Tree-based and GA tools for optimal sampling design

The R User Conference 2008

August 12-14, Technische Universität Dortmund, Germany

Marco Ballin, Giulio Barcaroli

Istituto Nazionale di Statistica (ISTAT)

Definition of the problem (1)

In a survey, the optimality of a stratified sample can be defined in terms of both the following elements:

- total cost (unit cost per interview, product the sample size);
- planned accuracy (expected sampling variance related to target estimates).

A sample design is acceptable if expected sampling errors are below pre-defined limits, and costs are sustainable.

Definition of the problem (2)

Bethel (1985) proposed an algorithm allowing to determine total sample size and allocation of units in strata, so to minimise costs under the constraints of defined precision levels of estimates, in the multivariate case (more than one estimate).

Under this approach, population stratification, i.e. the partition of the sampling frame obtained by cross-classifying units by means of stratification variables, is given.

But ***stratification*** has a great impact on sampling variance and, in general, it ***should not be considered as given***, but determined on the basis of the survey requirements.

Definition of the problem (3)

Our proposal is: given a **population frame**, with p **X** auxiliary variables, and a **sample survey,** with specific constraints on the accuracy of g **Y** target variables, then ***jointly determine***:

2. the **best stratification** (partition by means of auxiliary variables) of this frame, and
3. the minimum **sample size** and allocation of units in strata, required to satisfy constraints on estimates accuracy.

This can be done by using **search techniques** (***tree*** or ***genetic algorithm***) to explore the possible solutions, i.e. the different possible stratifications, that are evaluated by means of the **Bethel algorithm**.

Bethel algorithm

The optimal multivariate allocation problem can be defined as the search for the solution of the minimum (with respect to n_h) of linear function C under the convex constraints

$$V(Y_g) \leq U_g \qquad g = 1,\ldots,G$$

Bethel suggested that by introducing the variable $x_h = \begin{cases} 1/n_h \text{ if } n_h \geq 1 \\ \infty \ \text{ otherwise} \end{cases}$

the problem is equivalent to search the minimum of the convex function $C(x_1,\ldots,x_H)$ under the set of linear constraints

$$\sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \leq U_g$$

An algorithm, that is proved to converge to the solution (if it exists), is provided by Bethel (and Chromy) by applying Lagrange multipliers method to this problem.

Optimal stratification: the tree-based approach (1)

The tree-based approach has been ideated by Benedetti, Espa, Lafratta: “A tree-based approach to form strata in multi-purpose business surveys”, *Discussion Paper n.5/2005*, Università degli Studi di Trento.

The proposed procedure searches the best stratification by generating a tree with a splitting rule such that, at any given level, the generating node is chosen in such a way that the decrease of the overall sample size from one level to the other, is maximised.

Optimal stratification: the tree-based approach (2)

Given *p* auxiliary variables $X_1,\ldots,X_p$ in the frame, with domain sets $D_i = \{x_{i1},\ldots,x_{im_i}\}$ $(i = 1,\ldots,p)$ we can represent a solution by means of a vector $v = [v_1,\ldots,v_M]$ of cardinality

$$M = \sum_{k=1}^{p} m_k$$

whose elements v_j can assume 1 or 0 values.

If we set $j = (\sum_{k=1}^{i-1} m_k) + q$

then we have $v_j = \begin{cases} 1 \text{ if the q-th value of the } i\text{-th variable is activated} \\ 0 \text{ otherwise} \end{cases}$

Optimal stratification: the tree-based approach (3)

The tree-based algorithm is a sequence of four different steps.

Step 0 (initialisation): the node associated to the stratification characterised by a unique stratum, coinciding with the whole population, is the *root* of the tree (level k = 0), and is set as *generating node*.

Step 1: from the generating node at level k, “child” nodes of level (k+1) are generated, by on turn activating a single value of the vector $v = [v_1, ..., v_M]$ among those not yet activated..

Optimal stratification: the tree-based approach (4)

Step 2: at level (k+1), the overall sample size *n* is calculated with the Bethel-Chromy algorithm for each node in the level. The node with the minimum *n* is set as *generating node*.

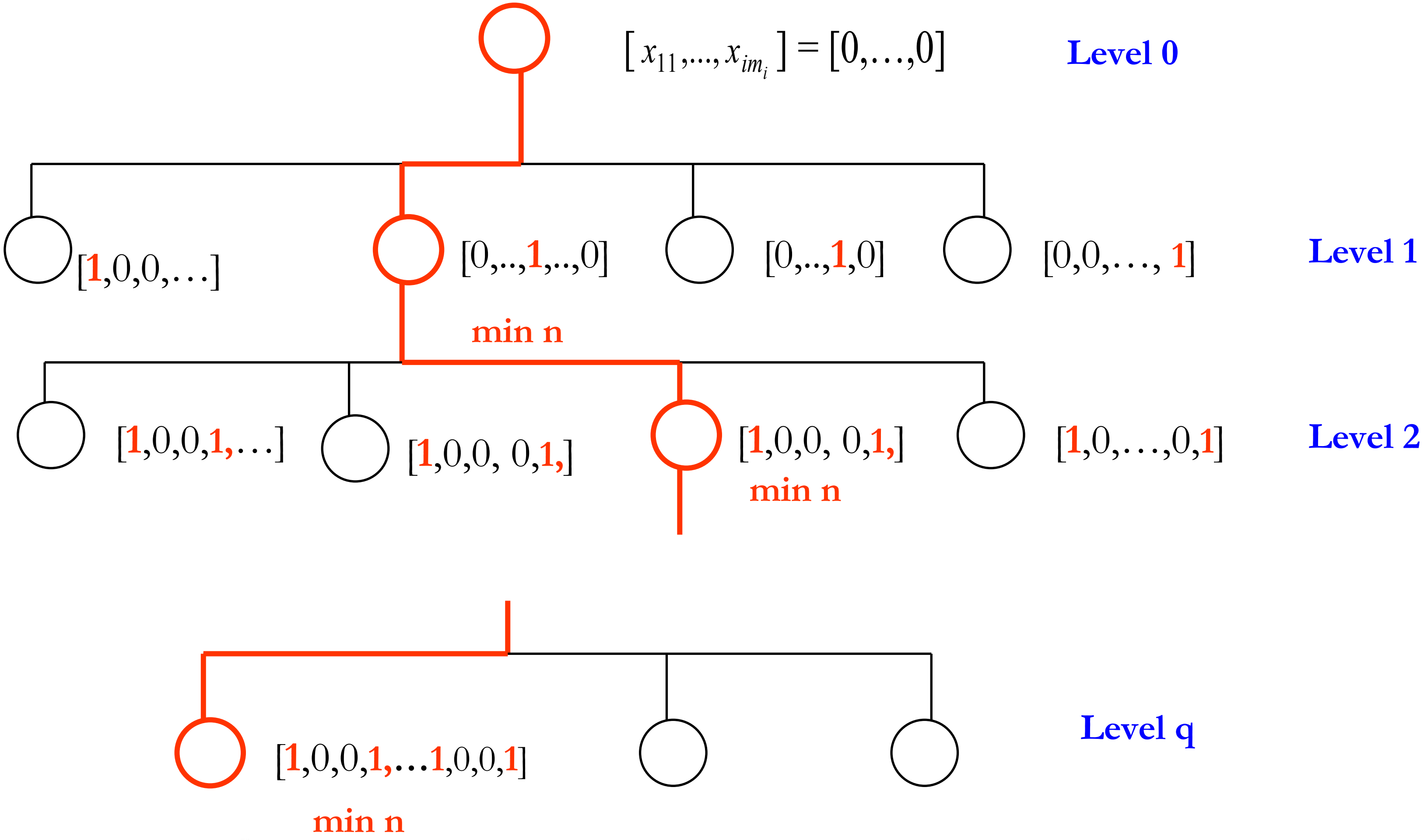
Step 3 (stopping rule): steps 1 and 2 are repeated until

(c) the maximum acceptable number of strata has been reached (the activation of new values in X's domains increases the number of resulting strata)

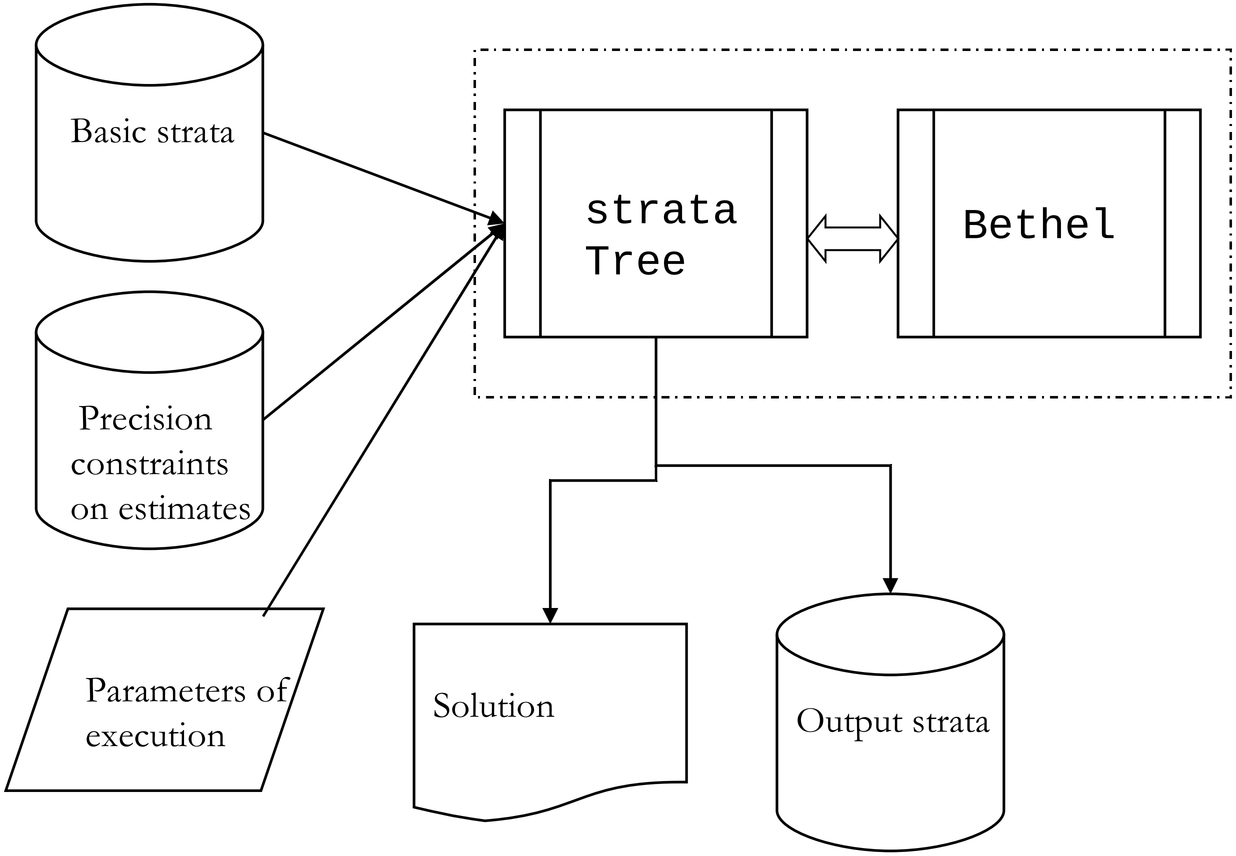
(d) the gain in terms of reduction of the overall sample size becomes negligible.

Best solution is then selected by considering the one associated to the generating node of the previous level.

Optimal stratification: the tree-based approach (5)



Optimal stratification: the tree-based approach (6)



Optimal stratification: the evolutionary approach (1)

The application of the tree-based algorithm, previously introduced, allows to obtain a (relatively) fast solution.

This approach, however, may be subject to local minima.

It is therefore convenient to verify (and possibly improve) the resulting solution by sequentially applying a different algorithm, which is of the evolutionary type, i.e. based on the genetic algorithm.

Optimal stratification: the evolutionary approach (2)

To be applied, a genetic algorithm requires two basic elements to be defined:

- ❑ a *genetic representation* of the solution domain;
- ❑ a *fitness function* to evaluate each solution.

In our problem, each ***solution*** can be represented by the $v = [v_1, \ldots, v_M]$ vector already introduced in the tree-based approach, that identifies a particular stratification (partition) of the population frame.

The ***fitness*** of any given solution is evaluated by means of the Bethel algorithm, and it is given by the minimum sample size required to satisfy precision constraints to sampling estimates.

Optimal stratification: the evolutionary approach (3)

The implemented genetic algorithm makes use of **`genalg`** package (Willighagen 2005), and is based on the following steps.

Step 0 (initialisation): an initial set of t individuals (possible solutions) are randomly generated, possibly containing (as a "suggestion") the solution found by the tree-based approach; the fitness of each individual is evaluated.

Step 1: the next generation of individuals is generated by ***selecting*** the fittest ones of the current generation, and by applying the genetic operators ***crossover*** and ***mutation***

Step 2 (stopping rule): step 1 is iterated k times, then the best solution (the fittest, i.e the one with the minimum sample size) is outputted

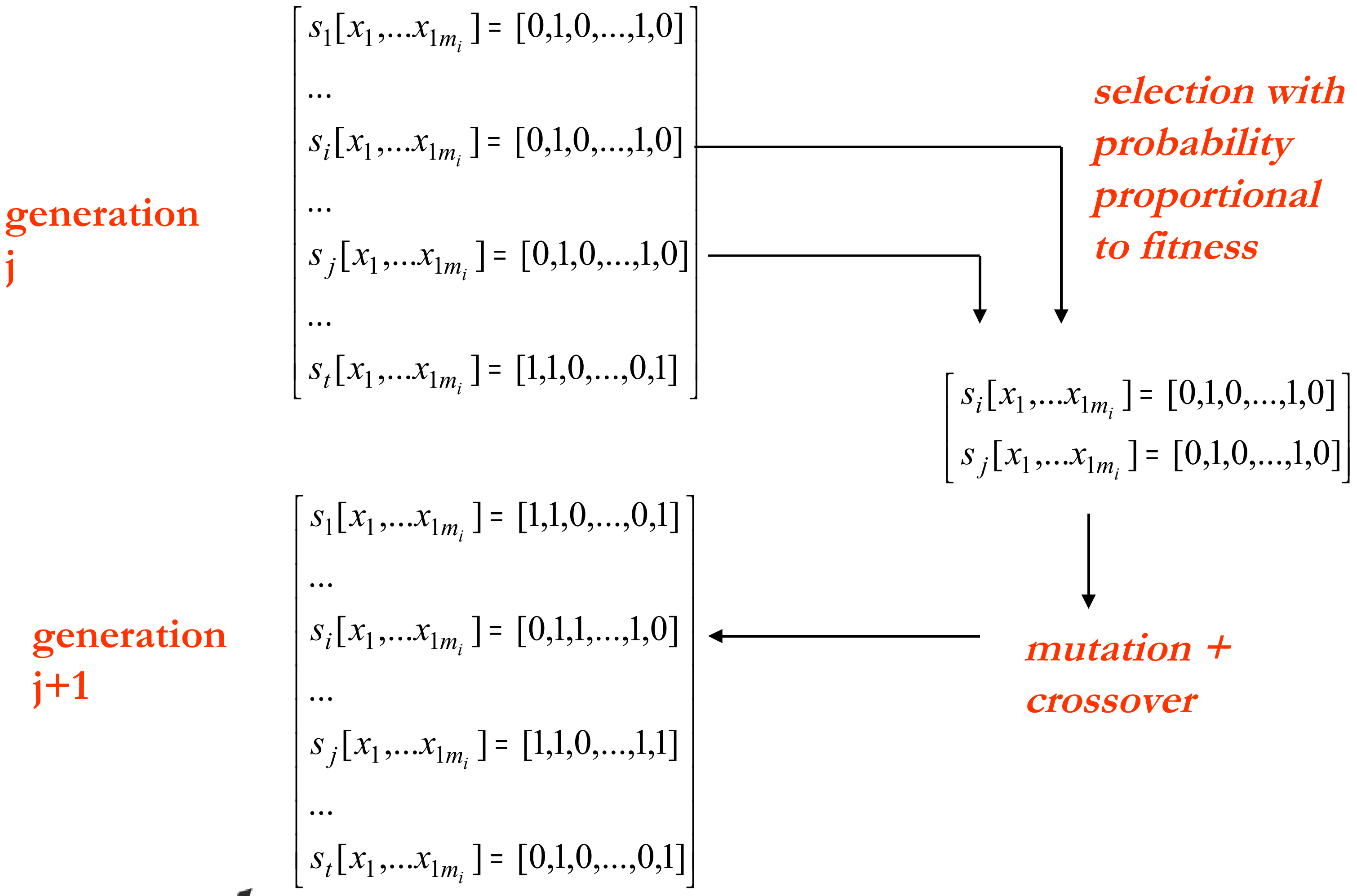
Optimal stratification: the evolutionary approach (4)

crossover : given two parents, a subset of chromosomes are exchanged between them

mutation: given the probability that an arbitrary chromosome may change from its original state to another (*mutation chance*), for each chromosome in an individual, a random value is drawn in order to decide to change or not

Mutation is very important to decide the rapidity of the convergence: too rapid, risk of local minima

Optimal stratification: the evolutionary approach (5)



generation j

$$\begin{bmatrix} s_1[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,1,0] \\ \ldots \\ s_i[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,1,0] \\ \ldots \\ s_j[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,1,0] \\ \ldots \\ s_t[x_1,\ldots x_{1m_i}] = [1,1,0,\ldots,0,1] \end{bmatrix}$$

selection with probability proportional to fitness

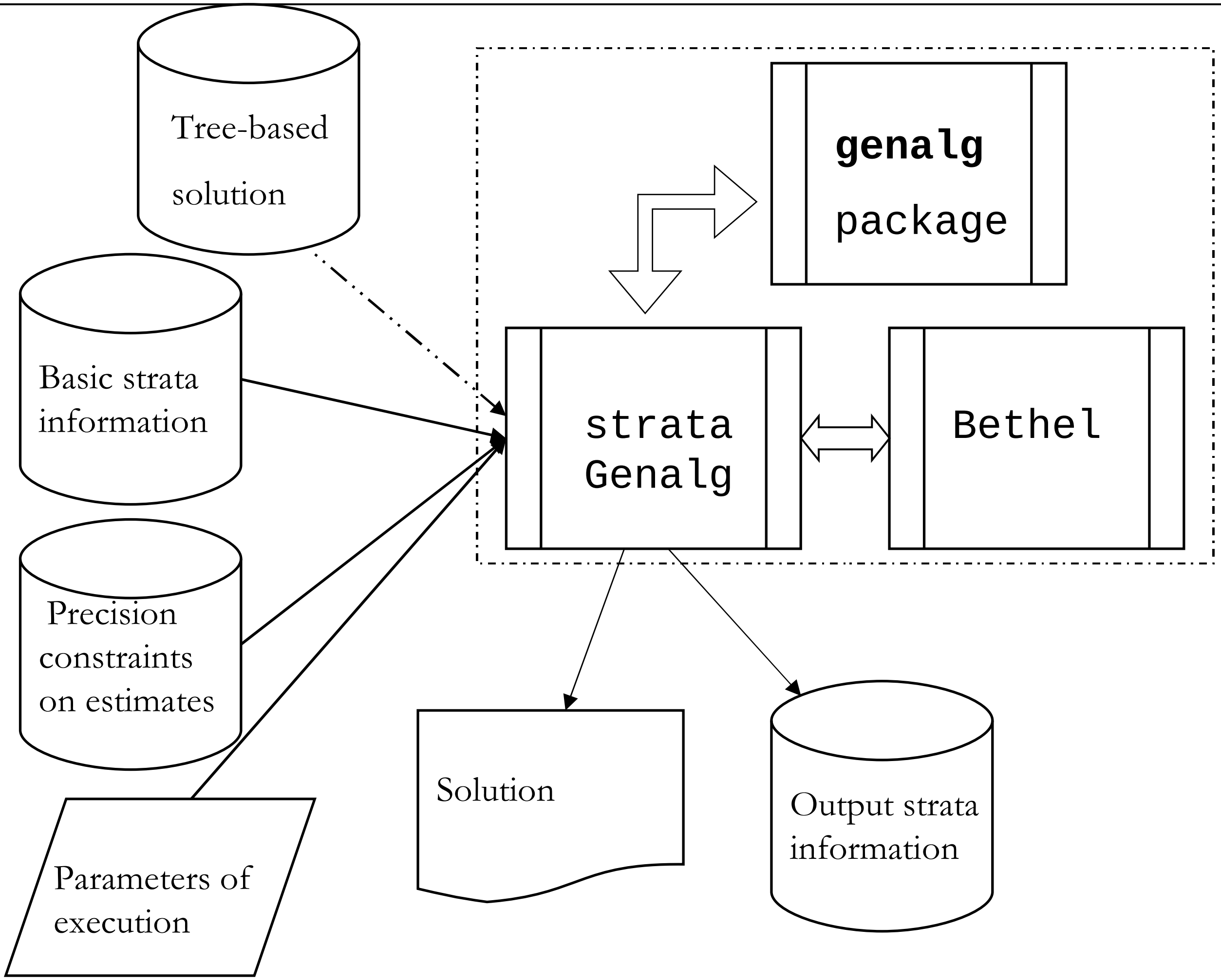
$$\begin{bmatrix} s_i[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,1,0] \\ s_j[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,1,0] \end{bmatrix}$$

mutation + crossover

generation j+1

$$\begin{bmatrix} s_1[x_1,\ldots x_{1m_i}] = [1,1,0,\ldots,0,1] \\ \ldots \\ s_i[x_1,\ldots x_{1m_i}] = [0,1,1,\ldots,1,0] \\ \ldots \\ s_j[x_1,\ldots x_{1m_i}] = [1,1,0,\ldots,1,1] \\ \ldots \\ s_t[x_1,\ldots x_{1m_i}] = [0,1,0,\ldots,0,1] \end{bmatrix}$$

Optimal stratification: the evolutionary approach (6)



An application: the Italian Farm Structure Survey

The sampling frame used for the selection of FSS sample contains 2,153,710 farms, each one characterised by the following X variables:

- provinces (103 different values);
- legal status (2 values);
- sector of economical activity (9 values);
- dimension in terms of production (3 values);
- dimension in terms of agricultural surface (3 values);
- dimension in terms of owned cattle (3 values)
- altimetry class (5 values).

14 different Y variables have been considered as the main target of FSS, on which required precision (in terms of maximum coefficient of variation) has been fixed at regional levels (domains of interest).

	(1) Current sample size	(2) Tree-based solution	% diff.
Italia	**52,713**	**29,726**	**-43.61**
Piemonte	3,560	1,546	-56.57
Valle d'A.	409	384	-6.11
Lombardia	5,125	2,237	-56.35
Bolzano	687	540	-19.94
Trento	667	638	-4.35
Veneto	3,873	2,299	-40.64
Friuli V.G.	1,262	619	-50.95
Liguria	1,327	777	-41.45
Emilia R.	3,117	1,966	-36.93
Toscana	2,833	1,341	-52.67
Umbria	1,363	858	-37.05
Marche	1,188	508	-57.24
Lazio	3,710	2,620	-29.38
Abruzzo	1,222	950	-22.26
Molise	1,183	867	-26.71
Campania	3,163	2,154	-31.90
Puglia	6,595	2,326	-64.73
Basilicata	965	684	-29.12
Calabria	2,846	2,080	-26.91
Sicilia	5,011	3,182	-36.50
Sardegna	2,607	1,140	-36.50

	(2) Tree-based solution	(3) evolutionary solution	% diff.
Italia	29,726	28,955	-2.59
Piemonte	1,546	1,546	0.00
Valle d'A.	384	376	-2.08
Lombardia	2,237	2,237	0.00
Bolzano	540	540	0.00
Trento	638	638	0.00
Veneto	2,299	2,138	-7.00
Friuli V.G	619	619	0.00
Liguria	777	657	-15.44
Emilia R.	1,966	1,933	-1.68
Toscana	1,341	1,310	-2.31
Umbria	858	858	0.00
Marche	508	498	-1.97
Lazio	2,620	2,620	0.00
Abruzzo	950	876	-7.79
Molise	867	719	-17.07
Campania	2,154	2,040	-5.29
Puglia	2,326	2,272	-2.32
Basilicata	684	684	0.00
Calabria	2,080	2,072	-0.38
Sicilia	3,182	3,182	0.00
Sardegna	1,140	1,140	0.00

Conclusions

In a sample survey design, the joint adoption of a consolidated algorithm for determining best sample size and units allocation, together with search techniques, as tree-based and genetic algorithm, to explore different possible stratifications, can be very convenient in situations where many different stratifications of a sampling frame are possible.

A limitation of this approach is in the constraint on the nature of auxiliary variables X, that must be categorical. An open problem is in the treatment of continuous X variables.