Revisiting Sample Allocation Methods: A Simulation-Based Comparison

Paola Maddalena Chiodini[*]
Giancarlo Manzi[†]
Bianca Maria Martelli[‡]
Flavio Verrecchia[§]

March 20, 2019

[*] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy. E-mail: paola.chiodini@unimib.it

[†] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy. E-mail: giancarlo.manzi@unimi.it

[‡] Former Head of Central Region Office, National Institute of Statistics, Italy. E-mail: biancamaria.martelli@gmail.com

[§] (*Corresponding Author*), Lombardy Office, National Institute of Statistics, Italy. E-mail: verrecchia@istat.it

Abstract

In stratified sampling the problem of optimally allocating the sample size is of primary importance, especially in business surveys when reliable estimates are required both for the overall population and for the domains of studies. To this purpose, in this paper we compare allocation methods via a simulation engine highlighting the effects on the reliability of the estimates due only to the sample allocation design. Allocation methods considered in this comparison are: the Neyman allocation, the uniform and proportional allocations, the Costa allocation, the Bankier allocation, the Interior Point Non Linear Programming allocation and the Robust Optimal Allocation with Uniform Stratum Threshold, an allocation method recently adopted by the Italian National Statistical Institute. The last two methods outperform the others at the stratum level. At the overall sample level they perform better than the others together with the Neyman allocation method.

Keywords: Business Surveys; Stratified Sampling; Compromise Allocation; Interior Point Non Linear Programming; Monte Carlo Simulation

1 Introduction

In stratified sampling, the optimal allocation of sample units is an ubiquitous problem, especially in business surveys when the survey frame changes continuously due to highly frequent firm inclusions and exclusions, or when new economic sectors are inserted in the official classification (Hidiroglou & Srinath 1993, Khan et al. 2015). Stratification is often based on predetermined rules and constraints according to some fixed geographical administrative strata and economic classifications (e.g. firm sizes and sectors of activity); therefore, it is not always possible to adjust the stratification process *ad libitum* in order to increase the stratum homogeneity so as to optimize a survey plan, as it is the case in business surveys when the firm size tends to have a positively skewed distribution (Smith et al. 2003).

The Neyman allocation for stratified sampling (Neyman 1934) is a popular method often used in business surveys (Smith et al. 2003, Smith & James 2017) and is often regarded as the most important benchmark in this field. This popularity stems from the fact that information from the sample is sufficiently retained and estimate efficiency is guaranteed also when facing challenging sampling issues, for example when both the overall and the stratum sample sizes are small (Kozak et al. 2007, Särndal et al. 2013, Hidiroglou & Kozak 2018).

However, when a given level of precision cannot be reached or the stratum sample size is greater than the corresponding stratum population size, the resulting allocation may be sub-optimal (Kozak 2006). In recent years, many authors have proposed flexible solutions when reliable estimates are required both for the overall population and domain level (Särndal et al. 2013, pp. 386-390), or when composite estimators for small strata are required (Molefe & Clark 2015, Bankier 1988, Choudhry et al. 2012).

Many authors have implemented comparative studies on allocation methods under particular constraints, among others Er (2012), Kozak (2014), Keto & Pahkinen (2017) and Clark et al. (2017).

This paper can be framed in this area of research, being its main aim the comparison of allocation methods used in a business survey framework. To this purpose we introduce the concept of *compromise allocation* intending the use of a mixture of allocation methods. Among others, we considered the Neyman allocation, the Costa allocation (Costa et al. 2004), the Bankier allocation (Bankier 1988), the proportional and the uniform allocation methods. In particular, we focus on the problem of finding suitable allocation methods both for the overall population and for domain level[1] estimation for the Italian Business Confidence Survey (IBCS), a survey conducted by the Italian National Statistical Institute (ISTAT). Furthermore, we want to define an optimum allocation strategy

[1] Throughout this paper we consider domain levels as equivalent to strata.

when facing the following constraints[2].:

1 The adopted stratification yields many small strata having a high variability;

2 The distribution of the target variable is highly skewed;

3 High levels of enterprise birth/death rate are present, with consequent frequent changes in the frame and delay in its updating.

For these reasons in our comparative studies we include the Interior Point Non Linear Programming (IPNLP) method and the Robust Optimal Allocation with Uniform Stratum Threshold (ROAUST) method, a method that has been recently adopted by ISTAT. One way to perform this comparison is through a simulation study as if we were a national statistical agency having to conduct a business confidence survey, like the IBCS in the case of ISTAT, and give national economic confidence indicators.

2 Data and Frame

An exhaustive list of all the Italian firms is provided by the Statistical Archive of Active Enterprises - ASIA (ISTAT 2007). This is the frame used in the IBCS (Table 1) and also the frame used in our comparison.

[2] Recently, many statistical agencies have dealt with some sort of 'stratum sample size constrains' (see Chiodini et al. (2017), pp. 3-4 for details)

Table 1: Main features of the ICBS

Starting Year:	1961
Timing:	Monthly, fieldwork in the first half of the reference month; dissemination of results within the end of the reference month
Frame:	Active Business Integrated Statistical Register - A.S.I.A. Each year t the archive is updated with reference to year $t-2$
Target population:	Manufacturing businesses with more than 10 p.e.
Sample unit:	Firm
Variables/Questions:	Qualitative questions mainly on a 3 points Likert scale; Assessment and/or expectation on order book, production, stocks, prices, employment, etc.
Data collection mode:	Mixed, mainly CATI; sometimes fax
Sample design:	Stratified random sample by 4 geographical regions, 19 economic sectors and 3 size classes (10-49; 50-249; 250 and more)
Allocation method:	ROAUST for units with less than 1,000 p.e.; remaining all included
Sample size/coverage	About 4,000 monthly interviews. Coverage: 4.6% of firms.
Estimation/weighting	Data estimated in two steps: 1) applying size weights to each unit and each question: persons employed declared by firms to transform categorical replies in percentages according to relative importance of the firms; 2) applying stratum weight according to value added. Balances for each question are calculated in each stratum as simple differences between favorable and unfavorable reply options. Series stemming from the survey waves are seasonally adjusted.

A lower cut-off is applied in forming the IBCS frame by excluding firms with less than 10 persons employed (Ellison & Elvers 2001). Therefore, the selected frame comprises details for just less than 90 thousand manufacturing firms (about 18% of all Italian manufacturing firms), accounting for almost 77% of the economic activity in terms of number of persons employed in manufacturing.

Strata refers to three variables: firm size (generally in terms of persons employed), economic classification and geographical areas. Firm size refers to 3 classes: small firms (10 - 49 persons employed), medium-sized firms (50 - 249 persons employed) and large firms (with at least 250 persons employed). These three classes, the economic sector classification and the geographical classification are in line with specific European Commission recommendations (European Union Commission 2003). The final frame consists of 226 strata (i.e. 228 strata minus two empty strata) and is described in details in Table 2.

The variable at the center of our study will be the number of employees Y. Units sampled from the ASIA archive refers to time *t-2* as stated in Table 1. In particular the employees are used as:

1 A proxy of the economic confidence reflecting the importance of the enterprises, therefore as an auxiliary variable (for example, bigger enterprises must have more weight in the construction of the economic indicator) at time *t-2*.

Table 2: Number of firms by stratum (Region, size and economic sector)

						Regions							
	North-West			North-East			Centre			South and islands			
	Firm size			Firm size			Firm size			Firm size			
Sectors*	10-49	50-249	250+	10-49	50-249	250+	10-49	50-249	250+	10-49	50-249	250+	Total
10-12.	1,566	228	54	1,782	278	39	1,156	86	15	1,978	177	15	7,374
13	1,503	328	52	557	75	9	983	80	3	273	25	-	3,890
14	1,398	128	22	1,817	138	26	1,217	99	8	1,317	86	6	6,262
15	309	45	-	879	117	14	2,095	136	9	628	46	3	4,283
16-17.	1292	148	20	1524	163	14	898	84	10	780	48	4	4985
18	952	79	11	739	58	5	507	34	-	298	19	-	2,704
19	31	10	7	17	5	.	21	7	4	74	6	4	186
20-21.	668	291	89	346	115	15	230	63	32	231	33	.	2,113
22	1,607	288	41	1,003	189	16	553	89	4	447	61	7	4,305
23	919	127	19	1,199	229	50	878	103	16	1,236	95	-	4,872
24	641	202	42	277	115	16	162	32	8	161	34	5	1,695
25	6,428	635	42	4,799	443	31	2,074	177	9	2,086	214	11	16,949
26	738	146	29	430	103	15	277	57	13	122	23	4	1,957
27	1,060	204	36	842	162	28	353	60	15	198	23	3	2,984
28	3,247	665	88	2,823	608	105	750	128	8	521	57	4	9,004
29-30.	571	183	79	373	99	31	333	71	15	245	87	19	2,106
31	898	90	5	1,659	234	19	925	101	8	475	49	6	4,469
32	598	80	12	692	107	7	492	36	4	194	6	-	2,229
33	1,408	87	5	958	43	3	674	27	4	802	63	6	4,080
Total	25,834	3,964	655	22,716	3,281	443	14,578	1,470	186	12,066	1,152	102	86,447

Notes:
***** 10-12: Manufacture of food, beverages and tobacco products; 13: Manufacture of textiles; 14: Manufacture of wearing apparel; 15: Manufacture of leather and related products; 16-17: Manufacture of wood and paper products; 18: Printing and reproduction of recorded media; 19: Manufacture of coke and refined petroleum products; 20-21: Manufacture of chemical and pharmaceutical products; 22: Manufacture of rubber and plastic products; 23: Manufacture of other non-metallic mineral products; 24: Manufacture of basic metals; 25: Manufacture of fabricated metal products, except machinery and equipment; 26: Manufacture of computer, electronic and optical products; 27: Manufacture of electrical equipment; 28: Manufacture of machinery and equipment n.e.c.; 29-30: Manufacture of transport vehicles; 31: Manufacture of furniture; 32: Other manufacturing; 33: Repair and installation of machinery and equipment.
. Missing data.

- Less than 3 units.

2 A measure of the stratum variability at time *t-2* as used in some of the allocation methods included in the comparison.

3 The most important constituent part of the quantities (i.e. the bias and the RMSE) used to evaluate their performance with data at time t .

3 Allocation Methods

In this section, allocation methods used in the simulation study and some modifications needed to comply with the constraints above highlighted are presented.

In IBCS a predetermined sample size of around $n \approx 4,000$ is assumed and kept identical for each allocation method throughout the simulation study. This simulation setting implies that gains in estimate precision will be a consequence of the efficiency of the allocation methods only and not of the estimators' efficiency.

3.1 An Adjusted Uniform Allocation

The uniform allocation implies a constant stratum sample size n_h for stratum h, which is set independently on the population stratum size N_h. Therefore, if $n = \sum_h n_h$ is the total sample size and H is the total number of strata, the sample size for stratum h is given by:

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

The uniform allocation ensures a non-null sample size in each stratum also in those strata where the population size is very small, regardless, for example, proportionality criteria. In our case, n/H is about 18. However, for the frame presented in Table 2, fifty-seven strata have size lower than 18. Therefore, we apply a slightly adjusted uniform allocation method in the simulation when $n/H > N_h$. This is performed as follows. Let $\mathcal{A}$ be the set of strata where $n/H \leq N_h$, and $\mathcal{B}$ the set of m strata where $n/H > N_h$. Compute:

$$n_h^* = \frac{n - \sum_{h \in \mathcal{B}} N_h}{H - m}.$$

Then, n_h becomes:

$$n_h = \begin{cases} n_h^* \text{ if } h \in \mathcal{A} \\ \\ N_h \text{ if } h \in \mathcal{B} \end{cases} . \tag{1}$$

The adjusted uniform allocation will be implemented using the pseudo code in the following box.

Pseudo code 1: Adjusted Uniform Allocation

```
Step 1. Set the array of population stratum sizes $N_h$ $(N_1, \ldots, N_H\ )$, the population size $N = \sum_h N_h$, the sample size $n = \sum_h n_h$,
the uniform stratum sample size $n_h = n_0/H$.

Step 2,. Initialize $n = 0$, $n^{\mathcal{A}} = 0$, $n^{\mathcal{B}} = 0$, $n^{(r)} = 0$, $m = 0$, $n_0 = 4,000$, $(n_1, \ldots, n_H) = (0, \ldots, 0)$.

Step 3. For $h = 1$ to $H$

        If $N_h < \lfloor n_h \rfloor$ then do:
            Set $h \in \mathcal{B}$
            Set $n_h^{\mathcal{B}} = N_h$
            $n^{\mathcal{B}} = n^{\mathcal{B}} + n_h^{\mathcal{B}}$
            $m = m + 1$
        Next $h$

Step 4. $H^{\mathcal{A}} = H - m$; $n^{\mathcal{A}} = \lfloor n_h \rfloor H^{\mathcal{A}}$; $n^{(r)} = n_0 - n^{\mathcal{B}} - n^{\mathcal{A}}$

Step 5. Do until $\lfloor n^{(r)}/H^{\mathcal{A}} \rfloor = 0$:

        $n_h = n_h + \lfloor n^{(r)}/H^{\mathcal{A}} \rfloor$
        Do Steps 2, 3 and 4

Step 6. For $h = 1$ to $H$

        If $h \notin \mathcal{B}$ then do:
            Set $h \in \mathcal{A}$
            End
            Next $h$

Step 7. $n = n^{\mathcal{A}} + n^{\mathcal{B}}$; $n_h^{\mathcal{A}} = n^{\mathcal{A}}/H^{\mathcal{A}}$
```

In Pseudo code 1, $n^{(r)}$ is the number of units not assigned to any stratum (residual units) and $n_h^{\mathcal{A}}$ is the adjusted stratum sample size for those strata for which $N_h \geq n_h$. In our simulation two iterations will be enough to reach the target sample size. In the first iteration the sample size for the uniform allocation is $n = 3,486$ ($n^{\mathcal{B}} = 444$; $n^{\mathcal{A}} = 3,042$) and the number of residual units is $n^{(r)} = 514$. In the second iteration the sample size for the equal allocation is $n = 3,984$ ($n^{\mathcal{B}} = 540$; $n^{\mathcal{A}} = 3,444$) and the number of residual units is $n^{(r)} = 16$.

3.2 An Adjusted Proportional Allocation

The proportional allocation represents the simplest methodology to build a self-weighting sample, although it could be unsatisfactory at the stratum level when the strata size is small. In the proportional allocation the sample size for stratum h is given by:

$$n_h = nW_h, h = 1, \ldots, H$$

where $W_h = N_h/N$. We use a slightly modified proportional allocation to ensure a non-null sample size in each stratum through a uniform stratum threshold equal to 1 (corresponding to a PAUST$_1$ allocation, see below), for $N_h > 0$:

$$n_h = 1 + (n - H)W_h, h = 1, \ldots, H.$$

For the data considered here, $n/N = 4.63\%$ and therefore in 41 strata we have $n_h = 0$. This adjusted proportional allocation will be implemented using the pseudo code provided in the following box, where first a uniform threshold is set to 1, and then a proportional quota is set for those strata where $nW_h > 1$.

Pseudo code 2: Adjusted Proportional Allocation

```
Step 1. Set the array of the population stratum weights $W_h$, $(W_1, \dots, W_H)$, the sample size $n = \sum_h n_h$, the initial proportional stratum sample size $n_h = n_0 W_h$

Step 2,. Initialize $n = 0$, $n^{\mathcal{A}} = 0$, $n^{\mathcal{B}} = 0$, $n^r = 0$, $n_0 = 4,000$, and $(n_1, \dots, n_H) = (0, \dots, 0)$

Step 3. For $h = 1$ to $H$

        If $N_h > 0$ then $n^{\mathcal{B}} = 1$ End
        $n^{\mathcal{B}} = n^{\mathcal{B}} + n_h^{\mathcal{B}}$
        Next $h$

Step 4. $n_0^{\mathcal{A}} = n_0 - n^{\mathcal{B}}$

Step 5. For $h = 1$ to $H$

        If $n_0 W_h > 1$ then $n_h^{\mathcal{A}} = \lfloor n_0^{\mathcal{A}} W_h \rfloor$ End
        $n^{\mathcal{A}} = n^{\mathcal{A}} + n_h^{\mathcal{A}}$
        Next $h$

Step 6. For $h = 1$ to $H$

        $n_h = n_h^{\mathcal{B}} + n_h^{\mathcal{A}}$
        $n = n + n_h$
        Next $h$

Step 7. $n^{(r)} = n_0 - n$
```

In Pseudo code 2, $n^{(r)}$ is the number of units not assigned to any stratum (residual units) and $n_h^{\mathcal{A}}$ is the adjusted stratum sample size for those strata for which $n_0 W_h > 1$. $n^{\mathcal{B}} = 226$, $n_0^{\mathcal{A}} = 3,774$ and $n^{\mathcal{A}} = 3,760$. In this algorithm, the final resulting sample size is $n = 3,984$ ($n^{\mathcal{B}} = 226$; $n^{\mathcal{A}} = 3,760$) and the number of residual units is $n^{(r)} = 16$.

3.3 A Neyman 'Compromise' Allocation

Given the target sample size n, the Neyman allocation method allows for an increase in the precision of the estimates by assigning different sampling fractions to the strata. This is performed by letting the sampling fractions depend on both the stratum standard deviation and the population size within each stratum (Kish 1965). The standard formula for the Neyman allocation is:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}.$$

To allow for cases where $\frac{n_h}{N_h} > 1$, we adopt the slightly modified version of the allocation method as proposed in Cochran (1977) by setting $n_h = N_h$ in these cases.

By letting $\mathcal{A}$ be the set of strata for which $n_h < N_h$ and $\mathcal{B}$ the set of strata for which $n_h \geq N_h$, this adjusted allocation will be implemented using the pseudo code provided in the following box.

Pseudo code 3: Neyman 'Compromise' Allocation

```
Step 1. Let $(\sigma_1, \ldots, \sigma_h)$ be the vector of the population stratum standard deviations, $(N_1, \ldots, N_h)$ the vector of the population stratum sizes, $N = \sum_{h=1}^{H} N_h$ the population size, $n = \sum_{h=1}^{H} n_h$ the sample size, $n_h = n_0 \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}$ the initial optimal stratum size, $h = 1, \ldots, H$. Construct the following two macros:

    Set_A: If $N_h \geq \lfloor n_h \rfloor$ then Do:
        Set $h \in \mathcal{A}$
        $n^{\mathcal{A}} = n^{\mathcal{A}} + n_h$
        End Do
    Set_B: If $N_h < \lfloor n_h \rfloor$ then Do:
        Set $h \in \mathcal{B}$
        $n_h = N_h$
        $n^{\mathcal{B}} = n^{\mathcal{B}} + N_h$
        $m = m + 1$
        End Do

Step 2. Initialize $n = 0$, $n^{\mathcal{A}} = 0$, $n^{\mathcal{B}} = 0$, $n^{(r)} = 0$, $m = 0$ ,$n_0 \leq 3,984$ (i.e., the total sample size obtained in the uniform allocation), and $(n_1, \ldots, n_H) = (0, \ldots, 0)$.
Step 3. For $h = 1$ to $H$
        Call $Set\_A$
        Call $Set\_B$
        Next $h$
Step 4. $n^{(r)} = n_0 - n^{\mathcal{A}} - n^{\mathcal{B}}$
Step 5. Do until $\lfloor n^{(r)} \rfloor \leq 1$
        For $h \in \mathcal{A}$ then Do:
            $n_h = (n^{\mathcal{A}} + n^{(r)}) \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}$
        Next $h$
        Initialize $n^{\mathcal{A}} = 0$
        For $h \in \mathcal{A}$ then Do:
            Call $Set\_A$
            Call $Set\_B$
        Next $h$
        Do Step 4. End
End
```

In this algorithm $n^{(r)}$ is the number of units not assigned to any stratum. After performing the algorithm, the final resulting sample size is $n = 3,983$ ($n^{\mathcal{B}} = 897; n^{\mathcal{A}} = 3,086$), the number of residual units is $n^{(r)} = 1$.

3.4 A Power 'Compromise' Allocation

The Power 'Compromise' Allocation (PCA), a method proposed by Bankier (1988), is a compromise between the Neyman allocation and the CV-allocation. PCA takes the relative stratum CV as the weight for each stratum, see Berry (1974) and Page Shapiro (1985). Let $CV_h = \frac{\sigma_h}{\bar{Y}_h}$ be the stratum CV, with $\bar{Y}_h$ being the stratum population mean. The stratum size is:

$$n_h = n \frac{CV_h X_h^q}{\sum_{h=1}^{H} CV_h X_h^q}. \tag{2}$$

X_h represents some measure of size or importance for stratum h and q is a constant in

$[0, 1]$. When $q = 0$ we have the CV allocation, whereas with $q = 1$ and $X_h = N_h\bar{Y}_h$ we have the Neyman allocation. Furthermore, we consider the values for q proposed by Bankier (1988), p.176, i.e. $q = 0.3$ and $q = 0.5$. With respect to the values for X_h we will consider both $X_h = N_h\bar{Y}_h$ (i.e., the so-called CV-Neyman family) and $X_h = N_h$.

In general, this procedure can be formulated in algorithmic form as follows:

Pseudo code 4: Power Compromise Allocation

```
    Step 1. Use formula (2), define $n_h$ as in equation (1), then apply Pseudo code 3.
```

3.5 A PAUST 'Compromise' Allocation

The Proportional Allocation with Uniform Stratum Threshold (PAUST) is a compromise between the proportional allocation and the uniform allocation (Costa et al. 2004). Let $W_h = \frac{N_h}{N}$ be the stratum weight h $(h = 1, \dots, H)$. The stratum size is:

$$n_h = k(nW_h) + (1 - k)(\frac{n}{H}), \tag{3}$$

where k is a constant in $[0, 1]$. The values of k provide a family of compromise allocations. In particular, when $k = 0$ we have the uniform allocation, when $k = 1$ we have the proportional allocation. Using the notation $n_1 = (1 - k)n$ and $n_2 = kn$ (with $n = n_1 + n_2$) in the simulation we will write, for example, $PAUST_3$, in order to highlight the uniform threshold.

Formula (3) needs to be modified when $\frac{n}{H} > N_h$. Therefore in the simulation study we will use a modified algorithm with the following stratum size:

$$n_h = k(nW_h) + (1-k)(n_h^*),$$

where

$$n_h^* = \begin{cases} \frac{n - \sum_{hin\mathcal{B} N_h}}{N-m} \text{ if } h \in \mathcal{A} \\ \\ N_h \text{ if } h \in \mathcal{B}, \end{cases} .$$

and m is the number of strata in set $\mathcal{B}$.

In this case the procedure can be formulated in algorithmic form as follows:

Pseudo code 5: PAUST Compromise Allocation

```
Step 1. Use Pseudo code 1 for the uniform allocation part, then Pseudo code 2 for the proportional allocation part.
```

3.6 The ROAUST 'Compromise' Allocation

The ROAUST allocation (Chiodini et al. 2008) is a compromise between the Neyman and the uniform allocation. Let $n = \sum_h n_h$ be the desired total sample size. ROAUST first applies the uniform allocation by sampling n_1 units ($n_1 = \alpha n$ with $\alpha \in [0,1]$) so that the uniform stratum sample size becomes $n_{1h} = \frac{n_1}{H}$. The Neyman allocation is then applied to the remaining n_2 units, such that $n_2 = n - n_1$. Hence, the stratum sample size is given by:

$$n_h = n_{1h} + n_2 \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}.$$

When $\alpha = 0$ the Neyman allocation is obtained; when $\alpha = 1$ the uniform allocation is obtained.

Among all the possible values n_{1h} can assume in each stratum, in this paper two values are proposed, namely $n_{1h} = 3$ and $n_{1h} = 9$ [3]. The first is a constraint deriving from the minimum stratum size. With the latter, corresponding to $\alpha = 0.5$ a balanced stratum size is achieved by assigning equal importance to the stratum information (i.e. a fixed number of units required within each stratum) and by allocating the remaining 50% of units proportionally to the size and strata heterogeneity. This procedure can be formulated in algorithmic form as follows.

```
Pseudo code 6: ROAUST Compromise Allocation
    Step 1. Use Pseudo code 1 for the uniform allocation part, then Pseudo code 3 for the optimal allocation part.
```

3.7 The Adjusted Nonlinear Programming Allocation

In the simulation study we will use an adjusted version of the Non-Linear Programming (NLP) allocation: the Interior Point NLP (IPNLP). NLP obtains an allocation to strata that minimizes the total sample size n subject to specified tolerances on the CV of the

[3] In the simulation we will denote these two cases as $ROAUST_3$ and $ROAUST_9$, respectively.

strata and population mean estimators. The NLP allocation class can be considered essentially a general constrained and non-linear optimized allocation (Choudhry et al. 2012). IPNLP uses a Quasi-Newton Interior Point (IPQN) method. IPNLP-IPQN can efficiently solve medium size optimization problems (SAS 2010). In details, this method works as follows.

Let $\mathbf{n} = (n_1, n_2, \ldots, n_H)$ be the vector of strata sizes. With IPNLP we aim at minimizing the total sample size:

$$g(\mathbf{n}) = \sum_{h=1}^{H} n_h, \tag{4}$$

subject to some constraints. Similarly to (Choudhry et al. 2012), we use the CV of $\bar{y}_h$, i.e.:

$$CV(\bar{y}_h) = CV_h \sqrt{\frac{N_h - n_h}{n_h N_h}}, h = 1, \ldots, H,$$

with $CV_h = \frac{\sigma_h}{\bar{Y}_h}$, and the CV of $\bar{y}_{est}$, the estimated population mean:

$$CV(\bar{y}_{est}) = \sqrt{\sum_{h=1}^{H} \frac{N_h - n_h}{n_h N_h} W_h^2 \frac{\sigma_h^2}{\bar{Y}^2}}, h = 1, \ldots, H,$$

where $W_h = \frac{N_h}{N}$ is the stratum weight.

We use SAS IPNLP solver with the IPQN option to find the optimal n_h that minimizes

(4) subject to:

$$CV(\bar{y}_h) \leq CV_{0h}, h = 1, \ldots, H; \tag{5}$$

$$CV(\bar{y}_{est}) \leq CV_0; \tag{6}$$

and

$$1 \leq n_h \leq N_h, h = 1, \ldots, H, \tag{7}$$

where CV_{0h} and CV_0 are specific tolerances on the CV for the stratum sample mean $\bar{y}_h$ and the estimated population mean $\bar{y}_{est}$, respectively. Note that the constrain $n_h \geq 1$ implies a uniform threshold equal to 1.

For the ease of comparison, given $n = 3,981$, and assumed CV_0 equal to the Relative Root Mean Square Error (RRMSE) in $ROAUST_9$, that is:

$$CV_0 = RRMSE(ROAUST_9) = 0.69\%,$$

we fix at 11.60% the corresponding value for CV_{0h}.

This procedure can be formulated in algorithmic form as follows.

Pseudo code 7: IPNLP

```
Step 1. Set tolerances $CV_{0h}$ and $CV_0$.
Step 2. Use SAS IPNLP-IPQN to find the minimum $n_h$ $(h = 1, \dots, H)$ that minimizes (4) subject to (5), (6), and (7).
```

4 Simulation Engine and Results

4.1 Simulation Engine

The implementation of the simulation study performed in this work is linked to the Permanent Random Number (PRN) technique proposed by Ohlsson (1995). PRN allows to optimize the process of selection of the units in repeated sampling from the same frame, maximizing the proportion of overlapping units between the compared allocation methods.

In the context of the data we have considered, let P be a population subdivided in H strata, $N_1, \dots, N_h$ the population strata sizes, $n_1, \dots, n_h$ the sample strata sizes[4].

A general pseudo code version of the MC-SSA engine with references to the toy example works as follows.

MC-SSA engine

```
Step 1. Replicate with replacement R=1,000 times the original population subdivided in $H$ strata. Call these replicates $P_1, \dots, P_r, \dots, P_R$.
Step 2. For each $P_r$:
    1 assign distinct PRNs to each unit in each stratum of the population;
    2 sort units in each stratum in ascending order with respect to the assigned PRNs;
```

[4] In our case, $N = \sum N_h = 86,447$, $n = \sum n_h \approx 4,000$.

```
        3  select the sample units in each population stratum according to each allocation method to be compared. Sample sizes from
           each allocation method are almost the same. By doing this, in each strata and across all the considered allocation methods, the maximum number of overlapping units is guaranteed.
        4  obtain $R$ sample estimates ${}_{MC}\bar{y}_r$ of the mean number of persons employed;
        5  obtain $H$ non-empty mean stratum estimates ${}_{MC}\bar{y}_h$;
Step 3. From the estimates ${}_{MC}\bar{y}_r$ obtain the overall RRMSE.
Step 4. From the estimates ${}_{MC}\bar{y}_h$, obtain the strata RRMSEs.
```

With this simulation engine we aim at guaranteeing that the difference in the estimate quality is scarcely influenced by:

- the non-overlapping units across different allocation methods;
- the different sample sizes, because they differ slightly when implementing the allocation methods.

These two characteristics allow highlighting the *effects due to the sample allocation design* only (see Chiodini et al. 2017 for a numerical example). Monte Carlo simulations are therefore needed to verify the properties of the estimators derived from the compromise allocation methods in comparison with the estimators from other methods proposed in the literature, or usually adopted in business surveys, as analytical check seems not easy to be obtained.

4.2 Results and Discussion

In order to compare the allocation methods, we used data on persons employed as reported by the ASIA frame, and, as a measure for comparison, the relative root mean square error (RRMSE) defined by:

$$\text{RRMSE} = \sqrt{(\frac{\text{Bias}}{{}_{MC}\bar{Y}})^2} + {}_{MC}CV^2.$$

Furthermore, we used the following criterion to refer each method to the Neyman method, which is our benchmark method:

$$\frac{\text{RRMSE}(\hat{\theta}_M)}{\text{RRMSE}(\hat{\theta}_{Neyman})} + \frac{\max_h(\text{RRMSE}([\hat{\theta_h}]_M))}{\max_h(\text{RRMSE}([\hat{\theta_h}]_{Neyman}))} = \min,$$

where RRMSE($\hat{\theta}_M$) is the overall RRMSE of the estimate $\hat{\theta}$ for a given allocation method M, RRMSE$\hat{\theta}_{Neyman}$ is the overall RRMSE of the estimate $\hat{\theta}$ for the Neyman allocation method, $\max_h(\text{RRMSE}([\hat{\theta_h}]_M))$ is the maximum stratum RRMSE of the estimate $\hat{\theta}$ for a given allocation method M, and $\max_h(\text{RRMSE}([\widehat{\theta_h)}]_{Neyman})$ is the maximum stratum RRMSE of the estimate $\hat{\theta}$ for the Neyman allocation method.

The *Bias* term was estimated as $\bar{Y} - {}_{MC}\bar{Y}$, where $\bar{Y}$ was the true population mean, ${}_{MC}\bar{Y}$ the empirical mean obtained across the MC replicates, and ${}_{MC}CV = \frac{{}_{MC}\sigma}{{}_{MC}\bar{Y}}$, with ${}_{MC}\sigma$ being the empirical standard error across the replicates. We also used the relative bias with respect to ${}_{MC}\bar{Y}$:

$$\text{RB} = |\frac{\text{Bias}}{{}_{MC}\bar{Y}}|.$$

Table 3: MC-SSA simulation results for the overall population: Relative errors (1,000 replicates)

Method	Sample size	$\lvert\frac{\text{Bias}}{_{MC}\bar{Y}}\rvert$	Overall RRMSE
IPNLP	3,981	0.0001	0.0069
BANKIER ($q = 1$;$X_h = \bar{Y}_h$)	3,983	0.0008	0.0144
NEYMAN	3,983	0.0000	0.0063
ROAUST$_3$	3,981	0.0000	0.0064
ROAUST$_9$	3,981	0.0000	0.0069
UNIFORM	3,984	0.0003	0.0179
PAUST$_1$ (adj. proportional)	3,984	0.0007	0.0520
PAUST$_3$	3,983	0.0001	0.0355
PAUST$_9$	3,983	0.0006	0.0248
BANKIER ($q = 0.3$; $X_h = N_h$)	3,982	0.0000	0.0099
BANKIER ($q = 0.5$; $X_h = N_h$)	3,982	0.0000	0.0130
BANKIER ($q = 0.3$; $X_h = N_h\bar{Y}_h$)	3,982	0.0001	0.0079
BANKIER ($q = 0.5$; $X_h = N_h\bar{Y}_h$)	3,982	0.0001	0.0070

Table 3 shows the results for the overall population estimates. The Neyman compromise allocation had the lowest RRMSE, as expected. This is a well-known result under normality as it provides the highest precision for estimating a population mean, given a fixed total sample size (Chen 2011), and is still valid in our case study under non-normality since it takes into account the stratum heterogeneity and the stratum size at the same time. On the other hand, the proportional allocation and the PAUST allocation were not suitable in terms of RRMSE. The Uniform allocation behaved better than PAUST even if it obtained RRMSE values which almost doubled those of the Neyman allocation. Finally, almost all Bankier allocation's RRMSEs were generally larger than those of the Neyman allocation, whereas the IPNLP and the ROAUST allocation performed similarly.

Table 4: MC-SSA simulation results for the strata: Relative errors (1,000 replicates)

	No. of null sample strata	Max stratum $\lvert\frac{\text{Bias}}{_{MC}\bar{Y}}\rvert$	Max stratum RRMSE
IPNLP	0	0.0109	0.1978
BANKIER ($q = 1;X_h = \bar{Y}_h$)	4	0.0167	0.2181
NEYMAN	11	0.0155	0.5809
ROAUST$_3$	0	0.0155	0.2495
ROAUST$_9$	0	0.0115	0.1588
UNIFORM	0	0.0112	0.3966
PAUST$_1$ (adj. proportional)	0	0.0550	1.6587
PAUST$_3$	0	0.0292	0.6743
PAUST$_9$	0	0.0362	1.0105
BANKIER ($q = 0.5$; $X_h = N_h$)	5	0.0155	0.3184
BANKIER ($q = 0.3$; $X_h = N_h$)	5	0.0112	0.1520
BANKIER ($q = 0.3$; $X_h = N_h\bar{Y}_h$)	5	0.0103	0.1526
BANKIER ($q = 0.5$; $X_h = N_h\bar{Y}_h$)	5	0.0116	0.2495

Table 4 shows the results for the stratum estimates. Considering the maximum stratum RRMSE for the ROAUST allocation with $n_{1h} = 9$ as the reference index value (base=100), an index more than three times larger (equal to 366) is obtained for the Neyman allocation. Moreover, within the strata the Neyman allocation - but also the Bankier allocations - does not seem suitable for domain estimation as they presents strata with null sample sizes. With respects to other methods with an overall RRMSE similar to the Neyman allocation, we observed an index half time larger (equal to 157) for the ROAUST allocation with $n_{1h} = 3$ and an unexpected index a quarter larger for IPNLP (equal to 125) resulted (see Chiodini et al. 2017, for details).

We can summarize these results as follows.

1. Using the firm size as stratum variable - as is customary in European business surveys - implies a decrease in stratum size corresponding to an increase in stratum variability. Moreover, the uniform allocation method, notwithstanding its known drawbacks, performs better than the proportional (and the PAUST) allocation both at domain and overall levels. With this data structure, the proportional allocation method, although widely used by sectorial operators, is not particularly suitable. Contrary to our results, Choudhry et al. (2012) found that the proportional allocation has a better performance than the uniform allocation. The discrepancy with respect to our results can be due to a positive correlation (equal to +0.8) between the stratum size and the stratum variability. However, in their experiment they used only geographical stratification variables. Furthermore, the negative correlation between stratum size and stratum variability (we found in our results that it is -0.2) is a common characteristic of the European business surveys, as the firm size is usually requested as a stratification variable.

2. The original Neyman allocation formula cannot allocate units in some strata. A further weakness of the Neyman allocation is that it allows for $n_h > N_h$.

3. The PAUST compromise allocation performs similarly to the proportional allocation. However the proportional allocation, if not adjusted, cannot allocate units in

some strata.

4. The Bankier compromise allocation performs optimally in many cases. But also this method cannot allocate units in some strata.

5. The IPNLP method minimizes the total sample size, subject to a specified tolerance on the CVs, performs optimally. However, it can be noted that the MC-SSA simulation for the IPNLP method highlighted a range of the stratum RRMSE larger than that of the ROAUST method (see Chiodini et al. 2017, for details).

6. The ROAUST allocation performs optimally, since it allows both for optimal allocation and stratum information (i.e., by construction this method requires a number of units in each stratum).

5 Conclusion

This paper evaluated the contribution provided by allocation methods to the goodness of the estimates when performing a business survey, in particular in the case of the Italian business confidence survey. This evaluation was conducted via an extensive simulation study performed with the aim of isolating the effects due to the sample allocation design only. The choice of evaluating the allocation methods via simulation comes from a pragmatic approach, and could lead to methodological developments to overcome problems like those regarding imperfect frames and the presence of heterogeneity in the strata. We tried to find suitable 'compromise' allocation methods, i.e. allocations devoted to blend together two or more methods to improve estimation. In summary, results highlighted some positive aspects of certain methods with respect to others, together with some critical points affecting the reliability of the majority of the allocation methods considered. All in all, both the IPNLP and ROAUST allocations are the most efficient methods for the Italian business survey data structure. The former is of interest for its capability of minimizing the total sample size subject to a specific tolerance on the CVs and different constraints (e.g. $n_h \leq N_h$), whereas the latter is also interesting for its simplicity.

References

Bankier, M. D. (1988), 'Power allocations: Determining sample sizes for subnational areas', *The American Statistician* **42**(3), 174–177.

Berry, D. (1974), 'Optimal sampling schemes for estimating system reliability by testing components-i: Fixed sample size', *Journal of the American Statistical Association* **69**(346), 485–491.

Chen, H. (2011), Neyman allocation, *in* P. J. Lavrakas, ed., 'Encyclopedia of Survey Research Methods', Sage, pp. 509–510.

Chiodini, P. M., Manzi, G. & Verrecchia, F. (2008), 'Allocazione ottimale robusta con soglia uniforme di strato [robust optimal allocation with uniform stratum threshold]', *Esec Working Papers* **1**(5).

Chiodini, P., Manzi, G., Martelli, B. M. & Verrecchia, F. (2017), 'Divide, allocate et impera: Comparing allocation strategies via simulation', *DEMM Working Paper Series* **2017**(9), 1–42.

Choudhry, G. H., Rao, J. N. K. & Hidiroglou, M. A. (2012), 'On sample allocation for efficient domain estimation', *Survey Methodology* **38**(1), 23–29.

Clark, R. G., Kokic, P. & Smith, P. A. (2017), 'A comparison of two robust estimation methods for business surveys', *International Statistical Review* **85**(2), 270–289.

Cochran, W. G. (1977), *Sampling Techniques*, third edn, Wiley, New York, US.

Costa, A., Satorra, A. & Ventura, E. (2004), 'Improve both domain and total area estimation by composition', *SORT* **28**(1), 69–86.

Ellison, H. & Elvers, E. (2001), Cut-off sampling and estimation, statistics canada international symposium series, *in* 'Proceedings of Statistics Canada Symposium'.

Er, Ş. (2012), 'Comparison of the efficiency of the various algorithms in stratified sampling when the initial solutions are determined with geometric method', *International Journal of Statistics and Applications* **2**(1), 1–10.

European Union Commission (2003), 'Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises', *Official Journal of the European Union* **46**, 36–41.

Hidiroglou, M. A. & Kozak, M. (2018), 'Stratification of skewed populations: A comparison of optimisation-based versus approximate methods', *International Statistical Review* **86**(1), 87–105.

Hidiroglou, M. A. & Srinath, K. (1993), 'Problems associated with designing subannual business surveys', *Journal of Business & Economic Statistics* **11**(4), 397–405.

ISTAT (2007), 'Struttura e dimensione delle imprese. Archivio statistico delle imprese attive (ASIA) [Structure and dimension of enterprises. Statistical archive of enterprises]', Database.

Keto, M. & Pahkinen, E. (2017), 'Sample allocation for efficient model-based small area estimation', *Survey Methodology* **43**(1), 93–106.

Khan, M. G. M., Reddy, K. G. & Rao, D. K. (2015), ‘Designing stratified sampling in economic and business surveys’, *Journal of Applied Statistics* **42**(10), 2080–2099.

Kish, L. (1965), *Survey Sampling*, Wiley, New York, US.

Kozak, M. (2006), ‘Multivariate sample allocation: Application of random search method’, *Statistics in Transition* **7**(4), 889–900.

Kozak, M. (2014), ‘Comparison of random search method and genetic algorithm for stratification’, *Communications in Statistics–Simulation and Computation* **43**(2), 249–253.

Kozak, M., Verma, M. R. & Zieliński, A. (2007), ‘Modern approach to optimum stratification: Review and perspectives’, *Statistics in Transition - New Series* **8**(2), 223–250.

Molefe, W. B. & Clark, R. G. (2015), ‘Model-assisted optimal allocation for planned domains using composite estimation’, *Survey Methodology* **41**(2), 377–387.

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**(4), 558–625.

Ohlsson, E. (1995), Co-ordination of samples using permanent random numbers, *in* B. G.

Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge & P. S. Kott, eds, 'Business Survey Methods', Wiley, New York, US, pp. 153–169.

Page Shapiro, C. (1985), 'Allocation schemes for estimating the product of positive parameters', *Journal of the American Statistical Association* **80**(390), 449–454.

Särndal, C., Swensson, B. & Wretman, J. (2013), *Model Assisted Survey Sampling*, Springer, New York, US.

SAS (2010), *SAS/OR® 9.22 User' s Guide: Mathematical Programming*.

Smith, P. A. & James, G. G. (2017), 'Changing industrial classification to sic (2007) at the uk office for national statistics', *Journal of Official Statistics* **33**(1), 223–247.

Smith, P., Pont, M. & Jones, T. (2003), 'Developments in business survey methodology in the Office for National Statistics, 1994-2000', *The Statistician* **52**(3), 257–295.