

Replicated Environmental Sampling Designs and Monte Carlo Integration Methods: Two Sides of the Same Coin

*Disegni Campionari Replicati di Tipo Ambientale e
Strategie di Integrazione con il Metodo di Monte Carlo:
Due Facce della Stessa Medaglia*

Lucio Barabesi

Dipartimento di Metodi Quantitativi, Università di Siena,
P.zza S.Francesco 17, 53100 Siena, Italy, barabesi@unisi.it

Riassunto: Nei protocolli campionari con applicazioni ambientali basati sul cosiddetto approccio da "popolazione continua" vengono considerate varie repliche di un opportuno disegno al fine di stimare il parametro obiettivo (usualmente il totale di una certa variabile nell'area di studio). Dal momento che questo parametro può essere espresso sotto forma di integrale, è immediato riconoscere una dualità fra disegni campionari replicati e strategie di integrazione con il metodo di Monte Carlo. Questa connessione permette di analizzare da un punto di vista generale i disegni campionari replicati nell'approccio da "popolazione continua" e di introdurre tecniche campionarie molto efficaci.

Key words: Replicated environmental designs; Horvitz-Thompson estimator; linear homogeneous estimator; Monte Carlo integration; importance sampling; control variate method; regression estimator.

1. Introduction

Ecological studies involve designs which may be dealt with on the basis of two paradigms (Williams and Eriksson, 2002, Thompson, 1992, p.4). The first approach - the so-called *tessellated plane* paradigm - relies on the assumption that the study region is partitioned into a finite number of sub-regions which constitute an area sampling frame. Thus, the inference under this paradigm may be drawn by adopting the usual sample theory setting. The other approach - the so-called *continuous plane* paradigm - implies that the design is carried out by selecting a point on a *continuum*. In ecological applications, the *continuous plane* paradigm is very useful (Williams and Eriksson, 2002). When large study regions are considered, the two approaches may be coupled in two-stage sampling: in the first stage some sub-regions are sampled - *i.e.* the *tessellated plane* approach is adopted - while in the second stage these sub-regions are in turn sampled using a design under the *continuous plane* approach (Pisani, 2002, Fattorini, Marcheselli and Pisani, 2004, Fattorini and Pisani, 2004).

When the *continuous plane* paradigm is considered, practical environmental designs may be divided into two families (see *e.g.* Barabesi and Fattorini, 1998, Barabesi and

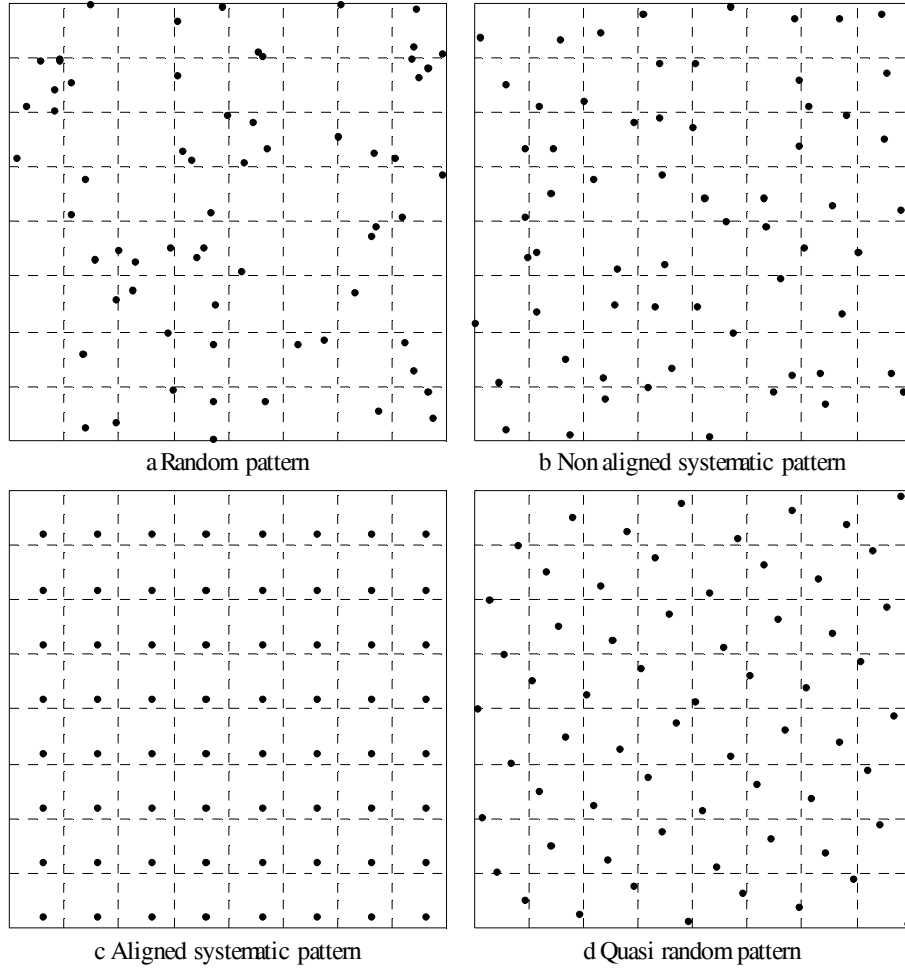
Pisani, 2002): designs which are implemented by selecting a point on the baseline (the projection of the study region onto a line of arbitrary direction) and designs which are carried out by selecting a point on the study region. On one hand, the first class (say Family 1) encompasses Thompson's (1992, p.225) line-intercept sampling, strip sampling (Fattorini and Pisani, 1999, p.106) and Thompson's (1992, p.176) line-transect sampling, among others. On the other hand, the second class (say Family 2) comprises plot sampling, Bitterlich sampling (Schreuder *et al.*, 1993, p.117), Kaiser's (1983) line-intercept sampling, point sampling (Brus and De Gruijter, 1997), line transect sampling (Barabesi, 2000, Barabesi *et al.*, 2002) and point-to-plant sampling (Barabesi, 2001a, Barabesi and Marcheselli, 2002), among others.

Under the *continuous plane* approach, environmental sampling protocols are commonly based on the replication of a suitable design (Thompson, 1992, p.188, Schreuder *et al.*, 1993, p.114). In this case, the aim of a quantitative ecological study is the estimation of the target parameter - usually the population total of the variable under study. Accordingly, if n replications of the selected design are performed, the focus is the choice of an appropriate strategy for the placement of the n points on the reference set - *i.e.* the baseline (Family 1 case) or the study region (Family 2 case). The U.S. Environmental Protection Agency actually suggests four strategies for the placement of the n points on the reference set (U.S. EPA QA/G-5S Guidance, 2002, p.36, p.39, p.63, Gilbert, 1987, p.21). The first strategy involves generating n independent random points on the reference set - a method commonly adopted but likely to produce undetected parts in the reference set (Fig.1a). Hence, field researchers prefer a systematic search over the study region. Therefore, the second strategy - so-called non-aligned systematic sampling - involves partitioning the reference set into n equal subsets and generating n independent random points in these subsets (Fig.1b). The third strategy - so-called aligned systematic sampling of points - is based on the same partition considered for the second strategy, but solely a random point is generated in the first subset and this point is systematically replicated in the other subsets (Fig.1c). The fourth strategy is obtained by generating a quasi-random sequence (or a randomized quasi-random sequence) of points on the reference set (Fig.1d). A quasi-random sequence aims to cover the space as uniformly as possible (L'Ecuyer and Lemieux, 2002).

If the target parameter is the population total of the variable under study, Barabesi (2003b) has shown that this parameter may be expressed as a suitable integral of the Horvitz-Thompson estimator function. Accordingly, Barabesi (2003b) has emphasized the connection between Monte Carlo integration and Horvitz-Thompson estimation in replicated designs under the *continuous plane* approach. Thus, the first strategy leads to the basic MC integration method (Rubinstein, 1981, p.118), the second strategy to the so-called Modified Monte Carlo (MMC) integration scheme (Haber, 1966), the third strategy to the Random Grid (RG) quadrature method (Cruz-Orive, 1993) and the fourth strategy to the Quasi-Monte Carlo (QMC) - or to the Randomized Quasi-Monte Carlo (RQMC) - quadrature method (L'Ecuyer and Lemieux, 2002). A related setting is considered in Valentine *et al.* (2001) which underlines the link between Monte Carlo methods and the sampling methods used in forestry.

In this paper, the connection between replicated environmental sampling designs under the *continuous plane* approach and Monte Carlo integration methods is exploited in a general perspective and some new procedures are suggested.

Figure 1. Four patterns of 64 points in the unit square generated according to the considered strategies.



2. The target parameter representation

When a quantitative environmental study is carried out, the target parameter is usually represented by the total of the variable in a delineated region. The variable under study may be defined over a finite set of points in the region (the *discrete variable* case) or over the whole region (the *continuous variable* case).

In the *discrete variable* case, let us consider a set of N units scattered at fixed locations in the region. In addition, let $y_1, \dots, y_N$ be the values of the variable corresponding to each unit. In this case, the target parameter is given by

$$T_y = \sum_{l=1}^N y_l .$$

First, let us consider a Family 1 design. Without loss of generality, let us suppose that the baseline corresponds to the set $(0, 1)$. Thus, let u be the location of the selected point on the baseline. In this setting, the inclusion set $\mathcal{P}_l$ for the l -th unit is a suitable interval contained in the baseline (see *e.g.* Barabesi, 2003a) and the l -th unit is selected - and y_l is measured - if $u \in \mathcal{P}_l$. As an example, let us consider a population of plants and

let y_l be the biomass of the l -th plant, in such a way that the target parameter is the total biomass in the forest. If line intercept sampling is adopted, the inclusion sets are the projections of the plant crowns on the baseline (Barabesi, 2003a). Indeed, a plant is selected if the corresponding crown is intercepted by a line perpendicular to the baseline at location u .

In order to obtain a suitable representation for T_y , it is worth noting that, if solely the l -th unit is considered, the *intensity* of the variable of interest over the l -th inclusion set is given by y_l/π_l , where $\pi_l = \int_0^1 \mathbf{I}_{\mathcal{P}_l}(u) du$ is the length of $\mathcal{P}_l$. Hence, the *intensity* of the variable at location u is given by

$$y(u) = \sum_{l=1}^N \frac{y_l}{\pi_l} \mathbf{I}_{\mathcal{P}_l}(u) . \quad (1)$$

Incidentally, it is at once apparent that $y(u)$ is no more than the Horvitz-Thompson estimate of T_y when location u is selected. Therefore, the total *intensity* of the variable of interest over the study region turns out to be

$$\int_0^1 y(u) du = \int_0^1 \sum_{l=1}^N \frac{y_l}{\pi_l} \mathbf{I}_{\mathcal{P}_l}(u) du = T_y ,$$

i.e. an integral representation is achieved and hence the estimation of T_y reduces to an integration problem. In turn, the same representation holds if the Horvitz-Thompson estimate is more generally substituted by the linear homogeneous estimate (for a review of these estimates in line-intercept sampling, see Barabesi and Fattorini, 1997)

$$y(u) = \sum_{l=1}^N \frac{y_l}{w_l} \psi_l(u) , \quad (2)$$

where $w_l = \int_0^1 \psi_l(u) du$ and ψ_l is a function vanishing if $u \notin \mathcal{P}_l$. Hence, the strategies adopted for the MC quadrature of an integral can be adopted in order to choose n locations $u_1, \dots, u_n$ on the baseline and to implement an overall estimate of type

$$\bar{T}_y = \bar{T}_y(u_1, \dots, u_n) = \frac{1}{n} \sum_{i=1}^n y(u_i) . \quad (3)$$

A similar framework may be introduced for a Family 2 design. Let us suppose that the study region corresponds to the unit square $(0, 1)^2$. If the study region is not the unit square, a box that closely fits the region is considered and it is scaled to the unit square: a point placed inside the unit square is selected only if it belongs to the study region. Hence, let $\mathbf{u}$ be the coordinates of the selected point on the study region. For a Family 2 design, the inclusion set $\mathcal{S}_l$ for the l -th unit is a suitable connected compact set contained within the study region (see *e.g.* Barabesi, 2003a). Hence, the l -th unit is selected - and y_l is measured - if $\mathbf{u} \in \mathcal{S}_l$. As an example, let us consider a population of trees and let y_l be the wood of the l -th tree. Therefore, the target parameter is the total quantity of wood in the forest. If circular plot sampling is adopted, the inclusion sets are circles of equal sizes centered at the tree locations (Barabesi, 2003a). Hence, a tree is selected if $\mathbf{u}$ is located in the corresponding circle.

In this setting, if solely the l -th unit is considered, the *intensity* of the variable of interest over the l -th inclusion set is given by y_l/π_l , where $\pi_l = \int_0^1 \int_0^1 \mathbf{I}_{S_l}(\mathbf{u}) d\mathbf{u}$ is the area of S_l . Hence, the *intensity* of the variable of interest at location $\mathbf{u}$ is

$$y(\mathbf{u}) = \sum_{l=1}^N \frac{y_l}{\pi_l} \mathbf{I}_{S_l}(\mathbf{u}) \quad (4)$$

and hence the integral representation of T_y turns out to be

$$\int_0^1 \int_0^1 y(\mathbf{u}) d\mathbf{u} = \int_0^1 \int_0^1 \sum_{l=1}^N \frac{y_l}{\pi_l} \mathbf{I}_{S_l}(\mathbf{u}) d\mathbf{u} = T_y .$$

Obviously, the Horvitz-Thompson estimate (4) may be more generally replaced by the linear homogeneous estimate

$$y(\mathbf{u}) = \sum_{l=1}^N \frac{y_l}{w_l} \psi_l(\mathbf{u}) , \quad (5)$$

where $w_l = \int_0^1 \int_0^1 \psi_l(\mathbf{u}) d\mathbf{u}$ and ψ_l is a suitable function vanishing if $\mathbf{u} \notin S_l$. It is worth noting that the Family 1 notations are also adopted in the Family 2 setting with slight abuse. In turn, the strategies adopted for the numerical quadrature of a bivariate integral can be used to choose n locations $\mathbf{u}_1, \dots, \mathbf{u}_n$ on the study region and to implement an estimate of type

$$\bar{T}_y = \bar{T}_y(\mathbf{u}_1, \dots, \mathbf{u}_n) = \frac{1}{n} \sum_{i=1}^n y(\mathbf{u}_i) . \quad (6)$$

In the *continuous variable* case, the parameter representation reduces to a straightforward issue. Indeed, for a Family 1 design, if $y(u)$ is the *intensity* of the variable at location u on the baseline, the target parameter obviously turns out to be

$$T_y = \int_0^1 y(u) du .$$

Accordingly, an estimate of type (3) is appropriate if n locations are selected on the baseline. As an example, in the case of line-transect sampling, if $y(u)$ is the total quantity of a pollutant along a transect placed at location u in the baseline, the target parameter is the total quantity of pollutant over the region. In this framework, an estimate of type (3) is usually implemented by selecting n points on the baseline as starting points for the lines (U.S. EPA QA/G-5S Guidance, 2002, p.38).

In the case of a Family 2 design, if $y(\mathbf{u})$ is the value of the variable at location $\mathbf{u}$ on the study region, the target parameter is obviously given by

$$T_y = \int_0^1 \int_0^1 y(\mathbf{u}) d\mathbf{u}$$

and an estimate of type (6) is suitable. Again, it is worth noting that slight abuse occurs in order to simplify the notation. As an example, let us suppose that $y(\mathbf{u})$ represents the presence of a certain habitat at location $\mathbf{u}$ in the study region, *i.e.* $y(\mathbf{u})$ is a two-valued

function taking value one if the habitat is present and zero otherwise. Therefore, the target parameter is the total coverage of the habitat over the region. In this case, point sampling is usually carried out by selecting n points on the study region and implementing an estimate of type (6).

3. The strategies

In this section, the main features of the four strategies considered are briefly reviewed (for a more comprehensive account, see Barabesi, 2003b). In order to implement estimate (3), the basic MC strategy requires n realizations $u_1, \dots, u_n$ of n independent random variables uniformly distributed on the baseline. Alternatively, in order to implement estimate (6), the basic MC strategy is based on n realizations $\mathbf{u}_1, \dots, \mathbf{u}_n$ of n independent bivariate random vectors uniformly distributed on the study region. As is well known (Rubinstein, 1981, p.118), the MC strategy produces unbiased estimators with a normal large-sample distribution, as well as consistent variance estimation. However, the MC estimators are inefficient since they display low variance convergence rates, whichever class function y belongs to (Table 1).

In order to implement estimate (3), the MMC strategy requires that the baseline be partitioned into n segments of equal lengths, in such a way that the u_i s are the realizations of independent random variables uniformly distributed in each segment. In order to implement estimate (6), the MMC strategy requires that the study region be partitioned into n rectangles of equal sizes, in such a way that the $\mathbf{u}_i$ s are the realizations of independent bivariate random vectors uniformly distributed in each rectangle. Interestingly enough, the genesis of the MMC strategy is linked to ranked-set sampling, as has recently been shown by Al-Saleh and Samawi (2000) (for more details on ranked-set sampling, see Barabesi and El-Sharaawi, 2001, Barabesi, 2001b, Barabesi and Fattorini, 2002, Barabesi and Marcheselli, 2004). The MMC strategy produces unbiased estimators which are more accurate (for each n) than the estimators obtained under the MC strategy. In addition, the MMC estimators display an elevated variance convergence rate, whichever class function y belongs to (Table 1). Moreover, if function y obeys a Lipschitz condition, the MMC estimators display a normal large-sample distribution and consistent variance estimation is available (Barabesi and Marcheselli, 2003, Barabesi and Pisani, 2004).

In order to implement estimate (3), the RG strategy requires the same partition as in the MMC strategy, but u_1 is the realization of a random variable uniformly distributed in the first segment, while $u_2, \dots, u_n$ are systematically replicated in the other segments. Similarly, in order to implement estimate (6), the RG strategy involves the same partition as in the MMC strategy, but $\mathbf{u}_1$ is the realization of a bivariate random vector uniformly distributed in the first rectangle, while $\mathbf{u}_2, \dots, \mathbf{u}_n$ are systematically replicated in the other rectangles. The RG strategy produces unbiased estimators, which however may be less accurate than the estimators obtained under the basic MC strategy. The RG estimators may display an elevated variance convergence rate, which increases with the "regularity" of the function y (Garcia-Fiñana and Cruz-Orive, 2000). For example, if a Family 1 design is considered, a $O(n^{-2q-2})$ variance rate is obtained where q is a positive "smoothness" parameter (with $0 < q < 1$ in practical stereological applications). However, in order to achieve a high variance rate (*i.e.* for $q > 0$), y must

be a periodic function. Moreover, no results on the large-sample distribution of the RG estimators are available, even if consistent variance estimation is possible (Kiêu *et al.*, 1999, Garcia-Fiñana and Cruz-Orive, 2000).

In order to implement estimate (3) and (6), the RQMC strategy requires n realizations $u_1, \dots, u_n$ (equivalently $\mathbf{u}_1, \dots, \mathbf{u}_n$) of the first n terms of a shuffled one-dimensional (equivalently two-dimensional) quasi-random sequence (L'Ecuyer and Lemieux, 2002). The RQMC strategy produces consistent estimators (which may be unbiased for some n), which however may be less accurate than the estimators obtained under the basic MC strategy. The RQMC estimators display an elevated variance convergence rate and a large-sample normal distribution if the function y is in the class C^2 (Table 1). However, no results are available for less smooth functions. In addition, consistent variance estimation is solely possible by using a set of randomized quasi-random sequences (L'Ecuyer and Lemieux, 2002).

Table 1. The variance rate for the class of the function y . E represents the class of elementary functions. The question mark indicates that the rate is not known. See the text for values indicated with an asterisk.

Strategy	Family	E	C^0	C^1	C^2
MC	1	n^{-1}	n^{-1}	n^{-1}	n^{-1}
	2	n^{-1}	n^{-1}	n^{-1}	n^{-1}
MMC	1	n^{-2}	n^{-3}	n^{-3}	n^{-3}
	2	$n^{-3/2}$	n^{-2}	n^{-2}	n^{-2}
RG	1	n^{-2}	*	*	*
	2	$n^{-3/2}$	*	*	*
RQMC	1	?	?	?	n^{-3}
	2	?	?	?	$n^{-3} \log n$

On the basis of the methodological remarks contained in this section and on the basis of the simulation performed in Barabesi (2003b), in the author's opinion the MMC strategy may be preferable in the setting of an elementary function y (as in the case of Horvitz-Thompson estimation). In addition, the MMC strategy might perform well also if y is in the class C^0 (as in the case of the linear homogeneous estimation considered by Barabesi and Fattorini, 1997), even if in this setting the RG strategy may be competitive. Further analysis would be welcome for smoother y functions, especially from a practical viewpoint.

4. Advanced methods

For a Family 1 design, it is worth noting that the parameter may be expressed as

$$T_y = \int_0^1 \frac{y(u)}{g(u)} g(u) du = \int_0^1 \frac{y(G^{-1}(u))}{g(G^{-1}(u))} du ,$$

where $g(u)$ is a probability density function on $(0, 1)$ with corresponding distribution function $G(u)$. This parameter reformulation gives rise to the well-known importance sampling method (Evans and Swartz, 2000, p.166), which leads to the estimate

$$\bar{T}_{y,\text{IMP}} = \bar{T}_{y,\text{IMP}}(u_1, \dots, u_n) = \frac{1}{n} \sum_{i=1}^n \frac{y(G^{-1}(u_i))}{g(G^{-1}(u_i))}, \quad (7)$$

where the u_i s may be selected according to a strategy described in the previous section. Hence, if some information about the structure of $y(u)$ is known, $g(u)$ may be chosen approximately proportional to $y(u)$. Therefore, the estimator variance may be considerably reduced when $g(u)$ is suitably chosen (Evans and Swartz, 2000, p.171). In the environmental framework, importance sampling may actually be interpreted as a stratification technique, in the sense that it allows for a more elevated density of points in some parts of the baseline according to the structure of $g(u)$. Finally, since it obviously turns out that

$$\bar{T}_{y,\text{IMP}} = \frac{1}{n} \sum_{i=1}^n \psi(u_i),$$

where $\psi(u) = y(G^{-1}(u))/g(G^{-1}(u))$, it is worth noting that estimate (7) is actually an analog of estimate (3). Hence, the same comments made in Section 3 may be applied in order to choose an appropriate strategy for the selection of $u_1, \dots, u_n$.

As an example of the previous considerations, if line intercept sampling is carried out, more information may be obtained if the procedure favors large length transects on the study region. Thus, it may be convenient to select transects with probabilities proportional to their length, *i.e.* so-called L-weighted transects (Barabesi and Fattorini, 1997). In this setting, $g(u)$ is chosen as

$$g(u) = \frac{l(u)}{A} \mathbf{I}_{(0,1)}(u),$$

where $l(u)$ is the length (on the study region) of a transect placed at location u on the baseline, while $A = \int_0^1 l(u) du$ is the area of the study region. The estimates considered in Barabesi and Fattorini (1997) and Valentine *et al.* (2001) may be obtained by coupling this choice of $g(u)$ with linear homogeneous estimates of type (2).

Furthermore, it is worth noting that Cordy's (1993) proposal - usually cited in the point sampling framework - is based on the mixture

$$g(u) = \frac{1}{n} \sum_{l=1}^n g_l(u), \quad (8)$$

where the $g_l(u)$ s are probability density functions. This special choice (8) is remarkable since it leads to an interpretation of estimate (7) in terms of Horvitz-Thompson estimation (Cordy, 1993). Moreover, it allows for an automatic stratification of the baseline. As an example, let us consider the classical stratification technique which requires the partition of the baseline into k strata $\mathcal{C}_1, \dots, \mathcal{C}_k$, in such a way that n_l points are uniformly selected in the l -th stratum. This setting may be obtained from expression (8) when $g(u)$ is set to

$$g(u) = \frac{1}{n} \sum_{l=1}^k \frac{n_l}{c_l} \mathbf{I}_{\mathcal{C}_l}(u), \quad (9)$$

where $c_l = \int_0^1 I_{C_l}(u) du$ is the size of the l -th stratum. Incidentally, it is worth noting that the MMC strategy is actually based on the choices $n_l = 1$ and $C_l = [(l-1)/n, l/n)$ in (9). Therefore, the MMC strategy may also be interpreted as an optimal stratification technique. Finally, it is very important to observe that the automatic stratification reached by means of (8) avoids the problem of additional boundary corrections (for a review of the boundary effect problem, see Schreuder *et al.*, 1993, p.297). In contrast, if the study region is partitioned into strata, in such a way that the aim is to estimate the total as a sum of the totals in each stratum, the boundary corrections increase with the number of strata. In this case, environmental sampling may become practically unfeasible in the field.

The issues contained in this section may be easily extended to a Family 2 design. An additional comment solely involves the two-dimensional generalization of the mixture (9). Indeed, in this case the MMC strategy is achieved if the strata are based on a regular grid of squares, but similar procedures may be obtained by considering regular grids of triangles or hexagons (Stevens, 1997, U.S. EPA QA/G-5S Guidance, 2002, p.72). In turn, these procedures may be easily adapted to the RG strategy.

5. Auxiliary variables

Let us suppose that an auxiliary variable is available. Hence, for a Family 1 design, let us suppose that a known function $z(u)$ - which resembles $y(u)$ - is at our disposal. Accordingly, the target parameter may be expressed as

$$T_y = \int_0^1 (y(u) - z(u)) du + \int_0^1 z(u) du = \int_0^1 (y(u) - z(u)) du + T_z ,$$

where

$$T_z = \int_0^1 z(u) du .$$

Obviously, this parameter reformulation gives rise to the well-known control-variate method (Evans and Swartz, 2000, p.206). In this case, an estimate (which closely resembles the usual difference estimate considered in sample theory) is given by

$$\bar{T}_{y, \text{CON}} = \bar{T}_{y, \text{CON}}(u_1, \dots, u_n) = \frac{1}{n} \sum_{i=1}^n (y(u_i) - z(u_i)) + T_z . \quad (10)$$

However, since it is difficult to specify in practice the function $z(u)$, it is useful to assume a model of type $z(u) = a + bx(u)$, where $x(u)$ is a known function, such that

$$T_x = \int_0^1 x(u) du ,$$

while a and b are unknown parameters. In the environmental sampling setting, this choice of $z(u)$ actually leads to the regression estimator (Brus, 2000). Thus, on the basis of the previous expression of T_y , by choosing n locations $u_1, \dots, u_n$ on the study region, the regression estimate is given by

$$\bar{T}_{y,\text{REG}} = \bar{T}_{y,\text{REG}}(u_1, \dots, u_n) = \bar{T}_y - \hat{b}(\bar{T}_x - T_x), \quad (11)$$

where $\bar{T}_y$ is defined in (3), while

$$\bar{T}_x = \bar{T}_x(u_1, \dots, u_n) = \frac{1}{n} \sum_{i=1}^n x(u_i)$$

and

$$\hat{b} = \frac{\sum_{i=1}^n (x(u_i) - \bar{T}_x)(y(u_i) - \bar{T}_y)}{\sum_{i=1}^n (x(u_i) - \bar{T}_x)^2}.$$

For large n , stochastic Taylor series expansion provides the following linearization of $\bar{T}_{y,\text{REG}}$

$$\bar{T}_{y,\text{REG}} \sim \bar{T}_y - b(\bar{T}_x - T_x),$$

where b is the population parameter value obtained with the least-square method (Brus, 2000)

$$b = \frac{\int_0^1 (x(u) - T_x)(y(u) - T_y) du}{\int_0^1 (x(u) - T_x)^2 du}.$$

On the basis of the previous expansion, it is worth noting that

$$\bar{T}_{y,\text{REG}} \sim \frac{1}{n} \sum_{i=1}^n \phi(u_i),$$

where $\phi(u) = y(u) - b(x(u) - T_x)$. Accordingly, (11) actually displays a structure similar to (3) - at least in a large-sample setting. Hence, the same comments made in Section 3 may be applied in this case in order to choose an appropriate strategy for the selection of $u_1, \dots, u_n$. Moreover, the regression estimator may be coupled with the importance sampling method. In addition, more than one auxiliary variable may be considered, *i.e.* the model may be expressed as $z(u) = a + \sum_{l=1}^m b_l x_l(u)$, where $x_1(u), \dots, x_m(u)$ are known functions, while a and $b_1, \dots, b_m$ are unknown parameters. Finally, it should be remarked that the regression estimators may be easily extended to a Family 2 design by means of suitable (and straightforward) changes in the notation.

6. Conclusions and future directions

When the *continuous plane* paradigm is considered, the replicated environmental sampling design setting may be embedded into a simple homogeneous theory by using Monte Carlo integration methods. This theory makes it possible to observe a *plethora* of different techniques commonly adopted in environmental studies in a unique perspective. Moreover, it is at once apparent that this link may produce new and effective methods. As an example, the MMC (alternatively the RG) strategy may be coupled with importance sampling estimates or regression estimates in order to obtain estimators with an elevated variance convergence rate to be used in forestry and in soil sciences. In

addition, more effective Monte Carlo integration strategies - such as an improved MMC strategy - may be introduced. The author is planning to explore these topics in subsequent papers.

References

- Al-Saleh, M.F. and Samawi, H.M. (2000) On the efficiency of Monte Carlo methods using steady state ranked simulated samples, *Communications in Statistics - Simulation and Computation* **29**, 941-954.
- Barabesi, L. (2000) Local likelihood density estimation in line transect sampling, *Environmetrics* **11**, 413-422.
- Barabesi, L. (2001a) A design-based approach to the estimation of plant density using point-to-plant sampling, *Journal of Agricultural, Biological and Environmental Statistics* **6**, 89-98.
- Barabesi, L. (2001b) The unbalanced ranked-set sample sign test, *Journal of Nonparametric Statistics* **13**, 279-289.
- Barabesi, L. (2003a) Sampling designs for monitoring ecological diversity, in *Environmetrics* (editor A. El-Sharaawi), Encyclopedia of Life Support Systems (EOLSS), developed under the auspices of the UNESCO, Eolss Publishers, Oxford.
- Barabesi, L. (2003b) A Monte Carlo integration approach to Horvitz-Thompson estimation in replicated environmental designs, *Metron* **LXI**, 355-374.
- Barabesi, L. and El-Sharaawi, A. (2001) The efficiency of ranked set sampling for parameter estimation, *Statistics and Probability Letters* **53**, 189-199.
- Barabesi, L. and Fattorini, L. (1997) Line intercept sampling with finite populations: a stereological approach, *Metron* **LV**, 23-37.
- Barabesi, L. and Fattorini, L. (1998) The use of replicated plot, line and point sampling for estimating species abundancies and ecological diversity, *Environmental and Ecological Statistics* **5**, 353-370.
- Barabesi, L. and Fattorini, L. (2002) Kernel estimators of probability density functions by ranked-set sampling, *Communications in Statistics - Theory and Methods* **31**, 597-610.
- Barabesi, L., Greco, L. and Naddeo, S. (2002) Density estimation in line transect sampling with grouped data by local least squares, *Environmetrics* **13**, 167-176.
- Barabesi, L. and Marcheselli, M. (2002) Species abundance estimation by using point-to-plant sampling in a design-based setting, *Environmental and Ecological Statistics* **9**, 393-403.
- Barabesi, L. and Marcheselli, M. (2003) A modified Monte Carlo integration, *International Mathematical Journal* **3**, 555-565.
- Barabesi, L. and Marcheselli, M. (2004) Design-based ranked set sampling using auxiliary variables, *Environmental and Ecological Statistics*, in the press.
- Barabesi, L. and Pisani, C. (2002) Ranked set sampling for replicated sampling designs, *Biometrics* **58**, 586-592.
- Barabesi, L. and Pisani, C. (2004) Steady-state ranked set sampling for replicated environmental sampling designs, *Environmetrics* **15**, 45-56.
- Brus, D.J. (2000) Using regression models in design-based estimation of spatial means of soil properties, *European Journal of Soil Sciences* **51**, 159-172.

- Brus, D.J. and De Gruijter, J.J. (1997) Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion), *Geoderma* **80**, 1-59.
- Cordy, C.B. (1993) An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe, *Statistics and Probability Letters* **18**, 353-362.
- Cruz-Orive, L.M. (1993) Systematic sampling in stereology, *Bulletin of the International Statistical Institute*, Proceedings of 49th Session, Florence 1993, **53**, 451-468.
- Evans, M. and Swartz, T. (2000) *Approximating integrals via Monte Carlo and deterministic methods*, Oxford University Press, Oxford.
- Fattorini, L., Marcheselli, M. and Pisani, C. (2004) Two-phase estimation of coverages with second-phase corrections, *Environmetrics*, in the press.
- Fattorini, L. and Pisani, C. (2004) Variance decomposition in two-stage plot sampling: theoretical and empirical results, *Environmental and Ecological Statistics*, in the press.
- Fattorini, L. and Pisani, C. (1999) *Metodi di campionamento per le indagini ambientali*, Stamperia della Facoltà di Economia, Università di Siena.
- Garcia-Fiñana, M. and Cruz-Orive, L.M. (2000) Fractional trend of the variance in Cavalieri sampling, *Image Analysis and Stereology* **19**, 71-79.
- Gilbert, R.O. (1987) *Statistical methods for environmental pollution monitoring*, Van Nostrand Reinhold, New York.
- Haber, S. (1966) A modified Monte-Carlo quadrature, *Mathematics of Computation* **20**, 361-368.
- Kaiser, L. (1983) Unbiased estimation in line-intercept sampling, *Biometrics* **39**, 965-976.
- Kiêu, K., Souchet, S. and Istas, J. (1999) Precision of systematic sampling and transitive methods, *Journal of Statistical Planning and Inference* **77**, 263-279.
- L'Ecuyer, P. and Lemieux, C. (2002) Recent advances in randomized quasi-Monte Carlo methods, in *Modeling uncertainty: an examination of stochastic theory, methods and applications*, P.L.M. Dror, P. L'Ecuyer and F. Szidarowski (eds.), Kluwer, Dordrecht, 419-474.
- Pisani, C. (2002) The estimation of biological population size at large scale by incomplete area surveys and replicated counts, *Environmetrics* **13**, 155-166.
- Rubinstein, R.Y. (1981) *Simulation and the Monte Carlo method*, Wiley, New York.
- Schreuder, H.T., Gregoire T.G. and Wood, G.B. (1993) *Sampling Methods for Multiresource Forest Inventories*, Wiley, New York.
- Stevens, D.L. (1997) Variable density grid-based sampling designs for continuous spatial populations, *Environmetrics* **8**, 167-195.
- Thompson, S.K. (1992) *Sampling*, Wiley, New York.
- U.S. Environmental Protection Agency (2002) *Guidance on choosing a sampling design for environmental data collection*, EPA QA/G-5S, Washington DC, 1-166.
- Valentine, H.T., Gove, J.H. and Gregoire, T.G. (2001) Monte Carlo approaches to sampling forest tracts with lines or points, *Canadian Journal of Forestry Research* **31**, 1410-1424.
- Williams, M.S. and Eriksson, M. (2002) Comparing the two paradigms for fixed area sampling in large-scale inventories, *Forest Ecology and Management* **168**, 135-148.