Between Theoretical and Applied Approach: Which Compromise for Unit Allocation in Business Surveys?[1]

Paola M. Chiodini, Bianca M. Martelli, Giancarlo Manzi, Flavio Verrecchia

Abstract Neyman's algorithm for the allocation of sample units in business sampling can result unsatisfactory in domain analysis with imperfect frames and sectorial and/or regional data. Improved estimates can be obtained using stratified estimators combined with an optimal unit allocation. We achieve this outcome by an interdisciplinary approach which leads to a methodological improvement. Starting from Martini's approach which considers an empirical view of the statistical analysis, we propose the *Robust Optimal Allocation with Uniform Stratum Threshold* (ROAUST) class of stratified estimators and prove their reliability by using a simulation approach inspired by Magagnoli's work on this issue. In particular, contrary to Neyman's stratified estimator with optimal allocation and stratum threshold, our class guarantees better domain representativeness.

Key words: Business survey, Imperfect frames, Heterogeneous strata, Simulation methods, Computational techniques, Sample design and estimation, Register based statistics

---

[1] The present paper is financially supported by ESeC.

[2] Paola M. Chiodini, Department of Quantitative Methods for Business and Economic Sciences, University of Milano-Bicocca; e-mail: paola.chiodini@unimib.it
Bianca M. Martelli, ISAE; e-mail: b.martelli@isae.it
Giancarlo Manzi, Department of Economics, Business and Statistics, Università degli Studi di Milano; e-mail: giancarlo.manzi@unimi.it
Flavio Verrecchia, ESeC; e-mail: flavio.verrecchia@gmail.com

1 Introduction and Aims

Among several aspects that are part of the sampling design process, this paper will refer to planned domains and focus on facets related to strata allocation and tools for validating their efficiency in business surveys. The need of stratum representativeness from one side and the optimum allocation principle from the other are often in conflict because of the strata definition (based on administrative settings and economic classification). These facts make impossible to subdivide a population into homogeneous strata so as to optimize a survey plan (e.g. by maximizing the precision of estimates and reducing the problem of empty strata which mainly affects the largest firms). This paper presents and discusses a joint research motivated by two different survey experiences which shared similar empirical bounds, and proposes the validation of a possible solution for the above issues.

The ISAE (Institute for Studies and Economic Analysis) is the Italian referent of the Joint Harmonised Business and Consumer Survey (BCS) program of the European Commission [10, 11]. Over the years, ISAE has been developing, for the manufacturing sector, a stratified sampling (by sectors, and later on by regions, and size), firstly built in 1995 according to the Neyman's allocation technique [9, 18]. However, the allocation of units to strata, together with some operative constraints, implied some empirical adjustments that could hardly rely on strong theoretical support. Among these bounds, a growing importance has to be attributed to the need of detailed sectorial (domain) information, that is to the strata representativeness which is not often guaranteed.

ESeC (Economic Statistics e-Center) experienced similar difficulties in Information Technology (IT) sector sampling and had previously proposed a class of allocation techniques (ROAUST – [1, 2]) for dealing with the problems arising from the Neyman's allocation method. Simulation as a validation tool and a methodological approach suggested by Magagnoli were used [3].

This common interest led the authors to form a research group and to apply their competences to the revision of the ISAE Manufacturing Tendency Survey sample based on the new Nace (Statistical Classification of Economic Activities in the European Community) Rev.2 sectorial classification. The first findings of this research group has been already presented in a conference in Poland [5]: several allocation methods were compared and their efficiency evaluated by means of a unique simulation experiment.

The reminder of this paper is organised as follows. In Section 2, two case studies are presented as examples of imperfect frames, namely the Assinform (Italian Association of Information and Communication Technology Companies) IT sector survey and the ISAE Manufacturing Tendency survey; in Section 3 different approaches and methods are introduced and the simulation technique is described; in Section 4 the main findings are discussed; Section 5 concludes the paper.

2 Imperfect Frames and Cluster Heterogeneity

A problem of great interest in sampling theory is that of imperfect frames. In practice, it is difficult to find out archives (frames) with no errors such as the presence of

incomplete or not updated information. In business surveys this problem is even more cogent. For estimation purposes, Rao [21] classifies frame problems in under-coverage, over-coverage and multiplicity. Under-coverage problems arise when some of the population units are not included in the population list for some reason. Over-coverage problems arise when some units not part of the target population are mistakenly included in the sample. Multiplicity problems arise when a target unit is included k times in the frame. A second problem relates to the between-strata heterogeneity (in terms of population size), as the strata come from pre-defined administrative designs and classifications of economic activities. A further problem is related to the within-strata variability (in terms of business size, usually measured in terms of workforce or turnover), linked to the presence of industrial polarization.

In the ISAE survey, only the heterogeneity issue in terms of population size and stratum variance arises, as the survey list is based on the Archive of Italian Active Firms (ASIA), whereas in the Assinform survey also the problem of imperfect frames is present as its list is based on the Chamber of Commerce data base.

The IT Sector

The economic analysis of the IT sector requires the detailed description of the survey perimeter in terms of classification of economic activities. To this purpose it is possible to use standard classifications adopted by national and international offices of statistics (e.g. the NACE classification). The analysis of the IT sector was performed on the results from the Assinform survey for the Italian IT sector [1]. The complete list of IT enterprises as of 31/12/2007 was provided by the Chamber of Commerce of Milan. The survey perimeter is given by 44,700 enterprises which form the list of the *Computer software and related services* class. The sampling plan is stratified by regions (and some smaller areas) and by enterprises legal forms. Possible non-sampling errors are (i) the presence of a relevant time gap between the survey and the use of the sampled data, (ii) a possible presence of missing data, (iii) the presence of wrong data in the storage process, and (iv) a wide range of professional skills claimed by the respondents. As for the survey, some of its main information features are summarized in the following: number of strata: 48; overall sample size: 996; data collection mode: telephone calls with the CAWI (about 73%) and CATI systems; questionnaire: semi-structured. For inferential purposes, weights proportional to strata as they are in the list of the Chamber of Commerce were used. Sample data were finally brought back to the universe. By applying the different allocation methods, heterogeneity among strata is manifest in terms of sizes and variability (see Table 1). For example, Milano and its province have the highest frequency of legal forms (14% in terms of total number of enterprises). On the other hand, this percentage for the Valle d'Aosta region is only 0.2%. Hence, if the *Uniform Allocation (UA)* is not representative for those strata with a high number of cases, then paradoxically also the *Proportional Allocation (PA)* is not representative, since the reduced sample size for some strata implies non-significant results (see, for example, the Valle d'Aosta and Molise regions). The *Optimal Allocation (OA)* is not at all representative (for some sample cells) due to the presence of some outliers (see, for example, the Sicilia region). The *Robust Optimal Allocation (ROA)* is useful both when the stratification is undertaken *ex-ante* (e.g. avoiding strata with no information) and when using a proxy variable for the stratum variability quantification in order to improve the precision of the estimates. In enterprise surveys on the IT sector the number of employees is useless since, although it can be retrieved from the Official National Register of

Enterprises, missing and wrong data are always present on a very huge scale. This is due to the fact that in Italy it is not compulsory to register this type of data into the official register of enterprises. However, if the distribution is α-Winsorized with regards to the size of the company, then this number can be regarded as a proxy variable for the estimation of the stratum variance. However, also the ROA does not ascertain a satisfactory representativeness, even if the allocation process is improved since the estimates of the stratum variability are better [23].

Table 1: Allocations of sample units

Regions	***Population***		***Sample allocation units (%)***			
	CV_{norm} (%)	***(#)***	***UA***	***PA***	***OA***	***ROA***
Valle d'Aosta	87%	106	4,2%	0,2%	0,8%	0,2%
Milano	17%	6280	4,2%	14,0%	21,7%	44,6%
Molise	20%	134	4,2%	0,3%	0,0%	0,1%
Sicilia	89%	1718	4,2%	3,8%	52,7%	3,3%
…	…	…	…	…	…	…

Source: [2].

The ISAE Business Tendency Survey

Although conceived in the sixties of the last century mainly as purposive panel among managers ("expert witnesses"), ISAE has been developing its Business Tendency (BT) Survey sample over the years in order to better match the methodological developments of sampling theory [20, 22 and 17]. The recent availability of the business frame ASIA provided by the Italian National Institute of Statistics (ISTAT), classified according to the Nace Rev.2 classification, let ISAE face up to the necessity of updating the sample design and consequently to revise the strata definition. In selecting the frame for the survey, a lower cut off is considered by excluding firms with less than 10 employees. Therefore the selected frame comprises details for just over 85 thousand of enterprises (about 20% of the total), accounting for about 90% of economic activity of the Italian manufacturing industry (i.e. in terms of firms turnover). The revised BT sample maintains the stratified design and the strata are as usually defined according to three variables: firm size (in terms of employees), economic classification and geographical areas. The classes, according to Eurostat definition, refers to 3 types of enterprises: *small enterprises* (10-49 employees), *medium-sized enterprises* (50-249 employees) and *large enterprises* (with at least 250 persons employed). The 19 economic sectors mainly reflect the Nace Rev.2 two digits classification with some few further grouping. The geographical detail refers to Nuts-1 (Nomenclature of Territorial Units for Statistics) classification that allows for reducing the number of the strata as compared to the Nuts-2 classification that was the usual classification in the past.

Table 2: Enterprises by stratum (Nuts1, firm size, economic sectors), Italy, 2006 (units)

Nace	**North-West**			**North-East**			**Centre**			**South – Islands**		
	10|-|49	**50|-|249**	**250 |-**	**10|-|49**	**50|-|249**	**250 |-**	**10|-|49**	**50|-|249**	**250 |-**	**10|-|49**	**50|-|249**	**250 |-**
19.	32	9	7	16	5	.	19	7	4	78	7	4
24.	652	208	41	275	119	12	160	35	6	147	34	4
…	…	…	…	…	…	…	…	…	…	…	…	…

Source: [5]. **Notes**: '.' Missing value; '-' less than 3 units. 19. Manufacture coke and refined petroleum products. 24. Manufacture basic metals.

This choice allows to nearly completely avoid the occurrence of empty strata in the frame and, although the within variance increases, this occurrence has not any substantial impact on sample designs with regards to optimality [5]. As already pointed out, these choices mainly derive from administrative settings (Table 2), and often do not respect the statistical and economic principles of stratum definition [12]. The allocation was performed according to the univariate Neyman x-optimal allocation, based on workforce variance, separately applied for each Nuts-1 area. A 5% cut off on variances has been applied and a requirement of a sampling fraction not higher than 50% was set.

3 Approaches and Methods

In this research, we want to combine two approaches: the one of Martini and the one of Magagnoli. It is from these two inheritances that the major contributions of this paper occur. From empirical observation (Section 2) new intuitions and formal representations follow. These first formal representations are verified by adapting the simulation approach to the context of sampling from finite populations.

Martini's Approach

This is an approach which originates from the observation of real phenomena. In Martini's words [19]: *Applied Statistics is the privileged place where the dialogue on changing reality between who speaks the language of economic and social sciences and who speaks the language of rigorous procedures, leading to numerical summaries from the reality itself.*

Magagnoli's Approach

In scientific research the empirical evidence is frequently invoked for supporting research hypotheses and developed theories. When the availability of real data is scarce the empirical evidence is supported by computational algorithms to perform simulation. The advantage of this method, through which data are random generated, relies upon (i) not wasting time and resources in finding reliable data for empirical validation, (ii) infinite (at least in theory) replications can be obtained with no additional costs, (iii) different scenarios - from applied to theoretical ones - can be evaluated and finally (iv) the robustness of the proposed method can be checked asymptotically.

This way of working is especially motivated in some particular fields such as quality control and system reliability, where the verification systems frequently lead to the elimination of product units or is time-consuming and entails unaffordable costs [8, 13, 14, 15 and 16].

The methodology here proposed can be used when the observation of reality induces to propose a new theoretical method (in our case the ROAUST method) which necessitate some empirical validation. More importantly, in survey sampling from finite populations, simulations allow for the checking of (i) the efficiency of estimators (at a lower level of costs and resources) even with small sampling fractions, (ii) the performance of the unit allocation methods (i.e. in case of stratified sampling) and (iii) the efficacy of the auxiliary information introduced in the sample. These are all issues

with no trivial solutions and a purposely built simulation is needed.

The Neyman's Domain Algorithm: the OAUST Class

In the context of the domain analysis the most important result is the *(R)OAUST - (Robust) Optimum Allocation with Uniform Stratum Threshold* [2]. We first define the sample size n, apply the *Uniform Allocation* by sampling n_1 units ($n_1=an$, with $a\in[0,1]$) and then the *Neyman Allocation* for the remaining n_2 units, such that $n_2=n-n_1$:

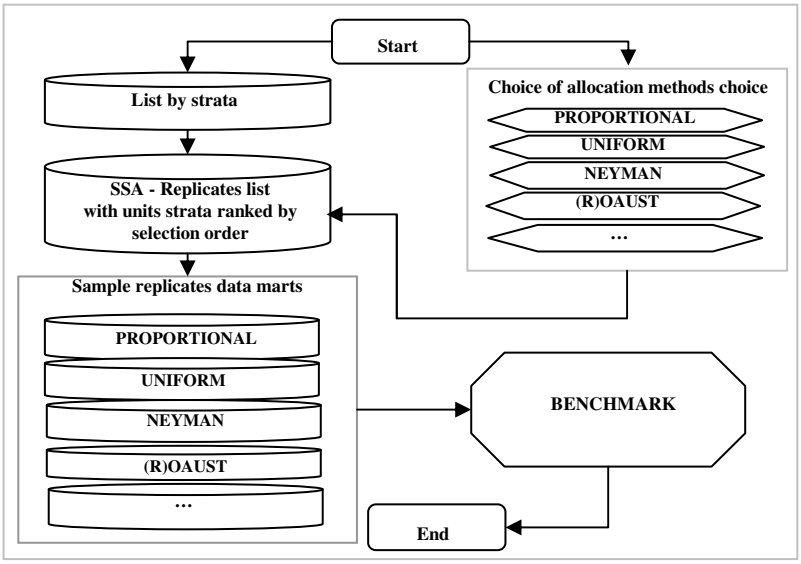
$$n_s=\frac{n_1}{S}+n_2\frac{N_s\sigma_s}{\sum_{s=1}^{S}N_s\sigma_s} \qquad [1]$$

In [1], n_s, N_s and σ_s are respectively the sample size, the number of units, the standard deviation in the s-th stratum, $s=1,\ldots,S$. When $a=0$ then $n_2=n$ (obtaining OA); when $a=1$ then $n_1=n$ (obtaining UA). Without loss of generality, with small sampling fractions, $a=0.5$

Sequential Selection-Allocation Computational Method

In order to empirically evaluate the performance of the various sampling allocation methods, simulation techniques are required. However, a computationally feasible general simulation method is hard to establish, especially when methods need to be compared *simultaneously*, that is when the sampling experiment has to be performed in a unique way, by separating the selection, the allocation and the inferential processes. This can be achieved via a *Sequential Selection-Allocation* (SSA) by constructing a new labelled *list* where population units are re-labelled within each stratum according to their selection rank after performing a Sampling WithOut Replacement (SWOR) of size equal to the stratum size [3, 6 and 7]. Then this process is replicated n times. From this new labelled population, all the allocation algorithms can be performed (Figure 1).

Figure 1: Flow chart of the allocation application



Source: [6].

4 Discussion and Conclusions

In the last sentence of his paper in the posthumous book in his honour, Marco Martini explains the sense of the interdisciplinary approach he proposed: *Statistics can not be used in reality if statisticians are not at the same time economists and sociologists and, above all, are not driven by solving problems from the entities which form the socioeconomic universes in their work.*

Sector market surveys often require solutions which are not extensively available in literature. In addition to the usual sample bias connected to the sample lists and the information retrieved via questionnaires, one has to deal with the estimation of population variables which are not normally distributed.

The ISAE and ESeC proposals come from observing the reality with a *substantial* approach and lead to a methodological development to deal with problems like imperfect frames and the heterogeneity of strata. For instance, from the research on IT sector a proposal comes: the (R)OAUST class can be considered as a Neyman's Domain Allocation, since it allow an optimal allocation and the stratum representativeness. However, the validation of this proposal is given by computational and *formal* statistical solutions. From the first findings of our simulation (see Table 3) the (R)OAUST method is more efficient than other methods and at the same time provides an overall population |RTE| similar to that of Neyman's algorithm [5].

Table 3: Maximum errors among strata (Total, 500 replicates)

Errors	***OA***	***OAUST***[a]	***UA***	***PA***
Max of stratum relative |Biases|	0.0315	0.0141	0.0226	0.1033
Max of stratum CVs	0.5659	0.1547	0.4038	1.6629
Max of stratum relative |TEs|	0.5974	0.1622	0.4148	1.6696
Overall population |RTEs|	0.0064	0.0072	0.0183	0.0562

Source: [5]. **Note**: OAUST with a = ~50%.

Innovative solutions can be brought forward by interdisciplinarity and multiple competence, especially in facing practical problems and when answers to problems are not *tout court* available in literature.

Acknowledgements. The authors are grateful to the participants of "The SSBS08 - Satellite RSS conference" (August 26-29, 2008 - Southampton, UK) and of "The 6th Conference on Survey Sampling in Economic and Social Research" (September 21-22, 2009 - Katowice, Poland) for their advice, comments and suggestions to previous papers on these topics. The authors wish to thank Professor Magagnoli for his advice and an anonymous referee whose comments served to substantially improve this paper.

References

1. Assinform (2009) Il settore IT in Italia, Promobit Srl, Milano, pp 1-3.
2. Chiodini P.M., Manzi G., Verrecchia F. (2008a) Allocazione ottimale robusta con soglia uniforme di strato ESeC. [Working paper: ESeC_WP005P_V20080912].

3. Chiodini P.M., Facchinetti S., Manzi G., Verrecchia F. (2008b) To be necessary or to be sufficient? The ratio estimator in enterprise market research. WP forthcoming.
4. Chiodini P.M., Facchinetti S., Manzi G., Nai Ruscone M., Verrecchia F. (2008c) Metodi e applicazioni per l'inferenza da popolazione finita: imprese e campionamento stratificato a selezione ordinata, ESeC. [Working paper: ESeC_WP004P_V20080714].
5. Chiodini P.M., Lima R., Manzi G., Martelli B.M., Verrecchia F. (2009a) Criticalities in Applying the Neyman's Optimality in Business Surveys: a Comparison of Selected Allocation Methods, in: The 6th Conference on Survey Sampling in Economic and Social Research (September 21-22, 2009), Katowice, Poland.
6. Chiodini P.M., Lima R., Manzi G., Martelli B.M., Verrecchia F. (2009b) On computational aspects of units selection for simulation on allocation methods. WP forthcoming.
7. Chiodini P.M., Lima R., Manzi G., Martelli B.M., Verrecchia F. (2009c) Strata optimization vs allocation methods. WP forthcoming.
8. Chiodini P.M., Magagnoli U. (2000), *Indici di capacità di processo. Modelli e procedure inferenziali: una rassegna e qualche comparazione statistica,* in "Valutazione della qualità e customer satisfaction: il ruolo della statistica", Vita e Pensiero, Milano, pp. 147-167.
9. Cochran W.G. (1977), *Sampling Techniques*, John Wiley & Sons, New York.
10. EC (2006), The Joint Harmonised EU Programme of Business and Consumer Surveys, *European Economy*, *Special Report* No. 5, Bruxelles.
11. EC (2007), The Joint Harmonised EU Programme of Business and Consumer Surveys User Guide (updated 4 July 2007), Bruxelles.
12. Kozac, M., Verma, M.R., Zieliński A., (2007), Modern Approach to Optimum Stratification: Review and Perspectives, *Statistics in Transition-new series*, vol. 8 (2), pp. 223-250, August.
13. Magagnoli U., Chiodini P.M. (2002), *Su una procedura iterativa per la stima di una funzione di regressione mediante il criterio dei minimi quadrati ponderati,* in "Studi in onore di Angelo Zanella" a cura di B.V Frosini, U. Magagnali, G. Boari, Vita e Pensiero, Milano, pp. 423-438.
14. Magagnoli U., Chiodini P.M. (2006), *Productive Systems Process Monitoring through Process Capability Indices*, Statistica Applicata, vol 18 (2), pp.215-230
15. Magagnoli U., Chiodini P.M. (2007a), *On Some Internal Auditing Procedures to Verify the Operating Risk Due to Accountancy Errors*, lavoro presentato in forma di Poster, Convegno Intermedio SIS (6 - 8 Giugno 2007) Isola di San Servolo, Venezia, pp. 553-554.
16. Magagnoli U., Chiodini P.M. (2007b), *Unilateral Hypothesis Tests of Efficiency Functions with Heteroscedastic Errors: an Iterative Procedure*, lavoro presentato in forma di Poster, Convegno SIS CLADAG 2007, 12-14 settembre 2007, Macerata, pp. 665-668.
17. Malgarini M., Margani P., Martelli B.M. (2005) New design of the ISAE Manufacturing Survey, *Journal of Business Cycle Measurements and Analysis*, 1, 125-142, OECD, Paris.
18. Martelli B.M. (1998), Le inchieste congiunturali dell'ISCO: aspetti metodologici" in "Le inchieste dell'ISCO come strumento di analisi della congiuntura economica", *Rassegna di lavori dell'ISCO*, Anno XV, n. 3, chap.1.
19. Martini M. (2004), La Statistica. In: Studi in Ricordo di Martini, Giuffrè, Milano, pp. 1-10.
20. OECD (2003) "Business Tendency Survey: A Handbook", Paris.
21. Rao, J.N.K. (2005) On Measuring the Quality of Survey Estimates, *International Statistical Review* (2005), 73, 2, 241–244.
22. Särndal C., Swensson B., Wretman J. (2003) *Model Assisted Survey Sampling*, Springer.
23. Verrecchia F., Chiodini P.M., Coin D., Facchinetti S., Nai Ruscone M. (2008), Bayesian Approach for Nonresponse, in: SSBS08 - Satellite RSS 2008 conference, Southampton, UK (26-29 August 2008).