Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

A tree-based approach to forming strata in multipurpose business surveys

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

by Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta

December 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2008
Vol. 34, No. 2, pp. 195-203
Statistics Canada, Catalogue No. 12-001-X

A tree-based approach to forming strata in multipurpose business surveys

Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta [1]

Abstract

The design of a stratified simple random sample without replacement from a finite population deals with two main issues: the definition of a rule to partition the population into strata, and the allocation of sampling units in the selected strata. This article examines a tree-based strategy which plans to approach jointly these issues when the survey is multipurpose and multivariate information, quantitative or qualitative, is available. Strata are formed through a hierarchical divisive 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 not to define empty or almost empty strata, thus avoiding the need for strata collapsing 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 multipurpose 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 units to become jumpers.

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 creating fewer classes 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 really be interpreted as an extreme solution to the problem of strata formation, since, between the cases of no stratification and using the atomised stratification, 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 hierarchical divisive 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 which variable is to be used to define the new, more disaggregated partition: each stratum in the current partition is split 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

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

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. The classical methods which deal with the univariate case (Dalenius and Hodges 1959; Singh 1971; Lavallée and Hidiroglou 1988; Hedlin 2000; Lu and Sitter 2002; Gunning and Horgan 2004; for a review see Horgan 2006) can't be easily extended to cover the case where one seeks to exploit multiple covariates for stratification. 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 unsplit strata and of the available split substrata.

According to what we have said before, our position in the grand picture of multivariate stratification follows the goal by Kish and Anderson (1978), namely bringing some results in the field of stratified sampling towards the needs of survey practice. Practitioners daily perform multivariate (several variables available for stratification) and multipurpose (several variables and many other statistics are the main objectives of survey efforts) surveys. Thus, the aim of our approach consists in giving the possibility of combining stratification and sample allocation. This means that we are concerned with the choice of the number of stratifying variables, of the number of class intervals for each variable and of the optimal Bethel's allocation to strata. As of this choice, our methodology cannot be reduced to the standard solution of the multivariate stratification problem, *i.e.*, the use of multivariate techniques such as cluster analysis and principal components (see, for example, Mulvey 1983, Pla 1991 and Jarque 1981). As a matter of fact, this branch of literature does not use (or uses only indirectly) the variables of interest, but only the auxiliary variables, and the allocation issue is neglected. It would be even less justifiable to reduce our approach to the ones reviewed by Särndal, Swensson and Wretman (1992, section 12.6 and 12.7): the techniques presented in section 12.6 combine stratification and multivariate sample allocation, but are not multipurpose, whereas the methods of section 12.7 are multipurpose but are based on predetermined strata.

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 F of H_F nonempty subpopulations, called strata, partitioning P. Our problem is how to select F in order to minimize the corresponding overall sample allocation n_F in a way such that, for $g = 1, \ldots, G$, the coefficient of variation (CV_g) corresponding to the g^{th} variate of interest is not greater than the desired level of precision, say $\varepsilon_g > 0$.

For a given F, such minimization is executed by computing the Bethel's (1985) sample allocation rule. More thoroughly, let us indicate by n_h, $h = 1, \ldots, H_F$, the sample allocation in stratum h. The global survey cost corresponding to F can thus be given as follows

$$f(n_1, \ldots, n_{H_F}) = c_F + \textstyle\sum_{h=1}^{H_F} c_h n_h,$$

where c_F is a fixed cost independent from $\mathbf{n}_F = (n_1, \ldots, n_{H_F})'$, and c_h represents the cost to sample one unit in stratum h. Furthermore, let Y_g be the total in P of the g^{th} response variable, N_h the size of the h^{th} stratum of F, and $S^2_{h,g}$ the variance of y_g in stratum h. Then the g^{th} constraint on the required precision can be expressed as:

$$(\mathrm{CV}_g)^2 \le \varepsilon_g^2 \equiv \sum_{h=1}^{H_F} \frac{N_h^2 S_{h,g}^2}{n_h} - \sum_{h=1}^{H_F} N_h S_{h,g}^2 \le Y_g^2 \varepsilon_g^2$$

$$\equiv \sum_{h=1}^{H_F} \frac{N_h^2 S_{h,g}^2}{\left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S_{h,g}^2\right) n_h} \le 1,$$

so that, if we consider the following quantities, referred to as the standardized precision units,

$$\xi_{h,g} = N_h^2 S_{h,g}^2 \Big/ \left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S_{h,g}^2 \right),$$

the problem of optimal allocation for F can be expressed as follows:

$$\begin{array}{lll} \min & f(\mathbf{n}_F) & \\ \text{subject to} & \sum_{h=1}^{H_F} \xi_{h,g} / n_h \leq 1, & g = 1, \ldots, G, \\ & 1/n_h > 0, & h = 1, \ldots, H_F. \end{array}$$

Bethel (1989) derived the solution to such problem, say n_h^*, $h = 1, \ldots, H_F$, as follows:

$$1/n_h^* =$$

$$\begin{cases} \sqrt{c_h} \Big/ \left(\sqrt{\sum_{g=1}^{G} \alpha_g^* \xi_{h,g}} \sum_{l=1}^{H_F} \sqrt{c_l \sum_{g=1}^{G} \alpha_g^* \xi_{l,g}} \right) & \text{if } \sum_{g=1}^{G} \alpha_g^* \xi_{h,g} > 0, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\alpha_g^* = \lambda_g / \sum_{g=1}^{G} \lambda_g$, and λ_g is the Lagrangian multiplier of the constraint on the maximum error allowed estimating the g^{th} surveyed variable, and indicates whether the g^{th} constraint is "active" in the allocation problem solution (namely, if $\alpha_g^* = 0$, then the constraint is not active). The corresponding global optimal allocation is thus given by setting $n_F = \sum_{h=1}^{H_F} n_h^*$.

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 $\mathbf{T} = (Y_{m,g})$ and estimation variances $\mathbf{V} = (S_{m,g}^2)$, and the sizes N_m, $m = 1, \ldots, M$. 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 $x_{i,k}$ be 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 $A_{m,k} = \{i \in P : x_{i,k} = a_{m,k}\}$. In the case where the covariates X_k are continuous, the set Ξ_k will contain N not empty atoms. As the algorithm is hierarchical divisive, the number of final atoms does not affect at all the steps of the algorithm. Only the initial phase of construction of the aggregate statistics and, at most, the memory allocation are impacted. Our empirical experience suggests that the computing times do not change much if the size of the atoms is equal to 1. On the contrary, continuous or ordered variables speed up the algorithm, as the number of possible binary partitions is, if the number of values is the same, much smaller with respect to the case of categorical variables. One verifies that this construction does yield a stratification: each unit of the population appears in one atom, and in one only. To illustrate our definitions, let us refer to the data shown in Table 1, where a simple example is described in which a set of $M = 9$ atoms (obtained exploiting $K = 2$ covariates both having 3 distinct values, namely 1, 2, and 3) is assumed to constitute the basic stratification to survey $G = 2$ variables, whose totals and estimation variances are also reported, together with atom sizes. In this context, we have $\Xi_1 = \Xi_2 = \{1, 2, 3\}$, and, for example, A_8 is the subpopulation whose elements i are such that $x_{i,1} = a_{8,1} \equiv 3$ and $x_{i,2} = a_{8,2} \equiv 2$.

Table 1
Example data for 9 atoms and 2 surveyed variates

Atoms				Surveyed Variates			
Id	Definition		Sizes	Totals		Variances	
m	$a_{m,1}$	$a_{m,2}$	N_m	$Y_{m,1}$	$Y_{m,2}$	$S_{m,1}^2$	$S_{m,2}^2$
1	1	1	1,000	10	10	16	25
2	1	2	1,000	10	10	16	4
3	1	3	1,000	10	10	16	4
4	2	1	1,000	10	10	16	25
5	2	2	1,000	10	10	16	4
6	2	3	1,000	10	10	16	4
7	3	1	1,000	10	10	4	25
8	3	2	1,000	10	10	4	16
9	3	3	1,000	10	10	4	16

The procedure we propose generates a sequence of stratifications which can be represented as a classification tree. Define the level l 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 the number of nodes sharing the same level l. Since only one node will be split at each level, we have $r_l = l + 1$ for every l. At each level $l \geq 0$ the procedure determines a class 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 F_l are all candidates for being split 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 unsplit strata in 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 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 $F_0 = \{P\}$, 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 obtained taking into account, separately, every single precision level ε_g set about the g^{th} surveyed variate:

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

where Y_g is the total estimate for y_g on P and S_g^2 is the corresponding variance (see, for the optimum allocation with only one item, Cochran 1977, pages 97-106, and Särndal *et al.* 1992, pages 104-109).

When $l > 0,$ the set of strata 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 $U \in F_{l-1}$ and every auxiliary variable X_k, the following algorithm is executed. Let A_U be the set of atoms contained in the current stratum U, so that $U = \bigcup A_U$ 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 for units contained in any atom of A_U as follows:

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

If X_k is an ordered variate, for every q in Q_k other than $\max(Q_k)$ the stratum U is partitioned into sets $U_1 = U_{q,1}$ and $U_2 = U_{q,2}$ as follows:

$$U_{q,1} = \bigcup \{A \in A_U: a_{m(A),k} \leq q\},$$

and $U_{q,2}$ is the relative complement of $U_{q,1}$ in U, *i.e.*, the set of all $i \in U$ which are not in $U_{q,1}$:

$$U_{q,2} = U \setminus U_{q,1}.$$

In our example, for a stratum U defined as $A_1 \cup A_2 \cup A_8$ we have $A_U = \{A_1, A_2, A_8\}$, $Q_1 = \{1, 3\}$ and $Q_2 = \{1, 2\}$ (see Table 1), so that our algorithm would try to split U in $U_1 = A_1 \cup A_2$ and $U_2 = A_8$ using X_1, and in $U_1 = A_1$ and $U_2 = A_2 \cup A_8$ using X_2. If, on the contrary, X_k is unordered, U is instead partitioned in sets U_1 and U_2 for every proper subset U_1 of U, with $U_2 = U \setminus U_1$.

We thus have a corresponding candidate stratification, namely

$$C = (F_{l-1} \setminus \{U\}) \cup \{U_1\} \cup \{U_2\},$$

which includes all the strata in F_{l-1} other than U, and, in addition, U_1 and U_2. For every stratum C in the collection C, the total estimates of Y_g, $g = 1, \ldots, G$,

$$Y_{C,g} = \sum\nolimits_{A \in A_C} Y_{m(A),g},$$

and their corresponding variances

$$S_{C,g}^2 = (N_C - 1)^{-1} \Big(\sum\nolimits_{A \in A_C} (N_A - 1) S_{m(A),g}^2 + \sum\nolimits_{A \in A_C} N_A (N_A^{-1} Y_{m(A),g} - N_C^{-1} Y_{C,g})^2 \Big),$$

are computed, and the sample allocation n_C is thus obtained applying the Bethel's rule. If $n_C < n_{b,l}$, then the split (U_1, U_2) becomes the current best one, the best stratification candidate C^* becomes C and $n_{b,l}$ is updated to n_C. In this way, the divisive procedure which achieves the best result, *i.e.*, the smallest sample size, is selected to generate the next optimal strata:

$$F_l = C^*.$$

In the framework of our example, for precision levels $\varepsilon_1 = \varepsilon_2 = 0.1$, let us describe the optimal split at level $l = 1$, *i.e.*, that one splitting the entire population in two strata. Using the data described in Table 1, the algorithm indicates that the best split of $U = P$ is based on variable X_2, and is obtained by setting

$$\begin{aligned} U_1 &= \bigcup \{A \in A_P: a_{m(A),k} \leq 2\} \\ &= A_1 \cup A_2 \cup A_4 \cup A_5 \cup A_7 \cup A_8, \end{aligned}$$

and correspondingly $U_2 = P \setminus U_1 = A_3 \cup A_6 \cup A_9$. Such optimal division is represented in Figure 1, where, for every stratum, its size, its definition in terms of included atoms, the current allocation, and the estimation statistics are thoroughly reported.

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 $\delta > 0$:

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

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 yield a sample size 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.

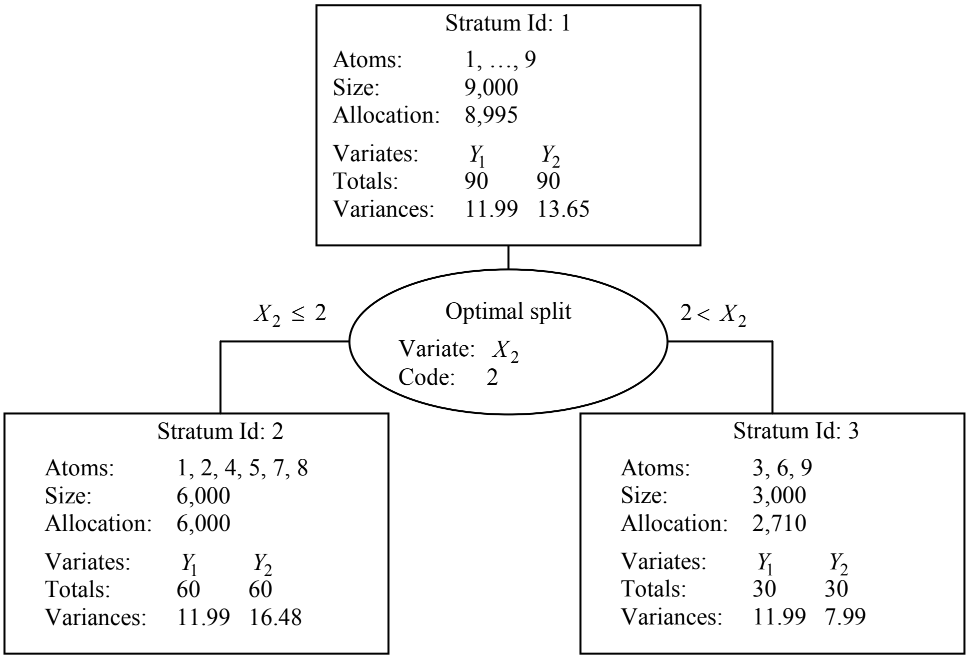

Figure 1 The first optimal split for the example data

This approach, performing an exhaustive search in each single split, guarantees that the corresponding stratification and allocation are optimal, but only conditionally to the splits previously executed. We know that monotonicity of solutions and conditional optimality of each sub-tree obtained by splitting recursively each node are necessary but not sufficient conditions for a binary tree to be optimal. In order to guarantee the overall optimality, to these conditions we should add the requirement that an optimal stratification in, say, H strata, can only be obtained by partitioning one of the nodes of the optimal stratification in $H - 1$ strata. In other words, we should assume that an optimal partition in H strata is a subspace of the optimal partition in $H - 1$ strata, which implies that partitioning a given stratum will not modify the objective function - *i.e.*, the allocation - in the remaining $H - 1$ strata. However, this assumption is rarely true in practical survey applications, since splitting a stratum usually induces a modification of the optimal allocations in all the remaining unsplit strata.

The proposed algorithm, inspired by the sequential and recursive nature of binary trees, can be considered as an heuristic approach to the problem of multivariate stratification, which enables us to detect good, nearly-optimal, strata at the cost of a reasonable computational burden. As a result, this technique is effective in partitioning populations making use of large sets of both continuous and qualitative auxiliary stratifying variables. In addition, the simple structure of binary trees implies a great flexibility in the introduction of any number of additional constraints, such as lower limits on the number of units in each stratum.

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. All the algorithms have been implemented by one of the authors in MATLAB language; a Win32 console application has also been developed in C++ to enable software execution in batch mode.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 were not available at the moment of this writing. For the procedure to be initialised, we need a set of atoms into 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 ($\mathrm{NP}=0,\ 1\leq \mathrm{NP}<500,\ 500\leq \mathrm{NP}<1{,}000,\ 1{,}000\leq \mathrm{NP}$), 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 2. 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 $l\geq 0$.

Table 2
Surveyed variables in the Italian farm structure survey and their precision levels

	Required CV		
Surveyed variable	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

Convergence was achieved since the maximum number of strata was reached and no other stopping rule was activated for $l<300$. Figure 2 shows the optimal allocations $n_{b,l}$ plotted as a function of the number of strata $r_l,\ l=1,\ldots,300$ 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 3 displays a diagram of the stratification tree generated up to level 7. In order to optimize 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 3), 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 $\mathrm{NBA}>10$

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 3) 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 utilized 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 3), on variables NP and NSG: more thoroughly, the procedure suggests to distinguish farms having no sheep or goat livestocks ($\mathrm{NSG} = 0$), or characterized by large livestocks of pigs ($\mathrm{NP} \geq 500$).

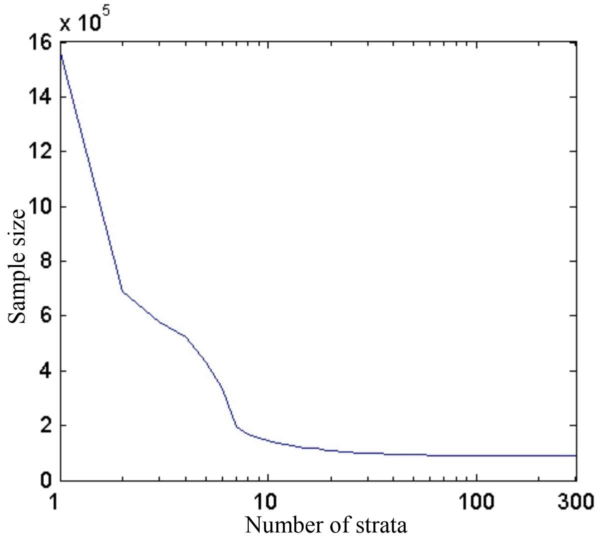

Figure 2 Step by Step Sample Sizes. The optimal allocations $n_{b,l}$ are shown as a function of the number of strata r_l exploited by the tree-based sampling design at steps $l = 0, \ldots, 299$. A logarithmic scale is applied to the horizontal axis, so that $n_{b,l}$ is plotted against $\log(r_l)$. 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

To evaluate the efficiency of the tree-based sampling design, we calculate the best allocation corresponding to the atomised stratification, which determined 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 2,964, the number of available atoms, henceforth enabling an easier organization of the survey. Another noticeable advantage of our procedure consists in avoiding unstable strata: it is worth noting that 1,618 of the 2,964 atoms have a size equal to or less 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 strata collapsing 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 2, 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 hierarchical divisive 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 creating 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, 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. In some situations, the use of other methods such as dynamic programming can be used for conducting an efficient exhaustive search for the globally optimal stratification (see Bühler and Deutler 1975, and Lavallée 1988).

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.

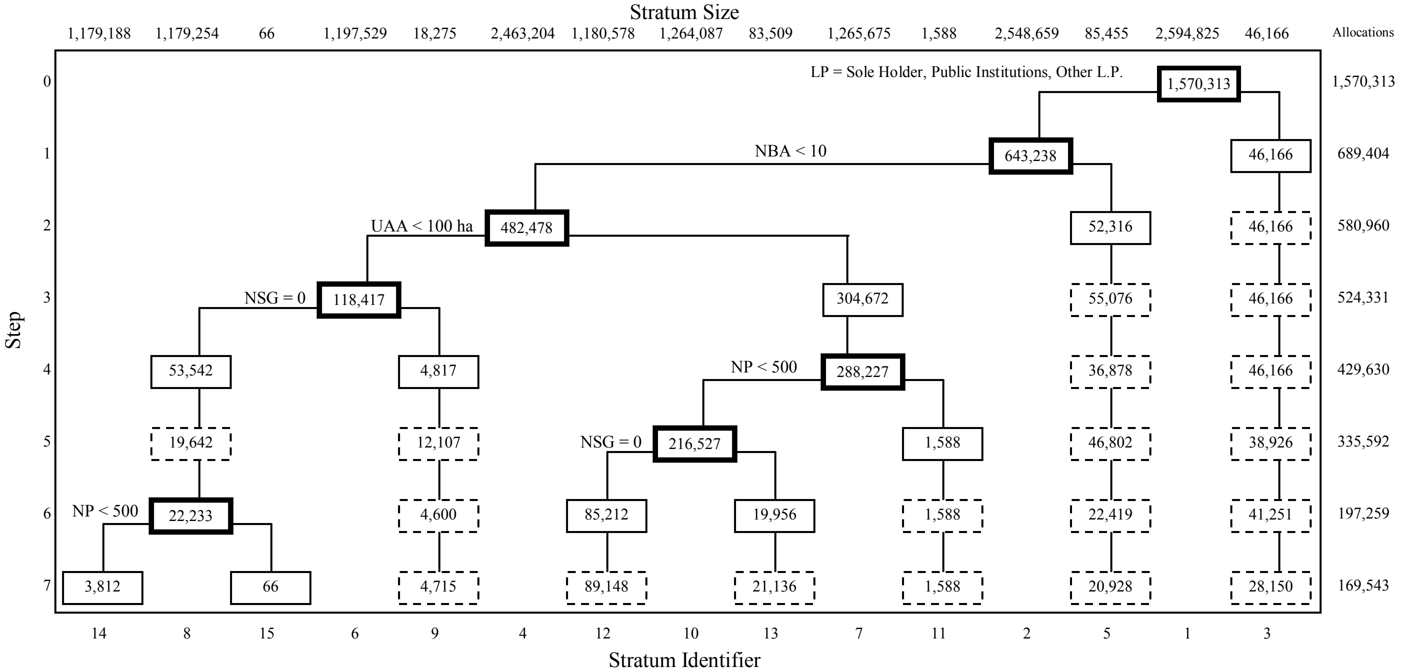

Figure 3 **Stratification Tree Diagram. The bottom side of the horizontal axis is labeled with the stratum identifier, a number that uniquely represents the corresponding subpopulation inside the stratification procedure. Sizes of such strata are reported on the top side. The left side of the vertical 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 split strata. Daughter strata are linked to their parents through elbow lines, and, when not further split in subsequent steps, they are shown as single bordered blocks. For left daughter 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 unsplit 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**

Acknowledgements

The authors wish to thank an Associate Editor and two referees for their valuable comments, which led to a significant improvement of the paper.

References

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. In *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 47-57.

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.

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

Bühler, W., and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Chromy, J. (1987). Design optimization with multiple objectives. In *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Dalenius, T., and Hodges, J.L. Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*. 54, 285, 88-101.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166.

Hedlin, D. (2000). A procedure for stratification by the extended ekman rule. *Journal of Official Statistics*, 16, 1, 15-29.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*. 74, 1, 67–76.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Applied Statistics*. 30, 2, 163-169.

Kish, L., and Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 361, 24-34.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Virginia, August 1988, 646-651.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Lu, W., and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. *Survey Methodology*, 28, 2, 199-207.

Mulvey, J.M. (1983). Multivariate stratified sampling by optimization. *Management Science*, 29, 6, 715-724.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, 47, 4, 1409-1422.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

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 and P.S. Kott). New York: John Wiley & Sons, Inc., 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association*, 66, 336, 829-833.

Valliant, R., and Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

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.