JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **12**, Number **2**, April **2016**

doi:10.3934/jimo.2016.12.xx

pp. **xx–xx**

A COMBINATORIAL OPTIMIZATION APPROACH TO THE SELECTION OF STATISTICAL UNITS

Renato Bruni

Università di Roma "Sapienza", Dip. di Ing. Informatica,
Automatica e Gestionale (DIAG), Via Ariosto 25, Roma, 00185 Italy

Gianpiero Bianchi and Alessandra Reale

Italian National Statistic Office "Istat"
Dip. per i Censimenti e gli Archivi Amm. e Statistici (DICA)
Viale Oceano Pacifico 171, Roma, 00144 Italy

(Communicated by Oleg Prokopyev)

Abstract. In the case of some large statistical surveys, the set of units that will constitute the scope of the survey must be selected. We focus on the real case of a Census of Agriculture, where the units are farms. Surveying each unit has a cost and brings a different portion of the whole information. In this case, one wants to determine a subset of units producing the minimum total cost for being surveyed and representing at least a certain portion of the total information. Uncertainty aspects also occur, because the portion of information corresponding to each unit is not perfectly known before surveying it. The proposed approach is based on combinatorial optimization, and the arising decision problems are modeled as multidimensional binary knapsack problems. Experimental results show the effectiveness of the proposed approach.

1. **Introduction.** The *scope* of a statistical survey is the set of statistical *units* that should be surveyed. In the case of many large surveys, the scope cannot be the list of all possible units, because otherwise the cost or the complexity of the survey would be prohibitive. On the contrary, the scope should be selected, also on the basis of economical consideration. A typical example of this situation is the case of a Census of Agriculture, where, differently from a more traditional censuary vision, one needs to exclude from the survey the farms that are too small or otherwise irrelevant to the survey itself. A main problem, in similar cases, is establishing criteria or tracing somehow boundaries for dividing what should be included in the survey from what should be excluded from it *before* surveying the units.

From a conceptual point of view, there is a very large set of statistical units (e.g. farms, companies, etc.) that could be surveyed. Surveying each of them has a *cost* and represents a different portion of the whole statistical *information* under investigation (e.g., the state of the agriculture, the industrial production, etc.). Also, some *coverage levels* on that whole statistical information are assigned. Hence, one

2010 *Mathematics Subject Classification.* Primary: 90C90, 90C06; Secondary: 05A99.

Key words and phrases. Data mining, knowledge management, discrete optimization.

Work developed during a biennial research collaboration between Italian National Statistic Office (Istat) and University of Roma "Sapienza" on the data processing of the 2010 Census of Italian Agriculture.

would like to choose only a subset of the whole set of units. This subset should have the minimum total cost for being surveyed but should represent a portion of the whole information large enough for respecting the above coverage levels.

An important additional difficulty is that the portion of information carried by each unit is not perfectly known before surveying it. This often happens because the units may have been last surveyed only during a previous Census, typically held several years before, and their characteristics may have changed in the meanwhile. We therefore need to establish *reliable criteria* for deciding whether to include or not a unit in the survey on the basis of the (possibly outdated) available data describing the unit.

This problem is often called Scope Selection, or statistical Universe Selection, see e.g. [9, 10], and evidently contains a Combinatorial Optimization structure, with the optimal solution being one of the feasible subsets of the ground set of all the units. Defining a scope has connections with the general statistical task of Population Definition, see e.g. [13, 22]. The problem of the selection of statistical units is also studied in [11], where the authors consider the measurement of the lowest night temperature and select with an efficient heuristic the subset of units that allows the best linear prediction of the temperature in the excluded units. That procedure can also be used for other types of measures, but there are a number of structural differences between the problem in [11] and our case. Indeed, in that work, strong correlation hypothesis should hold on the values of the measure under prediction (a covariance matrix constant in time should be obtainable), and the decision of including or excluding each unit is based on one single type of measures. On the contrary, in our case, each statistical unit is described by a set of different values that are generally not correlated to those of other units, and the decision of including or excluding each unit must consider all the different values simultaneously. On the other hand, other optimization models arising from the treatment of agricultural data are described in [3, 4, 19], or in [7, 8] for population data. The problem of selecting units from a list can also be viewed as a particular case of the very important problem of quota sampling, for which several approaches and techniques of purposive sampling have been proposed, see also [16, 18, 23]. After the selection of a set of units, if uncertainty occurs in the data, one should also evaluate the risk of undercoverage [2]. While similar problems of statistical units selection have been solved in practice by means of a variety of *ad hoc* techniques, whose features typically depend on the specific application but which in any case heavily rely on the contribution of experts of the field, we propose a more general approach based on the use of a binary linear model solved by means of Combinatorial Optimization techniques. This innovative approach to the problem overcomes the particularity of ad hoc approaches, and moreover allows to take advantage of effective algorithms already developed in that field.

More precisely, by using binary variables associated with the above units, the described selection problem is here modeled as a multidimensional binary knapsack problem (see e.g. [21, 25]). Since those models may reach in many cases very large dimensions, a Branch&Cut approach using a separation procedure based on covers [1, 15, 24] has been used. A solution to the above knapsack model will be referred to as an Optimal Selection. However, due to the above described uncertainty aspects, not just one but a sequence of Optimal Selection problems must be solved for selecting the Scope in practice. Indeed, in order to develop *inclusion criteria* based on thresholds, we need to evaluate some safety margins with respect to the

risk of undercoverage for different inclusion thresholds. The procedure has been implemented in c++ and tested in cooperation with the Italian National Statistic Office (Istat) on real data from the Italian Census of Agriculture 2010. Results are very encouraging both from the computational and from the statistical point of view. The main contribution of this work is therefore an innovative and effective approach based on Combinatorial Optimization for solving a challenging large-sized and economically important real-world problem.

The work is organized as follows. Section 2 describes the basic model proposed for the selection of an optimal subset of statistical units and techniques to improve this formulation and solve the overall model by means of a Branch&Cut approach. Section 3 explains how the solution of the Optimal Selection problem under different conditions leads to the determination of the inclusion criteria based on thresholds. Finally, in Section 4, we provide extensive results on real-world data from the Italian Census of Agriculture.

2. **Solving the optimal selection problem.**

The model proposed for the above problem will be hereinafter explained by referring to the specific case of a Census of Agriculture. This is probably the most important case, because it has a great economic relevance and a very large dimension. Moreover, in the case of EU countries, gathered information must be published and provided to the EU level, where it constitutes a basis for assigning financial resources, for planning production, and for several other economical European policies. However, the proposed model is not intrinsically limited to that case, but can be used for similar cases of Scope Selection problems.

In a Census of Agriculture, there is a very large list $U = \{u_1, \ldots, u_n\}$ of all the existing statistical units that could be surveyed. Each unit u_i represents a farm, which is described by the areas used for every cultivation (plus other data not relevant for this work). Units have therefore the following structure:

$$u_i = \{a_{i1}, \ldots, a_{im}, a_{i\mathrm{T}}\} \qquad \text{for } \; i = 1, \ldots, n$$

where a_{ij} is the area that farm i uses for cultivation j, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$, and $a_{i\mathrm{T}}$ is the total area of farm i used for cultivations, technically called Utilized agricultural Area (UA). Unfortunately, at least for the majority of the cases, the data in the list U are those that were surveyed during the last census, often held several years before, or were obtained by other possible sources that may still be outdated. Hence, the available data may very well be different from the current farm situation. In particular, the single cultivation areas a_{ij} may easily have changed, while total cultivation area $a_{i\mathrm{T}}$ of the farm is more stable.

For every cultivation j that we are interested in, including some types of livestock, a certain *coverage level* $q_j \geq 0$ and ≤ 1 is required, with $j = 1, \ldots, m$. Value q_j represents the minimum portion of the total area of cultivation j that must be surveyed: if this total area is $\sum_{i=1}^{n} a_{ij}$, we need to survey at least an area $q_j \sum_{i=1}^{n} a_{ij}$ of cultivation j (e.g. survey at least 0.8 of the total cultivation of oranges, at least 0.5 of the total cultivation of apples, etc.). A required coverage level q_{T} for the total cultivation areas is also given. For the case of EU countries, coverage levels are generally assigned by European regulations, e.g. [9]. The set of coverage levels is denoted by

$$\{q_1, \ldots, q_m, q_{\mathrm{T}}\}$$

Surveying unit u_i has a cost w_i (that can be evaluated either in terms of expense, or complexity added to the whole survey, or other) and produce, for each cultivation j, an amount of statistical information that, in absence of further elements, is estimated being equal to the available value of the cultivation area a_{ij}. By defining the cost of a set of units to be the sum of their individual costs w_i, we want to choose a subset $S \subseteq U$ of units producing the minimum total cost for being surveyed and simultaneously respecting all the above defined $m+1$ coverage levels. Hence, the problem cannot be decomposed in subproblems addressing a single cultivation type at a time. In order to represent whether to include or not a unit, we introduce a set of binary decision variables $\{x_i\}$, with $i = 1, \ldots, n$, such that

$$x_i = \begin{cases} 1 & \text{if unit } u_i \text{ is excluded from the scope;} \\ 0 & \text{if unit } u_i \text{ is included in the scope.} \end{cases}$$

Now, cost minimization can be expressed by maximizing the total cost of the units that we do not survey (i.e. the saving). Respecting the coverage levels, on the other hand, can be expressed by imposing that the area that is excluded cannot be more than the maximum area we are allowed to exclude. This latter condition should be imposed both for each cultivation and for the total area. The described Optimal Selection problem can be modeled as the following *multidimentional binary knapsack* problem.

$$\begin{cases} \max & \displaystyle\sum_{i=1}^{n} w_i x_i \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} a_{i1}\ x_i \le (1-q_1) \sum_{i=1}^{n} a_{i1} \\ & \cdots \\ & \displaystyle\sum_{i=1}^{n} a_{im}\ x_i \le (1-q_m) \sum_{i=1}^{n} a_{im} \\ & \displaystyle\sum_{i=1}^{n} a_{i\mathrm{T}} x_i \le (1-q_\mathrm{T}) \sum_{i=1}^{n} a_{i\mathrm{T}} \\ & x_i \in \{0,1\} \end{cases} \tag{1}$$

Multidimentional binary knapsack is a well-known combinatorial optimization problem [21]; in its optimization version it is NP-hard. Note that a complementary choice for the meaning of the x_i variables (1 if unit u_i is included in the scope, 0 otherwise), that may appear more straightforward, would have led to a *multidimensional packing problem* [21], that has the same complexity level. However, the proposed modeling choice will have less variables at 1, with consequent computational advantages.

Model (1) has a number of variables equal to the number n of units in list U and a number of constraints equal to the number of coverage levels $m+1$, so it may reach in practical cases a very large dimension. Therefore, solving such an integer linear program by means of a simple Branch&Bound approach can be excessively time consuming (see e.g. [12]), and we use a Branch&Cut approach based on the generation of covering inequalities, as follows. Given a single knapsack constraint $\sum_{i=1}^{n} a_{ij}\ x_i \le e_j$ from (1), a set $C \subseteq \{1, \ldots, n\}$ is a *cover* if $\sum_{i \in C} a_{ij} > e_j$.

Given a cover C, we can write the following *covering* inequality expressing that not all variables x_i , for $i \in C$, can be simultaneously one:

$$\sum_{i \in C} x_i \leq |C| - 1 \tag{2}$$

Covering inequalities of the above type are valid for a problem containing the knapsack constraint $\sum_{i=1}^{n} a_{ij}\ x_i \leq e_j$ ([1, 15, 24]). Therefore, an inequality of type (2) can be added to model (1) for improving the formulation given by its linear relaxation. In the case of the described Census, the meaning of the above inequality is that we simply cannot exclude from the survey a set C of units having a total area that is too large. It is evidently impracticable to generate all inequalities in the form (2), since their number can be too large. Therfore, we generate covering inequalities within a Branch&Cut scheme (see e.g. [21]). This requires to solve recursively the so called *separation problem*, that is, given a point $\bar{x} \in R^n$ and a polytope K, either to prove that $\bar{x} \in K$ or find an inequality, called *cut* or *cutting plane*, that is valid for K but cuts $\bar{x}$ away from K.

Denote by c the incidence vector of the generic subset C of $\{1, \ldots, n\}$, i.e. the binary n-vector whose i-th element is 1 if $i \in C$, 0 otherwise. By recalling that $a_{ij} \in R$ and ≥ 0, set C is a cover for the j-th knapsack constraint of problem (1) if and only if its incidence vector c satisfies the following condition

$$\sum_{i=1}^{n} a_{ij} c_i > (1 - q_j) \sum_{i=1}^{n} a_{ij}$$

that, by introducing $\epsilon > 0$ equal to the minimum possible difference in the a_{ij} values, can be rewritten as

$$\sum_{i=1}^{n} a_{ij} c_i \geq [(1 - q_j) \sum_{i=1}^{n} a_{ij}] + \epsilon \tag{3}$$

Among all possible covers, we want a cover C such that the components of $\bar{x}$ corresponding to elements of C sum to a value $> |C| - 1$, that means $\bar{x}$ can be cut away by the cutting plane generated by C. Cover C represents in practice a set of units that cannot be simultaneously excluded from the survey, but are actually excluded in solution $\bar{x}$. The condition of cutting away $\bar{x}$ can be expressed as follows:

$$\begin{aligned} &\sum_{i \in C} \bar{x}_i > |C| - 1 \Rightarrow \sum_{i=1}^{n} \bar{x}_i c_i > \sum_{i=1}^{n} c_i - 1 \Rightarrow \\ &\Rightarrow \sum_{i=1}^{n} (\bar{x}_i - 1) c_i > -1 \Rightarrow \sum_{i=1}^{n} (1 - \bar{x}_i) c_i < 1 \end{aligned} \tag{4}$$

Putting together condition (3) and (4) we have the following optimization problem encoding our separation procedure for the j-th knapsack constraint of problem (1).

$$\begin{cases} \min & \displaystyle\sum_{i=1}^{n} (1 - \bar{x}_i) c_i \\ s.t. & \displaystyle\sum_{i=1}^{n} a_{ij} c_i \geq [(1 - q_j) \sum_{i=1}^{n} a_{ij}] + \epsilon \\ & c_i \in \{0, 1\} \end{cases} \tag{5}$$

When model (5) is solved to optimality, we obtain vector $c^\star$ and the corresponding objective value $v^\star = \sum_{i=1}^{n} (1 - \bar{x}_i) c_i^\star$. If $v^\star$ is < 1, there exists a covering inequality

that is valid for K but cuts away $\bar{x}$ from K, and $c^\star$ is the incidence vector of the cover $C^\star$ generating that cover inequality. On the contrary, if $v^\star$ is ≥ 1, we cannot obtain from the j-th knapsack constraint a covering inequalities cutting away $\bar{x}$. In this latter case, we must try to obtain it from another one of the knapsack constraints of (1). If such a covering inequality cannot be obtained after all the knapsack constraints have been tested, it does not exist. Therefore, solving several problems in the form (5) may be needed. Although (5) is a binary problem with n variables, it can be solved quite easily, since any feasible solution $\hat{c}$ of of (5) such that the corresponding objective value $\hat{v}$ is < 1, even if sub-optimal, is the incidence vector of a cover $\hat{C}$ whose cover inequality cuts away $\bar{x}$ from K. Therefore, we may accept those kind of solutions, and search them with the following procedure. Recall that every variable c_i has a *cost* given by $(1 - \bar{x}_i)$ and a *value* a_{ij}. We denote by RHS the right-hand side of the only constraint in (5).

Greedy Algorithm for the solution of problem (5)

> **Input** An instance of separation problem (5), defined by the cost n-vector $(1 - \bar{x})$, the values n-vector a_j and a value RHS.
> **Output** A binary feasible (and possibly optimal) solution $\hat{c}$ to (5)

1. Order by increasing cost/value ratio the indices of the binary variables.
2. Following the above greedy order, put $\hat{c}_i = 1$ until the left-had side of the constraint becomes $\geq RHS$ (i.e. we have a feasible solution), and $\hat{c}_h = 0$ for all the rest of the indices. If the value of this solution is $\hat{v} < 1$, there exists a covering inequality that is valid for K but cuts away $\bar{x}$ from K, and $\hat{c}$ is the incidence vector of the cover generating it.

The above heuristic solution could be evaluated by using the lower bound given by the solution $r^\star$ of the linear relaxation of problem (5): we put $r_i = 1$ until the left-had side of the constraint remains $\leq RHS$ (i.e. we have a maximal infeasible solution), then $r_{(i+1)} = (RHS - LHS)/a_{(i+1)j}$, and finally $r_h = 0$ for all the rest of the indices. If the objective value corresponding to $r^\star$ is $v_r^\star \geq 1$, the value $v^\star$ of the integer solution of (5) is $\geq v_r^\star$, so we know that the j-th knapsack constraint cannot provide a covering inequalities cutting away $\bar{x}$. On the contrary, when $v_r^\star < 1$ but $\hat{v} \geq 1$, a covering inequalities cutting away $\bar{x}$ may exist, but was not found by the procedure. Nonetheless, we try to obtain it from another one of the knapsack constraints of (1), and if none of them can provide it, we simply branch following the mentioned Branch&Cut scheme. This technique guarantees to reach the optimal binary solution of model (1).

3. **Determining reliable inclusion criteria.** Since data are uncertain, an optimal solution $x^\star$ of model (1) cannot guarantee providing a set of units really respecting the required coverage levels. Indeed, as a trivial example, if the real cultivation areas of some of the selected farms have become smaller than what described by the available data a_{ij}, the risk of *undercoverage* (i.e. failing the required coverage levels q_j) is present. Hence,we need to distinguish between:

- solving the *Optimal Selection* problem, that is solving to optimality model (1);
- solving the *Scope Selection* problem, that is finding the set of units that we use as scope in practice.

For solving the Scope Selection problem, we need to determine a priori *inclusion criteria* for selecting the set of units respecting the coverage levels. A priori means here criteria that, for each unit u_i, could be checked *before* surveying u_i. A basic and mostly adopted criterion is using *thresholds*. Given a threshold value t_j, for each unit u_i one could determine whether to survey it or not: we survey u_i if $a_{ij} \geq t_j$, we do not survey it otherwise. Since in the analyzed case the total utilized area (UA) is the more reliable among the available informations, it was preferred to establish a threshold t only on that value.

The coverage levels, initially required by EU [9] and assigned for the whole Nation, were modified and slightly increased so as to determine more specific coverage levels assigned for each Region. Those new levels were determined by experts of the field according to specific features of the different regions, whose description goes beyond the aim of this work. The final established regional coverage levels, for Citrus plantations, Fruit trees cultivation, Olive cultivations, Arable land, Vineyard cultivation and UA, are reported in Table 1. A'-' denotes that the value is irrelevant because that cultivation is not used in that region.

TABLE 1. Regional coverage levels

Region	Citrus	Fruit	Olive	Arable land	Vineyard	UA
Piemonte	-	98.5	90.7	99.5	98.7	99.2
Valle d'Aosta	-	81.1	-	84.0	83.9	98.6
Lombardia	96.3	93.2	88.3	99.7	-	99.4
Trentino Alto Adige	-	99.3	66.1	95.8	97.8	98.8
Veneto	-	97.4	95.4	99.3	98.7	98.3
Friuli-Venezia Giulia	-	98.0	88.8	98.5	99.1	98.4
Liguria	68.4	84.8	89.5	92.6	82.0	92.7
Emilia-Romagna	-	98.7	91.6	99.6	99.5	99.4
Toscana	68.4	95.0	97.5	99.1	98.4	98.3
Umbria	-	94.8	96.9	98.8	97.1	98.5
Marche	80.3	94.2	94.3	99.1	98.6	98.8
Lazio	68.3	97.4	92.8	98.5	94.6	97.0
Abruzzo	85.1	94.6	96.0	98.2	99.1	98.5
Molise	-	96.3	96.2	99.1	97.2	98.7
Campania	82.4	97.2	95.3	96.8	94.5	96.7
Puglia	98.6	97.4	97.6	98.7	99.4	98.4
Basilicata	96.5	96.3	95.6	98.9	95.7	98.6
Calabria	98.0	97.6	97.1	96.3	95.0	97.3
Sicilia	97.4	97.0	94.6	97.8	99.2	97.6
Sardegna	93.6	93.9	95.8	99.4	97.4	99.3

In order to satisfy the above regional coverage levels, one may in general consider different options. A first one could be solving the regionals Optimal Selection problems, in order to determine, for each region, a selected set of units $U^\star$. After this, use $U^\star$ to determine the value a_{iT} of UA corresponding to the smallest included farm of the region, then compute how reliable that value is, possibly modify it for having a safety margin, and use it as inclusion threshold (at the regional level).

Another alternative would be fixing, according to some predetermined decision, a number of threshold values on the total utilized area (UA), and solve the regional Optimal Selection problems for each of them. Given a threshold t, denote by U_t the

set $\{u_i \in U : a_{i\mathrm{T}} \geq t\}$ and by $U_t^\star$ the result of the Optimal Selection on U_t (the set of units $u_i \in U_t$ having $x_i = 0$). Define $R_t = U_t - U_t^\star$ the set of units that satisfy threshold t but are not in the solution of the Optimal Selection (i.e. the set of units $u_i \in U_t$ having $x_i = 1$). Sets U_t and R_t can now be used for computing statistical indicators that evaluate the safety margin with respect to the risk of undercoverage obtained by using t as inclusion threshold. After this, the threshold value $t^\star$ that, among the predetermined ones, corresponds to the best compromise between risk estimation and list reduction for the region, is selected and adopted as inclusion criterion (again at the regional level). The solution to the Scope Selection problem would therefore be $U_{t^\star}$ This last option was preferred in the analyzed case, because it could provide more robustness in the procedure, in the sense of making it more stable and less prone to the changes that may have occurred in the data.

We also remark that a third alternative approach could be based on the computation, for each unit u_i, of some upper bound UB_{ij} and lower bound LB_{ij} of the amount of information (i.e., the currently used cultivation area) that unit u_i brings for cultivation j. When $\sum_{i=1}^n UB_{ij} \leq (1 - q_j) \sum_{i=1}^n LB_{ij}$, we could reasonably expect that the coverage level q_j is respected. Determining upper and lower bounds for the probability of an event logically related to a set of other events is a well-known problem in Statistics and Probability theory for which a number of techniques have been proposed (see, e.g., the seminal work [14], and the recent [6]). However, this option was not selected because of the heavy assumptions required for those computations.

We now describe the statistical indicators that were built by experts of the field for evaluating the safety margin from the risk of undercoverage corresponding to each threshold t. A basic index number is the percentage of farms taken in addition to $U_t^\star$ when using threshold t, denoted by $\beta(t)$ and computed as follows.

$$\beta(t) = 100 \frac{|R_t|}{|U_t|}$$

The larger the value of $\beta(t)$, the more threshold t is able to provide a set U_t that is bigger than the minimum set respecting the coverage levels, and consequently the more secure is the selection by using threshold t.

Another index number, for each cultivation j, is the percentage of cultivation area taken as safety margin when using threshold t, denoted by $\gamma_j(t)$ and computed as follows.

$$\gamma_j(t) = 100 \, \frac{\sum\limits_{u_i \in R_t} a_{ij}}{\sum\limits_{u_i \in U_t} a_{ij}}$$

Again, the larger the value of $\gamma_j(t)$, the more threshold t is able to provide an area $\sum_{u_i \in U_t} a_{ij}$ that is bigger than the minimum area respecting the coverage levels, and consequently the more secure is the selection by using threshold t. Clearly, the higher the values of t, the smaller the above safety margins become, but the larger the savings in the survey are. Therefore, we need to chose the higher t still having acceptable values for the above indicators, so as to obtain the maximum savings with negligible risk.

Given a set of farms U_t, define $S_j(U_t) = \{u_i \in U_t : a_{ij} > 0\}$ to be the subset of farms in U_t having cultivation j. Denote now by $\mu\{S_j(U_t)\}$ the average area of cultivation j over set $S_j(U_t)$. A third indicator, for each cultivation j, is the

average number of units in R_t that are needed to obtain a portion of information that is equivalent to the portion of information given by an average unit of $U_t^\star$, or, in other words, the average number of units in R_t needed to replace a unit in $U_t^\star$ (for example in case the latter one does not exist anymore). That will be denoted by $\omega_j(t)$ and computed as follows.

$$\omega_j(t) = \frac{\mu\{S_j(U_t^\star)\}}{\mu\{S_j(U_t) - S_j(U_t^\star)\}}$$

The $\omega(t)$ is clearly ≥ 1, and the smaller the values, the more robust is the choice of threshold t. Some values of β, γ and ω for the considered real-world case are reported in the following Table 4.

4. **Experimental results.** The described procedure has been implemented in C++ and tested for the treatment of data from the Italian Census of Agriculture 2010 (VI Censimento Generale dell'Agricoltura 2010). The experiments were conducted on a 16 cores server having 128Gb of RAM under Linux Operating System. The linear relaxations of (1) are solved by means of the open source solver Clp (Coin-or linear programming, available from https://projects.coin-or.org/Clp), which is a very good implementation of primal and dual simplex and barrier methods, written in C++ by a research group headed by Dr. John J. Forrest, from the IBM Watson Research Center, within a joint project among IBM, Maximal and Schneider called COIN-OR (COmputational INfrastructure for Operations Research, http://www.coin-or.org/index.html). This solver was selected because it appeared the most suitable open source LP solver in a previous study [5]. As described in the previous Section, a number of predetermined threshold levels on the total utilized area (UA) have been used. Their values were 0.0 (meaning that all farms are included, even the smallest ones); 0.1; 0.2; 0.3; 0.4 hectares.

Table 2 reports the detail of this analysis for one sample Italian region (Marche). Evidently, when increasing the inclusion threshold t, the cardinality of U_t decreases consistently (less farms satisfy that threshold). On the other hand, the cardinality of $U_t^\star$ does not decrease. It generally remains the same, even if it may occasionally slightly increase. This happens because, intuitively, when searching an optimal solution $U_t^\star$ within U_t, the smaller is U_t, the less choices we have for selecting the minimum-cost solution satisfying the coverage levels.

Consequently, the differences between U_t and $U_t^\star$ tend to become smaller, and the described β, γ tend to decrease. Therefore, when increasing t, there is a trade-off between the reduction in the cost and complexity of the Census, caused by the decreasing of $|U_t|$, and the rise in the risk of undercoverage, evaluated by the β and γ index numbers. In the case of a Census, the priority is having an acceptable risk level, so we are interested in maximizing the savings obtainable in correspondence to acceptable values of β and γ. Acceptable values for the β and γ indicators have been considered those respectively above 10% and 0.5%. On the contrary, values for the ω indicator should be as small as possible. Hence, for the case of Marche region, the best compromise is $t = 0.4$, corresponding to a very acceptable risk level (β and γ are considerably above the lowest acceptable values) but producing considerable savings: the cost of surveying 66,536 - 60,309 = 6,254 farms, that is about 10%.
Table 3 reports, for each Italian region, the number $|U|$ of all existing statistical units; the number $|U^\star|$ of units selected from U when solving model (1) to optimality; the value of threshold $t^\star$ selected among the predetermined values as the best compromise between list reduction and risk of undercoverage; the number $|U_t|$ of

TABLE 2. Results of the procedure applied to the Marche Region

Thershold	$\lvert U_t\rvert$	$\lvert U_t^\star\rvert$	β	γ_{vineyard}	ω_{vineyard}	γ_{olive}	ω_{olive}	γ_T
0.0	66,563	50,051	24.81	3.50	6.10	7.51	3.45	1.35
0.1	65,438	50,051	23.51	3.50	6.10	7.51	3.45	1.35
0.2	64,374	50,051	22.25	3.47	5.95	7.38	3.31	1.34
0.3	62,474	50,051	19.86	3.32	5.61	6.68	3.11	1.29
0.4	60,309	50,057	17.01	3.00	5.27	5.41	3.07	1.20

existing statistical units that are above threshold t; the number $|U_t^\star|$ of units selected from U_t when solving model (1) to optimality; computational time in seconds for the overall treatment of the region, including the solution of the five Optimal Selection problems and the evaluation of the described indicators (Time All); computational time in seconds for solving to optimality the single Optimal Selection problem (1) corresponding to $t^\star$.

TABLE 3. Results of the procedure applied to all Italian Regions

Region	$\lvert U\rvert$	$\lvert U^\star\rvert$	$t^\star$	$\lvert U_t\rvert$	$\lvert U_t^\star\rvert$	Time All	Time $t^\star$
Piemonte	120,965	78,651	0.3	103,347	78,651	605.2	116.0
Valle d'Aosta	6,595	4,050	0.4	5,441	4,051	33.5	7.6
Lombardia	74,867	56,949	0.3	69,890	56,949	374.5	73.8
Trentino Alto Adige	61,253	33,804	0.2	51,816	33,804	306.8	61.2
Veneto	191,085	118,204	0.3	176,251	118,204	955.2	186.0
Friuli-Venezia Giulia	34,963	25,455	0.3	32,953	25,455	175.6	34.1
Liguria	44,266	21,654	0.3	34,167	21,654	221.0	43.2
Emilia-Romagna	107,888	89,468	0.3	103,744	89.468	539.5	104.8
Toscana	139,872	77,823	0.3	119,788	77,823	699.0	136.8
Umbria	57,153	36,538	0.3	51,772	36,538	286.3	57.2
Marche	66,563	50,051	0.4	60,309	50,057	333.0	65.6
Lazio	214,666	123,026	0.3	189,906	123,026	1,073.3	210.6
Abruzzo	82,833	58,478	0.3	78,036	58,478	414.2	81.8
Molise	33,973	25,285	0.3	31,955	25,285	170.8	36.2
Campania	248,932	143,318	0.3	216,635	143,319	1,245.0	246.0
Puglia	352,510	229,118	0.2	348,380	229,118	1,763.0	346.6
Basilicata	81,922	58,460	0.3	76,307	58,460	410.5	82.1
Calabria	196,484	113,719	0.3	173,866	113,719	982.2	193.4
Sicilia	365,346	223,912	0.2	355,038	223,912	1,827.6	358.4
Sardegna	112,689	76,355	0.2	108,545	76,355	563.5	108.6

Table 4 reports, for each Italian Region, the values of some of the described statistical indicators corresponding to the selected threshold $t^\star$. In particular, we report

$$\beta, \gamma_{\text{vineyard}}, \omega_{\text{vineyard}}, \gamma_{\text{olive}}, \omega_{\text{olive}}, \gamma_{\text{T}}.$$

Vineyard and olive were selected because they are particularly important, being probably the two most typical Italian cultivations, and being subject to several EU regulations. In the real Census application several other cultivations were also considered.

Table 5, finally, summarizes the Italian situation. It reports, for the case of threshold $t = 0.0$ (all U) and for the threshold $t^\star$ reported for each region in Table 3, the total number of farms, their total utilized area UA, the number of farms obtained for the optimal solution, and their total UA. As showed in this last Table, when using as inclusion criterion the $t^\star$ values of Table 3, we have a reduction in the number of farms of 206,679, corresponding to a saving of 7.97%, that is worth of note (considering the large cost of a Census), and a reduction in the cultivation area (Total UA) of about 50,948 hectares, corresponding to a loss of only 0.39% of the

TABLE 4. Values of the indicators corresponding to the selected thresholds

Region	β	$\gamma_{vine.}$	$\omega_{vine.}$	γ_{olive}	ω_{olive}	γ_{T}
Piemonte	23.90	4.42	7.35	10.02	3.32	1.15
Valle d'Aosta	25.55	11.01	2.62	-	-	1.15
Lombardia	18.52	4.19	8.23	10.53	4.22	0.67
Trentino Alto Adige	34.76	6.59	7.37	33.88	2.43	1.71
Veneto	32.93	5.12	7.07	8.08	3.72	3.64
Friuli-Venezia Giulia	22.75	2.90	8.13	12.50	5.24	1.77
Liguria	36.62	16.25	2.94	12.24	4.09	7.18
Emilia-Romagna	13.76	3.03	5.25	9.86	2.83	0.74
Toscana	35.03	4.36	9.37	5.97	6.92	2.31
Umbria	29.43	5.26	6.02	6.29	5.44	1.98
Marche	17.02	3.00	5.27	5.41	3.07	1.20
Lazio	35.22	8.18	4.49	9.88	4.29	3.98
Abruzzo	25.06	4.19	5.35	7.20	3.82	2.16
Molise	20.87	4.42	4.27	6.47	3.30	1.52
Campania	33.84	8.47	3.45	8.34	3.94	5.22
Puglia	34.23	5.29	5.59	7.00	6.42	3.56
Basilicata	23.39	5.54	3.53	7.24	3.55	1.58
Calabria	34.59	7.79	3.98	6.83	6.38	4.53
Sicilia	36.93	5.11	7.33	9.48	5.16	3.81
Sardegna	29.66	6.48	4.64	7.96	5.08	1.09

total information, and this with a negligible risk of failing the required coverage levels, as observable from the values of β and γ in Table 4.

From the same Table 5 we also observe that, if the cultivation data were updated and reliable, the set $U_0^\star$ could have been surveyed directly, with a reduction in this case of 950,506 farms, corresponding to an even larger saving of 36.63%, and a reduction in the cultivation area (Total UA) of about 204,848 hectares, corresponding to a loss of only 1.55% of the total information, with the guarantee of respecting the coverage levels. These results were possible because of the existence, in the Italian territory, of a large number of very small farms that however constitute only a very small portion of the total national cultivation area. Such a set of farms would be quite expensive to survey but would not provide an amount of information justifying that spending, so their detection has allowed considerable savings. We moreover stress that the computational times required to solve those large-sized problems with the proposed procedure are extremely moderate.

TABLE 5. Aggregate results at the National level

Threshold	$\lvert U_t\rvert$	Total UA for U_t	$\lvert U_t^\star\rvert$	Total UA for $U_t^\star$
0.0	2,594,825	13,206,296.76	1,644,319	13,001,448.97
$t^\star$	2,388,146	13,155,349.09	1,644,315	13,003,198.84

5. **Conclusions.** We proposed an innovative approach to the Scope Selection problem based on Combinatorial Optimization. The proposed multidimensional knapsack model can be solved to optimality in short times by means of a Branch&Cut

algorithm based on the generation of cover inequalities. The procedure has been implemented and tested on the real-world case of an Italian national agricultural Census. By solving the Optimal Selection problem in different conditions, statistical indicators for the determination of reliable inclusion criteria based on thresholds have been computed. The proposed approach allows to considerably reduce costs and complexity of the survey while ignoring only a very small portion of the whole information that can be surveyed. The risk of failing the required coverage levels, i.e., the risk that such ignored portion was larger than the maximum admissible portion that we are authorized to ignore, is negligible.

Acknowledgments. The authors are grateful to Dr. Michela Di Lullo, from "Sapienza" University of Rome, for useful discussions on statistical aspects.

REFERENCES

[1] E. Balas, Facets of the knapsack polytope, *Mathematical Programming*, **8** (1975), 146–164.
[2] R. M. Bell and M. L. Cohen, *Coverage Measurement in the 2010 Census - Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census,* The National Academic Press, Washington, D.C., 2008.
[3] G. Bianchi, R. Bruni and A. Reale, Information reconstruction via discrete optimization for agricultural census data, *Applied Mathematical Sciences*, **6** (2012), 6241–6251.
[4] G. Bianchi, R. Bruni and A. Reale, Balancing of agricultural census data by using discrete optimization, *Optimization Letters*, **8** (2014), 1553–1565.
[5] G. Bianchi, R. Bruni and A. Reale, Open source integer linear programming solvers for error localization in numerical data, In *Advances in Theoretical and Applied Statistics* (eds. N. Torelli, F. Pesarin and A. Bar-Hen), Springer, New York, NY, 2012.
[6] E. Boros, A. Scozzari, F. Tardella and P. Veneziani, Polynomially computable bounds for the probability of the union of events, *Mathematics of Operations Research*, **39** (2014), 1311–1329.
[7] R. Bruni, Discrete models for data imputation, *Discrete Applied Mathematics*, **144** (2004), 59–69.
[8] R. Bruni, Error correction for massive data sets, *Optimization Methods and Software*, **20** (2005), 295–314.
[9] European Parliament, *Regulation of the European Parliament*, N. 1166/2008, 2008.
[10] Food and Agriculture Organization of the United Nations (FAO), *A System of Integrated Agricultural Censuses and Surveys,* Vol.1 - World Programme for the Census of Agriculture 2010. FAO Statistical Development Series (2005).
[11] M. Ferri and M. Piccioni, Optimal selection of statistical units: An approach via simulated annealing, *Computational Statistics & Data Analysis*, **13** (1992), 47–61.
[12] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness,* W.H. Freeman, San Francisco, CA, 1979.
[13] R. M. Groves, F. J. Jr. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, *Survey Methodology*, Wiley Series in Survey Methodology, John Wiley & Sons Inc., Hoboken, NJ, 2009.
[14] T. Hailperin, Best possible inequalities for the probability of a logical function of events, *The American Mathematical Monthly*, **72** (1965), 343–359.
[15] P. L. Hammer, E. L. Johnson and U. N. Peled, Facets of regular 0-1 polytopes, *Mathematical Programming*, **8** (1975), 179–206.
[16] W. K. Kremers, Completeness and unbiased estimation for sum-quota sampling, *Journal of the American Statistical Association*, **81** (1986), 1070–1073.
[17] H. Marchand, A. Martin, R. Weismantel and L. Wolsey, Cutting planes in integer and mixed integer programming, *Discrete Applied Mathematics*, **123** (2002), 397–446.
[18] C. A. Moser, Quota sampling, *Journal of the Royal Statistical Society*, **115** (1952), 411–423.
[19] A. Mucherino, P. Papajorgji and P. M. Pardalos, *Data Mining in Agriculture*, Springer, New York, NY, 2009.
[20] K. G. Murty, *Linear programming,* John Wiley & Sons Inc., New York, NY, 1983.

[21] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley, New York, NY, 1988.
[22] T. M. F. Smith, Populations and selection: Limitations of statistics, *Journal of the Royal Statistical Society* Series A (Statistics in Society), **156** (1993), 144–166.
[23] S. Sudman, Probability sampling with quotas, *Journal of the American Statistical Association*, **61** (1966), 749–771.
[24] L. A. Wolsey, Faces for a linear inequality in 0-1 variables, *Mathematical Programming*, **8** (1975), 165–178.
[25] Y. Zhang, F. Zhang and M. Cai, Some new results on multi-dimension Knapsack problem, *Journal of Industrial and Management Optimization*, **1** (2005), 315–321.

Received April 2014; 1st revision July 2014; 2nd revision November 2014.

E-mail address: bruni@dis.uniroma1.it
E-mail address: gianbia@istat.it
E-mail address: reale@istat.it