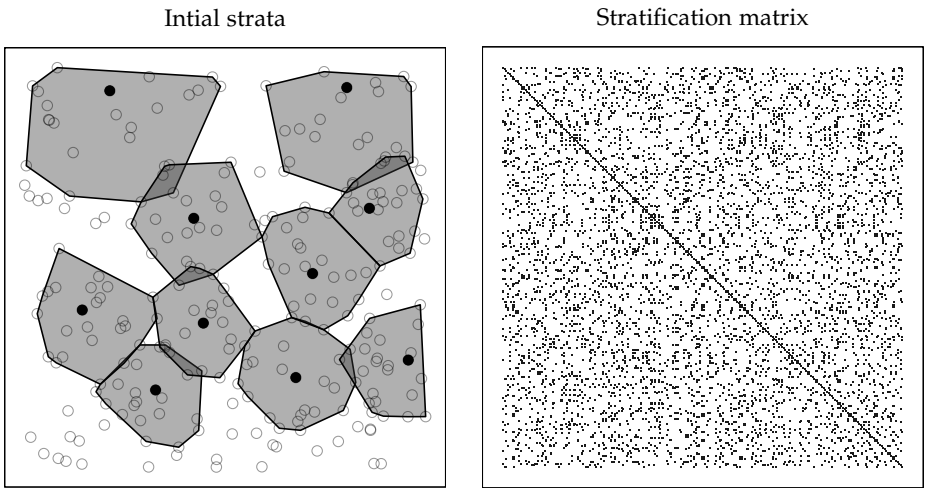


Figure 2.3 – *Representation of the strata defined by the spatial weights Equation* (2.8). *Spatial coordinates of the units are generated randomly from a uniform distribution on the square unit* $[0,1] \times [0,1]$. *The overall population size is equal to* $N = 250$ *and the inclusion probabilities are identical and equal to* $\pi_k = 1/25 = 0.04$. *Meaning that the sample size is equal to* $n = 10$. *With these parameters the expected number of units in each stratum is equal to* $1/0.04 = 25$. *The left graph shows the population and the selected units with its initial strata. On the right, it shows the sparsity pattern of the matrix* (2.9). *All entries of the matrix are equal to 1.*

the right null space is selected. If matrix $\mathbf{A}$ is full rank, we compute $\mathbf{v}$,$\mathbf{u}$ a left and a right singular vectors associated to the smallest singular value σ of $\mathbf{A}$ i.e,

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u}, \quad \mathbf{A}^\top\mathbf{u} = \sigma\mathbf{v}.$$

By choosing the modification vector $\mathbf{v}$, we ensure that we select the vector which remains closest to the set $\mathcal{A}$. Vector $\mathbf{v}$ is called the weakest associated vector to the matrix $\mathbf{A}$. Vector $\mathbf{v}$ is then centred to ensure the fixed sample size. By using these weakest associated vectors, the initial spatial configurations are the least modified. At each step, some inclusion probabilities $\boldsymbol{\pi}$ are modified and at least one component is set to 0 or 1. Matrix $\mathbf{A}$ is updated from the new inclusion probabilities. This step is repeated until there is only one component that is not equal to 0 or 1.

Algorithm 1 is implemented in a `R` package, which uses the Armadillo C++ library into the `R` interface (Eddelbuettel and Sanderson, 2014). The implementation uses the sparse matrix class. In fact, depending on the inclusion probabilities, matrix $\mathbf{A}$ given in (2.9) could be strongly sparse. Even if the function benefits from the C++ implementation, it could be quite time consuming as the size of the population N increases. Nevertheless, we will see in the next section that the algorithm performs better in terms of two spreading measures than those currently used for the spatial balanced sampling design.

2.4 Spatial balance

2.4.1 Voronoï polygons

Stevens Jr. and Olsen (2004) suggested the spatial balance of a sample consists of using

Algorithm 1: Algorithm for WAVE sampling

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

1. From $\boldsymbol{\pi}_t$, extract $\widetilde{\boldsymbol{\pi}}_t$ vector $\boldsymbol{\pi}_t$ restricted to the k such that $0 < \pi_k^{(t)} < 1$. Let J be the length of $\widetilde{\boldsymbol{\pi}}_t$.
2. Compute the $J \times J$ matrix $\mathbf{A}_t$ of Equation (2.9) using inclusion probabilities $\widetilde{\boldsymbol{\pi}}_t$.
3. Calculate the rank r of matrix $\mathbf{A}_t$.
   (a) If matrix $\mathbf{A}_t$ does not have full rank, choose $\mathbf{v}_t = (v_1^{(t)}, \ldots, v_J^{(t)}) \in \mathbb{R}^J$ a vector in the right null space of $\mathbf{A}_t$.
   (b) If matrix $\mathbf{A}_t$ has full rank, compute the singular value decomposition and seek for $\mathbf{v}_t$ a right singular vector associated to the smallest singular value σ_t.
4. Next in order to ensure the fixed sample size, vector $\mathbf{v}_t$ is centred:
$$\widetilde{\mathbf{v}}_t = \mathbf{v}_t - \frac{1}{J} \sum_{i \in J} v_i^{(t)} \mathbf{1}_J,$$
   where $\mathbf{1}_J$ is the $J \times 1$ vector of one.
5. Find λ_1 and λ_2 the largest positive real numbers such that all the $0 \leqslant \widetilde{\pi}_k^{(t)} + \lambda_1 \widetilde{v}_k^{(t)} \leqslant 1$ and $0 \leqslant \widetilde{\pi}_k^{(t)} - \lambda_2 \widetilde{v}_k^{(t)} \leqslant 1$, $k = 1, \ldots, J$.
6. Compute
$$\boldsymbol{\pi}_{t+1} = \begin{cases} \widetilde{\boldsymbol{\pi}}_t + \lambda_1 \widetilde{\mathbf{v}}_t & \text{with probability } \lambda_2/(\lambda_1 + \lambda_2) \\ \widetilde{\boldsymbol{\pi}}_t - \lambda_2 \widetilde{\mathbf{v}}_t & \text{with probability } \lambda_1/(\lambda_1 + \lambda_2). \end{cases}$$
7. Return at 1. with $\boldsymbol{\pi}_{t+1}$ until no units k remains such that $0 < \pi_k^{(t+1)} < 1$.

the Voronoï polygons. The Voronoï polygon associated to the sample unit k is the set of all units of the population that are closer to k than to any other sample units. Let b_k be the sum of inclusion probabilities of the units belonging to the Voronoï polygon associated with the sample unit k. If the sample is perfectly spread, b_k should be equal to 1 for each k. Indeed, n units are selected in the sample, then

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

and so

$$\frac{1}{n} \sum_{k \in S} b_k = 1.$$

The variance of the $\mathrm{E}[v_k]$ could be approximated and give a good measure of the spatial balance of the sample. The spatial balance measure based on the Voronoï polygons is

defined by

$$B(S) = \frac{1}{n} \sum_{k \in S} (b_k - 1)^2. \tag{2.10}$$

Two samples are compared in Fig. 2.4. The left one is selected with a simple random sampling without replacement and the right one is selected with WAVE sampling. The darker the Voronoï polygon, the more units it contains. An perfectly well-spread sample should have all polygons of the same colour.

The measure B has some limitations. It does not vary from a fixed finite range. This does not allow a clear understanding if the sample is balanced or clustered (Tillé et al., 2018). Moreover, the measure sometimes behaves incorrectly and suggests a well-spread sample although it is not the case. Examples are given in the Supplementary Material Section. For these reasons, we suggest using another measure based on Moran's I index.

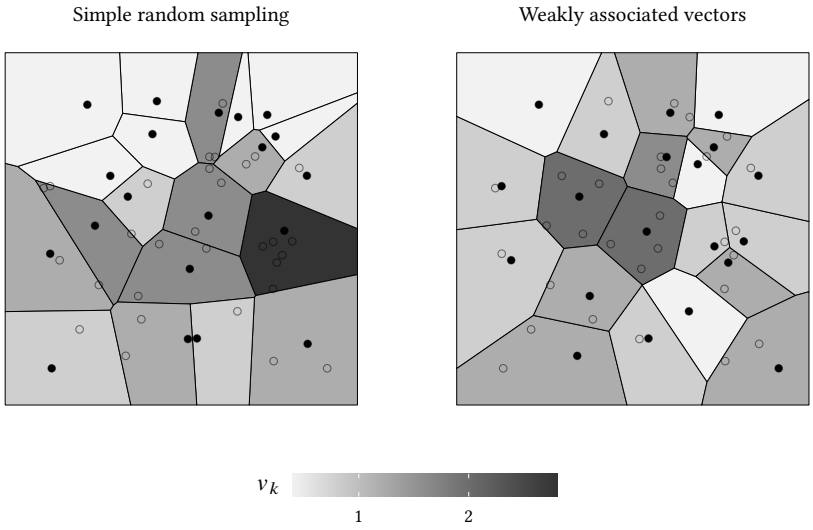


Figure 2.4 – *Illustrated example of how the spatial balance measure based on the Voronoï is performed. The population and sample sizes are respectively equal to* $N = 50$ *and* $n = 20$*, the inclusion probabilities are identical and equal to* $\pi_k = 0.4$*. The spatial coordinates are generated from two random uniform* $\mathcal{U}(0, 1)$*. Two sampling design are compared. The left one is the simple random sampling without replacement and the right one is the weakly associated vector sampling.*

2.4.2 Moran's I index

A second approach for measuring the spatial balance of a sampling design has been proposed by Tillé et al. (2018). Consider a $N \times N$ spatial weights matrix,

$$\mathbf{W} = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1N} \\ w_{21} & 0 & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & 0 \end{pmatrix}.$$

A large value of $w_{k\ell}$ indicates that ℓ is a neighbour of k. Matrix $\mathbf{W}$ is not necessarily

symmetric. The index proposed by Tillé et al. (2018) is defined by

$$I_B(\boldsymbol{\delta}) = \frac{(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{W} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)}{\sqrt{(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{Q} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{G} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)}}, \tag{2.11}$$

where $\boldsymbol{\delta}$ is the sample and

$$\bar{\boldsymbol{\delta}}_w = \frac{\boldsymbol{\delta}^\top \mathbf{W} \mathbf{1}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

$\mathbf{Q}$ is the diagonal matrix containing $w_{k.} = \sum_{\ell \in U} w_{k\ell}$ on its diagonal,

$$\mathbf{G} = \mathbf{C}^\top \mathbf{Q} \mathbf{C}, \quad \mathbf{C} = \mathbf{Q}^{-1} \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top \mathbf{W}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

and $\mathbf{1}$ is a column vector of N ones. Tillé et al. (2018) pointed out that I_B can be interpreted as weighted correlation between δ_k and the average of the δ_ℓ that are in the neighbourhood of k. We have that $-1 \leqslant I_B \leqslant 1$ and $I_B = -1$ when the sample is well-spread. Tillé et al. (2018) have proposed to use the inverse of the inclusion probability $h_k = 1/\pi_k$ to define the neighbours of the unit k. Specifically, if the unit k is selected it seems natural to consider $h_k - 1$ neighbours in the population. Let $\lfloor h_k \rfloor$ and $\lceil h_k \rceil$ be respectively the inferior and superior integers of h_k. Spatial weights are then defined as follows,

$$w_{k\ell} = \begin{cases} 1 & \text{if unit } \ell \text{ is in the set of the } \lfloor h_k \rfloor \text{ nearest neighbour of } k \\ h_k - \lfloor h_k \rfloor & \text{if unit } \ell \text{ is the } \lceil h_k \rceil \text{th nearest neighbour of } k \\ 0 & \text{otherwise .} \end{cases} \tag{2.12}$$

For example, if a unit k has an inclusion probability of $\pi_k = 0.35$ then $h_k \cong 2.857$. Meaning that the first nearest neighbour of k has a weight equal to 1 and the second has a weight of 0.857. In case there are units that are at equal distance from each other, Tillé et al. (2018) suggests to divide the spatial weights equally among them.

We propose a new way of defining the spatial weights. It consists of using spatial weights defined in (2.8) rather than the weights (2.12). We set $w_{kk} = 0$ for all $k \in U$. For the rest of the paper, I_{B1} will represent the measure based on the spatial weights (2.12) and I_B the one based on (2.8).

2.5 Variance estimation

If the sampling design is of fixed size, the variance of the Horvitz-Thompson estimator of the total (2.1) is defined by

$$\operatorname{var}(\widehat{Y}_{HT}) = -\frac{1}{2} \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell},$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ and $\pi_{k\ell} = \mathrm{E}(\delta_k \delta_l)$ is the joint inclusion probabilities. For complex sampling designs, quantities $\pi_{k\ell}$ are generally impossible to compute.

Many different estimators have been developed. Sen (1953) and Yates and Grundy (1953) proposed one classical estimator:

$$\widehat{\operatorname{var}}_{SYG}(\widehat{Y}_{HT}) = -\frac{1}{2} \sum_{k \in S} \sum_{\ell \in S} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

This estimator can take negative values, but it is non-negative when $\Delta_{k\ell} \leqslant 0$ for all $k \neq \ell \in U$. A common problem with spatially balanced sampling designs is that many joint inclusion probabilities are equal to zero. Indeed the probability of selecting two close units is generally zero or very close to zero. In this case, $\widehat{\text{var}}_{SYG}$ is not an unbiased estimator of $\text{var}(\widehat{Y}_{HT})$.

Tillé (2020, Chapter 5) gives a general estimator based on the variance estimator of the conditional Poisson sampling. It is equal to

$$\widehat{\text{var}}_{app}(\widehat{Y}_{HT}) = \sum_{k \in S} \frac{c_k}{\pi_k^2}(y_k - \hat{y}_k^\star)^2,$$

where

$$\hat{y}_k^\star = \pi_k \frac{\sum_{\ell \in S} c_\ell y_\ell / \pi_\ell}{\sum_{\ell \in S} c_\ell}.$$

Choosing $c_\ell = (1 - \pi_k)n/n - 1$ we obtain the Hájek-Rosén estimator Hájek (1981) defined by

$$\widehat{\text{var}}_{HAJ}(\widehat{Y}_{HT}) = \frac{n}{n-1} \sum_{k \in S} (1 - \pi_k) \left\{ \frac{y_k}{\pi_k} - \frac{\sum_{\ell \in S} y_\ell (1 - \pi_\ell)/\pi_\ell}{\sum_{\ell \in S} (1 - \pi_\ell)} \right\}^2. \tag{2.13}$$

This variance estimator is simple to compute and has the advantage of using only the first-order inclusion probabilities. It is a good estimator for maximum entropy sampling design and simple random sampling without replacement. Grafström et al. (2012) pointed out that the estimator seems to overestimate the variance for spread sampling design. Grafström and Schelin (2014) proposed an estimator based on the nearest neighbour in the sample. It is called variance estimator for spatially balanced samples and is defined as follows:

$$\widehat{\text{var}}_{SB}(\widehat{Y}_{HT}) = \frac{1}{2} \sum_{k \in S} \left(\frac{y_k}{\pi_k} - \frac{y_{\ell_k}}{\pi_{\ell_k}} \right)^2, \tag{2.14}$$

where ℓ_k is the nearest neighbour to the unit k in the sample. Stevens Jr. and Olsen (2003) proposed an estimator based on a local neighbourhood for each unit in the sample. It is called the local mean variance estimator and is given by

$$\widehat{\text{var}}_{LM}(\widehat{Y}_{HT}) = \sum_{k \in U} \sum_{\ell \in D_k} w_{k\ell} \left(\frac{y_k}{\pi_k} - \sum_{m \in D_k} w_{km} \frac{y_m}{\pi_m} \right)^2, \tag{2.15}$$

where the weights $w_{k\ell}$ are computed such that they vary inversely as π_ℓ and decrease as the distance between unit k and ℓ increases. Moreover, it satisfies the constraint $\sum_{k \in S} w_{k\ell} = \sum_{\ell \in S} w_{k\ell} = 1$. The set D_k is the neighbourhood of the unit k and is defined by the unit itself and the three neighbourhoods of the three nearest neighbours. Meaning that D_k contains at least four units and at most thirteen. This variance estimator is implemented by function `localmean.var` in the R package "spsurvey" Kincaid et al. (2019). It produces a good estimator for the GRTS method. For the rest of the manuscript, we will adopt the following notation: $\widehat{\text{var}}_{LM_j}(\widehat{Y}_{HT})$ where j is the number of neighbours used in the calculations. In Section 2.7, we compare the previous estimators for different sampling designs.

2.6 Simulations on artificial spatial configurations

In this section, we propose three artificial spatial configurations to study the performance of the WAVE sampling in terms of spreading measure. To generate the three population datasets, the expected size of the population is equal to $N = 144$.

1. The dataset is generated from the Complete Spatial Randomness (CSR) that is a Poisson process with intensity equal to N, meaning that the expected number of points in the unit square is equal to N.

2. A Neyman-Scott cluster process (Neyman and Scott, 1958) is generated with 12 circular discs of radius 0.055 with units uniformly distributed around the centre. Each cluster contains 12 units such that the population target size is equal to N.

3. Simple regular grid of size 12×12.

Figure 2.5 shows a sample selection by the WAVE sampling design on the three different datasets. For the three configurations, the sample size is equal to $n = 3$ and the inclusion probabilities are all equal to $\pi_k = n/N$ for all $k \in U$. When units are regularly dispersed in the space and when the inverse of inclusion probabilities is equal to an integer that is a divisor of the population size N, the selected sample can be systematic, which is the optimal solution.

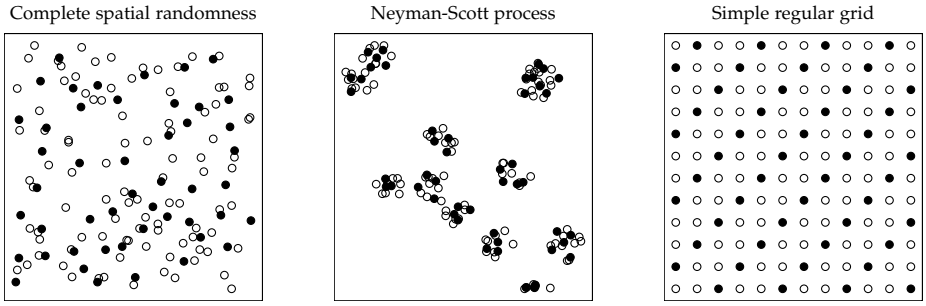


Figure 2.5 – *Example of a sample selection by the WAVE sampling on the three different spatial configurations, Complete Spatial Randomness, Neyman-Scott and regular grid. For each of them, the inclusion probabilities are equal to $\pi_k = n/N$ for all $k \in U$.*

For each population, 10,000 samples of size n respectively equal to 25, 50 and 100 are selected. Two cases are considered for the inclusion probabilities. In the first case, all inclusion probabilities are equal

$$\text{for all } k \in U, \pi_k = \pi = \frac{n}{N}.$$

For the second case, the inclusion probabilities are unequal and sum up to n,

$$\text{for all } k, \ell \in U, k \neq \ell \text{ we have } \pi_k \neq \pi_\ell \text{ and } \sum_{k \in U} \pi_k = n.$$

In each case we calculate the spatial balance based on the Voronoï polygons (2.10) and measures based on Moran's I index (2.11). The simulation results of the CRS dataset are given in the Table 2.1. For the measures based on Moran's I index, the WAVE sampling design performs better than the other algorithms. Moreover, for the

Table 2.1 – *Spreading measures results based on 10000 simulations on the Complete spatial randomness dataset. The population size is equal to 144.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
I_{B_1}											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.028	-0.009	-0.012	0.027	0.093
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.125	-0.095	-0.085	-0.059	0.016
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.436	-0.344	-0.318	-0.229	-0.020
I_B											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.459	-0.316	-0.331	-0.201	-0.028
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.548	-0.393	-0.373	-0.261	-0.013
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.621	-0.469	-0.424	-0.292	-0.029
B											
$n = 16$	0.115	0.117	0.108	0.164	0.135	0.338	0.123	0.124	0.118	0.177	0.345
$n = 32$	0.137	0.128	0.130	0.167	0.165	0.345	0.140	0.146	0.138	0.180	0.352
$n = 48$	0.158	0.137	0.149	0.177	0.195	0.337	0.165	0.151	0.158	0.189	0.319

classical measure based on the Voronoï polygons, the WAVE sampling design performs as well as and sometimes better than the local pivotal method. This can be explained by the fact that the spatial balance measure based on the Voronoï polygons is less sensitive to detect a well-spread sample and sometimes suggests a well-spread sample, although it is not the case (See Supplementary Material Section). For the equal probabilities designs the measures I_{B_1} and I_B coincide. Indeed, the strata based on the inverse inclusion probabilities are the same as the ones considered such that the inclusion probabilities sum to 1. For unequal sampling designs, the differences are less marked with the measure based on the inverse inclusion probabilities (2.12). This result comes from the heterogeneity of the strata and the randomness of the algorithm. If the inclusion probabilities of a unit are nearly zero, then the size of the strata will be very large. This effect can increase the spatial balance measure. Similar results for the two remaining datasets can be seen in the Supplementary Material Section. This analysis shows that the measure I_B should be preferred to I_{B_1}.

2.7 Application to the Meuse dataset

This section investigates the application of WAVE sampling on the dataset "Meuse" available in the R package "sp" of Pebesma and Bivand (2005). It is described as follows: "This data set gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations are from composite samples of an area of approximately 15 m x 15 m."

In order to see how the WAVE sampling performs in terms of spread measures, $m = 10,000$ samples of size respectively equal to 15, 30 and 50 are selected. As in the previous simulation with an artificial population, two cases are considered, equal and unequal probabilities. In the latter case, inclusion probabilities are set proportional to concentration of copper. Locations with high concentrations of copper were therefore more likely to be selected into the sample. Let Y be the total cadmium concentration over the whole population. To show that the variance of the estimated total with the WAVE sampling design is with the other method, we calculate the approximated

variance with the following quantity:

$$v_{SIM}(\widehat{Y}_{HT}) = \frac{1}{m}\sum_{s}\left\{\widehat{Y}_{HT}(s) - Y\right\}^2. \tag{2.16}$$

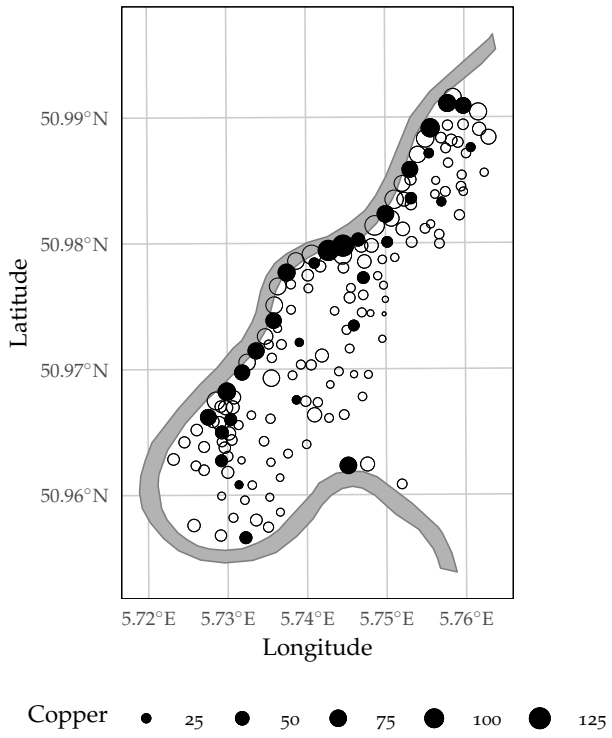


Figure 2.6 – *Example of WAVE sampling on the Meuse dataset. The overall population size is equal to 155. The inclusion probabilities are proportional to copper level variable and the sample size is equal to 30. Plotted sizes of the units are proportional to the copper concentration. The Meuse River is filled in light blue.*

Figure 2.6 shows a sample selected with the WAVE sampling. The filled black circles are selected units while the hollow circles are those that are not selected in the sample. We observe that the dataset is partially aggregated around the river showing a strong spatial correlation.

Results of the three spatial balanced measures on 10'000 simulated samples are given in Table 2.2. WAVE sampling performs better than other sampling designs in terms of I_B and I_{B_1}. In terms of spatial balance measure B, the algorithms are comparable to the artificial simulation, and the differences are less marked.

Results of the simulations on the variance estimator in Table 2.4 shows that the WAVE sampling strategy has a lower variance than the method currently used. This suggests that our method is more efficient in cases where there is a clear spatial correlation. A design-unbiased variance estimator does not exist for the Horvitz-Thompson estimator, but the spatially balanced estimator (2.14) seems to produce a good variance estimator for this dataset. Although the latter slightly overestimates the variance none of the other estimators seem to offer a better alternative. As there is no unbiased estimator we favour a slight overestimation of the variance. Table 2.3 shows the coverage rate as well as the ratio v_{SB}/v_{SIM} for all sampling methods.

Table 2.2 – *Spreading measures results based on 10000 simulations on the Meuse dataset. The population size is equal to 155.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
I_{B_1}											
$n = 15$	-0.518	-0.338	-0.351	-0.226	-0.230	-0.030	-0.340	-0.250	-0.246	-0.165	-0.003
$n = 30$	-0.664	-0.427	-0.427	-0.266	-0.259	-0.019	-0.407	-0.298	-0.288	-0.172	0.024
$n = 50$	-0.796	-0.519	-0.473	-0.302	-0.248	-0.011	-0.466	-0.326	-0.285	-0.204	0.038
I_B											
$n = 15$	-0.518	-0.338	-0.351	-0.226	-0.230	-0.030	-0.354	-0.244	-0.247	-0.153	0.009
$n = 30$	-0.664	-0.427	-0.427	-0.266	-0.259	-0.019	-0.427	-0.290	-0.283	-0.154	0.048
$n = 50$	-0.796	-0.519	-0.473	-0.302	-0.248	-0.011	-0.455	-0.305	-0.263	-0.181	0.060
B											
$n = 15$	0.119	0.125	0.118	0.170	0.160	0.379	0.115	0.121	0.120	0.170	0.387
$n = 30$	0.118	0.123	0.126	0.164	0.159	0.359	0.120	0.121	0.120	0.162	0.345
$n = 50$	0.139	0.132	0.143	0.174	0.194	0.329	0.138	0.133	0.141	0.160	0.281

Based on these simulation results, we are confident that we propose here a new method that allows selection of a sample with a really strong degree of spreading. It performs better than the other sampling method. It can be generalized to higher dimensions and respects the unequal inclusion probabilities.

Table 2.3 – *Results of 10000 simulations on Meuse dataset. The population size is equal to 155.* v_{SIM} *is equal to the variance approximated by the simulations* (2.16). $\widehat{var}$ *depends on the sampling design. For the srswor and maxent methods, we used the estimator* $\widehat{var}_{HAJ}$ (2.13) *while for the other sampling designs, we use* $\widehat{var}_{SB}$ (2.14). *Coverage rate of the 95% confidence intervals are computed as well as the ratio between averages of* $\widehat{var}$ *and* v_{SIM}.

	Sampling design										
	Equal probabilities						Unequal probabilities				
	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
v_{SIM}											
$n = 15$	1.232	1.387	1.309	1.517	1.315	1.774	0.250	0.287	0.260	0.330	0.361
$n = 30$	0.533	0.525	0.538	0.586	0.463	0.805	0.116	0.109	0.096	0.115	0.150
$n = 50$	0.250	0.250	0.222	0.284	0.200	0.413	0.052	0.049	0.039	0.049	0.065
$\widehat{\text{var}}$											
$n = 15$	1.847	1.670	1.635	1.596	1.701	1.784	0.393	0.362	0.371	0.333	0.365
$n = 30$	0.692	0.687	0.670	0.657	0.639	0.808	0.154	0.153	0.152	0.150	0.153
$n = 50$	0.380	0.375	0.385	0.353	0.337	0.403	0.081	0.078	0.080	0.080	0.066
Coverage of the 95% confidence interval											
$n = 15$	0.925	0.907	0.914	0.887	0.918	0.890	0.973	0.958	0.972	0.929	0.933
$n = 30$	0.953	0.943	0.942	0.929	0.963	0.924	0.971	0.972	0.983	0.966	0.942
$n = 50$	0.975	0.966	0.977	0.946	0.973	0.927	0.978	0.979	0.990	0.979	0.944
Ratio $\widehat{\text{var}}/v_{SIM}$											
$n = 15$	1.499	1.204	1.249	1.052	1.294	1.006	1.573	1.264	1.428	1.011	1.011
$n = 30$	1.298	1.307	1.246	1.121	1.380	1.003	1.323	1.400	1.588	1.308	1.016
$n = 50$	1.521	1.501	1.739	1.242	1.685	0.976	1.564	1.615	2.030	1.616	1.003

Table 2.4 – *Results of 10000 simulations on Meuse dataset. The population size is equal to 155.* v_{SIM} (2.16) *is equal to the variance approximated by the simulations.* $\widehat{var}_{SB}$ (2.14) *is the variance estimator based on the nearest neighbours in the sample.* $\widehat{var}_{LM_j}$ *is equal to the estimator* (2.15) *where the number of neighbouring units used is set to* $j = 2, 3, 4$. v_{HAJ} (2.13) *is the Hajek-Rosen estimator.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
$n = 15$											
v_{SIM}	1.232	1.387	1.309	1.517	1.315	1.774	0.250	0.287	0.260	0.330	0.361
$\widehat{\text{var}}_{SB}$	1.847	1.670	1.635	1.596	1.701	1.455	0.393	0.362	0.371	0.333	0.321
$\widehat{\text{var}}_{LM2}$	0.962	0.889	0.889	0.855	0.930	0.786	0.224	0.206	0.209	0.194	0.183
$\widehat{\text{var}}_{LM3}$	1.301	1.256	1.261	1.230	1.308	1.147	0.293	0.279	0.282	0.269	0.259
$\widehat{\text{var}}_{LM4}$	1.463	1.445	1.452	1.430	1.487	1.352	0.325	0.315	0.319	0.306	0.299
$\widehat{\text{var}}_{HAJ}$	1.808	1.824	1.829	1.826	1.854	1.784	0.375	0.370	0.373	0.369	0.365
$n = 30$											
v_{SIM}	0.533	0.525	0.538	0.586	0.463	0.805	0.116	0.109	0.096	0.115	0.150
$\widehat{\text{var}}_{SB}$	0.692	0.687	0.670	0.657	0.639	0.634	0.154	0.153	0.152	0.150	0.143
$\widehat{\text{var}}_{LM2}$	0.382	0.373	0.370	0.362	0.356	0.348	0.094	0.090	0.090	0.089	0.082
$\widehat{\text{var}}_{LM3}$	0.555	0.543	0.543	0.534	0.534	0.512	0.130	0.127	0.127	0.126	0.118
$\widehat{\text{var}}_{LM4}$	0.654	0.649	0.649	0.641	0.652	0.616	0.150	0.148	0.148	0.147	0.140
$\widehat{\text{var}}_{HAJ}$	0.808	0.805	0.806	0.808	0.814	0.808	0.153	0.154	0.155	0.154	0.153
$n = 50$											
v_{SIM}	0.250	0.250	0.222	0.284	0.200	0.413	0.052	0.049	0.039	0.049	0.065
$\widehat{\text{var}}_{SB}$	0.380	0.375	0.385	0.353	0.337	0.344	0.081	0.078	0.080	0.080	0.080
$\widehat{\text{var}}_{LM2}$	0.214	0.208	0.213	0.196	0.190	0.190	0.050	0.048	0.049	0.048	0.045
$\widehat{\text{var}}_{LM3}$	0.308	0.294	0.298	0.284	0.280	0.276	0.068	0.068	0.069	0.068	0.065
$\widehat{\text{var}}_{LM4}$	0.358	0.349	0.351	0.340	0.337	0.330	0.079	0.080	0.081	0.081	0.078
$\widehat{\text{var}}_{HAJ}$	0.406	0.407	0.407	0.404	0.405	0.403	0.065	0.066	0.066	0.066	0.066

Table 2.5 – *Results of 10000 simualtions of Moran's I_{B_1} spatial measure. The three spatial configurations of the Section 6 are taken into account.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
I_{B_1}	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
Complete spatial randomness											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.028	-0.009	-0.012	0.027	0.093
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.125	-0.095	-0.085	-0.059	0.016
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.436	-0.344	-0.318	-0.229	-0.020
Neyman-Scott process											
$n = 16$	-0.627	-0.411	-0.336	-0.301	-0.255	-0.034	-0.158	-0.133	-0.116	-0.092	0.030
$n = 32$	-0.716	-0.493	-0.416	-0.361	-0.306	-0.020	-0.433	-0.358	-0.297	-0.258	-0.016
$n = 48$	-0.837	-0.570	-0.489	-0.385	-0.276	-0.014	-0.460	-0.381	-0.303	-0.282	-0.046
Simple regular grid											
$n = 16$	-0.576	-0.345	-0.384	-0.251	0.000	-0.028	-0.088	-0.063	-0.066	-0.037	0.031
$n = 32$	-0.618	-0.400	-0.469	-0.274	0.000	-0.017	-0.124	-0.058	-0.077	-0.015	0.079
$n = 48$	-0.743	-0.444	-0.546	-0.345	0.000	-0.012	-0.505	-0.321	-0.383	-0.244	-0.037

2.8 Discussion

Environmental data are generally not uniformly distributed over a region of the space. Thus, it is generally justified to use unequal inclusion probabilities to overrepresent some parts of the population. As explained in Section 2.2.2, this reduces the variance of the Horvitz-Thompson estimator, a phenomenon also observed in Section 2.7 on the Meuse dataset.

In this manuscript, we present a sampling design that selects the units in a very well-spread configuration. We have shown on the Meuse dataset that the method behaves very well on measurements of spatial spreading. Moreover, the approximated variance of the Horvitz-Thompson estimator is lower with WAVE sampling than with the other methods. Some second-order inclusion probabilities are null, therefore it is impossible to estimate unbiasedly the variance of the estimator. However, we propose different estimators and compare their performance. We show that it is possible to estimate appropriately the variance and to construct confidence intervals that have good coverage rates, particularly when the sample size is large. All of these results indicate that our method is very efficient for the selection of a well-spread sample and has better properties than the usual spatial sampling designs.

2.9 Appendix

This section presents the additional results of the Section 6. For each spatial configuration presented in the paper, the three spatial measures are given. For each population, 10,000 samples of size respectively equal to 25, 50 and 100 are selected. Two cases are considered for the inclusion probabilities: equal and unequal. In terms of spatial measures, the WAVE sampling method performs better than the others do. Examples where the spatial balance B behaves badly are given in Fig. 2.7 and Fig. 2.8.

Table 2.6 – *Results of 10000 simulations of Moran's I_B spatial measure. The three spatial configurations of the Section 6 are taken into account.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
I_B	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
Complete spatial randomness											
$n = 16$	-0.530	-0.348	-0.370	-0.220	-0.259	-0.030	-0.459	-0.316	-0.331	-0.201	-0.028
$n = 32$	-0.693	-0.467	-0.464	-0.322	-0.392	-0.017	-0.548	-0.393	-0.373	-0.261	-0.013
$n = 48$	-0.807	-0.583	-0.506	-0.375	-0.373	-0.015	-0.621	-0.469	-0.424	-0.292	-0.029
Neyman-Scott process											
$n = 16$	-0.627	-0.411	-0.336	-0.301	-0.255	-0.034	-0.526	-0.383	-0.339	-0.285	-0.031
$n = 32$	-0.716	-0.493	-0.416	-0.361	-0.306	-0.020	-0.561	-0.410	-0.355	-0.288	-0.015
$n = 48$	-0.837	-0.570	-0.489	-0.385	-0.276	-0.014	-0.614	-0.436	-0.347	-0.301	-0.032
Simple regular grid											
$n = 16$	-0.576	-0.345	-0.384	-0.251	0.000	-0.028	-0.462	-0.317	-0.340	-0.236	-0.026
$n = 32$	-0.618	-0.400	-0.469	-0.274	0.000	-0.017	-0.523	-0.353	-0.389	-0.241	-0.018
$n = 48$	-0.743	-0.444	-0.546	-0.345	0.000	-0.012	-0.564	-0.336	-0.395	-0.252	-0.027

Table 2.7 – *Results of 10000 simulations of the spatial balance measure B. The three spatial configurations of the Section 6 are taken into account.*

	Sampling design										
	Equal probabilities						Unequal probabilities				
B	wave	lpm1	scps	grts	hip	srswor	wave	lpm1	scps	grts	maxent
Complete spatial randomness											
$n = 16$	0.115	0.117	0.108	0.164	0.135	0.338	0.123	0.124	0.118	0.177	0.345
$n = 32$	0.137	0.128	0.130	0.167	0.165	0.345	0.140	0.146	0.138	0.180	0.352
$n = 48$	0.158	0.137	0.149	0.177	0.195	0.337	0.165	0.151	0.158	0.189	0.319
Neyman-Scott process											
$n = 16$	0.127	0.126	0.170	0.162	0.168	0.467	0.124	0.110	0.140	0.160	0.457
$n = 32$	0.182	0.174	0.201	0.205	0.239	0.485	0.158	0.153	0.180	0.193	0.452
$n = 48$	0.187	0.161	0.178	0.206	0.240	0.452	0.189	0.173	0.206	0.220	0.415
Simple regular grid											
$n = 16$	0.050	0.074	0.059	0.096	0.000	0.275	0.069	0.086	0.073	0.107	0.287
$n = 32$	0.051	0.068	0.052	0.081	0.000	0.235	0.072	0.081	0.072	0.111	0.248
$n = 48$	0.040	0.059	0.046	0.065	0.000	0.196	0.073	0.086	0.077	0.102	0.206

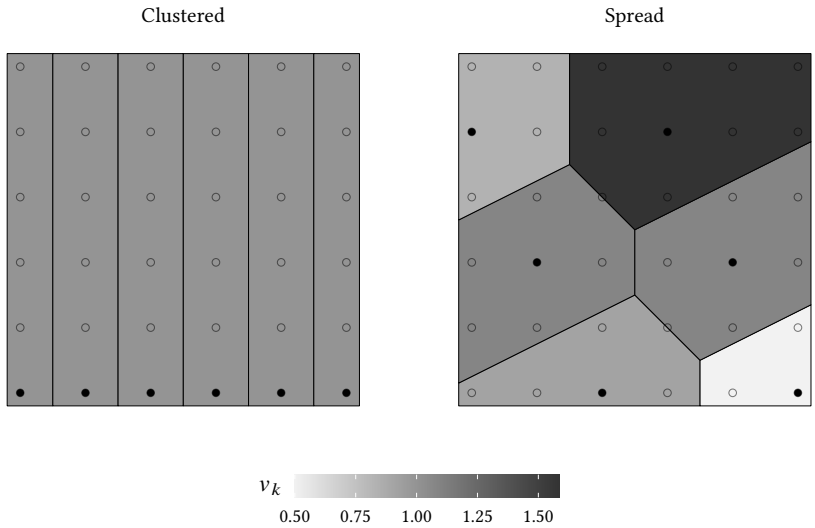


Figure 2.7 – *Let $U = \{1, \ldots, 36\}$ be on a regular grid 6×6, inclusion probabilities are all equal to $\pi_k = \pi = 1/6$. Measure B given in Equation (10) is equal to 0 for the clustered sample while it should be a great value. $B = 0.106$ for the spread sample. On the other hand the Moran index shows a correct behaviour with the two samples. For the clustered sample $I_B = 0.71$ while $I_B =$ -0.718 for the spread sample.*

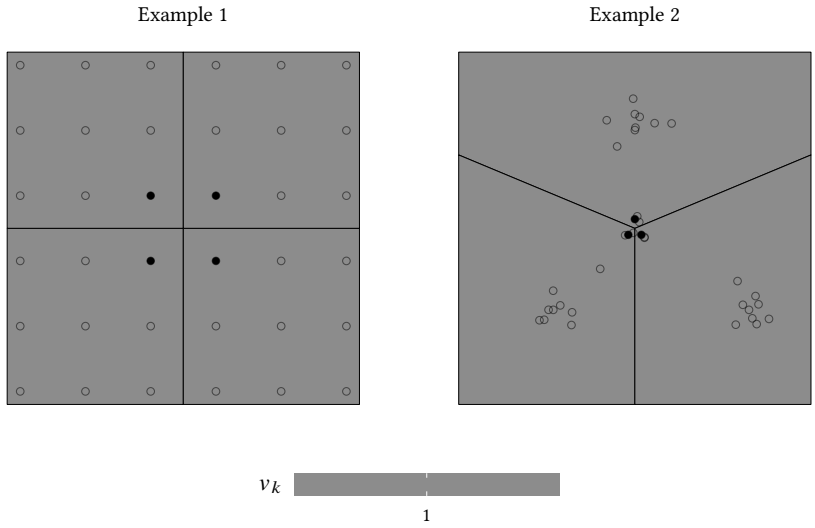


Figure 2.8 – *Let $U = \{1, \ldots, 36\}$. On the left figure, population is on a regular grid 6×6. On the right figure, population is clustered around four units. Inclusion probabilities are all equal to $\pi_k = \pi = 1/9$ (respectively $\pi_k = \pi = 1/12$). Measure B for both examples is equal to 0 while it should be a higher value. On the other hand, the Moran index shows correct behaviour with the two samples. For the Example 1 $I_B = 0.732$ while $I_B = 0.199$ for Example 2.*

Chapter 3

Sequential Spatially Balanced Sampling

Abstract

Sequential sampling occurs when an entire population is unknown in advance and data are received one by one or in groups of units. This manuscript proposes a new algorithm to sequentially select a balanced sample. The algorithm respects equal and unequal inclusion probabilities. The method can also be used to select a spatially balanced sample if the population of interest contains spatial coordinates. A simulation study is proposed, and the results show that the proposed method outperforms other methods. [1]

Keywords: inclusion probability, spread sampling, stream sampling, survey methods

3.1 Introduction

A sample is said to be balanced if the Horvitz-Thompson estimators of the totals calculated from the sample are equal or nearly equal to the population totals. The use of balancing constraints was first proposed by Gini and Galvani (1929). They selected a sample of 29 districts out of 214 to replicate certain population means. The selection of the sample was not random, and the method was strongly criticized by Jerzy Neyman. It is now well known that a sample can be selected randomly and balanced simultaneously. The cube method, which randomly selects a balanced sample, was proposed by Deville and Tillé (2004). Recently, Leuenberger et al. (2022) proposed an improvement. They showed that changing the order of the units before running the algorithm can significantly increase the quality of the sample balance.

In environmental studies particularly, data contain spatial coordinates. When the data are spatially autocorrelated, it is often more accurate to spread the sample in space. Moreover, well-spread samples in space are balanced on auxiliary variables (see Grafström and Lundström, 2013), even if the target parameters are nonlinear in the auxiliary variables. Many sampling methods are currently used to select a well-spread sample. One well-known algorithm is the Generalized Random Tessellation Sampling proposed by Stevens Jr. and Olsen (2004). It maps a multi-dimensional space into a real line and uses one-dimensionnal systematic sampling to select a sample. Grafström et al. (2012) proposed the local pivotal method that introduces repulsion between two nearby units to ensure that close units are not both selected. Benedetti and Piersimoni (2017) proposed recursive modification of the inclusion probability vector using the within sample distance of the spatial structure. Robertson et al. (2018) proposed using the properties of the Halton sequence to select a well-spread sample. More recently,

[1] This chapter is essentially a reprint of: Jauslin, R., Panahbehagh, B., & Tillé, Y., (2022). Sequential Spatially Balanced Sampling, *Environmetrics*, 33(8), e2776

Jauslin and Tillé (2020) proposed adaptation of the cube method using a contiguity matrix to select a well-spread sample.

Grafström and Lundström (2013) showed that well-spread samples could also be balanced on auxiliary variables. However, some methods can do even better. A combination of the cube method and the local pivotal method, named doubly balanced sampling, has been proposed by Grafström and Tillé (2013). This allows the selection of a sample which is simultaneously well-spread and balanced on auxiliary variables. Moreover, they showed that this method drastically diminishes the variance of the Horvitz-Thompson estimator. Furthermore, Vallée et al. (2015) presented an application of the doubly balanced sampling design for forest inventories. Throughout this manuscript, we write "spatially balanced sample" or "well-spread sample" to refer to the same interpretation.

All these methods apply to finite populations. This means that we must have access to the entire population before using the sampling algorithm. We speak of streaming data when the data arrive one by one or in groups of units. Examples are Internet network data, financial data, and environmental or biological studies over a long period of time. In these cases, the units may be distributed asymmetrically, meaning that unusual units may appear in the stream and have a significant impact on the estimator. Choosing balanced samples improves the precision of the estimator. Indeed, if an eccentric unit is selected, the balancing constraints ensure that the variance of the estimator is controlled. Having a sequential procedure for selecting a balanced sample can be useful if the data stream is large. Other methods that select a balanced sample need the entire population to select a sample, whereas with the proposed method, only a group of units needs to be known.

In this article, we test the method on a finite population, rather than a stream application. We have chosen to test the method on a finite population for two reasons. First, we believe it is more important to compare methods on comparable data sets. We can then measure the performance of the method against the usual methods. Secondly, this method is the only one known that selects a balanced sample sequentially, so we would have no point of comparison. The simulation results show that the finite population method with sequential implementation has better properties in terms of variance reduction than the usual methods.

In this manuscript, we propose a new method to select a balanced sample and a spatially balanced sample if spatial coordinates are available. Moreover, the method is sequential (i.e., the algorithm does not need to access the entire population to run). In section 3.2, we introduce the basic concept of survey sampling theory. In section 3.3, we expose the balancing equations formally and give insight into the interest of selecting a sample that is both balanced and well-spread. In Section 3.4, we introduce spread measure, while in Section 3.5, we present the outline of the method and explain how the method selects a spatially balanced sample. Section 3.7 contains simulation results on two datasets.

3.2 Notation

Throughout this manuscript, we consider a finite population U. The size of the population N can be known in advance or estimated, and especially in stream data, the population size is often unknown. A sample $s \subset U$ is a subset of the population U. A sampling design is defined by a probability function $p(.)$ on all possible samples such

that

$$p(s) \geq 0 \text{ and } \sum_{s \subset U} p(s) = 1.$$

Let $S \subset U$ denote a random sample, a random variable with probability distribution defined by the sampling design $P(S = s) = p(s)$. Theoretically, the inclusion probability $\pi_k \in (0, 1)$, $k \in U$ can be deduced from the sampling design

$$\pi_k = \mathrm{P}(k \in S) = \sum_{s \subset U | k \in s} p(s).$$

In practice, inclusion probabilities are predetermined by statisticians, and these could be equal with a fixed sample size (i.e. $\pi_k = n/N$, or unequal, for example proportional to an auxiliary variable). If it is not specified, in this manuscript, the population size N is supposed to be known. Not knowing the parameter N can be problematic to compute inclusion probabilities. In this particular case, a usual practice is to use two phase sampling, i.e. use a minimum approximation of the population and do a second step to adjust if the parameter N has been badly approximated for example if sample size is too large (Tillé, 2019). Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ denote the vector of inclusion probabilities and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_N)^\top$ the sample where

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

Let y denote a variable of interest, where y_k denotes the value of the variable for a particular unit $k \in U$. Let Y denote the total of the variable on the population U:

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

This total can be estimated unbiasedly by using the Horvitz-Thompson estimator of the total defined by

$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k \delta_k}{\pi_k}. \tag{3.1}$$

If data are coming in a stream, we suppose that the units $k \in U$ are coming one by one or by groups of units, for example, subpopulations $\{U_1, U_2, U_3, \ldots\}$. A necessary condition to sample in stream data is to have a sequential sampling method (i.e. the algorithm decides for one unit and then moves on to the next unit. It does not need the whole population to select a unit.). Indeed, as the size of the population might be unknown, it is generally impossible to wait for the full population data. Stream data are various; the reader may find a general outline in Tillé (2019).

Suppose $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top \in \mathbb{R}^p$ are auxiliary variables available for the unit $k \in U$. In an unequal probability sampling design, the aim is to select a sample with a fixed sample size proportional to the variable of interest y. Since the variable of interest is usually unknown for the whole population, we generally use an auxiliary variable. Suppose that an auxiliary variable highly correlated to the variable of interest exists, it could be used to establish unequal inclusion probabilities. As explained in Tillé (2020, Chapter 4), if inclusion probabilities π_k are proportional to the variable of interest y_k, the ratio y_k/π_k, $k \in U$ are approximately constant, which lead to a variance reduction. In the data stream context, the question of sample size with prescribed

inclusion probabilities is notably discussed by Cohen et al. (2009). Inclusion probabilities can then be set up directly during the stream in order to reach the fixed sample size. More specifically, for equal inclusion probabilities, Knuth (1981) proposed the reservoir method. The idea is to select the n first units as a sample and at each step where a unit is added in the stream, the algorithm decides whether this unit is exchanged in the reservoir with another unit. Chao (1982) proposed a generalisation of this method that can handle unequal inclusion probabilities. The particularity of these two algorithms is that the size of the population does not need to be known. In the algorithm of Chao (1982), it is explained that some inclusion probabilities can be modified greater than one at a transition step. A procedure is proposed to modify the inclusion probabilities to ensure that they remain between 0 and 1. This constraint need to be handled also for the proposed method, and we will see this in Section 3.5.

In addition to the auxiliary variables, let us suppose that the population contains spatial coordinates $\mathbf{z}_k \in \mathbb{R}^q$. Populations with spatial coordinates are often spatially correlated (Wang et al., 2012). In that case, selecting a well-spread sample will reduce the variance of the Horvitz-Thompson estimator (Grafström and Lundström, 2013). A stronger insight into this is presented in the next section.

3.3 Balanced Sampling

Suppose $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top \in \mathbb{R}^p$ are auxiliary variables available for the unit $k \in U$. A sample S is said to be balanced on auxiliary variables $\mathbf{x}_k$, $k \in U$ if it satisfies the balancing equations given by

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

Let $\mathbf{A} = (\mathbf{x}_1/\pi_1, \mathbf{x}_2/\pi_2, \ldots, \mathbf{x}_N/\pi_N)^\top = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)^\top$ denote the auxiliary variables expanded by the inclusion probabilities. The selection of a balanced sample can be written as the linear program

$$\begin{cases} \mathbf{A}^\top \boldsymbol{\delta} = \mathbf{A}^\top \boldsymbol{\pi}, \\ \boldsymbol{\delta} \in \{0, 1\}^N. \end{cases}$$

Deville and Tillé (2004) have proposed the cube method to select a balanced sample. The method first consists in performing a random walk inside the hypercube to approach a balanced sample. This first step is called the flight phase. Chauvet and Tillé (2006) have proposed a fast implementation, which modifies the inclusion probabilities $\boldsymbol{\pi}$ into a slightly different vector of inclusion probabilities $\widetilde{\boldsymbol{\pi}}$. This updated vector of inclusion probabilities $\widetilde{\boldsymbol{\pi}}$ verifies exactly the balancing constraints $\mathbf{A}^\top \widetilde{\boldsymbol{\pi}} = \mathbf{A}^\top \boldsymbol{\pi}$. Moreover, at most there are p units that have inclusion probabilities not equal to 0 or 1. The vector $\widetilde{\boldsymbol{\pi}}$ is almost a sample, as it remains only at most p entries that are not equal to 0 or 1 remain. Mathematically, we have that $0 \leq \widetilde{\pi}_k \leq 1$, $\mathrm{E}(\widetilde{\pi}_k) = \pi_k$, $k \in U$, and

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

A second phase is then launched on this updated vector to obtain a sample. This phase is called the landing phase and consists either of dropping balancing constraints one by one or using a linear program to discover the best solution. The latter method

can only be launched if the value of p is not too high, as it could lead to a combinatorial explosion. In the end, the random sample almost satisfies the constraint in the sense that

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}\approx\sum_{k\in U}\mathbf{x}_k.$$

In the presence of not only auxiliary variables but also spatial coordinates, Grafström and Tillé (2013) proposed a doubly balanced method that selects samples that are well-spread and balanced on auxiliary variables at the same time. If the population of interest has spatial coordinates, the units are generally spatially correlated. In fact, if we assume a simple model, we can see that applying both approaches reduces the variance of the estimator. Let us suppose a general linear model:

$$y_k=\mathbf{x}_k^\top\boldsymbol{\beta}+\varepsilon_k,\ \text{for all } k\in U,$$

where ε_k is a random variable that satisfies $\mathrm{E}_M(\varepsilon_k)=0$ and $\mathrm{var}_M(\varepsilon_k)=\varphi^2(\mathbf{x}_k)=\varphi_k^2$, with $\varphi^2(.)$ a Lipschitz continuous function and $\mathrm{E}_M(.)$, var_M being the expectation and the variance under the model, respectively. Spatial correlation is modelled by the function

$$\mathrm{cov}_M(\varepsilon_k,\varepsilon_\ell)=\varphi_k\varphi_\ell\rho_{k\ell},\ \text{with } k\neq\ell\in U$$

where $\rho_{k\ell}$ is a function that decreases when the distance between k and ℓ increases. Under this model, Grafström and Tillé (2013) show that the anticipated variance of the Horvitz-Thompson estimator is

$$\mathrm{E}_p\mathrm{E}_M(\widehat{Y}_{HT}-Y)^2=\mathrm{E}_p\left\{\left(\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}-\sum_{k\in U}\mathbf{x}_k\right)^\top\boldsymbol{\beta}\right\}^2+\sum_{k\in U}\sum_{\ell\in U}\varphi_k\varphi_\ell\rho_{k\ell}\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_k\pi_\ell},\tag{3.2}$$

where $\pi_{k\ell}=\mathrm{E}_p(\delta_k\delta_\ell)$ are the joint inclusion probabilities and $\mathrm{E}_p(.)$ is the expectation under the design. From the first term of Equation (3.2), the reduction of the variance is done by selecting a balanced sample, whereas the second term is reduced if the inclusion probabilities $\pi_{k\ell}$ are small, while $\rho_{k\ell}$ is large. This means that a sample must be selected in a spread and balanced manner to minimize the anticipated variance.

3.4 Spreading Measures

A question arises naturally in spatial sampling: are there measures to see whether a sample is well-spread? Stevens Jr. and Olsen (2003) proposed an index based on the Voronoï polygon. Let b_i, $i\in S$, be the sum of inclusion probabilities within the ith Voronoï polygons. The authors showed that the expected value of b_i is equal to 1 and proposed measuring how to vary these sums using the following quantity:

$$B=\frac{1}{n}\sum_{i\in S}(b_i-1)^2.\tag{3.3}$$

Moran (1950) proposes measuring spatial correlations in particular cases where the spatial coordinates are placed on a grid. Tillé et al. (2018); Jauslin and Tillé (2020) propose a normalized version that uses the notion of a spatial weight matrix $\mathbf{W}$. Different ways to compute a spatial weight matrix exist. A complete explanation on the spatial

weight matrix can be found in Jauslin and Tillé (2020) and Tillé et al. (2018). Moran's index I is given by

$$I = \frac{(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{W} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)}{\sqrt{(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{Q} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)^\top \mathbf{G} (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}_w)}}, \tag{3.4}$$

where $\boldsymbol{\delta}$ is the sample, and

$$\bar{\boldsymbol{\delta}}_w = \frac{\boldsymbol{\delta}^\top \mathbf{W} \mathbf{1}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

$\mathbf{Q}$ is the $N \times N$ diagonal matrix containing, for each k, $\sum_{\ell \in U} w_{k\ell}$ on its diagonal,

$$\mathbf{G} = \mathbf{C}^\top \mathbf{Q} \mathbf{C}, \quad \mathbf{C} = \mathbf{Q}^{-1} \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top \mathbf{W}}{\mathbf{1}^\top \mathbf{W} \mathbf{1}},$$

and $\mathbf{1}$ is a column vector of N ones. These two measures are the ones used to quantify the spread in Section 3.7.

3.5 Outline of the Proposed Method

In this section, we present the main idea of the method. Suppose a population U where units arrive sequentially, by groups of units or one by one. Suppose also that auxiliary variables $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top \in \mathbb{R}^p$ are available for each unit $k \in U$. Inclusion probabilities are supposed to be predetermined by the statistician. They are usually defined as unequal and proportional to an auxiliary variable. In this case, the fixed sample size is satisfied. If population size is unknown different solutions exist to set up inclusion probabilities. A two-phase sampling can be applied using a minimal approximation of the population size N to set up inclusion probabilities. Then a second phase is launched to adjust the sample size if it is too large at the end of the first phase. Another option would be to opt for a non-fixed sample size and set up the inclusion probabilities for each subpopulation U_i. Whichever option is used, in both cases, the following method will respect equal and unequal inclusion probabilities. This result is shown in the following paragraph.

In addition to the inclusion probabilities, we want to respect balancing equations and decide for the current unit observed in the stream. For simplicity, let us suppose that the current unit is the first one. The algorithm works in the following way: it will wait for a certain number of units, denoted J, so that it can decide on the considered unit while satisfying the balancing equations. The number of units to wait depends on the number of auxiliary variables and the inclusion probabilities. The variable J is tested at each step to see if a solution exists. J is therefore dynamic and changes after each decision. In general, J is much smaller than N, and the algorithm compensates the decision taken on the current unit to the $J-1$ remaining units. More formally, we try to find v_k, $k = 2, \ldots, J$, characterized by the following update, for $k = 2, \ldots, J$:

$$\begin{cases} \pi_1^1 = 0, \quad \pi_k^1 = \pi_k + v_k, & \text{with probability } 1 - \pi_1 \\ \pi_1^2 = 1, \quad \pi_k^2 = \pi_k - \dfrac{1 - \pi_1}{\pi_1} v_k, & \text{with probability } \pi_1 \end{cases} \tag{3.5}$$

where π_k^1 are the updated inclusion probabilities if the decision is to omit the current unit k, in this case $k = 1$, and the same goes for π_k^2, if we select the current unit k. The

balancing equations need to be satisfied, after which we obtain the following equality:

$$\sum_{k=1}^{J} \frac{\mathbf{x}_k}{\pi_k} \pi_k^1 = \sum_{k=1}^{J} \frac{\mathbf{x}_k}{\pi_k} \pi_k^2 = \sum_{k=1}^{J} \mathbf{x}_k,$$

which implies

$$\sum_{k=1}^{J} \frac{\mathbf{x}_k}{\pi_k} \pi_k^1 = 0 + \sum_{k=2}^{J} \frac{\mathbf{x}_k}{\pi_k} (\pi_k + v_k) = \frac{\mathbf{x}_1}{\pi_1} + \sum_{k=2}^{J} \frac{\mathbf{x}_k}{\pi_k} \left(\pi_k - \frac{1-\pi_1}{\pi_1} v_k \right) = \sum_{k=1}^{J} \frac{\mathbf{x}_k}{\pi_k} \pi_k,$$

and

$$\sum_{k=2}^{J} \frac{\mathbf{x}_k}{\pi_k} v_k = \frac{\mathbf{x}_1}{\pi_1} - \frac{1-\pi_1}{\pi_1} \sum_{k=1}^{J} \frac{\mathbf{x}_k}{\pi_k} v_k = \mathbf{x}_1,$$

and thus

$$\sum_{k=2}^{J} \frac{\mathbf{x}_k}{\pi_k} v_k = \mathbf{x}_1.$$

Note that by updating the inclusion probabilities as in Equation (3.5), independently of the value v_k, the method respects inclusion probabilities, for all $k = 2, \ldots, J$,

$$\begin{aligned} \mathrm{E}_p(\pi_k) &= (1-\pi_1)\pi_k^1 + \pi_1 \pi_k^2 \\ &= (1-\pi_1)(\pi_k + v_k) + \pi_1 \left(\pi_k - \frac{1-\pi_1}{\pi_1} v_k \right) \\ &= (1-\pi_1)\pi_k + (1-\pi_1)v_k + \pi_1 \pi_k - (1-\pi_1)v_k \\ &= \pi_k, \end{aligned}$$

where $\mathrm{E}_p(.)$ is the expectation under the design. To ensure that updated inclusion probabilities remain between 0 and 1, we must also have

$$0 \leq \pi_k^1, \pi_k^2 \leq 1, \text{ for } k = 2, \ldots, J,$$

which induces the following constraints:

$$\max \left\{ -\pi_k, (\pi_k - 1) \frac{\pi_1}{1-\pi_1} \right\} \leq v_k \leq \min \left(1 - \pi_k, \pi_k \frac{\pi_1}{1-\pi_1} \right), k = 2, \ldots, J. \tag{3.6}$$

In order to find v_k, we propose to solve the following program:

$$\begin{cases} \text{maximize} & \displaystyle\sum_{k=2}^{J} v_k c_k \\ \text{subject to} & \displaystyle\sum_{k=2}^{J} \frac{\mathbf{x}_k}{\pi_k} v_k = \mathbf{x}_1, \\ & v_k \geq -\min \left\{ \pi_k, (1-\pi_k) \frac{\pi_1}{1-\pi_1} \right\}, \\ & v_k \leq \min \left(1 - \pi_k, \pi_k \frac{\pi_1}{1-\pi_1} \right), \end{cases} \tag{3.7}$$

where c_k is a cost function that is supposed to penalize the units that are distant to the current one. A naive function is $c_k = (J - k)$, which decreases as k increases. If there is no solution, the value of J is incremented by one, and we recompute until a solution is found. The existence of the solution of the linear program (3.7) is not

guaranteed, mainly because if J is not large enough, the balancing constraints and the positivity of the updated inclusion probabilities cannot be satisfied jointly. However, as J increases, the feasible region expands. The initial value for the variable J must be larger than the dimension of $\mathbf{x}_k$. If we let $\mathbf{x}_k = \pi_k$, then the method only respects inclusion probabilities, and we obtain a fixed sample size.

Suppose now that some steps have proceeded and that the current step is the ith one. Let $P_i = \{\pi_i, \ldots, \pi_{i+M}\}$ denote the pool of available units at step i, where $M \in \mathbb{N}$. Note that the size of the pool M is not directly related to J. The pool P_i is actually all the available units for which we have waited to make a decision in the previous $i-1$ steps. The algorithm will choose a unit in P_i and decide on a certain amount of units J, with $J \leq M$. The constraint on v_k depends on π_i (i.e., the inclusion probability of the current unit i might have a serious impact on the potential value that v_k could take). Figure 3.1 shows possible v_k against π_k, for different values of π_i. The value π_k varies from 0 to 1, and the lower and upper bounds of the value v_k are plotted. When π_i is very small, it ends with a tight interval for v_k. On the other hand, as the inclusion probabilities π_i increase, we obtain a larger range for v_k. If the inclusion probability of the current unit i is very small, we might have a sharp increase in the value J. This is because the balancing equations are not symmetric for unequal probability sampling. If a balanced sample is selected with unequal inclusion probabilities, the complementary sample is not balanced.

To optimize the problem of a sharp increase in the value J, we propose, in all cases, selection of the unit that has the largest inclusion probability in the pool P_i of available units. By taking the one that has the largest inclusion probability and reordering the units with respect to it, we optimize the size of J.

If spatial coordinates are available, the proposed method could be used in an interesting way to select a well-spread sample. Let $P_{(i)} = \{\pi_{(i)}, \ldots, \pi_{(i+M)}\}$, such that $\pi_{(i+j)}$ is the jth closest unit to $\pi_{(i)}$. The pool is thus reordered with respect to the distance to the unit that has maximum inclusion probability. Reordering will then modify the primary inclusion probabilities of units closer to the current one. If a window has a sum of inclusion probabilities close to 1, and if the decision of the ith unit is to select it (i.e., the inclusion probability is transformed to 1), nearest neighbours of the ith unit will not be selected. This leads to a selection process that spreads very well the sample in the considered space.

If population units come from a stream, it is possible that some units have not yet appeared and that their spatial coordinates are in the neighbourhood of the ith unit. This side effect cannot be controlled, because, if we do not control which unit appears in the stream, it is not possible to know whether a future unit might appear within the considered window.

The proposed method is comparable to the doubly balanced cube method (Grafström and Tillé, 2013). Depending on the auxiliary variables, there is a possibility that the method does not end up directly with a sample. Some units might remain with inclusion probabilities not equal to an integer. In this case, we propose to land the method by launching the same process proposed in the landing phase of the doubly balanced method or cube method, either by suppression of variables or by linear programming. See Algorithm 2 for all details of the implementation.

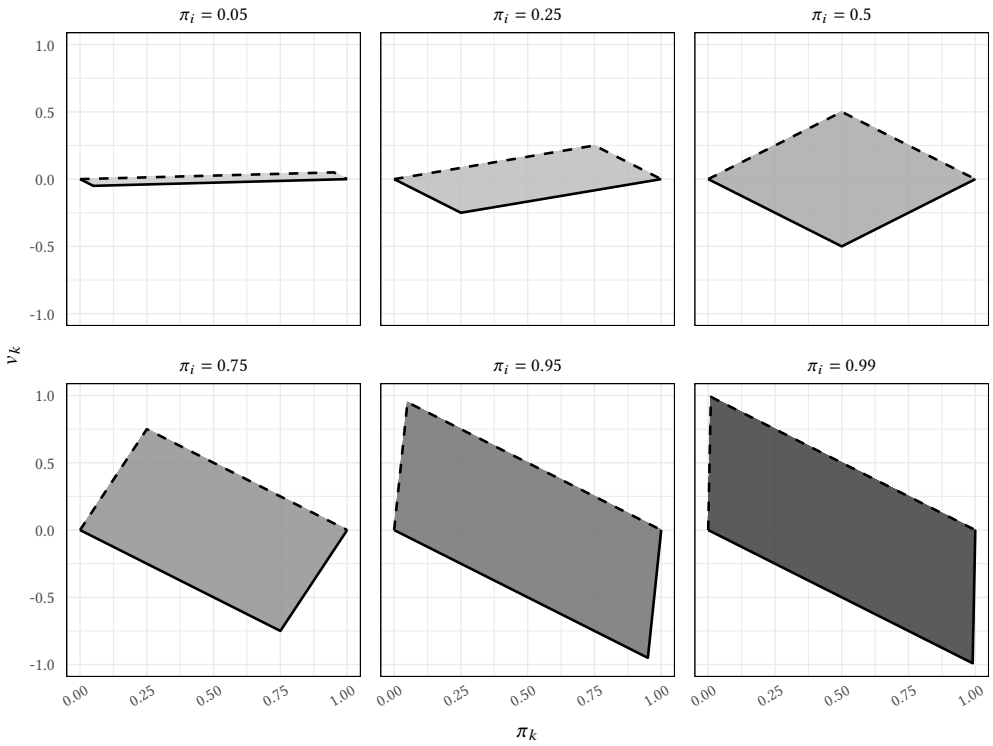


Figure 3.1 – *For six different inclusion probabilities π_i, we let π_k vary from 0 to 1. On the y-axis, the bounds of v_k in Equations (3.6) are calculated. The colored area represents eligible values of v_k, the shaded line is the upper bound, while the bottom line is the lower bound.*

3.6 Variance Estimation

Variance estimation of the Horvitz-Thompson estimator (3.1) requires second-order inclusion probabilities. In general, these probabilities are not known, therefore, different variance estimators have been proposed. For maximum entropy design such as conditional Poisson sampling, an appropriate estimator is the Hájek-Rosén estimator,

$$\widehat{\text{var}}_{HAJ}(\widehat{Y}) = \frac{n}{n-1}\sum_{k\in S}(1-\pi_k)\left\{\frac{y_k}{\pi_k} - \frac{\sum_{\ell\in S} y_\ell(1-\pi_\ell)/\pi_\ell}{\sum_{\ell\in S}(1-\pi_\ell)}\right\}^2. \tag{3.8}$$

Deville and Tillé (2005) show that the variance for balanced sampling methods can be computed as conditional variance with respect to balancing constraints. In particular, they propose a general formula for variance approximation:

$$\text{var}_{app}(\widehat{Y}) = \sum_{k\in U}\sum_{\ell\in U}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}\Delta_{k\ell},$$

where

$$\Delta_{k\ell} = \begin{cases} b_k - b_k\mathbf{a}_k^\top\left(\sum_{i\in U} b_i\mathbf{a}_i\mathbf{a}_i^\top\right)^{-1}\mathbf{a}_k b_k & k = \ell \\ -b_k\mathbf{a}_k\left(\sum_{i\in U} b_i\mathbf{a}_i\mathbf{a}_i^\top\right)^{-1}\mathbf{a}_\ell b_\ell & k \neq \ell. \end{cases}$$

Different values for parameter b_k can be chosen. To obtain the exact variance of the simple random sampling without replacement of fixed sample size, we must have

Algorithm 2: Algorithm of sequential balanced sampling

Let $\boldsymbol{\pi}$ be the inclusion probability vector, $\mathbf{z}_k \in \mathbb{R}^q$ the spatial coordinates, and $\mathbf{x}_k \in \mathbb{R}^p$ the auxiliary variables of the kth unit. For $i = 1, 2, \ldots,$ suppose a pool of units $P_i = \{\pi_i, \ldots, \pi_{i+M}\}$.

1. Find $\pi_{(i)} \in P_i$, the maximum inclusion probability value in the pool. Define $P_{(i)} = \{\pi_{(i)}, \ldots, \pi_{(i+M)}\}$ as the pool of units reordered with respect to the distance of $\pi_{(i)}$ calculated using the spatial coordinates $\mathbf{z}_k$ (i.e. $\pi_{(i+j)}$ is the jth closest unit to the unit $\pi_{(i)}$).

2. Find J and v_k, $k = i+1, \ldots, J$, such that the following linear program has a solution:

$$\begin{cases} \text{maximize} & \displaystyle\sum_{k=i+1}^{J} v_k (J - (k-i)) \\ \text{subject to} & \displaystyle\sum_{k=i+1}^{J} \frac{\mathbf{x}_{(k)}}{\pi_{(k)}} v_k = \mathbf{x}_{(i)} \\ & v_k \geq -\min\left\{\pi_{(k)}, (1-\pi_{(k)})\frac{\pi_{(i)}}{1-\pi_{(i)}}\right\}, \\ & v_k \leq \min\left(1-\pi_{(k)}, \pi_{(k)}\frac{\pi_{(i)}}{1-\pi_{(i)}}\right) \end{cases}$$

   At each step, J increases, and units are reordered with respect to the distance of the current unit.

3. Inclusion probabilities are modified on $P_{(i)}$. For $k = i+1, \ldots, J$, let $q = \pi_{(i)}$,

$$\begin{cases} \pi_{(i)} = 0 \quad \text{and} \quad \pi_{(k)} = \pi_{(k)} + v_k & \text{with probability } 1-q, \\ \pi_{(i)} = 1 \quad \text{and} \quad \pi_{(k)} = \pi_{(k)} - \dfrac{1-\pi_{(i)}}{\pi_{(i)}} v_k & \text{with probability } q. \end{cases}$$

   The decision is then taken for unit $\pi_{(i)}$.

4. The current unit $\pi_{(i)}$ and units that have inclusion probabilities transformed into an integer are removed from the pool, and new units are included.

5. Repeat steps 1-4 until it is no longer possible to find a solution to the linear program or if all inclusion probabilities are modified to an integer value.

6. In general, Some inclusion probabilities might not be equal to 0 or 1. A landing phase is launched by using either the doubly balanced sampling algorithm or the landing phase of the cube method if there is no spatial coordinate.

$b_k = \pi_k(1-\pi_k)\frac{N}{N-p}$. This approximated variance, calculated on the population using the same formula on the random sample S, gives an estimator of the variance:

$$\widehat{\text{var}}_{BAL}(\widehat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \widehat{\Delta}_{k\ell}, \tag{3.9}$$

where

$$\widehat{\Delta}_{k\ell} = \begin{cases} c_k - c_k \mathbf{a}_k^\top \left(\sum_{i \in S} c_i \mathbf{a}_i \mathbf{a}_i^\top \right)^{-1} \mathbf{a}_k c_k & k = \ell \\ -c_k \mathbf{a}_k \left(\sum_{i \in S} c_i \mathbf{a}_i \mathbf{a}_i^\top \right)^{-1} \mathbf{a}_\ell c_\ell & k \neq \ell \end{cases}$$

and $c_k = (1 - \pi_k)\frac{n}{n-p}$. Equation (3.9) can be rewritten by using residuals of the linear model,

$$e_k = y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}},$$

where

$$\widehat{\boldsymbol{\beta}} = \left\{ \sum_{\ell \in S} (1 - \pi_\ell) \mathbf{a}_\ell \mathbf{a}_\ell^\top \right\}^{-1} \sum_{\ell \in S} (1 - \pi_\ell) \mathbf{a}_\ell \frac{y_\ell}{\pi_\ell},$$

giving another expression for the Equation (3.9),

$$\widehat{\text{var}}_{BAL}(\widehat{Y}) = \frac{n}{n-p} \sum_{k \in S} (1 - \pi_k) \left(\frac{e_k}{\pi_k} \right)^2.$$

More details about variance estimation under balanced sampling designs can be found in Deville and Tillé (2005). Estimator (3.9) does not consider the spatial structure of the population. Grafström and Tillé (2013) propose another estimator based on a combination of the variance estimator (3.9) and the purely spatial variance estimator proposed by Stevens Jr. and Olsen (2003),

$$\widehat{\text{var}}_{DBS}(\widehat{Y}) = \frac{n}{n-p} \frac{p+1}{p} \sum_{k \in S} (1 - \pi_k) \left(\frac{e_k}{\pi_k} - \bar{e}_k \right)^2, \tag{3.10}$$

where

$$\bar{e}_k = \frac{\sum_{\ell \in G_k} (1 - \pi_\ell) e_l / \pi_\ell}{\sum_{\ell \in G_k} (1 - \pi_\ell)},$$

and G_k is the set of the $p + 1$ closest units of k in the sample. Grafström and Schelin (2014) also developed a generalized local mean variance estimator,

$$\widehat{\text{var}}_{SB}(\widehat{Y}) = \frac{1}{2} \sum_{k \in S} \left(\frac{y_k}{\pi_k} - \frac{y_{\ell_k}}{\pi_{\ell_k}} \right)^2, \tag{3.11}$$

where ℓ_k is the nearest neighbour to the unit k in the random sample S. This expression of the estimator is a simple version where no equal distance exists. In its more general expression, the estimator can handle equal distance between units in the population and is shown as appropriate for purely spatial sampling designs, such as in the local pivotal method (Grafström et al., 2012), the proportional within distance method (Benedetti and Piersimoni, 2017), Halton iterative partitioning (Robertson et al., 2018), and the weakly associated vectors (Jauslin and Tillé, 2020). The following section is contains a complete analysis of these methods and estimators.

3.7 Simulations

3.7.1 Motivation on an artificial dataset

In this section, two artificial datasets are generated to analyse the performance of the proposed method. The two datasets are generated using functions of the package `spatstat` developed by Baddeley and Turner (2005). The first dataset is generated from a complete spatial random process (CSR) using the function `rpoispp`, while the second dataset comes from a Neyman-Scott (NS) process using the function `rNeymanScott`. Figure 3.2 shows the two datasets generated; the CSR process is completely random on the considered space, while the NS is clustered. The population size N of each of the two datasets is equal to 300.

Different methods are compared with the proposed method. Firstly, we compare with the doubly balanced sampling design proposed by Grafström and Tillé (2013), and as this method selects a well-spread balanced sample, the performance should be at least as good as our method. To see if we gain by incorporating the spatial structure of the population into the sampling method, we compare it with the cube method (Deville and Tillé, 2004; Chauvet and Tillé, 2006). Next, we compare purely spatial sampling designs, namely, the local pivotal method (Grafström et al., 2012), the proportional within distance method (Benedetti and Piersimoni, 2017), the Halton iterative partitioning method (Robertson et al., 2018) and the weakly associated vectors method (Jauslin and Tillé, 2020). For each dataset, two sets of inclusion probabilities are computed, one with equal inclusion probabilities and the other with unequal inclusion probabilities. Five auxiliary variables are generated using different probability distributions, a Gaussian distribution $X_1 \sim \mathcal{N}(0,1)$, an exponential distribution $X_2 \sim \mathcal{E}(1)$, a gamma distribution $X_3 \sim \Gamma(3,1)$, a beta distribution $X_4 \sim \mathcal{B}(2,5)$, and uniform distribution $X_5 \sim \mathcal{U}(0,3)$. In addition, variable of interest y is computed using the generated auxiliary variables

$$y = f(z_1, z_2) + X_1 + X_2 + X_3 + X_4 + X_5 + \mathcal{N}(0,1), \tag{3.12}$$

where

$$f(z_1, z_2) = 15\exp\left[-\left\{\left(z_1 - \frac{1}{2}\right)^2 - \frac{3}{5}\left(z_1 - \frac{1}{2}\right)\left(z_2 - \frac{1}{2}\right) - \left(z_2 - \frac{1}{2}\right)^2\right\}\right].$$

The quantity $f(z_1, z_2)$ is a spatial autocorrelation computed from the density of a bivariate Gaussian distribution. The variable y is linear in the auxiliary variables. Figure 3.2 shows the set of unequal inclusion probabilities proportional to the quantity $f(z_1, z_2)$. Table 3.2 shows the relative deviation

$$RD^j = 100 \times \frac{|\,\widehat{t}_{X_i}(s) - t_{X_i}\,|}{t_{X_i}}, \tag{3.13}$$

where t_{X_i} is the true total of the variable X_i, $i \in \{1, \ldots, 5\}$, $\widehat{t}_{X_i}(s) = \sum_{k \in s} X_{i,k}/\pi_k$ are the estimated totals of the auxiliary variables $\{X_1, \ldots, X_5\}$ calculated on the sample s, which is a realization of the random sample S specified by the sampling design, and finally, j is a notation to specify the sampling design used, for example RD^{srs} for simple random sampling. The relative deviation shows how well the totals of the auxiliary variables are respected compared to the true totals. We observe that the three

methods that select a balanced sample considerably decrease the relative deviation. All the methods respect the sample size.

Let Y denote the total of the variables y. For each sampling design, we compute the simulated variance based on m simulations. Let us denote this quantity by

$$\text{var}^{j}_{sim}(\widehat{Y}) = \frac{1}{m}\sum_{s}\left\{\widehat{Y}_{HT}(s) - Y\right\}^2. \tag{3.14}$$

To measure the accuracy of the method, we compute the ratio between the simulated variance of the sampling design and a reference simulated variance, namely, simple random sampling with a fixed sample size for equal inclusion probabilities and a conditional Poisson sampling for the set of unequal inclusion probabilities. We call this quantity the relative variance reduction denoted by

$$RV^{j} = 100 \times \frac{\text{var}^{j}_{sim}(\widehat{Y})}{\text{var}^{srs}_{sim}(\widehat{Y})}. \tag{3.15}$$

This quantity gives the percentage of variance reduction compared to the maximum entropy sampling design. In other words, it measures the accuracy of the sampling design. The smaller this value is, the better we reduce the variance compared to the simple random sampling (respectively for unequal designs, the conditional Poisson sampling).

To see if the variance estimators discussed in Section 3.6 can retrieve the true variance, we compute the ratio between the variance estimator and the simulated variance. This quantity is named the relative efficiency of the variance estimator and is denoted by

$$RE^{j} = 100 \times \frac{\widehat{\text{var}}^{j}(\widehat{Y})}{\text{var}^{j}_{sim}(\widehat{Y})}. \tag{3.16}$$

Note that the variance estimator $\widehat{\text{var}}^{j}(\widehat{Y})$ depends on the sampling design j. For the proposed method and the doubly balanced sampling design, we use estimator (3.10), while for the local pivotal design, the proportional within distance, the Halton iterative partitioning, and the weakly associated vectors methods, we use the variance estimator (3.11). Finally, for the simple random sampling design and conditional Poisson sampling design, we use the Hajek-Rosen estimator (3.8). From Table 3.1, based on the results of the 10,000 simulations, we note that in terms of spread measures, the proposed method is slightly better compared to the doubly balanced sampling design. Of course, a purely spatial sampling design gives a better spreading measure. However, in terms of variance, the proposed method is the one that decreases the variance of the Horvitz-Thompson estimator the most compared to the simple random sampling design or condition Poisson sampling. The variance estimators show acceptable performance except for the unequal balanced sampling design. As the variable of interest is correlated to the balancing variables, the variance depends more on the rounding problem. The variance estimators discussed in Section 3.6 are known to fail to consider the rounding problem. This effect is discussed further in Leuenberger et al. (2022).

3.7.2 Real example on amphibians dataset

In this section, the performance of the proposed method is analysed on a real dataset. The “Centre de coordination pour la protection des amphibiens et des reptiles de Su-

Table 3.1 – *Results of 10,000 simulations on the variables of interest* (3.12). *The first column represents relative variance reduction* (3.15). *The second columns contains the relative variance estimator efficiency* (3.16). *The third and fourth columns correspond to the two spatial measures* (3.3) *and* (3.4), *respectively.*

	Simulated Variances	Variance Estimators	Spread measures	
	RV	*RE*	*B*	*I*
Neyman-Scott process				
Equal				
Proposed method	15.303	99.981	0.276	-0.109
Doubly balanced	20.427	74.581	0.316	-0.070
Cube method	28.439	80.492	0.470	-0.016
Local pivotal	75.684	112.498	0.145	-0.328
Proportional within distance	77.837	115.138	0.091	-0.456
Halton iterative partitioning	56.588	140.473	0.200	-0.216
Weakly associated vector	77.830	109.194	0.145	-0.524
Simple random sampling	100.000	103.678	0.465	-0.017
Unequal				
Proposed method	6.976	2.680	0.277	-0.115
Doubly balanced	8.567	2.188	0.317	-0.070
Cube method	7.315	2.543	0.469	-0.015
Local pivotal	98.123	110.137	0.149	-0.323
Weakly associated vector	97.282	109.907	0.148	-0.516
Conditional Poisson sampling	100.000	99.177	0.473	-0.015
Complete Spatial Randomness				
Equal				
Proposed method	15.848	112.268	0.202	-0.115
Doubly balanced	22.457	77.951	0.237	-0.068
Cube method	30.954	79.184	0.355	-0.015
Local pivotal	83.738	108.654	0.109	-0.312
Proportional within distance	82.931	94.118	0.067	-0.433
Halton iterative partitioning	80.341	111.220	0.117	-0.262
Weakly associated vector	81.762	112.643	0.099	-0.487
Simple random sampling	100.000	99.514	0.353	-0.018
Unequal				
Proposed method	8.696	1.682	0.202	-0.117
Doubly balanced	9.126	1.599	0.240	-0.064
Cube method	7.558	1.906	0.348	-0.014
Local pivotal	92.683	107.609	0.110	-0.309
Weakly associated vector	98.593	99.883	0.101	-0.479
Conditional Poisson sampling	100.000	97.904	0.349	-0.014

Table 3.2 – *Results of 10,000 simulations of the relative deviation (3.13) on the artificial dataset presented in Section 3.7.1.*

	X_1	X_2	X_3	X_4	X_5
Neyman-Scott process					
Equal					
Proposed method	72.004	4.689	1.984	2.005	2.176
Doubly balanced	81.825	5.151	2.380	2.226	2.432
Cube method	75.079	4.383	2.103	2.050	2.318
Local pivotal	342.078	15.724	8.238	8.387	9.049
Proportional within distance	364.628	14.762	8.306	7.886	9.003
Halton iterative partitioning	306.922	14.245	10.358	7.924	7.620
Weakly associated vector	344.991	15.149	8.004	8.501	9.115
Simple random sampling	335.823	15.640	8.086	8.117	9.015
Unequal					
Proposed method	82.261	5.503	2.234	2.328	2.439
Doubly balanced	86.840	5.613	2.405	2.409	2.562
Cube method	75.682	4.807	2.131	2.101	2.376
Local pivotal	346.160	15.770	8.121	8.326	9.017
Weakly associated vector	348.579	15.658	8.396	8.331	9.003
Conditional Poisson sampling	339.662	15.917	8.325	8.378	9.035
Complete Spatial Randomness					
Equal					
Proposed method	70.456	3.457	1.902	2.019	2.426
Doubly balanced	80.290	4.807	2.348	2.201	2.494
Cube method	75.396	4.468	2.090	2.032	2.301
Local pivotal	332.964	15.811	8.066	8.379	8.935
Proportional within distance	325.536	15.990	7.955	7.997	9.319
Halton iterative partitioning	306.923	11.961	8.593	7.502	9.828
Weakly associated vector	327.274	15.868	8.158	8.236	8.851
Simple random sampling	337.471	15.548	8.035	8.270	8.964
Unequal					
Proposed method	74.921	3.877	2.016	2.479	2.733
Doubly balanced	83.685	4.773	2.319	2.539	2.743
Cube method	75.100	4.367	2.083	2.197	2.434
Local pivotal	332.398	15.636	8.111	8.486	9.294
Weakly associated vector	330.964	15.606	7.995	8.477	9.123
Conditional Poisson sampling	339.515	15.720	8.100	8.485	9.117

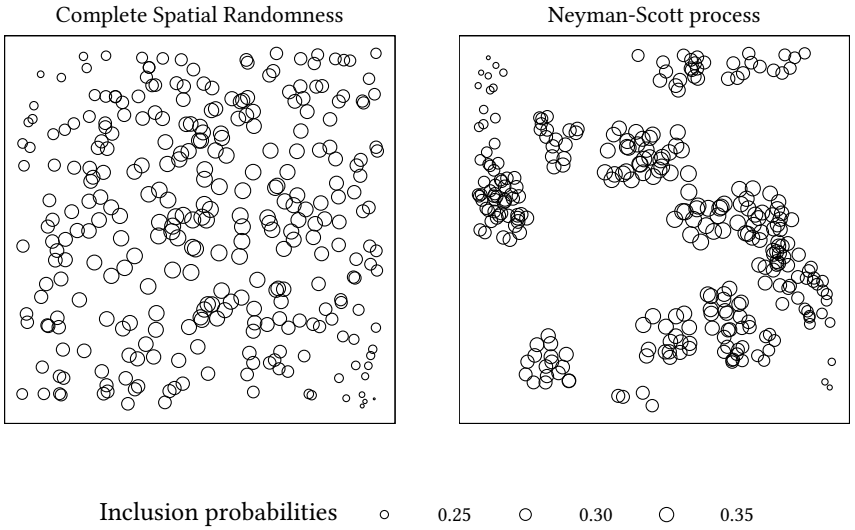


Figure 3.2 – *Simulated dataset used for the analysis of Section 3.7.1. The two datasets each contain 300 units.*

isse" provided a dataset containing spatial coordinates and auxiliary variables on amphibians. This dataset contains a list of 959 sites on which 19 species of amphibians are potentially observed. Figure 3.3 shows two plots: the upper plot shows biogeographical regions of Switzerland, while the lower plot shows the different sites. The sizes of the sites are displayed as well as the diversity score, which is the count of the number of species observed on the site divided by the rarity score of the species. The rarity score is an ordinal variable that determines the scarcity level of the species {1 = "endangered", 2 = "critically endangered", 3 = "threatened", 4 = "potentially threatened"}.

Other auxiliary variables are available in the dataset. The ones used for our analysis are the altitude of the sites, the indicator variables of the biogeographical region, and the area of the sites. Different sampling designs are analysed; the first one uses equal inclusion probabilities, while the unequal design uses inclusion probabilities proportional to the area variable. Table 3.4 shows the relative deviation of the auxiliary variables. Table 3.3 shows the results of 10,000 simulations. As explained in Section 3.7.1, the performance of the proposed method is analysed using the ratios (3.15) and (3.16). We note that the performance of our method is comparable to that of the doubly balanced sampling. The proposed method has good properties: it reduces the variance; shows good spread measures; and the variance estimator fails to catch the rounding problem for unequal balanced sampling designs.

3.8 Conclusion

The increasing amount of data in our century is going to make stream sampling very useful in a few decades. In some experiments in environmental studies, it is impossible to know the entire population, and so sequential sampling is crucial. In this manuscript, we propose a completely new sequential method to select a balanced sample. This method respects equal and unequal inclusion probabilities. We showed through simulations that the proposed method is comparable with the doubly balanced

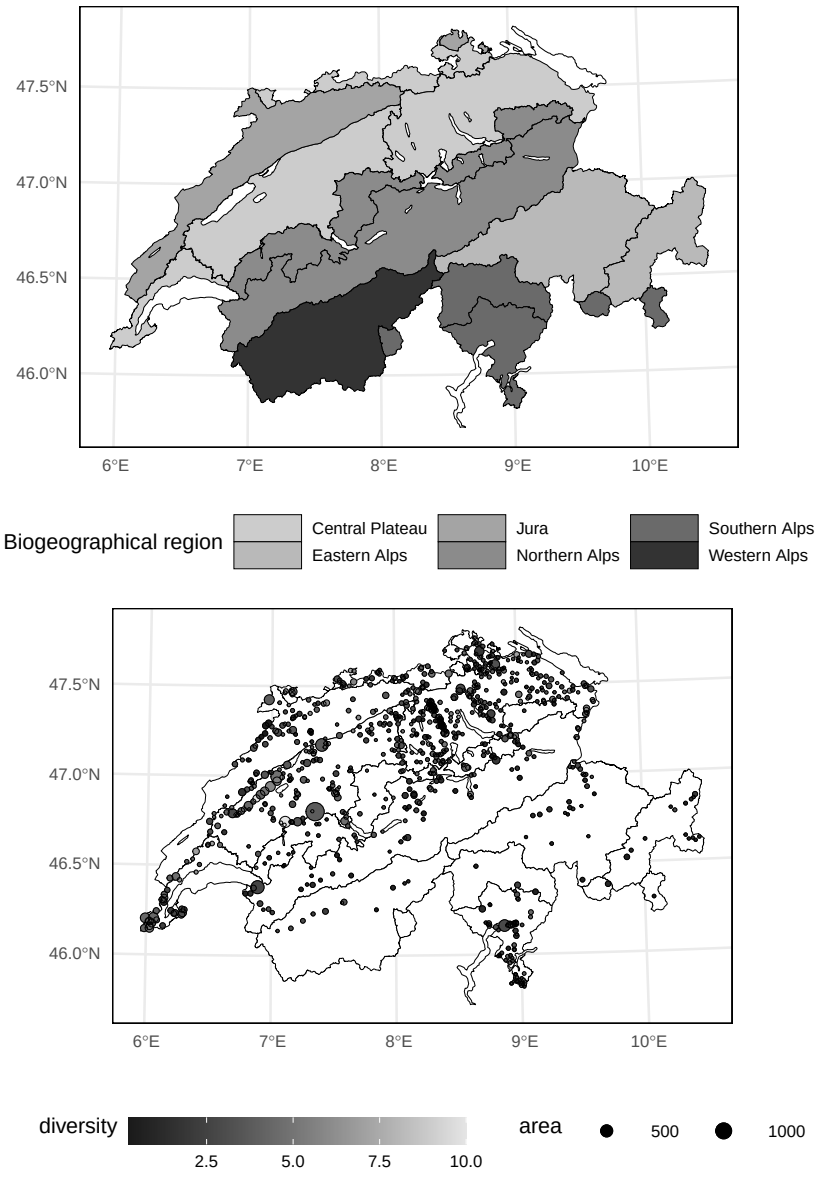


Figure 3.3 – *The upper plot displays the biogeographical region of Switzerland, while the lower plot gives the different sites of the amphibians dataset. The size of the point shows the area and the gradient the diversity score.*

Table 3.3 – *Results of 10,000 simulations on the diversity score of the amphibian dataset. The first column represents the relative variance reduction* (3.15). *The second columns contains the relative variance estimator efficiency* (3.16). *The third and fourth columns correspond to the two spatial measures* (3.3) *and* (3.4) *respectively.*

	Simulated Variance	Variance Estimator	Spread measures	
	RV	*RE*	*B*	*I*
Equal				
Proposed method	71.376	106.592	0.280	-0.160
Doubly balanced	69.231	110.391	0.196	-0.243
Cube method	73.937	103.168	0.400	-0.019
Local pivotal	70.240	119.056	0.132	-0.387
Proportional within distance	96.908	86.323	0.089	-0.479
Halton iterative partitioning	95.734	86.451	0.204	-0.223
Simple random sampling	100.000	95.826	0.435	-0.005
Unequal				
Proposed method	68.453	6.842	0.225	-0.104
Doubly balanced	50.478	9.238	0.219	-0.092
Cube method	38.209	12.221	0.397	0.014
Local pivotal	90.041	103.984	0.154	-0.148
Conditional Poisson sampling	100.000	98.740	0.407	0.016

Table 3.4 – *Results of 10,000 simulations of the relative deviation* (3.13) *for the amphibians dataset presented in Section 3.7.2.*

	Area	Jura	Central Plateau	N Alps	W Alps	E Alps	S Alps
Equal							
Proposed method	0.765	4.068	0.352	3.208	33.531	11.482	5.085
Doubly balanced	0.795	4.057	0.352	3.219	33.883	11.379	5.149
Cube method	0.729	4.065	0.350	3.215	32.942	11.159	4.933
Local pivotal	2.832	10.403	2.731	8.255	42.388	17.321	5.691
Proportional within distance	4.064	9.891	2.836	7.221	37.176	13.419	4.595
Halton iterative partitioning	2.705	12.844	3.335	8.861	39.506	19.568	9.369
Simple random sampling	4.510	20.969	6.096	17.747	66.740	41.436	24.637
Unequal							
Proposed method	14.574	28.955	12.366	31.117	96.362	99.002	56.868
Doubly balanced	14.805	28.697	12.142	32.201	96.722	97.855	57.298
Cube method	14.423	27.578	10.995	29.972	98.300	98.165	56.205
Local pivotal	23.154	39.719	20.170	40.282	101.595	100.094	60.179
Conditional Poisson sampling	24.605	44.360	22.206	46.113	110.398	111.400	68.156

sampling design in terms of variance of the Horvitz-Thompson estimator for a finite population. To conclude, the method is an effective improvement for stream sampling and proposes a real enhancement in this field.

3.9 Acknowledgements

We would like to thank the “Centre de coordination pour la protection des amphibiens et des reptiles de Suisse” for providing us with a dataset to evaluate our method. The author gratefully thanks the associate editor and two reviewers for their conscientious reading and positive comments, which improved the quality of this manuscript.

Part II

Highly stratified sampling and statistical matching

Chapter 4

Enhanced Cube Implementation For Highly Stratified Population

Abstract

A balanced sampling design should always be the adopted strategy if auxiliary information is available. In addition, integrating a stratified structure of the population in the sampling process can considerably reduce the variance of the estimators. We propose here a new method to handle the selection of a balanced sample in a highly stratified population. The method improves substantially the commonly used sampling designs and reduces the time-consuming problem that could arise if inclusion probabilities within strata do not sum to an integer. [1]

Keywords: balanced sampling, clustered sampling, auxiliary information, unequal probability sampling

4.1 Introduction

In survey statistics, balanced sampling is a particularly efficient method when values of auxiliary variables are available for all units in the population. The idea is to select the sample so that the totals of the Horvitz-Thompson estimators of some auxiliary variables equal the population totals. There are different methods for selecting a balanced sample. Deville and Tillé (2004) have proposed the cube method which successively transforms the vector of inclusion probabilities into a sample. The method has been improved by Chauvet and Tillé (2006) by reducing the computation time.

In many areas, it is very useful to use stratified sampling designs. As already indicated by Neyman (1934), the variance of the Horvitz-Thompson estimator can be reduced by constructing strata such that the variables are homogeneous within the strata. Besides, Chauvet (2009) proposed a specific algorithm to obtain balanced samples in the strata of a population. However, this method becomes cumbersome when the number of strata is large.

A highly stratified population is very common in survey sampling. For example, it may be necessary to select individuals from a population while requiring that at most only one individual from each household in a population is taken. Each household is then a stratum. In spatial statistics, we can also construct small strata of neighbouring units to obtain well-spread samples. Highly stratified sampling is also necessary for some donor imputation methods: the objective is to select a respondent for each non-respondent to impute its values. Each non-respondent then defines a stratum in which a respondent must be selected (Hasler and Tillé, 2016).

[1] This chapter is essentially a reprint of: Jauslin, R., Eustache, E., & Tillé, Y., (2020). Enhanced Cube Implementation For Highly Stratified Population, *Japanese Journal of Statistics Data Science*, 4:783-795

The balanced and stratified sampling method of Chauvet (2009) has been improved by Hasler and Tillé (2014) to partially resolve the disadvantage of the time required to process a highly stratified population. When the sum of the inclusion probabilities in the strata is not an integer, the computation time can become problematic. This problem arises, for example, when the objective is to select less than one individual per household. Neither of the two methods already proposed solves the computational time problem in these situations.

In this paper, we propose a new method to obtain a stratified balanced sample. This new method is particularly interesting when the population is highly stratified and the inclusion probabilities do not sum to an integer within the strata. We refer readers to Tillé (2020) and Hankin et al. (2019) for more information on the general settings on stratified balanced sampling design.

The document is organized as follows. Section 4.2 gives the basic notations and settings. Section 4.3 presents the problem of selecting a balanced sample. In the section 4.4, we review the cube method and how it is used to select a balanced sample. In section 4.5, we discuss the issue of a highly stratified population and review the methods used to select a sample in this case. In the section 4.6, we present the new method and the section 4.7 is devoted to variance estimation. In the section 4.8, we give the simulation results of the different algorithms on an artificial dataset while the section 4.9 gives a conclusion on the new method.

4.2 Basic sampling notations

Consider a finite population U of size N whose units can be defined by labels $k \in \{1, 2, \ldots, N\}$. Let define a variable of interest y. Suppose that we are trying to estimate the following unknown total:

$$Y = \sum_{k \in U} y_k. \tag{4.1}$$

A sampling design is defined by the probability $p(s)$ of selecting each possible subset $s \subset U$ such that $\sum_{s \subset U} p(s) = 1$. Consider a vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^\top$ that maps elements of a subset s to an N vector of 0s and 1s such that:

$$\delta_k = \begin{cases} 1 & \text{if } k \in s, \\ 0 & \text{otherwise,} \end{cases}$$

for $k \in U$. For each unit of the population, the inclusion probability π_k, with $0 \leqslant \pi_k \leqslant 1$, is defined as the probability of selecting k into a sample s:

$$\pi_k = \mathrm{P}(k \in s) = \mathrm{E}(\delta_k) = \sum_{s \subset U | k \in s} p(s).$$

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ be the vector of all the inclusion probabilities. Let also $\pi_{k\ell}$ be the probability of selecting units k and ℓ together in the sample, with $\pi_{kk} = \pi_k$. Assuming that $\pi_k > 0$ for all $k \in U$, the total (4.1) can be estimated using the classical unbiased Horvitz-Thompson estimator defined by

$$\widehat{Y} = \sum_{k \in U} \frac{y_k \delta_k}{\pi_k}. \tag{4.2}$$

4.3 Stratified balanced sampling

Usually, some auxiliary information is available for each unit $k \in U$ in a vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kq})^\top \in \mathbb{R}^q$, with $q \in \mathbb{N}$. A sampling design is said to be balanced on the q auxiliary variables if and only if it satisfies the following balancing equation:

$$\widehat{\mathbf{X}} = \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \frac{\mathbf{x}_k \delta_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k = \mathbf{X}.$$

Sometimes, selecting a sample that exactly satisfies the constraints is not possible due to the rounding problem.

In many applications, inclusion probabilities are such that the selected sample has a fixed size. In order to obtain a sampling design with fixed sample size, a linear combination of the auxiliary variables must be proportional or equal to the vector of inclusion probabilities, i.e. there exists $\boldsymbol{\psi} \in \mathbb{R}^q$ such that $\boldsymbol{\psi}^\top \mathbf{x}_k = \pi_k$, for all $k \in U$. This gives

$$\sum_{k \in s} \frac{\boldsymbol{\psi}^\top \mathbf{x}_k}{\pi_k} = \sum_{k \in s} \frac{\pi_k}{\pi_k} = n.$$

The size of the sample will be fixed only if n is an integer. If it is not the case, the sample size will be equal to the higher or lower integer to n.

More generally, the problem of selecting a balanced sample is written as the following linear system :

$$\begin{cases} \mathbf{A}^\top \boldsymbol{\delta} = \mathbf{A}^\top \boldsymbol{\pi}, \\ \boldsymbol{\delta} \in \{0, 1\}^N, \end{cases} \tag{4.3}$$

where $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_N/\pi_N)^\top$. The aim consists then of obtaining a sample $\boldsymbol{\delta}$ that satisfies (or approximately satisfies) the constraints.

Suppose that the population U is divided into H strata $U_1, \ldots, U_H$, with respective sizes of $N_1, \ldots, N_H$. The strata form a partition and respect the following properties:

$$U = \bigcup_{h=1}^{H} U_h, \; N_h > 0, \; U_h \cap U_\ell = \emptyset, \text{ for all } h, \ell \in \{1, \ldots, H\}.$$

Then, this implies that $\sum_{h=1}^{H} N_h = N$. The inclusion probabilities sum to a value n_h in each stratum h, i.e. $n_h = \sum_{k \in U_h} \pi_k$. Let $\mathbf{h} = (h_1, \ldots, h_N)^\top$ be a categorical vector that specifies the stratum to which each unit belongs. For example, $h_k = \ell$ means that unit k belongs to strata U_ℓ, with $k \in U$ and $\ell \in \{1, \ldots, H\}$. Another way to express the stratum of each unit is to use the disjunctive form. Let $\mathbf{H}$ be the disjunctive matrix of the corresponding vector $\mathbf{h}$ of size $N \times H$, such that:

$$\mathbf{H} = \big(\mathbf{1}(U_1), \ldots, \mathbf{1}(U_H)\big),$$

where $\mathbf{1}(U_h) \in \mathbb{R}^N$ is a column vector such that its kth element is equal to 1 if the unit k belongs to the stratum U_h and 0 otherwise.

Obtaining a balanced sample in a stratified population is equivalent to adding stratification constraints to the previous linear system (4.3). These constraints are contained in the matrix $\mathbf{H}$, so the modification of the linear problem gives:

$$\begin{cases} (\mathbf{H}\, \mathbf{A})^\top \boldsymbol{\delta} = (\mathbf{H}\, \mathbf{A})^\top \boldsymbol{\pi}, \\ \boldsymbol{\delta} \in \{0, 1\}^N. \end{cases} \tag{4.4}$$

The number of constraints in the linear problem is then $(q + H)$. In the next section, a method to select a balanced sample is presented.

4.4 Cube Method

Deville and Tillé (2004) developed the cube method that selects a balanced sample respecting the inclusion probabilities. The method can deal with equal or unequal inclusion probabilities. The algorithm is separated into two phases.

- The first phase is called the flight phase. It modifies recursively and randomly the vector of inclusion probabilities $\boldsymbol{\pi}$ into a sample by respecting exactly the balancing constraints of the problem. The subspace induced by the linear system (4.3) could be rewritten using the following notation:

$$\mathcal{A} = \left\{\boldsymbol{\delta} \in \mathbb{R}^N | \mathbf{A}^\top \boldsymbol{\delta} = \mathbf{A}^\top \boldsymbol{\pi}\right\} = \boldsymbol{\pi} + \text{Null}(\mathbf{A}^\top),$$

  where $\text{Null}(\mathbf{A}^\top) = \left\{\mathbf{u} \in \mathbb{R}^N | \mathbf{A}^\top \mathbf{u} = 0\right\}$. The idea is then to use a vector $\mathbf{u}$ of the null space of $\mathbf{A}^\top$ in order to update randomly the vector $\boldsymbol{\pi}$. The whole procedure of the update can be found in Deville and Tillé (2004). At each step, at least one component is set to 0 or 1. Matrix $\mathbf{A}$ is updated with the new inclusion probabilities. This step is repeated until the null space of $\mathbf{A}^\top$ is empty. At the end of the flight phase, the final updated vector of $\boldsymbol{\pi}$ contains at most q elements that are still not equal to 0 or 1.

- The second phase is called the landing phase. This phase results in the sample $\boldsymbol{\delta}$ that respects as much as possible the balancing constraints. There are two different ways to achieve it, by relaxing the q constraints one by one, or by linear programming.

In the flight phase, the major computational cost comes from the research of a vector in the null space of $\mathbf{A}^\top$. Chauvet and Tillé (2006) have improved this time-consuming inconvenience using a submatrix of $\mathbf{A}$ rather than the entire matrix. The idea is to consider a submatrix that has one more row than the number of columns to ensure that it has at least one vector in its null space. This submatrix, denoted by $\mathbf{B}$, has then a size of $(q+1) \times q$, with respect to $q < N$ and $\text{Rank}(\mathbf{B}) \leqslant q$.

The interest of using this submatrix comes from the following result: a vector $\mathbf{u}$ of $\text{Null}(\mathbf{B}^\top)$ completed by $(N - (q+1))$ zeros is a vector of $\text{Null}(\mathbf{A}^\top)$. With this idea, all the computations can be done using only a submatrix $\mathbf{B}$. Usually, N is much larger than q, the size of $\mathbf{B}$ is then much smaller than $\mathbf{A}$. This implies obviously an important gain of computational time. The method proposed in this paper uses the same idea. In the next section, the particular case of highly stratified sampling is considered.

4.5 Highly stratified population

It is always preferable to consider a stratified population in order to estimate the total (4.1). The variance of the Horvitz-Thompson estimator (4.2) can be considerably reduced compared to the non-stratified estimator (4.1). However, when the population is highly stratified (i.e. H is very large), the selection of a balanced sample with classical methods becomes difficult due to the too large number of constraints in $\mathbf{H}$. In order to decrease the time-consuming problem, different approaches have already been proposed.

Chauvet (2009) has developed an algorithm to select a balanced sample in a highly stratified population. Firstly, a flight phase is applied inside each stratum. This allows

modification of the inclusion probabilities such that these are as balanced as possible in each stratum. Next, a flight phase is applied to the whole population. Finally, a landing phase is carried out on units that are not still selected or rejected. This procedure has the advantage of being simple to implement. Its major deficiency is when the number of strata H becomes too large. The procedure remains then very slow and often cannot even be used.

Hasler and Tillé (2014) have proposed another method to deal with highly stratified population. As the previous method, it begins by applying the flight phase of the cube method to each stratum of the population. Next, it carries out a flight phase on a union of strata by adding another stratum at each step. By doing this, strata are managed one after the other and the inclusion probabilities of certain strata are set to 0 or 1 during this step. The idea behind this procedure is to reduce the matrix $\mathbf{H}$ considered because some strata are removed from the matrix when all its units are selected or rejected. At the end, a landing phase is applied. However, if n_h is not equal to an integer for a stratum U_h, this method also remains very time-consuming. Indeed, some strata are never completely removed during the procedure and then the submatrix of $\mathbf{H}$ considered becomes too large.

The properties of the cube method imply that the inclusion probabilities are satisfied and that the sample is balanced on the auxiliary variables in these two methods. However, they still have difficulty in dealing with all the situations of highly stratified sampling. In the next following section, a new method is presented in order to completely resolve these drawbacks.

4.6 Proposed method

In the fast implementation of the cube method proposed by Chauvet and Tillé (2006), the main modification was to use a matrix smaller than $\mathbf{A}$ to update $\boldsymbol{\pi}$. This allows considerable reduction of the computational cost. The idea of our method is similar but adapted to a stratified population. It considers a matrix of constraints $\mathbf{B}$ smaller than $(\mathbf{H}\,\mathbf{A})$ during the use of the cube method. The submatrix $\mathbf{B}$ must be found at each step of the flight phase of the cube method. As explained in Section 4.3, the number of balancing constraints depends on the number of strata H when the population is stratified. Since all the strata must be considered, the number of balancing equations is equal to $(q+H)$. So, in the classical flight phase presented in Section 4.4, the considered matrix $\mathbf{B}$ is of size $(q+H+1)\times(q+H)$. The columns of $\mathbf{B}$ corresponding to strata which do not contain any unit in the rows of $\mathbf{B}$ are only composed of 0. To update the vector $\boldsymbol{\pi}$, we find a vector in the nullspace of $\mathbf{B}$. Since columns that are only equal to 0 are inefficient for this, they only increase the size of the matrix $\mathbf{B}$ and are irrelevant for finding a null vector.

The idea of the proposed method is then to compute a matrix $\mathbf{B}$, with still one more row than the number of columns, but that considers a smaller number of strata. By considering a matrix $\mathbf{B}$ with fewer rows, the corresponding vector of strata $\mathbf{h}$ will be reduced. This subvector of $\mathbf{h}$ will contain fewer categories and then the corresponding matrix $\mathbf{H}$ will have fewer columns. This is why obtaining the matrix $\mathbf{B}$ with exactly one row more than its number of columns is not as easy as with an unstratified population. Algorithm 3 shows how to find the number of rows to consider in order to obtain the smaller matrix $\mathbf{B}$ such that $\mathbf{B}$ has exactly one row more than its number of columns.

Algorithm 3: Find the submatrix $\mathbf{B}$ of $(\mathbf{H}\ \mathbf{A})$

Let q be the number of auxiliary variables of $\mathbf{A}$. Initialize q^1 by q. For $t = 1, 2, 3, \ldots$ repeat the following steps:

1. Extract the first q^t rows of the vector $\mathbf{h}$ and denote it $\mathbf{h}^t$.
2. Denote H^t the number of different strata in $\mathbf{h}^t$.
3. Update $q^{t+1} = q + H^t + 1$.

while $q^{t+1} > q^t$.
Finally, $\mathbf{B}$ is defined as the q^t first rows of the concatenated matrix $(\mathbf{H}^t\ \mathbf{A}^t)$, where $\mathbf{A}^t$ and $\mathbf{H}^t$ are the submatrix containing only its q^t first rows.

Example 4.6.1 *Suppose that $q = 2$ and that the categorical vector is equal to $\mathbf{h} = (1, 1, 2, 2, 3, 3, 3, 4, 4, \ldots)^\top$. We obtain*

$$
\begin{array}{llllll}
t = 1: & q^1 = 2, & \mathbf{h}^1 = (1,1)^\top, & H^1 = 1 & \rightarrow & q^2 = 2+1+1 = 4, \\
t = 2: & q^2 = 4, & \mathbf{h}^2 = (1,1,2,2)^\top, & H^2 = 2 & \rightarrow & q^3 = 2+2+1 = 5, \\
t = 3: & q^3 = 5, & \mathbf{h}^3 = (1,1,2,2,3)^\top, & H^3 = 3 & \rightarrow & q^4 = 2+3+1 = 6, \\
t = 4: & q^4 = 6, & \mathbf{h}^4 = (1,1,2,2,3,3)^\top, & H^4 = 3 & \rightarrow & q^5 = 2+3+1 = 6, \\
t = 5: & q^5 = q^4 & & & &
\end{array}
$$

$\mathbf{B}$ contains then $q^4 = 6$ rows and $5 = 2 + 3$ columns. So it is a matrix with only one more row than the number of columns as desired.

The matrix $\mathbf{B}$ is found after having computed its number of rows q^t using Algorithm 3. The first q^t elements of $\mathbf{h}$ composed the strata membership vector $\mathbf{h}^t$. The disjunctive matrix $\mathbf{H}^t$ can then be found using $\mathbf{h}^t$. The matrix $\mathbf{B}$ is equal to $(\mathbf{H}^t\ \mathbf{A}^t)$, with $\mathbf{A}^t$ the submatrix of $\mathbf{A}$ containing only its q^t first rows. The units must be ordered in such a way that the strata are clustered. If it is not the case, Algorithm 3 ends up with a matrix that could be large. Indeed, the larger the value $\mathbf{H}^t$ is, the larger the size of the matrix $\mathbf{B}$ will be. The same procedure proposed by Chauvet and Tillé (2006) can then be applied. If the population is highly stratified and the number of auxiliary variables is acceptable, our procedure can be very efficient. Moreover, it handles inclusion probabilities that do not sum to an integer inside strata. Algorithm 4 presents the whole method and is implemented in an R package (Jauslin et al., 2022a).

The Algorithm can be decomposed into two flight phases and two landing phases. The first step is to apply a flight phase within each stratum. This step modifies the inclusion probabilities so that they are balanced in each stratum. In the second step, we first check that there are no strata containing only one unit. In this case, the balancing equations of these strata cannot be balanced, so we remove these units from the treatment in order to manage them afterwards. Next, we apply the flight phase on the other units by using the matrix $\mathbf{B}$ found by the Algorithm 3. This second flight phase is repeated until the matrix $\mathbf{B}$ is no longer found or the null space is empty. Note that the dimension of the computed matrix $\mathbf{B}$ can change depending on the stratum vector $\mathbf{h}$ and the inclusion probabilities set to 0 or 1. At the end of this flight phase, we can have H units that are alone in their stratum and also have maximum q units for which the balancing equations could not be satisfied. It is then no longer possible to find a matrix $\mathbf{B}$ such that its kernel is not empty. Note that at the end of this step, the aux-

iliary variables are still completely satisfied, i.e. the equation (4.4) with the modified inclusion probabilities is perfectly satisfied.

The objective is to respect the strata sizes as much as possible. Therefore, the constraints on the strata are prioritized. We then perform a landing phase by suppression of variables on the units that are in the strata containing at least two units. At the end of this first landing phase, we have at most H units which are alone in their strata and thus the corresponding disjunctive matrix is a diagonal matrix. We can drop the variables corresponding to the strata and apply a landing phase only on the auxiliary variable to finally find the sample.

If the vector $\mathbf{h}$ is reordered in a different way before the procedure is started, the matrix $\mathbf{B}$ could change during the execution of the Algorithm 4. This has no impact on the computational cost if the $\mathbf{h}$ vector is still clustered. Figure 4.1 shows an example of what the matrix $\mathbf{B}$ looks like in the Chauvet and Hasler methods and the one computed by the Algorithm 3.

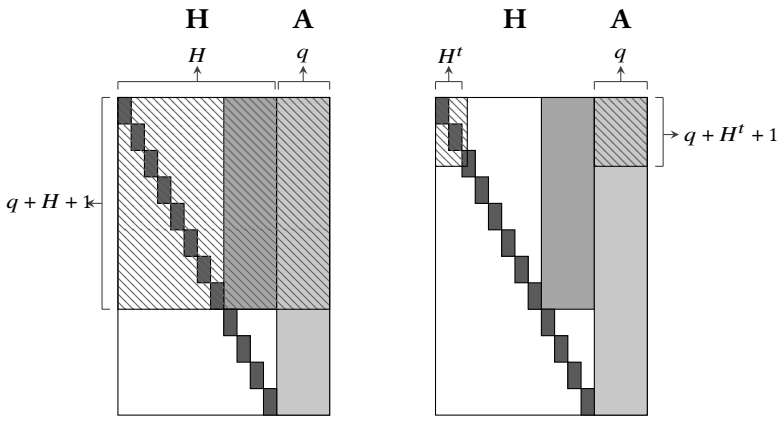


Figure 4.1 – *Illustration of how the* $\mathbf{B}$ *matrix is calculated within the flight phase in the discussed methods. The light grey area represent the auxiliary variable matrix* $\mathbf{A}$ *while the dark grey area is the* $\mathbf{H}$ *matrix. The* $\mathbf{B}$ *matrix is illustrated by the hatched area. The medium grey area is the one inside of the matrix* $\mathbf{B}$ *where all values are equal to zero. On the left, we see how the matrix* $\mathbf{B}$ *is computed in Chauvet and Hasler methods, while on the right, it is the matrix proposed by Algorithm 3. We see that the matrix* $\mathbf{B}$ *on the left contains some columns that are only equal to zero.*

4.7 Variance estimation

The variance can be approximated by using the method proposed by Deville and Tillé (2005). Let the vector

$$\mathbf{z}_k = (\mathbf{H}\ \mathbf{A})_k,$$

where $(\mathbf{H}\ \mathbf{A})_k$ denotes the kth row of the matrix $(\mathbf{H}\ \mathbf{A})$. The variance of the Horvitz-Thompson estimator of the total $\widehat{Y}$ can be approximated by

$$\mathrm{var}_{app}(\widehat{Y}) = \sum_{k \in U} c_k \left(\frac{y_k}{\pi_k} - \boldsymbol{\alpha}^\top \mathbf{z}_k \right)^2, \tag{4.5}$$

where

$$c_k = \pi_k(1-\pi_k)\frac{N}{N-(H+q)} \text{ and } \boldsymbol{\alpha} = \left(\sum_{\ell\in U} c_\ell \mathbf{z}_\ell \mathbf{z}_\ell^\top\right)^{-1} \sum_{\ell\in U} c_\ell \mathbf{z}_\ell \frac{y_\ell^\top}{\pi_\ell}.$$

There are many different ways to express the quantity c_k and then this leads to various approximations of the variance. Value c_k can in particular be approximated by

$$\widetilde{c}_k = (1-\pi_k)\frac{n}{n-(H+q)}.$$

We define the estimator of the approximated variance as the following equation:

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k\in s} \widetilde{c}_k \left(\frac{y_k}{\pi_k} - \widetilde{\boldsymbol{\alpha}}^\top \mathbf{z}_k\right)^2. \tag{4.6}$$

Note that the sum in the equation (4.6) is on the sample s using $\widetilde{c}_k$ and $\widetilde{\boldsymbol{\alpha}}$ instead of c_k and $\boldsymbol{\alpha}$.

4.8 Simulations

In this section, the performance of the method is evaluated on real data produced by the Swiss Federal Statistical Office (2020). The dataset contains information on Swiss establishments. We restrict the study to the Swiss region called Espace Mittelland (a region of the second degree of the Nomenclature of Territorial Units for Statistics (NUTS) of Switzerland). This region contains 5 cantons (a region of the third degree of the NUTS) and 675 municipalities. For confidentiality reasons, the units considered are the hectares of land in which at least one establishment is located. In order to be able to estimate the variance, only 3 hectares of land per municipality are included in the study. This implies that the dataset contains information from 2025 hectares including at least one establishment.

We stratify the units in two different ways: by cantons and by municipalities. The number of strata is then respectively equal to $H_c = 5$ and $H_m = 675$. Figure 4.2 shows the dataset with the two proposed stratifications. The idea behind this procedure is to compare the execution time for a stratified population with a low number of strata versus a high one. To compare the method, we will use balancing variables $\mathbf{x}_j$ containing the number of women employed in a sector j, with $j = 1, 2, 3$. Each sector represents a type of activity: sector $j = 1$ involves the natural environment and agriculture, sector $j = 2$ is manufacture and sector $j = 3$ is related to services. Table 4.1 shows the mean time execution of three methods for highly stratified sampling: the methods of Hasler and Tillé (2014) and Chauvet (2009), detailed in Section 4.5, and the new one presented in this article.

Inside each stratum, the inclusion probabilities are equal. For each stratification, we carry out two different types of sampling: one with inclusion probabilities that sum to an integer number within each stratum and one with a non-integer sum. Then, we consider $n_h = 80$ and $n_h = 80.4$, $h \in \{1, \ldots, H_c\}$, for the first stratification. For the second one, we take $n_h = 2$ and $n_h = 1.4$, $h \in \{1, \ldots, H_m\}$. We choose to deal also with non-integers n_h in order to compare the impact of this situation on the mean sampling time.

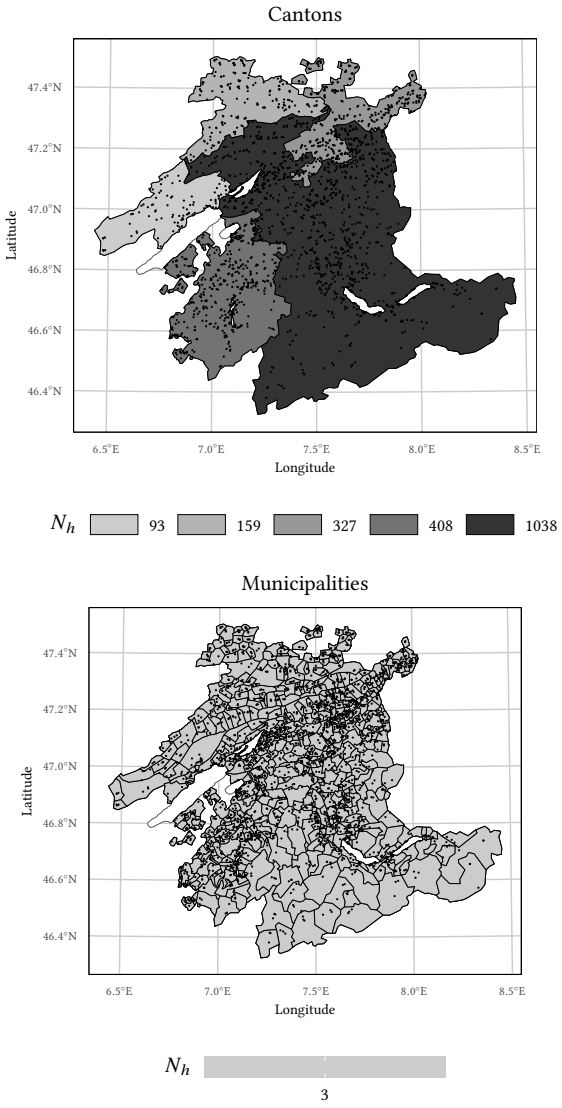


Figure 4.2 – *Data extracted from the Swiss establishments data base of the Swiss Federal Statistical Office (2020). The data are restricted to the NUTS region 2. The upper plot is showing the separation by Cantons $H_c = 5$, the lower one the separation by Municipalities $H_m = 675$. The grey gradient scale gives the number of units considered in each Canton. The data are selected such that each municipality contains 3 units.*

Chauvet's method cannot be compared because its execution time is too long and should be avoided for highly stratified population. However, it remains very efficient if the number of strata is acceptable. If inclusion probabilities in a stratum sum to an integer, Hasler's method performs very well. However, the execution time increases strongly when n_h is not an integer. The proposed method is very well-behaved and the execution time is considerably reduced for a highly stratified population.

In order to compare the variance of the method with the others, we estimate the variance using some variables of interest y_j that contain the total number of employees of the sector j, $j = 1, 2, 3$. In Table 4.2, we compare the approximated variance (4.5),

the estimated variance and the simulated variance computing using the equation:

$$v_{sim} = \frac{1}{m} \sum_{s} \{\widehat{Y}(s) - Y\}^2, \tag{4.7}$$

where m is the number of simulations.

For each method, we vary the number of selected units within each stratum by taking n_h equals to 2 for the stratification with H_c strata and 80 for the stratification with H_m strata. This implies samples of size equal 400 and 1350 respectively. The variance estimator seems to be unbiased for the approximated variance. However, we see that the approximated variance and estimator are slightly biased to the v_i. This comes from the landing phase of each method. We can conclude that the proposed method is comparable in terms of variance to other methods.

4.9 Conclusion

The stratified sampling procedure is a well-known and appropriate procedure to reduce the variance of the Horvitz-Thompson estimator. In this paper, we propose a new method and implementation that provide an excellent executing time and a flexibility that the existing methods did not allow. In many surveys where the population is stratified, the sum of inclusion probabilities within each stratum is not an integer. Other methods are not directly applicable in this case. We have shown by means of simulations that the variance of the estimator is not impacted by our method. All of these results indicate that our proposed algorithm is very efficient for selecting a sample in a stratified and highly stratified population.

Algorithm 4: Complete proposed algorithm for highly stratified population

Consider $\boldsymbol{\pi}$ the N vector of inclusion probabilities such that $0 < \pi_k < 1$, for $k \in \{1, \dots, N\}$.

I. Perform a flight phase on each stratum according to the inclusion probabilities $\boldsymbol{\pi}$ and the balancing constraints in $\mathbf{A}^\top$. The vector $\boldsymbol{\pi}$ is updated by $\boldsymbol{\pi}^1$ such that some of its elements are set to 0 or 1. Compute the set of indices $\mathbf{i}^1 \subset \{1, \dots, N\}$ containing the unit indices that have an inclusion probability still not equal to 0 or 1 and that are not alone in their strata. Define the set of units that are alone in their strata and denote this set $\widetilde{\mathbf{i}^t}$.

II. Initialize t by 1. Repeat step 1. to 7. until it is no more possible to find the matrix $\mathbf{B}$ or until the vector $\mathbf{u}$ is null.

1. Update the set $\widetilde{\mathbf{i}^t}$ by adding the indices of the units that are contained in only one strata and remove it from $\mathbf{i}^t$.
2. In $\mathbf{A}$, $\mathbf{h}$ and $\boldsymbol{\pi}$, consider only units with indices in $\mathbf{i}^t$.
3. Apply the Algorithm 3 to find the submatrix $\mathbf{B}$ of $(\mathbf{H}\ \mathbf{A})$.
4. Compute $\mathbf{u}$, a vector of the null space of $\mathbf{B}$ completed by 0s to obtain a vector with the same size as $\mathbf{i}^t$.
5. Compute $\lambda_1 > 0$ and $\lambda_2 > 0$, the two larger values such that

$$\begin{array}{rcl} 0 & \leqslant & \pi_k^t + \lambda_1 u_k \leqslant 1 \\ 0 & \leqslant & \pi_k^t - \lambda_2 u_k \leqslant 1 \end{array}, \text{ for all } k \in \mathbf{i}^t.$$

6. Update $\boldsymbol{\pi}^t$ by:

$$\boldsymbol{\pi}^{t+1} = \begin{cases} \boldsymbol{\pi}^t + \lambda_1 \mathbf{u} & \text{with probability} \quad \lambda_2/(\lambda_1 + \lambda_2), \\ \boldsymbol{\pi}^t - \lambda_2 \mathbf{u} & \text{with probability} \quad \lambda_1/(\lambda_1 + \lambda_2). \end{cases}$$

7. Update t by $t + 1$ and update $\mathbf{i}^t$ the set of indices containing the unit that have an inclusion probability still not equal to 0 or 1.

III. Pool the set $\widetilde{\mathbf{i}^t}$ and the remaining units $\mathbf{i}^t$ and do a landing phase until we have only one unit alone in their strata. This step is done by suppression of variables on the balancing matrix $(\mathbf{H}\ \mathbf{A})$, but only dropping the variables that are in $\mathbf{A}$.

IV. Perform a landing phase by suppression of variables on the balancing variables $\mathbf{A}$ on the remaining units.

Table 4.1 – *Results of 1000 simulations on the Swiss establishments dataset. The population size is equal to 2025. We compute the mean time execution in seconds of each sampling procedure. We vary the number of strata H and the number of units selected within each stratum n_h.*

	Algorithm		
	Proposed method	Hasler's method	Chauvet's method
Cantons ($H = 5$)			
$n_h = 80$	0.24	0.25	0.24
$n_h = 80.4$	0.24	0.24	0.24
Municipalities ($H = 675$)			
$n_h = 2$	0.42	0.4	418.07
$n_h = 1.4$	0.53	400.55	701.77

Table 4.2 – *Results of 1000 simulations on a population of size 2025. The number of strata is equal to 5 for Cantons and 675 for Municipalities. For each variable of interest y_j, $j = 1, 2, 3$ and for each sampling methods, we compute the ratio between the different estimators (i.e. approximated variance (4.5) as well as the variance estimator (4.6)) and the variance approximated by the simulations (4.7).*

	Algorithm								
	Proposed method			Hasler's method			Chauvet's method		
	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$
Cantons ($H = 5$)									
y_1	1	0.945	0.959	1	0.994	1.019	1	1.064	1.08
y_2	1	0.905	0.969	1	0.932	1.012	1	0.927	0.983
y_3	1	0.891	0.927	1	0.864	0.911	1	0.84	0.879
Municipalities ($H = 675$)									
y_1	1	1.042	1.049	1	0.952	0.957	1	0.978	0.981
y_2	1	1.024	1.028	1	1.035	1.036	1	0.951	0.954
y_3	1	0.995	0.994	1	0.933	0.937	1	0.849	0.853

Chapter 5

An Efficient Approach for Statistical Matching of Survey Data Through Calibration, Optimal Transport and Balanced Sampling

Abstract

Statistical matching aims to integrate two statistical sources. These sources can be two samples or a sample and the entire population. If two samples have been selected from the same population and information has been collected on different variables of interest, then it is interesting to match the two surveys to analyse, for example, contingency tables or covariances. In this paper, we propose an efficient method for matching two samples that may each contain a weighting scheme. The method matches the records from the two sources. Several variants are proposed in order to create a directly usable file integrating data from both information sources. [1]

Keywords: auxiliary information, balanced sampling, data integration, distance, unequal probability sampling

5.1 Introduction

Integrating data from different sources represents a major challenge in statistics. Yang and Kim (2020) and Kim and Tam (2021) discussed a set of methods for data integration. Statistical matching is the field of statistics that deals with the best way to merge two different files by matching units based on a group of common variables (D'Orazio et al., 2006; D'Orazio, 2019).

Renssen (1998) distinguishes between two types of analysis. The macro approach focuses on estimating a full joint distribution between the respective variable of interest in the two samples. This can be, for example, a covariance matrix, a correlation or a contingency table. The objective of the micro approach is to complement one file with information from the other, imputation to correct non-response is an example related to this approach (Haziza, 2009; Chen and Haziza, 2019). Kim et al. (2016) use the technique of fractional imputation to perform statistical matching.

In this paper, we propose an efficient method for statistical matching. Units are matched based on the proximity of a group of variables measured in both surveys.

[1] This chapter is essentially a reprint of: Jauslin, R., & Tillé, Y., (2023). An Efficient Approach for Statistical Matching of Survey Data Through Calibration, Optimal Transport and Balanced Sampling, *Journal of Statistical Planning and Inference*, 225:121–131

Moreover, both sources can either have common units or have an empty intersection. One of the two sources may even contain the entire population. In addition, we impose a set of constraints in order to take advantage of all the information available in the two sources. This method can also be used for imputations and so can also be used for micro approach analyses.

Both sources of information may contain a weighting system that allows the files to be extrapolated to the full population. These weights are usually calculated to take into account inclusion probabilities, non-response treatment and calibration. In official statistics, the calibration methods have been proposed by Deville and Särndal (1992) and Deville et al. (1993) to adjust survey data on census data or a register. Calibration can also be used to adjust or harmonize several surveys from different sources (see Guandalini and Tillé, 2017, and references therein). Dudoignon (2018) through an unpublished paper, proposed using the idea of optimal transport on internet audience data. More recently, Garès et al. (2020); Garès and Omer (2022) have proposed optimal transport for merging dataset by considering joint distribution of covariates and outcomes.

We have set out a series of recommendations that a matching method should follow: The method should match common units as a priority. The result of the matching must integrate information from both sources. The matching should also take into account the weighting system. After matching, the estimated totals of the variables common to both sources must be identical to the totals before matching. Optimal matching should take advantage of all the information available in both sources.

The proposed methods therefore allow the matching of two data files but also the imputation of one file on another. First, calibration theory is used to harmonize the two samples. Then a linear program is used to perform an optimal matching while taking into account the weights. This program can be written as an optimal transport problem. Finally, the values to be matched can be selected using a balanced stratified sampling technique as presented in Jauslin et al. (2021) and implemented in the `R` package 'StratifiedSampling' (Jauslin et al., 2022a). The methods either perform matching based on a division of weights, produce a prediction, or impute a value from one source to another.

5.2 Problem and notation

Consider a population $U = \{1, \ldots, k, \ldots, N\}$ of size N from which two random samples S_1 and S_2 of size respectively equals to n_1 and n_2 have been selected. It is assumed that three groups of variables can be measured on the population units. The vectors of variables $\mathbf{x}_k \in \mathbb{R}^p, k \in U$ are measured on both units selected in S_1 and S_2. The vectors of variables $\mathbf{y}_k \in \mathbb{R}^q, k \in U$ are measured only on the units selected in S_1. The vectors of variables $\mathbf{z}_k \in \mathbb{R}^r, k \in U$ are measured only on the units selected in S_2.

The variables $\mathbf{x}_k, k \in U$ are called the matching variables. These variables are supposed to be known for both samples S_1 and S_2 and could be discrete, such as sex or citizenship, or continuous, for example, income or age. We also supposed that if some units are common between sample S_1 and sample S_2 it is because they shared the same values on these matching variables. The set of variables $\mathbf{y}_k, k \in S_1$ and $\mathbf{z}_\ell \in S_2$ are called the variables of interest. Figure 5.1 illustrates the matching problem. Hatched areas are the unknown quantities of the two samples. A more general question occurs on the accuracy of a specific conclusion made on the statistical match compared to

the population $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k), k \in U$. Supposing empty intersection between samples, as the variables $\mathbf{y}_k, k \in S_1$ are only observed in S_1 and variables $\mathbf{z}_k \in S_2$ only in S_2, it might be the case that the model is not identifiable. Some model assumption must be made on the population $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k), k \in U$ to ensure that a specific conclusion can be made on the statistical match. Conditionally to the variables $\mathbf{x}_k$, the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ must be independent. This assumption is called the conditional independence assumption. D'Orazio et al. (2006) discuss this assumption in more detail. Throughout this manuscript we suppose that we are in the conditional independence assumption context.

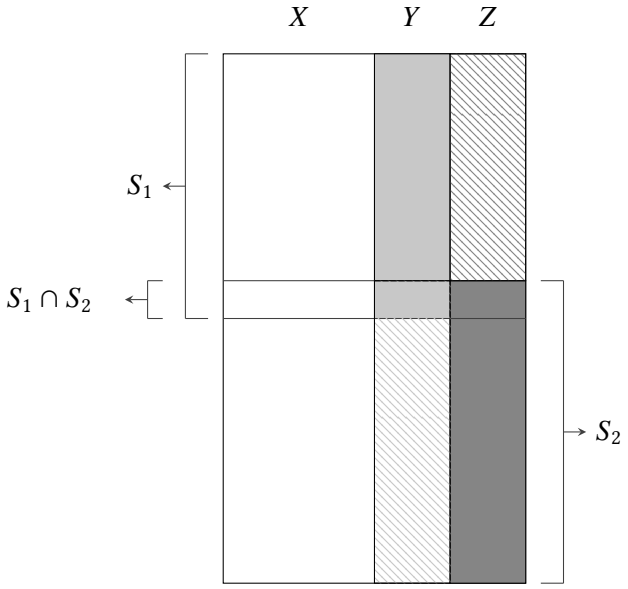


Figure 5.1 – *Representation of the statistical matching with intersection. Hatched areas are unknown quantities in each sample.*

In survey sampling, samples are generally designed with complex weighting systems $v_{1k}, k \in S_1$ and $v_{2\ell}, \ell \in S_2$. These weights can take into account the inverses of the inclusion probabilities, a possible re-weighting to compensate questionnaire non-response and a possible calibration. Frequently, a statistical match is made when one of the two files is seen as the recipient file while the other one is seen as the donor file. Throughout this manuscript, we will suppose, without loss of generality, that the sample S_1 is the recipient file while S_2 is the donor file. Generally we have $n_2 > n_1$, note that one of the two samples might be the whole population.

The population totals on the common auxiliary variables are equal to:

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k .$$

They be estimated either by sample S_1 or sample S_2, which are mathematically written:

$$\widehat{\mathbf{X}}_{v1} = \sum_{k \in S_1} v_{1k} \mathbf{x}_k, \quad \widehat{\mathbf{X}}_{v2} = \sum_{\ell \in S_2} v_{2\ell} \mathbf{x}_\ell . \tag{5.1}$$

Following the same idea on the variables of interest, the totals

$$\mathbf{Y} = \sum_{k \in U} \mathbf{y}_k \text{ and } \mathbf{Z} = \sum_{k \in U} \mathbf{z}_k,$$

can be estimated using the following estimators:

$$\widehat{\mathbf{Y}}_{v1} = \sum_{k \in S_1} v_{1k} \mathbf{y}_k, \quad \widehat{\mathbf{Z}}_{v2} = \sum_{\ell \in S_2} v_{2\ell} \mathbf{z}_\ell.$$

In the micro approach, the two samples S_1 and S_2 are merged into a single usable file, while the macro approach focuses on the joint distribution of the variables of interest. Under the usual hypothesis that the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ are independent conditionally to the variables $\mathbf{x}_k$, the relationships between the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ can be analysed. For example, if the variables $\mathbf{y}_k$ and $\mathbf{z}_k$ are dummy variables with respectively q and r variables, we could estimate the contingency table

$$\mathbf{N}_{yz} = \sum_{k \in U} \mathbf{y}_k \mathbf{z}_k^\top.$$

If the variables of interest are continuous, we could compute the covariance matrix of the totals

$$\mathbf{\Sigma}_{yz} = \text{cov}(\mathbf{Y}, \mathbf{Z}).$$

5.3 Harmonization by calibration

The two samples are assumed to have their own sampling weights: $v_{1k}, k \in S_1$ and $v_{2\ell}, \ell \in S_2$. To perform the matching, we need to harmonize the weights so that, they sum to the same values and respect the totals given in Equations (5.1). We then look for a new weighting system

$$w_{1k}, k \in S_1 \text{ and } w_{2\ell}, \ell \in S_2, \tag{5.2}$$

such that we have the following results:

$$\widehat{\mathbf{X}}_{w1} = \sum_{k \in S_1} w_{1k} \mathbf{x}_k = \widehat{\mathbf{X}}_{w2} = \sum_{\ell \in S_2} w_{2\ell} \mathbf{x}_\ell.$$

One aspect that must be considered is the intersection between S_1 and S_2. Several cases can occur, the sample S_1 can be included in S_2 or vice versa, the intersection can also be empty. Let $n_{12} = \#(S_1 \cap S_2)$ denote the size of the intersection of the two samples. Guandalini and Tillé (2017) analyse estimators of the form $\widehat{\mathbf{X}}_\alpha = \alpha \widehat{\mathbf{X}}_{v1} + (1 - \alpha)\widehat{\mathbf{X}}_{v2}$. They showed that, when $p = 1$, to best estimate $\mathbf{X}$ using both S_1 and S_2, the value of α must be equal to

$$\alpha^{\text{opt}} = \frac{\text{var}(\widehat{\mathbf{X}}_{v2}) - \text{cov}(\widehat{\mathbf{X}}_{v1}, \widehat{\mathbf{X}}_{v2})}{\text{var}(\widehat{\mathbf{X}}_{v1}) + \text{var}(\widehat{\mathbf{X}}_{v2}) - \text{cov}(\widehat{\mathbf{X}}_{v1}, \widehat{\mathbf{X}}_{v2})}.$$

This optimal value minimizes the variance of $\widehat{\mathbf{X}}_\alpha$. However, it depends on unknown variances and on a covariance that must be estimated. Since variance estimators are particularly unstable, we may find ourselves far from the optimal estimator. Guandalini and Tillé (2017) suggest using a proxy value for α^{opt} that only depends on the sample sizes and the size of the overlapped sample:

$$\alpha^* = \frac{n_1 - n_{12}}{n_1 + n_2 - 2\, n_{12}}. \tag{5.3}$$

We can then construct the estimator $\widehat{\mathbf{X}}^* = \alpha^*\widehat{\mathbf{X}}_{v1}+(1-\alpha^*)\widehat{\mathbf{X}}_{v2}$. In particular, if $S_2 \subset S_1$, then $\alpha^* = 1$ and $\widehat{\mathbf{X}}^* = \widehat{\mathbf{X}}_{v1}$. Moreover, if $S_1 \cap S_2 = \emptyset$, then $\alpha^* = n_1/(n_1+n_2)$.

In order to compute the two new weighting systems (5.2) close to $v_{1k}, k \in S_1$ and $v_{2\ell}, \ell \in S_2$, the two samples are calibrated $\widehat{\mathbf{X}}^*$. If $G_k(w_{1k}, v_{1k})$ is one of the pseudo-distance defined in Deville and Särndal (1992), we can search the weighting systems that solve the following problem:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k\in S_1} G_k(w_{1k}, v_{1k}) \text{ and } \sum_{\ell\in S_2} G_\ell(w_{2\ell}, v_{2\ell}) \\ \text{subject to} & \displaystyle\sum_{k\in S_1} w_{1k}\mathbf{x}_k = \sum_{\ell\in S_2} w_{2\ell}\mathbf{x}_\ell = \widehat{\mathbf{X}}^*, \\ & \displaystyle\widehat{N}^* = \sum_{k\in S_1} w_{1k} = \sum_{\ell\in S_2} w_{2\ell} = \alpha^* \sum_{k\in S_1} v_{1k} + (1-\alpha^*)\sum_{\ell\in S_2} v_{2\ell}. \\ & w_{1k}, w_{2\ell} \geqslant 0 \text{ for all } k \in S_1, \ell \in S_2 \end{cases}$$

To verify the constraint of positivity of the weights, we can, for example, use the Kullback-Leibler divergence as pseudo-distance, i.e. $G_k(w_{1k}, v_{1k}) = w_{1k}\log w_{1k}/v_{1k}$. Thus, the new weights obtained have the same sum:

$$\sum_{k\in S_1} w_{1k} = \sum_{\ell\in S_2} w_{2\ell}.$$

The new system weights define estimators for $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$.

$$\widehat{\mathbf{X}}_{w1} = \sum_{k\in S_1} w_{1k}\mathbf{x}_k = \widehat{\mathbf{X}}_{w2} = \sum_{\ell\in S_2} w_{2\ell}\mathbf{x}_\ell,$$

$$\widehat{\mathbf{Y}}_{w1} = \sum_{k\in S_1} w_{1k}\mathbf{y}_k \text{ and } \widehat{\mathbf{Z}}_{w2} = \sum_{\ell\in S_2} w_{2\ell}\mathbf{z}_\ell.$$

The samples S_1 and S_2 are now linked by the matching variables that are both equal to the same totals and they sum to the same value. The latter property is essential for the optimal transport that follow in Section 5.5.

5.4 Renssen's methods

A method for estimating contingency table was developed in Renssen (1998) and more recently presented in D'Orazio et al. (2006). The general idea consists of harmonizing the weighting systems as explained in the Section 5.3 and then uses the matching variables $\mathbf{x}_k$ to create linear models to get an estimated contingency table. At the first step, regression coefficients $\boldsymbol{\beta}_{yx}$ and $\boldsymbol{\beta}_{zx}$ are computed from the samples S_1 and S_2, respectively, by using the weights, w_{1k}, $k \in S_1$ and $w_{2\ell}$, $\ell \in S_2$, respectively. Using a weighted linear model, the following coefficients are obtained:

$$\widehat{\boldsymbol{\beta}}_{yx} = \left(\sum_{k\in S_1} w_{1k}\mathbf{y}_k\mathbf{x}_k^\top\right)\left(\sum_{k\in S_1} w_{1k}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1},$$

$$\widehat{\boldsymbol{\beta}}_{zx} = \left(\sum_{\ell\in S_2} w_{2\ell}\mathbf{z}_\ell\mathbf{x}_\ell^\top\right)\left(\sum_{\ell\in S_2} w_{2\ell}\mathbf{x}_\ell\mathbf{x}_\ell^\top\right)^{-1}.$$

The contingency table is then estimated using the matrix product:

$$\widehat{\mathbf{N}}_{yz}^{ren} = \widehat{\boldsymbol{\beta}}_{yx}\left(\alpha^* \sum_{k\in S_1} w_{1k}\mathbf{x}_k\mathbf{x}_k^\top + (1-\alpha^*)\sum_{\ell\in S_2} w_{2\ell}\mathbf{x}_\ell\mathbf{x}_\ell^\top\right)\widehat{\boldsymbol{\beta}}_{zx}^\top,$$

where α^* is the coefficient (5.3) that depends on the value n_{12}. Renssen's method can be easily generalized to a continuous case, but some assumptions must be satisfied on variables $\mathbf{y}_k$, $k \in S_1$ and $\mathbf{z}_\ell$, $\ell \in S_2$. For more information, we refer the reader to the article by Renssen (1998) and the book by D'Orazio et al. (2006).

5.5 Matching by optimal transport

The main idea of our method uses the optimal transport to perform a statistical matching. Optimal transport is an old mathematical problem that consists of finding the best solution to minimize the cost of transporting some quantities of goods from a given set of locations to a given set of destinations. In its simple case, the optimal transport problem can be solved with a linear program. The optimal transport problem can be written as the minimization of a cost function (the output of goods in factories must be sent to the nearest sellers), subject to equality constraints (all goods produced by a factory must be sent to sellers). This corresponds exactly to the linear program method. Optimal transport has been a very prolific topic in statistics for the past 10 years and it is strongly related to the notion of Wasserstein distance. We refer the reader to Panaretos and Zemel (2020) for further reading in the field of optimal transport and Wasserstein distance.

In the particular case of statistical matching, we use the optimal transport problem to match the units of two different samples. Let S_1 and S_2 denote the recipient and donor samples, respectively. The idea is then to see which units of S_2 can be associated with a particular unit $k \in S_1$. We start by computing an $n_1 \times n_2$ matrix $\mathbf{D}$ containing the distances between the units of S_1 and the units of S_2. We can for example use the usual Euclidean distance or a Mahalanobis distance defined as follows:

$$d^2(k,\ell) = (\mathbf{x}_k - \mathbf{x}_\ell)^\top \widehat{\boldsymbol{\Omega}}_{xx}^{-1}(\mathbf{x}_k - \mathbf{x}_\ell),$$

where

$$\widehat{\boldsymbol{\Omega}}_{xx} = \frac{1}{\widehat{N}^*}\left\{\alpha^* \sum_{k\in S_1} w_{1k}(\mathbf{x}_k - \widehat{\overline{\mathbf{X}}})(\mathbf{x}_k - \widehat{\overline{\mathbf{X}}})^\top + (1-\alpha^*)\sum_{\ell\in S_2} w_{2\ell}(\mathbf{x}_\ell - \widehat{\overline{\mathbf{X}}})(\mathbf{x}_\ell - \widehat{\overline{\mathbf{X}}})^\top\right\},$$

and

$$\widehat{\overline{\mathbf{X}}} = \frac{\widehat{\mathbf{X}}^*}{\widehat{N}^*}.$$

Note that the choice of the distance function can have a significant impact on the optimal transport result. Depending on the type of matching variables, finding a distance that makes sense can be a complex task. For example, if the matching variables are only continuous variables, the Euclidean distance or the Mahalanobis distance is appropriate. However, if the matching variables are discrete or mixed discrete/continuous, some distance functions such as the Hamming distance can be used. Garès and Omer (2022) discuss the choice of distance functions in more detail. Then,

we search for weights $W_{k\ell}$ for each couple $k \in S_1, \ell \in S_2$. To do this, we solve the following linear program:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\ d(k,\ell) \\ \text{subject to} & \displaystyle\sum_{k\in S_1} W_{k\ell} = w_{2\ell} \text{ for all } \ell \in S_2, \\ & \displaystyle\sum_{\ell\in S_2} W_{k\ell} = w_{1k} \text{ for all } k \in S_1, \\ & W_{k\ell} \geqslant 0, \text{ for all } k \in S_1, \text{ and } \ell \in S_2, \end{cases}$$

where $W_{kk} = \min(w_{1k}, w_{2k})$, for all couples of identical units in S_1 and S_2. These constraints force the matching of identical units that can be selected from both samples. This linear program is nothing more than an optimal transport problem for which many efficient implementations exist.

Most of the $W_{k\ell}$ weights are zero. It is therefore not necessary to manipulate a large matrix of data. The completed calibration is not adversely affected in the linear program. Thus, we have

$$\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{x}_k = \sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{x}_\ell = \widehat{\mathbf{X}}^*.$$

The output of the linear program ends with a matrix of weights $\mathbf{W}$ of size $(n_1 \times n_2)$. The non-zero entries in the ith rows of the matrix $\mathbf{W}$ contain the corresponding weights of the matched units in the sample S_2. Figure 5.2 shows an illustration of the statistical matching idea using optimal transport and what it is the output of the linear program. We generally do not have a one-to-one match, which means that for each unit k in S_1 we have more than one unit with weights not equal to 0 in S_2. The next two sections propose two different ways to obtain, from the output of the optimal transport, a file where each unit from S_1 has only one imputed unit from S_2. Without loss of generality, in the following development, we suppose that sample S_1 is completed by realizing a prediction from S_2.

5.5.1 Matching by using prediction

We can do a prediction by computing the weighted averages of the $\mathbf{x}_\ell$ and $\mathbf{z}_\ell$ of S_2. Formally, this gives the following quantity to compute:

$$q_{k\ell} = \frac{W_{k\ell}}{\sum_{\ell\in S_2} W_{k\ell}} = \frac{W_{k\ell}}{w_{1k}}, \text{ for all } k \in S_1, \ell \in S_2.$$

By using these new weights, we can then compute a prediction of the $\mathbf{x}_k$ and the $\mathbf{z}_k$ $k \in S_1$,

$$\widehat{\mathbf{x}}_k = \sum_{\ell\in S_2} q_{k\ell}\mathbf{x}_\ell \text{ and } \widehat{\mathbf{z}}_k = \sum_{\ell\in S_2} q_{k\ell}\mathbf{z}_\ell, \text{ for all } k \in S_1.$$

Note that $\widehat{\mathbf{x}}_k$ is not equal to the observed matching variables $\mathbf{x}_k$. The matching quality can be evaluated by comparing the $\mathbf{x}_k$ with the predictions $\widehat{\mathbf{x}}_k$. For the predicted values $\widehat{\mathbf{x}}_k$, the calibration is always valid. Indeed that

$$\sum_{k\in S_1} w_{1k}\widehat{\mathbf{x}}_k = \sum_{k\in S_1} w_{1k}\mathbf{x}_k = \widehat{\mathbf{X}}^*.$$

However, the interest of the procedure is that we now have predicted values $\widehat{\mathbf{z}}_k$ for each unit of S_1 whereas these variables were only measured on S_2.

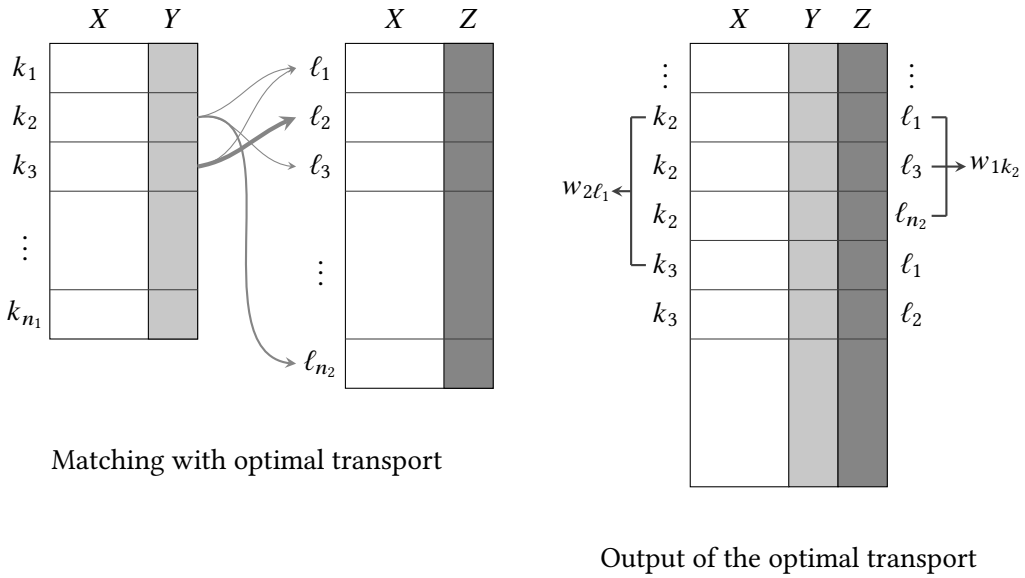


Figure 5.2 – *Outline of the statistical matching using optimal transport.*

5.5.2 Matching by using stratified balanced sampling

As explained in Section 5.5, the output of the optimal transport generally has repetition of some units. In some cases, we are interested in obtaining a synthetic file without any repetition of units. In this section, we propose an imputation method based on the optimal transport result. We propose to use a balanced sampling method to select from the repetition stratum of a unit k of S_1, only one unit of the units of S_2. This is in line with the hypothesis that S_2 is the donor file while S_1 is the recipient file. To do this, we randomly generate a matrix of Bernoulli random variables $a_{k\ell}, k \in S_1, \ell \in S_2$, where $a_{k\ell}$ is 1 if unit $\ell \in S_2$ is imputed to unit $k \in S_1$. Since each unit k of S_1 can only receive one imputation, we must have

$$\sum_{\ell \in S_2} a_{k\ell} = 1, \text{ for all } k \in S_1.$$

We now want to generate the random matrix of $a_{k\ell}$ with expectations $\mathrm{E}(a_{k\ell}) = q_{k\ell}$ in such a way that the following system of equations is satisfied at best

$$\sum_{k \in S_1} \sum_{\ell \in S_2} \frac{a_{k\ell} W_{k\ell}}{q_{k\ell}} \mathbf{x}_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} \mathbf{x}_\ell = \widehat{\mathbf{X}}^*,$$

$$\sum_{k \in S_1} \sum_{\ell \in S_2} \frac{a_{k\ell} W_{k\ell}}{q_{k\ell}} \mathbf{z}_\ell \approx \sum_{k \in S_1} \sum_{\ell \in S_2} W_{k\ell} \mathbf{z}_\ell$$

and with

$$\sum_{\ell \in S_2} a_{k\ell} = 1, \text{ for all } k \in S_1.$$

This sampling problem is known as 'stratified balanced sampling' (see Jauslin et al., 2021; Hasler and Tillé, 2014; Jauslin et al., 2022a). Each unit k of S_1 can be seen as a stratum for which a unit ℓ of S_2 must be selected.

The imputed values are then

$$\mathring{\mathbf{x}}_k = \sum_{\ell \in S_2} a_{k\ell}\mathbf{x}_\ell \text{ and } \mathring{\mathbf{z}}_k = \sum_{\ell \in S_2} a_{k\ell}\mathbf{z}_\ell, \text{ for all } k \in S_1.$$

Again, we have

$$\sum_{k \in S_1} w_{1k}\mathring{\mathbf{x}}_k \approx \sum_{k \in S_1} w_{1k}\mathbf{x}_k = \widehat{\mathbf{X}}^*.$$

However, the interest of the procedure is that we now have values $\mathring{\mathbf{z}}_k$ for each unit of S_1 whereas these variables were only measured on S_2.

If $\mathrm{E}_q(.)$ is the expectation to the $a_{k\ell}$ conditionally to S_1 and S_2, then, for all $k \in S_1$, we have,

$$\mathrm{E}_q(\mathring{\mathbf{x}}_k) = \sum_{\ell \in S_2} \mathrm{E}_q(a_{k\ell})\mathbf{x}_\ell = \sum_{\ell \in S_2} q_{k\ell}\mathbf{x}_\ell = \widehat{\mathbf{x}}_k,$$

and

$$\mathrm{E}_q(\mathring{\mathbf{z}}_k) = \sum_{\ell \in S_2} \mathrm{E}_q(a_{k\ell})\mathbf{z}_\ell = \sum_{\ell \in S_2} q_{k\ell}\mathbf{z}_\ell = \widehat{\mathbf{z}}_k.$$

5.6 Analysis of the data

Once the optimal transport is performed, different analyses are possible. We can either work directly with the optimal transport result or use the prediction or the imputations methods. This gives us five possibilities to analyse the data.

1. Use the full results of the optimal transport problem ($\mathbf{x}_k, \mathbf{x}_\ell, \mathbf{y}_k, \mathbf{z}_\ell, W_{k\ell}, k \in S_1, \ell \in S_2$).
2. Use the predicted values ($\mathbf{x}_k, \widehat{\mathbf{x}}_k, \mathbf{y}_k, \widehat{\mathbf{z}}_k, w_{1k}, k \in S_1$) by predicting the values of $k \in S_1$.
3. Use the imputed values ($\mathbf{x}_k, \mathring{\mathbf{x}}_k, \mathbf{y}_k, \mathring{\mathbf{z}}_k, w_{1k}, k \in S_1$) by imputing the values of $k \in S_1$.
4. Use the predicted values ($\mathbf{x}_\ell, \widehat{\mathbf{x}}_\ell, \widehat{\mathbf{y}}_\ell, \mathbf{z}_\ell, w_{2\ell}, \ell \in S_2$) by predicting the values of $\ell \in S_2$.
5. Use the imputed values ($\mathbf{x}_\ell, \mathring{\mathbf{x}}_\ell, \mathring{\mathbf{y}}_\ell, \mathbf{z}_\ell, w_{2\ell}, \ell \in S_2$) by imputing the values of $\ell \in S_2$.

The estimations of the means are completely consistent for the five possibilities. We obtain

$$\widehat{\overline{\mathbf{Z}}} = \frac{\sum_{k \in S_1}\sum_{\ell \in S_2} W_{k\ell}\mathbf{z}_\ell}{\sum_{k \in S_1}\sum_{\ell \in S_2} W_{k\ell}} = \frac{\sum_{\ell \in S_2} w_{2\ell}\mathbf{z}_\ell}{\sum_{\ell \in S_2} w_{2\ell}} = \frac{\sum_{k \in S_1} w_{1k}\widehat{\mathbf{z}}_k}{\sum_{k \in S_1} w_{1k}} \approx \frac{\sum_{k \in S_1} w_{1k}\mathring{\mathbf{z}}_k}{\sum_{k \in S_1} w_{1k}},$$

and

$$\widehat{\overline{\mathbf{Y}}} = \frac{\sum_{k \in S_1}\sum_{\ell \in S_2} W_{k\ell}\mathbf{y}_k}{\sum_{k \in S_1}\sum_{\ell \in S_2} W_{k\ell}} = \frac{\sum_{k \in S_1} w_{1k}\mathbf{y}_k}{\sum_{k \in S_1} w_{1k}} = \frac{\sum_{\ell \in S_2} w_{2\ell}\widehat{\mathbf{y}}_\ell}{\sum_{\ell \in S_2} w_{2\ell}} \approx \frac{\sum_{\ell \in S_2} w_{2\ell}\mathring{\mathbf{y}}_\ell}{\sum_{\ell \in S_2} w_{2\ell}}.$$

If the variables are categorical, we can then estimate a contingency table using the results of the optimal transport matching, or the prediction on S_1 (respectively on S_2),

$$\widehat{\mathbf{N}}_{yz} = \sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}\mathbf{y}_k\mathbf{z}_\ell^\top = \sum_{k\in S_1} w_{1k}\mathbf{y}_k\widehat{\mathbf{z}}_k^\top = \sum_{\ell\in S_2} w_{2\ell}\widehat{\mathbf{y}}_\ell\mathbf{z}_\ell^\top.$$

We can also use the imputed values that give slightly different results,

$$\widehat{\mathbf{N}}^1_{yz} = \sum_{k\in S_1} w_{1k}\mathbf{y}_k\mathring{\mathbf{z}}_k^\top \text{ and } \widehat{\mathbf{N}}^2_{yz} = \sum_{\ell\in S_2} w_{2\ell}\mathring{\mathbf{y}}_\ell\mathbf{z}_\ell^\top.$$

If the variables are continuous, we can estimate the covariances between the $\mathbf{y}_k$ and the $\mathbf{z}_\ell$ variables. We can also work indifferently from $S_1 \times S_2$, S_1 or S_2. In fact, we have

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{yz} &= \frac{1}{\widehat{N}^*}\sum_{k\in S_1}\sum_{\ell\in S_2} W_{k\ell}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top \\ &= \frac{1}{\widehat{N}^*}\sum_{k\in S_1} w_{1k}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\widehat{\mathbf{z}}_k - \widehat{\overline{\mathbf{Z}}})^\top \\ &= \frac{1}{\widehat{N}^*}\sum_{\ell\in S_2} w_{2\ell}(\widehat{\mathbf{y}}_\ell - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top.\end{aligned}$$

As previously seen, it is also possible to use the imputed values that give slightly different results

$$\widehat{\mathbf{\Sigma}}^1_{yz} = \frac{1}{\widehat{N}^*}\sum_{k\in S_1} w_{1k}(\mathbf{y}_k - \widehat{\overline{\mathbf{Y}}})(\mathring{\mathbf{z}}_k - \widehat{\overline{\mathbf{Z}}})^\top$$

and

$$\widehat{\mathbf{\Sigma}}^2_{yz} = \frac{1}{\widehat{N}^*}\sum_{\ell\in S_2} w_{2\ell}(\mathring{\mathbf{y}}_\ell - \widehat{\overline{\mathbf{Y}}})(\mathbf{z}_\ell - \widehat{\overline{\mathbf{Z}}})^\top.$$

Since $\mathrm{E}_q(\mathring{\mathbf{y}}_k) = \widehat{\mathbf{y}}_k$ and $\mathrm{E}_q(\mathring{\mathbf{z}}_k) = \widehat{\mathbf{z}}_k$, then $\mathrm{E}_q(\widehat{\mathbf{\Sigma}}^1_{yz}) = \mathrm{E}_q(\widehat{\mathbf{\Sigma}}^2_{yz}) = \widehat{\mathbf{\Sigma}}_{yz}$. The three estimators are thus very close to each other. One can thus use in an undifferentiated way $S_1 \times S_2$, S_1 or S_2.

5.7 Simulations

In this section we propose two simulations, the first one on a simulated normal dataset where the conditional independence assumption holds and a second one on the Austrian data EU-SILC (European Union Statistics on Income and Living Conditions).

5.7.1 Gaussian example

In this section we discuss a generated dataset such that the conditional independence assumption (CIA) is satisfied. Let U be a population of 10 000 units generated such that the matching variables $\mathbf{x}_k \in \mathbb{R}^p$, $k \in U$, the variables recorded only in S_1 $\mathbf{y}_k \in \mathbb{R}^q$, $k \in U$ and the variables recorded only in S_2, $\mathbf{z}_k \in \mathbb{R}^r$, $k \in U$ are normally distributed with joint distribution equal to

$$f(\mathbf{x},\mathbf{y},\mathbf{z}|\boldsymbol{\mu},\mathbf{\Sigma}) = \frac{1}{\sqrt{2\pi^{p+q+r}|\mathbf{\Sigma}|}}\exp\left[-\frac{1}{2}\left\{(\mathbf{x},\mathbf{y},\mathbf{z})^\top - \boldsymbol{\mu}\right\}^\top \mathbf{\Sigma}^{-1}\left\{(\mathbf{x},\mathbf{y},\mathbf{z})^\top - \boldsymbol{\mu}\right\}\right],$$

where the parameters are defined as follow

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zx} & \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{pmatrix}.$$

To ensure that the CIA holds, the quantity $\boldsymbol{\Sigma}_{yz}$ must be determined by the following equality (D'Orazio et al., 2006):

$$\boldsymbol{\Sigma}_{yz} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xz}$$

Let us suppose that $p = 3$, $q = 2$ and $r = 2$. We simulate 10 000 samples of recipient sample S_1 and donor S_2 by using simple random sampling without replacement with sizes fixed to $n_1 = 600$ and $n_2 = 3000$. The parameters of the matching variables are equal to

$$\boldsymbol{\mu}_x = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{xx} = \begin{pmatrix} 7.364 & 2.579 & -0.475 \\ 2.579 & 5.694 & -0.021 \\ -0.475 & -0.021 & 7.864 \end{pmatrix}.$$

By using the properties of the Gaussian distribution, if $\mathbf{y}_k = (y_{k1} \;\; y_{k2})^\top \in \mathbb{R}^2, k \in U$ and $\mathbf{z}_k = (z_{k1} \;\; z_{k2})^\top \in \mathbb{R}^2, k \in U$ are linear combination of the matching variables, the CIA holds and we can estimate the covariance matrix $\boldsymbol{\Sigma}_{yz}$. Linear combinations are then equal to

$$\begin{aligned} y_{k1} &= 0.2x_{k1} - 0.3x_{k2} + x_{k3} + \varepsilon_k, \\ y_{k2} &= 1.2x_{k1} + 0.4x_{k2} - 0.5x_{k3} + \varepsilon_k, \\ z_{k1} &= -0.4x_{k1} + x_{k2} - 0.3x_{k3} + \varepsilon_k, \\ z_{k2} &= -1.4x_{k1} + 0.3x_{k2} - 0.6x_{k3} + \varepsilon_k, \end{aligned}$$

where $\varepsilon_k \sim \mathcal{N}(0, 1)$ and the variance-covariance matrix is given by

$$\boldsymbol{\Sigma}_{yz} = \begin{pmatrix} -3.667 & -5.368 \\ 2.627 & -9.772 \end{pmatrix}.$$

Table 5.1 – *Mean squared errors and their bias-variance decomposition on 10 000 simulations of the estimation of the variance-covariance matrix* $\boldsymbol{\Sigma}_{yz}$. *Matrix notation* $\mathbf{B}$ *stands for the squared bias while* $\mathbf{V}$ *denote the variance of the estimator* $\widehat{\boldsymbol{\Sigma}}_{yz}$.

Optimal transport

$$\mathbf{B}_{yz}^{opt} = \begin{pmatrix} 0.014 & 0.022 \\ 0.013 & 0.046 \end{pmatrix} \quad \mathbf{V}_{yz}^{opt} = \begin{pmatrix} 0.032 & 0.09 \\ 0.042 & 0.162 \end{pmatrix} \quad \mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}_{yz}^{opt}) = \begin{pmatrix} 0.046 & 0.113 \\ 0.056 & 0.208 \end{pmatrix}$$

Renssen

$$\mathbf{B}_{yz}^{ren} = \begin{pmatrix} 0 & 0 \\ 0.011 & 0 \end{pmatrix} \quad \mathbf{V}_{yz}^{ren} = \begin{pmatrix} 0.104 & 0.255 \\ 0.162 & 0.568 \end{pmatrix} \quad \mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}_{yz}^{ren}) = \begin{pmatrix} 0.104 & 0.255 \\ 0.173 & 0.568 \end{pmatrix}$$

Balanced imputation

$$\mathbf{B}_{yz}^{bal} = \begin{pmatrix} 0.011 & 0.021 \\ 0.013 & 0.045 \end{pmatrix} \quad \mathbf{V}_{yz}^{bal} = \begin{pmatrix} 0.055 & 0.127 \\ 0.086 & 0.235 \end{pmatrix} \quad \mathrm{MSE}(\widehat{\boldsymbol{\Sigma}}_{yz}^{bal}) = \begin{pmatrix} 0.066 & 0.148 \\ 0.099 & 0.279 \end{pmatrix}$$

Table 5.1 shows the mean squared errors of the simulations of the three different methods. The mean squared error result shows that optimal transport procedure and balanced imputation method are better. It is interesting to point out that even in the linear case where the CIA holds, as in the previous Gaussian case, the optimal transport methods give a slightly better MSE. Nevertheless, it is important to note here that the optimal transport is slightly biased. This comes from the fact that optimal transport and balanced imputation do not use a model prediction to create the match. This means that for this particular case, the Renssen method has a bias that converges to zero faster. In this Gaussian case, the CIA holds and could be written mathematically, but in general, this assumption is not testable. The next section presents an example where the CIA does not hold. Optimal transports methods will produce better estimates in that case.

5.7.2 EU-SILC example

This section proposes a simulation study to see how the proposed method performs compared to the method proposed by Renssen (1998) on the dataset `eusilc` available in the R package Alfons and Templ (2013). This dataset contains 14 827 observations and 28 variables. It is based on real Austrian data EU-SILC (European Union Statistics on Income and Living Conditions). We slightly modified the dataset to remove the missing values. It represents thus a dataset of 12 107 observations. Table 5.2 shows a summary of the different variables used for the simulations. In particular, `pl030` is the categorical variable representing the economic status while `eqIncome` represents the continuous variable of household income. `pl030` is the variable of interest recorded only in S_1 while `eqIncome` is recorded only in S_2. Figure 5.3 shows the household income by economic status. Each simulation will estimate the average income by category.

We run simulations using a stratified balanced sampling design (Jauslin et al., 2021) with sample size $n_1 = 1000$ for each sample S_1 and $n_2 = 4000$ for each sample S_2. Inclusion probabilities are selected such that the design respects an optimal stratification, i.e., the number of unit selected in each stratum is proportional to the product of the stratum size and the standard error of `eqIncome`. In addition, the samples are balanced on the totals of the matching variables.

Let $\mathbf{m} = (m_1, \ldots, m_7)$ be the average income (of the population) by economic status, whereas $\widehat{\mathbf{m}}^j(r) = \{\widehat{m}_1^j(r), \ldots, \widehat{m}_7^j(r)\}, j = \{opt, ren, bal\}$ denote the estimation of the averages on a particular simulation sample r. Let $M = 10\,000$ denote the number of simulations. To measure the effectiveness of the methods, we define the root mean squared error relative to the true total

$$RRMSE_i^j = \frac{\sqrt{\frac{1}{M}\sum_{r=1}^{M}\left\{\widehat{m}_i^j(r) - m_i\right\}^2}}{m_i}, i = 1, \ldots, 7, j = \{opt, bal, ren\},$$

as well as the relative bias

$$RB_i^j = \frac{\frac{1}{M}\sum_{r=1}^{M}\left\{\widehat{m}_i^j(r) - m_i\right\}}{m_i}, i = 1, \ldots, 7, j = \{opt, bal, ren\}.$$

Table 5.3 shows the results of the estimation of the average income within each category based on 10 000 simulations. The table also displays the relative root mean

squared errors as well as the relative bias. We can observe that the matching based on optimal transport is the best in terms of mean squared error. Indeed, in the case of non-linear variables, the model might have difficulties in creating a good match using the linear model. However, if the distance matrix is well defined, the optimal transport will find real relations between the variables more easily. In fact, the optimal transport is more robust due to the two axes of the method. On the one hand, distance minimization means that associated units are really close to each other. On the other hand, the totals of the auxiliary variables are used as a regularized constraint to ensure that the solution keeps the expected totals. These two constraints give us a method that is more robust under the CIA compared to other methods.

Table 5.2 – *Selected variables of the* `eusilc` *dataset of the* `R` *package developed by Alfons and Templ (2013). The first five variables are the ones used for the matching while the two last ones are the variables of interest.*

Matching variables	
`hsize`	The number of persons in the household.
`db040`	The federal state in which the household is located.
`age`	The person's age.
`rb090`	The person's gender. (male or female)
`pb220a`	The person's citizenship (AT, EU and Other).
Variables of interests	
`pl030`	1 : working full time. 2 : working part time. 3 : unemployed. 4 : pupil, student, further training, unpaid work experience, in compulsory military or community service. 5 : in retirement or early retirement or has given up business. 6 : permanently disabled or/and unfit to work or other inactive person. 7 : fulfilling domestic tasks and care responsibilities.
`eqIncome`	Slightly simplified version of the household income.

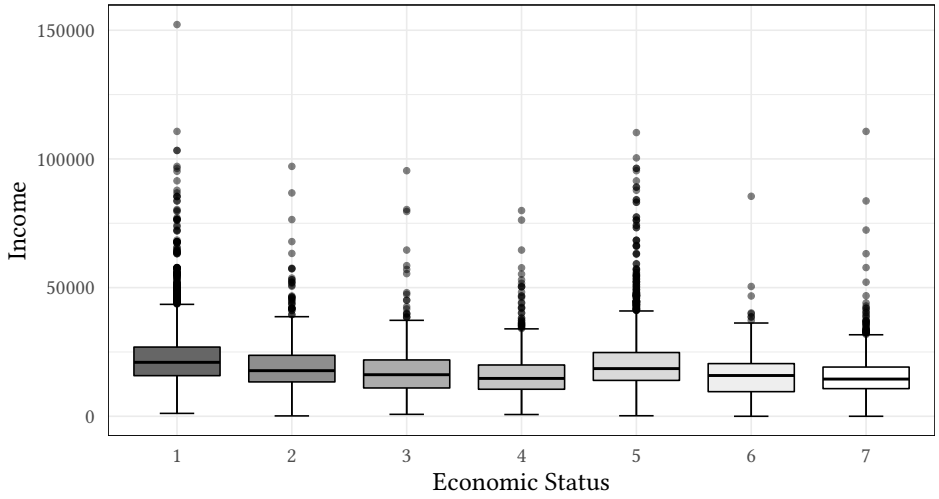


Figure 5.3 – *Boxplot of the household income by economic status.*

Table 5.3 – *Results of 10 000 simulations for the estimation of average income per category. Relative root mean squared errors as well as the relative bias are calculated for each economic status. The overall mean squared errors are equal to 47.443 for the optimal transport, 61.154 for the balanced imputation and 53.065 for the method of Renssen.*

	Economical Status						
Method	1	2	3	4	5	6	7
RRMSE							
opt	0.072	0.056	0.154	0.151	0.021	0.243	0.242
bal	0.074	0.076	0.175	0.168	0.034	0.296	0.252
ren	0.089	0.027	0.15	0.226	0.014	0.23	0.237
RB							
opt	-0.072	0.049	0.146	0.146	-0.016	0.225	0.24
bal	-0.072	0.049	0.147	0.145	-0.015	0.226	0.239
ren	-0.089	0.025	0.149	0.226	0.009	0.23	0.237

5.8 Conclusion

Statistical matching is set to become a valuable tool with the increasing amount of data created in this century. In this article, we propose new methods for matching two complex surveys. The proposed statistical matching methods are flexible, depending on the type of analysis we want to perform. We can either have a one-to-one unit matching using stratified balanced sampling, or use the optimal output of the linear program, or finally use prediction employing weighted averages.

Based on simulations, we observe that the proposed methods have lower cumulative mean squared error. A major assumption that persists in statistical matching is the conditional independence. Since in most cases this hypothesis is not satisfied, it is generally difficult to do anything other than assuming this postulation. Our method returns a mean squared error smaller in both cases whether conditional independence assumption is satisfied or not. Thus, our results show that the proposed methods are less sensitive to a conditional independence defect. This suggests that they are more efficient and give a better quality statistical match.

Chapter 6

Conclusion

In this thesis, we address three main areas of survey theory; spatial statistics, stratification and statistical matching. In Chapter 2, we show that well-spread methods are effective in reducing the error of estimators. Especially in an environmental context, the possibility of using well-spread methods should always be investigated. Chapter 2 also investigates the measure to assess the spread of a sample. Tillé et al. (2018) adjusted the Moran I index to remain between zero and one. The spatial weights are modified in Chapter 2 to better consider the spatial pattern and inclusion probabilities. We show that the commonly used measure, based on Voronoï polygons, lacks accuracy in some cases. More generally, it is not always so simple to apply a well-spread pure method in a survey. In many environmental studies, the area of study is so large that the use of these methods can be very expensive. An enlightening example of a huge region using a balanced sampling design is provided by Vallée et al. (2015).

During my thesis, it was fascinating to study the problem of systematic sampling in several dimensions. This problem is simple to present but not so simple to solve. Even the definition of a systematic sample in multiple dimensions with unequal inclusion probabilities is ambiguous. Chapter 2 shows that, if a distance without edges is used, a better sample can be selected using the weakly associated vector method. A future research area would be to clarify the definition of a systematic sample in multiple dimensions with prescribed unequal inclusion probabilities. Chapter 3 is closely related to Chapter 2. It also shows that a well-spread sample increases the quality of the estimator. Moreover, if auxiliary information is available, balancing increases the quality of the estimators even more. This has already been shown in Grafström and Tillé (2013), but the sequential feature is interesting for a large population. Nevertheless, the method has some limitations. If the number of auxiliary variables becomes large, the number of units needed in the linear program increases rapidly. In fact, the tendency in survey theory is to increase the database size and the auxiliary information available.

Stratification is a powerful sampling tool to increase the quality of estimators. As said in Tillé (2020, Chapter 4), if auxiliary information is available, it is almost always a good idea to stratify. In Chapter 4, we show a method for the efficient selection of a balanced sample from a highly stratified population. Indeed, we have significantly reduced the execution time of the commonly applied method, namely Hasler and Tillé (2014); Chauvet (2009). In particular, we reduce the execution time in its limiting cases, i.e., unequal inclusion probabilities in the strata and a sum in the strata not equal to an integer. These cases will be less rare in the future due to the increasing amount of data. Our method appears also to be useful for other methods. Indeed, the method is used in Chapter 5 if a synthetic file is desired. Stratification is a keystone of survey sampling. Since the early days of random sampling, stratification has been a powerful tool to increase the quality of estimators. Future research continues to be related to stratification, for example Eustache et al. (2022b, 2020); Jauslin and Tillé (2023).

In Chapter 5 we study the statistical matching domain of survey theory. Using the optimal transport problem, we propose an elegant way to merge two surveys. We

propose a harmonization of the two sets of sampling weights using calibration. The method can be used either for a micro or macro approach depending on if we are interested more in a synthetic file or a joint distribution between the variables of interest. Optimal transport has also been a prolific topic in the last decade. For example, the recent paper by Garès and Omer (2022) uses optimal transport to merge data sets. Optimal transport is now strongly related to the notion of Wasserstein distance. In this thesis, we have not gone into the general theory related to Wasserstein distance, but this could be an area of future research.

Throughout this thesis, we also explore the concept of balanced sampling. Balanced sampling is actually present in every chapter of this thesis. In Chapter 2, the method proposed is a balanced selection of the spatial strata. In Chapters 3 and 4 as it is the objective of the methods. Finally, a stratified balanced sampling method is applied to create a synthetic file using the output of the optimal transport in Chapter 5. This thesis, therefore, shows the importance of balanced sampling in survey theory and how it is currently being used in the research field.

Finally, I have ensured that all the methods discussed in this thesis are implemented and available for testing. All algorithms and methods presented in this thesis are available in the two R packages Jauslin and Tillé (2022); Jauslin et al. (2022a). Some of them are implemented in `R` while others are encapsulated as `C++` functions using `Rcpp` and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014).

Bibliography

Alfons, A. and Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, 54(15):1–25. (Cited pages xviii, 82, and 83.)

Baddeley, A. J. and Turner, R. (2005). `spatstat`: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42. (Cited page 48.)

Bellhouse, D. R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika*, 64(3):605–611. (Cited page 15.)

Benedetti, R. and Piersimoni, F. (2017). A spatially balanced design with probability function proportional to the within sample distance. *Biometrical Journal*, 59(5):1067–1084. (Cited pages 37, 47, and 48.)

Benedetti, R., Piersimoni, F., and Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85(3):439–454. (Cited page 15.)

Brown, J. A., Robertson, B. L., and McDonald, T. (2015). Spatially balanced sampling: application to environmental surveys. *Procedia Environmental Sciences*, 27:6–9. (Cited page 16.)

Chao, M.-T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69:653–656. (Cited page 40.)

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115–119. (Cited pages 2, 59, 60, 62, 66, and 87.)

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21:9–31. (Cited pages 40, 48, 59, 62, 63, and 64.)

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: a critical review. *International Statistical Review*, 87:S192–S218. (Cited page 71.)

Cohen, E., Duffield, N., Kaplan, H., Lund, C., and Thorup, M. (2009). Stream sampling for variance-optimal estimation of subset sums. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1255–1264. Society for Industrial and Applied Mathematics. (Cited page 40.)

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382. (Cited pages 72 and 75.)

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020. (Cited page 72.)

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85:89–101. (Cited page 16.)

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912. (Cited pages 8, 16, 18, 22, 37, 40, 48, 59, and 62.)

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591. (Cited pages 45, 47, 65, and 96.)

Dickson, M. M. and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics*, 31(4):1359–1372. (Cited page 16.)

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232. (Cited page 16.)

D'Orazio, M. (2019). *StatMatch: Statistical Matching or Data Fusion*. R Foundation for Statistical Computing, Vienna, Austria. R package version 1.4.0. (Cited page 71.)

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and practice*. John Wiley & Sons, Hoboken (New Jersey). (Cited pages 71, 73, 75, 76, and 81.)

Dudoignon, L. (2018). Fusion statistique de données d'enquêtes : dernières avancées pour les mesures d'audience. Médiamétrie, 70 rue Rivay, 92532 Levallois. (Cited page 72.)

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance c++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063. (Cited pages 23 and 88.)

Eustache, E., Jauslin, R., and Tillé, Y. (2020). *The R Package 'SpotSampling'*. CRAN project. (Cited page 87.)

Eustache, E., Jauslin, R., and Tillé, Y. (2022a). Spatiotemporal sampling with spatial spreading and rotation of units in time. *Spatial Statistics*, 47:100613. (Cited page 2.)

Eustache, E., Vallée, A.-A., and Tillé, Y. (2022b). Balanced donor imputation handling swiss cheese nonresponse. *Statistica Sinica*, accepted. (Cited page 87.)

Garès, V., Dimeglio, C., Guernec, G., Fantin, R., Lepage, B., Kosorok, M. R., and Savy, N. (2020). On the use of optimal transportation theory to recode variables and application to database merging. *The International Journal of Biostatistics*, 16(1):20180106. (Cited page 72.)

Garès, V. and Omer, J. (2022). Regularized optimal transport of covariates and outcomes in data recoding. *Journal of the American Statistical Association*, 117(537):320–333. (Cited pages 72, 76, and 88.)

Gini, C. and Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica*, Series 6, 4:1–107. (Cited page 37.)

Grafström, A. (2011). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142:139–147. (Cited page 16.)

Grafström, A. and Lisic, J. (2019). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.5. (Cited page 16.)

Grafström, A. and Lundström, N. L. P. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1):36–41. (Cited pages 8, 15, 18, 37, 38, and 40.)

Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520. (Cited pages 16, 27, 37, 47, and 48.)

Grafström, A. and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41:277–290. (Cited pages 9, 15, 27, 47, and 96.)

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14(2):120–131. (Cited pages 2, 16, 18, 38, 41, 44, 47, 48, 87, and 96.)

Guandalini, A. and Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85:250–269. (Cited pages 72 and 74.)

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York. (Cited page 27.)

Hankin, D., Mohr, M., and Newman, K. (2019). *Sampling Theory: For the Ecological and Natural Resource Sciences*. Oxford University Press, New York. (Cited pages 15 and 60.)

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81–94. (Cited pages 2, 60, 63, 66, 78, and 87.)

Hasler, C. and Tillé, Y. (2016). Balanced k-nearest neighbour imputation. *Statistics*, 105:11–23. (Cited page 59.)

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C. R., editors, *Sample Surveys: Design, Methods and Applications*, pages 215–246, New York, Amsterdam. Elsevier/North-Holland. (Cited page 71.)

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685. (Cited pages 2 and 6.)

Jauslin, R., Eustache, E., Panahbehagh, B., and Tillé, Y. (2022a). *StratifiedSampling: Different Methods for Stratified Sampling*. R Foundation for Statistical Computing, Vienna, Austria. R package version 0.4.1. (Cited pages 64, 72, 78, and 88.)

Jauslin, R., Eustache, E., and Tillé, Y. (2021). Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science*, 4:783–795. (Cited pages 2, 72, 78, and 82.)

Jauslin, R., Panahbehagh, B., and Tillé, Y. (2022b). Sequential spatially balanced sampling. *Environmetrics*, 33(8):e2776. (Cited page 2.)

Jauslin, R. and Tillé, Y. (2020). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):431–451. (Cited pages 2, 38, 41, 42, 47, and 48.)

Jauslin, R. and Tillé, Y. (2022). *WaveSampling: Weakly Associated Vectors (WAVE) Sampling*. R package version 0.1.3. (Cited page 88.)

Jauslin, R. and Tillé, Y. (2023). An efficient approach for statistical matching of survey data through calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference*, 225:121–131. (Cited pages 2 and 87.)

Kim, J. K., Berg, E., and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, 42(1):19–40. (Cited page 71.)

Kim, J.-K. and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2):382–401. (Cited page 71.)

Kincaid, T. M., Olsen, A. R., and Weber, M. H. (2019). *spsurvey: Spatial Survey Design and Analysis*. R package version 4.1.0. (Cited pages 16 and 27.)

Knuth, D. E. (1981). *The Art of Computer Programming (Volume II): Seminumerical Algorithms*. Addison-Wesley, Reading, MA. (Cited page 40.)

Leuenberger, M., Eustache, E., Jauslin, R., and Tillé, Y. (2022). Balancing a sample almost perfectly. *Statistics & Probability Letters*, 180:109229. (Cited pages 37 and 49.)

Marker, D. A. and Stevens Jr., D. L. (2009). Sampling and inference in environmental surveys. In *Sample surveys: design, methods and applications*, volume 29 of *Handbook of Statistics*, pages 487–512. Elsevier/North-Holland, New York, Amsterdam. (Cited page 15.)

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23. (Cited pages 11, 16, and 41.)

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174. (Cited pages 2 and 6.)

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606. (Cited page 59.)

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29. (Cited page 28.)

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing. (Cited page 76.)

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13. (Cited pages 16 and 29.)

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20:355–375. (Cited page 15.)

Renssen, R. H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, 24(2):171–183. (Cited pages 71, 75, 76, and 82.)

Robertson, B., McDonald, T., Price, C., and Brown, J. (2018). Halton iterative partitioning: spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, 25:305–323. (Cited pages 2, 16, 37, 47, and 48.)

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127. (Cited pages 7, 26, and 96.)

Stevens Jr., D. L. and Olsen, A. R. (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:415–428. (Cited page 15.)

Stevens Jr., D. L. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6):593–610. (Cited pages 15, 27, 41, 47, and 96.)

Stevens Jr., D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278. (Cited pages 2, 10, 11, 15, 16, 23, and 37.)

Swiss Federal Statistical Office (2020). *Statistique structurelle des entreprises (STATENT) Description des données GEOSTAT*. Neuchatel, Switzerland. (Cited pages xvi, 66, and 67.)

Theobald, D. M., Stevens Jr., D. L., White, D. E., Urquhart, N. S., Olsen, A. R., and Norman, J. B. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management*, 40(1):134–146. (Cited page 16.)

Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York. (Cited page 16.)

Tillé, Y. (2019). A general result for selecting balanced unequal probability samples from a stream. *Information Processing Letters*, 152:1–6. (Cited page 39.)

Tillé, Y. (2020). *Sampling and Estimation From Finite Populations*. Wiley, Hoboken. (Cited pages 7, 8, 15, 27, 39, 60, and 87.)

Tillé, Y., Dickson, M. M., Espa, G., and Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on Moran's I index. *Spatial Statistics*, 23:182–192. (Cited pages 2, 16, 25, 26, 41, 42, and 87.)

Tillé, Y. and Wilhelm, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science*, 32(2):176–189. (Cited page 15.)

Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P., and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8):557–570. (Cited pages 15, 38, and 87.)

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York. (Cited page 5.)

Wang, J.-F., Stein, A., Gao, B.-B., and Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, 2(0):1–14. (Cited pages 15 and 40.)

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650. (Cited page 71.)

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15:235–261. (Cited pages 7, 26, and 96.)

Symbols and notation

Throughout this work, matrices appears in uppercase boldface letters, vectors in lowercase boldface. All vectors are considered to be column vectors.

General notation

$\mathbf{a}_k$	weighted auxiliary variable $\mathbf{x}_k/\pi_k$
$\mathbf{A}$	matrix of the vector $\mathbf{a}_k$ such that $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)^\top$
$\mathcal{A}$	affine subspace $\{\boldsymbol{\delta} \in \mathbb{R}^N \mid \mathbf{A}\boldsymbol{\delta} = \mathbf{A}\boldsymbol{\pi}\}$
b_k	sum of inclusion probabilities within the kth Voronoï polygons
β	regression coefficient
$\boldsymbol{\beta}$	vector of regression coefficient
B	spatial balance measure based on the Voronoï polygons
$\mathbf{B}$	submatrix of constraint
$\mathrm{cov}(.,.)$	covariance between two random variables
d_k	$d_k = 1/\pi_k$ survey weights
$\mathbf{D}$	distance matrix
$\delta_k(S)$	indicator variable that is equal to 1 if $k \in S$ and 0 otherwise.
$\mathrm{E}_p(.)$	expected value under the sampling design $p(.)$
$\mathrm{E}_M(.)$	expected value under the model M
$G_k(.,.)$	pseudo-distance for calibration
h	strata index
$\mathbf{h}$	vector of h
H	number of strata
$\mathbf{H}$	disjunctive matrix of the corresponding vector $\mathbf{h}$
I_B, I	Moran index (See (3.4) or (2.11))
J	length of subvector $\boldsymbol{\pi}$
k oder ℓ	index of population or sample unit $k \in U$ or $\ell \in U$
n	sample size
n_S	sample size in the sample S if the size is random
n_h	sample size in the strata U_h
N	population size
N_h	number of units in the strata U_h
$\mathbb{N}$	set of natural numbers
$p(s)$	probability selection of the sample s
$P(.)$	probability measure
$\varphi^2(.)$	Lipschitz continuous function to express a variance
q	number of auxiliary variable
$\mathbf{Q}$	diagonal matrix containing for each k, $\sum_{\ell \in U} w_{k\ell}$ on its diagonal
ρ_{kl}	function that decreases when the distance between k and ℓ increases
$\mathbb{R}$	set of real numbers
RD	relative deviation
RV	relative variance
RE	relative efficiency

π_k	first order inclusion probability
$\pi_{k\ell}$	second order inclusion probability
$\sigma(.)$	Lipschitz continuous function
s	sample or subset of the population, $s \subset U$
S	random sample such that $P(S = s) = p(s)$
$\mathbf{S}$	estimator of the variance-covariance matrix
$\mathcal{S}$	power set of U
$\mathcal{S}_n$	set of all samples with fixed sample size n
σ	singular value of singular value decomposition
U	population
$\mathbf{u}$	left singular vector of singular value decomposition
$\mathbf{v}$	right singular vector of singular value decomposition
var	variance
var_{app}	approximated variance calculated on the population
$\widehat{\text{var}}_{app}$	estimator of the approximated variance calculated on the sample
$\widehat{\text{var}}_{SYG}$	Sen (1953), Yates and Grundy (1953) variance estimator
$\widehat{\text{var}}_{HAJ}$	Hájek-Rosén variance estimator
$\widehat{\text{var}}_{LM}$	Local mean variance estimator developed by Stevens Jr. and Olsen (2003)
$\widehat{\text{var}}_{SB}$	Spatially balanced estimator presented by Grafström and Schelin (2014)
$\widehat{\text{var}}_{BAL}$	variance estimator for balanced sampling design presented by Deville and Tillé (2005)
$\widehat{\text{var}}_{DBS}$	Doubly balanced sampling variance estimator developed by Grafström and Tillé (2013)
$w_{k\ell}$	spatial weight between k and ℓ
$\mathbf{W}$	spatial weights matrix
x	auxiliary variable
x_k	value of the auxiliary variable for the kth unit
$\mathbf{X}$	vector of the total of the auxiliary variables
$\widehat{\mathbf{X}}$	estimator of the total of the auxiliary variables
y	variable of interest
y_k	value of the variable of interest for the kth unit
Y	total of the variable of interest
$\widehat{Y}, \widehat{Y}_{HT}$	Horvitz-Thompson estimator
z	spatial coordinate