A review of spatial sampling

Jin-Feng Wang [a,*], A. Stein [b], Bin-Bo Gao [a], Yong Ge [a]

[a] *State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, A11 Datun Road, Beijing 100101, China*
[b] *Twente University, Faculty of Geo-information Science and Earth Observation (ITC), Enschede, The Netherlands*

ARTICLE INFO

ABSTRACT

The main aim of spatial sampling is to collect samples in 1-, 2- or 3-dimensional space. It is typically used to estimate the total or mean for a parameter in an area, to optimize parameter estimations for unsampled locations, or to predict the location of a movable object. Some objectives are for populations, representing the "here and now", whereas other objectives concern superpopulations that generate the populations. Data to be collected are usually spatially autocorrelated and heterogeneous, whereas sampling is usually not repeatable. In various senses it is distinct from the assumption of independent and identically distributed (i.i.d.) data from a population in conventional sampling. The uncertainty for spatial sample estimation propagates along a chain from spatial variation in the stochastic field to sample distribution and statistical tools used to obtain an estimate. This uncertainty is measured using either a design-based or model-based method. Both methods can be used in population and superpopulation studies. An unbiased estimate with the lowest variance is thus a common goal in spatial sampling and inference. Reaching this objective can be addressed by sample allocation in an area to obtain a restricted objective function.

© 2012 Elsevier B.V. All rights reserved.

Contents

* Corresponding author.
  E-mail address: wangjf@lreis.ac.cn (J.-F. Wang).

1. Introduction

Sampling concerns selection of a subset of individuals from within a population to estimate characteristics of the whole population. The characteristics could be the total or mean parameter value for a random field (Haining, 2003; Christakos, 2005), values at unsampled sites (Goovaerts, 1997), or location of target(s) (Rogerson et al., 2004).

1.1. A brief history of spatial sampling

Intuitive application of the principles of sampling in science has occurred for a long time from very early human history in Egypt, China and other places throughout the world. The first known attempt to make statements about a population using information about only part of it was by the English merchant John Graunt (1620–1674). His famous tract describes a method for estimating the population of London on the basis of partial information. Since then, sampling theory has developed separately from the mainstream of classical statistics (Neyman, 1934) and has evolved into an extensive body of theory, method, and operation used daily all over the world. "Sampling Techniques", a landmark book by Cochran (1946, 1977), widely used in modern sampling practices. To deal with two-dimensional spatial sampling, the regionalized variable theory, often referred to as geostatistics, was well built and is widely applied in geoscience (Matheron, 1963, 1971). This approach uses spatial autocorrelation to improve the sampling efficiency in terms of the estimator error variance in relation to sample design and sample size (Ripley, 1981; Haining, 1988; Christakos, 2005; Stein and Ettema, 2003). Spatial stratified heterogeneity was considered to achieve more efficient spatial sampling and inference (Goovaerts, 1997; Li et al., 2008; Wang et al., 2009a, 2010).

1.2. Contents of spatial sampling

Compared with complete enumeration, a statistically based sampling process has great advantages. It leads to reduced costs, higher speed, greater scope and a loss of accuracy that is negligible compared to complete enumeration (Cochran, 1977). Moreover, statistical sampling results in designs that can be used repeatedly, hence leading to small changes in estimates, that in turn provide measures of uncertainty. The aims of spatial sampling methods are to get results of a higher quality at a lower cost. Commonly applied constraints are (a) a cost constraint that the costs do not exceed a given budget and (b) a quality constraint that the quality of the result meets a given minimum requirement.

For a formal description we consider the triple $(\Re, \Im, \psi)$ consisting of a random field $\Re$, a design $\Im$ and an estimator ψ. The triple is also called a sampling strategy trinity. The symbol $\Re$ refers to characteristics of a random field such as the random error, trends, autocorrelation or stratification, which are determinants of the sample estimation efficiency. The design $\Im$ refers to the procedure determining sample selection in terms of density and location, whereas the estimator ψ defines the procedure used to calculate an estimate of a population parameter from the sample data. It is useless

to discuss the merits of a sampling design without considering the estimator to be used and the characteristics of the random field, or vice versa. It is the combination of the three that determines the accuracy of the estimates (Wang et al., 2010).

Spatial sampling and inference are divided into the following six principal steps (Cochran, 1977; Wang et al., 2009b).

(1) Clarification of the sampling objectives to estimate the mean or total for a population, to predict values at unsampled sites, to map an area, or to identify the position of a target. It should also be clear that the objective is about the here-and-now population (Haining, 2003) or about the superpopulation that generates the cross-sectional population.
(2) Definition of the population to be sampled and sampling units. Population donates the aggregate from which the sample is chosen, and the population is then divided into non-overlapping parts called sampling units or elements. In the spatial context, the population from which sample units are drawn and the population to be estimated may be spatially inconsistent; the former is called the sampling population and the latter the estimating population. Note that the two are identical for non-spatial sampling.
(3) Choice of the sampling method and computation of the sample size under specific budget and precision constraints. A correct decision can be made using the sampling design trinity $(\Re, \Im, \psi)$.
(4) Design of a sampling plan that describes where, when and how sampling units are to be taken.
(5) Drawing of the sample and taking measurements on the sampling units obtained. The collection of measurements is commonly called the sample.
(6) Analysis of the sample in light of prior knowledge to obtain the features of the stochastic field, such as randomness, trend, autocorrelation, or stratification. These features and the sampling objectives provide the basis for choosing the best statistic to make inferences for population parameters.

Two theories are commonly distinguished for sampling and inference: design-based and model-based sampling. They differ in their element of randomness in assigning a stochastic structure to inferences (Särndal, 1978; Särndal et al., 1992; de Gruijter and ter Braak, 1990; Haining, 2003). In a spatial context, the "here and now" characteristics of a random field are obtained once the entire area is exhaustively surveyed. The uncertainty of the sample estimation arises from the randomness of the sampling sites, which can be reduced by increasing the sample size until the area is fully covered. Such an approach is called design-based sampling, which can be implemented by traditional sampling (Cochran, 1977) applied to random spatial fields. Alternatively, if a value was once a realization of an underlying stochastic process (a superpopulation) it might not be obtained even if the entire area is exhaustively surveyed. This approach is called model-based sampling (Särndal, 1978) and the methods involved are mainly related to spatial statistics (Cliff and Ord, 1981; Haining, 2003) and geostatistics (Matheron, 1963), where the random field observed is regarded as a single realization of a spatial process or a superpopulation. Some assumptions have to be made and the process model has to be well defined to estimate the superpopulation because only a cross-sectional observation is available, which is often not sufficient to support inference for the superpopulation.

Clearly, the two major distinct objectives of sampling result in two major categories of estimation techniques.

2. Design-based sampling

In design-based sampling, the population of values in the interested region are regarded as unknown but fixed. The primary source of randomness arises by drawing a sample randomly. Repeated sampling according to a given scheme such as random sampling will generate a distribution of sample values, and parameters of the population and their uncertainty are evaluated using this distribution. For example, when calculating weighted averages, the data are assigned weights determined by the selection probabilities for the sampling locations, not by their geographical coordinates. Design-based sampling and statistical methods can be used for both populations (a single realization of a stochastic field) and superpopulations (a stochastic field). As an example of a population, we may consider the spatial disease incidence in 1 year, whereas the disease prevalence,

the cumulative number of cases divided by the susceptible population in a range years reflecting the risk of disease occurrence, is an example of a superpopulation. In the population example, incidence can be mapped as cases in that single year; in the superpopulation example, prevalence can be mapped by averaging the cases in many years.

2.1. Sampling of i.i.d. populations

Classical sampling theory is based on probability sampling. It includes methods for sample selection and estimation that provide, at the lowest costs, estimates that are precise enough for our purpose (Cochran, 1977) and therefore sampling is efficient. Samples can be selected with an equal or an unequal probability. Equal probability sampling is often apparently feasible for sampling in practical surveys. Simple random sampling (SRS) and systematic sampling (SYS) serve as the starting point for an understanding of sampling. In SRS or SYS, sample units are drawn independently from each other with equal probability. It is theoretically a relatively simple approach, but it is rarely used in practical sampling because of its low efficiency. More cost-effective sampling methods include stratified sampling, cluster sampling, multistage sampling, two-phase sampling, and sequential sampling (Neyman, 1934, 1938; Cochran, 1946; Horvitz and Thompson, 1952; Royall, 1968).

(1) *Simple random sampling.*

Although this category is rarely applied in practice because of its low efficiency, all design-based sampling techniques originate from it and can only be fully understood with proper knowledge of SRS. In SRS it is assumed that the population is independent and identically distributed (i.i.d.). The sample size required to reach a prespecified precision is based on the dispersion variance of the population and the survey precision required. Then the sample units are chosen from the population independently with equal probability, and inferences are conducted using the sample. The population mean $\overline{Y}$ is estimated using the sample mean $\overline{y}$, and the variance of the estimate is proportional to the dispersion variance of the sample and inversely proportional to the number of sampling units. The confidence interval for the mean is $\overline{y} \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{v}$ if the population is not only i.i.d. but also $N(0, \sigma^2)$-distributed, where σ^2 is the dispersion variance of the population, v is the variance of $\overline{y}$ and α is the level of significance. SRS is not often applied in spatial studies because it completely neglects prior knowledge for the object and leads to relatively large gaps in the sampled area. Moreover, access to locations is often prohibitive or difficult to find, despite recent developments in global positioning systems.

Horvitz and Thompson (1952) developed a general theory for construction of unbiased estimates. Regardless of the selection probabilities, as long as they are known and positive it is always possible to construct a useful estimate. Horvitz and Thompson completed the classical theory, and random sampling was almost unanimously accepted. Most of the classical books on sampling were also published by then (Cochran, 1946; Hansen et al., 1953).

(2) *Systematic sampling.*

In SYS it is again assumed that the population is i.i.d. Once the first sample unit is drawn from the population, the other sampling units are drawn in a given preset order relative to this first point. SYS is relatively easy to implement. It has the advantage over SRS that the sampling units are spread more evenly over the sampled area and it may produce more precise estimates than SRS. SYS requires the use of systematic sampling zones. The mean $\overline{y}_{sy}$ is the same as in SRS. The variance is $v(\overline{y}_{sy}) = \frac{1}{n}[\frac{N-1}{N}s^2 - \frac{k(n-1)}{N}s_{wsy}^2]$, where s^2 is the dispersion variance of the pooled data, $s_{wsy}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{y}_i)^2$, N is the number of units of the population, n is the sample size, y_{ij} is the value for the jth sample unit in the ith systematic zone, and $\overline{y}_i$ is the mean for the ith systematic zone if the population is not only i.i.d. but also $N(0, \sigma^2)$-distributed and σ^2 is the dispersion variance for the population. The confidence interval for the population mean is $\overline{y}_{st} \pm t\sqrt{v}s$, where t is the t-statistic.

(3) *Stratified random sampling.*

In stratified random sampling we assume that the population is spatially stratified and that the subpopulation within each stratum is i.i.d. The population is divided into subpopulations called strata,

which are non-overlapping and together comprise the whole of the population (Cochran, 1977). SRS or SYS is then applied in each stratum. By dividing the population into subpopulations, the heterogeneity in each stratum is reduced and it is easier to collect samples are highly representative (Wang et al., 2010). The mean for each stratum is computed as $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$, where n_h is the sample size of stratum h and y_{hi} is the value for the ith sample unit in stratum h. The sample mean is $\bar{y}_{st} = \sum_{k=1}^{L} W_k \bar{y}_h$, where L is the number of strata and W_k is the weight for stratum k, which can be computed as the fraction of stratum k within the population. The dispersion variance for each stratum can be computed as $s_h^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$. The variance for the full sample is

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h},$$ where N_h is the number of sampling units in stratum h and the standard deviation is $s = \sqrt{v(\bar{y}_{st})}$. The confidence interval for the sampling mean is $\bar{y}_{st} \pm ts(\bar{y}_{st})$, where t is the t-statistic.

(4) *Cluster sampling.*

For cluster sampling we again assume that the population is i.i.d. In each cluster the first sample unit is selected randomly and the rest are drawn in an order relative to it. The population mean is estimated as $\bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$, where n is the number of clusters and $\bar{y}_i$ is the mean for the ith cluster computed using the same method as in SYS. The variance is $v(\bar{y}_{cl}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{cl})^2$ and the standard deviation is $s = \sqrt{v(\bar{y}_{cl})}$. Note that SYS is a special case of cluster sampling in which several clusters are chosen.

(5) *Two-step random sampling.*

In two-step random sampling we again assume that the population is i.i.d. and that the population is divided into subpopulations (called primary units). The first step is to decide on subpopulations to be sampled, which is usually done using random selection. In the next step, samples are selected, for example using SRS within each selected subpopulation. Two-step random sampling is different from stratified sampling in the sense that only some of the subpopulations are sampled. The population mean is estimated as $\bar{y}_{ts} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$, where n is the number of primary units and $\bar{y}_i$ is the mean for the ith primary unit computed using the same method as in SRS. The variance is $v(\bar{y}_{ts}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{ts})^2$ and the standard deviation is $s = \sqrt{v(\bar{y}_{ts})}$. If the subpopulations are unequal in size, the means and variances are calculated by embedding a weighting factor proportional to the size of the subpopulation.

2.2. Sampling considering spatial autocorrelation

Spatial autocorrelation is ubiquitous in spatial populations. It is one possibility for taking into account the well-known geographical principle that observations close to each other are more likely to be similar than observations further apart. This principle is also known as Tobler's law in geography. Spatial autocorrelation violates the first assumption of population i.i.d., namely that of independence. Recognizing autocorrelation has implications for the efficiency with which sampling is carried out. The variance of the estimation error may, for example, be incorrectly estimated with respect to choices of sampling design and sample size (Ripley, 1981; Haining, 1988; Christakos, 2005). If the objective is to estimate population parameters, then we do not need as many sample sizes as in a population survey in which the sampling units are independent from each other. In fact, $n = \frac{1}{v} [\sigma^2 - R] = \frac{\sigma^2}{v} [1 - \frac{R}{\sigma^2}] = n_{\text{Classic}}[1 - r]$, where $R = \overline{\text{cov}(y_i, y_j)}$ and $r = \frac{R}{\sigma^2}$; v is the error allowed for the spatial mean estimate and is defined by the variance $E(\bar{y}_{ss} - E\bar{y}_{ss})^2$; σ^2 is the dispersion variance of the population and is defined by $E(y - Ey)^2$; y or y_i refers to a sample element. $\bar{y}_{ss}$ is the mean value estimated using a spatial sample; the estimation is the same as for the simple random statistic, but their variances are different as described above. n_{Classic} is the sample size computed using classic sampling theory (Ripley, 1981; Dunn and Harrison, 1993, Griffith et al., 1994; Wang et al., 2009b). If the objective is to estimate the superpopulation, then the sample size should be greater than that for an i.i.d. population, because the variance of the sample estimation is inflated by the spatial autocorrelation (Haining, 1988). On one hand, the number of sampling units can be decreases as the

spatial autocorrelation increases if the aim is to estimate the population mean. On the other hand, more sampling units are needed to estimate a superpopulation mean if sampling is constrained by a required estimation error.

In the case of stratified random sampling, the error variance for estimating the population mean is given by $\frac{1}{n}[\sigma^2 - \overline{\text{cov}(y_{i,}y_j)}]$, where the second term within the square brackets is the average autocovariance between all possible pairs (k, l) within a stratum (Ripley, 1981; Dunn and Harrison, 1993). The sample size in each stratum can also be computed as for spatial SRS. In the case of SYS, the error variance for estimating the population mean is given by $\overline{\text{cov}(y_{v,}y_u)} - \overline{\text{cov}(y_{i,}y_j)}$, where $\overline{\text{cov}(y_{v,}y_u)}$ is the average autocovariance between members of the systematic sample (Ripley, 1981; Dunn and Harrison, 1993).

Early work that evaluated spatial sampling strategies in two dimensions includes contributions from Quenouille (1949), Das (1950), Zubrzycki (1958) and Matern (1960). These theoretical studies were based on random fields that were the outcomes of spatially homogeneous processes. Ripley (1981) showed that SYS outperforms other sampling schemes when no prior knowledge is available about random fields except if there is periodicity in the attribute. These findings have generally been endorsed by empirical studies (Mátern, 1986; Milne, 1959; Payandeh, 1970).

Dunn and Harrison (1993) showed that while SYS was the most efficient of the three methods for sampling (and random sampling consistently the least efficient), the gains in efficiency, relative to stratified random sampling, were highly variable. They also compared two different methods for estimating the error variance for the mean from a single systematic sample (Ripley, 1981). Their work was based on sampling of real land-use maps with complex and varied spatial autocorrelation structures and their findings suggested that the presence of non-stationarity and anisotropy in spatial data could have a severe effect on the SYS efficiency.

Map stratification, whether applied to stratified random sampling or SYS, usually involves the definition of strata and the sampler needs to decide on stratum size as well as the orientation and starting point for the grid overlay. Typically, the starting point is chosen at random and the effect of the starting position on the calculated sample variance is assumed to be small. This is also supposed to be true for the grid orientation, unless there are strong directionalities in the random field, as in the case of a repetitive ridge and valley topography for example. The stronger the spatial autocorrelation on the map, with high levels over long distances, the greater will be the redundancy in a sample if the inter-point sampling distance is small (Griffith, 2005). Ripley (1981) pointed out that the gain from stratification will be high when spatial autocorrelation is large for all distances up to the scale of the strata, beyond which the gain becomes negligible. This suggests that for monotonically decreasing correlation functions we should take small strata and hence the number of sample points per strata will be small (Ripley, 1981). If spatial autocorrelation structures vary in different sections of the map, then presumably there will be efficiency gains to be achieved by adapting stratum size to such heterogeneity (Berry and Baker, 1968).

2.3. Sampling considering spatial heterogeneity

Spatial heterogeneity is another characteristic of natural phenomena and is becoming increasingly important in different areas of geographical data analysis and environmental sciences (Csillag et al., 1996; Goodchild and Haining, 2004; Green and Plotkin, 2007), representing the more practical side of spatial statistics. The heterogeneity of attributes of a geographical random field comprises two elements of second-order variation: global variance (population variance) and the spatial structure of that variation (population spatial autocorrelation) (Wang et al., 2010). Both elements of geographical variation need to be recognized in designing and evaluating sample designs, including the sampling plan and choice of the estimator (Cochran, 1977; Haining, 1988; Griffith, 2005).

Although the pervasiveness and importance of heterogeneity have been recognized and studied, there are only a few methods for dealing with spatial heterogeneity using design-based sampling. Purposive sampling is generally more efficient than probability sampling (Brus and de Gruijter, 1997; de Gruijter et al., 2006). However, purposive sampling cannot satisfy the sampling randomness required by design-based theory. Two of the main methods for dealing with spatial heterogeneity

are spatial stratified sampling in design-based sampling and MSN (means of surfaces with nonhomogeneity) sampling in model-based sampling (Wang et al., 2009a).

In spatial stratified sampling, the heterogeneous area is divided into several subareas that are each more homogeneous than the area as a whole, which reduces the total variance. The division may be based on prior knowledge, pre-sampling, an effective proxy variable (Rodeghiero and Cescatti, 2008), or the distribution of other ancillary variables that are known to affect the target variable. After the division, the number of sampling sites is assigned to each subarea according to the area and/or variance proportion and then SRS is applied to each subarea. Wang et al. (2010) suggested that there are usually no "true" or "correct" divisions of the whole area of interest, but some divisions will be better than others for any given scale of geographical detail. The authors investigated the effect of first-order heterogeneity on the performance of different methods for estimating the population mean and its error variance. They found that zoning improves the estimator efficiency when sampling a heterogeneous random field and compared the efficiency of different stratified methods.

3. Model-based sampling

In model-based sampling, the values in a region of interest are regarded as unfixed and a set of values observed across the region represents a single realization of a stochastic model of the variation in the universe. The randomness in this method is introduced using a set of stochastic models also called superpopulations. The sampling sites are fixed and inference is based on assuming the validity of the stochastic model. To compute an estimator or a predictor, the weights of the sample data are determined by covariances between the observations, which are given by the model as a function of the coordinates of the sampling locations (van Groenigen and Stein, 1998). In contrast to design-based sampling, model-based sampling features a predefined objective function that is minimized and thus leads to a single optimal sampling scheme for that specific objective. Different objective functions may lead to different optimal sampling schemes. A model-based sampling scheme and model-based statistics can be applied to estimation of both population parameters (a single realization of a stochastic field) and superpopulation (a stochastic field) parameters. Kriging, for example, can be considered as a model-based method. The incidences in a single year are considered as a population and a map can be produced by kriging by interpolation of incidences at the nodes of a fine-meshed grid. A prevalence map for a range of years is a superpopulation and can similarly be produced by kriging interpolation of prevalence at the grid nodes.

In this paper we distinguish three specific objectives that are relevant in a spatial context: minimization of the estimation error variance, achieving equal spatial coverage for irregular polygons, and achieving equal coverage in feature space.

3.1. Objective 1: minimizing the estimation error variance

Kriging is the main estimation method in geostatistics and was first used by D.J. Krige and H.S. Sichel in mining theory in the 1950s and 1960s and was further developed and applied by Matheron (1963) as a fully fledged scientific theory. It provides unbiased estimation of values of a variable at unvisited locations with a minimum mean-squared estimation error (Isaaks and Srivastava, 1989; Pardo-Iguzquiza and Dowd, 1998; Olea, 1999; Bayraktar and Turalioglu, 2005; Saito et al., 2005; Wang et al., 2009a). Kriging sets requirements for the random field, assuming, for example, that it is second-order stationary and thus can be represented as a regionalized variable. Furthermore, the random field is potentially autocorrelated in space, and is then modeled using the covariance function or the variogram. Kriging is a generic name and consists of many variants, such as normal kriging, universal kriging, block kriging, co-kriging, indicator kriging, factorial kriging and regression kriging, among others. MSN is another example of a model-based approach that takes both spatial autocorrelation and spatial stratified non-homogeneity into account to minimize the estimation error (Wang et al., 2009a,b).

To minimize the estimation error variance, a sample of a given size should be optimally distributed over space (Hughes and Lettenmaier, 1981; Russo, 1984; Warrick and Myers, 1987; Sacks and Schiller,

1988; van Groenigen and Stein, 1998; Bertolino et al., 1983; Wang et al., 2009a,b). To achieve this purpose, a frequent requirement is that the spatial variation as provided by the variogram is known beforehand. If this is the case, then the optimal model, which is a function of the Kriging variance, and sampling sites are allocated initially at random in a region and are then shifted using an optimization technique. Examples are spatially simulated annealing and particle swarm optimization (Hu and Wang, 2011). The technique has been applied in many areas, such as in soil surveys, environmental pollution monitoring and agricultural modeling (van Groenigen et al., 2000; Zimmerman and Holland, 2005).

3.2. Objective 2: equal spatial coverage

It can be important to have good coverage over a spatial area for several reasons. First, it may serve as a general type of optimality, there may be good reasons for not missing out any subregion, and there could be a good economic argument to have sparse data equally distributed. Moreover, the first objective may not be realistic since a variogram may not be available. As an alternative in this case, spatial coverage distributes sampling sites to cover the space as uniformly as possible over a region. This is analogous to Horvitz–Thompson sampling, which weights a sample by its inclusion probability to cover unsampled units.

The criterion to apply is then minimization of the mean for the shortest distances (MMSD). The target quantity to be minimized is the mean of the distances from unsampled sites to the nearest sampling site. To achieve this, the area is often discretized into a finite grid and the criterion is applied to the center of each grid cell (van Groenigen et al., 1999). An alternative is the weighted mean of the shortest Distances (WMSD), which is the mean of distances multiplied by weights to give varying emphasis to different subregions. Another criterion is the mean squared distance based on Thiessen polygons. The basic idea of this criterion is that if the sampling sites are evenly distributed, the variance of the area constructed by Thiessen polygons should equal zero, whereas the variance should be large for unevenly distributed sampling sites. As before, sampling locations are moved at random to reach the objective. Other methods have been proposed to disperse the distribution of the sample points. A typical example is repulsion simulation, which distributes sampling sites evenly within a given irregular polygon via simulation of the movements of some ideal homogeneous point charges (Chen et al., 2012). For other purposes, grid sampling, transect sampling, sequential sampling and nested sampling are often used in practical applications (Pettitt and McBratney, 1993; Oliver and Webster, 1986).

3.3. Objective 3: equal coverage in feature space

The objective of equal coverage in feature space is to distribute sampling units such that the samples reveal the population distribution as well as possible. If the distribution of a population is unknown, a postulated distribution from experience or ancillary data is used. Two main criteria used in the literature are the Warrick–Myers (WM) and Latin hypercube sampling (LHS) criteria.

The WM criterion optimizes sample locations for variogram estimation (Warrick and Myers, 1987). It approximates the distribution of sampling units such that pairs of the units follow a prespecified distribution as closely as possible. The LHS criterion is a stratified random sampling procedure that provides an efficient way of sampling variables from their multivariate distributions. It fully utilizes ancillary variables to ensure each of the variables is represented in a fully stratified feature space (McKay et al., 1979; Iman and Conover, 1980). In a recent study, a stratified conditional LHS (scLHS) scheme based on the variance quadtree technique (VQT) and LHS using ancillary variables were combined to coordinate both spatial coverage and feature space coverage (Lin et al., 2011).

4. Population or superpopulation?

Originally, design-based and model-based methods were developed to estimate populations and superpopulations, respectively (Cochran, 1977; Särndal, 1978). A major characteristic of design-based

sampling is that it is a combination of probability sampling and design-based inference, whereas model-based sampling is a combination of purposive sampling and interpolation (de Gruijter et al., 2006). Design-based sampling is more suitable for tackling "how much" questions, for example estimation of global quantities, including frequency distribution of the target variable and parameters of this distribution, such as the mean, standard deviation or a quantile. Model-based sampling is more suitable for tackling "where" questions, such as prediction of values for unvisited locations, mapping and estimation of the parameters of the underlying stochastic model. If a medium sample size is affordable, strong autocorrelations exist and a reliable model of the spatial variation is available, model-based sampling should be considered (Haining, 2003; Wang et al., 2010). If parameters of a small domain or within a local area are to be estimated, both methods may be suitable. de Gruijter et al. (2006) pointed out that the suitability of the two methods can be expressed as functions of the spatial resolution at which estimates are required. A decreasing function points to the appropriateness of design-based sampling; in other words, the higher the spatial resolution, the more suitable is design-based sampling. Conversely, the relative suitability of model-based sampling is an increasing function of the spatial resolution of the estimator.

It can easily be determined whether a sampling objective concerns a population or a superpopulation. Population sampling is involved if the attribute to be estimated is about "here and now". Superpopulation sampling is involved if enumeration of the sampling units is a single realization of a stochastic process and parameters of the latter are of interest. A population is one realization of a superpopulation. Table 1 summarizes the objectives of sampling and inference, as well as several methods and examples.

Table 1 lists several aspects of sampling in the spatial domain. In principle, these could also be expanded to the spatiotemporal domain (for monitoring), but we restrict the discussion to spatial sampling as the topic in focus. Table 1 shows that a population might consist of several superpopulations. For example, observations of temperature across a plain on a single day (spatial population) comprise a single realization of the meteorological process (spatial superpopulation) of daily variations within a year (M1 in Table 1); alternatively, the population could consist of temperature variations on that day over many years. Similar remarks apply to disease observations (M3 in Table 1), for which the prevalence could be the human population. The number of cases of a disease within a cross-section (human population) in China is a summation of incidences over previous years that have not recovered yet, and hence there is a combination of prevalence and population. Prevalence, however, could also be a superpopulation. For example, the prevalence of neural tube defects (NTDs) in Heshun county is stable over decades or a hundred years (Gu et al., 2007), reflecting the physical environmental risk. NTDs are rare (low probability) and the incidences for a single or several years (within the human population) do not reflect the baseline risk in the area. Both spatial (design-based) and temporal (model-based) sampling are highly varied. The boundary of the two methods vanishes as time elapses. In the same way, incidence could be either a population or a superpopulation.

Design-based or model-based sampling methods actually provide a full sampling strategy, that is, the components of the triple $(\Re, \Im, \psi)$.

Design-based or model-based ideas also apply to non-sampling aspects and kriging is a good example. For kriging a form of stationarity is assumed, such as second-order stationarity, intrinsic stationarity or strict stationarity. It is assumed that the data originate from a random field and therefore sampling within a kriging context can be categorized as a model-based method (Särndal, 1978; Brus and de Gruijter, 1997). For spatial interpolation using kriging, values at unsampled sites are predicted, and in this sense the approach is closer to a population than to a superpopulation technique. In this sense it is also a design-based method. This demonstrates that model does not need to be categorized as a specific type (design-based or model-based), because that might lead to unnecessary logical confusion.

This can also affect the sampling strategy. It is best to first judge whether the objective of using a technique concerns a population or a superpopulation. Next, appropriate models should be identified that match the objective and a strategy on how to reach the objective should be decided on. For example, both the incidence and prevalence could be considered from a population and a superpopulation point of view. Before deciding on where to sample, samplers must decide which

Table 1
Population and superpopulation.

Objective		Population (here-and-now distribution, single realization of a process or superpopulation)		Superpopulation (process that generates a time series of spatial distributions)	
		Areal mean	Distribution	Areal mean	Distribution
M1	Assumption	i.i.d.	i.i.d.	i.i.d.	i.i.d.
	Model	Cochran (1977)	Cochran (1977)	Sum of i.i.d.	Sum of i.i.d. at each site
	Examples	Average temp. on 1 June 2012 in a plain area	Map of temp. on 1 June 2012 in a plain area	Daily average temp. in 2012 in a plain area; Yearly average temp. on 1 June in North China	Maps of daily temp. in 2012 in a plain area; Maps of yearly temp. on 1 June in a plain area
M2	Assumption	2nd-order stationary	2nd-order stationary	Slow	Slow
	Model	Block kriging with spatial semivariogram	Kriging with spatial semivariogram	Block kriging with S–T semivariogram	Kriging with S–T semivariogram
	Examples	Average temp. in 2012 in a plain area	Map of annual temp. in 2012 in a plain area	Yearly annual temp. in a plain area	Maps of yearly annual temp. in a plain area; Maps of daily temp. in 2012 in a plain area
M3	Assumption	Strata and autocorrelation	Strata and autocorrelation	Temporal process[a]; Time series	ST process[b]; BME (Christakos, 1992)[e]
	Model	MSN Wang et al. (2009a)[c]	Sandwich Wang et al. (2002)[d]	Yearly net accumulation of antibiotic in soil; Poisson process	Map of yearly net accumulation of antibiotic in soil; Maps of determinants; Map of prevalence
	Examples	Soil antibiotic residue in a town surveyed in 2011 (Xie et al., 2012); Incidence of NTD[f] in 2005 in Heshun county	Map of soil antibiotic residue in a town surveyed in 2011; Map of NTD incidence in 2005 in Heshun county; Cross-sectional observation of all cases	NTDs induced by physical environment	Map of NTDs induced by physical environment
⋮	Assumption	⋯			
	Model	⋯			
Method		Design-based (enumeration or statistics) or model-based (spatial process)		Model-based (spatiotemporal process) or design-based (summation over long term)	

Notes:

[a] Temporal process: the attribute changes over time.

[b] ST process: the attribute changes in both spatial and temporal dimensions.

[c] MSN: a model to estimate the mean of stratified nonhomogeneous surfaces (Wang et al., 2009a,b).

[d] Sandwich: a model for multi-unit reporting of a stratified spatial attribute (Wang et al., 2002),

[e] BME: Bayesian maximum entropy, a model for spatiotemporal interpolation (Christakos, 1992).

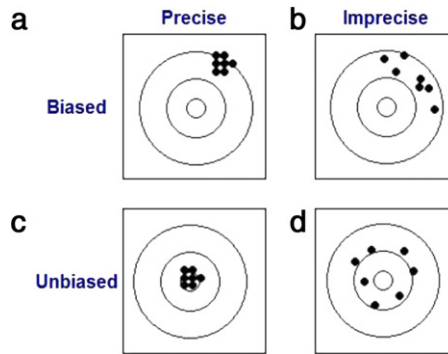[f] NTD: neural tube defect (Gu et al., 2007).

Fig. 1. Biased sampling.

sampling strategy is to be adopted. The appropriate strategy then depends on the objective and relevant constraints.

5. Biased sampling and remedy

An estimate of a parameter is biased if its expectation differs from the true value of the parameter, whatever the true value is, and the estimate is unbiased if its expectation is identical to the true value. The expectation is a value for a population or superpopulation parameter, such as its total or mean or the value of an individual unsampled unit. An estimate $\hat{y}$ is obtained by applying a statistic ψ to a sample $\boldsymbol{y}$:

$$\hat{y} = \psi(\mathbf{y}).$$

Therefore, the bias of an estimate $\hat{y}$ depends on the interaction between the sample $\boldsymbol{y}$ and the statistic ψ. A sample $\boldsymbol{y}$ is often said to be unbiased if it is drawn randomly (thus each sample has an equal probability of being obtained) from its population. However, this would lead to a conflict, as elaborated in the following section. Therefore, a sample is unbiased if it results in an estimate unbiased to the population under a simple random statistic, a so-called SRS unbiased sample. An unbiased estimation can be achieved using either a sample randomly drawn (with equal probability) from its population or a sample biased to its population under SRS or remedied by a particular statistic.

For Fig. 1 it is assumed that the population is the area within the outermost circle, the population parameter to be estimated is the innermost circle, and four samples from the population are investigated, with the sampling units indicated by points. An SRS unbiased estimate can be obtained either by drawing randomly from a population (Fig. 1d), or by non-randomly drawing (unequal probability) from a population (Fig. 1c). A biased sample, however, does not necessarily lead to a biased estimate, because several statistical approaches, such as the Heckman method (Heckman, 1979) and the B-shade model (Wang et al., 2011), can remedy the bias of a sample in some circumstances to yield an unbiased estimate. In other words, an estimate $\psi(\mathbf{y})$ may be biased under SRS, but may be unbiased under some other statistical approaches, such as the Heckman method and the B-shade method. Fig. 1a,b shows a biased sample with low and high variations, whereas Fig. 1c,d shows unbiased sampling of low and high variation. Fig. 1c is an SRS unbiased sample while the sample is biased to population. The above discussion is summarized in Table 2.

6. Discussion and conclusion

The efficiency of spatial sampling can be increased by embedding prior knowledge about random fields. From this perspective, spatial sampling theories have developed along the following path: sampling a population that is i.i.d. (Cochran, 1977), spatially autocorrelated (Haining, 2003; Stein and Ettema, 2003; Christakos, 2005), spatially heterogeneous, and spatially autocorrelated (Goovaerts,

Table 2
Combination of sampling and inference for Fig. 1.

Bias of an estimate			Statistical method	
			SRS	Remedy
Sample	Biased to both population and SRS[*]: Fig. 1a,b		A. Biased	B. Unbiased
	Unbiased to SRS	Biased to population: Fig. 1c	C. Unbiased	D. N/A
		Unbiased to population: Fig. 1d		

[*] SRS, simple random sampling.

1997; Wang et al., 2009b, 2010). The properties of the latest methods need to be studied in more detail; the relationship between distinct random fields, sampling and inference methods, and the estimation precision deserve to be investigated thoroughly. Although the division between design- and model-based methods is relevant, some methods are difficult to categorize into either of these two classes. In fact, population and superpopulation approaches are easily justified, and a research question can be defined flexibly. Finally, spatial sampling and inference methods can be quickly and appropriately chosen from a broad spectrum of methods and techniques available, taking all the relevant constraints into consideration.

The uncertainty of the final estimation originates, propagates and accumulates in the trinity of spatial sampling and statistical inference. This consists of (1) the stochastic field $\Re$, defined by its distribution such as variability aspects; (2) the spatial sampling $\Im$, including its unbiasedness and density; and (3) the spatial statistic ψ for statistical inference. Note that the bias of $\Im$ may be overcome by proper choice of ψ. In practice, the determinants X of the sampling attribute may be helpful in reducing uncertainty, especially when the sample is small or biased. (4) Determinants X: when the original sample is small a large sample of direct determinants of the original sampling attribute or proxy variables strongly associated with the direct determinants can be used to estimate parameters in spatial statistics, such as stratification, auto- and cross- correlation, and ratios between sample and population values. This was illustrated by a small sample survey of occupational diseases in China in which data on the number and size of coalmining factories in each county in China, available from the statistical year book published by the local government, served as a large sample of determinants. In addition, production data for the mining industry in each county were available from the statistical year book published by the central government and served as proxy variables.

This brings us finally to the role of a spatial statistician. S/he may contribute to implementation of a spatial statistic ψ if the sample has already been collected, or to both spatial sampling $\Im$ and the spatial statistic ψ if the data (at least in part) still have to be collected. Disciplinary scientists such as epidemiologists and other experts in the domain represented by the field $\Re$ could be responsible for choosing determinants X mostly related to the sampling attribute.

Acknowledgments

References

Bayraktar, H., Turalioglu, F.S., 2005. A kriging-based approach for locating a sampling site in the assessment of air quality. Stochastic Environmental Research and Risk Assessment 19, 301–305.

Berry, B.L., Baker, A.M., 1968. Geographic sampling. In: Berry, B.L., Marbles, D.F. (Eds.), Spatial Analysis. Prentice-Hall, Englewood Cliffs, NJ, pp. 91–100.

Bertolino, F., Luciano, A., Racugno, W., 1983. Some aspects of detection networks optimization with the kriging procedure. Metron 41, 91–107.

Brus, D.J., 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, 1–44.

Chen, B.S., Pan, Y.C., Wang, J.H., 2012. Even sampling designs generation by charges repulsion simulation. Environmental Monitoring and Assessment 184, 3545–3556.

Christakos, G., 1992. Random Field Models in Earth Sciences. Academic Press, San Diego, CA, 474p.

Christakos, G., 2005. Random Field Models in Earth Sciences. Dover Publications, New York.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London.

Cochran, W.G., 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals of Mathematical Statistics 17, 164–177.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. Wiley, New York.

Csillag, F., Kertesz, M., Kummert, A., 1996. Sampling and mapping of heterogeneous surfaces: multi-resolution tiling adjusted to spatial variability. International Journal of Geographical Information Systems 10, 851–875.

Das, A.C., 1950. Two-dimensional systematic sampling and the associated stratified and random sampling. Sankhyā 10, 95–108.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag, New York.

de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Mathematical Geology 22, 407–415.

Dunn, R., Harrison, A.R., 1993. Two-dimensional systematic sampling of land use. Applied Statistics 42, 585–601.

Goodchild, M.F., Haining, R.P., 2004. GIS and spatial data analysis: converging perspectives. Papers in Regional Science 83, 363–385.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.

Green, J.L., Plotkin, J.B., 2007. A statistical theory for sampling species abundances. Ecology Letters 10, 1037–1045.

Griffith, D.A., 2005. Effective geographic sample size in the presence of spatial autocorrelation. Annals of the Association of American Geographers 95, 740–760.

Griffith, D.A., Haining, R.P., Arbia, G., 1994. Heterogeneity of attribute sampling error in spatial data sets. Geographical Analysis 26 (4), 300–320.

Gu, X., Lin, L.M., Zheng, X.Y., Zhang, T., Song, X.M., Wang, J.F., Li, X.H., Li, P.Z., Chen, G., Wu, J.L., Wu, L.H., Liu, J.F., 2007. High prevalence of NTDs in Shanxi province: a combined epidemiological approach. Birth Defects Research, Part A—Clinical and Molecular Teratology 79, 702–707.

Haining, R.P., 1988. Estimating spatial means with an application to remote sensing data. Communications in Statistics—Theory and Methods 17, 537–597.

Haining, R.P., 2003. Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge.

Hansen, M.H., Hurwitz, W.N., Madow, W.G., 1953. Sample Survey Methods and Theory. Wiley, New York.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Hughes, J.P., Lettenmaier, D.P., 1981. Data requirements for kriging: estimation and network design. Water Resources Research 17, 1641–1650.

Hu, M.G., Wang, J.F., 2011. A spatial sampling optimization package using MSN theory. Environmental Modelling & Software 26, 546–548.

Iman, R.L., Conover, W.J., 1980. Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. Communications in Statistics A9, 1749–1874.

Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, New York.

Lin, Y.P., Chu, H.J., Huang, Y.L., 2011. Monitoring and identification of spatiotemporal landscape changes in multiple remote sensing images by using a stratified conditional Latin hypercube sampling approach and geostatistical simulation. Environmental Monitoring and Assessment 177, 353–373.

Li, L.F., Wang, J.F., Cao, Z.D., Feng, X.L., Zhang, L.L., Zhong, E.S., 2008. An information-fusion method to regionalize spatial heterogeneity for improving the accuracy of spatial sampling estimation. Stochastic Environmental Research and Risk Assessment 22, 689–704.

Matern, B., 1960. Spatial Variation, 2nd ed. In: Lecture Notes in Statistics, vol. 36. Springer, Berlin.

Mátern, B., 1986. Spatial Variation. In: Lecture Notes in Statistics, vol. 36. Springer-Verlag, Berlin.

Matheron, G., 1963. Principles of geostatistics. Economic Geology 58, 1246–1266.

Matheron, G., 1971. The Theory of Regionalized Variables and its Application. Ecole Nationale Supérieure des Mines de Paris, 211p.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Milne, A., 1959. The centric systematic area-sample treated as a random sample. Biometrics 15, 270–297.

Neyman, J., 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97, 558–606.

Neyman, J., 1938. Contribution to the theory of sampling human populations. Journal of the American Statistical Association 33, 101–106.

Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists. Kluwer, Boston, MA.

Oliver, M.A., Webster, R., 1986. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. Geographical Analysis 18, 227–242.

Pardo-Iguzquiza, E., Dowd, P.A., 1998. The second-order stationary universal kriging model revisited. Mathematical Geology 30, 347–378.

Payandeh, B., 1970. Relative efficiency of two-dimensional systematic sampling. Forest Science 16, 271–276.

Pettitt, A.N., McBratney, A.B., 1993. Sampling designs for estimating spatial variance components. Applied Statistics 42, 185–209.

Quenouille, M.H., 1949. Problems in plane sampling. Annals of Mathematical Statistics 20, 355–375.

Ripley, B., 1981. Spatial Statistics. Wiley, New York.

Rodeghiero, M., Cescatti, A., 2008. Spatial variability and optimal sampling strategy of soil respiration. Forest Ecology and Management 255, 106–112.

Rogerson, P.A., Delmelle, E., Batta, R., Akella, M., Blatt, A., Wilson, G., 2004. Optimal sampling design for variables with varying spatial importance. Geographical Analysis 36, 177–194.

Royall, R.M., 1968. An old approach to finite population sampling theory. Journal of the American Statistical Association 63, 1269–1279.

Russo, D., 1984. Design of an optimal sampling network for estimating the variogram. Soil Science Society of America Journal 48, 708–716.

Sacks, J., Schiller, S., 1988. Spatial designs. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics IV, Vol. 2. Springer Verlag, New York, pp. 385–399.

Saito, H., McKenna, S.A., Zimmerman, D.A., et al., 2005. Geostatistical interpolation of object counts collected from multiple strip transects: ordinary kriging versus finite domain kriging. Stochastic Environmental Research and Risk Assessment 19, 71–85.

Särndal, C.E., 1978. Design-based and model-based inference in survey sampling. Scandinavian Journal of Statistics 5, 27–52.

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

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agriculture, Ecosystems and Environment 94, 31–47.

van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27, 1078–1086.

van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

van Groenigen, J.W., Pieters, G., Stein, A., 2000. Optimizing spatial sampling for multivariate contamination in urban areas. Environmetrics 11, 227–244.

Wang, J.F., Liu, J.Y., Zhuang, D.F., Ge, Y., 2002. Spatial sampling design for monitoring the area of cultivated land. International Journal of Remote Sensing 23 (2), 263–284.

Wang, J.F., Christakos, G., Hu, M.G., 2009a. Modeling spatial means of surfaces with stratified nonhomogeneity. IEEE Transactions on Geoscience and Remote Sensing 47, 4167–4174.

Wang, J.F., Jiang, C.S., Li, L.F., Hu, M.G., 2009b. Spatial Sampling and Inferences. Science Press, Beijing.

Wang, J.F., Haining, R.P., Cao, Z.D., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. International Journal of Geographical Information Science 24, 523–543.

Wang, J.F., Reis, B.Y., Hu, M.G., Christakos, G., Yang, W.Z., Sun, Q., Li, Z.J., Li, X.Z., Lai, S.J., Chen, H.Y., Wang, D.C., 2011. Area disease estimation based on sentinel hospital records. PLoS One 6, e23428.

Warrick, A.W., Myers, D.E., 1987. Optimization of sampling locations for variogram calculations. Water Resources Research 23, 496–500.

Xie, Y.F., Li, X.W., Wang, J.F., Christakos, G., Hu, M.G., An, L.H., Li, F.S., 2012. Spatial estimation of antibiotic residues in surface soils in a typical intensive vegetable cultivation area in China. Science of the Total Environment 430, 126–131.

Zimmerman, D.L., Holland, D.M., 2005. Complementary co-kriging: spatial prediction using data combined from several environmental monitoring networks. Environmetrics 16, 219–234.

Zubrzycki, S., 1958. Remarks on random, stratified and systematic sampling on a plane. Colloquium Mathematicum 6, 251–264.