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

Survey Methodology

Optimizing a mixed allocation

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

by Antoine Rebecq and Thomas Merly-Alpa

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

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

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

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 253-267
Statistics Canada, Catalogue No. 12-001-X

Optimizing a mixed allocation

Antoine Rebecq and Thomas Merly-Alpa[1]

Abstract

This article proposes a criterion for calculating the trade-off in so-called "mixed" allocations, which combine two classic allocations in sampling theory. In INSEE (National Institute of Statistics and Economic Studies) business surveys, it is common to use the arithmetic mean of a proportional allocation and a Neyman allocation (corresponding to a trade-off of 0.5). It is possible to obtain a trade-off value resulting in better properties for the estimators. This value belongs to a region that is obtained by solving an optimization program. Different methods for calculating the trade-off will be presented. An application for business surveys is presented, as well as a comparison with other usual trade-off allocations.

Key Words: Sampling; calculation of allocation; optimization; dispersion of weights; Neyman allocation.

1 Introduction

In this article, we present a framework that replicates part of the surveys carried out in official statistics, specifically business surveys, for which the sampling design is most often a one-stage stratified simple random sampling. A design is created to estimate the total $T(y)$ or the mean $\overline{Y}$ of a continuous key variable of interest $y,$ where y_k designates the value of y for individual k^{th} in the population. Survey data are also used to estimate a collection of other variables, which are sometimes decorrelated or anti-correlated with $y.$

When a stratified design is used, the choice of an allocation generally serves a specific purpose, based on the classic "one objective, one sample" rule. In order to estimate the total $T(y)$ of the variable of interest with maximum precision, the Neyman allocation (1934) can be used. The specific allocations meet a precise need relative to $y.$ Where survey data are used to estimate quantities from other variables, it is desirable that the design used not deteriorate the quality of the estimators. For example, Cochran (1963) and Chatterjee (1967) propose a specific allocation for a collection of variables of interest. However, this does not solve the case of variables that cannot be included in the creation of the sampling design.

If a variable is decorrelated or anti-correlated to the variables used to calculate a specific allocation, it is known that the variance of the estimate of its total can be very strong (for example, see Ardilly, 2006). Therefore, using a proportional allocation, even when auxiliary information is available, can be advantageous. It enables us to be "agnostic" and to avoid constructing a design that will be harmful to estimate certain variables or certain parameters other than totals or means, or to estimate specific domains. We can also refer to Chiodini, Martelli, Manzi and Verrecchia (2010a, 2010b) for a more extensive discussion on the interest of proportional allocation in the trade-off.

We are interested in a certain type of allocation for stratified samplings with H strata (of respective sizes $N_h,$ for which the sum is equal to the size of the population N) of fixed size $n.$ The choice of an

[1] Antoine Rebecq, MODAL'X - Université Paris Nanterre, 200 avenue de la République, 92000 Nanterre, France. E-mail: antoine.rebecq@parisnanterre.fr; Thomas Merly-Alpa, Insee, 88 Avenue Verdier, 92120 Montrouge, France. E-mail: thomas.merly-alpa@insee.fr.

allocation consists in determining a vector $\mathbf{n} = (n_1, \ldots, n_H)$ verifying constraint $\sum_{h=1}^{H} n_h = n.$ We are specifically studying a "mixed" allocation, which consists of a trade-off between the proportional allocation and another specific allocation, responding to a specific need on one or more variables of interest in a survey:

$$\mathbf{n}_\alpha = \alpha \mathbf{n}_{\text{prop}} + (1 - \alpha)\, n_{\text{specific}} \tag{1.1}$$

where $\mathbf{n}_\alpha$, $\mathbf{n}_{\text{prop}}$ and $\mathbf{n}_{\text{specific}}$ are vectors of size H and $\alpha \in [0, 1]$. This trade-off allocation corresponds to the ROAUST method (Chiodini, Manzi, Martelli and Verrecchia, 2017) when the specific allocation chosen is the Neyman allocation (1934). Proportional allocation is defined by:

$$n_h = n \frac{N_h}{N}, \;\; h = 1, \ldots, H.$$

The purpose of this article is to propose a method for determining α. As a result, we would like to calculate a parameter that satisfies a certain optimality criterion that we will detail in Section 2.1. In this article, we will not discuss the composition of the strata, a subject that has been widely explored in the literature, such as in Baillargeon, Rivest and Ferland (2007) and Dalenius and Hodges Jr. (1959). Moreover, we are not trying to account for the phenomena of non-response here.

Note that proportional allocation is one that minimizes the dispersion of weights. Choosing a proportional allocation therefore comes down to the more general logic of choosing a design that minimizes the dispersion of design weights. The design of the INSEE master sample was designed with this objective in mind (Christine and Faivre, 2009), in a design-based logic. In a model-based logic, if we seek to estimate parameters (coefficients of the regression line) and the sampling design is non-informative, then constant design weights minimize variance of the estimate (see Davezies and D'Haultfoeuille, 2009 and Solon, Haider and Wooldridge, 2015).

The trade-off allocation involves reconciling two opposite objectives: creating an effective sampling design for a variable of interest, while keeping the weights as close as possible so as not to deteriorate estimation on very diverse variables. In the following, we will formalize the optimization program corresponding to these constraints. We will present a theorem that will define the criterion of optimality that we seek to resolve. Finally, we will analyze the performances of the allocation determination method that we are proposing and compare them with some other existing methods in the literature on a practical case, particularly a survey of businesses conducted by INSEE (French National Institute of Statistics and Economic Studies).

Several known allocations are already used to perform trade-offs between several objectives. An allocation frequently used at INSEE is a Neyman allocation under local precision constraints, presented in Koubi and Mathern (2009). A better-known allocation in the literature is the Bankier power allocation (1988), which makes a trade-off between the Neyman allocation and an allocation that produces a consistent coefficient of variation of the estimate of the total of a variable of interest on each stratum. This allocation is written as follows:

$$n_h = n \frac{S_h \left(T_h (x)\right)^q \big/ \bar{Y}_h}{\sum_h S_h \left(T_h (x)\right)^q \big/ \bar{Y}_h}, \quad h = 1, \ldots, H$$

where q is a parameter in $[0; 1]$, $T_h(x)$ is a measure of the size or importance of stratum h (for example, the size of the stratum or its economic importance), S_h^2 is the empirical variance of y in stratum h and $\bar{Y}_h$ its mean.

In the expression of the Bankier allocation, q is a parameter that, like parameter α of the allocation we are proposing, arbitrates between the two contrary objectives of the allocation: when q is close to 1, the allocation is very close to a Neyman allocation, but when q tends toward 0, the allocation approaches an allocation guaranteeing equal coefficients of variation in all strata. However, the article by Bankier (1988) does not propose a method for choosing this parameter; we will present such a method in this article for our family of mixed allocations.

In this article, we propose to accomplish this trade-off by solving the following program:

$$\min_{\mathbf{n}=(n_h)_{h\in[\![1,\,H]\!]}} \sum_{h=1}^{H} n_h \left(\frac{N_h}{n_h} - \frac{N}{n}\right)^2 + \lambda \left\| \mathbf{n} - \mathbf{n}_{\text{specific}} \right\|_p \tag{1.2}$$

with $\lambda \in [0, +\infty[$, $p \geq 1$ and $\|\cdot\|_p$ denoting the standard p of a vector of size H (in this equation, the term on the right represents a distance between the trade-off allocation and the specific allocation chosen for the survey). We also observe that N/n is the average weight for the sampled units. As in a stratified design, the sampling weight for a unit in stratum h is N_h / n_h; the first term of the optimization program therefore corresponds to the mean square deviation of the weight vector, or the weight dispersion. This program therefore corresponds to a trade-off between the two desired objectives. In part 3, we will see that the interest of the method consists of the choice of an adapted value λ; this choice is decisive for finding the most appropriate balance between the two contrary objectives we are targeting with the allocation, i.e., optimality for certain variables brought by the specific allocation and by equal weighting.

The optimization program used for this paper is inspired by the program used in the CURIOS algorithm (*Curios Uses Representativity Indicators to Optimize Samples*, Merly-Alpa and Rebecq, 2015), which performs an arbitration to establish a prioritization operation for the collection of face-to-face surveys by determining a second-wave allocation. In this paper, we will consider only the problem of determining *ex ante* allocations, and therefore we will not use the algorithm in the context of its introduction.

In Section 2, we present the optimization program that solves the satisfaction of these constraints. In Section 3, we explain how the crucial λ parameter should be chosen. In Section 4, we present a practical application of the mixed allocation on data from French businesses. We conclude in Section 5 by discussing how we could extend the mixed allocation to other designs than the Neyman allocation.

2 Optimization program

The program (1.2) is difficult to resolve and analyze, which is why we will simply look for a solution on a segment between the proportional allocation and a given specific allocation, the Neyman allocation, the one most frequently used. Often, the choice of an $\alpha = 1/2$ is a good trade-off. For example, this is proposed in Chiodini et al. (2010a), or in some INSEE business survey designs.

This method combines the benefits of both methods at a low cost. However, we can question the arbitrary choice of the factor $1/2$. In this paragraph, we will present a method based on a minimization program involving the dispersion of weights as well as the distance to the Neyman allocation to choose a parameter α such as the "optimal" mixed allocation between proportional allocation and the Neyman allocation:

$$\mathbf{n}_{\alpha}^{\text{opt}} = \alpha \mathbf{n}_{\text{prop}} + (1-\alpha)\, \mathbf{n}_{\text{Neyman}}. \tag{2.1}$$

We situate ourselves here in the context of stratified sampling with H strata, ignoring the influence of non-response. This could be integrated by considering anticipated response rates or a second Poisson phase, but this unnecessarily complicates the form of the results. We will focus here on a set of allocations $(\mathbf{n}_{\alpha})$ that go through a segment between the proportional allocation $\left(\mathbf{n}_{\text{prop}}\right)$ and the Neyman allocation $\left(\mathbf{n}_{\text{Neyman}}\right)$, as indicated in equation (2.1). We therefore limit ourselves to achieving the following minimization program, a simplified form of that in equation (1.2):

$$\min_{\alpha \in [0,\,1]} \sum_{h=1}^{H} n_{\alpha,h} \left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n} \right)^2 + \lambda \alpha. \tag{2.2}$$

The term on the right corresponds to the distance between the desired allocation and the Neyman allocation, up to a constant, integrated in λ: this result is shown in Appendix A.

This minimization program depends on the chosen constant $\lambda \geq 0$. It is clear that when λ is large enough, the term of distance becomes preponderant and we obtain $\alpha = 0$ and therefore $\mathbf{n}_{\alpha} = \mathbf{n}_{\text{Neyman}}$. Similarly, when λ tends toward 0, the factor representing the dispersion of weights becomes preponderant and the allocation tends toward the proportional allocation.

3 Choosing λ

As mentioned in part 1, we must choose an adapted value for λ, which represents the importance we want to give to each term of the trade-off. We will see in this part that the choice of this value is crucial for obtaining a good trade-off parameter. For this, we will focus on the variance of the Horvitz-Thompson estimator of the total of a survey variable of interest obtained with a given allocation when the sampling design applied in each stratum is a simple random sampling.

The idea is to use a key property of the Neyman allocation, which is its flatness (for example, see Ardilly, 2006). This means that in the vicinity of the allocation, the variance of the estimator of the total for the survey variable of interest is close to its minimum value, which is satisfactory from both a theoretical and an empirical standpoint. The issue is properly defining this vicinity: if we succeed in choosing a value λ with which we can produce an allocation sufficiently close to the proportional allocation while belonging to the flat area around the Neyman allocation, we will have succeeded in obtaining weights guaranteeing an estimate with near-optimal precision for the survey's key variable of interest with minimal weight dispersion.

Actually, the choice of λ and the choice of α are interchangeable. For a fixed value of $\lambda,$ we can solve the optimization program of equation (2.2) and obtain a value $\alpha(\lambda),$ and therefore an allocation $\mathbf{n}_{\alpha(\lambda)}.$ Conversely, directly choosing a value of α favours one of the two aspects of the optimization program (distance to the Neyman allocation or equal weighting), similar to the choice of $\lambda.$ Choosing to conserve parameter λ maintains a broader application framework.

We will focus on the variance of the estimator obtained when λ varies. From the allocation $\mathbf{n}_{\alpha(\lambda)},$ it is possible to study the variance of the Horvitz-Thompson estimator of the total of a variable of interest $\hat{T}_{\text{HT}}(y) = \sum_{i \in s} \frac{y_i}{\pi_i},$ with π_i the probability of inclusion in the sample of unit i (i.e., equal to n_h / N_h if h is the stratum that contains i). We then show that there is a "flat" region in the vicinity of the precision optimum (that is, the Neyman allocation, obtained when $\lambda \rightarrow +\infty$). Therefore, choosing a λ on this flat region ensures that the precision is only slightly deteriorated from the optimum, while significantly reducing the variance of the survey weights. Mathematically, it is a matter of choosing as λ the torsion point of the curve, whose existence is ensured by the following theorem:

Theorem 1. *Let $V(\lambda)$ be the variance function of $\hat{T}_{\text{HT}}(y)$ for the allocation obtained for the α solution of the minimization program of equation (2.2) for such a $\lambda.$ Therefore, there exists a segment $S \subset [0, +\infty[$ such that:*

- *$\alpha(S) = [0, 1],$ where $\alpha(\lambda)$ associates with λ the solution of program 2.2.*
- *$V(\lambda)$ is decreasing over $S.$*
- *Its second derivative admits a maximum in $S,$ which we call the torsion point.*

This theorem is illustrated in Appendix B.

We therefore want to take λ at the torsion point of the curve, which is also a point of inflection of its derivative; this amounts to situating the "elbow" of the curve, that is, being right at the limit of the variance plateau due to the proximity of the Neyman allocation, linked to the flatness of the optimum. The intuition that justifies this choice is that, on the one hand, the variance of the estimator of the total of the key variable of interest used to calculate the Neyman allocation decreases when λ increases, because the mixed allocation then approaches the Neyman allocation, and on the other hand, beyond a certain threshold, this variance varies little and is very close to its limit, the variance obtained with the Neyman allocation. This threshold corresponds intuitively to the moment when the variance ceases to decrease significantly when λ increases. This point, whose existence is proven by the theorem, is adequately identified by analyzing the variations of the evolutions of the variance with $\lambda,$ i.e., by studying the second derivative of the variance as a function of λ and the point where this derivative reaches a maximum, the derivatives of the variance being negative. Moreover, placing ourselves at the edge of the plateau allows us to limit the maximum of the value of λ and therefore the dispersion of weights. Figure 3.1 illustrates this choice.

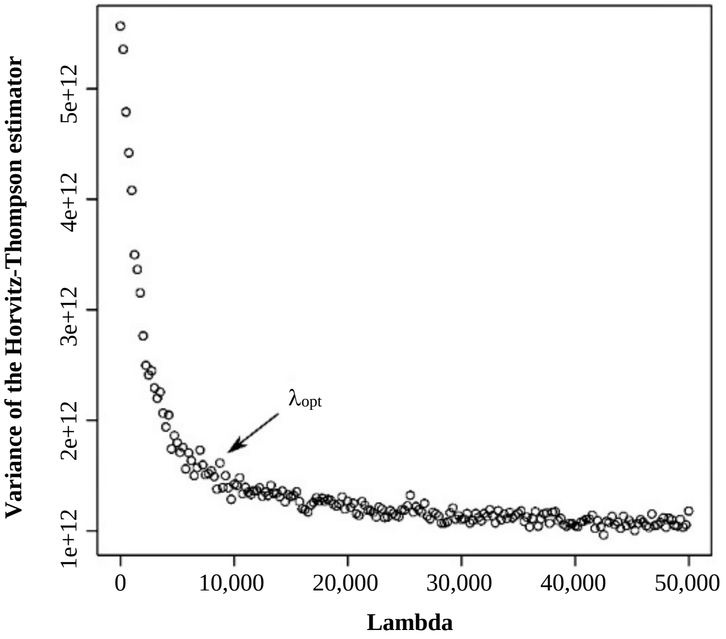

Figure 3.1 Example of a torsion point of the function $V(\lambda)$ for a sampling design explained in Merly-Alpa and Rebecq (2015).

In the simplest cases, meaning once all the information is available, and when sampling takes place in one stage, without considering other parameters, the simplest solution to determine the λ is to analytically study the curve of Figure 3.1 using the classic variance calculation formulas of $\hat{T}_{\text{HT}}(y)$ in a stratified sampling design. The torsion point is obtained by searching for the maximum of the second derivative of the curve $V(\lambda)$. This derivative is generally difficult to calculate analytically, but it is quite possible to find a numerical maximum when we have an analytical formula (or, failing that, a sufficiently smooth curve) for $V(\lambda)$.

Unfortunately, it is not always possible to analytically calculate the variance, such as when other constraints (combining strata, etc.) come into play, or if all the information is not available at the sampling stage. In this case, we replace the curve $V(\lambda)$ with a version estimated by Monte Carlo method:

1. We choose λ in $[0, 1]$.
2. The available data are used to simulate the variance of $\hat{T}_{\text{HT}}(y)$. For this, we calculate the allocation resulting from equation (2.2) for λ and we perform K independent sample draws

based on this sampling design. For each of the $k = 1, \ldots, K$ draws performed, we calculate $\hat{T}_{\text{HT}}^{(k)}(y),$ the Horvitz-Thompson estimator of $T(y)$ obtained with the data from sample $k.$ Next, we calculate the quantity:

$$V_{\text{MC}}(\lambda) = \frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{T}_{\text{HT}}^{(k)}(y) - \frac{1}{K} \sum_{j=1}^{K} \hat{T}_{\text{HT}}^{(j)}(y) \right)^2 .$$

This quantity is a Monte Carlo estimator of $V(\lambda).$

Note that these simulations require a proxy variable of the variable of interest available for all population units. For business surveys, the turnover available in the tax bases can be a good substitute variable for the actual turnover.

3. We restart for other values of λ covering $[0, 1]$ with a certain step $\eta.$ The values of η and K should be chosen by considering the calculation time, which can be quite long depending on the original population, but also to ensure that the variance due to the simulations is not too great, which would invalidate the results obtained.
4. Once these results are obtained for different values of $\lambda,$ we plot the curve of $V_{\text{MC}}(\lambda),$ which we hope is sufficiently smooth. We can then display the curve and visually place the elbow, which allows us to choose the final value of $\lambda_{\text{MC}}.$ Another possibility is to search for the maximum of the second derivative of $V_{\mathbf{MC}}(\lambda)$ using an optimization algorithm sufficiently robust to noise. For example, the algorithm of Nelder and Mead (1965) is implemented in the vast majority of optimization software (e.g., in R or in Python), and Rebecq and Merly-Alpa (2015) show that it gives good practical results for this type of problem.

In all cases, if determining λ at the elbow is difficult, a value should be chosen that ensures that we are to the right of the actual elbow on the curve. This conservative method ensures that we are on the flat region of the curve and that the precision of the estimator of the variable of interest is not impaired.

4 Practical application

We are interested in drawing a sample of 1,000 businesses in the industry based on different stratified sampling designs to learn the total turnover of the sector. The exact field is defined as follows:

- Active businesses located in France.
- Businesses with a workforce between 1 and 100.
- Businesses whose activity sector, measured using the principal activity code, belongs to one of the industry divisions in the Statistical classification of economic activities in the European Community (NACE, whose divisions are identical to the 88 divisions of the International Standard Industrial Classification of All Economic Activities–called ISIC, or CITI in French),

i.e., in divisions 10 to 33, except 12 (Manufacture of tobacco products) and 19 (Manufacture of coke and refined petroleum products), which have a structure too atypical for our study.

The initial population is 102,172 businesses. In general, businesses with a large workforce, i.e., more than 100 employees, are often surveyed exhaustively. Here, we limit ourselves to the non-exhaustive part of a survey.

This population is stratified according to two criteria:

1. The principal activity, at the division level (first two digits).
2. The employee size group, as follows: 1 to 9 employees; 10 to 19 employees; 20 to 49 employees; 50 or more employees.

this constitutes 88 strata, which will be denoted as (A, B), where A is the sector of activity and B the workforce.

We then calculate the proportional and Neyman allocations relative to the dispersion of turnover in each stratum, for $n = 1{,}000.$ Table 4.1 summarizes the characteristics of these two allocations, as well as the strata where the allocation is maximal, both in division 10 (Manufacture of food products).

Table 4.1
Distribution of sample sizes by stratum for both allocations, and sample sizes for strata corresponding to maximum sample sizes

Allocation	Min.	Median	Max.	Stratum	Proportional allocation	Neyman allocation
Proportional	1	3	278	(10, 1-9)	*278*	80
Neyman	1	5	162	(10, 20-49)	18	*162*

We want to choose the optimal mixed allocation for the problem presented in the previous paragraph. For the distance function, we choose the Euclidean distance. Equation 2.2 therefore becomes:

$$\min_{\alpha \in [0,\,1]} \sum_{h=1}^{H} n_{\alpha,h} \left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n} \right)^2 + \lambda \sqrt{\sum_{h=1}^{H} \left(n_{\alpha,h} - n_{\text{Neyman},h} \right)^2}. \tag{4.1}$$

We then apply the following method to calculate the optimal allocation:

- Calculate, for different values of $\lambda,$ the value of α solution of the minimization program for equation (4.1).
- For each $\alpha,$ calculate the corresponding allocation.

- For each allocation, analytically calculate the variance of the Horvitz-Thompson estimator for the total turnover. This is possible because we have the turnover of the businesses in the directory used as the survey frame.

The curve represented in Figure 4.1 is finally obtained. We note that its general shape corresponds to what was expected by applying Theorem 1. We visually determine the torsion point, which seems to be located around $1 \cdot 10^7$. So we place $\lambda_{\text{elbow}} = 1 \cdot 10^7$, which is slightly to the right of the elbow, on the flat part of the curve $V(\lambda)$.

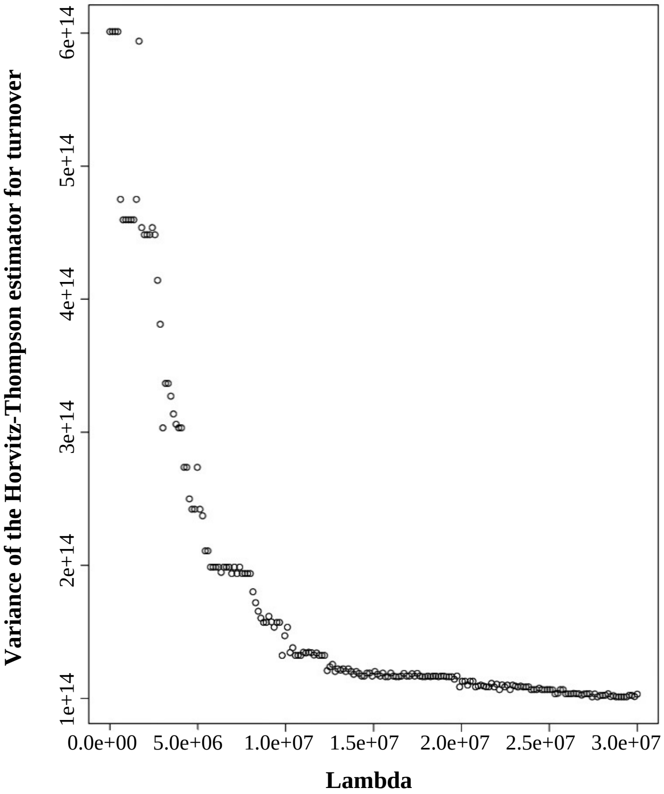

Figure 4.1 Variance of the Horvitz-Thompson estimator for total turnover as part of a trade-off with the Neyman allocation.

We can then use the value of λ_{elbow} to determine α_{elbow}, using the optimization program of equation (4.1). Here, we obtain $\alpha_{\text{elbow}} = 0.644$. This value of α can be interpreted directly. It is close enough to 0.5, which shows that the final allocation is also close enough to what is called the classically mixed allocation, but it is greater than 0.5, which shows that the program optimum is significantly approaching the proportional allocation. The allocation obtained is described in Table 4.2 and is compared with the usual mixed allocation using the arithmetic mean between the two initial allocations.

Table 4.2
Distribution of sample sizes by stratum for the allocation obtained, and for the two strata corresponding to the maximum sample sizes for the Neyman allocation and proportional allocation

Allocation	Min.	Median	Max.	α	Stratum (10, 1-9)	Stratum (10, 20-49)
Proportional	1	3	278	*1*	278	18
Elbow	1	4	208	*0.644*	208	69
Mixed	1	4	179	*0.5*	179	90
Neyman	1	3	162	*0*	80	162

In terms of sample sizes in the strata for the various allocations, we can see that a maximum is obtained for the same stratum as the proportional allocation (10, 1-9), but with a less extensive distribution. Furthermore, stratum (10, 20-49), which has the largest workforce in the Neyman allocation, actually increases in size relative to the proportional allocation, but still remains well below the Neyman allocation. We see the appearance of a trade-off between the allocations, as in the usual mixed allocation.

However, we still have to look at the two criteria that motivate this analysis, namely the standard deviation of the Horvitz-Thompson estimator for the total turnover (in billions of euros), and the dispersion of weights and its influence on the precision of estimators related to other concepts: to evaluate it, we introduce a variable z that is not correlated to turnover. Here, we choose the variable z related to the geographic location of the business defined as follows:

$$z_i = \begin{cases} 1 & \text{if the company } i \text{ is located in Ile-de-France} \\ 0 & \text{otherwise.} \end{cases}$$

We will use these three criteria to compare our method with the initial allocations (proportional, Neyman), but also with the classic mixed allocation (with a factor of 0.5), with Bankier power allocations (1988) for different values of q (where $T_h(\alpha)$ is taken as the sum of turnover in stratum h) and with the Neyman allocation under the local precision constraints from Koubi and Mathern (2009). The results obtained are presented in Table 4.3. In this table, $\hat{T}_{\text{HT}}(\text{CA})$ refers to the Horvitz-Thompson estimator of turnover, and $\hat{T}_{\text{HT}}(z)$ the Horvitz-Thompson estimator of the variable z.

Table 4.3
Dispersion of weights and variance of estimators of turnover and of z for several allocations

Allocation	Parameter	Standard deviation of $\hat{T}_{\text{HT}}(\text{CA})$	Dispersion of weights	Standard deviation of $\hat{T}_{\text{HT}}(z)$
Proportional	$\alpha = 1$	24.7	47	10.7
Elbow	*0.644*	12.5	1,929	11.6
Mixed	*0.5*	11.4	3,473	12.3
Neyman	*0*	9.8	18,585	17.9
Bankier	$q = 0.25$	13.1	36,250	22.2
	0.5	11.2	25,922	19.7
	0.75	10.1	20,187	18.2
Koubi-Mathern	·	12	35,680	22.7

We observe here that the allocation obtained using λ_{elbow} has a precision for the estimate of total turnover that is quite close to the Neyman allocation, while the proportional allocation leads to a much larger standard deviation of the Horvitz-Thompson estimator for total turnover. However, this slight loss in precision is largely offset by the gain in weight dispersion compared with the Neyman allocation and by a significant gain in terms of precision over the total of the geographic variable z. Note that the dispersion of weights is not nil in the proportional allocation because of rounding. When we compare the allocation obtained with the "mixed" strategy using the factor $\alpha = 1/2$, we observe that the loss of a factor 1.1 in the precision of the total turnover is compensated by the gain of a factor 1.8 in weight dispersion and of 1.1 in the precision of the total number of businesses located in the Île-de-France region. The final allocation satisfies our constraints and meets our specification: to have good precision and low dispersion of weights.

Comparison with the methods in the literature illustrates the contribution of the trade-off on the dispersion of weights. For the power allocations, we find that by choosing high values of q corresponding to allocations close to the Neyman allocations, we obtain better precision for the estimate of total turnover than for our allocation. We note that for all the Bankier allocations and for the Neyman allocation under constraints, the weight dispersion is greater than for the Neyman allocation, and therefore much greater than for our allocation. Symmetrically, and as expected, all these allocations contribute to weaken the precision of the estimated total of the variable z.

As the objective of these competing methods is to obtain better local precision, we will examine several subdomains of our field (statistical classification A17 of the French economy):

- Domain C1: Manufacture of food products, beverages;
- Domain C3: Manufacture of electrical, electronic and computer equipment; Manufacture of machines;
- Domain C4: Manufacture of transport equipment;
- Domain C5: Manufacture of other industrial products.

We then compare the precision of the total turnover estimator for each sector. The results are compiled in Table 4.4.

Table 4.4
Local precisions of the total turnover estimator for several allocations

Allocation	Parameter	C1	C3	C4	C5
Proportional	$\alpha = 1$	0.29	0.30	0.46	0.16
Elbow	*0.644*	0.16	0.20	0.35	0.07
Mixed	*0.5*	0.15	0.18	0.30	0.07
Neyman	*0*	0.12	0.15	0.25	0.06
Bankier	$q = 0.25$	0.21	0.13	0.18	0.07
	0.5	0.17	0.13	0.19	0.06
	0.75	0.14	0.14	0.22	0.06
Koubi-Mathern	.	0.11	0.11	0.11	0.09

We observe here that the allocation we propose gives slightly worse results than the classic mixed allocation on the local precision of the total turnover estimator. However, it is much better than the proportional allocation and, slightly less so, less effective than the Neyman allocation. Our method of choosing α is thus an effective trade-off for reducing the dispersion of weights without overly impacting the overall and local precision of the estimators.

In contrast, and as expected, the allocations with the trade-off objective of maximizing or standardizing local precision are better than the proposed allocation for the majority of the sectors. Therefore, choosing between the trade-off we propose and the one proposed by Bankier (1988) comes down to choosing between better precision for variables not correlated with the variable of interest y (via weight dispersion), like variable z defined here, for our family of mixed allocations, or choosing better local precision for only this variable y in the case of the power allocation. However, the advantage of our method is being able to propose a value of the optimal trade-off parameter α on a certain criterion, which the Bankier method does not do with parameter q.

5 Conclusion

For the stratified designs, we have studied a trade-off allocation situated on a segment between the proportional allocation and the Neyman allocation. A theorem guarantees the existence of a flat region in the vicinity of the optimum and of a particular point that gives an optimal trade-off parameter according to a certain criterion. As part of a survey of businesses in the industry, simulations are conducted showing how the calculation can be done in practice and that the usual choice of a parameter of $1/2$ is not always the most effective. A comparison with other trade-off allocations, such as the classic Bankier allocation, shows that our weight dispersion goal produces more equal weighting at the expense of lower precision for the variable of interest on subdomains of the field. However, it illustrates the variability of the results obtained for the trade-off allocations according to the value of the parameter used; our method for determining parameter α remedies this problem often encountered in the study of these allocation families.

It is possible to replace the Neyman allocation in the trade-off with other specific ad hoc allocations. We postulate that the method remains applicable to obtain the same desirable properties. Different applications of this work were carried out at INSEE with other specific allocations. In the case of the annual Survey on the cost of labour and wage structure (ECMOSS), the specific allocation used for drawing the surveyed businesses is part of a two-stage design where, in each establishment sampled in the first stage, a sample of employees is drawn. The allocation used in the first stage is then optimized to obtain the lowest estimate variance on the estimated total net pay on the final sample of employees, given the dispersion of wages in each establishment. The allocation also integrates precision constraints on certain dissemination domains. Curves of the desired shape are still obtained and the trade-off allocation can be implemented.

Appendix A

Distance term in equation (2.2)

The choice of distance (i.e., a value for p) in the second term of optimization program (1.2) is not crucial in the proposed context, because we will be able to rewrite the second term as follows where C_p is a strictly positive constant dependent only on the choice of p:

$$\left\| \mathbf{n}_{\alpha} - \mathbf{n}_{\text{Neyman}} \right\|_p = \alpha C_p. \tag{A.1}$$

Let us demonstrate this result. By definition (2.1), we have in each stratum h:

$$n_{\alpha,h} = \alpha n_{\text{prop},h} + (1-\alpha)\, n_{\text{Neyman},h}$$

and therefore,

$$n_{\alpha,h} - n_{\text{Neyman},h} = \alpha \left(n_{\text{prop},h} - n_{\text{Neyman},h} \right).$$

We therefore have for any choice of p:

$$\begin{aligned}
\left\| \mathbf{n}_{\alpha} - \mathbf{n}_{\text{Neyman}} \right\|_p &= \left(\sum_{h=1}^{H} \left| n_{\alpha,h} - n_{\text{Neyman},h} \right|^p \right)^{\frac{1}{p}} \\
&= \left(\sum_{h=1}^{H} \alpha^p \left| \left(n_{\text{prop},h} - n_{\text{Neyman},h} \right) \right|^p \right)^{\frac{1}{p}} \\
&= \alpha \left(\sum_{h=1}^{H} \left| n_{\text{prop},h} - n_{\text{Neyman},h} \right|^p \right)^{\frac{1}{p}} \\
&= \alpha C_p.
\end{aligned}$$

We will then integrate C_p, a strictly positive constant, into λ.

Appendix B

Demonstration of Theorem 1

For a $\lambda \geq 0$, the minimization function of program (2.2) is written as follows:

$$\begin{aligned}
f(\alpha) &= \sum_{h=1}^{H} n_{\alpha,h} \left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n} \right)^2 + \lambda\alpha \\
&= \sum_{h=1}^{H} \left(\frac{N_h^2}{n_{\alpha,h}} - 2\frac{N}{n} N_h + \frac{N^2}{n^2} n_{\alpha,h} \right) + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h^2}{n_{\alpha,h}} - 2\frac{N^2}{n} + \frac{N^2}{n} + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h^2}{\alpha \frac{nN_h}{N} + (1-\alpha)\, n_{\text{Neyman},h}} - \frac{N^2}{n} + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h}{\alpha \frac{n}{N} + (1-\alpha) \frac{n_{\text{Neyman},h}}{N_h}} - \frac{N^2}{n} + \lambda\alpha.
\end{aligned}$$

We now pose for all $h \leq H$:

$$\beta_h = \frac{n}{N} - \frac{n_{\text{Neyman},h}}{N_h}.$$

For each stratum, the β_h represent the difference between the uniform and Neyman sampling fractions. When $\beta_h < 0,$ this means that the Neyman allocation is greater than the proportional allocation; the variable of interest is more dispersed in this stratum. Let us now derive f:

$$f'(\alpha) = \sum_{h=1}^{H} \frac{-N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2} + \lambda. \tag{B.1}$$

We deduce from equation (B.1) that the derivative cancels out when:

$$\lambda = \sum_{h=1}^{H} \frac{N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2} =: g(\alpha).$$

Now function g_h defined as follows:

$$g_h \colon \alpha \rightarrow \frac{N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2}$$

is decreasing. So:

- If β_h is negative, the denominator decreases when α increases. In this case, its inverse increases with α. Therefore, we multiply by β_h to obtain g_h, which implies that g_h is decreasing.
- If β_h is positive, the denominator increases when α increases. By inverting and then multiplying by β_h, we find that g_h is decreasing.

So, if $\lambda \in [g(1), g(0)]$, we know that there is an α_0 that cancels the derivative. As f' evolves inversely to g, f' is increasing and therefore α_0 is the minimum of f on $[0, 1]$.

Furthermore, as $g(\alpha_0) = \lambda$ by definition, the decrease of g implies that when λ increases in $[g(1), g(0)]$, then α_0 decreases. We therefore use the following lemma, admitted because it is relative to a classic property of the Neyman allocation:

Lemma 1. *The function that at α associates the variance of the Horvitz-Thompson estimator of the variable of interest X for the allocation $n_{\alpha,h}$ is increasing.*

We deduce that $V(\lambda)$ is decreasing over S. Finally, by continuity, V'' admits a maximum over S.

References

Ardilly, P. (2006). *Les Techniques de Sondage*. Editions Technip.

Baillargeon, S., Rivest, L.-P. and Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42.3, 174-177.

Chatterjee, S. (1967). A note on optimum allocation. *Scandinavian Actuarial Journal*, 1-2, 40-44.

Chiodini, P.M., Manzi, G., Martelli, B.M. and Verrecchia, F. (2017). *Sampling Allocation Strategies: A Simulation-Based Comparison*. url: https://events.unibo.it/itacosm2017/abtracts-of-invited-papers/manzi-et-al_presentation_itacosm17.pdf/@@download/file/Manzi%20et%20al_Presentation_ITACOSM17.pdf.

Chiodini, P.M., Martelli, B.M., Manzi, G. and Verrecchia, F. (2010a). The ISAE manufacturing survey sample: Validating the Nace Rev. 2 sectorial allocation. *Economic Tendency Surveys and the Services Sector*. CIRET.

Chiodini, P.M., Martelli, B.M., Manzi, G. and Verrecchia, F. (2010b). Between theoretical and applied approach: Which compromise for unit allocation in business surveys? *SIS Conference*. Società italiana di statistica.

Christine, M., and Faivre, S. (2009). Le projet OCTOPUSSE de nouvel Échantillon-Maître de l'INSEE. *JMS*, 24.

Cochran, W. (1963). *Sampling Techniques*.

Dalenius, T., and Hodges Jr, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54.285, 88-101.

Davezies, L., and D'Haultfoeuille, X. (2009). Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à un problème de sondage. Working paper of INSEE.

Koubi, M., and Mathern, S. (2009). Résolution d'une des limites de l'allocation de Neyman. *JMS*, 1.

Merly-Alpa, T., and Rebecq, A. (2015). L'algorithme CURIOS pour l'optimisation du plan de sondage en fonction de la non-réponse. *Journées de la Statistique de la SFdS*, Lille.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7.4, 308-313.

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*, 558-625.

Rebecq, A., and Merly-Alpa, T. (2015). Algorithme CURIOS et méthode de « priorisation » pour les enquêtes en face-à-face. Application à l'enquête Patrimoine 2014. *Actes des Journées de Méthodologie Statistique*.

Solon, G., Haider, S.J. and Wooldridge, J.M. (2015). What are we weighting for? *Journal of Human Resources*, 50.2, 301-316.