Survey Research Methods (2012)
Vol.6, No.3, pp. 125–135
ISSN 1864-3361
http://www.surveymethods.org

Multivariate Boundaries of a Self Representing Stratum of Large Units in Agricultural Survey Design

Roberto Benedetti
"G. d'Annunzio" University of Chieti-Pescara

Federica Piersimoni
Istat, Italian National Institute of Statistics

In business surveys in general, and in multipurpose agricultural surveys in particular, the problem of designing a sample from a list frame usually consists of two different aspects. The first is concerned with the choice of a rule for stratifying the population when several size variables are available and the second is devoted to sample size determination and sample allocation to a given set of strata. The main property that is required of the sample design is that it delivers a specified level of precision for a set of variables of interest using as few sampling units as possible. This article examines how this can be achieved via a basic partition into two strata, one completely enumerated and the other sampled, defined in such a way as to achieve both these objectives.

The procedure was used to design the Italian Milk Products Monthly Survey on the basis of a set of auxiliary variables obtained from an annual census of the same target population. Given the combinatorial optimization nature of the problem, we use stochastic relaxation theory, and in particular, we use simulated annealing because of its flexibility. Our results indicate that in this situation the multivariate partition obtained by using this random search strategy is a suitable solution as it permits identification of boundaries of any shape. Furthermore, numerical comparisons between sampling designs obtained by using these procedures and some simple extensions of univariate stratification rules are made. The gain from using the proposed strategy is nontrivial as it achieves the required precision using a sample size that is notably smaller than that required by simple extensions to univariate stratification rules.

Keywords: skewed population distribution, sample design, sample allocation, stratification, combinatorial optimization, simulated annealing

1 Introduction

Agricultural statistics are generated within the framework of the European Union (EU) programme on surveys on farm structure conducted by all EU members in order to have up-to-date and comparable information on the Member States, with the purpose of being used as a basic tool for designing the Common Agricultural Policy (CAP). However, to perform a well defined analysis of the agricultural sector, it is important to follow the whole chain of agricultural products transformation. Thus the farms data are necessarily used jointly with the information arising from surveys aimed to estimate the input and output of firms belonging to the sector defined by agro-alimentary transformation and commercialization.

In both the farms and transformation firms populations, the structure of the European economy is extremely different among and within each Member State. The variability of farms and establishments sizes is a relevant structural characteristic at Community level. Most of these units have a small size and are not important in economic terms even if they are interesting for the analysis of rural development. On the other hand, a limited number of large units represents a relevant part of the population in standard gross margin (SGM) terms and so have to be always included in any sample survey.

This is a typical situation in any business survey, in which the population of interest is extremely positively skewed because of the presence of few "large" units and many "small" units. Thus, when estimating an unknown total of the population, many small observations give a negligible contribution, whereas few large observations have a dramatic impact on the estimates.

In sampling theory the large concentration of the population with respect to surveyed variables constitutes a problem which is difficult to handle without the use of selection probabilities proportional to a size measure or by use of a stratification or partition tool. The first strategy is quite difficult, even if possible, to be extended to situations in which multiple auxiliaries are available (Benedetti and Piersimoni, Submitted for publication), while the second strategy can be dealt with by the introduction of a take-all (Censused) stratum and of one or more take-some (Sampled) strata. This procedure is commonly used by National Statistical Institutes (NSI) to select samples, even if it is not easy to give a unique definition of the boundaries of such strata when they have to be based on a multivariate set of size measures. Roughly speaking, this solution consists in partitioning the population in two sets of units: a take-all stratum whose units are en-

Contact information: Roberto Benedetti, "G. d'Annunzio" University of Chieti-Pescara, Department of Economic Studies, Italy, e-mail: benedett@unich.it

tirely surveyed (Censused - *C*) and a take-some stratum from which a simple random sample is drawn (Sampled - *S*).

This approach is not new and has been widely employed by survey practitioners, often using a heuristic rule for determining the part of the population to be censused (for example, firms with more than one hundred employees). This way of proceeding, typically motivated by the desire to match administrative criteria, usually ignores the statistical implications on the precision of the estimates.

The univariate methodological framework for this problem was suggested by Hidiroglou (1986) who proposed an algorithm for the determination of the optimal boundary between the two strata *C* and *S*. In the literature several formal extensions to the univariate optimal determination of the boundaries between more than two strata have been proposed (Dalenius and Hodges 1959; Singh 1971; Lavallée and Hidiroglou 1988; Hedlin 2000; Lu and Sitter 2002; Gunning and Horgan 2004 and 2007; for a review see Horgan 2006) through the use of algorithms that usually derive simultaneously the sample size needed to guarantee a fixed accuracy level for the resulting estimates and the sample allocation to the strata. Rivest (2002) proposed a generalization of these algorithms, extended in Baillargeon and Rivest (2009), to be used when the survey variable and the stratification variable differ. However these classical methods deals only with the univariate case and cannot be easily extended when there are multiple covariates for stratification.

Consider a finite population $U = \{1, 2, ...N\}$ recorded on a list frame together with a set of k positive auxiliary variables $\mathbf{X} = \left\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_j}, ..., \mathbf{x_k}\right\}$ where $\mathbf{x_j} = \left\{\mathbf{x_{1j}}, \mathbf{x_{2j}}, ..., \mathbf{x_{ij}}, ..., \mathbf{x_{Nj}}\right\}$ is the generic *j-th* auxiliary. This is a typical situation in many business surveys, particularly in the agricultural sector, and NIS's usually make huge efforts to design surveys that are based on an efficient use of all the available auxiliary information in order to obtain more precise and reliable estimates (Bee et al. 2010; Hidiroglou and Laniel 2001; Hidiroglou and Srinath 1993).

A traditional approach to dealing with multivariate auxiliary variables in sampling design is to employ a stratification scheme such that the population units are classified in a stratum according to the values of their auxiliary variables (Benedetti et al. 2008; Sigman and Monsour 1995; Vogel 1995). In this context the common procedure is to perform a stratification by size by defining a set of univariate threshold levels for each auxiliary variable included in the sampling frame. Such an approach is equivalent to partitioning the population into strata that have "box-shaped" boundaries or that are approximated through the union of several such boxes. This constraint prevents the identification of irregularly shaped strata boundaries unless a grid constituted by several rectangles of different size are used to approximate the required solution.

Optimal data partitioning is a classical problem in the statistical literature, following the early work of Fisher on linear discriminant analysis (Fisher 1936; for a review see Izenman 2008). However our problem is more directly related to the use of unsupervised classification methods (Everitt et al. 2011) to cluster a set of units (in this case a population frame). The main difference between the two problems lies in the fact that the underlying objective functions are different: in sampling design the aim is usually to minimize the sample size while in clustering it is a common practice to minimize the within cluster variance. There is an intuitive connection between these two concepts even if the definition of sample size depends not only on the variance within each stratum but also on other parameters (see Section 2).

The purpose of this paper is to propose a more general and accurate solution for the identification of the optimal multivariate boundaries of a self representing stratum since a general methodological framework to deal with this situation does not currently exist.

Our approach is based on the use of a random search algorithm whose aim is to find a partition of the frame that will minimize the sample size needed to respect a given limit on the coefficients of variation of the estimates for the totals of a given set of auxiliary variables.

Of course, the sample size can be further reduced by increasing the number of sampled strata. However in this paper we focus on the basic situation in which only two strata are considered. Our proposed multivariate stratification solution must therefore be considered as providing a point of departure for solution of the more general problem of strata boundaries for more than two strata. In this context two additional problems arise, an increasing computational burden (see Section 3) and, even more complex, the assessment of the optimal number of strata to be used.

It is also worthwhile pointing out that estimation accuracy can also be increased through the use of well known model assisted estimation methods such as the calibration weighting (Deville and Särndal 1992). In this context an optimal, or at least well designed, sample selection should be considered as complementary to an appropriate estimator and not an alternative.

The paper is organized as follows. Section 2 introduces a theoretical framework for the multivariate census threshold determination in which we focus on the partition of the population in two strata, with an overall sample size constrained to respect an upper bound on the coefficient of variation of the estimate of each variable of interest. Our proposal starts by simply extending the univariate method of Hidiroglou (1986) to a multivariate framework and then, by removing the assumption of box-shaped partition rules, we delineate a general combinatorial optimization problem that is solved via simulated annealing. To avoid excessive computational burden when dealing with list frames with a large number of statistical units (it is not uncommon that the number of units is greater than one million) we also investigated the generalized version of simulated annealing suggested by Tsallis and Stariolo (1996), and its deterministic counterpart, the so-called Iterated Conditional Modes (ICM; Besag 1986) algorithm. Section 3 examines how our approach can be used to draw a monthly sample of firms in the Milk Products sector for which a huge amount of auxiliary information is available. This practical example is also used to evaluate the performances of the proposed solutions through numerical comparisons. Finally, Section 4 is devoted to some concluding remarks, focusing on issues regarding further research on this topic.

2 Optimal Multivariate Threshold for a Completely Enumerated Stratum

The stratification method that is introduced in this section is tailored to population with strongly positive asymmetry: for example when the population is composed by few large units and many small units. A possible sample design for such a population is to split it into two sets according to a partitioning criterion. If this rule has to be applied to a single scalar auxiliary variable X we then have to determine a threshold such that units whose X values lie below the threshold are randomly sampled, while the units with values over the threshold are completely enumerated. This auxiliary variable is assumed to be correlated with the variable of interest Y. Suppose that for a finite population of size N values of an auxiliary variable $x_1, x_2, \ldots, x_i, \ldots, x_N$ are given such that, without any loss of generality, $x_1 \leq x_2 \leq \cdots \leq x_i \leq \cdots \leq x_N$. This population is partitioned into two sets of large and small units, labelled C and S respectively, with cardinalities N_C and N_S. A sample of n units is then obtained, consisting of all the N_C large units and $n - N_C$ small units, by drawing a simple random sample without replacement from the subset S of the small units.

The Horwitz-Thompson estimator of the total of the auxiliary is:

$$\hat{t}_{HT,x} = \frac{N_S}{n - N_C} \sum_{i \in s} x_i + \sum_{i=N-N_C+1}^{N} x_i, \qquad (1)$$

where s is the sample selected from the stratum S.

For the precision of the estimate (1) to achieve the required limit c_x for the coefficient of variation, the number of sampled units must then be equal to:

$$n = N - \frac{N_S c_x^2 t_x^2}{c_x^2 t_x^2 + N_S V_{S,x}^2}, \qquad (2)$$

where t_x is the known total of the auxiliary variable and $V_{S,x}^2$ is its variance in the stratum S.

In Hidiroglou (1986) it is shown that, when c_x, t_x and N are fixed, n has only one local minimum. Equation (2) shows that n cannot be fixed in advance, but it depends on the threshold thr of units to be completely enumerated, on the required coefficient of variation c_x and on the variance $V_{S,x}^2$. A population unit i is considered to be a sampled unit if it is in the set $S = \{i : x_i < thr\}$ or to be a completely enumerated (ce) unit if it is in the set $C = \{i : x_i \geq thr\}$.

If the population is strongly asymmetric, for a generic iteration h, Hidiroglou (1986) proposed to evaluate thr by using the following rule:

$$thr_{h+1} = \mu_{S_h} + \sqrt{\frac{N_{S_h} - 1}{N_{S_h}^2} c_x^2 t_x^2 + V_{S_h,x}^2}. \qquad (3)$$

In some experimental results Hidiroglou (1986) shows that, when a population is strongly asymmetric, this procedure converges (as the index h increases) to the optimum threshold.

2.1 Extensions to Multivariate Auxiliaries

The extension of the algorithm (3) to a set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_j}, ..., \mathbf{x_k}\}$ of k auxiliary variables is simply obtained through the union of each univariate partition or, in other terms, by defining the set S={ i: $x_{i,j}$ <thr_j∀ j=1,...,k}. This is an extremely conservative solution because the threshold obtained for a certain variable j does not use the information that there may be units that are under the threshold for this variable but over the threshold for at least one of the other auxiliary variables. We refer to this method as "Union" from now on.

An implication of using Union is that each individual threshold is set too low, i.e. too many units are placed in C. A solution, when using the information in (3) to define the threshold for any single auxiliary variable, is to define the set C not just as the $N_C = N - N_S$ greatest units according to the ordering induced by that variable, but as all the units that are considered to be in C according to the thresholds of the other $k - 1$ auxiliary variables. This is equivalent to replacing (3) by the iterative scheme:

$$thr_{r,h+1,j} = \mu_{x_j,S_{r,h,j}} + \sqrt{\frac{N_{S_{r,h,j}} - 1}{N_{S_{r,h,j}}^2} c_{x_j}^2 t_{x_j}^2 + V_{x_j,S_{r,h,j}}^2}, \qquad (4)$$

where $N_{S_{r,h,j}}$ is the number of units in S for auxiliary variable j at the iteration h of the algorithm used to evaluate its threshold and at iteration r of this "conditional optimization" algorithm. The key component at each iteration of this iterative algorithm is the definition of the set $S_{r,h,j}$={ i: $x_{i,g}$ <$thr_{r,h,g}$∀ g=1,...,j}∩{ i: $x_{i,g}$ <$thr_{r-1,h,g}$∀ g=j+1,...,k} since this includes all units defined as being in C in previous iterations for variable j as well as those already identified as being in C for the remaining variables.

The procedure is iterated over the index r, and within each value of this index, over the index h for each of the k variables, until convergence.

This iterative search is quite efficient and usually finds a solution in a small number of iterations using as starting thresholds the maximum value of each auxiliary variable, i.e. $C = \{\emptyset\}$ initially. Its main drawback is that it still corresponds to a box-shaped partitioning into sampled and completely enumerated units, and hence does not necessarily lead to a global optimum. In what follows we refer to this method as "Iterated Conditional Union".

2.2 A Stochastic Relaxation Approach: the Simulated Annealing Algorithm

An alternative optimizing approach which does not assume any shape for the optimal partition is via simulated annealing (SA). This is a stochastic optimization method for finding a global minimum of a function (Kirkpatrick et al.

1983). The method is a generalization of the Metropolis-Hastings algorithm (Metropolis et al. 1953) and represents one of the most popular optimization strategies for solving complex combinatorial problems.

Let $\theta \in \Theta = \{S, C\}^N$ be a vector of size N whose $i-th$ element may assume two possible values: S indicating that the unit is to be sampled ($i \in S$) and C if it has to be enumerated ($i \in C$). In the approach of Geman and Geman (1984), the optimization problem can be viewed as a stochastic process described through a family of distributions:

$$\pi_T(\theta) = \frac{\exp\{-f(\theta)/T\}}{\sum_{\theta \in \Theta} \exp\{-f(\theta)/T\}}, \tag{5}$$

where $f(\theta)$ is the energy function, T is a positive parameter called *temperature* and the denominator of equation (5) is called the *normalization constant* of the model. It can be shown (Brémaund 1999) that $\lim_{T\to 0} \pi_T(\theta) = \pi_{\lim}(\theta)$ where $\pi_{\lim}(\theta)$ takes the same positive value at any configuration θ that corresponds to a global minimum of the energy function and $\pi_{\lim}(\theta) = 0$ otherwise.

The SA algorithm is an iterative random search procedure where at each step we generate a non-homogeneous Markov-Chain with a reduced value for the temperature T. Specifically, for k auxiliary variables, the energy function for each iteration h and for each visited unit i can be defined as $f(\theta_{h,i}) = \max(n_{x_1,h,i}, n_{x_2,h,i}, \ldots, n_{x_k,h,i})$ where each sample size is evaluated using (2) under the configuration θ. Given a configuration $\theta_{h,i}$ at the $h-th$ iteration, another configuration, say $\theta_{h+1,i}$, is then chosen on the basis of a visiting schedule for each unit i. Here we visit units sequentially following an initial random ordering of the list frame. The new configuration of the $i-th$ element of the vector θ following such a visit is defined by first exchanging the codes S and C or vice versa of the unit i, and then accepting this exchange if it leads to a reduction of the energy function defined by the sample size.

For any suitable choice of stopping criterion, a final configuration is therefore obtained. Generally, this corresponds to a local minimum. In order to avoid convergence to local minima, a stochastic decision rule is used, which allocates a positive probability to exchange of configuration even when an increase in the energy function is obtained. In particular, the algorithm replaces the solution obtained at the $h-th$ iteration $\theta_{h,i}$ with a new solution $\theta_{h+1,i}$ according to an acceptance rule known as Metropolis criterion that allows hill climbing of the objective function.

More formally, the algorithm can be summarized as follows. The procedure starts at the first iteration $h = 1$, with an initial value T_1 and randomly selects the initial configuration θ_1 from $\{S, C\}^N$. At step h the elements of θ_h are updated as follows:

- select a unit in the population according to the visiting schedule (we used a sequential criterion) and exchange its status (from S to C or vice versa);
- denote with θ_h^* the new proposed vector of codes and with $f(\theta_h^*)$ the sample size evaluated as the maximum of (2) obtained for each of the k auxiliary variables. Randomly decide whether or not to adopt θ_h^* according to the Boltzmann distribution:

$$\theta_h = \begin{cases} \theta_h^* \text{ with probability} \\ p = \min\left(1, \exp\left\{\frac{[f(\theta_h) - f(\theta_h^*)]}{T_h}\right\}\right) \\ \theta_h \text{ otherwise;} \end{cases} \tag{6}$$

- repeat these steps, say m times (in the case studies of section 3 we used $m = 5$), for all the units of the population, then update the temperature according to a very simple and widely used rule $T_{h+1} = \rho T_h = \rho^h T_1$ with $0 < \rho < 1$ representing a control parameter on the cooling speed of the algorithm. Values close to 1 increase the number of iterations needed to reach the optimum but also prevent convergence to local minima;
- stop when $|f(\theta_h) - f(\theta_{h+1})| / f(\theta_h) \leq \varepsilon$ or when the number of iterations exceeds a fixed maximum.

The most serious drawback for this algorithm is its computational burden. Partitioning populations with a huge number of units could be unfeasible since the procedure is applied to all units in the population. While this might not be a major drawback for relative small populations in the case of two strata, it could be a serious drawback for very large populations and potentially a severe drawback in the case of more than two strata. If there are more than two strata then optimal multivariate sample allocations would need to be determined for each unit in the population at each iteration, leading to huge computational burden.

A way of speeding up the annealing process is to use generalized simulated annealing (GSA; Tsallis and Stariolo 1996), in which the updating criterion (6) is again random but with an acceptance probability given by:

$$\theta_h = \begin{cases} \theta_h^* \text{ with } p = \min\left(1, \frac{1}{\left[1+(q_A-1)\left(\frac{f(\theta_h^*)-f(\theta_h)}{T_h}\right)\right]^{\frac{1}{q_A-1}}}\right) \\ \theta_h \text{ otherwise,} \end{cases} \tag{7}$$

and with a temperature that decreases according to the rule:

$$T_h = T_1 \frac{2^{q_V - 1} - 1}{(1+h)^{q_V - 1} - 1}. \tag{8}$$

The two constants q_A and q_V regulate the acceptance probability distribution and the rate of temperature decrease respectively. When $q_A = q_V = 1$ the distribution (7) corresponds to (6) and the temperature decreases logarithmically with the increase of the iteration number, while when $q_A = 2$ and $q_V = 1$ the algorithm is equivalent to the so called "Fast Simulated Annealing" or "Cauchy Machine" (Tsallis and Stariolo 1996).

If the size N of the list frame is such that even the GSA is unfeasible, then it is possible to speed up the convergence of the annealing process even further using "Iterated Conditional Modes" (ICM; Besag 1986). This replaces the stochastic rule (6) by the deterministic rule:

$$\theta_h = \begin{cases} \theta_h^* & \text{if } f(\theta_h) > f(\theta_h^*) \\ \theta_h & \text{otherwise.} \end{cases} \tag{9}$$

Note that this rule will lead to convergence to a local minimum but not necessarily to a global one. However, this theoretical drawback could be more than compensated by a dramatic decrease in the computational burden. In our empirical experience the ICM has always suggested solutions that are rarely equal to the global optimum but usually very close to it and with a number of iterations which is negligible in comparison to that required by random search procedures as SA and GSA.

2.3 Auxiliary and survey variables

Up to now, the coefficient of variation constraints have been imposed on the auxiliary variables **X** rather than on the survey variables. The typical assumption is that optimal sample design based on specifying target levels of precision for a set of auxiliary variables will lead to a design that achieves the required target precision c_j for each survey variable j. However if, as often happens, the survey variables and the auxiliary variables are not just the same variables recorded in two different periods, then there could be considerable differences among them. In such situations the solution developed above could be sub-optimal because it is well known that in practice this hypothesis is only an approximation to the true situation and that using the auxiliary variables to design the sample could therefore underestimate the sample size needed to reach a predetermined level of precision.

A standard alternative is to use a model for any of the unknown q survey variables $\mathbf{Y} = \left\{\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y_l}, \dots, \mathbf{y_q}\right\}$ in terms of the known vector of auxiliaries **X**. The solution that underpins the approach that we adopt in the present paper is to derive from past surveys a model that relates each y_l with its counterpart x_j observed in previous years. In our particular application thus $q = k$ and for each model $j = l$. The sample allocation to each stratum is then made on the basis of the anticipated moments of **Y** given **X**. This approach is discussed in Dayal (1985), Sigman and Monsour (1995) and Baillargeon and Rivest (2009). It is important to emphasise that there is considerable advantage to designing a survey that is repeated at two time points in which the variables collected at each time point have the same definition and the phenomenon being investigated is known to be highly dependent on its past values. In this context, the case used as an illustration in this paper is a classical survey repeated in time that is very similar to most NIS-run business surveys (Benedetti et al. 2010; Hidiroglou and Laniel 2001; Hidiroglou and Srinath 1993) and consequently represents the natural application of the method proposed in this paper.

An important issue relates to the implicit use of a linear model linking the auxiliaries and the variable of interest **Y** in our approach. Clearly, we may use a simple linear regression if we are in the case where each variable has its own counterpart within the auxiliaries or multiple regression if they represent a set of completely different information only related to the set of covariates. In these simple models a log-scale relationship should help reduce the effects of heteroscedastic errors and skewness of the population data.

A more complex issue that often arises when dealing with business surveys, whose statistical units are usually establishments, is that the observed phenomenon can also be equal to 0 with a non-null probability. Such a zero inflated situation, where $\mathbf{X} > 0$ and $\mathbf{Y} = 0$, may occur because a unit can go out of business between the collection of the **X** variables and the date of the survey (Baillargeon and Rivest 2009). The probability to be zero, i.e. to go out of business, or suspend or postpone the activity of interest, typically decreases with the increase of the size of the establishments. The proposed model for y_j given x_j can then be based on a log-scale mixture model with survival probabilities p_h that are assumed to be constant for each unit i belonging to the same stratum $h = \{S, C\}$:

$$y_{i,j} = \begin{cases} \exp\left(\alpha_j + \beta_j \log\left(x_{i,j}\right) + \varepsilon_{i,j}\right) & \text{with probability } p_h \\ 0 & \text{with probability } 1{-}p_h; \end{cases} \tag{10}$$

where $\varepsilon_{i,j} \sim N\left(0, \sigma_j^2\right)$. Such models, whose parameters can be estimated by using maximum likelihood (Liu and Chan 2010), are widely used for ecological count data and recently extended to the analysis of economic microdata (Cameron and Trivedi 2005). The anticipated moments under (10) can be derived from Baillargeon and Rivest (2009):

$$\mu_{y_j, S_{r,h,j}} = p_h e^{\alpha_j + \frac{\sigma_j^2}{2}} \left(\frac{\sum_{i \in h} x_{i,j}^{\beta_j}}{N_h} \right), \tag{11}$$

$$V^2_{y_j, S_{r,h,j}} = p_h e^{2\alpha_j + 2\sigma_j^2} \left(\frac{\sum_{i \in h} x_{i,j}^{2\beta_j}}{N_h} \right) - p_h^2 e^{2\alpha_j + \sigma_j^2} \left(\frac{\sum_{i \in h} x_{i,j}^{\beta_j}}{N_h} \right)^2, \tag{12}$$

$$t_{y_j}^2 = e^{\alpha_j + \frac{\sigma_j^2}{2}} \sum_{h \in \{S,C\}} p_h \left(\sum_{i \in h} x_{i,j}^{\beta_j} \right). \tag{13}$$

3 A case study: the milk products monthly survey

In this section we will apply the multivariate census threshold determination procedures proposed in Section 2 to the Milk Products monthly survey performed in Italy by Istat. The purpose of this survey is to collect data on the quantities of milk and the main milk products produced every month. This survey is based on a stratified sample, with a stratification of dairies by size and output specialization for a total of seven strata. On average, the population and the sample size are approximately N=2000 and n=400 establishments respectively and the desired level of precision c is set to 5% with the exception for the variable Cows Milk Collected which is set to 1% (Istat 2011).

In addition to the monthly survey, Istat carries out an annual census of the dairies. In this study we use eight variables enumerated completely in the frames obtained from the censuses conducted from 2004 to 2009. These are Cow's Milk Collected (V1) and Cream Used (V2) corresponding to production input, and six variables for production output: Drinking Milk (V3), Cream Produced (V4), Acidified Milk (V5), Concentrated Milk (V6), Powdered Dairy Products (V7) and Cheese From Cow's Milk (V8).

The multivariate partition procedures proposed in this paper were then implemented, with the aim of determining the boundaries of the two strata C and S for the monthly survey on dairies corresponding to a hypothetical sample for the year 2010, using as auxiliary variables the census data of the previous year. Our aim was to estimate the population totals of the variables measured by Istat in the monthly survey.

We start with a brief description of the frames. The scatter plot of two of the most important and correlated variables, Cow's Milk Collected (V1) and Cheese From Cow's Milk (V8), are shown in Figure 1 for the 2009 data; the graphs for the other years (2004–2008) are almost identical and therefore are not reported here. The main evidence is that within a given year the variables are highly skewed with some mild positive correlation which is quite stationary in time, as confirmed by the linear correlation coefficients (see table 1 and table 2). The scatter plots of Figure 2 show that when the auxiliaries are compared between two years, say 2004 and 2005 or 2004 and 2009, they exhibit a very high correlation ranging in the interval [0.77; 0.99] that, as expected, decreases as the time span between the two variables is higher (see table 3). Moreover, dairies are strongly specialized and most firms are small. In particular in 2009, 79.92% of the firms produced only one type of output (78.7% produced only Cheese From Cow's Milk), 6.62% two types and only 2.02% three types.

The anticipated moments used in this experiment, i.e. sums and sums of squares, are estimated for each variable of interest for the year 2010 by fitting a simple linear regression to the available frame data for the years 2008 and 2009. Given that this very simple model fitted well, we chose not to try any of the more complex solutions described in section 2.3. In order to estimate the 2010 values of the anticipated moments of the survey variables we therefore assume that for each variable the relationship between 2010 and 2009 is equivalent to that observed between 2009 and 2008.

Before attempting to partition the population in the eight dimensional space of all the available auxiliaries it is interesting to observe the results obtained by some of the proposed algorithms in only two dimensions.

We chose two variables: Cow's Milk Collected (V1) and Cheese From Cow's Milk (V8). The box-shaped boundaries identified through the union of univariate results and by the iterated conditional union (ICU) approach (this required just three iterations) described in section 2 are set out in table 4.

Starting from a threshold where all population units are sampled, the stopping criterion adopted is when the maximum difference between the thresholds obtained in two consecutive iterations is zero.

An interesting practical property of the ICU algorithm, observed in every design in which it was applied, is that it always converges to the same solution even when for each auxiliary the threshold used is a first guess and is not iterated as in the univariate procedure (3), i.e. the maximum number of iterations for h is assumed to be equal to 1. This characteristic could be a useful way of speeding up convergence when dealing with very large populations.

Table 1: Correlation matrix of the 2009 data

	V2	V3	V4	V5	V6	V7	V8
V1	0.08	0.29	0.16	0.14	-0.01	-0.01	0.49
V2	1.00	0.22	0.29	0.03	0.02	0.00	0.14
V3		1.00	0.73	0.18	0.00	0.00	0.02
V4			1.00	0.16	0.00	0.00	0.02
V5				1.00	0.00	0.00	0.03
V6					1.00	0.00	-0.01
V7						1.00	-0.01

Table 2: Correlation Coefficients between Cow's Milk Collected (V1) and Cheese From Cow's Milk (V8) from 2004 to 2009

2004	2005	2006	2007	2008	2009
0.6474	0.5829	0.5538	0.4668	0.5085	0.4955

The main evidence from table 4 is that, as suggested by the theoretical considerations in section 2, the two limits proposed by the iterative procedure are higher than those obtained by the simple extension of the univariate criterion, resulting in 28 units being moved out of the completely enumerated stratum and a saving of approximately 8 sampling units when c=1% (see Table 5).

However it can also be shown that these boundaries are not optimal and that a better stratification can be achieved (see Figure 4) through the use of Simulated Annealing, further reducing the number of the units to be completely enumerated (units belonging to the dark grey area). Note that, due to the complex nature of the algorithm, the boundaries for each single variable follow a curve and are no longer linear (see Figure 4).

Unfortunately, this procedure cannot be considered as a definitive solution to our problem because it suffers from a critical disadvantage: its convergence to a global optimum could require an unfeasible number of iterations. Note that this problem is due to the temperature decreasing schedule that we choose. It is evident that the smaller the value of the parameter ρ, the quicker the procedure reaches convergence (see Figure 3 for the convergence of the algorithm by using only the two variables V1 and V8). However, speeding up the procedure could lead to convergence to a local minimum, implying that the optimal sample size could be lower than that obtained using a low value of ρ. This is not the case when using only two auxiliaries as the algorithm always converges to the global optimum (when ρ=0.99 it needs 221 iterations to converge while for ρ=0.96 only 83 are required and for ρ=0.93 and ρ=0.90 it used 52 and 36 iterations respectively).

Table 5 shows the sample allocations to the Sampled and Completely Enumerated strata when strata boundaries were calculated based on the two variables V1 and V8 as well as on the complete set of auxiliaries using all the methods suggested in Section 2, i.e. Union, Iterated Conditional Union, Simulated Annealing, Generalized SA (with parameters q_A = 2 and q_V = 1.5) and Iterated Conditional Modes. The op-

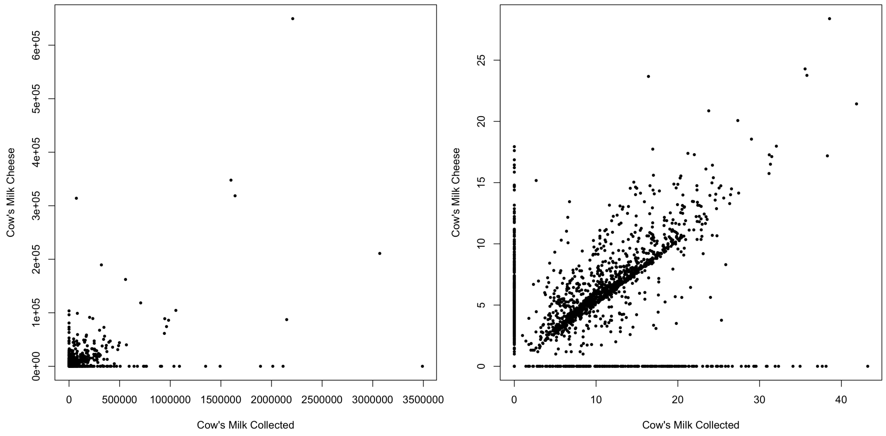
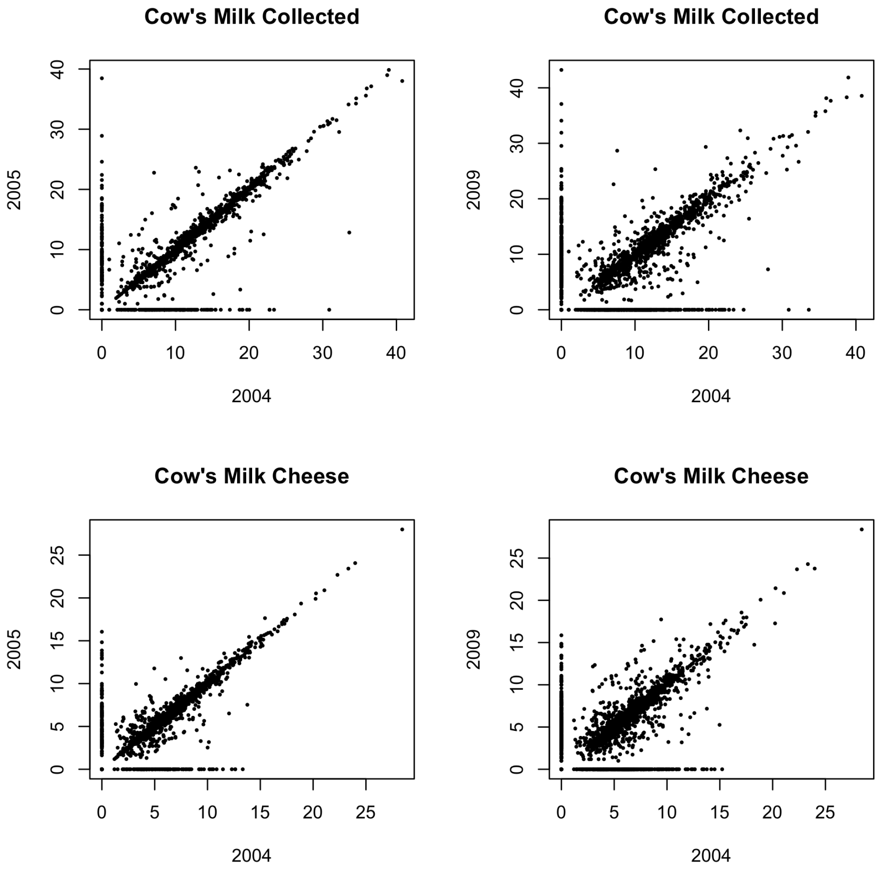


Figure 1. Scatter plots of the 2009 data; the units of measurement are quintal (left) and its $\sqrt[4]{}$ transformation (right)


Cow's Milk Collected

Cow's Milk Collected

Cow's Milk Cheese

Cow's Milk Cheese

Figure 2. Scatter plots of 2004 vs 2005 data and of 2004 vs 2009 data for the Cow's Milk Collected and for Cheese from Cow's Milk (data are a $\sqrt[4]{}$ transformation of quintals)

Table 3: Correlation Coefficients between the years 2004-2009 for Cow's Milk Collected (V1) and for Cheese From Cow's Milk (V8)

	Cow's Milk Collected					Cow's Milk Cheese				
Year	2005	2006	2007	2008	2009	2005	2006	2007	2008	2009
2004	0.91	0.85	0.80	0.79	0.77	0.98	0.98	0.97	0.96	0.95
2005	1.00	0.92	0.87	0.85	0.83	1.00	0.99	0.98	0.97	0.96
2006		1.00	0.97	0.95	0.94		1.00	0.99	0.98	0.96
2007			1.00	0.97	0.96			1.00	0.99	0.97
2008				1.00	0.99				1.00	0.98

Table 4: Complete enumeration thresholds based on application of the Iterated Conditional Union and Union algorithms, with *c=1%*. The thresholds of the Iterated Conditional Union are reported for each value of the parameters (r = among auxiliaries iteration number; h = univariate threshold iteration number for each auxiliary)

Algorithm	Max. Abs. Difference	Cow's Milk Collected	Cow's Milk Cheese	r	h
Iterated Conditional Union		37156.26	3994.10	1	6,6
	1628.44	38784.70	3997.40	2	5,1
	0.00	38784.70	3997.40	3	1,1
Iterated Conditional Union		75435.86	13660.61	1	1,1
	31321.38	44114.48	4843.24	2	1,1
	4542.59	39571.89	4074.67	3	1,1
	766.22	38805.67	4006.89	4	1,1
	33.10	38772.57	3999.74	5	1,1
	12.13	38784.70	3997.40	6	1,1
	0.00	38784.70	3997.40	7	1,1
Union		37156.26	3841.26		

timal sample size suggested by each method was evaluated for three different target sampling errors *c*=1%, *c*=5% and *c*=10% which were set to be the same for all the variables.

The main conclusion that can be drawn from the results set out in table 5 is that moving from Union to Iterated Conditional Union leads to a decrease in the sample size (approximately 1% for 2 variables and 1,6% for 8 variables for *c*=1% and *c*=5% while almost no reductions are present when *c*=10%). However, this reduction is not nearly as dramatic when the SA and the GSA are used, both for two and eight auxiliaries, and for the different errors. In this case reductions range from 3,1% when *c*=1% to 5,3% when *c*=10% with two auxiliaries and from 5% for *c*=1% to 15,8% for *c*=10% with eight auxiliaries. Thus, as the number of the auxiliaries increases, SA and GSA seem to increase their relative efficiency.

Note that although they require a different number of iterations SA and GSA always converged to the same solution in our experiments, while ICM, which requires a considerably lower number of iterations, typically reached a local optimum that was about 1% to 2% larger than that reached by SA and GSA. For many practical sample designs this could be a satisfactory price to pay to ensure a drastically reduced computational burden compared to that required by a random search method.

Furthermore, it is clear that there is a decrease in the number of completely enumerated units when one goes from a threshold based on a box-shaped boundary procedure such as Union or Iterated Conditional Union to one based on a random search criterion such as SA and GSA. This illustrates that as the target error level *c* increases, the boundaries returned by SA and GSA can reduce the number of the units to be completely enumerated compared with boundaries defined by box-shaped algorithms, and this reduction increases as we increase the number of the auxiliary variables.

Table 6 shows the CVs for the auxiliary variables generated by the different partitions defined by the various methods. Observe that for every variable and every method, the limits *c* are always respected but the irregularly shaped partitions (SA, GSA and ICM) always avoid unnecessarily low CV values. That is, in order to achieve the CV limits, a partitioning rule that is based on box-shaped boundaries tends to require a much greater number of population units to be completely enumerated compared with a rule that allows an irregularly shaped partition.

4 Conclusions

In the design of agricultural surveys it is very useful to have a completely enumerated stratum of large units defined in terms of a multivariate size measure. However, in spite of the fact that such multivariate auxiliary variables are now available, the current literature only addresses the definition of a completely enumerated stratum for the univariate case. The numerical solutions described in this paper represent an attempt to fill this gap. In particular we show that algorithms that use a random search strategy to find a better partition of the frame are a good alternative to those that are obtained as simple extensions of well known univariate solutions. In

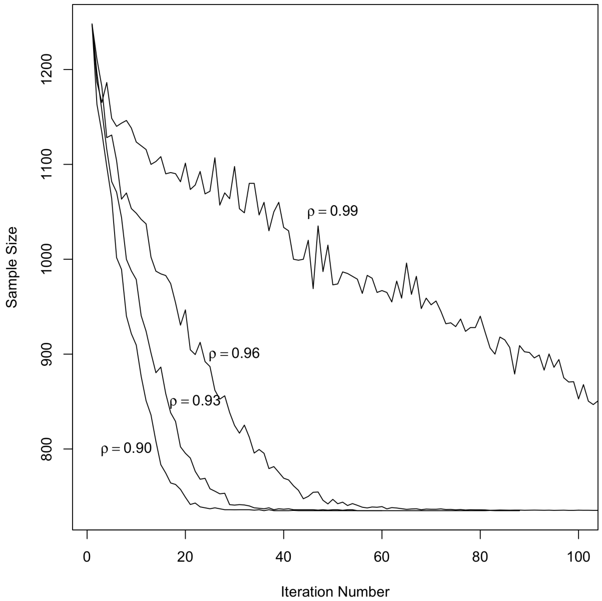


Figure 3. Two auxiliary variables: convergence to a global optimum of the sample size obtained by the simulated annealing partitions, varying the iteration number, for different values of the temperature decrease parameter ρ and *c=1%*

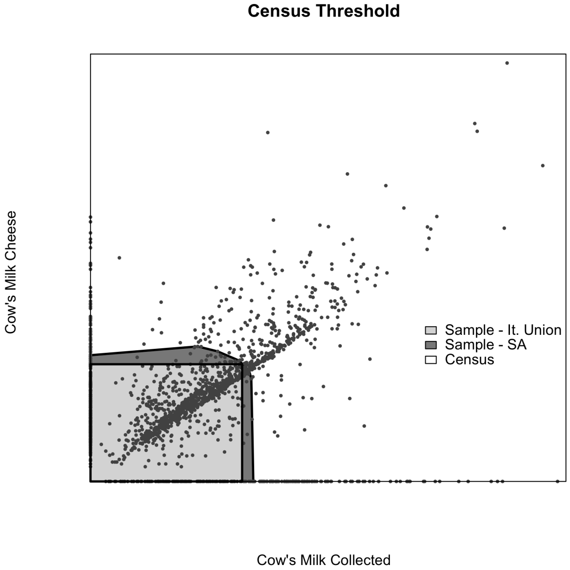


Figure 4. Two auxiliary variables: boundaries between the completely enumerated stratum (white) and the sampled stratum defined by Iterated Conditional Union (light gray) and Simulated Annealing (dark gray)

particular, we show that a solution to this type of combinatorial optimization problem, subject to an overall sample size sufficient to achieve a fixed level of error for each variable, can be implemented via simulated annealing.

The use of anticipated moments in our approach allows its extension to optimal design for multipurpose surveys. In particular, the use of statistical models to predict the future values of a set of survey variables, is a tool that could help to extend the results of this paper to situations more complex than those that we encountered in our design of the Milk Products Monthly Survey. However, how to calculate the predicted values and how to use these results in analytical calculations of the expected coefficients of variation are two issues that require further research.

It is important to emphasise that it is not appropriate to consider that the partition methods explored in this paper represent as an alternative to a selection with probability proportional to size (πps) design. Even though a πps design strategy is typically used when only one size measure in the list frame is available, it is our opinion that the best overall

Table 5: Sample sizes *n*, and numbers of units *NS* in the sampled stratum and *NC* in the completely enumerated stratum, for different methods, numbers of variables considered and target levels of the sampling error *c*

c=1%

	2 Auxiliary Variables			8 Auxiliary Variables		
Method	*NS*	*NC*	*n*	*NS*	*NC*	*n*
Union	1297	642	766.1	1255	723	833.6
It. Cond. Union	1325	614	758.7	1294	684	820.3
SA, GSA	1409	530	735.2	1399	579	778.7
ICM	1341	598	744.7	1342	636	786.5

c=5%

	2 Auxiliary Variables			8 Auxiliary Variables		
Method	*NS*	*NC*	*n*	*NS*	*NC*	*n*
Union	1718	221	309.1	1678	300	373.3
It. Cond. Union	1723	216	305.4	1687	291	367.3
SA, GSA	1769	170	291.5	1762	216	330.6
ICM	1736	203	297.9	1736	242	333.3

c=10%

	2 Auxiliary Variables			8 Auxiliary Variables		
Method	*NS*	*NC*	*n*	*NS*	*NC*	*n*
Union	1815	124	177.4	1782	196	242.7
It. Cond. Union	1817	122	176.5	1783	195	242.7
SA, GSA	1851	88	167.1	1849	129	204.3
ICM	1830	109	170.5	1829	149	208.7

Table 6: Ex post CVs for the eight auxiliary variables, for the different methods and sampling errors

Limit CV	Method	V1	V2	V3	V4	V5	V6	V7	V8
c=1%	Union	1.00	0.16	0.08	0.17	0.12	0.00	0.00	0.93
	It. C. Union	1.00	0.17	0.07	0.16	0.11	0.00	0.00	0.93
	SA, GSA	1.00	0.76	0.99	0.69	1.00	0.98	0.70	1.00
	ICM	1.00	0.88	0.94	0.85	0.35	0.00	0.00	1.00
c=5%	Union	5.00	1.15	1.20	1.24	0.75	0.00	0.00	4.86
	It. C. Union	4.98	1.25	1.18	1.22	0.74	0.00	0.00	5.00
	SA, GSA	5.00	3.53	4.99	4.46	1.65	1.55	1.11	5.00
	ICM	5.00	3.90	4.93	4.95	1.88	1.77	1.26	5.00
c=10%	Union	9.85	3.20	2.94	3.36	0.99	0.00	0.00	10.00
	It. C. Union	9.75	3.17	2.91	3.33	0.98	0.00	0.00	10.00
	SA, GSA	10.00	9.19	9.92	9.88	7.50	2.01	1.44	10.00
	ICM	9.97	9.22	9.91	9.40	8.51	2.27	1.62	10.00

strategy should be the one that makes joint use of these two important and efficient sample design tools.

Other issues that remain open for future research include extension of the algorithms developed in this paper to solve the more general problem of multivariate and multipurpose optimal partition of the population into more than two strata. Furthermore, it would be extremely useful if this multipurpose and multivariate approach could also deal with the practical problem of nonresponse, perhaps through the use of response probabilities estimated from previous surveys.

References

Baillargeon, S., & Rivest, L. P. (2009). A General Algorithm for Univariate Stratification. *International Statistical Review*, *77*(3), 331-344.

Bee, M., Benedetti, R., Espa, G., & Piersimoni, F. (2010). On the Use of Auxiliary Variables in Agricultural Surveys Design. In R. Benedetti, M. Bee, G. Espa, & F. Piersimoni (Eds.), *Agricultural survey methods.* Chicester: Wiley.

Benedetti, R., Bee, M., & Espa, G. (2010). A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics*, *26*(4), 651-671.

Benedetti, R., Espa, G., & Lafratta, G. (2008). A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology*, *34*, 195-203.

Benedetti, R., & Piersimoni, F. (Submitted for publication). *πps Selection Based On Multivariate Auxiliary Variables.* Submitted for publication.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, *48*(3), 259-302.

Brémaund, P. (1999). *Markov Chain: Gibbs Fields, Monte Carlo Simulation and Queues*. New York: Spinger-Verlag.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

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

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristics. *Journal of Statistical Planning and Inference*, *11*, 321-328.

Deville, J. C., & Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, *87*(418), 376-382.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis. Wiley series in probability and statistics* (5th ed.). Chichester, West Sussex: Wiley.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179-188.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Vol. PAMI-7*(6), 721-741.

Gunning, P., & Horgan, J. M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, *30*(2), 159-166.

Gunning, P., & Horgan, J. M. (2007). Improving the Lavalleé and Hidiroglou algorithm for stratification of skewed populations. *Journal of Statistical Computation and Simulation*, *77*, 277-291.

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

Hidiroglou, M. A. (1986). The construction of a self representing stratum of large units in survey design. *The American Statistician*, *40*, 27-31.

Hidiroglou, M. A., & Laniel, N. (2001). Sampling and estimation issues for annual and subannual Canadian business surveys. *International Statistical Review*, *69*, 487-504.

Hidiroglou, M. A., & Srinath, K. P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economics Statistics*, *11*, 397-405.

Horgan, J. M. (2006). Stratification of Skewed Populations: A Review. *International Statistical Review*, *74*(1), 67-76.

Istat. (2011). *Dati mensili sul latte di vacca raccolto e sui prodotti lattiero-caseari.* http://agri.istat.it.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. New York: Springer.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671-680.

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

Liu, H., & Chan, K. S. (2010). Introducing COZIGAM: An R Package for Unconstrained and Constrained Zero-Inflated Generalized Additive Model Analysis. *Journal of Statistical Software*, *35*(11), 1-26.

Lu, W., & Sitter, R. R. (2002). Multi-way Stratification by Linear Programming Made Practical. *Survey Methodology*, *28*(2), 199-207.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Rivest, L. P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, *28*, 191-198.

Sigman, R. S., & Monsour, N. J. (1995). Selecting Samples From List Frames of Businesses. In B. G. Cox, D. A. Binder, B. Nanjamma Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business Survey Methods* (p. 133-152). New York: Wiley.

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

Tsallis, C., & Stariolo, D. A. (1996). Generalized Simulated Annealing. *Physica A*, *233*, 395-406.

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.