Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 20

11-1-2016

Optimal Estimation and Sampling Allocation in Survey Sampling Under a General Correlated Superpopulation Model

Ioulia Papageorgiou
Athens University of Economics and Business (AUEB), ioulia@aueb.gr

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 293-321.
doi: 10.22237/jmasm/1478002680

ISSN 1538 − 9472

Optimal Estimation and Sampling Allocation in Survey Sampling Under a General Correlated Superpopulation Model

Ioulia Papageorgiou
Athens University of Economics and Business
Athens, Greece

Sampling from a finite population with correlated units is addressed. The proposed methodology applies to any type of correlation function and provides the sample allocation that ensures optimal efficiency of the population parameters estimates. The expressions of the estimate and its MSE are also provided.

Keywords: Superpopulation, systematic sampling, model-based sampling, sampling strategy, optimality, autocorrelation

Introduction

In classical sampling theory, the finite population under study is assumed to be a fixed vector of dimension N, where N is the number of population members. If U denotes the population set and Y the variable of interest, the population vector can be denoted as $U = \{Y_1, Y_2, \ldots, Y_N\}$, and is assumed to be fixed but is in general unknown. The superpopulation approach in sampling from a finite population is assumed in this work. According to this approach, the finite set of measurements U is a realization of a sample of size N drawn from an infinite population with common distribution ξ.

The superpopulation model was introduced by Cochran (1946 and 1977, 1953) and further developed by Godambe (1955), Cassel et al. (1977), Tam (1984), Blight (1973), Mukerjee & Sengupta (1989, 1990) and Bolfarine & Zacks (1992), among others. The problem of finding optimum sampling schemes under a superpopulation model is discussed by several authors including Blight (1973), Papageorgiou & Karakostas (1998), Arnab (1992), Mukerjee & Sengupta (1989, 1990), Nayak (2003) and Chao (2004). The superpopulation model assumes the population measurements are comprised of a deterministic and a non-

Dr. Papageorgiou is an Assistant Professor in the Department of Statistics. Email her at ioulia@aueb.gr.

deterministic element that can be attached to a variable. More analytically, the superpopulation model in its general form is

$$Y_i = \mu_i + \varepsilon_i,\ i = 1,2,\ldots,N \tag{1}$$

where μ_i is constant, representing the deterministic part, while ε_i are random variables also called errors. The random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ is assumed to have zero mean and variance covariance matrix $\boldsymbol{V}$. Various special models to describe more specific or realistic population assumptions can be derived from (1) by making assumptions on matrix $\boldsymbol{V}$ and relationships among μ_i. For example, model

$$E_x(Y_i) = \mu_i \text{ and } E_x(Y_i - \mu_i)(Y_j - \mu_j) = \begin{cases} \sigma^2, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

where the errors are uncorrelated and with constant variance is the model that describes a finite population with uncorrelated measurements and different superpopulation mean.

Another special case is the model where

$$E_x(Y_i) = \mu \text{ and } E_x(Y_i - \mu_i)(Y_j - \mu_j) = \begin{cases} \sigma^2, & \text{if } i = j \\ \rho\sigma^2, & \text{if } i \neq j \end{cases}$$

according to which the population units are correlated with a constant correlation ρ and constant superpopulation parameter of mean μ.

A more realistic autocorrelated superpopulation model results if one assumes that the degree of correlation among two population units depends on between-unit distance. This is also known as serial correlation. Populations that exhibit this characteristic can be encountered in applications where an order is assigned to each of the population members. The ordering can be according to time, space, magnitude or the serial number in a production line. The model with serial correlation was first introduced by Cochran (1946) and it can be written in mathematical terms as

$$E_x(Y_i) = \mu \text{ and } E_x(Y_i - \mu_i)(Y_j - \mu_j) = \sigma^2 \rho(|i - j|) \tag{2}$$

where $\rho(h)$ is the autocorrelation function of the population model for units at distance h.

All above models can also be seen as special cases of the more general superpopulation regression model where the deterministic part μ has been modeled as linear functions of a set of auxiliary nonstochastic variables that may be available for the population vector (Bolfarine and Zacks, 1992).

Madow & Madow (1944), Cochran (1977), Royall (1970), Blight (1973), Ramakrishnan (1975), Bellhouse & Rao (1975) and Graubard & Korn (2002), among others, assumed (2) or a special case of this. The results available from the literature aim to answer two questions: first, to estimate the superpopulation parameter μ; and second, to determine the optimal sampling design. The optimal sampling design is the selection process according to which the sample units are drawn from the population so that the derived estimate will achieve an assumed optimality criterion, such as minimum variance. Sampling strategy is the pair of the sampling design and estimator used towards the estimation problem (see for example Ramakrishnan, 1975). Often in practice, certain properties are attached to the autocorrelation function $\rho(h)$ such as positive, decreasing or convex. An outline of related results from the literature is presented in the following section.

In this current work the assumptions made on function $\rho(h)$ are extended. More specifically, $\rho(h)$ can be the autocorrelation function of any random process with second-order stationarity. The proposed methodology aims to determine the optimal allocation of the sampling units for a sample of size n, when the least squared estimator of the superpopulation mean is used as a criterion of optimality. The optimum is defined with respect to the mean squared error (*mse*) of the estimate. The proposed optimal sampling strategy is completed by providing the statistical inference of the assumed estimate when the sample is selected, according to the derived optimal sampling scheme. Both the derived optimal sample allocation and its *mse* depend on $\rho(h)$ and therefore take into account the specific autocorrelation of the population under study.

General notation and brief review

Denote by $s = \left\{ Y_{j_1}, Y_{j_2}, \ldots, Y_{j_n} \right\}$ the sample of size n that is selected from the complete vector U. Indexes j_i (i = 1, 2, …, n) in the notation indicate the positions of the selected units in the population U. $\theta = \sum_{i=1}^{N} Y_i$, the population sum, is considered as the parameter of interest. θ is a linear function of the population measurements. Dealing with the estimation of θ is equivalent with the estimation

of population mean $\bar{Y} = \sum_{i=1}^{N} Y_i / N$, as the two quantities differ only by a constant coefficient.

The aim of the sampling procedure is to estimate θ based on a set of measurements, *s*, selected from *U*. The assumption of selection without replacement is made, but sampling with replacement is equally possible. The sampling design is determined by the probability $p(s)$ that is assigned to each of all possible samples *s* selected from the population. Let P_n denote this set of all possible samples of size *n*. Important probabilities related to the design $p(s)$ are the first and second order inclusion probabilities π_i and π_{ij} respectively, defined as

$$\pi_i = \sum_{s' \ni i} p(s) \text{ and } \pi_{ij} = \sum_{s' \ni i,j} p(s).$$

By making use of this notation, simple random sample (that is, the simplest sampling design) is the design that assigns equal probability $p(s) = 1/\binom{N}{n}$ in every sample *s* that belongs in P_n, where P_n is the selection of all the possible combinations of *n* measurements chosen from *U* in this case. For the systematic scheme, the probabilities of selection are also equal, $p(s) = 1/k$, where $k = N/n$. The number of samples that belong in P_n is also *k* in the systematic case and if s_i, $i = 1, 2, \ldots, k$ is a representative sample, then $s_i = (Y_i, Y_{i+k}, Y_{i+2k}, \ldots, Y_{i+(n-1)k})$, $i = 1, 2, \ldots, k$ (see for example Cochran, 1977). If $N \neq nk$, a slight complication and need for modification arises, but the effect is negligible (Yates, 1960, 1948 1st ed.). The samples generated by a systematic procedure are equally spaced, and moreover if the start Y_i is chosen with *i* such that $2i = N + 1 - (n - 1)k$, the sample is a centrally located systematic sample (Blight, 1973). In this last case P_n contains only one sample *s* with $p(s) = 1$.

Blight in the previously mentioned work assumes that the deviations of population values from the superpopulation mean μ are generated by an autoregressive model of order one, AR(1), e.g.

$$Y_i - \mu = \lambda (Y_{i-1} - \mu) + \varepsilon_i. \quad (3)$$

where ε_i is uncorrelated normally distributed series with zero mean and constant variance σ^2. This yields $\rho(h) = \lambda^h$ at lag h ($h = 1, 2, \ldots, N-1$). Employing the sample mean as the estimator of the corresponding population mean, the effect of the autocorrelation is studied and the optimal sampling design when λ is positive

or negative is obtained. The optimality criterion is the conditional variance $Var_{\xi}\left(\bar{Y} \mid Y_{j_i} \in s\right)$. The sign of λ controls the monotonicity shape of $\rho(h) = \lambda^h$ and, as expected, the resulting optimal design is remarkably different among the two cases. More specifically, for $\lambda > 0$ the optimal sample is the centrally located systematic and for $\lambda < 0$ the optimal sampling design is concentrated towards the two ends of the population. This also verifies the fact that the optimal solution for the sampling scheme is not unique, but depends on the specific type of the autocorrelation. However, when the autocorrelation function $\rho(h)$ is not only λ^h, $h > 0$, but in general any positive, decreasing and convex function, the same result holds and the centrally located systematic design is the optimal (Papageorgiou & Karakostas, 1998).

Function $\rho(h)$ is defined in all positive integer numbers and therefore $\rho(h)$ is decreasing if $\rho(h + 1) - \rho(h) \leq 0$ ($\Delta\rho(h) \leq 0$), while convexity holds when

$$\mathrm{D}^2 r\left(h\right) = r\left(h+2\right) - 2r\left(h+1\right) + r\left(h\right) \geq 0 \text{ for } h = 0,1,2,...$$

Denote by K the class of all autocorrelation functions that satisfy the aforementioned properties (positive, decreasing and convex). AR(1) model assumed in (3) has an autocorrelation function that belongs in K when $\lambda > 0$ and since the optimality of the centrally located systematic scheme holds for the whole class K it also holds for this occasion as a special case. In fact, class K includes a wide range of correlation functions (Bellhouse, 1984).

Although the question about the optimum sampling scheme seems to have a unique answer when $\rho(h) \in K$ and it is closely related to the systematic scheme, under almost any combination of estimators and optimality criterions considered, the problem remains when $\rho(h)$ does not belong in K. The optimum sampling scheme in this case can be quite far from the systematic and it varies depending on the specific type of $\rho(h)$. In other words, there is no uniquely defined optimum sampling scheme that can cover any random process with respect to the sampling problem. In this direction, a practical and easy-to-implement methodology, that suggests the optimum sampling procedure once the specific type of $\rho(h)$ or $\boldsymbol{V}$ is provided, is proposed in this paper.

A related work is provided by Chao (2004), where a general known matrix $\boldsymbol{V}$ is assumed, and a similar to principal component analysis method is suggested in order to obtain the sampling procedure. More specifically, the idea is to choose as sampled units those population units pointed from the n most important components or the largest eigenvalues of matrix $\boldsymbol{V}$. Two algorithms are proposed,

called Design I and Design II, with the second being a slight modification of the former. Design I makes use of the n eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of $\boldsymbol{V}$ that correspond to the n first-in-magnitude eigenvalues of the matrix. If $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \ldots, e_{iN})$ is such an eigenvector and j, ($j = 1, \ldots, N$) is the index with the largest-in-absolute-magnitude component in $\boldsymbol{e}_i$, the population unit that corresponds in position j is the one selected in the sample according to this design. If the unit is already in the sample, the second-in-absolute-magnitude component is selected. Design II works in the same principle, with the difference that the sign of the components is also taken into account. From each eigenvector two components are selected, the largest-in-absolute-magnitude and the second-largest with opposite sign of the first. The approach for both designs is rather intuitive and the resulting designs do not hold any optimality criterion. Their performance is measured by the relative efficiency over the simple random sample as a general sampling scheme. They indicate improved efficiency with respect to the simple random most of the times, but their performance is not stable and the simple random sample itself is usually far from the optimal when a correlation exists.

Before dealing with the problem and proposing the solution of the optimal sampling design, a list of possible applications is provided. The range of applications is wide, and they cover any scientific area where the framework includes correlated measurements and a sample is selected from the population. A typical application of sampling from autocorrelated populations where the autocorrelation is not necessarily decreasing and convex is seen in the context of statistical process control in monitoring manufacture and industrial production lines. A variety of control charts or other statistical instruments can be constructed based on a set of measurements selected from the process, and help practitioners to derive information or warning if the process is out of control. Traditionally the statistical theory behind the control charts is based on the assumption that the sample measurements are independent. It is however quite common in practice—and especially in continuous manufacture or production lines—that this assumption is violated, and this produces misleading and unreliable control charts (Alwan, 1992; Montgomery and Mastrangelo, 1991) with tighter control limits than the true ones. A lot of attention has been drawn lately to this area of research; see for example Alwan and Roberts (1988), Harris and Ross (1991), Mastrangelo and Montgomery (1995), Apley and Lee (2003) and Lu and Reynolds (1999, 2001), and all proposed approaches make use of the present autocorrelation to either modify the existing control limits, or to model the process, identify the autocorrelation, and use the independent errors instead of the measurements for constructing any statistical tool. The models that have been assumed are AR(1)

(Autogregressive), MA(1) (Moving Average) and ARMA(1,1) (Autoregressive Moving Average) (Wardell et al., 1992) and efficiency in sampling and therefore construction of the control limits provided can be improved further if the specific type of correlation is taken into account.

Similarly, geostatistical data in spatial statistics very often exhibit a small-scale variation, typically a strong correlation between data at neighboring locations (Watson, 1972). If the population mean is the parameter of interest, failure to realize the presence of positive correlation in the data leads to very narrow confidence interval (Cressie, 1993), a result similar with this in quality control charts. The superpopulation model is therefore extensively used in modeling geostatistical data in order to accommodate this correlation (Cressie, 1993). In this context, let $\boldsymbol{s} \in \mathbb{R}^d$ be the data location in d-dimensional Euclidean space and $\boldsymbol{Y}(\boldsymbol{s})$ the measured data, assumed random, at location $\boldsymbol{s}$. Assuming that $\boldsymbol{s}$ takes values over an index set $D \subset \mathbb{R}^d$, the superpopulation model results as a realization $\{y(\boldsymbol{s}): \boldsymbol{s} \in D\}$ from the multivariate random field $\{\boldsymbol{Y}(\boldsymbol{s}): \boldsymbol{s} \in \mathrm{D}\}$. Land and agricultural surveys, ground-water monitoring, environmental statistics and socio-economic habitat surveys are some of the sampling applications in two dimensions with spatial dependence among population units.

Other applications of sampling from correlated populations include genetics and ecological statistics. In particular, the superpopulation model is often used to explain genetic or ecological patterns where the covariance in the genetic makeup of individuals or in the growth of populations can be assumed to be a function of the spatial distance separating the units (Lande, 1991; Bjørnstad et al., 1999).

Clustered data, often found in social, educational, psychometric and behavioral studies, also represent an application of sampling from correlated populations. Clustered data may result either because of repeated measurements in time such as in longitudinal studies or because of sub-sampling from a large primary unit: for instance, sampling graduates from the same educational institute or the same region/country for a large scale study. The existing intra-class correlation has to be taken into account during the analysis and the statistical inference in order to produce valid results (Neuhaus and Kalbfleisch, 1998). Moreover the knowledge of the intra-class correlation can contribute at the selection stage of the sub-sampling.

Another application of sampling in time series, apart from the serial correlation and the typical applications described already, is the use of composite marginal likelihoods in order to estimate the parameters of the model (Cox and Reid, 2004; Varin, 2008). Pairwise likelihoods, based only on the bivariate joined distributions of the measurements, produce estimates very close to those under the

full likelihood with respect to the dimension. The benefit of pairwise likelihood is on the computational demand that is required for the optimization. Moreover, further improvement in this direction can be achieved if not all possible pairs but only a selection of them will be used instead. Current work in this context shows that the same accurate estimates can be obtained if the correlation between observations is taken into account towards the selection procedure: for example, pairwise likelihood of order *m* (Hjort and Varin, 2008).

The general problem

Model (2) describes the population and $\rho(h)$ is assumed to be any autocorrelation function. Moreover, $\hat{q}$, the least squared estimator for the parameter θ, is assumed as the optimality criterion. The aim is to determine the sampling design p or the sample s that minimizes the mean square error of $\hat{q}$ under this model. The least squared estimator of the population mean is the sample mean and it is unbiased under model (2) (see Karakostas, 1984), which yields that

$$\hat{\theta} = \tfrac{N}{n} \sum_{i=1}^{n} Y_{j_i} \tag{4}$$

The mean square error of $\hat{q}$ when a sample $s = \left\{ Y_{j_1}, Y_{j_2}, \ldots, Y_{j_n} \right\}$ is provided is given by

$$\begin{aligned} mse\left(\hat{q} \middle| s\right) &= E\left[\left(\hat{q} - q\right)^2 \middle| s \right] \\ &= Var\left(q\right) + Var\left(\hat{q}\right) - 2Cov\left(\hat{q}, q\right) \\ &= Var\left(\sum_{i=1}^{N} Y_i \right) + \frac{N^2}{n^2} Var\left(\sum_{i=1}^{n} Y_{j_i} \right) - 2\frac{N}{n} Cov\left(\sum_{i=1}^{n} Y_{j_i}, \sum_{i=1}^{N} Y_i \right) \end{aligned}$$

Let $\boldsymbol{V}$ be the variance covariance matrix of the complete population vector under (2). The partition of matrix $\boldsymbol{V}$ according to the sampled part, *s*, is considered next and let $\boldsymbol{V}_s$ denote the part of $\boldsymbol{V}$ that corresponds to the sampled units and $\boldsymbol{V}_{s,U}$ the $n \times N$ matrix of $\boldsymbol{V}$ where its rows correspond to the sampled units while the columns to the whole population *U*. Under this notation the *mse* can be written as

$$mse\left(\hat{q} \middle| s\right) = \mathbf{1}_N' \boldsymbol{V} \mathbf{1}_N + \frac{N^2}{n^2} \mathbf{1}_n' \boldsymbol{V}_s \mathbf{1}_n - 2\frac{N}{n} \mathbf{1}_n' \boldsymbol{V}_{s,U} \mathbf{1}_N \tag{5}$$

where $\mathbf{1}_j'$ stands for the j–dimension vector of units. For the sampling problem it is necessary to minimize $mse(\hat{q})$ with respect to the sample s, or equivalently to find the minimum

$$\min_s \left\{ \tfrac{N}{n} \mathbf{1}_n' \boldsymbol{V}_{s,U} \mathbf{1}_n - 2 \to \mathbf{1}_n' \boldsymbol{V}_{s,U} \mathbf{1}_n \right\} \tag{6}$$

For any sample $s = \left\{ Y_{j_1}, Y_{j_2}, \ldots, Y_{j_n} \right\}$ let $h_i = j_{i+1} - j_i$, 1, 2, …, $n-1$ denote the distances of two successive sampled units with moreover $h_0 = j_1 - 1$ and $h_N = N - j_n$ the two end distances. Under this notation any sample s can also take the form $s = \left\{ Y_{h_0+1}, Y_{h_1+h_0+1}, \ldots, Y_{h_{n-1}+\cdots+h_1+h_0+1} \right\}$ and uniquely represented by the vector of distances $h = (h_0, h_1, h_2, \ldots, h_n)$ with $h_i, i = 0,1,\ldots,n$ to be integers with $h_0 + h_1 + \ldots + h_n = N - 1$. Using this equivalent notation for the sample s the minimization expression can finally be written as

$$\min_{h_i} \left[\begin{array}{l} N - 2n + \dfrac{2N}{n} \left(\begin{array}{l} \displaystyle\sum_{i=1}^{n-1} \rho(h_i) + \sum_{i=1}^{n-2} \rho(h_i + h_{i+1}) \\ \qquad + \displaystyle\sum_{i=1}^{n-3} \rho(h_i + h_{i+1} + h_{i+2}) + \cdots + \rho(h_1 + \cdots + h_{n-1}) \end{array} \right) \\ \quad -2 \left(\displaystyle\sum_{i=1}^{h_0} \rho(i) + \sum_{i=1}^{h_0+h_1} \rho(i) + \cdots + \sum_{i=1}^{h_0+\cdots+h_{n-1}} \rho(i) + \sum_{i=1}^{h_1+\cdots+h_n} \rho(i) + \cdots + \sum_{i=1}^{h_n} \rho(i) \right) \end{array} \right]$$

or equivalently

$$\min_{h_i} Q(h_0, h_1, h_2, \cdots, h_n)$$

where

$$Q=\frac{N}{n}\left(\begin{array}{l}\sum_{i=1}^{n-1}\rho\left(h_i\right)+\sum_{i=1}^{n-2}\rho\left(h_i+h_{i+1}\right)\\ \qquad+\sum_{i=1}^{n-3}\rho\left(h_i+h_{i+1}+h_{i+2}\right)+\cdots+\rho\left(h_1+\cdots+h_{n-1}\right)\end{array}\right) \tag{7}$$
$$-\left(\sum_{i=1}^{h_0}\rho\left(i\right)+\sum_{i=1}^{h_0+h_1}\rho\left(i\right)+\cdots+\sum_{i=1}^{h_0+\cdots+h_{n-1}}\rho\left(i\right)+\sum_{i=1}^{h_1+\cdots+h_n}\rho\left(i\right)+\cdots+\sum_{i=1}^{h_n}\rho\left(i\right)\right)$$

Finding the optimum sample (s) is now a constrained minimization problem of minimizing (7) with respect to the unknowns $h_i, i = 0,1,\ldots,n$. The parameter constrains, that mainly result from their definition, are

$$\begin{aligned}&0 \le h_0 \le N-1\\ &0 < h_i \le N-1,\ i = 1,2,\ldots,n-1\\ &0 \le h_n \le N-1 \text{ and}\\ &h_0 + h_1 + \ldots + h_n = N-1\end{aligned} \tag{8}$$

Therefore, the sampling problem is mathematically formulated as a constrained minimization problem. However the mathematical solution is not straightforward, due to the unknown integer function $\rho(h)$ involved in Q. Unless certain properties are assumed for $\rho(h)$ the problem cannot be solved in its general case. The difficulty is mainly caused from the upper bounds of the summations in the second parenthesis of Q that depend on the unknowns h_i and make the number of the terms in those summations a variable itself.

Methodology

A Solution Based On The Continuous Approximation

The objective function Q given by (7) is in general a sum of values of the function $\rho(h)$. Function $\rho(h)$ on the other hand represents the autocorrelation function of the population series and takes values at lag $h, h = 0,1,2,\ldots,N-1$, being therefore an integer defined function. The integer feature of $\rho(h)$ leads to the summations appearing in Q that in turn prevent from its minimization.

The idea is to use an approximate, but approachable towards its minimization expression instead of Q. The approximation consists of two stages, first to approximate every sum that appears in the second parenthesis of Q by an

integral and secondly to approximate the integer function $\rho(h)$ with a continuous function. Approximating a sum with an integral is a known practice in literature and departs from the Euler-Maclaurin formula. The aim is to use Euler-Maclaurin formula in order to obtain a continuous approximation of the objective function and provide bounds for the error in the approximation. Note however that the derivation of the point(s) $(h_0, h_1, h_2, \ldots, h_n)$ where the minimum is attained will suffice the sampling problem and will provide with the optimal sample. Once the optimal sample is determined the corresponding for the estimate exact *mse* under the optimal sample can be calculated by a single substitution in (5) and not through its continuous approximation. In other words, the approximate and the true versions of Q need only to share the same monotonicity and not coincide. The second condition is stronger and guarantees the first.

Euler-Maclaurin formula is a mathematical tool, an equality, where a finite sum of values of a function f at the left side part is expressed as a finite integral of the same function f plus an error term at the right side part. The error term involves all consecutive derivatives of f, the Bernoulli numbers and Bernoulli polynomials. More analytically, it holds

$$\sum_{k=a}^{b-1} f(k) = \int_a^b f(x)\,dx + \sum_{k=1}^{m} \frac{B_k}{k!} f^{(k-1)}(x)\Big|_a^b + R_m \tag{9}$$

where

$$R_m = (-1)^{m-1} \int_a^b \frac{B_m(\{x\})}{m!} f^m(x)\,dx, \text{ for integers } a \le b, m \ge 1$$

B_k, $k = 1, 2, \ldots$ stands for the Bernoulli numbers and $B_m(\{x\})$ is the Bernoulli polynomial with $\{x\} = x - \lfloor x \rfloor$ the fractional part of x. R_m is the remainder and m is chosen accordingly. Euler-Maclaurin expression is a fundamental result in algebra providing a link between a sum and the corresponding integral. A number of other important results can be derived from this formula. For more details see Graham et al, 1994, p. 469.

The integer number m that can be chosen accordingly in (9) will affect the remainder and consequently the error in this continuous approximation. The Bernoulli numbers are closely related with this choice. Recall the first few values:

$$B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_4 = -\frac{1}{30}, B_6 = \frac{1}{42}, B_8 = -\frac{1}{30}$$
$$\text{and } B_3 = B_5 = B_7 = B_9 = \cdots = 0$$

For $m = 3$, for example, the Euler-Maclaurin equation (9) for a function f studied in the interval $[a, b]$ is

$$\sum_{k=a}^{b-1} f(k) = \int_a^b f(x)dx + \left\{ -\frac{1}{2} f(x) + \frac{1}{12} f^{1}(x) \right\}\Bigg|_a^b + \int_a^b \frac{B_3(\{x\})}{6} f^{(3)}(x)dx \quad (10)$$

The remainder in general must always be considered, as often it diverts, depending on function f (Graham et al, 1994). Function $\rho(h)$ is playing the role of function f in this present application of Euler-Maclaurin. Consequently, the second stage of approximation in Q consisted of a continuous approximation of $\rho(h)$, and is also related to the remainder calculation. Such an approximation is needed because of the integer nature of $\rho(h)$ and the presence of integrals at the right hand side of formula (9).

Because equation (9) involves all the successive-in-order derivatives of f, a continuous extension of $\rho(h)$ through a spline interpolating technique is proposed. If the spline is selected within the broad group of cubic polynomial splines, the third derivative can always be constant and the fourth or higher equal to zero. There are a few alternative splines that preserve the cubic characteristics, with most popular (i) the piecewise cubic shape-preserving hermite interpolation and (ii) the cubic spline, both implemented in Matlab with routines *pchip* and *csaps* respectively. The characteristics that these two alternatives share in common are that they both produce a polynomial which passes though the provided data points, they are piecewise three degree polynomials and they have continuous first derivatives. The differences between them is that the *pchip* produces a function that in order to reserve the shape of the data has discontinuous second derivatives, while *csaps* leads to a smoother function with continuous second derivatives. Moreover, *csaps* allows a smoothing parameter p to be chosen, either manually or by default, which controls the smoothness of the resulting curve in contrast with how close this curve will be to the data points to which it will be fitted.

Let $r(h)$ denote a continuous piecewise cubic interpolation of $\rho(h)$, obtained by either *pchip* or *csaps*. Applying next Euler formula for $m = 4$ to a typical summation of those contained in Q, it can take the form

$$\begin{aligned}\sum_{k=a}^{b-1} r(k) &= \int_a^b r(x)\,dx + \sum_{k=1}^{4} \frac{B_k}{k!} r^{(k-1)}(x)\Big|_a^b + R_4 \\ &= \int_a^b r(x)\,dx + \left\{ -\frac{1}{2} r(x) + \frac{1}{12} r^{(1)}(x) \right\}\Big|_a^b \end{aligned} \tag{11}$$

This last expression is an equality and not an approximation because $R_4 = 0$ since $r(x)$ is a polynomial of third order and therefore $r^{(4)}(x)$. Moreover $B_3 = 0$ and also $r^{(3)}(x)$ is constant, not depending on x, and therefore it adds to zero when evaluated at the two ends of the interval. The only limitation for the exact equivalent and not an approximate expression is $r(x) = \rho(x)$ for all the discrete points between a and b. In other words, r has to be a continuous extension of $\rho(x)$. Under these conditions the error term in Euler-Maclaurin formula is zero and the two functions Q and the corresponding continuous will coincide for all possible points of $(h_0, h_1, h_2, \ldots, h_n)$.

Summarizing, the steps of the proposed methodology in order to determine the optimal sampling allocation and inference about the population parameter are

Step 1. Use (11) for every summation in the second parenthesis of Q in (7) and obtain the continuous equivalent expression given by

$$\begin{aligned} Q^c(h_0,\ldots,h_n) = \frac{N}{n}\Bigg(&\sum_{i=1}^{n-1} r(h_i) + \sum_{i=1}^{n-2} r(h_i + h_{i+1}) + \cdots + r(h_1 + \cdots + h_{n-1}) \Bigg) \\ -\Bigg(&\int_1^{h_0+1} r(x)\,dx + \int_1^{h_0+h_1+1} r(x)\,dx + \cdots + \int_1^{h_0+\cdots+h_{n-1}+1} r(x) \\ &+ \int_1^{h_1+\cdots+h_n+1} r(x)\,dx + \int_1^{h_2+\cdots+h_n+1} r(x)\,dx \\ &+ \cdots + \int_1^{h_n+1} r(x)\,dx + n r(1) \\ &- \frac{1}{2}\Big[r(h_0+1) + r(h_0+h_1+1) \\ &\qquad + \cdots + r(h_0+\cdots+h_{n-1}+1) \\ &\qquad + r(h_n+1) + r(h_n+h_{n-1}+1) \\ &\qquad + \cdots + r(h_n+h_{n-1}+\cdots+h_1+1) \Big] \\ &+ \frac{1}{12}\Big[r^{(1)}(h_0+1) + r^{(1)}(h_0+h_1+1) + \cdots \end{aligned}$$

$$
\begin{aligned}
&+r^{(1)}\left(h_0+\cdots+h_{n-1}+1\right)\\
&+r^{(1)}\left(h_n+1\right)+r^{(1)}\left(h_n+h_{n-1}+1\right)\\
&+\cdots+r^{(1)}\left(h_n+\cdots+h_1+1\right)\Big]\\
-\frac{n}{6}r^{(1)}(1)&\Bigg)
\end{aligned}
$$

Step 2. Minimize Q^c with respect to (h_0, h_1, h_2, …, h_n) and constrains (8), where n is the sample size. Numerical constrained optimization can be used since function Q^c is easily programmed. Function $r(h)$ is calculated by a cubic interpolation on the original discrete function $\rho(h)$.

Step 3. If $\boldsymbol{h}^* = \left(h_0^*, h_1^*, \ldots, h_n^*\right)$ is the vector where the minimum in step 2 is attained and $\left(\bar{h}_0, \bar{h}_1, \ldots, \bar{h}_n\right)$ is its closest integer vector, the optimal sample is the collection of units at positions

$$\bar{t}_1 = \bar{h}_0, \bar{t}_2 = \bar{h}_0 + \bar{h}_1 + 1, \bar{t}_3 = \bar{h}_0 + \bar{h}_1 + \bar{h}_2 + 1, \ldots, \bar{t}_n = \bar{h}_0 + \bar{h}_1 + \ldots + \bar{h}_{n-1} + 1$$

Step 4. The *mse* of the population mean estimate calculated on the optimal sample $s = \left(Y_{\bar{t}_1}, Y_{\bar{t}_2}, \ldots, Y_{\bar{t}_n}\right)$ is derived from (5) by single substitution.

For a small numerical example, a set of simulated N observations from a Moving Average (MA) process of order 2, with parameters −0.4 and 0.5 are assumed to represent the population. The autocorrelation function of the assumed MA model within the population range is listed in the first part of Table 1. The resulting set of values forming the population is U = (−0.52, −1.33, 0.19, 1.70, −1.37, −1.35, −0.22, −0.16) and let the aim of the experiment to be the selection of a sample of size n = 3 that minimizes the *mse*. The set of all possible samples P_n contains 56 samples, and in order to obtain the optimal s = (h_0, h_1, h_2, h_3), the quantity Q needs to be minimized with respect to (h_0, h_1, h_2, h_3). Function Q given by (7) for this example is

$$Q(h_0,h_1,h_2,h_3)=\frac{N}{n}\left(\rho(h_1)+\rho(h_2)\rho(h_1+h_2)\right)$$
$$-\left(\sum_{i=1}^{h_0}\rho(i)+\sum_{i=1}^{h_0+h_1}\rho(i)+\sum_{i=1}^{h_0+h_1+h_2}\rho(i)+\sum_{i=1}^{h_1+h_2+h_3}\rho(i)+\sum_{i=1}^{h_2+h_3}\rho(i)+\sum_{i=1}^{h_3}\rho(i)\right)$$
$$+\int_1^{h_1+\ldots+h_n+1}r(x)\,dx+\int_1^{h_2+\ldots+h_n+1}r(x)\,dx+\cdots+\int_1^{h_n+1}r(x)\,dx$$

and the corresponding Q^c is

$$Q^c(h_0,h_1,h_2,h_3)=\frac{8}{3}\left(r(h_1)+r(h_2)+r(h_1+h_2)\right)$$
$$-\left(\int_1^{h_0+1}r(x)\,dx+\int_1^{h_0+h_1+1}r(x)\,dx+\int_1^{h_0+h_2+h_3+1}r(x)\,dx\right.$$
$$+\int_1^{h_1+h_2+h_3+1}r(x)\,dx+\int_1^{h_2+h_3+1}r(x)\,dx+\int_1^{h_3+1}r(x)\,dx$$
$$+3r(1)-\frac{1}{2}\Big[r(h_0+1)+r(h_0+h_1+1)+r(h_0+h_1+h_2+1)$$
$$+r(h_3+1)+r(h_3+h_2+1)+r(h_3+h_2+h_1+1)\Big]$$
$$+\frac{1}{12}\Big[r^{(1)}(h_0+1)+r^{(1)}(h_0+h_1+1)+r^{(1)}(h_0+h_1+h_2+1)$$
$$+r^{(1)}(h_3+1)+r^{(1)}(h_3+h_2+1)+\cdots$$
$$\left.+(h_3+h_2+h_1+1)\Big]-\frac{1}{2}r^{(1)}(1)\right)$$

Note the complexity of Q^c does not depend on N. The number of the unknowns and consequently the efficiency of the numerical minimization depends only on n. Sizes N and n have been chosen small in order to proceed in an exhaustive enumeration of all samples in P_n and confirm both the approximation of Q and its minimum. Minimizing $Q^c(h_0, h_1, h_2, h_3)$ yields $(h_0, h_1, h_2) = (0, 1.80, 1.91)$ and $h_3 = N - 1 - (h_0, h_1, h_2)$. The closest discrete solution is $(h_0, h_1, h_2) = (0, 2, 2)$ and this corresponds to the sample $s = (Y_1, Y_3, Y_5)$.

Table 1a. Numerical example for a population with $N = 8$ generated from MA(2)

i	Y_i	lag h	$\rho(h)$	$r(h)$
1	-0.52	0	1.00	1.00
2	-1.33	1	-0.06	-0.06
3	0.19	2	-0.66	-0.66
4	1.70	3	0.02	0.02
5	-1.37	4	0.21	0.21
6	-1.35	5	0.02	0.02
7	-0.22	6	-0.03	-0.03
8	-0.16	7	0.00	0.00

Table 1b. Numerical example for a population with $N = 8$ generated from MA(2)

sample	Q	Hermite Q^c	Spline Q^c	sample	Q	Hermite Q^c	Spline Q^c
(0,1,1)	2.4533	2.4308	2.4531	(1,2,3)	3.3185	3.2853	3.3069
(0,1,2)	2.8326	2.8135	2.8284	(1,2,4)	3.8756	3.8564	3.8714
(0,1,3)	6.9570	6.9379	6.9528	(1,3,1)	8.1837	8.1634	8.1842
(0,1,4)	6.9570	6.9345	6.9568	(1,3,2)	3.3185	3.2853	3.3069
(0,1,5)	4.6993	4.6638	4.6870	(1,3,3)	6.1985	6.1794	6.1943
(0,1,6)	4.4978	4.4764	4.4929	(1,4,1)	7.0637	7.0271	7.0561
(0,2,1)	3.9526	3.9464	3.9605	(1,4,2)	3.8756	3.8530	3.8754
(0,2,2)	2.0089	2.0027	2.0168	(1,5,1)	4.6993	4.6638	4.6870
(0,2,3)	4.3319	4.3223	4.3437	(2,1,1)	4.8000	4.7960	4.8085
(0,2,4)	3.8756	3.8530	3.8754	(2,1,2)	5.1793	5.1719	5.1918
(0,2,5)	3.0104	3.0020	3.0176	(2,1,3)	8.1837	8.1634	8.1842
(0,3,1)	8.0770	8.0742	8.0809	(2,1,4)	8.0770	8.0709	8.0849
(0,3,2)	4.3319	4.3257	4.3397	(2,2,1)	5.1793	5.1719	5.1918
(0,3,3)	6.1985	6.1794	6.1943	(2,2,2)	2.1156	2.0953	2.1160
(0,3,4)	7.1348	7.1298	7.1380	(2,2,3)	4.3319	4.3257	4.3397
(0,4,1)	8.0770	8.0709	8.0849	(2,3,1)	8.1837	8.1600	8.1182
(0,4,2)	3.8756	3.8564	3.8714	(2,3,2)	4.3319	4.3223	4.3437
(0,4,3)	7.1348	7.1298	7.1380	(2,4,1)	6.9570	6.9345	6.9568
(0,5,1)	5.8193	5.7967	5.8191	(3,1,1)	4.8000	4.7960	4.8085
(0,5,2)	3.0104	3.0020	3.0176	(3,1,2)	4.0593	4.0424	4.0557
(0,6,1)	4.4978	4.4764	4.4929	(3,1,3)	8.0770	8.0742	8.0809
(1,1,1)	3.6800	3.6597	3.6805	(3,2,1)	4.0593	4.0390	4.0597
(1,1,2)	4.0593	4.0390	4.0597	(3,2,2)	2.0089	2.0027	2.0168
(1,1,3)	8.1837	8.1600	8.1882	(3,3,1)	6.9570	6.9379	6.9528
(1,1,4)	7.0637	7.0271	7.0561	(4,1,1)	3.6800	3.6597	3.6805
(1,1,5)	5.8193	5.7967	5.8191	(4,1,2)	3.9526	3.9464	3.9605
(1,2,1)	4.0593	4.0424	4.0557	(4,2,1)	2.8326	2.8135	2.8284
(1,2,2)	2.1156	2.0953	2.1160	(5,1,1)	2.4533	2.4308	2.4531

Table 1b provides a comparison of the arithmetic values of Q and Q^c for every sample in P_n. The 56 samples of P_n consist of all possible vectors

(h_0, h_1, h_2) that fulfill constrains (8); $h_3 = 7 - h_0 - h_1 - h_2$ and is not given. Two versions of Q^c are presented for the example, the first one using piecewise cubic hermite interpolation to construct r, noted as Hermite Q^c, and the second using smooth spline, noted as Spline Q^c. The differences compared to the true function Q are in the second decimal place, while the range of values is between 2.0089 and 8.1837. The differences among the function values are due to the use of numerical instead of analytical integration. It is also verified that the minimum *mse* value is achieved for the same sample $s = (0, 2, 2)$ for all methods, and agrees with the one derived from the numerical minimization. Since the optimal sample is found, the exact *mse* can be calculated from (5) and is 6.0267.

The smoothness characteristic of the spline $r(\cdot)$ improves the performance of the numerical integration and produces numerical values closer to the true ones. The smoothing parameter for the *csaps* routine, which has been used for this example, was chosen as 1. This means that a priority to the exact matching of the spline values with the initial was given, rather than the smoothness.

Experiments and Applications

Experiments with Simulated Data

Three numerical examples follow, with simulated data generated from three different ARMA models to represent the population values under study. The justification for the ARMA model is that its autocorrelation function is general enough to cover a wide range of types for the serial correlation, depending on the specific values of their order and parameters. The aim of the experiment is to obtain the optimal sampling allocation following the proposed methodology, and compare its efficiency with other competitive sampling designs, chosen for either their broad use, because they are standard sampling designs, or because the literature suggests their application is appropriate to the case of correlated populations. More specifically, the sampling designs chosen are simple random sampling (srs), systematic sampling (sy), an optimal design for correlated populations with positive correlation, and Designs I and II proposed by Chao (2004) for correlated populations.

A range of values for the sample size is taken in every population case for a more complete view of the sampling design performance. The corresponding mean square errors of the estimates are calculated for all examined sampling designs by simulation and assuming normality. More specifically, if K realizations from each population model are generated, and $\hat{\theta}_j^d$ is the estimate for the

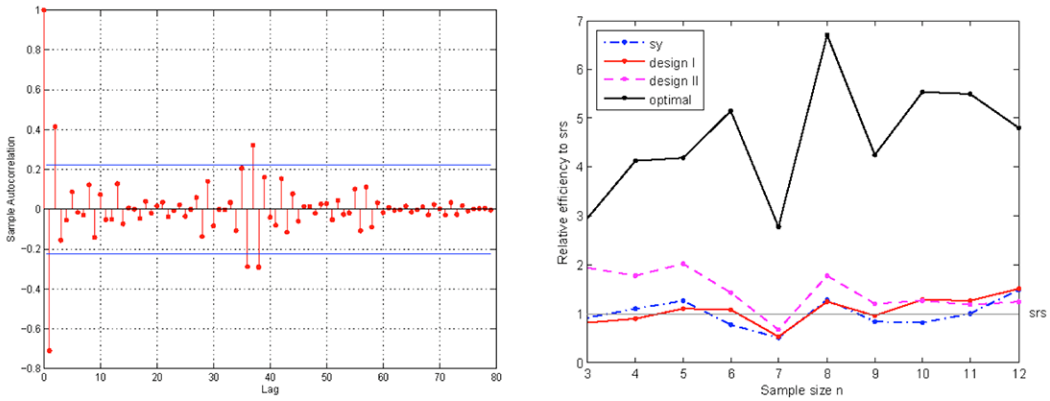
population total at the j^{th} realization according to the sampling design d, the *mse* for the estimate will be calculated by

$$E_d\left(\theta-\hat{\theta}\right)^2=\frac{1}{K}\sum_{j=1}^{K}\left(\theta_j-\hat{\theta}_j^d\right)^2$$

The number of iterations for each experiment is 15,000, while the common variance σ^2 is assumed unity. The optimal allocation of samples is derived by implementing the proposed methodology as previously described in Steps 1 to 5. For the numerical optimization, twenty different starting values have been used for each application and the smooth spline with $p = 1$ has been used as the interpolation function of ρ. The performance of the examined sampling designs is evaluated by the relative efficiency to the srs, defined as the ratio of the *mse* obtained with a sampling design to that obtained with the srs. Values of relative efficiency greater than one indicate efficiency of the examined design.

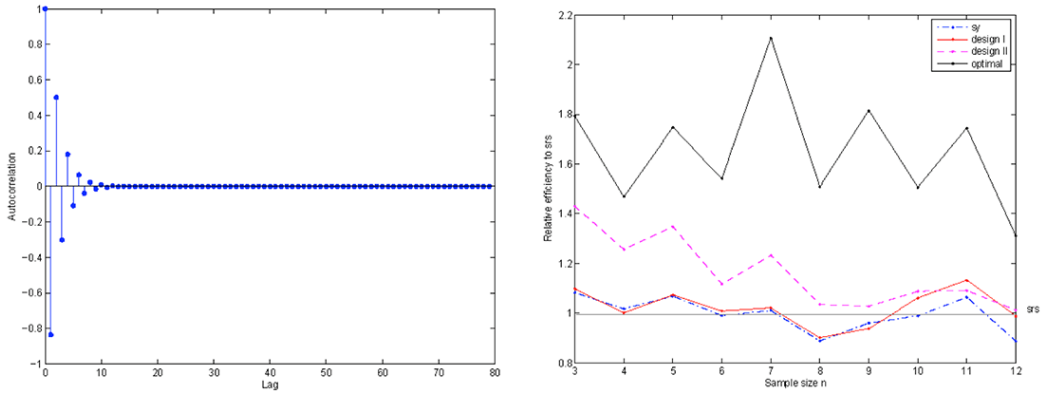
Model 1. The population measurements are generated from an ARMA(1,1) model with autocorrelation function plotted in Figure 1a. The degree of correlation is moderate for the assumed population occasion with the sign to alternate because of the negative sign of the parameter φ, the autoregressive part of the process. For population size $N = 80$ and sample size that ranges from $n = 3$ to $n = 12$ the calculated efficiencies of the examined designs compared to srs are plotted in Figure 1b. For better illustration the reciprocal of the design effect is plotted in Figure 1b. Systematic, Design I and Design II are comparable with srs with respect to their *mse*, while the optimal allocation derived by the proposed methodology is clearly more efficient.

As a specific example, for $n = 12$ the optimal sample determined from the solution of the numerical minimization problem is $s = [1\ 2\ 3\ 59\ 60\ 61\ 62\ 63\ 64\ 78\ 79\ 80]$. Sample s has the sampling units separated into three groups, two groups located at the two ends of the population and one in the middle. Moreover neighbor units of the population are selected within the groups. Its mean square error by using simulation is 113.79, and its exact value from expression (5) is 113.87. Design II, the second best with respect to the mean square error in this example, has *mse* of 441.87.



(a) (b)

Figure 1. Relative efficiencies using the empirical autocorrelation function on Model 1



(a) (b)

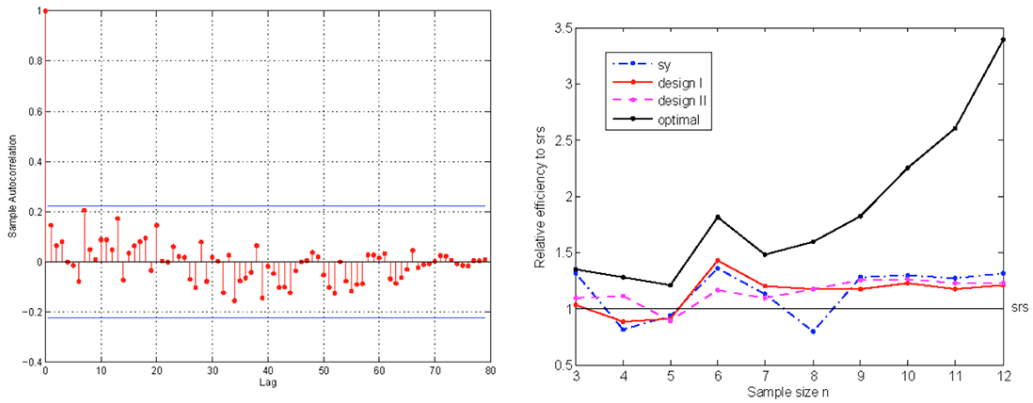
Figure 2. Relative efficiencies using the theoretical autocorrelation function on Model 1

The empirical autocorrelation function calculated from the population of size $N = 80$ has been used for implementing the methodology in this first example. If not the empirical but the exact autocorrelation function according to the assumed ARMA(1,1) model is used, the two resulting plots (corresponding to those in Figure 1) are presented in Figure 2. The sampling allocation according to the proposed method remains efficient. The assumed theoretical ARMA model has a negative φ parameter as it can be seen from Figure 2a and the sign of the

ACF alternates. For such cases, the systematic sampling is far from the optimal, and plots in Figures 1b and 2b verify this result from the literature.

Model 2. $N = 80$ is assumed for this example. The population vector, U, is generated from an ARMA(2,1) model with autocorrelation function plotted in Figure 3a. The serial correlation is not strong in this model, but the sign alternates and therefore it cannot be characterized as a positive, convex function. Following the same steps as in Model 1, the corresponding plot that presents the relative efficiencies of the sampling schemes under study with srs are presented in Figure 3b.

The optimal sample derived by the proposed methodology implemented here is the most efficient sample along all examined sample sizes, as shown in Figure 3b. The three other samples proposed from the remaining techniques exhibit similar performance, faintly better if not comparable with the srs. The comparable to srs performance of sy is explained from the fact that the correlation between observations is low. It is known in sampling literature that sy and srs are equivalent with respect to accuracy when the population measurements do not present a trend or correlation (Cochran, 1977). Figure 3b demonstrates that the efficiency of the optimal sample is increased as the sample size increases.

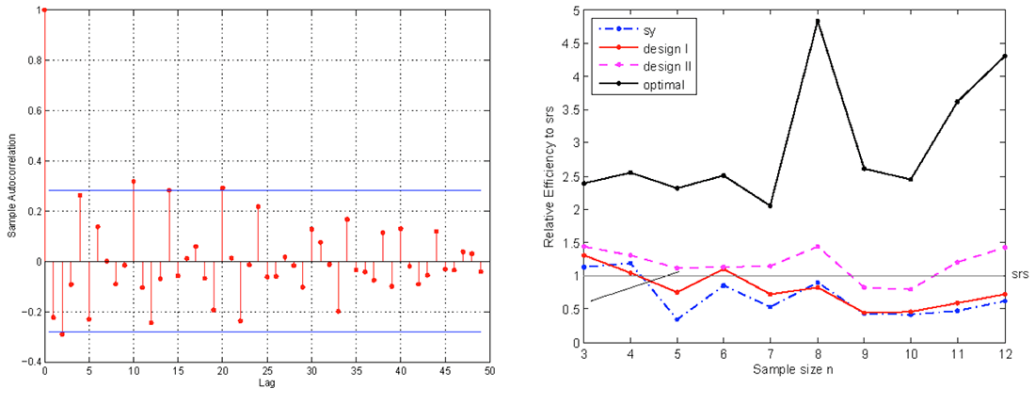


(a) (b)

Figure 3. Relative efficiencies using the autocorrelation function on Model 2

Model 3. Assume $N = 50$ and population values generated from an ARMA(2,4) model with autocorrelation function plotted in Figure 4a. The

autocorrelation in this model is not strong and alternates in sign. Again for sample size ranging between $n = 3$ and $n = 12$, the relative efficiencies are plotted in Figure 4b. The optimal sample as derived by the proposed methodology is clearly the sample with the minimum *mse*. Its relative efficiency is between 0.207 and 0.48, indicating a significant improvement in accuracy with respect to srs sampling scheme. Among the remaining competitive designs, Design II compares better than the srs, although not consistently, followed by Design I and systematic sampling, which produce higher than the srs *mse* and are not appropriate for this population case.



(a) (b)

Figure 4. Relative efficiencies using the autocorrelation function on Model 3

An application in Statistical Process Control

Consider an application of the proposed methodology in Statistical Process Control (SPC) based on a real data set. The data include 204 consecutive measurements of electrical resistance of insulation in megaohms and was first presented in Shewhart (1931, p. 20). The data set often serves as a typical example in SPC, where the existing autocorrelation can lead to incorrect conclusions about a process if it has not been detected or handled properly. The implementation of sampling in SPC happens during the construction of the statistical charts, which aim to provide some warning limits for the production line and detect a deviation in mean or variance of the process. Many forms of statistical charts are available, but the common basis for any chart is a sample

taken at an initial stage from the production line. Shewhart's control chart is one of the best known statistical control charts, and its basic components are presented during this application. Any statistical control chart can be evaluated by calculating the expected probability of false positive or negative alarms.

Shewhart's control chart of the $\overline{X}$, the mean of a sample taken from the process, was originally constructed for the electrical measurements and presented by taking successive groups of four. The resulting 51 subsamples were used to estimate the mean and the variance of the population towards the construction of the upper and lower control limits. The control limits provide a reference interval for a mean of a sample of four selected from the process if this is in control. The two limits are in mathematical terms

$$\left[\overline{x} - 3*\frac{\hat{\sigma}}{\sqrt{k}}, \overline{x} + 3*\frac{\hat{\sigma}}{\sqrt{k}}\right]$$

where $\overline{x}$ is the mean of the means of the 51 subsamples and is used as an estimate of the population mean, $\hat{\sigma}$ is the estimate of σ, the square root of the process units variance, and k is the size of the sub-samples. For $k = 4$ and the total of 51 subsamples in the data, the resulting control limits for the mean of the process are plotted in Figure 5, solid line. The means of the 51 samples are plotted together in the same Figure, and a large percentage of those means are outside the limits an indication that the process is not in control. The process is however in statistical control, as subsequent analyses of the same data set concluded (see for example, Alwan and Roberts, 1995). The variation that the data exhibit can be explained from the present autocorrelation not taken into account in the first application, and is not due to a special cause.

Yang and Hancock (1990) introduced the autocorrelation into the calculation of the control limits for Shewhart's control chart. The new control limits suggested by their methodology are given by

$$\rho = \frac{\sum_{i \neq j} r_{ij}}{k(k-1)}$$

where r_{ij} are the i, j elements of matrix R if assumed that the variance covariance matrix $\boldsymbol{V}$ of a sample can be written as $\boldsymbol{V} = \sigma^2 R$. Implementing this approach for the electrical measurements and using all 51 samples of four, the

resulting control limits are wider, as expected, and include all 51 sub-sample means, as can be seen in Figure 5 (dotted line).

Alternatively, the control limits can be calculated by implementing the methodology proposed in Steps 1–4. The implementation is possible in both stages of sampling. For the first stage of sub-samples of four, formula (5) can be used to estimate the variance of the sample mean. The resulting estimate is more accurate than the average correlation ρ because the exact matrix $\boldsymbol{V}$ according to the model, and not an average ρ, is used. For the second stage of the 51 sub-samples, either complete enumeration or sampling is possible. Sampling is more realistic in practice and can also be applied to continuous processes. Both scenarios are presented here using the proposed methodology to choose the sample in the case of sampling at the level of sub-samples. Moreover, in a real situation application, a sample instead of successive measurements could also be the case for the first stage of SPC.

A first-order autoregressive model has been fitted to the data with parameter φ equal to 0.549 (Alwan & Roberts, 1995), and this is the model used for the implementation. When all sub-samples are taken into account and the variance of the mean with sub-sample is calculated by (5), the resulting control limits are plotted in Figure 5 (dashed line). The use of the exact form of the model that describes the population units allows control limits that are wider than in the first analysis, but not as much as according to Yang and Hancock methodology. Note that too wide control limits lead to an increase of the probability of falsity in control conclusions.

If a sample of seven sub-samples is selected according to the proposed methodology, and the estimates of the mean as well as their standard errors in both stages are calculated from expression (4) and (5) respectively, the resulting control limits are plotted in Figure 5 (dash-dot line), and compare closely to the ones derived from the complete population of $N = 51$ sub-samples.

Therefore, identifying the model correctly and fully incorporating this information in the selection of samples procedure and the statistical inference allows us to accurately construct control limits using only 28 measurements instead of 204.

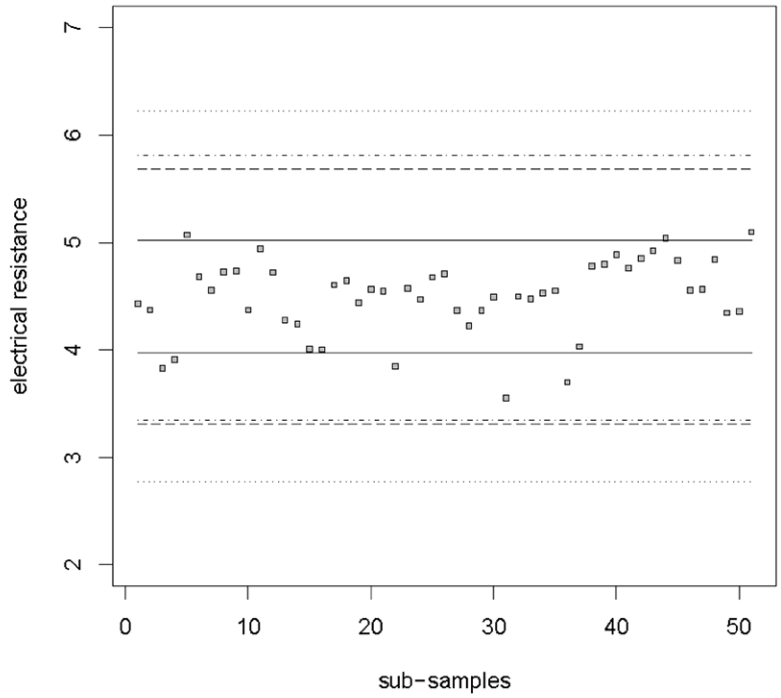


Figure 5. Control limits for electrical measurements.

The optimal sample of size seven for this application was found not to be equally spaced. An equally spaced scheme in SPC, also called fixed distance sampling, corresponds to a systematic design and is often the choice for selecting the sub-samples during the second stage for SPC applications, especially in cases where a positive correlation is detected. However, it has been verified that variable distance outperforms fixed distance sampling. The comparison has been conducted with simulation studies that calculate the average time to signal (ATS), a measure of efficiency of control charts. The advantage of variable distance sampling depends on the degree and type of correlation (Prybutok, et al., 2007).

Within the same framework, other models, more general than the AR, can also be treated with the proposed methodology.

Conclusion

A continuous approach was proposed for an intractable otherwise discrete optimization problem with primary application in sampling. During the process of

sampling from correlated populations, the specific type of the autocorrelation function $\rho(h)$ among the populations' units affects both the choice of the sample and the inference about the population parameters. If $\rho(h)$ has certain properties, such as constant, positive, decreasing, convex, etc., it is possible to derive conclusions about the optimal sampling designs even if $\rho(h)$ is not known in its exact form. In cases of a more general type of correlation (for example, a realization of a time series process), characterizing the optimal sampling designs or the class of the optimal samples is not possible and the results depend closely on the specific type of $\rho(h)$. A feasible and accurate way of deriving a sample that belongs to the class of optimal samples in such cases is proposed here. The estimate with its *mse* is also provided. The proposed technique uses continuous approximation of a finite sum from an integral. A continuous interpolation function $r(h)$ based on $\rho(h)$ is an important component for its implementation, and when $r(h)$ holds certain properties it is shown that the proposed approach is not an approximation but exact.

The method can be used in any case of correlated population, or not. It is fast, easily programmed and implemented, and computationally efficient. The dimensionality coincides with the sample size and therefore the computational efficiency remains unaffected from the population size. As a general approach, it can find applications in other than sampling context and facilitate the solution of a mathematical problem that depends on a function with a discrete nature.

The benefit for estimation is significant. Ignoring or incorrectly specifying the existing correlation within a population set can lead to misleading results, especially regarding the accuracy of the derived parameter estimate. The proposed methodology suggests a more sophisticated and informative sampling procedure, specialized for the population under study. This specialization has been incorporated into the *mse* calculation of the assumed estimator and the minimum *mse* is the criterion for the sampling procedure derivation. Therefore, the suggested sample is optimal with respect to the accuracy of the resulting estimate, and the improvement in *mse* is significant when compared to other known and widely used sampling schemes. Moreover, the simulation experiments suggest that the inclusion of the population model towards the correct calculation of the *mse* is necessary, and has a considerable impact on efficiency even if a small degree correlation occurs. Finally, the actual arithmetic value of both the estimate and its exact *mse* implemented for the optimal sampling allocation are provided.

The extension of the proposed methodology to continuous stationary processes is straightforward. The assumption of other than the least squared estimator is also possible. The least squared estimator for the population

parameter has been considered here because of its frequent use in practice, due to its simplicity and ease of implementation. Assumption of the best linear unbiased or the best unbiased estimators are some possible extensions, along with the assumptions of model (2). The constant mean parameter μ may be assumed dependent on population unit *i*. Under this model the least squared estimator (4) is not unbiased for the population total. The bias depends on the sample *s*, but not on the type of autocorrelation. The new expression of the estimator's *mse* needs to be minimized following a procedure similar to that proposed here.

References

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. *Communications in Statistics – Theory and Methods*, *21*(4), 1025−1049. doi: 10.1080/03610929208830829

Alwan, L. C. & Roberts, H. V. (1995). The problem of misplaced control limits. *Applied Statistics*, *44*(3), 269−278.

Alwan, L. C. & Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business and Economic Statistics*, *6*(1), 87−95. doi: 10.1080/07350015.1988.10509640

Apley, D. W. & Lee H. C. (2003). Design of exponentially weighted moving average control charts for autocorrelated processes With model uncertainty. *Technometrics*, *45*(3), 187−198. doi: 10.1198/004017003000000014

Arnab, R. (1992). Estimation of a finite population mean under superpopulation models. *Communications in Statistics – Theory and Methods*, *21*(6), 1717−1724. doi: 10.1080/03610929208830874

Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, *12*(1), 53−65. doi: 10.2307/3314724

Bellhouse, D. R. & Rao, J. N. K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, *62*(3), 694−697. doi: 10.1093/biomet/62.3.694

Bjørnstad, O. N., Ims, R. A., & Lambin, X. (1999). Spatial population dynamics: Analysing patterns and processes of population synchrony. *Trends in Ecology and Evolution*, *14*(11), 427−431. 10.1016/S0169-5347(99)01677-8

Bolfarine, H. & Zacks, S. (1992). *Prediction theory for finite population*. NY: Springer Varlag.

Blight, B. J. N. (1973). Sampling from an autocorrelated finite population. *Biometrika*, *60*(2), 375−385. doi: 10.1093/biomet/60.2.375

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). *Foundations of inference in survey sampling*. New York: Wiley.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, *17*(2), 164−177. doi: 10.1214/aoms/1177730978

Cochran, W. G. (1953). *Sampling techniques*, 1st Ed. NY: Wiley & Sons

Cochran, W. G. (1977). *Sampling techniques*, 3rd Ed. NY: Wiley & Sons.

Cox, D. R. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, *91*(3), 729−737. doi: 10.1093/biomet/91.3.729

Chao, C-T. (2004). Selection of sampling units under a correlated population based on the eigensystem of the population covariance matrix. *Environmetrics*, *15*(8), 757−775. doi: 10.1002/env.655

Cressie, N. A. (1993). *Statistics for spatial data,* Revised version. NY: Wiley.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *17*(2), 269−278.

Graham, R. L., Knuth, D. E. & Patashnik, O. (1994) *Concrete mathematics*, 2nd Ed. Boston, MA: Addison-Wesley.

Graubard, B. I. & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*(1), 73−96.

Harris, T. J. & Ross, W. H. (1991). Statistical process control procedures for correlated observations, *Canadian Journal of Chemical Engineering*, *69*(1), 48−57. doi: 10.1002/cjce.5450690106

Hjort, N. L. & Varin, C. (2008). ML, PL and QL for Markov chain models. *Scandinavian Journal of Statistics*, *35*(1), 64−82. doi: 10.1111/j.1467-9469.2007.00559.x

Karakostas, K. X. (1984). *Optimum systematic sampling for autocorrelated superpopulations*. (Unpublished thesis). Imperial College, University of London.

Lande, R. (1991). Isolation by distance in a quantitative trait. *Genetics*, *128*(2), 443−453.

Lu, C. W. & Reynolds, M. R. (1999). Control charts for monitoring the mean and variance of autocorrelated processes. *Journal of Quality Technology*, *31*(3), 259−274.

Lu, C. W. & Reynolds, M. R. (2001). CUSUM control charts with variable sample sizes and sampling internals. *Journal of Quality Technology*, *33*(1), 66−81.

Madow, L. H. & Madow, W. G. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, *15*(1), 1−24. doi: 10.1214/aoms/1177731312

Mastrangelo, C. M. & Montgomery, D. C. (1995). SPC with correlated observations for the chemical and process industries. *Quality and Reliability Engineering International*, *11*(2), 79−89. doi: 10.1002/qre.4680110203

Montgomery, D. C. & Mastrangelo, C. M. (1991) Some statistical process control chart methods for autocorrelated data. *Journal of Quality Technology*, *23*(3), 179-193.

Mukerjee, R. & Sengupta, S. (1989). Optimal estimation of finite population total under a general correlated model. *Biometrika*, *76*(4), 789−794. doi: 10.1093/biomet/76.4.789

Mukerjee, R. & Sengupta, S. (1990). Optimal estimation of finite mean in the presence of linear trend. *Biometrika*, *77*(3), 625−630. doi: 10.1093/biomet/77.3.625

Nayak, T. K. (2003). Finding optimal estimators in survey sampling using unbiased estimators of zero. *Journal of Statistical Planning and Inference*, *114*(1-2), 21−30. doi: 10.1016/S0378-3758(02)00460-3

Neuhaus, J. M. & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638−645. doi: 10.2307/3109770

Papageorgiou, I., & Karakostas, K. X. (1998). On optimal sampling designs for autocorrelated finite populations. *Biometrika*, *85*(2), 482−486. doi: 10.1093/biomet/85.2.482

Prybutok, V. R., Clayton H. R., & Harvey, M. M. (1997). Comparison of fixed versus variable sampling interval Shewhart $\bar{X}$ control charts in the presence of positively autocorrelated data. *Communications in Statistics – Simulation and Computation*, *26*(1), 83−106. doi: 10.1080/03610919708813369

Ramakrishnan, M. K. (1975). Choice of an optimum sampling strategy–1. *Annals of Statistics*, *3*(3), 669−679. doi: 10.1214/aos/1176343129

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377−387. doi: 10.1093/biomet/57.2.377

Shewhart, W. A. (1931). *The economic control of manufactured products*. NY: Van Nostrand.

Tam, S. M. (1984). Optimal estimation in survey sampling under a regression superpopulation model. *Biometrika*, *71*(3), 645−647. doi: 10.1093/biomet/71.3.645

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*(1), 1−28. doi: 10.1007/s10182-008-0060-7

Wardell, D. G., Moskowitz, H. & Plante, R. D. (1992) Control charts in presence of data correlation. *Management Science*, *38*(8), 1084−1105. doi: 10.1287/mnsc.38.8.1084

Watson, G. S. (1972). Trent-surface analysis and spatial correlation. *Geological Society of America Special Papers*, 146, 39−46. doi: 10.1130/SPE146-p39

Yang, K., & Hancock, W. M. (1990). Statistical quality control for correlated samples. *International Journal of Production Research*, *28*(3), 595−608. doi: 10.1080/00207549008942738

Yates, F. (1960). *Sampling methods for censuses and surveys*, 3rd Ed. London, England: Griffin.