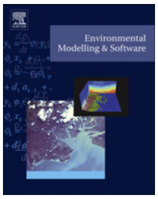
Environmental Modelling & Software 40 (2013) 280–288

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft



Design-based spatial sampling: Theory and implementation

Jin-Feng Wang [a,*], Cheng-Sheng Jiang [a], Mao-Gui Hu [a], Zhi-Dong Cao [a,b], Yan-Sha Guo [a], Lian-Fa Li [a], Tie-Jun Liu [a], Bin Meng [a]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, PR China
[b] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 11 April 2012
Received in revised form
29 August 2012
Accepted 29 September 2012
Available online 13 November 2012

Keywords:
Spatial sampling
Sampling trinity relationship
Prior knowledge
Design-based sampling

ABSTRACT

Various sampling techniques are widely used in environmental, social and resource surveys. Spatial sampling techniques are more efficient than conventional sampling when surveying spatially distributed targets such as CO_2 emissions, soil pollution, a population distribution, disaster distribution, and disease incidence, where spatial autocorrelation and heterogeneity are prevalent. However, despite decades of development in theory and practice, there are few computer programs for spatial sampling. We investigated the three-fold relationship between targets, sampling strategies and statistical methods in spatial contexture. Accordingly, the information flow of the spatial sampling process was reconstructed and optimized. SSSampling, a computer program for design-based spatial sampling, has been developed from the theoretical basis. Three typical applications of the software, namely sampling design, optimal statistical inference and precision assessment, are demonstrated as case studies.

Software availability

Name: SSSampling
Hardware requirement: Windows-compatible PC
Program language: Visual C++
License type: free
Availability information: http://www.sssampling.org

1. Introduction

Spatial sampling and statistical inference are becoming fundamental elements of surveys in broad physical and social disciplines, including surveys of soil (Webster, 1985), ecology (Müller et al., 2012), atmospheric pollutants (Pozo et al., 2006), population health (Kumar, 2007), remote sensing (Stein et al., 1999), etc. Spatial sampling uses a smaller sample to make a more precise estimation relative to conventional sampling (Cochran, 1977), by taking spatial autocorrelation (Haining, 2003) and spatial heterogeneity (Wang et al., 2009, 2010) into account. In the next decade or so, we should see great advances in real-time environmental monitoring technologies. Spatial sampling techniques are crucial in this regard, particularly with respect to the design of monitoring networks, making inferences based on the observed sample, and assessing the posterior precision of the estimate. Compared with exhaustive surveys, sampling techniques have the advantage of being quicker, cheaper, and more precise (Cochran, 1977). Given a limited budget for a survey, higher precision can be attained by locating people who are more experienced and using specific instruments at appropriate sampling sites, rather than having people who are less experienced and inadequate instruments at all sites.

Sampling techniques evolved centuries ago from probability and statistics. In recent decades, characteristics of spatially referenced phenomena have been recognized and they have stimulated the development of spatial statistics and sampling methodology. There is a vast literature on spatial sampling techniques, which can be roughly divided as design based (e.g., Cochran, 1946; Rodriguez-Iturbe and Mejia, 1974; Bellhouse, 1977; Matérn, 1986; Haining, 1988; de Gruijter and Ter Braak, 1990; Overton and Stehman, 1993; Opsomer and Nusser, 1999; Stein and Ettema, 2003; Stevens and Olsen, 2004; Rogerson et al., 2004; Gallego, 2005; de Gruijter et al., 2006; Lister and Scott, 2008; Wang et al., 2010), model based (e.g., Olea, 1984; Cressie, 1991; Christakos, 1992; Olken and Rotem, 1995; Caeiro et al., 2003; Wang et al., 2009; Hu and Wang, 2011; Spöck, 2012), and both (for example, Griffith, 2005). The choice of the distinct approaches should be based on the objective of the survey (Haining, 2003; de Gruijter and Ter Braak, 1990). The model-based approach acknowledges that the observed population is one realization of a probability process and

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

1364-8152/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2012.09.015

aims at estimating the parameters underpinning the process, or a superpopulation; the design-based approach acknowledges that the value is fixed at each sampling location and aims at estimating the observed (here and now) population using a sample. A practical guide to distinguish a population and superpopulation is as follows. If users want an enumerated survey result then a sampling to this end relates to a population; if an enumerated survey was only one realization of a process then a sampling to estimate the process relates to a superpopulation. For example, birth defects are low-probability events, and a cross-sectional survey over an area relates to the population, which can be estimated using a design-based approach; i.e., conventional sampling (Cochran, 1977). In contrast, a long-term time series of the spatial distribution of a disease is a superpopulation of the disease, which can be estimated using a model-based approach with some assumptions of the spatiotemporal process of the birth defects or be estimated using a design-based approach with a long-term cohort survey. Although there has been great progress in the development of spatial sampling theories, there is little open computer software for spatial sampling (Spöck, 2012), because prior knowledge, spatial autocorrelation, and spatial heterogeneity are not easily implementable in software. Thus, developing software for this purpose is seen solution way to promote the use of these sophisticated techniques.

In this study, we develop software for design-based spatial sampling. We clarify the tasks involved in spatial sampling in the real world and review existing software in Section 2. In Section 3, we summarize the mechanics of spatial sampling. Accordingly, in Section 4, we design a computer program for design-based spatial sampling. In Section 5, we demonstrate three typical applications of the software, namely distributing a sample optimally over space; making an optimal inference using an existing sample; and assessing the precision of an existing statistical report. Finally, conclusions are drawn in Section 6.

2. Sampling surveys in the real world and existing software

An example of a question that arises during sampling in the real world is as follows. To achieve relative error less than 20%, how many villages, and which villages, should be drawn from the 326 villages in a county to estimate the proportion of birth defects in live births? Its dual question is, given a budget for the survey or a cap on the number of villages, which villages should be drawn from the 326 villages and how precise can the estimate be?

The key idea of a spatial sampling method is to infer the properties of a population using a sample that is distributed over space using a suitable statistic. The resulting estimate of a population could be its total, mean value (Haining, 2003; Griffith, 2005; Wang et al., 2009), values at unsampled sites (Spöck, 2012) or spatial maxima (Rogerson, 2005), spatial patterns (Dungan et al., 2002), statistical hypotheses (Stein and Ettema, 2003), semi-variograms, or the precision together with its confidence interval of the estimates. The theory of spatial sampling addresses the following dual tasks.

- For a given precision, with the confidence interval of the estimate, project the number of sample units or the budget of the survey to meet the precision requested. This is conditional upon the properties of the target domains and prior information available.
- For a given number of sample units or the budget of a survey, forecast the precision of an estimate and its confidence interval. Again, this is conditional upon the properties of the target domains and prior information available.

Although a wide range of sampling techniques have been developed (Cochran, 1977; Li et al., 2005), it is almost the case that only random sampling is implemented in open computer packages (Lwange and Lemeshow, 1991). For instance, G*Power, Macorr, PASS, Raosoft, and nQuery Advisor are sampling packages that deal with issues such as power values for given sample sizes, effect sizes, and alpha levels (post hoc power analyses); sample sizes for given effect sizes, alpha levels, and power values (*a* priori power analyses); and alpha and beta values for given sample sizes, effect sizes, and beta/alpha ratios (compromise power analyses). Spöck (2012) recently developed spatial sampling software based on a spectral model to reduce kriging variance.

Spatial autocorrelation and heterogeneity, usually inherent in spatial data, can seriously impede the efficiency of conventional sampling techniques (Cochran, 1977; Haining, 1988; Griffith, 2005) and should therefore be implemented in spatial sampling software (e.g., van Groenigen and van Stein, 2000). In addition, mapping is a necessary function in a package handling spatial data. The software SPSS allows users to choose a sample from a given population framework, randomly, systematically or in a stratified manner. The sampling handbook of the World Health Organization (Lwange and Lemeshow, 1991) greatly facilitates field surveys by epidemiologists. However, spatial autocorrelation and spatial stratification are difficult to account for in conventional sampling (Cochran, 1977), if not impossible. Flexibility, robustness, and a user-friendly interface are critical qualities needed for the success of a sophisticated package. We consider all of the above requirements in developing our geographical information system (GIS)-based and design-based spatial sampling and statistic software, SSSampling, an open and freely downloadable package (www.sssampling.org).

3. Mechanics of spatial sampling

Spatial sampling is to sample a target population, which involves drawing a number of sample units from the geographically distributed target, and then using the sample to infer the properties of the target. The performance of a sampling survey is measured by both the variance (v) of the sample estimate and the number (n) of sample units used, denoted as (v, n). More intense sampling gives a better reconstruction of the variable of interest, but is expensive, time-consuming and sometimes redundant. Conversely, although sparse sampling is cheap, it may miss important features. A good sampling survey has a small variance of the estimate using a small sample, considering the budget for sampling or required precision of the estimate.

3.1. Trinity relationship among the target domain, sampling frame and statistics

The performance of a sampling (v, n) is controlled by the trinity relationship $\Re$, $\Im$, Ψ of the target domain with its features $\Re$, geographical distribution of a sample $\Im$, and the statistical method Ψ (i.e., the model used to calculate the mean and variance of samples) (Wang et al. 2010). The target domain $\Re$ may or may not be identical to the study area Ω. For example, in surveying the human population in China, the country is the study area Ω, while the places that humans inhabit makes up the target domain $\Re$. In another example of mapping or estimating the annual mean air temperature in China, the whole geographical territory of the country is the study area Ω and is identical to the target domain $\Re$ because the target $\Re$ covers the whole country Ω. The features of a target domain $\Re$ could be identified independent distribution or i.i.d., dispersion variance, spatial autocorrelation, spatial heterogeneity, trend, and periodicity; sampling $\Im$ = random sampling, systematic sampling, and stratified sampling; statistic Ψ = random or spatial random statistic, systematic or spatial systematic statistic, and stratified or spatial stratified statistic, where the term ‘spatial random statistic’ refers to a random statistic considering spatial autocorrelation. The element(s) of the sampling trinity may be modified if sites of the target domain $\Re$ have different importance (Rogerson et al. 2004) or sample units $\Im$ have various sizes (Journel and Huijbregts, 1978), and these will be transferred into sampling practice through the sampling trinity ($\Re$, $\Im$, Ψ). For example, a seismic surveillance network should be much denser in population-dense areas, rather than being distributed purely considering the physical activity of the local crust.

Small variance and a lack of unbiased are pursued in a spatial sampling and estimation by optimizing the sampling trinity relationship. For example, for a stratified target domain ($\Re$ = spatial heterogeneity), applying stratified sampling ($\Im$ = stratified sampling) and then using a stratified statistic (Ψ = stratified statistic) will result in a sample estimate with small variance (Cochran, 1977; Wang et al., 2010). i.i.d. $\Re$, any sampling $\Im$ and statistic Ψ will result in an unbiased estimate (Cochran, 1977); for a stratified target domain $\Re$, a random sampling $\Im$ with random statistic Ψ will still result in an unbiased estimate (Horvitz and Thompson, 1952) but its variance can be reduced if stratified sampling $\Im$ and/or stratified statistic Ψ were employed; for a stratified target domain $\Re$ and a sample $\Im$ biased with respect to $\Re$ (a sample not randomly drawn from the target domain), employing a random statistic Ψ will result in a biased estimate but the sample bias $\Im$ may be remedied and an unbiased estimate reached using specific statistics Ψ (Heckman, 1979; Wang et al., 2011); a sample $\Im$ biased with respect to the target domain $\Re$ cannot result in a biased estimate under a simple random statistic Ψ. A random statistic (Ψ) of the human population may be biased if the locations of the sample ($\Im$) are chosen randomly across a territory in a manner that is inconsistent with the geographical distribution of the target population ($\Re$) (Schwanghart et al., 2008). The bias estimate arises from the spatial heterogeneity of the surveyed population $\Re$, a sample $\Im$ random to the territory but nonrandom to the geographical distribution of target human population, and a random statistic Ψ; an estimate can be unbiased either using a sample $\Im$ drawn randomly for the geographical distribution of human density (target domain $\Re$) with a random statistic Ψ, or using a sample $\Im$ drawn randomly for the territory (study area Ω) but followed by the use of a statistic Ψ weighted by population.

Although Cochran (1946) and Haining (1988) explicitly considered spatial autocorrelation and the probability distribution of the target domain in deducing spatial sampling models, there have been few systematic studies on the role of the various heterogeneities of surfaces in sampling design. This is despite target domain characteristics seriously affecting the efficiency of sampling design and statistical inference (Lin et al., 2008; Wang et al., 2010).

3.2. Multi-unit reporting problem

Often, there is a need to simultaneously estimate the values of an attribute in multiple reporting units (i.e., domains (Sarndal et al., 1992) or subpopulations (Cochran, 1977)) in the existing literature. We use the term ‘reporting units’ here because it is easily understood. For instance, central government may want to know the current population size in each of the 2700 counties in China, or the organism concentration in soil in each cell of a fine-grid system. In both instances, an estimate can be obtained under the condition that at least two sample units are drawn in each of the many reporting units, according to design-based framework. Therefore, the total cost of the survey is proportional to the number of reporting units, and the survey would be expensive when there are many reporting units. Alternatively, a model-based approach such as kriging can make a projection but fails if the surfaces are stratified non-homogeneously, which is often the case in the real world (Goodchild and Haining, 2004).

A design-based statistic called the sandwich statistic has been developed for the areal interpolation of a heterogeneously stratified surface (Wang et al. 2002), where the reporting units and target domain are completely independent of each other, and are viewed as two totally separate layers. The reporting layer could be a census system, physical units such as watersheds, or an artificial delimited grid system. This layer is subjectively defined. Stratified heterogeneous surfaces can be reflected by zonation constructed from prior knowledge of the area or physical laws (Wang et al., 1997, 2010). Stratified sampling is conducted over the zonation layer. In this way, the sample size is completely independent of the number of reporting units. An information propagation function chain has been established to model the propagation of information from the target domain layer, to the zonation layer, to the sample layer and finally to the reporting unit layer. This is accompanied by a propagation of uncertainty of the estimate (Wang et al. 2002).

The advantage of the sandwich statistic can be readily understood in an extreme situation: a target domain is perfectly stratified into several flat patches, a number of grid units over the surface have to be reported, and the number of strata of the target domain is much less than the number of reporting units. In this case, stratified sampling in a small area relative to the zoned target surface will be sufficient to estimate the mean value and variance for each of the zones. The information is then transferred by the sandwich model onto each of the multi reporting units with high precision, while in conventional design-based approaches, each of the reporting cells has to contain at least two sampling points (Cochran, 1977; Sarndal et al., 1992; Rao, 2003) and a large sample has to be drawn for multi-unit reporting. Sandwich sampling reduces to stratified sampling if there is only one reporting unit, and the efficiency of the sandwich reduces if surfaces are stratified less heterogeneously.

3.3. Prior knowledge and implementation

The uncertainty of a spatial sampling estimate is proportional to the gap between the human recognized features of the target domain and the real features of the target domain (Wang et al., 2010; Bueso et al., 2005). Prior knowledge of the target domain can dramatically reduce the uncertainty in the sampling estimate. No structure can be detected if the structure is smaller than the interval between sampling units or larger than the extent of the study (Dungan et al., 2002), and efficient sampling should thus consider prior knowledge.

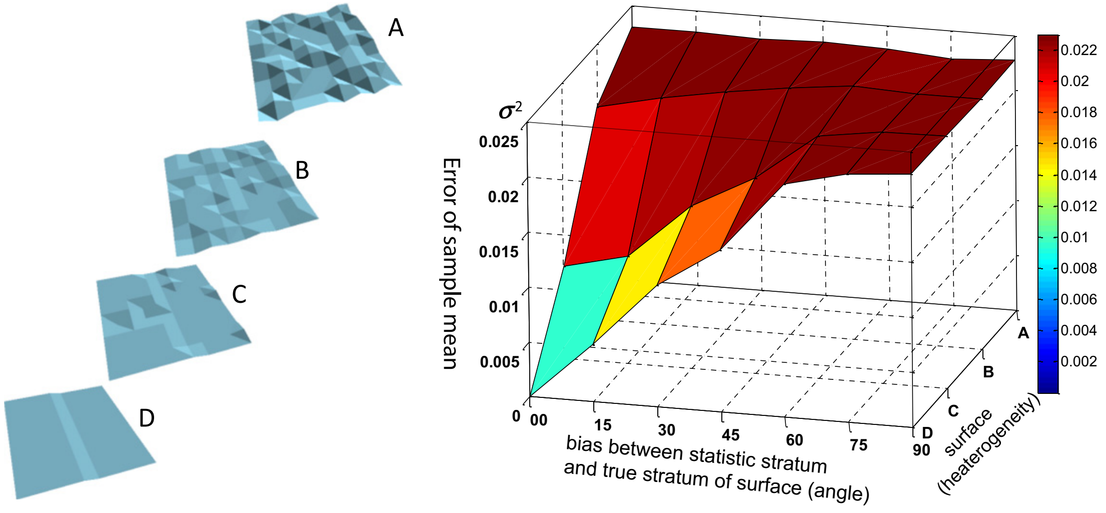
The spatial features of a target domain that affect the efficiency of the sampling estimate are dispersion variance (Cochran, 1977) (the variance within a population, not the variance of a sample mean), spatial autocorrelation (Cochran, 1946; Rodriguez-Iturbe and Mejia, 1974; Haining, 1988; Griffith, 2005), and spatial heterogeneity (Wang et al., 1997). Knowledge of these population properties could be in the form of maps, semi-variograms, zonations, or even stochastic field models (e.g., SAR, MAR, CAR) (Fischer and Wang, 2011) and physical laws (Christakos, 2010). Additionally, process models (Christakos, 1992; Paola et al., 2006) can potentially be integrated into the sampling model to allow the two disciplines of sampling modeling and process modeling to add value to one another (Sarndal et al., 1992).

Prior knowledge contributes to sampling design in three ways. The first is that the target domain features determine the optimal sample geographical distribution and choice of statistic, according to the trinity relationship $\Re$, $\Im$, Ψ described in Section 3.1. The second is that prior knowledge of the target domain is beneficial in delimiting the stratum for sampling and statistics so as to improve the efficiency of sampling. Finally, known and unknown target domain features lead to different choices in sampling methods. For instance, we can only choose simple random or systematic sampling if we know nothing about the target domain, although this is at the risk of low sampling efficiency when the target domain has heterogeneous stratification. Alternatively, we can use spatial stratified sampling and statistics if we know the dispersion variance, spatial heterogeneity and spatial autocorrelation of the target domain, which will be accompanied by a lesser loss of precision of the estimate. Therefore, knowledge of the target domain is developed as an independent module (Li et al., 2008), which formulates the prior information relevant to spatial sampling (Wang et al., 2002).

Prior knowledge can be obtained from general knowledge of physical or human processes; from previous surveys or expert knowledge in the same or similar areas; or from observed determinants of the target domain. The prior knowledge is either transformed into strata on a map by a classification algorithm and then drawn by GIS mapping or transformed into strata on a map by hand using the synthetic and qualitative knowledge of experts. Several specific cases are considered as follows. (1) If there is neither historical data about the surface nor prior knowledge or experience available for zoning, administrative units or physical units are sometimes used as zones for sampling. No specialist computation is required. This sampling suffers from low efficiency if the region has different geophysical, geographical or socioeconomic environments. (2) If detailed data are not available but there is relevant prior knowledge or experience, the prior knowledge or experience may be sufficient to produce useful zoning, drawn up by a panel of experts. For example epidemiologists may be able to suggest ways of partitioning a city or region into areas with similar health risks, on the basis of their work experience and knowledge of race, occupation, income, age, and environment of a residential area. (3) If we have relevant data (e.g., in the form of a pilot survey or historical data) but uncertain contemporary expert knowledge, k-means or shared-nearest-neighbor clustering can be used to construct the zones when sufficient data are available covering the area. The quality of any zoning is usually improved by integrating diverse sources of information (Wang et al., 2010; Li et al., 2008). (4) If we have both relevant good-quality data and expert prior knowledge then the zoning will reflect the output of the classification algorithm modified or endorsed by the insights of field experts. Semi-supervised methods (Li et al., 2008) use decision-tree or rough-set rules to allow the output from a formal algorithm to be adjusted by engaging with experts. The significance of the stratification of a target domain can be detected by the geographical detector (www.sssampling.org/geogdetector).

3.4. Sensitivity analysis of the specification of strata

Four surfaces from completely random (A) to perfectly stratified (D) are illustrated in Fig. 1a and denoted on the oblique axis in Fig. 1b. Samplers demarcate the surfaces to represent the real stratification according to the sampler's knowledge of the target domain. The bias between the real stratification of the target domain and the sampler's zonation is indicated by the horizontal axis of Fig. 1b, where a value of zero denotes perfect coincidence of the two strata; the bias increases from 0 to 90; 90 denotes that the zonation completely fails to reflect the real stratification of the target domain. The vertical axis in Fig. 1b denotes the error in the sample mean, which changes with stratification of the target domain and the bias between the stratification of the target domain and the zonation for sampling for a given sample size. The sensitivity simulation shows that (1) the sampling efficiency increases (error reduces from 0.025 to 0 along the vertical axis in Fig. 1b) with an increase in the spatially stratified heterogeneity of the target domain (from target domain A to D on the oblique axis in Fig. 1b) and with a reduction in the bias between the sampling zonation and the true stratum of the target domain (from 90 to 0 along the horizontal axis in Fig. 1b); (2) stratified sampling loses efficiency and is no different from simple random sampling if the target domain is completely random (see target domain A in Fig. 1b); and (3) stratified sampling obtains a sample estimate with very small error if the target surface is perfectly stratified (surface D) and is well reflected by the sampler's zonation (the bias is zero). Samplers would have no choice but to carry out random sampling if there is no knowledge of the real target domain no matter whether the domain was random of stratified. Therefore, samplers should take every effort to collect prior knowledge to approximate the stratification of the target domain.



a target surfaces ℜ with four spatial heterogeneities **b** sample mean variance v

Fig. 1. Reduction of sample mean variance change with surface heterogeneity and zonation bias.

3.5. *Information flow of spatial sampling strategies*

If users have no prior knowledge of the target domain $\Re$, sample units have to be distributed randomly or systematically $\Im$ over the target domain $\Re$ (Fig. 2); e.g., see the simple random sampling in case study 1 in Section 5.2. Alternatively, if information on the features of the target domain $\Re$ is available, users can sample randomly for a random or trend surface, or sample systematically for a random target domain or systematic target domain, or use stratified sampling for a heterogeneous (stratified) or trend target domain (see the dark solid arrows from sampling $\Im$ to target domain $\Re$ in Fig. 2). An example is the spatial stratified sampling in case study 1 in Section 5.2. No matter whether the users know the target domain $\Re$, users can explore the features of the target domain after sampling using the sample data acquired (see the box between target domain $\Re$ and statistic Ψ) and choose the optimal statistic that most closely fits the target domain features (see the solid dark arrows from $\Re$ to Ψ in Fig. 2) according to the sampling trinity theory (refer to Section 3.1), irrespective of how the sample was already distributed over space (see the distribution of sample units $\Im$ in Fig. 2). Case study 2 in Section 5.2 is an example of the posterior stratification strategy. Possible non-correspondence between the statistical model Ψ and the distribution of sample units $\Im$ may arise when users do not know the features of the target domain $\Re$ before sampling $\Im$ and thus have to choose random or systematic sampling. This posterior exploratory data analysis after sampling $\Im$ can identify features of the target domains that can be used to improve the subsequent statistics. For statistic Ψ in Fig. 2, the term ‘spatial random’ refers to a random statistic considering spatial autocorrelation (Haining, 2003), and the terms ‘spatial systematic’ and ‘spatial stratified’ have corresponding meanings. The gray lines in the figure are not recommended selections because the spatial autocorrelation existing in target domain $\Re$ is not accounted for by statistic Ψ and the sampling would be inefficient.

Uncertainty in the final sample estimation inevitably arises and propagates through the whole process, from the very beginning of sampling to the final sample statistic. The causes of this uncertainty are varied, and include uncertainty in an individual sample unit, target domain random features of spatial autocorrelation and heterogeneity, intrinsic uncertainty in sampling due to not enumerating the population, bias between the real stratum of the actual target domain and the sampling stratum due to incomplete human knowledge or physical inaccessibility of some sites, bias between the real stratum of the actual target domain and the statistical stratum, and the fact that statistical models are not perfect. There is a good chance of high uncertainty in the sample estimation if we do not have prior knowledge of target domain $\Re$ (see the right-hand side of the judgment diamond in Fig. 2). In this case, we can only use random or systematic sampling, and the estimate would differ greatly from the real target domain if the target domain was strongly heterogeneous or stratified, and consequently result in a significant loss of efficiency of the sampling.

4. SSSampling

We have developed a computer software package called SSSampling (Sandwich Spatial Sampling and Inference Software) that implements spatial sampling procedures. It has functions for the identification of a study area Ω, for the distribution of sample units $\Im$ over a target domain $\Re$, and for statistical inference Ψ using

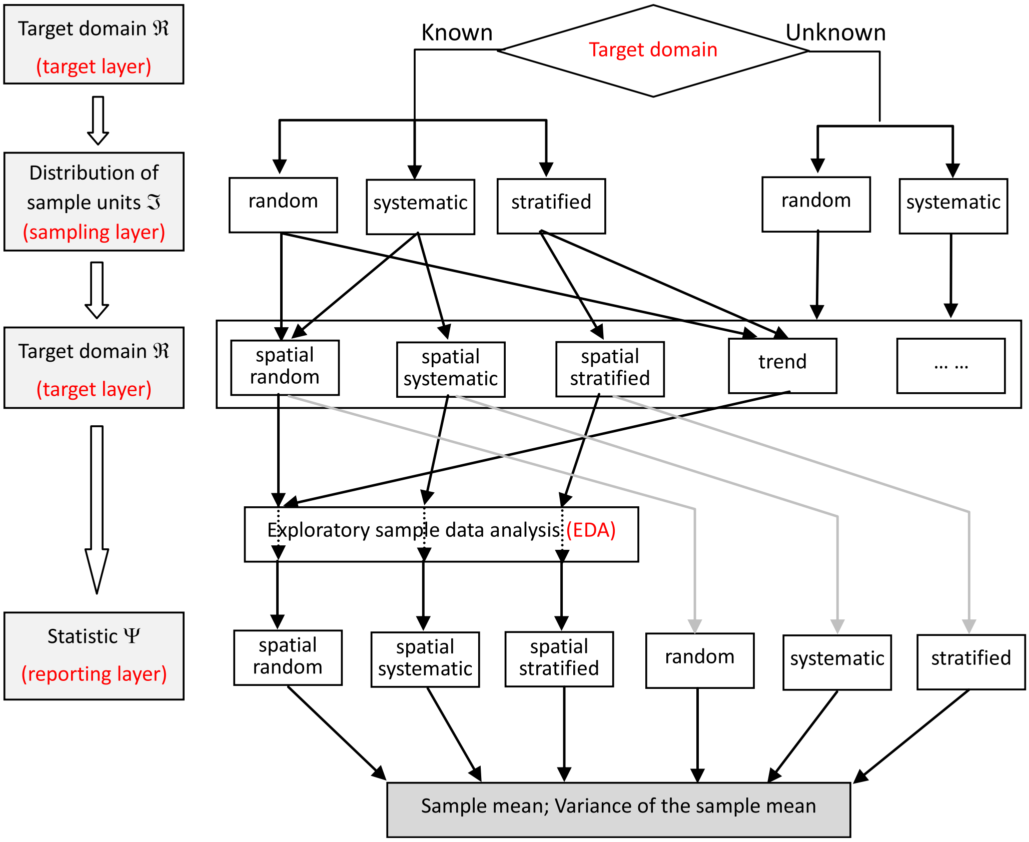


Fig. 2. Information flow in spatial sampling (the solid arrows are the optimal choices of sampling $\Im$ and statistic Ψ most suited to the surface features $\Re$; the gray solid lines are not optimal statistics, and can be avoided once the sample data have been acquired and the surface features have been explored using the sample data).

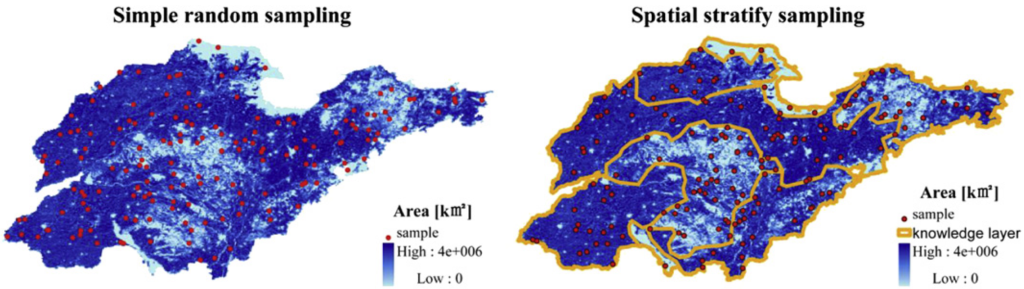


Fig. 3. Two sampling schemes in Shandong Province.

the sample, and GIS functions providing a platform for I/O and management of the spatial data. The SSSampling website is www.sssampling.org.

The SSSampling infrastructure is based on sampling trinity theory and supports three processes. The first of these is sampling design ($\Re \to \Im \to \Psi$). The second is statistical inference ($\Im \to \Psi$), which considers the target surface properties $\Re$ and sampling distributions $\Im$, which are not necessarily consistent with the distribution of the sample as implied in conventional sampling techniques (Cochran, 1977; Kreyszig, 1999). The inconsistency between the sample distribution $\Im$ and statistic Ψ allows improved estimation after the sample has been collected. For example, users have no alternative but to distribute samples randomly over target domain $\Re$ if there is no prior knowledge of the target domain $\Re$. However, the target domain $\Re$ may be found to be spatially stratified by exploration of the collected sample, and a stratified statistic instead of a random statistic should then be used to estimate the sample mean and its variance. The third process assesses the precision of a reported statistic (Ψ) by investigating the properties of the target surface $\Re$, the distribution of the sample $\Im$ and the statistic Ψ used to calculate the reported quantity. SSSampling provides the following options (ref. www.sssampling.org).

- Sampling with prior or posterior precision. The option allows the user to estimate the size of a sample or the precision of a sample estimate prior to sampling, or to make a statistical inference Ψ for an existing sample $\Im$.
- Precision of estimate or size of sample. The option allows the user to estimate the precision and confidence level of the estimate given the number of sample units, or to estimate the number of sample units required to meet a pre-specified precision and confidence level for an estimate.
- Sampling design. The option allows the user to distribute a sample randomly, systematically, or in a stratified manner on a map.
- Statistical method for estimation. Example options include simple random, spatial random, spatial stratified, and spatial sandwich mapping.
- Zonation. Spatial heterogeneity is prevalent in geo-phenomena, and is usually reflected by zonation. Two methods for zonation are provided in SSSampling: k-means clustering if users have relevant data while currently there is no direct way to specify the number of zones, and manual coding if the prior knowledge is categorical.
- Intelligent parameter setting. Users who are unfamiliar with sampling theory sometimes have difficulty in choosing a sampling procedure. SSSampling provides a user interface that allows users to enter their parameters, and the software then finds a model(s) that matches the given parameters.

5. Case studies

5.1. Case study 1: optimal design for a sampling plan

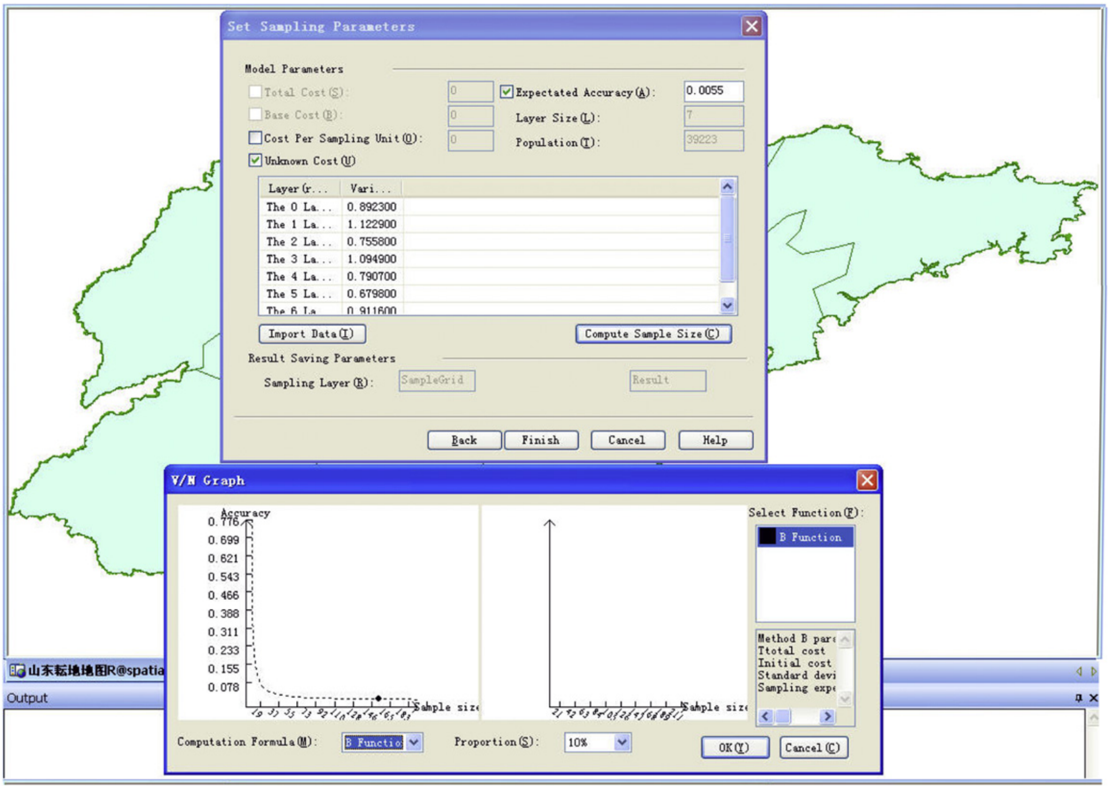
To design a monitoring network or field sampling survey, we need to calculate the optimal sample size or prior precision of the sample estimate.

(1) Aim. To design a sampling plan to survey the area of cultivated land in the Shandong Province.
(2) Data. The province is divided into 39,223 cells, each having dimensions of 2 km × 2 km; the area of cultivated land within a cell is taken as the sampling unit. The province is stratified into seven strata, according to the principle of minimizing the variance of the attribute within each stratum, and maximizing the variance between strata. In this study, the stratification was drawn by a panel of experts on land use to delimit the province into areas with a homogeneous proportion of cultivated area per cell, according to determinants of land use such as elevation, climate zone, and cultivation culture in the province (Fig. 3).
(3) Two sampling plans. We conducted both simple random sampling (Cochran, 1977, p. 18) and spatial stratified sampling (Wang et al., 2002) by drawing 160 sample units in each case. Fig. 4 shows a user interface of SSSampling. The B function shown in Fig. 4 is one of the groups of input parameters.
(4) Conclusion. Table 1 presents the averaged area and its relative errors of cultivated land in each of the cells, as estimated by each of the two sampling plans. The interval of the sample mean acquired from the spatial stratified sampling is narrower than that of the simple random sampling. Additionally, according to the relative errors, the spatial stratified sampling has greater precision than the simple random sampling.

5.2. Case study 2: improving inference given an existing sample

For an existing monitoring network or sample dataset, such as an existing weather observation network or epidemic surveillance network, we need to recommend the best statistic or suggest improvements to the monitoring network.

(1) Aim. To improve the estimates of the existing national weather observation network.
(2) Data. Given a distribution of 720 national meteorological stations in China and the mean annual temperature for each of the stations averaged over the period 1991–2000, we need to estimate the mean annual temperature for the whole country. A simple random statistic adds all sample values together and then divides by the number of sample units (in this case, 720). This estimate can, however, be improved using a spatial stratified statistic.



Set Sampling Parameters

Model Parameters

☐ Total Cost (S): 0

☐ Base Cost (B): 0

☐ Cost Per Sampling Unit (O): 0

☑ Unknown Cost (U)

☑ Expectated Accuracy (A): 0.0055

Layer Size (L): 7

Population (T): 39223

Layer (r...	Vari...
The 0 La...	0.892300
The 1 La...	1.122900
The 2 La...	0.755800
The 3 La...	1.094900
The 4 La...	0.790700
The 5 La...	0.679800
The 6 La...	0.911600

Result Saving Parameters

Sampling Layer (R): SampleGrid Result

V/N Graph

Select Function (F): B Function

Method B pars
Ttotal cost
Initial cost
Standard devi
Sampling expe

Computation Formula (M): B Functio

Proportion (S): 10%

Fig. 4. Parameter settings and calculation of the estimate precision in spatial stratified sampling.

Table 1
Means and errors of two sampling schemes.

Sampling models	Sample mean	95% confidence of sample mean		Relative errors = (1/true mean) × (true mean − sample mean)
		Low	Upper	
Simple random sampling	0.700	0.660	0.740	0.070
Spatial stratified sampling	0.700	0.690	0.700	0.050

Note: the true mean is for all 39,223 cells of the population.

(3) Spatial stratified statistic (Wang et al., 2002). Annual temperature is obviously affected by elevation. We use SRTM DEM data with spatial resolution of 90 m × 90 m to stratify the country into nine strata, according to the principle of minimizing the dispersion variance of the annual temperature within each stratum and maximizing the values between strata (see Fig. 5). We then calculate the mean annual temperature using stratified statistics.

(4) Conclusion. The spatial stratified statistic gives a smaller standard variance of the sample mean (0.21) than the simple random statistic (0.26). Thus, the stratified statistic is more accurate than the simple random statistic for estimation of the annual average temperature in China. This also means that, given the precision of the estimate, a smaller number of sample units (meteorological stations) are needed if a much more advanced statistic is employed.

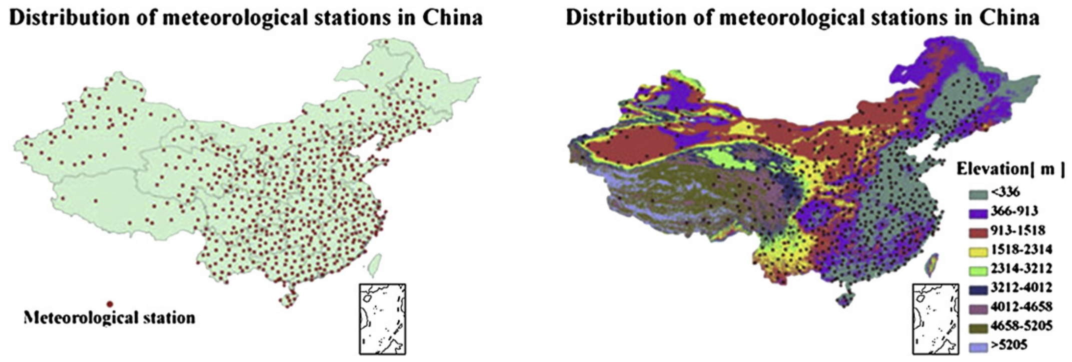


Fig. 5. Improving estimates based on data recorded by existing meteorological stations.

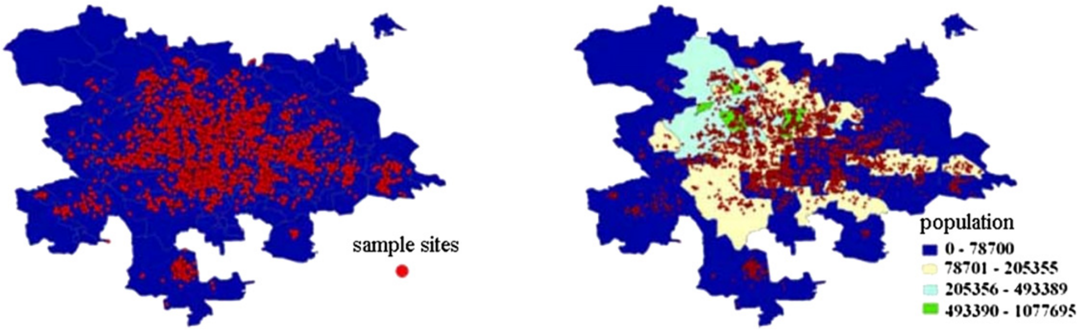


Fig. 6. Assessment of the precision of a reported residential satisfaction survey in 2005.

5.3. *Case study 3: evaluation of the precision of a statistic report*

Given a published statistic, such as regional greenhouse gas emissions, prevalence of a disease in a region, or the amount of contaminated soil in a region, we need to evaluate its precision.

(1) Aim. To assess the precision of a published report.
(2) Data. 3797 people were randomly drawn from 7,848,000 citizens in eight core districts of Beijing in 2004 (see Fig. 6) and asked to give their residential satisfaction as a score between 0 and 100. The reported mean of the survey was 65.9 (Zhang et al., 2006) while the standard variance of the sample mean using simple random sampling was 0.197.
(3) Assessment of the precision of the reported value. In 2000, an enumerate survey of the population was conducted in Beijing. The size of the population in each of the statistical units is seen, to some extent, as a proxy variable of the citizens' expression of their residential satisfaction, and is thus used to stratify the city into four strata (see Fig. 6). The sample mean and its standard variance using the spatial stratified statistic (Wang et al., 2002) are estimated as 65.9 and 0.156, respectively.
(4) Conclusion. The mean residential satisfaction of the citizens in Beijing in 2005 is $65.9 \pm 0.156 \times 1.96 = 65.9 \pm 0.3$ with 95% confidence.

6. Conclusions and discussion

We have developed the SSSampling computer program to facilitate design-based spatial sampling design and statistical inference. SSSampling can be used for prior sampling design before field work, and posterior precision assessment in sample estimation. The prominent features of SSSampling are summarized as follows. (1) According to the trinity relationship between the target domain ($\Re$), sampling ($\Im$) and statistical inference (ψ) (Fig. 1), the software distributes a sample optimally over space, then makes optimal inferences using the sample collected, and assesses the precision of an existing statistical report. (2) The sandwich framework of a target domain ($\Re$), intelligent layer ($\Im$) and reporting layer (ψ) is able to integrate various forms of prior knowledge, enable multi-unit reporting and facilitate operation in GIS environments because of the shared concept of layers. (3) The zonation module, an intelligent engine used to formalize prior knowledge from diverse sources, is used to sample spatially stratified heterogeneous target domains, given the different prior information available (Huang et al., 2006; Li et al., 2008). (4) Intelligent setting of parameters for various models is supported.

The inferential framework invoked throughout (Cochran, 1977; Sarndal et al., 1992) is design-based inference, in which repeated and random sampling is employed to obtain an estimate of the population, to approximate the population without bias (de Gruijter and Ter Braak, 1990). Model-based inference, in which an observation is regarded as one realization of an underlying mechanism with some probability distribution, is the other major inference framework (more the focus in Cressie's book). Obviously, the two inference frameworks have different objectives, bias, variance, and 'optimality' criteria. Model-based inference is appropriate for the projection of parameters of a superpopulation or over a cross section space, while the design-based approach is more appropriate for "here and now" statistics, and is also applicable to a superpopulation when the observation is long enough to reflect its underlining process. Some estimation objectives are addressed best by an inference framework (e.g., estimating a mean, total or proportion), and others are strongly associated with the model-based framework (e.g., estimating a variogram, predicting values at unsampled locations, estimating a maximum, or estimating a spatial pattern). For example, birth defects in a population are low-probability events in villages of Heshun County, with around 0–10 cases from 0 to 150 live births annually (Gu et al., 2007). However, observations are only available for a few years. A design-based approach reflects new occurrences of the disease in the observed period, and thus denotes the amount of healthcare resources that should be allocated to handle the disease burden in the period. Alternatively, the model-based approach reflects the baseline level and the risk of the disease in the long-term, and as such, would suggest healthcare resources should be reserved for the county to handle the potential long-term burden. The model-based incidence should be taken as the response variable if one wants to model the disease and its determinants. A model-based approach has to be employed for spatial interpolation because spatial interpolation usually has to be based on an assumption of the probability process. If the observations are long-term or large enough so that a design-based estimate of incidence is reliable, the design-based incidence gives a proper indication of the risk of the disease in the area.

The SSSampling program has been designed so that alternate and future advances in spatial sampling and spatial statistics can be easily implemented. Example techniques include wavelet sampling (Atkinson and Emery, 1999), high-dimension and oscillatory surface modeling (Benedetto et al., 2001), model-assisted survey sampling (Sarndal et al., 1992), inverse kriging to distribute the sample (Spöck, 2012) such that the mean error or maximum error of prediction at unsampled sites is minimize, importance sampling (Rogerson, 2005), and a module for the uncertainty of an individual sample unit. The infrastructure of the software can also be revised to adopt new state-of-art software techniques.

Acknowledgments

This study was supported by NSFC (41023010; 41271404), MOST (2012CB955503; 2012ZX10004-201; 2011AA120305) and CAS (XDA05090102) grants.

References

Atkinson, P.M., Emery, D.R., 1999. Exploring the relation between spatial structure and wavelength: implications for sampling reflectance in the field. International Journal of Remote Sensing 20, 2663–2678.

Bellhouse, D.R., 1977. Optimal designs for sampling in two dimensions. Biometrika 64, 605–611.

Benedetto, J.J., Paulo, J.S., Ferreira, G., 2001. Modern Sampling Theory. Birkhauser, Boston.

Bueso, M.C., Angulo, J.M., Alonso, F.J., Ruiz-Medina, M.D., 2005. A study on sensitivity of spatial sampling designs to a priori discretization schemes. Environmental Modelling & Software 20 (7), 891–902.

Caeiro, S., Painho, M., Goovaerts, P., Costa, H., Sousa, S., 2003. Spatial sampling design for sediment quality assessment in estuaries. Environmental Modelling & Software 18 (10), 853–859.

Christakos, G., 1992. Random Field Models in Earth Sciences. Dover Publications, Inc., New York.

Christakos, G., 2010. Integrative Problem-solving in a Time of Decadence. Springer, Berlin.

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, third ed. John Wiley & Sons, USA.

Cressie, N.A.C., 1991. Statistics for Spatial Data. Wiley, 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.

de Gruijter, J.J., Brus, D., Bierkens, M., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, New York, NY.

Dungan, J.L., Perry, J.N., Dale, M.R.T., Legendre, P., Citron-Poustym, S., Fortinm, M.-J., Jakomulskam, A., Miritim, M., Rosenberg, M.S., 2002. A balanced view of scale in spatial statistical analysis. Ecography 25, 626–640.

Fischer, M.M., Wang, J.F., 2011. Spatial Data Analysis: Models, Methods and Techniques. Springer, Berlin.

Gallego, F.J., 2005. Stratified sampling of satellite images with a systematic grid of points. ISPRS Journal of Photogrammetry Remote Sensing 59, 369–376.

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

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

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 (10), 702–707.

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

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

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 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.

Hu, M.G., Wang, J.F., 2011. A spatial sampling optimization package using MSN theory. Environmental Modelling & Software 26 (4), 546–548.

Huang, B., Liu, N., Chandramouli, M., 2006. A GIS supported ant algorithm for the linear feature covering problem with distance constraints. Decision Support Systems 42 (2), 1063–1075.

Journel, A., Huijbregts, C.H., 1978. Mining Geostatistics. Academic Press Inc. LTM, London.

Kreyszig, E., 1999. Advanced Engineering Mathematics, eighth ed. John Wiley & Sons Inc, New York.

Kumar, N., 2007. Spatial sampling design for a demographic and health survey. Population Research and Policy Review 26, 581–599.

Li, L.F., Wang, J.F., Liu, J.Y., 2005. Optimal decision-making model of spatial sampling for survey of China's land with remotely sensed data. Science in China – Series D 48 (6), 752–764.

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.

Lin, Y.P., Yeh, M.S., Deng, D.P., Wang, Y.C., 2008. Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. Global Ecology and Biogeography 17, 175–188.

Lister, A.J., Scott, C.T., 2008. Use of space-filling curves to select sample locations in natural resource monitoring studies. Environmental Monitoring and Assessment 149, 71–80.

Lwange, S.K., Lemeshow, S., 1991. Sample Size Determination in Health Studies: A Practice Manual. World Health Organization, Geneva.

Matérn, B., 1986. Spatial Variation, Volume 36 of Lecture Notes in Statistics. Springer-Verlag, Berlin.

Müller, W.G., Rodríguez-Díaz, J.M., López, M.J.R., 2012. Optimal design for detecting dependencies with an application in spatial ecology. Environmetrics 23, 37–45.

Olea, R.A., 1984. Sampling design optimization for spatial functions. Mathematical Geology 16, 369–392.

Olken, F., Rotem, D., 1995. Sampling from spatial databases. Statistics and Computing 5, 43–57.

Opsomer, J.D., Nusser, S.M., 1999. Sample designs for watershed assessment. Journal of Agricultural, Biological, and Environmental Statistics 4, 429–442.

Overton, W.S., Stehman, S.V., 1993. Properties of designs for sampling continuous spatial resources from a triangular grid. Communications in Statistics – Theory and Methods 22, 2641–2660.

Paola, C.E., Foufoula-Georgiou, W.E., Dietrich, M., Hondzo, D., Mohrig, G., Parker, M.E., Power, I., Rodriguez-Iturbe, V., Wilcock, P., 2006. Toward a unified science of the Earth's surface: opportunities for synthesis among hydrology, geomorphology, geochemistry, and ecology. Water Resources Research 42, W03S10. http://dx.doi.org/10.1029/2005WR004336.

Pozo, K., Harner, T., Wania, F., Muir, D.C.G., Jones, K.C., Barrie, L.A., 2006. Toward a global network for persistent organic pollutants in air: results from the global atmospheric passive sampling study. Environmental Science and Technology 40, 4867–4873.

Rao, J.N.K., 2003. Small Area Estimation. John Wiley & Sons, New York.

Rodriguez-Iturbe, I., Mejia, J.M., 1974. The design of rainfall networks in time and space. Water Resources Research 10, 713–728.

Rogerson, P., 2005. Monitoring spatial maxima. Journal of Geographical System 7, 101–114.

Rogerson, P.A., Delmelle, E., Batta, R., et al., 2004. Optimal sampling design for variables with varying spatial importance. Geographical Analysis 36, 177–194.

Sarndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, Berlin Heidelberg.

Schwanghart, W., Beck, J., Kuhn, N., 2008. Measuring population densities in a heterogeneous world. Global Ecology and Biogeography 17, 566–568.

Spöck, G., 2012. Spatial sampling design based on spectral approximations to the random field. Environmental Modelling & Software 33, 48–60.

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.

Stein, A., van der Meer, F., Gorte, B., 1999. Spatial Statistic for Remote Sensing. Kluwer Academic Publishers, New York.

Stevens Jr., D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99, 262–278.

van Groenigen, J.W., van Stein, A., 2000. Constrained optimization of spatial sampling in a model-based setting using SANOS software. In: Heuvelink, G.B.M., Lemmens, M.J.P.M. (Eds.), Accuracy 2000: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Nature Resources and Environmental Sciences, Amsterdam, 2000. Delft University Press, Delft, pp. 679–687.

Wang, J.F., Wise, S., Haining, R., 1997. An integrated regionalization of earthquake, flood and drought hazards in China. Transactions in GIS 2, 25–44.

Wang, J.F., Liu, J.Y., Zhuang, D.F., Li, L.F., Ge, Y., 2002. Spatial sampling design for monitoring the area of cultivated land. International Journal of Remote Sensing 13, 263–284.

Wang, J.F., Christakos, G., Hu, M.G., 2009. Modeling spatial means of surfaces with stratified non-homogeneity. IEEE Transactions on Geoscience and Remote Sensing 47 (12), 4167–4174.

Wang, J.F., Haining, R., 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 (4), 523–543.

Wang, J.F., Reis, B.Y., Hu, M.G., Christakos, G., Yang, W.Z., et al., 2011. Area disease estimation based on sentinel hospital records. PLoS ONE 6 (8), e23428.

Webster, R., 1985. Quantitative spatial analysis of soil in the field. Advances in Soil Science 3, 1–70.

Zhang, W.Z., Yi, W.H., Zhang, Q.J., Meng, B., Gao, X.L., 2006. China Livable Cities Report. Social Sciences Academic Press, Beijing, Beijing (in Chinese).

Further reading

nQuery Advisor (http://www.statsol.ie/nquery/demo/index.html)

PASS (http://www.ncss.com/pass.html)

G*Power (http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/)

Sample size calculator/The Survey System (http://www.surveysystem.com/sscalc.htm)

Sample size calculator/Macorr (http://www.macorr.com/ss_calculator.htm)

Sample size calculator/Raosoft (http://www.raosoft.com/samplesize.html)

Matlab (http://www.mathworks.com/)

SPSS (http://www.spss.com/spss/)

GeoDA (https://www.geoda.uiuc.edu)

GeoBUG (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml)

Crimestat (http://www.icpsr.umich.edu/CRIMESTAT/)

SatScan (http://www.satscan.org/)

SSSampling (www.sssampling.org) design-based spatial sampling software

MSN (www.sssampling.org/MSN): model-based spatial sampling software

B-shade (www.sssampling.org/B-shade): software to remedy biased sampling

Geographical detector (www.sssampling.org/geogdetector): software to detect spatial stratification and health risk