UNIVERSITA' DEGLI STUDI DI TRENTO - DIPARTIMENTO DI ECONOMIA

A TREE-BASED APPROACH TO FORMING STRATA IN MULTIPURPOSE BUSINESS SURVEYS

Roberto Benedetti
Giuseppe Espa
Giovanni Lafratta

Discussion Paper No. 5, 2005

The Discussion Paper series provides a means for circulating preliminary research results by staff of or visitors to the Department. Its purpose is to stimulate discussion prior to the publication of papers.

Requests for copies of Discussion Papers and address changes should be sent to:

Dott. Stefano Comino
Dipartimento di Economia
Università degli Studi
Via Inama 5
38100 TRENTO ITALY

A Tree-Based Approach to Forming Strata in Multipurpose Business Surveys

Roberto Benedetti
"G. d'Annunzio" University,
Department of Business, Statistical, Technological and Environmental Sciences (DASTA),
Viale Pindaro 42, Pescara, IT-65127.

Giuseppe Espa
University of Trento, Department of Economics,
Via Inama, 5 - Trento, IT-38100.

Giovanni Lafratta
"G. d'Annunzio" University,
Department of Business, Statistical, Technological and Environmental Sciences (DASTA),
Viale Pindaro 42, Pescara, IT-65127.

March 15, 2005

Author's footnote: Roberto Benedetti is Professor (email: benedett@unich.it), and Giovanni Lafratta is Assistant Professor (email: giovanni.lafratta@unich.it), Department of Business, Statistical, Technological and Environmental Sciences (DASTA), "G. d'Annunzio" University, Viale Pindaro 42, Pescara, IT-65127. Giuseppe Espa is Professor (email: giuseppe.espa@economia.unitn.it), Department of Economics, University of Trento, Via Inama 5, Trento, IT-38100.

Abstract: The design of a stratified sample from a finite population deals with two main issues: the definition of a rule to partition the population, and the allocation of sampling units in the selected strata. This article examines a tree-based strategy which plans to solve jointly these issues when the survey is multipurpose and multivariate information, quantitative or qualitative, is available. Strata are formed through a scissorial algorithm that selects finer and finer partitions by minimizing, at each step, the sample allocation required to achieve the precision levels set for each surveyed variable. In this way, large numbers of constraints can be satisfied without drastically increasing the sample size, and also without discarding variables selected for stratification or diminishing the number of their class intervals. Furthermore, the algorithm tends to not define empty or almost empty strata, so avoiding the need for ex post strata aggregations. The procedure was applied to redesign the Italian Farm Structure Survey. The results indicate that the gain in efficiency held using our strategy is nontrivial. For a given sample size, this procedure achieves the required precision by exploiting a number of strata which is usually a very small fraction of the number of strata available when combining all possible classes from any of the covariates.

Key Words: Multivariate stratification, Optimal sample allocation, Farm Structure Survey, Sample design.

1. INTRODUCTION

Many business surveys employ stratified sampling procedures in which simple random sampling without replacement is executed within each stratum (see, e.g., Sigman and Monsour 1995, and, for farm surveys, Vogel 1995). Usually the list frame from which units are selected is set up using administrative or census information, represented by a rich data base of auxiliary variables, each of which can be potentially exploited to form strata. Furthermore, such surveys are often also multipurpose, and given precision levels must be achieved in estimating multiple variables under study.

The goal of satisfying such a large number of constraints without drastically increasing the sample size is commonly considered as strictly related to the choice of the number of stratifying variables and of their class intervals (Kish and Anderson 1978). This is due to the well known fact that finer partitions of the population introduce more information useful for the reduction of estimation variances, but, on the other hand, their application implies higher risks for each unit to pass from one stratum to the other, and to produce – because of non-responses – empty or almost empty strata, so that, when computing estimates, some strata collapsing procedure have to be introduced.

Let us indicate as the *atomised* stratification that one obtained forming strata by combination of all possible classes from any of the covariates in use. If the corresponding number of such basic strata, or atoms, exceeds a given threshold imposed by practical restrictions, it seems unavoidable to redesign the survey selecting a smaller number of stratifying variables or fewer classes available from each of them. Notwithstanding, it can be noted that another way of obviating such an unsatisfactory situation can be based on the following argument: the atomised stratification can be really intepreted as a starting solution to the problem of strata formation, since, besides the condition of no stratification and the one making use of the atomised one, there exists a full range of opportunities to select a stratification whose subpopulations can be obtained as unions of atoms.

Our proposal is to accomplish this selection through the definition of a tree-based stratified design. We form strata by means of a scissorial algorithm that selects finer and finer partitions by minimizing, at each step, the sample allocation required to achieve the precision levels set for each surveyed variable. The procedure is sequential, and determines a path from the null stratification, i.e. that one whose single stratum matches the population, to the atomised one. At each step, we select what variable is to be used to define the new, more disaggregated partition: every stratum in the current partition is splitted on any covariate, using in turn all of its available classes, and the one that better decreases the global allocation size is selected.

Bloch and Segal (1989) discussed the application of classification tree methods (see, e.g., Breiman, Friedman, Olshen and Stone 1984) to strata formation, but their focus was mainly on strata interpretation about the relationships between the covariates and a unique outcome variable. Instead, our rules to partition the population are directly oriented to the optimal allocation of sampling units in the selected strata. Unfortunately, classical methods of optimal stratification (Dalenius and Hodges 1959; Ekman 1959; Hess, Sethi and Balakrishnan 1966; Singh 1971; Lavallee and Hidiroglou 1988; Hedlin 2000) pertain to the univariate problem of estimating the total (or the mean) of a study variable, given a quantitative stratifying covariate divided in classes, by minimizing the variance of the standard expansion estimator, where the number of strata and the sample size are taken as given. The solutions proposed in this literature are, as a consequence, of poor practical value if the survey is multipurpose and information on multiple covariates is available. In such a context, methods to satisfy a large number of constraints on errors when minimizing the sample size were proposed by Bethel (1985, 1989) and Chromy (1987). Valliant and Gentle (1997) also approached the problem for two-stage sampling frameworks. For a given stratification, we choose to apply the Bethel's allocation rule and henceforth the procedure selects subsequent partitions by minimizing the survey cost function corresponding to the stratifications consisting of the currently unsplitted strata and of the available splitted substrata.

The paper is organized as follows. Section 2 introduces the procedure we propose for the computation of stratification trees. We thoroughly describe the algorithm used to generate the sequence of stratifications, and we show how it can be represented as a classification tree. Stopping criteria are also discussed to determine how they can affect the optimal number of strata. In Section 3 we examine how a stratification tree can be exploited to design the European Community survey on the structure of agricultural holdings, also known as Farm Structure Survey (FSS). We illustrate our stratification technique identifying a tree-based set of strata and allocations using a basic set of atoms defined by means of multivariate information collected during the fifth Agricultural General Census held in Italy in the year 2000. Finally, Section 4 is devoted to some concluding remarks, focusing on issues regarding the practice of forming strata by trees and discussing how the procedure can be used to better manage multipurpose surveys based on stratified designs.

2. A PROCEDURE TO GENERATE MULTIVARIATE STRATIFICATION TREES

Consider a finite population P of N units, on which variables $Y_1, \ldots, Y_g, \ldots, Y_G$ are to be surveyed to estimate their totals using a stratification on P, i.e. a collection $\mathcal{S}$ of $|\mathcal{S}|$ nonempty subpopulations partitioning P. Our problem is how to select $\mathcal{S}$ in order to minimize the corresponding overall sample allocation $n_{\mathcal{S}}$ in a way such that, for $g = 1, \ldots, G$, the coefficient of variation (CV) corresponding to the g-th variate of interest is not greater than the desired level of precision, say $\varepsilon_g > 0$.

For a given $\mathcal{S}$, such minimization is executed by computing the Bethel's (1985) sample allocation rule. More thoroughly, let us indicate by n_h, $h = 1, \ldots, |\mathcal{S}|$, the sample allocation in stratum h, and define

$x_h = 1/n_h$ if $n_h > 0$, and $x_h = +\infty$ otherwise. The global survey cost corresponding to $\mathcal{S}$ can thus be expressed as

$$f\left(x_1, \ldots, x_{|\mathcal{S}|}\right) = c_{\mathcal{S}} + \sum\nolimits_{h=1}^{|\mathcal{S}|} \frac{c_h}{x_h},$$

where $c_{\mathcal{S}}$ is a fixed cost independent from $\mathbf{x}_{\mathcal{S}} = \left(x_1, \ldots, x_{|\mathcal{S}|}\right)'$, and c_h represents the cost to sample one unit in stratum h. To take into account the G constraints on the required precision, let us consider the following quantities, referred to as the standardized precision units,

$$u_{h,g} = N_h^2 \; v_{h,g}^2 / \left(t_g^2 \left(\varepsilon_g^2 - \sum\nolimits_{h=1}^{|\mathcal{S}|} N_h \; v_{h,g}^2 \right) \right),$$

where t_g is the total in P of the g-th response variable, N_h is the size of the h-th stratum of $\mathcal{S}$, and $v_{h,g}^2$ is the variance of Y_g in stratum h. Using these units, the problem of optimal allocation for $\mathcal{S}$ can be expressed as follows:

$$\begin{array}{lll} \text{min} & f\left(\mathbf{x}_{\mathcal{S}}\right) & \\ \text{sub} & \sum\nolimits_{h=1}^{|\mathcal{S}|} u_{h,g} \; x_h < 1, & g = 1, \ldots, G, \\ & x_h > 0, & h = 1, \ldots, |\mathcal{S}|. \end{array}$$

Bethel (1985) derived the solution to such problem, say x_h^*, $h = 1, \ldots, |\mathcal{S}|$, as follows:

$$x_h^* = \begin{cases} \sqrt{c_h} \Big/ \left(\sqrt{\sum\nolimits_{g=1}^{G} \alpha_g^* \; u_{h,g}} \sum\nolimits_{l=1}^{|\mathcal{S}|} \sqrt{c_l \sum\nolimits_{g=1}^{G} \alpha_g^* \; u_{l,g}} \right) & \text{if } \sum\nolimits_{g=1}^{G} \alpha_g^* \; u_{h,g} > 0, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\alpha_g^* = \lambda_g \Big/ \sum\nolimits_{g=1}^{G} \lambda_g$, and λ_g is the Lagrangian multiplier of the constraint on the maximum error allowed estimating the g-th surveyed variable. The corresponding global optimal allocation is thus given by setting $n_{\mathcal{S}} = \sum\nolimits_{h=1}^{|\mathcal{S}|} 1 / x_h^*$.

Let us now assume that total estimates and their variances are available for any of a given set of $M > 1$ basic strata $A_1, \ldots, A_m, \ldots, A_M$, so that we can rely on two $M \times G$ matrices, respectively of totals $\mathbf{T} = \left(t_{m,g}\right)$ and estimation variances $\mathbf{V} = \left(v_{m,g}^2\right)$. The definition of such strata, which in the sequel will be referred to as *atoms,* is based on a set of covariates $X_1, \ldots, X_k, \ldots, X_K$ as follows. Let $x_{i,k}$ is the value of X_k measured on unit $i \in P$, and consider the set of distinct values observed for X_k in P, $\Xi_k = \{x \in \mathbb{R} : \exists i \in P : x = x_{i,k}\}$. We build $M = \prod_{k=1}^{K} |\Xi_k|$ atoms, one for every vector $(a_{m,1}, \ldots, a_{m,K})$ in the Cartesian product $\Xi = \otimes_{k=1}^{K} \Xi_k$, by setting for $m = 1, \ldots, M$

$$A_m = \bigcap_{k=1}^{K} A_{m,k},$$

where $A_{m,k} = \{i \in P : x_{i,k} = a_{m,k}\}$.

The procedure we propose generates a sequence of stratifications which can be represented as a classification tree. Define the *level* of a given node ν in the tree as the number of arcs in the (unique) chain connecting node ν to the root node, and let us indicate with $r_l = l + 1$ the number of nodes sharing the same level l. At each level $l \geq 0$ the procedure determines a class $\mathcal{F}_l$ of r_l nonempty subpopulations in which P can be partitioned, putting them in a one-to-one correspondence with the nodes of level l. The strata in $\mathcal{F}_l$ are all candidated to be splitted on any given covariate X_k, and, following Bethel (1989), the sample allocation is computed which optimally minimizes the survey cost function for the stratification consisting of the unsplitted strata in $\mathcal{F}_l$ and the two substrata which define the current split. The best split at level l is identified as the most favorable in terms of decreasing sample allocation, with respect to that characterizing $\mathcal{F}_l$, than any other possible split on any of the covariates in use. The optimal allocation corresponding to the stratification defined by such best split, indicated by $n_{b,l+1}$, is taken as the optimal sample size at level $l+1$, and is considered as an upper bound value constraining allocations in the successive level of classification. At initialization, we set $\mathcal{F}_0 = \{P\}$, whose single stratum is thus equivalent to the entire population, and the best sample size $n_{b,0}$ is computed as the maximum among those optimal sizes (see, e.g. Cochran 1977) obtained taking into account, separately, every single precision level ε_g set about the g-th surveyed variate :

$$n_{b,0} = \max_{g=1,\ldots,G} \frac{N^2 v_g^2}{\varepsilon_g^2 \tau_g^2 + N v_g^2},$$

where τ_g is the total estimate for Y_g on P and v_g^2 is the corresponding variance.

When $l > 0$, the set of strata $\mathcal{F}_{l-1}$, optimal at step $l-1$, is analyzed. The best sample allocation at step l, $n_{b,l}$, is initially set equal to $n_{b,l-1}$, and, for each stratum $S \in \mathcal{F}_{l-1}$ and every auxiliary variable X_k, the following algorithm is executed. Let $\mathcal{A}_S$ be the set of atoms included in the current stratum S, so that

$S = \bigcup \mathcal{A}_S$ holds true, and let $m(A)$ be a function returning the index assigned to any atom A ($m(A) = m_0$ if and only if $A = A_{m_0}$), then we can express the set of values taken on by X_k at units in any atom included in $\mathcal{A}_S$ as follows:

$$Q_k = \{q \in \mathbb{R} : \exists A \in \mathcal{A}_S : q = a_{m(A),k}\}.$$

If X_k is an ordered variate, for every q in Q_k other than $\max(Q_k)$ the stratum S is partitioned in sets S_1 and S_2 as follows:

$$S_1 = \bigcup \{A \in \mathcal{A}_S : a_{m(A),k} \leq q\},$$

and S_2 is the relative complement of S_1 in S, i.e. the set of all $i \in S$ which are not in S_1:

$$S_2 = S \backslash S_1.$$

If, on the contrary, X_k is unordered, S is instead partitioned in sets S_1 and S_2 for every proper subset S_1 of S, with $S_2 = S \backslash S_1$. For every stratum C in the candidate stratification

$$\mathcal{C} = (\mathcal{F}_{l-1} \backslash \{S\}) \cup \{S_1, S_2\},$$

the total estimates of Y_g, $g = 1, \ldots, G$,

$$\tau_{C,g} = \sum\nolimits_{A \in \mathcal{A}_C} t_{m(A),g},$$

and their corresponding variances

$$v_{C,g}^2 = (|C| - 1)^{-1} \left(\sum\nolimits_{A \in \mathcal{A}_C} (|A| - 1) \, v_{m(A),g}^2 + \sum\nolimits_{A \in \mathcal{A}_C} |A| \left(|A|^{-1} t_{m(A),g} - |C|^{-1} \tau_{C,g} \right)^2 \right),$$

are computed, and the sample allocation $n_{\mathcal{C}}$ is thus obtained applying the Bethel's rule. If $n_{\mathcal{C}} < n_{b,l}$, then the split (S_1, S_2) becomes the current best one, the best stratification candidate $\mathcal{C}^*$ becomes $\mathcal{C}$ and $n_{b,l}$ is updated to $n_{\mathcal{C}}$. In this way, the scissorial procedure which achieves the best result, i.e. the smallest sample size, is selected to generate the next optimal strata:

$$\mathcal{F}_l = \mathcal{C}^*.$$

Issues concerning the optimal number of strata are taken into account by defining the stopping criteria of the tree generating procedure. We decide to stop the algorithm if the relative difference between the optimal sample size at the current level and the optimal one at the previous level is smaller than a given parameter $\delta > 0$:

$$\delta > \left(n_{b,l-1} - n_{b,l}\right) / n_{b,l-1}. \tag{1}$$

Since the Bethel's algorithm converges to a vector whose range is $]0, +\infty[^{l+1}$, its entries must be rounded to the corresponding nearest integers towards infinity; as a consequence, especially in presence of many small strata, a given allocation is likely to be greater than the previous one: also in this case we decided to stop our procedure. To avoid too small and henceforth statistically unstable strata, additional rules can be set to avoid further disaggregations of current strata if the corresponding substrata have cardinalities smaller than a predefined minimum stratum size. Complexities in survey management can also be easily mitigated by imposing a maximum number of strata.

3. FORMING STRATA FOR THE ITALIAN FARM STRUCTURE SURVEY

For the requirements of European Community agricultural policies, the Farm Structure Survey (FSS) is executed, every two years, as a census update (Council Regulation (EEC) No 70/66), collecting data on techno-economic variables characterizing EU farms. It represents the primary source of information for the EUROFARM project (Council Regulation (EEC) No 571/88), a set of data banks to be used for processing Community surveys on the structure of agricultural holdings. Member States are responsible for taking all appropriate steps to carry out the FSS in their territories, and they are also free to select a sampling criterion, but the questionnaire and the precision required, at a national level, for the estimates of the study variables are fixed by Community regulations (see EC Regulations No 837/90 and No 959/93, and subsequent Commission Decisions 1998/377/EC and 2000/115/EC).

To illustrate our stratification technique, we execute the algorithm described in Section 2 to design the italian FSS and identify a tree-based set of strata and allocations using multivariate information. The design exploits the frame of farms listed during the fifth Agricultural General Census held in Italy in the fall of 2000. ISTAT, the Italian national statistical institute, is responsible for updates of such frame based on integration of administrative records, but they are not available at the moment of this writing. For the procedure to be initialised, we need a set of atoms in which the population of the italian agricultural holdings must be partitioned. This set of basic strata is obtained by aggregation of farms sharing the same classes of seven covariates. We select four variables related to land use and livestocks, namely utilised agricultural area (UAA), number of bovine animals (NBA), number of pigs (NP), and number of sheep and goats (NSG). To take into account the geographical characteristics of the holdings, we also added, as a stratification variable, the altitude of the farm (ALT). Finally, we collected information about holding administration and organization by means of two variables referred to as legal personality of the holder (LP), and type of tenure of the holding (TT).

Ranges of the covariates concerning the farming structure are divided into four classes for number of bovine animals ($\mathrm{NBA} = 0$, $1 \leq \mathrm{NBA} < 10$, $10 \leq \mathrm{NBA} < 50$, $50 \leq \mathrm{NBA}$), number of pigs ($\mathrm{NP} = 0$, $1 \leq \mathrm{NP} < 500$, $500 \leq \mathrm{NP} < 1000$, $1000 \leq \mathrm{NP}$), and number of sheep and goats ($\mathrm{NSG} = 0$, $1 \leq \mathrm{NSG} < 250$, $250 \leq \mathrm{NSG} < 500$, $500 \leq \mathrm{NSG}$), and into seven classes for utilised agricultural area ($\mathrm{UAA} = 0$, $0 < \mathrm{UAA} < 1$, $1 \leq \mathrm{UAA} < 5$, $5 \leq \mathrm{UAA} < 10$, $10 \leq \mathrm{UAA} < 50$, $50 \leq \mathrm{UAA} < 100$, $\mathrm{UAA} \geq 100$ ha). The range of altitude values is divided into five classes: inland mountains, coastal mountains, inland hills, coastal hills, and flat lands. Classes for the legal personality of the holder are defined in order to discriminate among sole holders, legal persons (companies) and groups of physical persons (partnership) in a group holding, cooperative enterprises, associations of holders, public institutions, and, finally, legal personalities other than the previous ones (e.g., consortia), which will be referred to as the residual ones. Holdings are also stratified taking into account their type of tenure, by discerning among owner-farmed (with further subclasses based on farm labour force categories: family labour, prevalent family labour, prevalent non-family labour), tenant-farmed, shared-farmed agricultural areas, and modes of tenure other than the previous ones. Combining all possible classes from any of the selected covariates leads to 2,964 nonempty atoms, the starting point of the procedure.

We put under study 12 land use variables, whose list is reported in Table 1. For every surveyed variable, totals and variances in each atom are computed elaborating the available Census data, enabling us to execute the Bethel's algorithm at each step of our procedure. Additional parameters needed to identify our stopping criteria are set as follows. The maximum number of strata is defined as 300, and we decide to disallow strata having a size smaller than 10. A tolerance about the relative difference between optimal sample sizes at subsequent levels is introduced setting $\delta = 0$ in equation (1), so the algorithm is stopped if $n_{b,l-1} < n_{b,l}$ for some level $l \geq 0$.

Convergence was achieved since the maximum number of strata was reached and no other stopping rule was activated for $l < 300$. Figure 1 shows the optimal allocations $n_{b,l}$ plotted as a function of the number of strata $r_l = 1, \ldots, 300$ on a logarithmic scale, i.e. against $\log(n_{b,l})$. It can be noted that the relative difference between subsequent allocations rapidly decreases, with the first ten splits being the more important with respect to such behaviour: in fact, by setting $\delta = 10\%$ the procedure would reach convergence at step $l = 7$.

Figure 2 displays a diagram of the stratification tree generated up to level 7. In order to optimise the global allocation, our splitting criterion recursively created smaller and smaller strata. The first split is on the legal personality of the holder, LP, and atoms have been included in the left daughter stratum if the class of variable LP they assume was sole holder, public institution, or a residual one. Such split is the optimal split at level 1, since it corresponds to a partition of the entire population, the only stratum available at level 0, that best decreases the sample allocation. This mainly indicates that farms organized by sole holders behave differently from those managed by more complex legal persons, such as companies, partnerships, associations, or cooperative enterprises. The second split is on the number of bovine animals, NBA. It creates two new substrata of stratum 2 (see the bottom side of Figure 2), namely strata 4 and 5, as follows: the new stratum 4 is defined as the union of such atoms in stratum 2 for which condition NBA > 10 holds true, while stratum 5 is the relative complement of stratum 4 in stratum 2. In this way, the algorithm detects the best decrement of the overall sample size (passing from 1,570,313 to 689,404 sampled units, see the right side of Figure 2) by recognizing that farms characterized by medium or large bovine livestocks need to be treated separately for sole held farms. The third split is instead on the utilised agricultural area, UUA. Here, stratum 4 is partitioned between atoms for which variable UUA is less than 100 ha (stratum 6) and remaining ones (stratum 7). Both these new strata are also divided, in successive steps, namely steps 4 to 7 (see the left side of Figure 2), on variables NP and NSG: more thoroughly, the procedure suggests to distinguish farms having no sheep or goat livestocks (NSG $= 0$), or characterized by large livestocks of pigs (NP ≥ 500).

To evaluate the efficiency of the tree-based sampling design, we calculate the best allocation corresponding to the atomised stratification, which happened to determine a sample of 89,522 units. By inspecting the stratification tree, it can be noted that a very similar overall allocation corresponds to the best stratification obtained at level $l = 102$: in fact, for such partition of 103 strata the sample size is equal to 89,509. This means that, for the same sample size, our algorithm achieves the precision requested for the survey by exploiting a number of strata, 103, which is a very small fraction of 2964, the number of available atoms, henceforth enabling an easier organization of the survey. Noticeably, another advantage of our procedure consists in avoiding unstable strata: it is worth noting that 1618 of the 2964 atoms have a size equal to or lesser than 5, while the minimum size of any of the optimal strata at level 102 is 16, so that, as a consequence, there is no need to introduce any ex post strata aggregation procedure. Further comparisons can be obtained contrasting the levels of precision achieved implementing, respectively, the atomised stratification and the stratification tree at step 102. Such levels, as reported in Table 1, can be considered very similar for the two designs. In fact, we observed that, for the atomised stratification, the Bethel's allocation was actively constrained on the precision regarding three surveyed variates, namely Cereals, Vegetables and Number of Sheep. With respect to the strata corresponding to level 102 of the tree, the previous constraints also happened to be active, even if another constraint, that on variable Number of Goats, also resulted tight for the optimization, with achieved precision levels increased from 1.92% to 1.98%. Such findings suggest that, with respect to the atomised partition, the tree can be used to detect a more compact stratification of the population, still preserving the achieved precision levels and the overall sample size.

4. CONCLUDING REMARKS

The tree-based strategy for multipurpose surveys examined in this article is planned to jointly define a rule to partition the population and to allocate sampling units in strata formed exploiting multivariate information, quantitative or qualitative. A scissorial algorithm selects finer partitions by minimizing, at each step, the sample allocation needed to achieve the required precision levels. In this way, large numbers of constraints can be satisfied without drastically increasing the number of strata. In addition, variables selected for stratification are not discarded merely on the basis of practical considerations, nor the number of their class intervals is diminished. Furthermore, the algorithm avoids the definition of empty or almost empty strata, thus excluding the need for ex post strata aggregations aimed at a better evaluation of in stratum estimation variances.

Notwithstanding, some points of criticism can be raised about our proposal. Theoretically, our procedure cannot be considered as a multiresponse generalization of the well known classification regression tree method, where the aim is that of exploiting the relationships between the covariates and a unique outcome variable. In fact, even if we deal with multipurpose surveys, nevertheless our approach consists in partitioning the available information so as to optimise only one variable, namely the sampling allocation in strata. Furthermore, the sampling strategy obtained through our methodology does not necessarily represent a global optimum: in fact, the procedure constitutes a forward strata selection algorithm, and, as a consequence, the search for optimality at a given step is conditioned on the stratification currently in use, i.e. that one based upon the splits previously executed: there is no guarantee that the stratification selected by the procedure at a certain step l will be the optimal one, even solely among all the possible partitions in $l+1$ subsets of the population. However, this situation seems to be considered as unimportant for applications, since the combinatorial nature of the problem excludes the possibility of efficient exhaustive searches for the globally optimal stratification.

The procedure was applied to redesign the Italian Farm Structure Survey. The results indicate gains in efficiency held using our strategy: for a given sample size, our procedure achieves the requested precision by exploiting a number of strata which is usually a very small fraction of the number of strata available when combining all possible classes from any of the covariates. In addition, allowing for more strata, the algorithm detects further sampling strategies for which the constraints are satisfied with sample sizes smaller than the one corresponding to the atomised stratification. The final sampling choice obviously depends upon the survey overall cost function. For this purpose, stratification trees can be applied to take into consideration the fact that an increasing number of strata usually implies larger costs due to survey organization issues, but also corresponds to smaller sample sizes, which lead to decreasing unitary costs. Forming strata by trees can thus be useful to manage the survey in an easier way, as a tool to assist the selection of the stratified sampling design which is suited to collect information about the multivariate phenomenon under study.

References

[1] Bethel, J. (1985), "An Optimum Allocation Algorithm for Multivariate Surveys," in *American Statistical Association Proceedings of the Surveys Research Methods Section*, pp. 209-212.

[2] Bethel, J. (1989), "Sample Allocation in Multivariate Surveys," *Survey Methodology*, 15, 47-57.

[3] Bloch, D. A., and Segal, M. R. (1989), "Empirical Comparison of Approaches to Forming Strata – Using Classification Trees to Adjust for Covariates," *Journal of the American Statistical Association*, 84, 408, 897-905.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth International Group.

[5] Chromy, J. (1987), "Design Optimization With Multiple Objectives," in *American Statistical Association Proceedings of the Surveys Research Methods Section*, pp. 194-199.

[6] Cochran, W. G. (1977), *Sampling Techniques*, New York: Wiley.

[7] Dalenius, T., and Hodges, J. L. Jr. (1959), "Minimum Variance Stratification," *Journal of the American Statistical Association*, 54, 285, 88-101.

[8] Ekman, G. (1959), "An Approximation Useful in Univariate Stratification," *Annals of Mathematical Statistics*, 30, 1, 219-229.

[9] Hedlin, D. (2000), "A procedure for Stratification by the Extended Ekman Rule," *Journal of Official Statistics*, 16, 1, 15-29.

[10] Hess, I., Sethi, V. K., and Balakrishnan, T. R. (1966), "Stratification: A Practical Investigation," *Journal of the American Statistical Association*, 61, 313, 74-90.

[11] Kish, L., and Anderson, D. W. (1978), "Multivariate and Multipurpose Stratification," *Journal of the American Statistical Association*, 73, 361, 24-34.

[12] Lavallee, P., and Hidiroglou, M. A. (1988), "On the Stratification of Skewed Populations," *Survey Methodology*, 14, 33-43.

[13] Sigman, R. S., and Monsour, N. J. (1995), "Selecting Samples From List Frames of Businesses," in *Business Survey Methods*, eds. B. G. Cox, D. A. Binder, B. Nanjamma Chinnappa, A. Christianson, M. J. Colledge, P. S. Kott, New York: Wiley, pp. 133-152.

[14] Singh, R. (1971), "Approximately Optimum Stratification on the Auxiliary Variable," *Journal of the American Statistical Association*, 66, 336, 829-833.

[15] Valliant, R., and Gentle, J. E. (1997), "An Application of Mathematical Programming to Sample Allocation," *Computational Statistics & Data Analysis*, 25, 337-360.

[16] Vogel, F. A. (1995), "The evolution and Development of Agricultural Statistics at the United States Department of Agriculture," *Journal of Official Statistics*, 11, 2, 161-180.

TABLES

Table 1: Surveyed Variables and Their Precision Levels

Surveyed variable	Precision Level		
	Requested by FSS	Achieved by	
		Atomised stratification	Stratification tree
Cereals	1.00	0.98	0.98
Vineyards	3.00	1.38	1.38
Olive plants	3.00	1.11	1.11
Fodder roots and brassicas	3.00	2.39	2.40
Industrial plants	3.00	2.22	2.23
Forage plants	3.00	1.37	1.39
Vegetables	3.00	3.03	3.03
Fallow land	3.00	2.69	2.78
Number of Bovine Animals	1.00	0.99	1.00
Number of Pigs	2.00	0.80	0.82
Number of Sheep	2.00	1.99	2.01
Number of Goats	2.00	1.92	1.98

FIGURE CAPTIONS

Figure 1. Step by Step Sample Sizes. The optimal allocations $n_{b,l}$ are shown as a function of the number of strata r_l exploited by the tree-based sampling design at steps $l = 0, \ldots, 299$. A logarithmic scale is applied to the horizontal axis, so that $n_{b,l}$ is plotted against $\log(r_l)$. As the number of strata increases, the tree-based stratification design attains its goals using a rapidly decreasing global sample size, since the procedure greatly improves the sampling efficiency in its first ten steps of execution.

Figure 2. Stratification Tree Diagram. The bottom side of the horizontal axis is labeled with the stratum identifier, a number that uniquely represents the corresponding subpopulation inside the stratification procedure. Sizes of such strata are reported on the top side. The left side of the vertical axis displays the sequence of steps from 0 to 7, while the right side accounts for the global optimal allocations corresponding to such steps. Double bordered blocks represent splitted strata. Doughter strata are linked to their parents through elbow lines, and, when not further splitted in subsequent steps, they are shown as single bordered blocks. For left doughter strata, the covariate on which the split happened and the condition it satisfied when defining the left substratum are reported above the corresponding elbow line. The number inside a given block is the sample allocation the procedure assigns, to the corresponding stratum, during the step at which the block is positioned. Since a stratum can remain unsplitted in steps successive to that in which it is created, but its sample allocation can vary from one step to the other, dashed blocks are used to report modifications of stratum sample sizes.

FIGURE ARTWORK

Figure 1

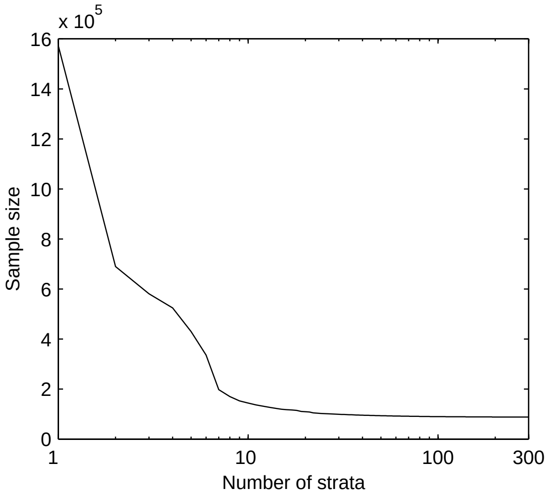
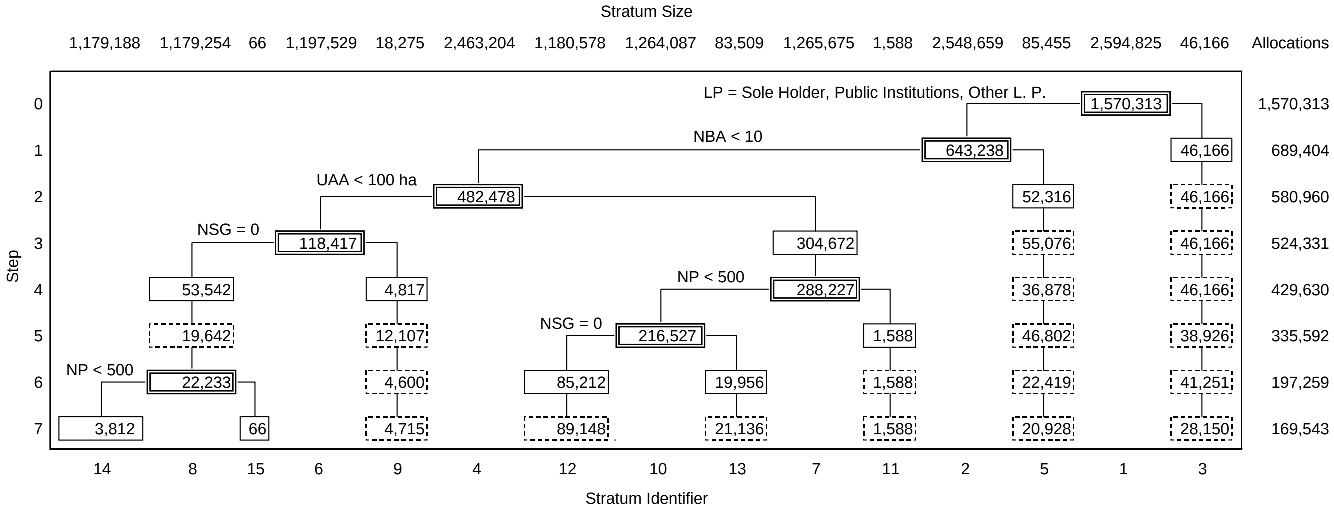


Figure 2



Elenco dei papers del Dipartimento di Economia

2000.1 *A two-sector model of the effects of wage compression on unemployment and industry distribution of employment*, by Luigi Bonatti

2000.2 *From Kuwait to Kosovo: What have we learned? Reflections on globalization and peace,* by Roberto Tamborini

2000.3 *Metodo e valutazione in economia. Dall'apriorismo a Friedman* , by Matteo Motterlini

2000.4 *Under tertiarisation and unemployment.* by Maurizio Pugno

2001.1 *Growth and Monetary Rules in a Model with Competitive Labor Markets,* by Luigi Bonatti.

2001.2 *Profit Versus Non-Profit Firms in the Service Sector: an Analysis of the Employment and Welfare Implications,* by Luigi Bonatti, Carlo Borzaga and Luigi Mittone.

2001.3 *Statistical Economic Approach to Mixed Stock-Flows Dynamic Models in Macroeconomics,* by Bernardo Maggi and Giuseppe Espa.

2001.4 *The monetary transmission mechanism in Italy: The credit channel and a missing ring,* by Riccardo Fiorentini and Roberto Tamborini.

2001.5 *Vat evasion: an experimental approach,* by Luigi Mittone

2001.6 *Decomposability and Modularity of Economic Interactions,* by Luigi Marengo, Corrado Pasquali and Marco Valente.

2001.7 *Unbalanced Growth and Women's Homework,* by Maurizio Pugno

2002.1 *The Underground Economy and the Underdevelopment Trap,* by Maria Rosaria Carillo and Maurizio Pugno.

2002.2 *Interregional Income Redistribution and Convergence in a Model with Perfect Capital Mobility and Unionized Labor Markets,* by Luigi Bonatti.

2002.3 *Firms' bankruptcy and turnover in a macroeconomy,* by Marco Bee, Giuseppe Espa and Roberto Tamborini.

2002.4 *One "monetary giant" with many "fiscal dwarfs": the efficiency of macroeconomic stabilization policies in the European Monetary Union,* by Roberto Tamborini.

2002.5 *The Boom that never was? Latin American Loans in London 1822-1825,* by Giorgio Fodor.

2002.6 *L'economia senza banditore di Axel Leijonhufvud: le 'forze oscure del tempo e dell'ignoranza' e la complessità del coordinamento,* by Elisabetta De Antoni.

2002.7 *Why is Trade between the European Union and the Transition Economies Vertical?,* by Hubert Gabrisch and Maria Luigia Segnana.

2003.1 *The service paradox and endogenous economic gorwth,* by Maurizio Pugno.

2003.2 *Mappe di probabilità di sito archeologico: un passo avanti,* di Giuseppe Espa, Roberto Benedetti, Anna De Meo e Salvatore Espa.
(*Probability maps of archaeological site location: one step beyond*, by Giuseppe Espa, Roberto Benedetti, Anna De Meo and Salvatore Espa).

2003.3 *The Long Swings in Economic Understianding,* by Axel Leijonhufvud.

2003.4 *Dinamica strutturale e occupazione nei servizi,* di Giulia Felice.

2003.5 *The Desirable Organizational Structure for Evolutionary Firms in Static Landscapes,* by Nicolás Garrido.

2003.6 *The Financial Markets and Wealth Effects on Consumption* An Experimental Analysis, by Matteo Ploner.

2003.7 *Essays on Computable Economics, Methodology and the Philosophy of Science,* by Kumaraswamy Velupillai.

2003.8 *Economics and the Complexity Vision: Chimerical Partners or Elysian Adventurers?*, by Kumaraswamy Velupillai.

2003.9 *Contratto d'area cooperativo contro il rischio sistemico di produzione in agricoltura,* di Luciano Pilati e Vasco Boatto.

2003.10 *Il contratto della docenza universitaria. Un problema multi-tasking*, di Roberto Tamborini.

2004.1 *Razionalità e motivazioni affettive: nuove idee dalla neurobiologia e psichiatria per la teoria economica?* di Maurizio Pugno.
(*Rationality and affective motivations: new ideas from neurobiology and psychiatry for economic theory?* by Maurizio Pugno.

2004.2 *The economic consequences of Mr. G. W. Bush's foreign policy. Can th US afford it?* by Roberto Tamborini

2004.3 *Fighting Poverty as a Worldwide Goal* by Rubens Ricupero

2004.4 *Commodity Prices and Debt Sustainability* by Christopher L. Gilbert and Alexandra Tabova

2004.5 *A Primer on the Tools and Concepts of Computable Economics* by K. Vela Velupillai

2004.6 *The Unreasonable Ineffectiveness of Mathematics in Economics* by Vela K. Velupillai

2004.7 *Hicksian Visions and Vignettes on (Non Linear) Trade Cycle Theories* by Vela K. Velupillai.

2004.8 *Trade, inequality and pro-poor growth: Two perspectives, one message?* by Gabriella Berloffa and Maria Luigia Segnana.

2004.9 *Worker involvement in entrepreneurial nonprofit organizations. Toward a new assessment of workers? Perceived satisfaction and fairness* by Carlo Borzaga and Ermanno Tortia.

2004.10 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part I): Rational Bargaining and Justification* by Lorenzo Sacconi

2004.11 *A Social Contract Account for CSR as Extended Model of Corporate Governance (Part II): Compliance, Reputation and Reciprocity* by Lorenzo Sacconi

2004.12 *A Fuzzy Logic and Default Reasoning Model of Social Norm and Equilibrium Selection in Games under Unforeseen Contingencies* by Lorenzo Sacconi and Stefano Moretti

2004.13 *The Constitution of the Not-For-Profit Organisation: Reciprocal Conformity to Morality* by Gianluca Grimalda and Lorenzo Sacconi

2005.1 *The happiness paradox: a formal explanation from psycho-economics* by Maurizio Pugno

2005.2 *Euro Bonds: in Search of Financial Spillovers* by Stefano Schiavo

2005.3 *On Maximum Likelihood Estimation of Operational Loss Distributions* by Marco Bee

2005.4 *An enclave-led model growth: the structural problem of informality persistence in Latin America* by Mario Cimoli, Annalisa Primi and Maurizio Pugno

2005.5 *A tree-based approach to forming strata in multipurpose business surveys,* Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta.

PUBBLICAZIONE REGISTRATA PRESSO IL TRIBUNALE DI TRENTO