Journal of Official Statistics, Vol. 26, No. 4, 2010, pp. 651–671

A Framework for Cut-off Sampling in Business Survey Design

Roberto Benedetti[1], *Marco Bee*[2], *and Giuseppe Espa*[3]

In sampling theory skewed distributions of many of the survey variables in a population make use of classical tools difficult. One possible solution is cut-off sampling, which discards a part of the population from the sampling frame. Although cut-off sampling is common among practitioners, its theoretical foundations are weak because the inclusion probabilities of some of the units are zero. In this article we propose a framework that justifies cut-off sampling and provides a means for determining census and cut-off thresholds. We use an estimating model that assumes that the sizes of the discarded units for each variable are known. We compute the variance of the resulting estimator and its bias. We develop a mean-squared-error-minimizing algorithm as a function of multivariate auxiliary information at the population level. Due to the multivariate nature of the model, we employ the theory of stochastic relaxation and use the simulated annealing algorithm.

Key words: Cut-off sampling; skewed populations; model-based estimation; optimal stratification; simulated annealing.

1. Introduction

Cut-off sampling is a procedure commonly used by national statistical institutes to select samples, but it is not easy to give a unique, clear-cut definition of the methodology. Roughly speaking, the population is partitioned in two or three strata such that the units in each stratum are treated differently. In particular, part of the target population is usually excluded a priori from sample selection.

A short introduction to cut-off sampling is found in Knaub (2008b). The basic formulation (Hansen et al. 1953, pp. 486–490; Särndal et al. 1992, pp. 531–533), frequently employed in the field of price collection, is characterized by a threshold such that the units above this threshold are included in the sample with a positive probability. The units below this threshold are discarded, their probability of being included in the sample being zero. In this case, as noted by de Haan et al. (1999), the sampling variance is zero by definition. This does not imply, however, the solution of all of the accuracy problems. It is well known (see, for example, Särndal et al. 1992, p. 531) that cut-off

[1] Department of Business, Statistical, Technological and Environmental Sciences, University "G. d'Annunzio" of Chieti-Pescara, Pescara, Italy. Email: benedett@unich.it
[2] Department of Economics, University of Trento, Trento, Italy. Email: marco.bee@unitn.it
[3] Department of Economics, University of Trento, Trento, Italy. Email: guiseppe.espa@economia.unitn.it
Acknowledgments: The authors would like to thank the Associate Editor, an anonymous referee and F. Piersimoni (ISTAT, Servizio Agricoltura) for helpful comments and suggestions that contributed to considerably improving a first draft of this article.

sampling produces biased estimators. Therefore, the error measure typically used is the mean squared error (that is, the sum of variance and squared bias). It follows that cut-off sampling might be a good choice where the variance reduction more than offsets the introduction of a small bias (Knaub 2007).

An alternative interpretation is proposed by Hidiroglou (1986), who considers two strata. In the first one all the observations are included in the sample, whereas in the second one the units are not discarded but sampled. In this context, the algorithm proposed by Lavallée and Hidiroglou (1988) is often used to determine the stratum boundaries and the stratum sample sizes. For the setup where the survey variable and the stratification variable differ, Rivest (2002) proposed a generalization of the Lavallée and Hidiroglou algorithm. The Rivest algorithm includes a model that takes into account the differences between the survey and the stratification variable and allows one to find the optimal sample size and the optimal stratum boundaries for a take-all/take-some design.

Finally, the most general approach (the one adopted in this article) considers three strata whose units are respectively enumerated completely, sampled and discarded. As pointed out by Sigman and Monsour (1995), this type of stratification is particularly appropriate in business surveys, because businesses tend to have skewed distributions. Thus, size has a considerable impact on the precision of survey estimates, and failure to notice that such populations should be stratified in the aforementioned manner may cause an underestimation of the population characteristics. When the distribution of the selection variable is concentrated in a few large establishments, this methodology provides the investigator with a sample whose size is rather small but whose degree of coverage is high.

The problem treated in this article is a generalization of the standard cut-off sampling. As is usual in business surveys, we assume the population of interest to be positively skewed, because of the presence of few "large" units and many "small" units. If one is interested in estimating the total of the population, a considerable percentage of the observations makes a negligible contribution to the total; on the other hand, the inclusion in the sample of the largest observations is essentially mandatory.

In such situations, practitioners often use partitions of the population in three sets: a take-all stratum whose units are surveyed entirely (U_C), a take-some stratum from which a simple random sample is drawn (U_S) and a take-nothing stratum whose units are discarded (U_E). In other words, survey practitioners decide a priori to exclude part of the population from the analysis (for example, firms with less than five employees). However, this choice is often motivated by the desire to match administrative rules: in this case, the partition of firms into small, medium and large. This strategy is employed so commonly in business surveys that its use is "implicit" and "uncritical", so that the inferential consequences of the restrictions caused to the archive by this procedure are mostly ignored.

The problem of determining the optimal take-all threshold, i.e., the partition of the population into strata U_C and U_S, is relatively straightforward both from the technical and from the methodological point of view (Hidiroglou 1986). On the other hand, finding a criterion that assigns each unit to exactly one of the three strata tends to be considered as a nonviable alternative, mainly because some inclusion probabilities are set equal to zero. It follows that cut-off sampling is, in some sense, in an intermediate position between probabilistic and nonprobabilistic sampling schemes, a feature that is not appreciated by

experts in this field. As a result, in the literature there are very few papers concerning its methodological foundations.

Nonetheless, in applications it is frequently used. It is the case, for example, when it comes to the monthly survey of manufacturing performed by Statistics Canada (see, for example, Statistics Canada 2001), that implicitly uses cut-off sampling, without paying too much attention to methodological implications: "The sampling frame for the Canadian Monthly Survey of Manufacturing (MSM) is determined from the target population after subtracting establishments that represent the bottom 2% of the total manufacturing shipments estimate for each province. These establishments were excluded from the frame so that the sample size could be reduced without significantly affecting quality." Similar procedures are employed in surveys performed by other national statistical institutes (for a thorough review see Knaub 2007, Section II): cut-off sampling is widely used but methodological aspects are not documented. Two exceptions are the book by Särndal et al. (1992, pp. 531–533), and the paper by de Haan et al. (1999): the latter presents successful applications of cut-off sampling in the field of consumer price indexes. As pointed out by Knaub (2007, p. 2), cut-off sampling for estimation of unit prices may be useful: "If a cut-off sample is used for revenues and another is used for sales volume, then the ratio will tend to be more accurate than either the numerator or the denominator".

Finally, Elisson and Elvers (2001) performed a univariate analysis that compares cut-off sampling with simple stratified sampling. They conclude that cut-off sampling deserves more consideration and suggest its use in applications; however, they find that the dimensional variable that determines the cut-off threshold has a relevant impact on the results, so they stress that great care must be employed in choosing this variable. Moreover, they point out the need for an appropriate model for the estimation of the fraction of population excluded from the sample.

In any case, it is worth mentioning the practical advantages of cut-off sampling as concerns the costs of a survey:

i) building and updating a sampling frame for small business units could be too costly, considering that the gain in efficiency of the estimators would probably be small;
ii) excluding the units of the population that make little contribution to the aggregates to be estimated usually implies a large decrease in the number of units that have to be surveyed in order to get a predefined accuracy level of the estimates;
iii) putting a constraint on the frame population and, as a consequence, on the sample, makes it possible to reduce the problem of empty strata that mainly affects the smallest firms. Regarding this issue, several empirical analyses shown that some difficulties, such as the nonresponse rate, the turnover rate of economic units and the errors of under- or over-coverage of the frame, become more relevant as the size of the units gets smaller;
iv) cut-off sampling may be demonstrably more practical in terms of accuracy when total survey error is taken into account. Knaub (2004) shows a way to gauge total survey error in the context of nonsampling errors, such as measurement error.

Given that practitioners are in favor of such partitions of the population and there are technical reasons that justify their use, the basic question is: is it possible to consider

cut-off sampling as a valid sampling scheme? If the answer is positive, the issue is to define a statistical framework for cut-off sampling.

In this connection we try to develop an easily implementable solution to the problem of the construction of the three strata U_C, U_S, and U_E in a multipurpose and multivariate setup. In other words, similarly to what happens in practical applications, we assume an interest in surveys with more than one target variable, using auxiliary information contained in multiple variables.

The case when a single measure of size is available is, however, not that uncommon. For example, in business surveys, the only auxiliary information is often the number of employees. Furthermore, when the surveys are voluntary, the rate of participation of small firms is mostly very low. In this instance a cut-off sampling procedure based on a dimensional variable (Bailar et al. 1983, Section 5.1) is undoubtedly convenient. The situation should simplify substantially in that the take–all categories would be the units with the biggest size measures and the take–none would be the units with the smallest. It would remain to determine the boundary points. For this issue there is probably an analytical, instead of algorithmic, resolution to be found. However, the present problem is multivariate, so that we leave to future research the study of the univariate problem.

The structure of the article is as follows. In Section 2 we will define an estimation model that assumes, for each variable, the weight of the units excluded from the analysis to be known and constant. However, this hypothesis is not, in general, under the control of the investigator, so that this estimator is biased, and we will have to find its bias and mean squared error (MSE). The model will be developed for the estimation of a total. Section 3 will be devoted to the derivation of the sample size for the cut-off scheme, focusing on its optimization and, consequently, on the construction of the optimal design. The problem will be tackled by defining the sample size as a function of the partition U_C, U_S, and U_E determined on the basis of multivariate auxiliary information that will be assumed to be known for the whole population. In view of the combinatorial nature of this problem, we will use the theory of stochastic relaxation and, in particular, the simulated annealing (SA) algorithm. In Section 4 we will show some empirical evidence about the bias of the estimator when using data from surveys concerning slaughtering firms in Italy. In the same section we will present the main results of the application of the sampling scheme to this dataset. Finally, Section 5 shall conclude the article and point out some open problems.

2. An Estimator of the Total for Cut-off Sampling Schemes

The problem of stratifying in two strata (take-all and take-some) and finding the census threshold was first treated by Dalenius (1952) and Glasser (1962). The former author determined the census threshold as a function of the mean, the sampling weights and the variance of the population. Glasser (1962) derived the value of the threshold under the hypothesis of sampling without replacement a sample of size n from a population of N units. Hidiroglou (1986) reconsidered this problem and provided both exact and approximate solutions under a more realistic hypothesis. He found the census threshold when a level of precision concerning the mean squared error of the total was set a priori, without assuming a predefined sample size n. It is worth noticing that he considered a case with only a take-all and a take-some stratum, so that he developed a method for finding

a "census threshold" (defined "cut-off threshold" in the paper). In some important applications, mostly in conjunctural business surveys, it may be convenient to use the so−called "census threshold" (namely strata U_C and U_E only), especially when the data element of interest is a ratio of other data elements (such as cost per unit volume). For such surveys the sample data are collected monthly or quarterly and it may be difficult (or impossible) to obtain accurate data from the smallest members of the population (Royall 1970; Knaub 2007; Knaub 2008a). However, all these authors limit their attention to a single purpose and univariate setup.

This work stems from Hidiroglou's (1986) idea but extends it substantially. We stratify the target population by means of a criterion that defines the belonging of each observation to one of the three strata U_C, U_S, and U_E in a multipurpose and multivariate framework. The solution is based on the identification of appropriate estimators for the quantities in Table 1.

In this article we consider the estimator of the total $\hat{t}_{y_j}$ of the jth surveyed variable ($j = 1, \ldots, J$). This estimator is the sum of three independent components, corresponding respectively to the take-all, take-some and take-nothing strata. Thus, omitting for simplicity the index of the variables (the same way of reasoning can be applied to all the J variables), we can write $\hat{t}_y = \hat{t}_C + \hat{t}_S + \hat{t}_E$. As for the take-all stratum, it is clear that $\hat{t}_C = t_C = \sum_{k \in U_C} y_k$. In the take-some stratum, we use the classical π−estimator of the total $t_S = \sum_{k \in U_S} y_k$:

$$\hat{t}_S = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k \tag{1}$$

that is the expansion formula known in the literature as the Horvitz-Thompson estimator, from now on the HT estimator (Horvitz and Thompson 1952). In (1), the π_ks are the inclusion probabilities, that are assumed to be strictly positive; the same assumption holds for the second-order probabilities π_{kl}, that are necessary for the computation of the variance of the estimator. The quantities $d_k = 1/\pi_k$ are the design weights of each unit $k \in s$, namely the original weights resulting from the sampling scheme.

The sample s is a probabilistic sample drawn from the subpopulation U_S; in the following we will always assume that it is a simple random sample from U_S. It is worth stressing that in this article we deal with a sample design issue. When doing this, simplifying the problem from the formal point of view is a classical strategy. However, there would be no reason against the use, for drawing a sample from U_S, of any selection criterion different from simple random sampling. Considering the special features of the problem at hand, a natural candidate would be the *pps* (probability proportional to size) criterion. An excellent reference on *pps* is Särndal et al. (1992, pp. 97–100). On the other hand, in general, the *pps* method only works in the univariate setup. Thus, in order to

Table 1. Estimators and error measures. $b(\cdot)$ is the bias function, f and g are functions that shall be defined in the following

Stratum	U_C	U_S	U_E
Estimator	$\hat{t}_C$	$\hat{t}_S$	$f(\hat{t}_C, \hat{t}_S)$
MSE	0	$\text{var}(\hat{t}_S)$	$g(\text{var}(\hat{t}_S) + b^2(\hat{t}_E))$

maintain the generality of the multivariate auxiliary information approach used here, we delay the study of the *pps* case to future research.

According to the setup of our problem, the Hidiroglou-type estimator $t_C + \hat{t}_S = \sum_{k \in U_C} y_k + \sum_{k \in s} d_k y_k$ has to be augmented by a model-based component that takes into account the discarded fraction of the population, U_E. As concerns this issue, we can write

$$t_E = (t_C + t_S)\delta \tag{2}$$

i.e., the total of the discarded population is a fraction of $t_C + t_S$. In (2) the quantity δ, that is usually unknown, can be evaluated by means of external sources (i.e., auxiliary variables $\boldsymbol{x}$); thus

$$\tilde{\delta} = \frac{\sum_{k \in U_E} x_k}{\sum_{k \in U_C} x_k + \sum_{k \in U_S} x_k} \tag{3}$$

For notational simplicity and without loss of generality, in the following we will always assume that each auxiliary variable is the target variable as known from the last census: $x_k = y_{k,t-1}$. Using these hypotheses we obtain the following identity:

$$\hat{t}_y = t_C + \hat{t}_S + \hat{t}_E = (1 + \tilde{\delta})(t_C + \hat{t}_S) = (1 + \tilde{\delta}) \left(\sum_{k \in U_C} y_k + \sum_{k \in s} d_k y_k \right) \tag{4}$$

The assumptions introduced to obtain (4) are slightly different from those introduced Särndal et al. by (1992, p. 532), who use a ratio estimator in the domain S as a "compensation" for the fraction of population discarded. As we are concerned with a sampling design, in this article we find it more convenient to employ, as a starting point for the part of the population to be sampled, the "neutral" HT estimator. It is always possible to implement, in the estimation procedure, a second step. One could indeed use the auxiliary information *ex post*, in order to correct t_C and $\hat{t}_S$ either by means of a ratio estimator or by means of a more general approach to the use of auxiliary information such as the so-called calibration estimators (Deville and Särndal 1992). In any case, assuming the ability to correct for the bias in the design phase seems to be a very strong hypothesis, because in this step the researcher usually employs simple estimators, such as HT. When the survey is completed and all the data are available, it is possible to resort to more complicated estimators.

It is well-known (see, for example, Särndal et al. 1992, p. 531) that cut-off sampling produces biased estimators. Using (4) and the independence of the three strata U_C, U_S and U_E, the mean squared error of $\hat{t}_y$ is given by:

$$\begin{aligned} \mathrm{MSE}(\hat{t}_y) &= \mathrm{var}(\hat{t}_y) + b^2(\hat{t}_y) = \mathrm{var}(t_C + \hat{t}_S + \hat{t}_E) + b^2(\hat{t}_y) \\ &= \mathrm{var}[(1 + \tilde{\delta})(t_C + \hat{t}_S)] + b^2(\hat{t}_y) = (1 + \tilde{\delta})^2 \mathrm{var}(t_C + \hat{t}_S) + b^2(\hat{t}_y) \\ &= (1 + \tilde{\delta})^2 \mathrm{var}(\hat{t}_S) + b^2(\hat{t}_y) = (1 + \tilde{\delta})^2 \mathrm{var}(\hat{t}_S) + b^2(\hat{t}_E) \end{aligned} \tag{5}$$

Note that (5) is the conditional MSE given that $\tilde{\delta} = \delta$. Moreover, we put $b(\hat{t}_y) = b(\hat{t}_E)$ to stress that the bias, that represents the price to pay for discarding part of the population,

only depends on the excluded stratum. It is indeed clear that $\tilde{\delta} \in \mathbb{R}^+$ in (4) introduces a bias because the true ratio δ is unknown and different from the value $\tilde{\delta}$ computed in the current survey. Note that, when only the "census threshold" (that is, only the strata U_C and U_E) has to be determined, (5) reduces to the bias–related component.

It is therefore crucial to concentrate on the bias $b(\hat{t}_E)$. It is not difficult to see that:

$$\begin{aligned} b(\hat{t}_E) &= E(\hat{t}_y) - t_y = E(t_C + \hat{t}_S + \hat{t}_E) - t_y \\ &= \sum_{k \in U_C} y_k + \sum_{k \in U_S} y_k + E[\tilde{\delta}(t_C + \hat{t}_S)] - t_y \\ &= \tilde{\delta}(t_C + t_S) - t_E \end{aligned} \tag{6}$$

Putting $t_E = \delta(t_C + t_S)$, (6) can be conveniently rewritten as follows:

$$b(\hat{t}_y) = (\tilde{\delta} - \delta)(t_C + t_S) \tag{7}$$

From (7) it appears that the source of the bias of Estimator (4) is the mismatch between the numerical value $\tilde{\delta}$ used in the survey and the true value δ. In particular, the magnitude of the bias is determined by the difference $|\tilde{\delta} - \delta|$.

As will become clearer in the next section, (7) is a fundamental ingredient of the sample design proposed here. In Section 4 we will show some empirical evidence concerning the functional form of the bias.

In the next sections we will express the sample sizes as functions of the MSE because the two components of the MSE vary as the census and cut-off thresholds change. As a consequence, neither of the components can be held fixed.

3. The Optimal Design

3.1. Sample Size

In the preceding section we showed that the MSE of the estimator $\hat{t}_y$ of the total for cut-off designs is equal to $\text{MSE}(\hat{t}_y) = (1 + \tilde{\delta})^2 \text{var}(\hat{t}_S) + b^2(\hat{t}_y)$, where $\text{var}(\hat{t}_S)$ is the variance of the HT estimator used for estimating the total of the target variables in the subpopulation U_S. The well-known expression for this variance in simple random sampling without replacement (Särndal et al. 1992, p. 46) is given by

$$\text{var}(\hat{t}) = N^2 \frac{1 - f}{n} S^2 \tag{8}$$

where S^2 is the variance of the target variable. However, in our setup this formula needs to be modified. The HT estimator is indeed only used in U_S, so that (Cochran 1977, Theorem 5.3)

$$\begin{aligned} \text{var}(\hat{t}_S) &= \frac{(N - N_C - N_E)[N - N_C - N_E - (n - n_C)]}{n - n_C} S^2 = N_S \frac{N_S - n_S}{n_S} S^2 \\ &= N_S^2 \frac{1 - \dfrac{n_S}{N_S}}{n_S} S^2 \end{aligned} \tag{9}$$

where $N_S = N - N_C - N_E$, $n_S = n - n_C$ and $f = n/N$ is the sampling fraction. In (9) the variance S^2 is equal to

$$S^2_{U_S} = \frac{1}{N_S - 1}\sum_{k \in U_S}(y_k - \mu)^2$$

where $\mu = \mu_{U_S} = (1/N_S)\sum_{k \in U_S} y_k$.

In applications, the MSE is usually required to satisfy the following equality:

$$\text{MSE}(\hat{t}_y) = c^2 t_y^2 \tag{10}$$

where c is the desired level of precision for the estimation of the total. In addition to (10), another common cost equation is $\text{MSE}(\hat{t}_y) = k^2$. This expression is equivalent to (10) with $k = ct_y$.

If we substitute for $\text{MSE}(\hat{t}_y)$ in (10) the next to last expression in (5) we get:

$$(1 + \tilde{\delta})^2 \text{var}(\hat{t}_S) + b^2(\hat{t}_y) = c^2 t_y^2$$

from which we easily derive the variance of the estimator:

$$\text{var}(\hat{t}_S) = N_S \frac{N_S - n_S}{n_S} S^2 = \frac{c^2 t_y^2 - b^2(\hat{t}_y)}{(1 + \tilde{\delta})^2} \tag{11}$$

We now focus on Expression (11) in order to derive the total sample size. Here, the size is defined to be "total" because it includes both the size of the stratum completely enumerated and of the simple random sample without replacement from Stratum U_S. In the following it obviously holds that $n_C = N_C = N - N_S - N_E$; for notational simplicity, we first put $\psi \stackrel{\text{def}}{=} \left[c^2 t_y^2 - b^2(\hat{t}_y)\right]/(1 + \tilde{\delta})^2$, so that (11) can be rewritten as

$$N_S \frac{N_S - n_S}{n_S} S^2 = N_S \frac{N_S - n + n_C}{n - n_C} S^2 = \psi$$

from which we get

$$n\psi + nN_S S^2 = N_S^2 S^2 + n_C N_S S^2 + n_C \psi$$

Solving with respect to n, with some algebra we obtain the following result:

$$n = n_C + \frac{1}{\frac{1}{N_S} + \frac{\psi}{N_S^2 S^2}} = N - N_E - \frac{1}{\frac{1}{N_S} + \frac{S^2}{\psi}} \tag{12}$$

It is worth noting that the solution $n = N - N_E - 1/\left(\frac{1}{N_S} + \frac{S^2}{\psi}\right)$ corresponds to Formula (2.4) of Hidiroglou (1986) if we substitute the quantities $c^2 Y^2$ and N respectively with ψ and $(N - N_E)$. In other words, if we limit ourselves to single purpose and univariate surveys, the sampling design proposed here is an extension of Hidiroglou's (1986) take-all/take-some design to the case where a cut-off stratum is added. As said before, in many practical applications including business surveys this is a reasonable strategy.

3.2. Optimal Partition

In this subsection we deal with the problem of optimal partitioning of the population U in the three strata U_C, U_S, and U_E. In (12) the sample size n depends on c (that is chosen *a priori* by the researcher) on the bias $b(\hat{t}_E)$, on the total t_y and on the partition in the three strata. The partition determines four additional quantities, namely $\tilde{\delta}$, N_S, N_E, and S^2 (see the application in Section 4 for details about the computation of the quantities used in (12)).

Thus, if we denote with $\Phi = \{k_1, k_2, \ldots, k_N\}$ ($k_i \in \{C, S, E\}$) the generic element of the set Θ of the possible partitions of the population (whose cardinality is equal to 3^N), we conclude that n is a function of Φ and write

$$n = n(\Phi) \tag{13}$$

because all the other quantities listed above are either chosen by the researcher or computed using the auxiliary variables once a partition has been determined.

At this point it is quite clear that the problem consists in finding the partition Φ^* that minimizes (13) given the desired level of precision c. In particular, as our aim is the estimation of the totals $t_{y_j}, j = 1, \ldots, J$, of J variables by means of the same number J of auxiliary variables (see Section 2), the optimal sample size can be defined as follows:

$$n(\Phi^*) = \min \left\{ \max_{j=1,\ldots,J} n_j(\Phi) \right\} \tag{14}$$

The term $\max_{j=1,\ldots,J} n_j(\Phi)$ in (14) means that the optimization concerns, at each iteration, the largest of the sample sizes n_j corresponding to each auxiliary variable. Now (14) is the formalization of a combinatorial optimization problem, and the simulated annealing (Metropolis et al. 1953; Kirkpatrick et al. 1983; Geman and Geman 1984) is an appropriate tool for solving it. This algorithm enjoys several desirable properties (see Casella and Robert 2004, Section 5.2.3, for a review).

The implementation of the SA algorithm to the problem at hand can be summarized as follows.

(1) Choose an initial "temperature" T_0 and number of subiterations N_{sub}. These quantities will be used below.
(2) Stratify the population by means of a random uniform partition Φ_0, that is, assign to each of the N units of the population a label ϕ from the set $\{C, S, E\}$, where $P(\phi = C) = P(\phi = S) = P(\phi = E) = 1/3$. Let $\phi_i^{(0)}$ ($i = 1, \ldots, N$) be these labels.
(3) Perform substeps (a) and (b) for $i = 1, \ldots, N$.
    (a) Visit the ith unit of the population and put $\phi_i^{(1)} = \xi$, where ξ is a label drawn with uniform probability from the set $\{C, S, E\}$ and is the update of the label assigned to the ith unit at the 0-th iteration.Obviously, $\phi_j^{(1)} = \phi_j^{(0)}$ $\forall j \neq i$, so that the vector of labels $\phi^{(1)}$ at the first iteration differs from $\phi^{(0)}$ at most by one element.
    (b) Let $\Delta^{(1)} = n(\Phi^{(1)}) - n(\Phi^{(0)})$. If $\Delta < 0$, put $\phi_i^{(1)} = \xi$; otherwise, put $\phi_i^{(1)} = \xi$ with probability $\exp\{\Delta^{(1)}/T_0\}$ or $\phi_i^{(1)} = \phi_i^{(0)}$ with probability $1 - \exp\{\Delta^{(1)}/T_0\}$.

(4) Repeat step (3) N_{sub} times.
(5) Replace T_0 with $T_1 = f(T_0)$, where $f(\cdot)$ is a decreasing function that satisfies the conditions of the *annealing theorem* (Geman and Geman 1984). The function originally proposed by Geman and Geman (1984) was $T_{t+1} = f(T_t) = (\log(1+t)/\log(2+t))T_t$. Here we follow Sebastiani (2003) and use the so-called geometric temperature schedule $T_{t+1} = f(T_t) = \rho T_t$, with $\rho \in (0,1)$. The choice of f in applications has been the object of a lot of interest and some controversy in the literature: see Stander and Silverman (1994) and Casella and Robert (2004, p. 201), and the references therein. As for the numerical value of ρ, it is well-known that it has to be "large" enough to avoid a too rapid decrease of the temperature and "small" enough to keep the computation time reasonably short. Numerical experiments showed $\rho = 0.985$ to be a reasonable compromise.
(6) Repeat Steps (3)–(5) until some convergence criterion is met. We found it convenient to stop the algorithm the first time that one of the following conditions was satisfied: (i) in two successive iterations no labels are switched; (ii) $n_{\text{iter}} = 300$ iterations are reached. The numerical value of n_{iter} was again found by numerical experiments: in the present application (see Table 3 below) results are approximately stable after the first 100 iterations, so that $n_{\text{iter}} = 300$ seems to be large enough to guarantee an appropriate level of precision.

Notice that at Step 6, the tth iteration is just obtained by replacing (0) with (t) and (1) with ($t+1$) in Steps 2–5 above. At convergence, the algorithm determines the optimal partition Φ^*, that minimizes the total sample size n for a given precision level c.

4. A Case Study: The Slaughtering Monthly Survey

In this section we will apply the optimal cut-off design proposed above to the red meat slaughtering monthly survey performed by ISTAT (Italian National Institute of Statistics).[1] The goal consists in obtaining information concerning the number and the weight of the animals slaughtered monthly in Italy. This survey is based on a stratified sampling, with a stratification by kinds of slaughterhouse and geographical division, for a total of 5 strata, two of them with geographical references. Geographical divisions are North–West (1), North–East (2), Center (3), South (4) and Islands (5). Strata are the following:

- Stratum 1 (always totally observed): private slaughterhouses with European Economic Community (EEC) stamp in the geographical division 1 or 2;
- Stratum 2: private slaughterhouses with EEC stamp in the geographical division 3, 4 or 5;
- Stratum 3: private slaughterhouses with low capacity (regardless of geographical division);
- Stratum 4: private slaughterhouses in neglect, public with EEC stamp and public in derogation (apart from geographical division);
- Stratum 5: public slaughterhouses with low capacity.

[1] The code that was used in the study can be obtained by sending an email request to the first author.

The stratification also acts as a dimension-based criterion that assigns to Stratum 1 the enterprises with more than 10,000 sheep and goat or more than 50,000 pig slaughterings.

On average, the sample is of about 460 units for a population of 2,211 units with the desired level of precision c set to 5% (ISTAT 2007). In the following we will compare the ISTAT stratification, prepared by experts on the basis of their knowledge of the variables under investigation, to our optimal stratification.

In addition to the monthly survey, Istat performs yearly the census of the slaughterhouses. Thus, our frame contains $N = 2{,}211$ slaughterhouses for which we know four variables enumerated completely in 1999, 2000, and 2001. They are respectively the total number of slaughtered (i) cattle, (ii) pigs, (iii) sheep and goats and (iv) equines. We will first consider the complete dataset (for each of the three years) in order to assess the behavior of the bias $b(\hat{t}_E)$ and to look for possible regularities. Recalling that δ is defined as the ratio of the number of population units discarded to the number of population units sampled and enumerated completely, it is necessary to know the complete list of the lagged auxiliary variables.

The cut-off design proposed in this article will then be implemented, with the aim of setting up a monthly survey on slaughtering for the year 2002, using as auxiliary variables only the data enumerated completely in 2001. Our exercise consists in estimating the same totals estimated by ISTAT in its monthly survey.

We start with a brief description of the archive at hand. The scatterplots of all the pairs of the four variables in 2001 are shown in Figure 1; the same graphs for the years 1999 and 2000 are almost identical and therefore not reported here. The main evidence is that the variables are essentially uncorrelated, as confirmed by the linear correlation coefficient, that ranges in the interval $[-0.0096, 0.0566]$. Moreover, slaughterhouses are strongly specialized and most firms are small. In particular, 38.9% of the firms slaughter only one type of animals, 24.06% two types, 21.85% three types and only 15.2% all the four types of animals.

4.1. Bias Assessment

In order to implement the design developed in the preceding sections, it is crucial to analyze the bias $b(\hat{t}_E)$ of the estimator $\hat{t}_y$ given by (4), because the algorithm described in the preceding section requires as an input a starting value for $b(\hat{t}_E)$. We solved this problem with the help of empirical evidence concerning real data. Note that the "bias assessment" we give estimates absolute value of "bias" as an increasing function of increasing volumes of missing data. That would generally be true, but a clarification is in order about the results obtained by means of "bias assessment". They will usually be good, on average, if the massive amounts of historical data (good data) required are available, but, as in small area estimation, we are relying on a general effect, which may not be very good on a case-by-case application.

Figure 2 shows the absolute value of the bias $b(\hat{t}_{E_i})$, where the quantity $\hat{t}_{E_i}$ is defined as the total of the discarded population observed in 1999 and 2000, which in turn is given by the i smallest observations of the population. In other words, the ith point of the graph is the absolute value of the bias corresponding to $\hat{t}_{E_i}$, where E_i contains the i smallest observations of the population.

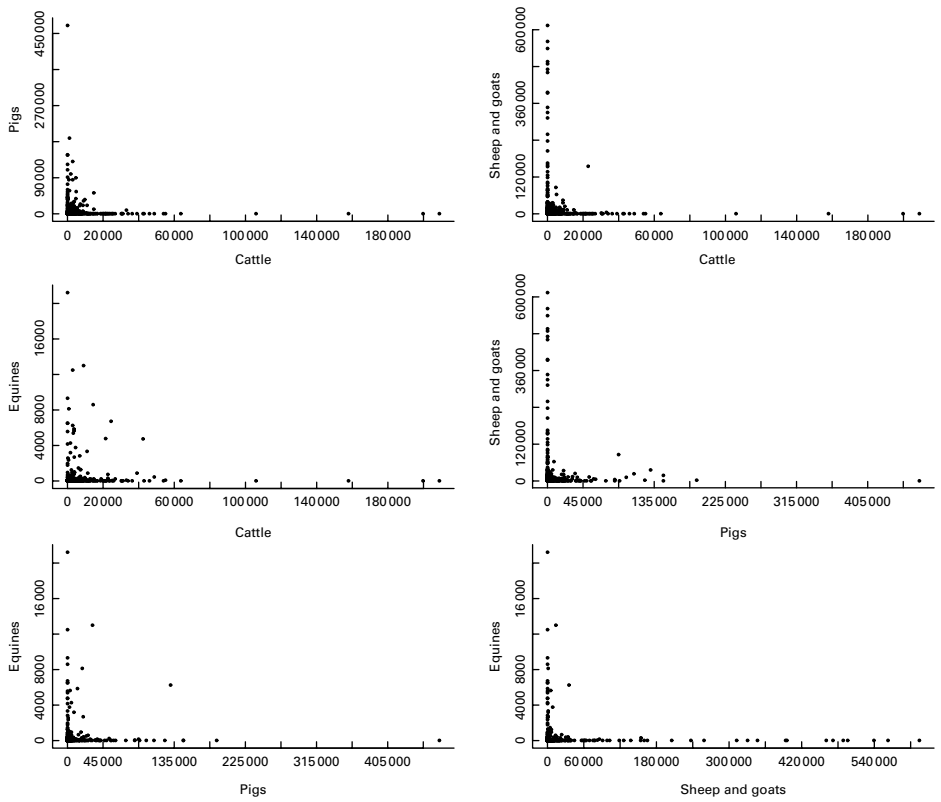

Fig. 1. Scatterplots of the data; the unit of measurement is number of animals

The procedure used to estimate the bias works as follows. If a complete enumeration of both the auxiliary variable x and the objective y (usually they are the same variable relevant to two different periods) is available, they can be ordered on the basis of the values of x:

$$x_{(1)}, \ldots, x_{(N)},$$

$$y_{(1)}, \ldots, y_{(N)},$$

where the (i) codes are such that $x_{(i)} \leq x_{(i+1)}$ for $i = 1, 2, \ldots, N-1$. Let now $C_{x,(i)}$ and $C_{y,(i)}$ be the respective cumulative sums:

$$C_{x,(i)} = \sum_{j=1}^{i} x_{(j)}, \quad C_{y,(i)} = \sum_{j=1}^{i} y_{(j)}$$

and $O_{x,(i)}$ and $O_{y,(i)}$ be the corresponding countercumulative sums:

$$O_{x,(i)} = t_x - C_{x,(i)} = \sum_{j=i+1}^{N} x_{(j)}, O_{y,(i)} = t_y - C_{y,(i)} = \sum_{j=i+1}^{N} y_{(j)}$$

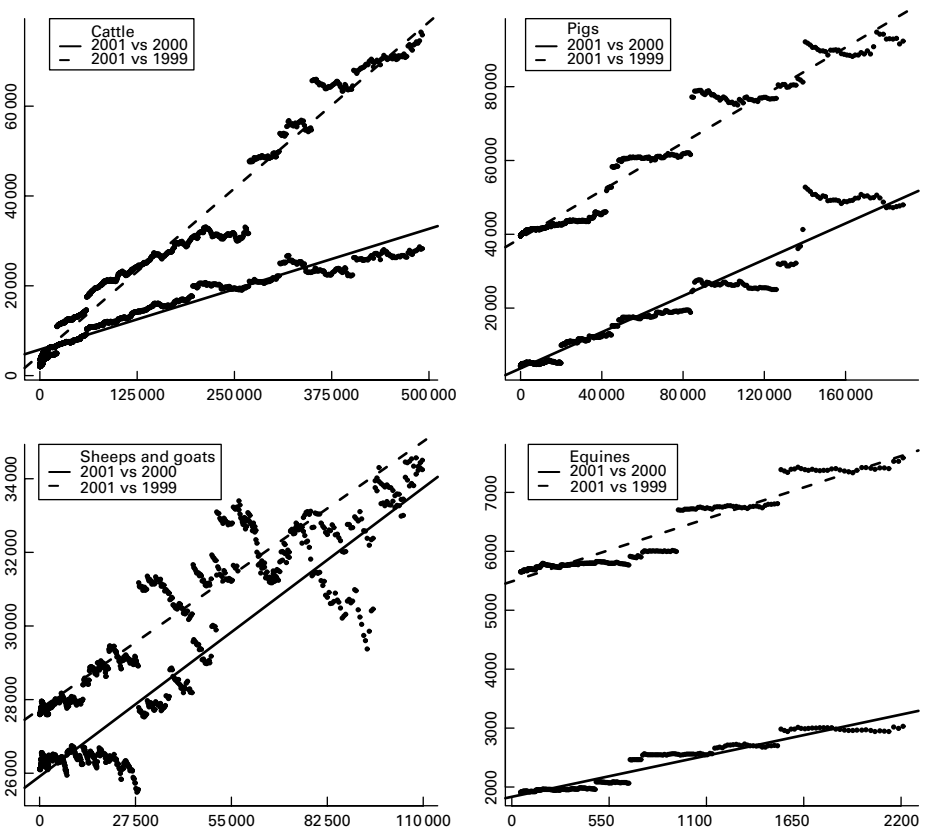

Fig. 2. The relationship between $\left|b(\hat{t}_{E_i})\right|$ *(y-axis) and* $C_{x,(i)}$ *(x-axis)*

Thus, if i is used as a threshold, according to (6), the absolute value of the bias obtained by using Estimator (4) can be written as:

$$\left|b\left(\hat{t}_{E_i}\right)\right| = \left|C_{x,(i)}\frac{O_{y,(i)}}{O_{x,(i)}} - C_{y,(i)}\right|$$

where the excluded part of the population is defined as $E_i = \{1, 2, \ldots, i\}$.

The $\left|b\left(\hat{t}_{E_i}\right)\right|$'s can be either entered directly in the optimization algorithm or used to construct a model, in order to simplify calculations and to obtain more stable results, i.e., not depending on particular discontinuities in the data frame. In our experiment we found satisfactory fits for the simple linear regression model:

$$\left|b\left(\hat{t}_{E_i}\right)\right| = \alpha + \beta C_{x,(i)} + \epsilon_i \tag{15}$$

Moreover we decided to exclude from the analysis the tails of the ordered distributions because the fit turned out to be better. In practical applications a threshold is usually neither a very small nor a very large value, so that this way of proceeding does not cause any problem.

The four graphs in Figure 2, corresponding to each auxiliary variable, have been obtained using respectively the complete 1999 and 2000 frame as a basis for the construction of the cut-off design in 2001. As expected, a larger temporal lag of the auxiliary information causes a significant modification of the bias: the bias for 1999 is in most cases larger than the bias for 2000. Furthermore, from the graphs it can be seen that the function that formalizes the relationship between $\left|\hat{b}\left(t_{E_i}\right)\right|$ and $C_{x,(i)}$ is well fitted by the linear model (15). Therefore we used the following estimated regressions:

$$\left|\hat{b}\left(\hat{t}_{E_i}\right)\right|_{j,2001} = \hat{\alpha} + \hat{\beta} C_{j,2000(i)}, \quad j = 1, \ldots, 4, \tag{16}$$

$$\left|\hat{b}\left(\hat{t}_{E_i}\right)\right|_{j,2001} = \hat{\alpha} + \hat{\beta} C_{j,1999(i)}, \quad j = 1, \ldots, 4. \tag{17}$$

Equations (16) and (17) actually give an estimate of the absolute value of the bias, but this is not a drawback because (12) only uses the square of this estimate. Detailed results are displayed in Table 2 respectively for cattle, pigs, sheep and goats, and equines. To assess the existence of a pattern of bias over time in the data series at hand, we also regress data from 2000 onto 1999, even though this regression will not be used for the determination of the optimal sample size. The estimates of the model $\left|\hat{b}\left(\hat{t}_{E_i}\right)\right|_{j,2000} = \hat{\alpha} + \hat{\beta} C_{j,1999(i)}$ $(j = 1, \ldots, 4)$ are shown in Table 2 as well.

The fit is extremely good in all cases. In particular, the values of the R^2 statistics are always large; that is not surprising if we consider that the variables used in the regression are cumulative sums. Thus, they are very likely correlated. As a result, standard statistical tests fail and are not reported here. We restrict ourselves to displaying the R^2 statistics, which in this setup should be interpreted as a descriptive measure. Notice that the regressions from 2001 onto 2000 and from 2000 onto 1999 are similar and show some evidence of temporal pattern.

We now apply our cut-off procedure in order to re-design, with respect to the year 2001, the ISTAT red meat slaughtering monthly survey. To this aim we use, as auxiliary information, the aforementioned frame. Thus, at each iteration of the algorithm and for

Table 2. Estimates and R^2-values for the three linear regressions

	Years	$\hat{\alpha}$	$\hat{\beta}$	R^2
	2000–1999	2,690	0.0681	0.8742
Cattle	2001–1999	4,621	0.1482	0.9662
	2001–2000	5,810	0.0538	0.9494
	2000–1999	37,740	0.1889	0.9082
Pigs	2001–1999	38,930	0.3230	0.9600
	2001–2000	3,664	0.2455	0.9541
	2000–1999	25,127	0.0804	0.9128
Sheep and goats	2001–1999	27,740	0.0661	0.9158
	2001–2000	25,910	0.0713	0.8517
	2000–1999	2,707	0.2144	0.8221
Equines	2001–1999	5,488	0.9674	0.9169
	2001–2000	1,834	0.6344	0.9286

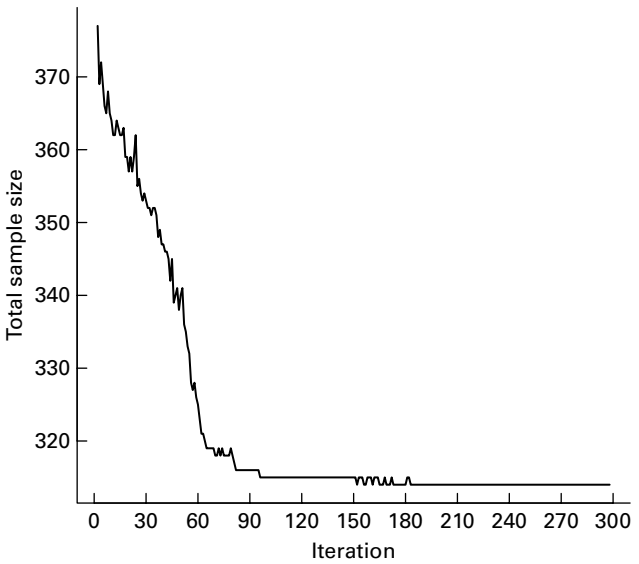

Fig. 3. Total sample size $n = N_C + n_S$ as a function of the SA iterations

each auxiliary variable, in $n(\Phi^{(t)})$ (see (14)) we substitute for the bias $b(\hat{t}_E)$ contained in (12) the estimate obtained via the estimated linear regression (16).

4.2. Sampling Design

Let us now finally turn to the results of the implementation of the cut-off design for the estimation of the same totals of the ISTAT red meat slaughtering monthly survey. Hence, in the objective function (14) the sample sizes $n_j(\Phi)$ ($j = 1, \ldots, 4$) are given by (12).

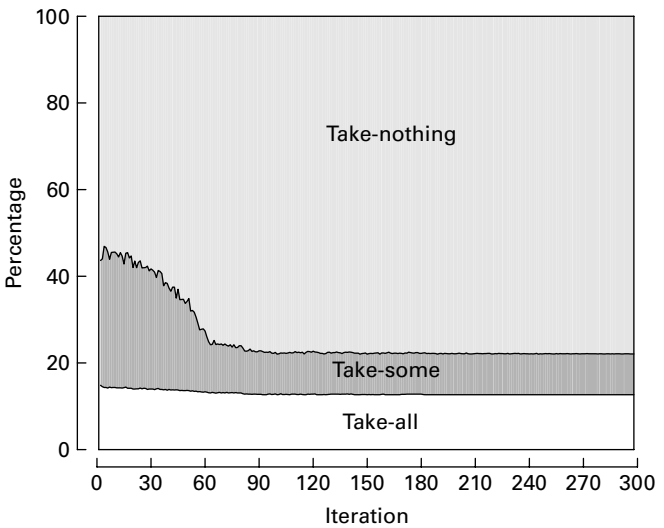

Fig. 4. Percentage composition of Strata U_C, U_S, and U_E as a function of the SA iterations

Table 3. Cut–off sampling results as a function of selected iterations of SA

Iter	n	N_C	N_S	N_E	# changes
1	797	738	736	737	–
2	377	328	637	1,246	2,544
3	369	321	652	1,238	2,095
4	372	317	720	1,174	2,139
5	369	317	714	1,180	2,192
6	366	314	694	1,203	2,179
7	365	319	652	1,240	2,133
8	368	316	690	1,205	2,151
9	365	317	690	1,204	2,101
10	364	317	691	1,203	2,089
20	357	309	620	1,282	1,827
30	353	309	613	1,289	1,613
40	347	307	517	1,387	1,286
50	340	302	454	1,455	940
60	325	295	310	1,606	489
100	315	280	207	1,724	117
150	315	280	207	1,724	61
200	314	280	210	1,721	25
250	314	280	209	1,722	5
298	314	280	208	1,723	2

In the algorithm we use a desired level of precision $c = 1\%$. The reason why we employ a desired level of precision of 1% instead of 5% as done by ISTAT is that cut-off sampling is considerably more efficient than standard stratified sampling. Thus, it is practically impossible to reach a 1% level of precision by means of the standard stratified approach used by ISTAT, unless the sample size is unrealistically large. However, at the end of this section we discuss the consequences of the use of different numerical values of c.

Figure 3 shows the total optimal sample size as a function of the number of iterations of the simulated annealing. It is immediately evident that the "largest decrease" in the sample size takes place in the first few iterations; the remaining iterations seem to provide just an adjustment towards the global optimum. More precisely (see Figure 4), approximately after the first 100 iterations, the algorithm just moves some observations from U_E to U_S; to these label-switching operations correspond very small decreases in the total sample size.

Table 4. Mean and coefficient of variation for each stratum and type of animal

		U_C	U_S	U_E
Cattle	μ	9,248	1,334	160
	CV	2.42	0.61	1.36
Pigs	μ	13,019	317	206
	CV	2.89	1.50	3.72
Sheep and goats	μ	34,393	302	583
	CV	2.83	1.96	3.84
Equines	μ	670	8.38	9.59
	CV	3.27	1.95	6.66

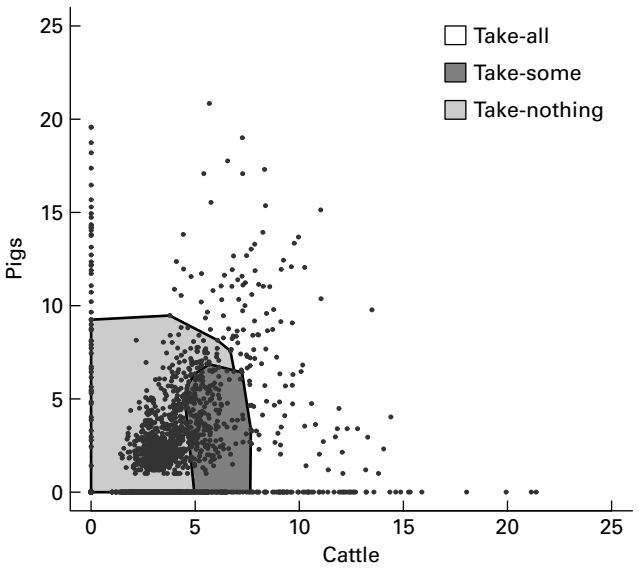

Fig. 5. Optimal partition of the population for the cattle and pigs (fourth roots of the data)

Table 3 gives some details about the implementation of the algorithm. The quantity N_S is the size of stratum U_S; the number of units actually sampled from this stratum can be computed as $n - N_C$; for example, at the 298th iteration (namely when the algorithm converges) we sample $n - N_C \approx 314 - 280 = 34$ units. It is worth adding that the algorithm is rather slow: one iteration takes almost five seconds, so that convergence is reached in approximately 22 minutes on a Pentium 4, 3.00 GHz. According to our computational experience, the convergence time increases very quickly as N gets large, so that the application of the method in large populations may be difficult.

The sampling scheme developed in this article produces the partition of the population shown in Table 4 and Figure 5. Only the partition concerning cattle and pigs is displayed. The graphs of the remaining pairs are very similar and therefore we omit them. This figure gives the scatterplot of the fourth roots of the two auxiliary variables. Observations in the light-grey portion of the graph are excluded (take-nothing), those in the white portion are enumerated completely (take-all), and those in the dark-grey portion are sampled (take-some).

The stratification is very clear-cut, with two strata (U_C and U_E) whose sizes are larger than U_S. The take-some stratum is nested into the take-nothing stratum, with a sampling

Table 5. Sizes of the strata and of the sample for various values of the bias

$b(\hat{t}_E)$	N_C	N_S	N_E	n
0.01	6	22	2,183	9
0.05	205	110	1,896	226
Case study	280	208	1,723	314
0.25	372	562	1,277	449
0.50	414	928	869	520
0.75	427	1,117	667	546

Table 6. Sizes of the strata and of the sample for various values of the level of precision

c	N_C	N_S	N_E	n
Case study	280	208	1,723	314
0.02	186	294	1,731	193
0.03	131	333	1,747	142
0.04	90	356	1,765	99
0.05	61	373	1,777	68

fraction equal to 16%. This means that in our application the sampling scheme is fairly similar to a take-all/take-nothing design. In fact, in the case study, 1,723 units are cut off, 280 units are completely enumerated and only 208 belong to the genuine sampling stratum (Table 3). It may seem surprising that approximately $3/4$ of the population can be left out without detriment to the survey. In addition, more than 50% of the units with positive inclusion probability belong to the completely enumerated stratum. However, this is not unexpected, because typically it happens if a stratification variable is highly correlated with the study variable, which is the case here given the short time span between censuses.

Moreover, according to the theoretical results derived in the preceding section, such a small sampling fraction was expected. Considering the large concentration of the population, stratum U_S contains mostly the firms with values of all the four auxiliary variables are different from zero, namely the least specialized ones.

To complete this discussion, consider Tables 5 and 6. Table 5 shows a comparison of the outcomes just presented and some results obtained by means of cut-off sampling as the bias $b(\hat{t}_E)$ changes. The position of our case study in the table is explained by the fact that, even though we have used four numerical values for the bias (as many as the surveyed variables), the average value of the bias is approximately equal to 0.1.

Table 6 has the same structure as the preceding one, but the results of cut-off sampling vary as a function of the desired level of precision c.

From Table 5 it is clear that, if the bias is negligible (i.e., $b(\hat{t}_E) = 0.01$), the auxiliary variables are essentially identical to the target variable. It follows that the survey is almost unnecessary ($n = 9$). As the bias increases, N_E gets smaller. This does not happen as c decreases (see Table 6). Thus, the size of U_E appears to be mainly a function of the bias. On the other hand, Table 6 shows that the subdivision of $(N - N_E)$ in U_C and U_S depends mostly on c. The smaller the desired level of precision, the larger the number of completely enumerated units.

Recall that the desired level of precision c was set at 1%. This value has also been employed to perform the following comparisons, that show the considerable advantages of our approach in terms of sample size corresponding to the predetermined level of precision. Table 7 displays detailed results concerning some direct competitors of the cut-off design; in particular, Table 7(a) shows the sample sizes corresponding to the Hidiroglou approach, Table 7(b) gives the sizes obtained stratifying the population with the *K*-means algorithm (Rencher 2002, Section 14.4.1a) used as a minimizer of the variance. Finally, the sample size corresponding to the ISTAT design introduced at the beginning of this section (but setting $c = 1\%$ instead of 5% used by ISTAT for its official

Table 7(a). Sample sizes using Hidiriglou's approach

	n	N_C	N_S
$y_{1,2001}$: cattle	476.97	332	1,879
$y_{2,2001}$: pigs	301.03	246	1,965
$y_{3,2001}$: sheep and goats	291.29	229	1,982
$y_{4,2001}$: equines	227.11	180	2,031
Union	744.26	663	1,548

Table 7(b). Sample sizes using the K-means algorithm

# strata	2	3	4	5	6	7	8	9	10
n	2,193	2,094	1,800	1,689	1,259	1,206	1,145	990	649

survey) is equal to $n = 866$ with 5 strata. In the last row of Table 7(a), N_C is the size of Stratum U_C obtained as the union of the four strata enumerated completely with respect to each auxiliary variable (reported in the first four rows of the table). This is one way of rendering Hidiroglou's approach, that is single purpose and univariate, comparable to our technique, that is multipurpose and multivariate.

5. Conclusions

The goal of this article consisted in proposing a framework for cut-off sampling where a model-based estimator of the unobserved part of the population plays a crucial role in introducing a bias into the final estimates. The rationale for this proposal is based on the assumption that often the population distributions are highly skewed, with a huge number of small units whose weight on the population total is negligible. We have discussed a formal approach for combining estimation and optimal partition of the population in three strata: census, sample and exclusion. We view this issue jointly with the multipurpose allocation of sampling units in the case where multivariate partitioning variables are available.

We have used the SA algorithm to minimize the number of observed units necessary to satisfy a required precision. This is expressed in terms of MSE of the estimates of the population total. The results are encouraging. For example, for $c = 1\%$, the sample size obtained using the present approach is approximately 50% to 60% less than its direct competitors.

These outcomes also shed some light on the directions of future research in this field. In particular, we believe that attention should be focused on the bias of the estimator with the purpose of tackling at least two issues:

- assess the robustness of the design with respect to variations of the functional form of the bias function (that here was assumed to be linear);
- use the estimated value of the bias not only for finding the optimal sample size but also for correcting the bias of the chosen estimator (whatever it is).

Finally, the last problem is related to the fact that the SA algorithm is rather slow, so that the computational burden may become heavy when the population is large.

6. References

Bailar, B.A., Isaki, C.T., and Wolter, K.M. (1983). A Survey Practitioner's Viewpoint, Proceedings of the American Statistical Association, Survey Research Methods Section, 16–25. http://www.amstat.org/sections/SRMS/proceedings/papers/1983_004.pdf.

Casella, G. and Robert, C.P. (2004). Monte Carlo Statistical Methods, (second edition). New York: Springer.

Cochran, W.G. (1977). Sampling Techniques, (third edition). New York: Wiley.

Dalenius, T. (1952). The Problem of Optimum Stratification in a Special Type of Design. Skandinavisk Aktuarietidskrift, 35, 61–70.

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

Elisson, H. and Elvers, E. (2001). Cut-off sampling and Estimation. Proceedings of Statistics Canada Symposium.

Geman, S. and 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.

Glasser, G.J. (1962). On the Complete Coverage of Large Units in a Statistical Study. Review of the International Statistical Insitute, 30, 28–32.

de Haan, J., Opperdoes, E., and Schut, C.M. (1999). Item Selection in the Consumer Price Index: Cut-off versus Probability Sampling. Survey Methodology, 25, 31–41.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. II. New York: Wiley.

Hidiroglou, M.A. (1986). The Construction of a Self Representing Stratum of Large Units in Survey Design. The American Statistician, 40, 27–31.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by Simulated Annealing. Science, 220, 671–680.

Knaub, J.R., Jr. (2004). Modeling Superpopulation Variance: Its Relationship to Total Survey Error. InterStat, August, http://interstat.statjournals.net/.

Knaub, J.R., Jr. (2007). Cutoff Sampling and Inference. InterStat, April, http://interstat.statjournals.net/YEAR/2007/abstracts/0704006.

Knaub, J.R., Jr. (2008a). Cutoff vs. Design–Based Sampling and Inference for Establishment Surveys. InterStat, June, http://interstat.statjournals.net/YEAR/2008/abstracts/0806005.php.

Knaub, J.R., Jr. (2008b). Cutoff Sampling. In Encyclopedia of Survey Research Methods, P.J. Lavrakas (ed.). London: Sage.

ISTAT (2007). Dati mensili sulla macellazione delle carni rosse. http://www.istat.it/agricoltura/datiagri/carnirosse. [In Italian]

Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. Survey Methodology, 14, 33–43.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21, 1087–1092.

Rencher, A.C. (2002). Methods of Multivariate Analysis, (second edition). New York: Wiley.

Rivest, L.P. (2002). A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys. Survey Methodology, 28, 191–198.

Royall, R.M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. Biometrika, 57, 377–387.

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

Sebastiani, M.R. (2003). Markov Random-field Models for Estimating Local Labour Markets. Applied Statistics, 52, 201–211.

Stander, J. and Silverman, B.W. (1994). Temperature Schedules for Simulated Annealing. Statistics and Computing, 4, 21–32.

Sigman, R.S. and Monsour, N.J. (1995). Selecting Samples from List Frames of Businesses. In Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds). New York: Wiley, 133–152.

Statistics Canada (2001). Monthly Survey of Manufacturing (MSM), Statistical Data Documentation System, Reference Number 2101, Statistics Canada.

Received October 2008
Revised October 2009