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.

© 2012 Elsevier Ltd. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/6964548

Download Persian Version:

https://daneshyari.com/article/6964548

Daneshyari.com