Tree-based and GA tools for optimal sampling design

Marco Ballin, Giulio Barcaroli
(ballin@istat.it, barcarol@istat.it)
ISTAT, via Cesare Balbo 16, 00184 Roma Italy

Keywords: sampling design, multivariate allocation,
optimal stratification, genetic algorithm, multivariate allocation

The optimality of a sample design can be defined in terms of costs (associated to fieldwork: number of units to be interviewed) and accuracy (sampling variance related to target estimates). Bethel proposed an algorithm (Bethel, 1985) able 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). Input to this algorithm is given by the information on distributional characteristics (total and variance) of target variables in the population strata. Under this approach, population stratification, i.e. the partition of the sampling frame obtained by cross-classifying units by means of potential stratification variables, is given. But stratification has a great impact on the optimal solution determined by Bethel algorithm and, in general, it must be defined in the first steps of a survey planning.

If a frame with a set of potential variables for stratification is available, the survey planner has to choose the "best" auxiliary variable cross product (partition of the frame). Among the possible partitions, the one with the maximum number of strata, given by the Cartesian product of all auxiliary variables, does not always yield the optimal sample size. In fact, organisational considerations, and the necessity to define a minimum amount of units per stratum, oblige not to increase the number of strata beyond a certain limit. In that case, how to determine the best partition among all partitions obtainable combining the auxiliary variables (what auxiliary variables? what values for each of them to take into consideration?) has to be considered as a part of the whole problem.

Until recently, on the contrary, the problem of determining the optimal size and allocation of units in strata has been solved considering the stratification of population as given; and, conversely, the definition of an optimal stratification has been investigated independently by the optimisation problem of sampling size and allocation.

An interesting proposal has been advanced in the recent past (Benedetti et. al 2005), offering a joint solution to both problems: it is based on a tree search in the space of possible strata configurations, solving for each visited node the corresponding multivariate allocation problem accordingly to Bethel algorithm. At each level, the node that is the best in terms of sample size reduction, is chosen as the branching node. This tree-based approach is deterministic and very fast, but it may heavily suffer for the presence of local minima and, consequently, solutions can be far from optimality.

Together with this tree-based approach, we propose a non deterministic evolutionary approach, based on the genetic algorithm (GA) paradigm. Under the GA approach, each solution (i.e. a particular partition in strata of the sampling frame) is an individual in a population, whose fitness is evaluated by calculating the sampling size satisfying accuracy constraints on the target estimates; crossover and mutation carried out along each iteration ensure an increase of average fitness.

In general, the characteristic of GA are such that the risk of local minima is lower than in the tree search, though processing time is noticeably higher. Our proposal is the following: in complex situations (characterised by a high number of stratification alternative configurations and/or a high number of target variables and domains), first the tree-based algorithm is applied, in order to individuate a solution. This solution is then introduced in the GA initial population, in order to speed its convergence to a better solution. Our experiments show an improvement of the tree-based solution, and encourage the adoption of this procedure.

The whole system can be thought of as a "toolkit", composed by a series of instruments, all implemented and operating in the R environment. Main scripts are:

1. strataTree.R implementing the tree-based algorithm;
2. strataGenalg.R that implements the GA approach, making use of "genalg" package (Willighagen, 2002) in a slightly modified version;
3. Bethel.R implementing Bethel algorithm.

These programs can be run directly in the R environment, but, as and an additional facility, a simple web interface has been developed using Rwui that enables the user to carry out the processing without being obliged to be acquainted with R language or even R environment.

References

Benedetti R., Espa G., Lafratta G. (2005), "A Tree-based Approach to Forming Strata in Multipurpose Business Surveys", *Discussion Paper No 5, 2005*, Università degli Studi di Trento – Dipartimento di Economia

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

Willighagen E. (2005), "*genalg: R Based Genetic Algorithm*". R package version 0.1.1. URL http:// cran.r-project.org/