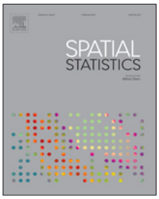
Spatial Statistics 13 (2015) 106–122

Contents lists available at ScienceDirect

Spatial Statistics

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



# Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability

Mareike Ließ

*University of Bayreuth, Department of Geosciences/ Soil Physics Division, Universitaetsstrasse 30, 95447 Bayreuth, Germany*

ARTICLE INFO

*Article history:*
Received 24 February 2015
Accepted 10 June 2015
Available online 17 June 2015

*Keywords:*
Sampling design
Digital soil mapping
Regression

ABSTRACT

With respect to sampling for regression-based digital soil mapping (DSM), the above all aim is to ensure that the spatial variability of the soil is well-captured without introducing any bias, while the design remains feasible according to operational constraints such as accessibility, man power and cost. Representativeness of the sample concerning the population to be sampled needs to be guaranteed in any regression-based modelling approach. Four selected sampling designs were adapted to show that basically any design may be optimised to represent the *n*-dimensional predictor space of a particular area, while selecting points is only permitted from a small accessible sub-area or from outside the area. Sampling efficiency may be evaluated based on the representation of the predictor space. However, not only each predictor's probability function but also the interaction between predictors may have to be considered, to select a representative sample. Instead of sampling a previously un-sampled area with limited accessibility, the four sampling designs may also be used to subsample an existing dataset and, thereby, optimise a suboptimal dataset based on the predictor space of the area which shall be mapped by DSM.

*E-mail address:* mareike.liess@uni-bayreuth.de.

http://dx.doi.org/10.1016/j.spasta.2015.06.002
2211-6753/© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

Within the field of digital soil mapping (DSM) special emphasis has to be given to the application of an appropriate sampling scheme in order to receive unbiased predictions by the later developed model. Representativeness of a sample concerning the population to be sampled is of primary importance. Simple random sampling is one of the most popular sampling techniques. It gives each sample the same probability of being selected and the selected samples should, therefore, well-represent the considered population. However, particularly with a large number of selectable points and a small sample size, this representativeness is left to chance. Furthermore, in large areas which are difficult to access this approach is not feasible. Accordingly, this causes the often observed discrepancy between statistical theory and the applicability to the soil scientist collecting data in the field.

While with interpolation methods such as kriging, emphasis has to be given to a good spatial coverage and hence a uniform spreading of the observations in geographical space (Brus et al., 2007), which is also not feasible in inaccessible or arduous terrain, regression-based DSM does not need a good spatial coverage. There are two general approaches which are common to select sampling positions for model training in regression-based DSM: (1) Samples are taken based on stratified random sampling, i.e. the area is subdivided into a number of strata or subareas and then random samples are taken from each of these strata, giving either equal or different selection probabilities to the strata. The strata can be based on e.g. *xy*-coordinates, terrain parameters, land use classes, etc. (e.g. McKenzie and Ryan, 1999; Dobermann and Simbahan, 2007) or (2) samples are chosen according to the probability density function of each predictor within the research area as is done e.g. in Gessler et al. (1995) and Hengl et al. (2003) or particularly within conditioned latin hypercube sampling (clhs, Minasny and McBratney, 2006, 2007).

So far there are only few proposals regarding sampling for DSM in inaccessible areas or areas which are difficult to access (DTA-areas). These include transect sampling along hill-slopes and thereby covering terrain units (Ließet al., 2009) and the approach by Cambule et al. (2013) to validate a model established for an accessible area with few samples from the inaccessible area. Clifford et al. (2014) made a very promising approach to introduce some flexibility into clhs, so that one can easily select an alternative sampling point in case a point occurs to be inaccessible in the field. However, in general the approaches still assume that any of the points within an area is accessible depending on efforts and cost. This is of course correct, but in reality nobody will walk for several days to take a single soil sample or even risk death in areas contaminated by landmines. Sometimes even other research interests prohibit access to the area of interest.

In order to map DTA-areas, a general design is developed for the complete investigation area while samples according to the design are then only taken from the accessible part of the area. This means (1) randomly selected points are replaced by accessible points which represent similar landscape positions, i.e. similar combinations of terrain parameters, vegetation cover, etc. and (2) a landscape classification is applied to the whole area, while samples are then randomly selected only from those parts within each unit which are accessible. The basic assumption on which all strategies are based is of course the concept that similar landscape positions carry similar soils with similar soil properties, the basic concept of all DSM regression approaches, i.e. Jenny's factor model of soil formation (Jenny, 1941).

According to Brus and de Gruijter (1997) the search for cost-effective sampling strategies considering the optimal use of ancillary information has been a major task of sampling theorists since the 1940s. It still remains an up-to-date topic particularly while cost-effectiveness and feasibility are considered. Usually, we do not know the probability function or the spatial specification of a soil property in a particular area. We also do not know if the map created based on a sampled dataset is indeed representing the area's spatial reality. Whereas, we may try as much as possible to select a sample which covers the spatial variability of the factors of soil formation and their interaction and thereby guarantee that we also capture the spatial variability of any soil property caused by these factors. Differences between the various sampling designs in accounting for this "representativeness" of the sample according to the predictor space, while feasibility of the design due to accessibility, man power and cost is given, shall be demonstrated in this study. It is a concern which has so far only been discussed to be relevant for DSM model validation. But then, who wants to build a biased model?