MCMC selection of spatially and doubly balanced samples

Roberto Benedetti*
“G. d’Annunzio” University, Pescara, Italy – benedett@unich.it

M. Simona Andreano
Universitas Mercatorum, Rome, Italy – s.andreano@unimercatorum.it

Federica Piersimoni
Istat, Rome, Italy – piersimo@istat.it

Paolo Postiglione
“G. d’Annunzio” University, Pescara, Italy – postigli@unich.it

Abstract
A method for selecting samples from a spatial finite population that are well spread over the population in every dimension, without the use of any spatial stratification, is presented. The within sample distance matrix is summarized in a descriptive index that is used to define the probability of each sample to be selected. A set of units with higher within distance will be selected with higher probability than a set of nearby units. Through the standardization of the distance matrix, the method can be used to produce equal and unequal probability samples either exact when a linear index is used to summarize the matrix or approximate when products and powers of the mean are used. The high flexibility of the selection algorithm can make possible numerous extensions to deal with some practical topics that are usually met in spatial surveys, such as the sample coordination and the spread of units belonging to different domains. Some examples on real and simulated data show that the method gives estimates that are better than those obtained by using a classical solution as the Generalized Random Tessellation Stratified (GRTS) design and that often are even slightly better than those obtained by using recently proposed selection procedures as the Spatially Correlated Poisson Sampling (SCPS), the Pivotal method and the doubly balanced design.

Keywords: anticipated variance; empirical inclusion probabilities; correlated Poisson sampling; Generalized Random Tessellation Stratified design.

1. Introduction
In business surveys in general, and in multipurpose agricultural surveys in particular, the problem of designing a sample from a frame usually consists of three different aspects (Särndal et al. 1992; Benedetti et al. 2010). The first is the choice of a rule for stratifying the population when several size variables are available; the second is the definition of the selection probabilities for each unit in the frame; and the third is devoted to sample size determination and sample allocation to a given set of strata.

The main property required of the sample design is that it should provide a specified level of precision for a set of variables of interest, using as few sampling units as possible. Stratification is introduced into sampling designs for a number of different reasons: for example, to select the sample from a given frame, to obtain reliable estimators for sub-populations (i.e. domains), or to improve the estimators’ efficiency of global population parameters.

In most cases, populations are either naturally stratified or can be stratified easily, on the basis of practical considerations such as administrative subdivisions. In other circumstances, strata are established to satisfy the interest in identifying characteristics of sub-populations. When such straightforward definitions are not possible, then a decision should be taken on the number of strata and their respective boundaries.
These typical design issues should be considered together with that of the importance of bearing geographical position in mind when selecting samples of statistical units. Today, this issue is recognized more than ever when measuring many phenomena, due to several reasons. First, because evidence exists that statistical units are defined by using purely spatial criteria, as occurs in most agricultural and environmental studies. Second, in several countries, it is common practice for the National Statistical Institute (NSI) to geo-reference the typical sampling frames of physical or administrative bodies not only according to the codes of a geographical nomenclature, but also by adding information on the exact or estimated, position of each record.

Often, spatial units are also artificially defined, and made available over a domain partitioned into a number of predetermined regularly- or irregularly-shaped sets of spatial objects. This may occur, for example, when the original data lie over a continuous spatial domain and, to simplify the problem, researchers decide to observe them only in a selection of fixed points, potentially made at random or averaged over a selection of predefined polygons.

Although the monitoring and the estimation of infinite populations covers an important part of sampling problems in the context of natural resources, agricultural surveys deal mainly with finite populations.

Indeed, in this latter context, the spatial distribution of the frame is a constraint. For this reason, it is supposed that it could have considerable impact on the performance of a random sampling method. For example, the traditional solution of extending the systematic sampling to multidimensional data by simply overlaying a grid of points to a spatial domain may not be feasible, if the population cannot be considered as distributed on a regular grid because it is clustered or displays different intensities of units across the domain.

Assume that it is sought to estimate a parameter of a set \( Y = \{ y_1, y_2, \ldots, y_i \} \) of \( v \) variables of interest, generally denoted as survey variables. Let \( U = \{1, 2, \ldots, N\} \) be a finite population recorded on a frame together with a set of \( k \) auxiliary variables \( X = \{ x_1, x_2, \ldots, x_l \} \) and a set of \( h \) (usually \( h=2 \)) coordinates \( C = \{ c_1, c_2, \ldots, c_h \} \), obtained by geo-coding each unit, where \( x_i = \{ x_{i1}, x_{i2}, \ldots, x_{il}, \ldots, x_{in} \} \) is the generic \( l \)-th auxiliary and \( c_i = \{ c_{i1}, c_{i2}, \ldots, c_{ih}, \ldots, c_{in} \} \) is the generic \( l \)-th coordinate. From \( C \) we can always derive, for any distance definition, a matrix \( D_U = \{ d_{ij}; i = 1, \ldots, N, j = 1, \ldots, N \} \), which specifies how far apart are all the pairs of units in the population.

In several agricultural surveys, the geographical position is an intrinsic characteristic of the unit and, given the particular nature of this information, its efficient use in sample design often requires methods that cannot be adapted from those used when dealing with classical auxiliary variables.

This is not only a consequence of its multivariate nature and of traditional design solutions, as the \( \pi ps \) (i.e. inclusion probability proportional to size) can handle only one auxiliary (Bee et al. 2010). To use certain covariates, we always assume that there is, at least approximately, a certain degree of correlation between a survey variable \( y \) and the set \( X \). With regard to the use of set \( C \), the commonly used distance matrix as a synthesis of the spatial information emphasizes the importance of the sample’s spread over the region under study. This feature may be related to this dependence, but also to some form of similarity between adjacent units.

Usually, \( X \) and \( C \) in agricultural surveys play different roles according to the definition of the statistical unit:

1 - When \( U \) is a list of agricultural households, \( C \) is rarely obtainable, depending on the availability of accurate cadastral maps, and should be constituted by a map of polygons representing the parcels of land used by each holding. \( X \) is usually filled with administrative data sources, previous census data and, only if \( C \) is available, remotely sensed data obtained through the overlay of the polygon map with a classified image;

2 - If \( U \) is a list of regularly- or irregularly-shaped polygons defined ad hoc for the agricultural survey, \( C \) is always available, since it represents the very definition of each statistical unit and \( X \), unless an
overlay of C with a cadaster is possible, can be constituted only by some geographical coding and by summarizing a classification arising from remotely sensed data within each polygon;

3 - Another possible choice often made in agricultural surveys is for U to represent a list of points, usually the corners of the regular grid, overlaid over the survey geographical domain. This, therefore, represents a non-exhaustive population of the study area and only a first stage of sampling. In this case, X can be only represented with a geographical nomenclature and with a design matrix of codes for land use classification obtained, with previous land use maps or with a classification of remotely sensed data, while C represents the coordinates of each point.

In the first type of survey, the relevant structural characteristic to be controlled is that the population under investigation is highly skewed, as the concentration of farms and agricultural households’ sizes is very high. Most of these units have a small size, and are not important in economic terms, although they may be interesting for a rural development analysis.

On the other hand, a limited number of large units represents a significant part of the population, and must, therefore, always be included in any sample survey. This is a typical situation arising in any business survey in which the population of interest is extremely positively skewed, due to the presence of a few large units and several small units. Thus, when estimating an unknown total of the population, many small observations make a negligible contribution, whereas a few large observations have a dramatic impact on the estimates.

2 Spatially balanced samples

In recent decades, the spatial balancing of samples has become so peculiar that several researchers and survey practitioners have suggested sampling algorithms to achieve it (Benedetti et al. 2015; Wang et al. 2012). Surprisingly, this balancing is mainly based on intuitive considerations, and it is not so clear when and to what extent it could have an impact on estimate efficiency. Moreover, it is also useful to note that this feature was not defined with sufficient adequacy. As a consequence, a range of possible interpretations exists that make it unfeasible to perform any comparison between different methods.

In design-based sampling theory, if it is assumed that there is no measurement error, the potential observations concerning each unit of the population cannot be considered as dependent. However, an inherent and fully recognized feature of spatial data is that it is dependent; this much was succinctly expressed in Tobler’s First Law of Geography, according to which everything is related to everything else, but near things are more related than distant things.

Thus, it is clear that sampling schemes for spatial units can be given reasonable treatment by introducing a suitable model of spatial dependence into a model-based or at least a model-assisted framework. In literature (Benedetti and Palma, 1995; Rogerson and Delmelle, 2004), this approach proved to be helpful in rationalizing the intuitive procedure to spread the selected units over the space, because closer observations will provide overlapping information as an immediate consequence of dependence.

However, based on this assumption, the concern is, necessarily, how to find the sample configuration that can best represent the entire population. This leads to define our selection as a problem of combinatorial optimization. Indeed, if the sample size is fixed, the aim is to minimize an objective function defined over the whole set of possible samples, which is a measure of the loss of information due to dependence.

An optimal sample selected with certainty is of course not acceptable, if we assume the randomization hypothesis that is the background for design-based inference. Thus, it is necessary to depart from the concept of dependence in favour of that of spatial homogeneity, measured in terms of the local variance of the observable variables, where all the units of the population within a given distance could be defined as local units.

An intuitive method of producing samples that are well spread across the population, and that is widely used by practitioners, is to stratify the population units on the basis of their location. The problems arising from the adoption of this strategy lie in the evidence that it fails to have direct and substantial impact on the second-order inclusion probabilities, certainly not within a given stratum; and that the way to obtain a good partition of the study area is often unclear.
These drawbacks are related to a certain extent; for this reason, they are usually approached in combination by defining a maximal stratification, i.e. partitioning the study in as many strata as possible, and selecting one or two units per stratum. However, this straightforward and fast scheme to ensure that the sample is well spread across the population is somewhat arbitrary, because it greatly depends upon the stratification criterion, which should be general and efficient.

The basic principle is to extend the use of systematic sampling to two or more dimensions, a concept on which the Generalized Random Tessellation Stratified (GRTS) design is based (Stevens and Olsen 2004). In this case to select the units systematically, the design maps the two-dimensional population into one dimension, while attempting to preserve some multi-dimensional order. This approach is essentially based on the use of Voronoi polygons, which are used to define an index of spatial balance. Let denote with S the set of all possible random samples of fixed size n, which can be selected from U, where its generic element is \( s = \{s_1, s_2, \ldots, s_N\} \), and \( s_i \) is equal to 1 if the unit with label \( i \) is in the sample, and 0 otherwise. For any unit \( i \), let \( \pi_i (i \in U) \) be the first-order inclusion probability, and for any couple \( \{i,j\} \), let \( \pi_{ij} (i,j \in U) \) be the second order-inclusion probability. For a generic sample \( s \), the Voronoi polygon for the sample unit \( s_i \) includes all population units closer to \( s_i \) than to any other sample unit \( s_j \). If we let \( v_i \) be the sum of the inclusion probabilities of all units in the \( i \)-th Voronoi polygon, for any sample unit \( u_i \), we have \( E(v_i)=1 \); and for a spatially balanced sample, all \( v_i \) should be close to 1. Thus, the index \( V(v_i) \) (i.e. the variance of the \( v_i \)) can be used as a measure of spatial balance for a sample.

Note that this concept is rather distant from that of balanced sampling, introduced in model-based sampling (Deville and Tillé, 2004) and that is reasonably accepted even in the design-based approach, through the introduction of the cube method (Chauvet and Tillé, 2006; Tillé, 2006; Tillé, 2011), as a restriction of the support \( S \) of samples that can be selected by imposing a set of linear constraints on the covariates. These restrictions represent the intuitive requirement that the sample estimates of the total, or of the average, of a covariate should be equal to the known parameter of the population. In a spatial context, this plan could be applied by imposing the requirement that any selected sample should respect the first \( p \) moments for each coordinate, assuming implicitly that the survey variable \( \gamma \) follows a polynomial spatial trend of order \( p \) (Breidt and Chauvet, 2012). However, these selection strategies do not use the concept of distance, which is a basic tool to describe the spatial distribution of the sample units, which leads to the intuitive criterion that units that are close, seldom appear simultaneously in the sample. This condition can be considered reasonable under the assumption that, if the distance between two units \( i \) and \( j \) increases, the difference \( |y_i - y_j| \) between the values of the survey variable always increases. In these situations, it is clear that the Horvitz-Thompson variance of estimates will necessarily decrease, if high joint inclusion probabilities are assigned to couples with very different \( y \) values due to their great distance from one another, to the detriment of couples that are expected to have similar \( y \) values due to their closeness.

Following this idea, Arbia (1993), inspired by purely model-based assumptions on the dependence of the stochastic process generating the data, suggested a draw-by-draw scheme: the Dependent Areal Units Sequential technique (DUST), which began with a unit selected at random, say \( i \), and in any step \( t<n \) updates the selection probabilities according to the rule \( \pi_i^{(t)} = \pi_i^{(t-1)} \left( 1 - e^{-\lambda d_{i}} \right) \), where \( \lambda \) is a tuning parameter that is useful in controlling the sample’s distribution over the region under study. This algorithm, or at least the sampling design that it implies, can be interpreted and analysed easily in a design-based perspective, especially referring to a careful estimation and analysis of its first and second-order inclusion probabilities.

Recently, some advances have been proposed for list sequential algorithms whose updating rules present the crucial property of preserving the fixed first-order inclusion probabilities (Grafström, 2012; Grafström et al. 2012; Grafström and Tillé, 2013). In particular, Grafström (2012) suggested a list sequential algorithm that, for any unit \( i \), in any step \( t \) updates the inclusion probabilities according to a \( \pi_i^{(t)} = \pi_i^{(t-1)} \left( I_i - \pi_i^{(t-1)} \right) w_{i}^{(t)} \) rule, where \( w_{i}^{(t)} \) are weights that depend on \( I_1, I_2, \ldots, I_{t-1} \) but not on \( I_t \),
and \( I_t \) is an indicator function set equal to 1, if the unit \( t \) is included in the sample and otherwise equal to 0.

The weight determines how the inclusion probabilities for the unit \( i \) should be affected by the outcome of unit \( t \). These are defined such that, if they satisfy upper and lower bounds, the initial \( \pi \)'s are not modified. The suggested maximal weights criterion places as much weight as possible upon the closest unit, then to the second closest unit, and so on.

Two alternative procedures to select samples having a fixed \( \pi \) and correlated inclusion probabilities were derived (Grafström et al. 2012), as an extension of the Pivotal method introduced to select \( \pi \)'s.

These are essentially based on an updating rule of the \( \pi \) and \( \pi \) probabilities, that should, at each step, locally keep the sum of the updated probabilities as constant as possible, and differ from each other so that the two nearby units \( i \) and \( j \) can be chosen. These two methods are known as the Local Pivotal Method 1 (LPM1), which, according to the authors’ suggestion, is the more balanced of the two; and the Local Pivotal Method 2 (LPM2), which is simpler and faster.

To understand the circumstances in which it is efficient to spread the selected units over the population, it must be supposed that the distance matrix summarizes all the features of the spatial distribution of the population and, as a consequence, of the sample. Within a model-based perspective, this general hypothesis is equivalent to assuming that the data-generating process is stationary and isotropic (i.e. its distribution does not change the coordinates’ space as we shifted or rotated).

Focusing on set \( C \), without using any other information from \( X \), this assumption implies that the problem in selecting spatially balanced samples lies in the definition of a design \( p(s) \), with a probability proportional to some synthetic index \( M(d_s) \) of the within sample distance matrix \( d_s \), when it is observed within each possible sample \( s \) by using an MCMC algorithm to select the sample (Traat et al. 2004).

There are several circumstances in which caution should be taken when selecting samples, so that they are spatially well-distributed:

1. \( y \) has a linear or monotone spatial trend;
2. there is spatial homogeneity (i.e. the data relating to close units is more similar than those relating to distant units);
3. \( y \) appears to follow zones of local stationarity of the mean and/or of the variance; in other words, the observed phenomenon displays spatial stratification;
4. the population’s units present a spatial pattern, which can be clustered; in other words, the units’ intensity varies across the region under study.

It is noteworthy that, while the distance between a pair is a basic concept in all these features of the phenomenon, the index \( I(v) \) of spatial balance appears to be directly related to the third aspect, but only indirectly with the other three. This consideration and the practical impossibility of using the index \( I(v) \), because it involves the \( \pi_i \), motivates the use of a rule that sets the probability \( p(s) \) of selecting a sample \( s \) proportionally, or more than proportionally, to a synthesis of the distance matrix within the sample \( d_s \).

5. Conclusions

Many populations in environmental, agricultural, and forestry studies are distributed over space, but it is almost clear nowadays that spatial units cannot be sampled as if they were generated within the classical urn model. This is mainly due to the impact on the sample design that a set of inherent structures which characterizes spatial data present clustering of the coordinates, homogeneity and/or dependence, spatial trends and local stationarity.

The question is how to incorporate these spatial aspects into the design following an efficient approach and to understand to which limits these aspects can be exploited to reduce the variance of the estimators. The common and widely methods of spatial systematic sampling and spatial stratified sampling use these features only partially, thus in this paper we treat a framework for sampling from a spatial population that is based on the within sample distance which could give some explanation to these questions and therefore propose a broad and flexible class of sampling designs which may differ from each other by changing the definition of the index used to summarize the distances.
The main strength of selecting samples according to this criterion lies in its ability to produce samples that are well spread over the population and that take advantage of the presence of any of the peculiar spatial structures that can be met in the analysis of geo-coded populations. The possible extensions of these methods to stratified designs and to sample coordination also deserve to be mentioned.

References