Geoderma 315 (2018) 160–169

Contents lists available at ScienceDirect

Geoderma

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

Accounting for access costs in validation of soil maps: A comparison of design-based sampling strategies

Lin Yang[a,b], Dick J. Brus[c,d,*], A-Xing Zhu[b,c,e,f,g], Xinming Li[b], Jingjing Shi[b]

[a] School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing 210023, China
[b] State Key Laboratory of Resources and Environment Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China
[d] Biometris, Wageningen University and Research, Wageningen 6708 PB, The Netherlands
[e] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, School of Geography, Nanjing Normal University, Nanjing 210023, China
[f] State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China
[g] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Probability sampling
Sampling with probabilities-proportional-to-size
Stratified random sampling
Optimal stratification
Digital soil mapping
Mean squared error

ABSTRACT

The quality of soil maps can best be estimated by collecting additional data at locations selected by probability sampling. These data can be used in design-based estimation of map quality measures such as the population mean of the squared prediction errors (MSE) for continuous soil maps and overall accuracy for categorical soil maps. In areas with large differences in access costs it can be attractive to account for these differences in selecting validation locations. In this paper two types of sampling design are compared that take access costs into account: sampling with probabilities proportional to size (pps) and stratified simple random sampling (STSI). In pps the inverse of the square root of the access costs is used as a size variable. Two estimators of MSE are applied, the Hansen-Hurwitz and Hajek estimator. In STSI optimal strata are constructed based on access costs. Simple random sampling (SI) is taken as a reference design. The sampling strategies were compared on the basis of: 1) the variance of the estimated MSE; 2) the variance of the total pointwise access costs; 3) the 95-percentile of the sampling distribution of the total access costs. The comparison was done at equal expected total pointwise access costs. The sampling strategies were compared in a simulation study and a real-world case study in Anhui, China. In the case study car travel and hiking costs were considered in computing access costs per point. The results showed that the variance of estimated MSE with pps(Hansen-Hurwitz) was larger than with pps(Hajek) and STSI. The variances of estimated MSE of pps(Hajek) and STSI were about equal and smaller than that of SI. The gain in precision compared to SI depends on the cost distribution. The larger the coefficient of variation of the costs, the larger the gain. The 95 percentile of the sampling distribution of the total pointwise access costs with STSI was smaller than with pps and SI. The gain in precision of pps(Hajek) and STSI was about 30% accounting for hiking costs only, and about 10% accounting for the sum of car travel and hiking costs in the case study. The proposed sampling strategies are of interest for surveying any soil property in areas with marked differences in access costs, not just for validation of soil maps.

1. Introduction

In recent years, there has been an upsurge of digital soil mapping due to a global need of soil information for environmental modelling and developed quantitative methodologies (Minasny and McBratney, 2016). Many soil maps have been produced all over the world (McBratney et al., 2003; Zhu et al., 2010; Yang et al., 2017; Hengl et al., 2017). Validation of those maps is important and necessary for producers and users to understand the quality of the soil maps (Brus et al., 2011; Bishop et al., 2015; Simo et al., 2015).

Map quality measures such as the population mean of the squared prediction error (MSE) for continuous soil maps and overall accuracy (purity) of categorical sol maps, commonly are determined by data-splitting or cross validation (Brus et al., 2011). A serious drawback of these validation methods is that a model of the prediction error is needed to estimate the map quality measures and the uncertainty about

[*] Corresponding author at: Biometris, Wageningen University and Research, Wageningen 6708 PB, The Netherlands.
E-mail address: dick.brus@wur.nl (D.J. Brus).

https://doi.org/10.1016/j.geoderma.2017.11.028
Received 23 August 2017; Received in revised form 9 November 2017; Accepted 23 November 2017

these estimates. Knotters and Brus (2013) showed that different models can lead to largely different estimates of map quality. The quality of the estimated map quality measures depends on the quality of the model. Preferably, map quality is estimated by model-free, design-based inference, so that discussions on the validity of the estimated map quality are avoided (Stehman, 1999; Brus et al., 2011). Design-based estimation requires that the validation locations are selected by probability sampling. Numerous sampling designs are available for selecting probability samples (de Gruijter et al., 2006). This paper is about how differences in access costs can be accounted for in the random selection of the validation locations. The access costs of a point in the field are the costs of traveling from a starting point to that point by car or/and by foot. Sampling at lower densities in poorly-accessible and/or remote areas may increase the sampling efficiency in terms of cost effectiveness.

In practice, selected sampling points can be difficult to access or even inaccessible, in areas with poor road networks, dense vegetation or rough terrain conditions. Researchers thus developed methods to select alternative sampling locations to replace the locations that in the field appear to be inaccessible or poorly accessible, see e.g. Thomas et al., 2012; Kidd et al., 2015, Clifford et al., 2014, and Stumpf et al., 2016. However, when a map of the accessibility is available, it can be attractive to account for accessibility constraints already from the beginning, at the design-stage, when selecting the sampling locations.

In recent years a couple of papers were published on how to account for accessibility constraints or operational costs in selecting the sampling locations. In all these papers the proposed methods are adaptations of conditioned Latin Hypercube Sampling, cLHS (Minasny and McBratney, 2006). Roudier et al. (2012) proposed to use a second objective function equal to the sum of the costs over the sampling points along with the usual objective function of standard cLHS. In the simulated annealing algorithm both objective functions are evaluated, leading to two acceptance probabilities. A new sample is accepted if a draw from the uniform distribution is smaller than at least one of the acceptance probabilities. Yin et al. (2016) also introduced a second objective function for the access costs, but multiplied this objective function with the objective function of cLHS. Mulder et al. (2013) adapted standard cLHS in a different way: they added the sum of the pointwise off-road travel costs over all sampling points to the O_1 criterion of standard cLHS.

In most cases cost-constrained cLHS sampling is used for selecting a sample for calibration of a model for digital soil mapping. Exceptions are Silva et al. (2014) and Yin et al. (2016): they used cost-constrained cLHS for validation of maps. However, as argued by Brus (2015), cLHS is not a probability sampling design, despite the randomness in the selection of the sampling locations, and therefore we think cLHS is not a suitable design for map validation.

This paper presents design-based sampling strategies for validation of soil maps that account for differences in access costs. We work out two types of probability sampling design: stratified simple random sampling (STSI) and sampling with probabilities-proportional-to-size (pps). Simple random sampling (SI) that does not account for differences in access costs, is taken as a reference design. The sampling strategies are compared in a simulation study and a real-world case study in Anhui, China. In the simulation study the pointwise costs are simulated by drawing from an exponential distribution. In the case study both hiking costs and car travel costs from a basepoint to the sampling point were considered in computing access costs per point.

2. Sampling strategies accounting for access costs

A sampling strategy is a combination of a sampling design and an estimator (de Gruijter et al., 2006). A sampling design assigns a selection probability to each possible sample, i.e. combination of units in the population. For instance in simple random sampling without replacement of n units out of a population of N units there are $\binom{N}{n}$ possible samples. The selection probability of each sample is $1/\binom{N}{n}$ (Lohr, 1999). The sum of the selection probabilities of all samples containing a specific unit is referred to as the inclusion probability of that unit. With simple random sampling (with or without replacement) this inclusion probability equals n/N for each unit in the population. Note that the inclusion probabilities need not be equal. For instance, in stratified random sampling they usually differ between the strata. As long as the inclusion probabilities are known, these can be used in estimation, so that unbiased estimates of the population parameter of interest can be obtained. In the sampling designs of this paper the inclusion probabilities are adapted to the location-specific access costs, the larger these costs the smaller the inclusion probabilities.

An estimator is a formula that is used to estimate a population parameter, such as the population total or mean. In this paper interest is not in the population mean of a soil property, but in the population mean of the squared prediction errors, i.e. the mean of the squared differences between the predicted values at unobserved points and the true (observed) values. This unusual choice of the variable of interest has no implications for the estimator. We simply replace in the formulas the values of the soil property measured at the sampling locations by the squared prediction errors.

In this paper, we present two sampling designs that take the location-specific access costs into account: 1) sampling with probabilities-proportional-to-size (pps), and 2) stratified simple random sampling (STSI). Simple random sampling (SI) in which differences of access costs between locations are not accounted for, is used as a reference design. The sampling designs were compared on the basis of: 1) the sampling variance of the estimated mean of the squared prediction errors (variance of estimated MSE), 2) the sampling variance of the total access costs (sum of pointwise access costs over all sampling points), and 3) the 95-percentile of the sampling distribution of the total access costs. The sampling variance of an estimated population parameter such as the mean is the variance of the estimated parameter over repeated sampling with a given sampling design. It quantifies our uncertainty about the estimated parameter. The smaller the variance of the estimated MSE of a sampling strategy, the more precise the sampling strategy, the less uncertain we are about the population MSE. The smaller the variance of total access costs of a sampling design, the better the control of the total access costs. We want to avoid the risk of selecting with a given design a probability sample leading to much higher total access costs than on average with that design. For a fair comparison the sampling variances are computed for cost-equivalent sample sizes, i.e. for sample sizes such that the expected total access costs are equal for all three designs. In the next subsection the three sampling designs are described in detail. Estimators for the sampling variance of the estimated MSE and of the variance of the total access costs are given.

2.1. Simple random sampling

In the reference design SI sampling points are selected independently from each other and with equal probability. Differences in access costs between points are not accounted for. The population mean of the squared prediction error (MSE) is estimated by

$$\hat{\bar{e}}_{SI} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{1}$$

with n the sample size (number of selected sampling points) and e_i the squared prediction error. Note that for notational convenience, the square is not added to the symbol e, so e represents squared prediction errors.

The sampling variance of the estimated mean squared errors equals