Geoderma 267 (2016) 123–136

Contents lists available at ScienceDirect

Geoderma

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

# An heuristic uncertainty directed field sampling design for digital soil mapping

Shu-Jie Zhang [a,e], A-Xing Zhu [b,c,d,a,*], Jing Liu [d], Lin Yang [a], Cheng-Zhi Qin [a], Yi-Ming An [a]

[a] State Key Lab. of Resources and Environment Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, 1 Wenyuan Road, Nanjing, Jiangsu 210023, China
[c] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, 1 Wenyuan Road, Nanjing, Jiangsu 210023, China
[d] Department of Geography, University of Wisconsin, Madison, WI 53706, USA
[e] Policy Research Center for Environment and Economy, Ministry of Environmental Protection, Beijing 100101, China

ARTICLE INFO

*Article history:*
Received 24 December 2014
Received in revised form 3 December 2015
Accepted 13 December 2015
Available online 8 January 2016

*Keywords:*
Legacy samples
Individual representativeness
Prediction uncertainty
Stepwise sampling scheme

ABSTRACT

Legacy samples are a valuable data source for digital soil mapping. However, these sample sets are often small in size and ad hoc in spatial distribution. Constrained by the limited representativeness of such a sample set, the obtained soil maps are often incomplete in spatial coverage with "gaps" at the locations which cannot be well represented by these samples. The maps may also contain areas of high prediction uncertainty. In order to extend the predicted area and reduce prediction uncertainty, additional samples are needed. This paper presents a sampling design based on prediction uncertainty to select samples which will effectively complement the sparse and ad hoc samples, and maximize the spatial coverage of prediction and minimize prediction uncertainty. A case study in China shows that this sampling scheme was effective in achieving these goals. Compared with stratified random sampling scheme, when the number of additional samples is the same, the produced map using uncertainty directed samples has larger predicted area, and the accuracy of the produced map is higher than that of the maps using stratified random samples. The finding of this study suggests that prediction uncertainty is a useful indicator to aid field sample selection and to complement the legacy data. Furthermore, the mapping accuracy produced using this method can be quantitatively related to the number of additional samples needed which opens a new horizon for digital soil mapping.

© 2015 Published by Elsevier B.V.

## 1. Introduction

Legacy samples which were accumulated through historical national soil surveys and/or specific field studies are a valuable data source for digital soil mapping. However, in most areas especially in developing countries, the number of legacy samples could be so limited that it may not be appropriate to apply traditional methods (such as regression or kriging interpolation methods) to map soils using these samples. For example, the land area of China is about $9.6 \times 10^6$ km$^2$, but there are only 7292 legacy profiles available across the country (about 1 sample per 1316 km$^2$ in average) in the Soil Attributes Database from the Soil Series of China (volumes 1–6) and Soil Series of Provinces (total 34 volumes) (Yu et al., 2007; Shi et al., 2007). Some legacy samples were collected without following any conventional sampling designs (e.g. stratified random sampling, regular sampling). These samples, which are limited in number and do not provide a good spatial coverage of the study area, are referred to as sparse and ad hoc samples in this paper. Sparse and ad hoc samples are especially common for large area but cannot be used with the conventional soil mapping methods (such as regression or kriging) for mapping the spatial distribution of soil properties. Besides, it is difficult to collect a large set of samples following a well-defined sampling design due to restrictions in sampling under complex field conditions and limited sampling budget.

However, each of these sparse and ad hoc samples does contain the local relationship between soil and environmental conditions although these samples do not represent the entire area very well. To make full use of legacy samples, Zhu et al. (2015) presented an individual predictive soil mapping (iPSM) method to predict soil properties using sparse and ad hoc samples. Under the assumption that the more similar the environment conditions between two locations the more similar the soil property values (Hudson, 1992), they used similarity in environmental conditions between an unvisited location and a field sample to approximate the similarity between the soil at the unvisited location and that at the sample. Thus, the soil property value at the unvisited location can be computed based on the similarities to a set of samples and the property values at these samples (Zhu, 1997; Qi et al., 2007; Zhu et al. (2010a)). The method also produces uncertainty associated with each predication based on the similarities of the unvisited location to a set of samples. A "No Data" value will be assigned to locations where uncertainty values exceed a certain threshold (user-specified) under the

* Corresponding author at: School of Geography, Nanjing Normal University, No. 1, Wenyuan Road, Xianlin University District, Nanjing 210023, China.
*E-mail address:* azhu@wisc.edu (A.-X. Zhu).

http://dx.doi.org/10.1016/j.geoderma.2015.12.009
0016-7061/© 2015 Published by Elsevier B.V.

notion that it is not reasonable to predict soil property values for locations which the current set of samples cannot represent well. Therefore, it is very possible to end up with an incomplete soil property map with areas of "No Data" values using this method and it is necessary to collect additional samples to cover these unmapped areas and to reduce the overall uncertainty of the predicted map.

The question then is how to effectively use the computed uncertainty information in designing a sampling scheme which can integrate existing samples and provide as few samples as possible but will maximize the reduction of the unmapped areas and minimize the overall uncertainty in the predicted map.

Design-based sampling (regular grids, simple random sampling, etc.) is very difficult to integrate legacy samples, because the legacy samples are always ad hoc (Brus and de Gruijter, 1997; Walvoort et al., 2010). For model-based sampling scheme, the semi-variance function which quantifies the spatial auto-correlation is always estimated from a large amount of existing samples (Brus et al., 2006; Isaaks and Srivastava, 1989; Goovaerts, 1999). But, the legacy samples are usually sparse, which makes the model-based sampling scheme not suitable for integrating legacy samples.

The conditioned Latin hypercube sampling (cLHS) as proposed by Minasny and McBratney (2006) has the advantage that the distribution of the designed sample locations replicates the distribution of environmental covariates. cLHS is used widely at present (Worsham et al., 2012; Taghizadeh-Mehrjardi et al., 2014; Reza Pahlavan Rada et al., 2014; Kidd et al., 2015). But each cLHS sampling scheme is designed independently, thus cLHS is hardly used for designing additional sample and it cannot provide the sampling order of the additional samples.

Spatial Simulated Annealing (SSA) method could be used to optimize placement of the individual observations by meeting some criterion (the minimal average or maximum Kriging variance), and this optimization sampling method also could include use of previous samples to direct additional sampling (van Groenigen et al., 1999; Van Groenigen, 2000; Brus and Heuvelink, 2007). However, SSA method cannot provide the sampling order of the additional samples also. When the sampling resource is limited, sampling order is very important information for allowing investigators to effectively plan sampling resources.

Purposive sampling intends to collect samples which are typical of soil types or soil mapping units. Purposive samples are usually designed by local soil experts based on their knowledge during conventional soil mapping. But this type of purposive sampling highly depends on experience or personal judgment of soil experts. An integrative hierarchical stepwise sampling strategy has been proposed to design representative samples with assistance of environmental covariates through a fuzzy clustering approach (Yang et al., 2013). Although this method is effective for predicting soil maps in digital soil mapping for initial sampling, this method has not been used in additional sampling, and it also does not have any mechanism to include uncertainty in the design.

This paper presents an effective and stepwise sampling scheme to identify additional sample locations based on the prediction uncertainty information quantified by the iPSM method proposed by Zhu et al. (2015). The method not only effectively extends the mapped area and reduce overall uncertainty using as few samples as possible, but also integrate all the legacy soil samples. Section 2 presents the details of this uncertainty directed and stepwise sampling scheme, which is followed by a case study illustrating the effectiveness of the proposed method. Section 4 presents the result and discussion. Conclusions are drawn in Section 5.

## 2. Methods

### 2.1. Basic idea

The basic idea of the uncertainty directed sampling reported in this paper is to use the uncertainty derived from the method (iPSM) by Zhu et al. (2015) to identify as few samples as possible to map soil spatial distribution below a user specified level of prediction uncertainty and to maximize the reduction of the prediction uncertainty.

Zhu et al. (2015) proposed the individual predictive soil mapping (iPSM) method which can make full use of limited soil sample data for predictive soil mapping and provide the prediction uncertainty at every location where a prediction is made. iPSM uses the soil–environment relationship at each individual soil sample location to predict soil properties at unvisited locations and estimate prediction uncertainty. The more similar the soil environmental conditions between an unvisited location and the locations of legacy soil samples, the lower the prediction uncertainty on the unvisited location because the legacy data could represent the unvisited location well. Prediction uncertainty can be measured before a prediction is made for a given location and a given set of soil sample locations. Thus, with iPSM a user can specify a level of prediction uncertainty under which the soil property at an unvisited location can be predicted. Locations with uncertainty values higher than the specified level are assigned "No Data" to signify that the current set of samples is insufficient to make soil property estimation at the given level of acceptability (as specified by the uncertainty level) (Zhu et al., 2015). These locations form holes or gaps on the output soil property map. Fig. 4 shows the map with unmapped area (gray area) which cannot be represented by legacy samples well. The uncertainty on the gray area is higher than a specified threshold, and "No Data" would be assigned to these locations when mapping soil property.

To examine the effects of sampling on gap filling and on uncertainty reduction, the proposed method were divided into two stages. The first is "Gap filling", which is to identify as few samples as possible to complete the prediction for areas previously labeled as "No Data". The second is to select samples which can maximize the reduction of overall prediction uncertainty. To fill the gaps of soil map and at the same time to minimize the number of samples needed, we adopt the following process: 1) in the unmapped region ("No Data" data), the location which can extend the predicted area most is chosen as the first additional sample and add this sample into the pool of existing samples; 2) we update the soil map based on the existing samples (including the newly selected additional sample); 3) if the resulted soil map still contains "No Data" locations or the "No Data" areas are still too large to be acceptable, select the location which has most incremental predicted area as the second additional sample. Repeat this process till the soil map is complete or the "No Data" areas are below an acceptable level (user-specified).

To maximize the reduction of the overall prediction uncertainty we adopt the following process: 1) based on the assumption that the more similar the environment conditions between two locations the more similar the soil property values, we select the additional sample representing the largest area in high uncertainty region and add this sample into the pool of existing samples; 2) update the prediction uncertainty map using the existing samples (including the additional sample just selected). Repeat the above process till the overall uncertainty is below a certainty threshold or the sampling is beyond the project budget. These additional samples for reducing the overall uncertainty are also stepwise because each sample is based on the uncertainty map generated from the last round.

The above uncertainty directed sampling method not only integrates the legacy samples but also provides as few stepwise samples as needed to fill the map gaps and reduce the overall prediction uncertainty. In addition, at the time of selecting a sample location using this design does not mean actually filed sampling at this location, because the selection for maximization of area covered or minimization of uncertainty is based on environmental similarity only. Calculation of environment similarity does not need the soil property value at the sample. In fact, the user(s) can wait till the locations of all the additional samples (based on the budget and or area coverage and or uncertainty reduction defined by users) are determined through the above process before starting the field sampling campaign so that samples can be collected at once in the field.

Download English Version:

https://daneshyari.com/en/article/6408381

Download Persian Version:

https://daneshyari.com/article/6408381

Daneshyari.com