ARTICLE IN PRESS
GEODER-12659; No of Pages 11
Geoderma xxx (2017) xxx–xxx
Contents lists available at ScienceDirect
Geoderma
journal homepage: www.elsevier.com/locate/geoderma

Identification of representative samples from existing samples for digital soil mapping

An Yiming [a,b], Yang Lin [a,c,*], Zhu A-Xing [a,d,e,f], Qin Chengzhi [a], Shi JingJing [a,b]

[a] State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing, China
[d] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China
[e] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[f] Department of Geography, University of Wisconsin-Madison, Madison, USA

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 19 October 2016
Accepted 11 March 2017
Available online xxxx

Keywords:
Digital soil mapping
Similarity based method under soil-landscape inference model (SoLIM)
Representative samples

ABSTRACT

Existing sample data are important for digital soil mapping. Different sample points possess different representativeness. The representativeness of samples influences the soil mapping result greatly. However, few study focus on assessing the representativeness of single sample. In this paper, we proposed a method to identify representative samples from existing samples collected from multiple resources. The basic idea of the method was to use clusters of environmental covariates to approximate types of soil variations, and check the occupancy of the existing samples in centroids of environmental clusters. Those samples locating in typical locations or centroids of environmental clusters were considered as representative samples. In this paper, the proposed method was used to discern representative samples in 282 soil samples in Anhui Province, China. SOM content was mapped using a similarity based mapping method. Two cases with different training samples (representative samples, non-representative samples, and training samples including representative and non-representative samples) and validation samples were set to compare the mapping results and accuracies. The results showed that the SOM content maps predicted using representative training samples had generally higher accuracy than the results produced using non-representative samples, and comparative accuracies with the results produced using full training samples. To discern representative samples is helpful for understanding the soil-landscape relationships in an area and the proposed method can be used to design supplementary samples for a better soil mapping result. Mapping results and accuracies showed that different training and validation sample sets impacted the mapping results and accuracies greatly, which indicates that researchers should be cautious when using randomization to obtain training and validation subsets for soil mapping.

1. Introduction

With soil surveys and related investigations in the past decades, point soil measurement data have been generated in many areas. These data are important for digital soil mapping (McBratney et al., 2003; Carré et al., 2007; Minasny and McBratney, 2016). Usually the existing sampling points are used directly without any assessment (Carré et al., 2007; Vaysse and Lagacherie, 2015). A current common way to soil mapping is to randomly split the total sample set into training and validation subsets without evaluating the representativeness of the samples (Gomez et al., 2016; Rodríguez-Lado and Martínez-Cortizas, 2015; Somarathna et al., 2016). Each soil sample point, as a product of complex interaction of soil forming factors, reflects the in situ relationship of soil to the environmental "niche" that it is in (McBratney et al., 2003; Zhu et al., 2015). Different sample points possess different representativeness, in terms of soil-landscape relationships. There exist locations where the soil-landscape relationships are typical and those of others' that are not (Zhu et al., 2008). As shown in Fig. 1, three rectangles represent three soil classes A, B and C in the environmental covariates domain (feature domain), respectively. Location E_2 is very close to the center of class A and E_1 is at the transitional area of three classes. It is obviously that E_2 represent a more typical soil-landscape relationship than E_1.

The representativeness (typicality) of samples influences the soil mapping result greatly. Take Fig. 1 as an example, it is possible to get a worse result to classify soil classes by using E_1 than using E_2. Therefore, it is necessary to assess the representativeness of the existing soil samples before using them, especially when the existing soil samples are from multiple sources using different sampling strategies.

There are few study to assess the representativeness of each existing sample. Carré et al. (2007) assessed the quality of legacy soil samples by

* Corresponding author.
E-mail address: yanglin@lreis.ac.cn (L. Yang).

http://dx.doi.org/10.1016/j.geoderma.2017.03.014
0016-7061/© 2017 Elsevier B.V. All rights reserved.

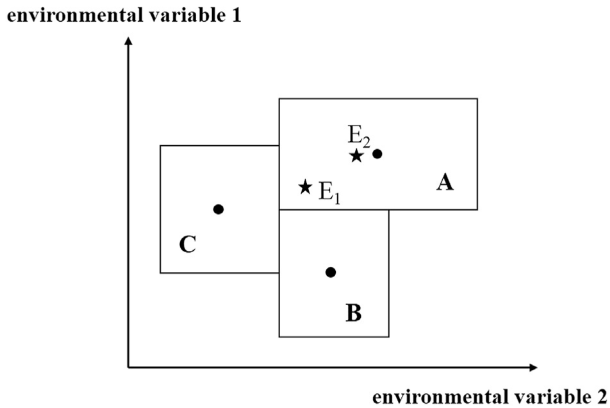

Fig. 1. Soil samples with different representativeness, modified based on Fig. 2 in Zhu, 1997.

checking if the density of samples in each covariate strata. The research evaluated the coverage of samples in the feature space but hardly paid attention to the representativeness of single sample. In addition, the impact of representativeness of samples on mapping results is hardly known.

It is usually difficult to discern the representative locations. Due to the high correlation between soils and environmental covariates, environmental covariates can be helpful to discern those typical locations (Zhu et al., 2008; Yang et al., 2012). This idea can also be used to select representative samples from the existing sample data.

The main objectives of this study are 1) to propose a method to discern representative samples from the existing samples, and 2) to compare the prediction accuracies by using representative samples, non-representative samples and training samples including representative and non-representative samples. A similarity-based mapping method was employed to map soil organic matter (SOM) in Anhui province (with an area of 1.34×10^5 km^2), China.

2. Material and methods

2.1. Study area

The study area is Anhui Province in China (29°23′44″N – 34°39′5″ N, 114°52′35″E – 119°39′37″) (Fig. 2). It is located in the central-south of China and covers an area of 1.34×10^5 km^2. The elevation ranges from 0 to 1806 m, where northern areas have a low relief, while southwestern and southern areas have a relatively high and variable terrain. The climate is warm and humid in summer and relatively cool and dry in winter, with an average annual temperature between 14 and 16 °C and an average annual precipitation between 750 and 2000 mm, mainly occurring in summer. There are five main soil orders in this area, Semi-hydromorphic soils in the northern area, Primitive soils, Anthropogenic soils and Eluvial soils in the central area, and Ferro-allitic soils in the southern areas (National Soil Survey Office, 1992). The parent materials in this area are various and complex, composed of granite, basalt, schist, perknite, diorite, sandstone, shale, limestone, conglomerate, tuff, mudstone and other types. The geological information shows a similar pattern from north to south with variation of terrain. Land use of this area includes cultivated land, forest, shrub and grass. The main crops in the cultivated area are rice, wheat, cotton, and rape. The study area has experienced intensive human activities, especially in flat areas.

Mapping soil organic (SOM) content of the top soil is the objective soil property in our study. SOM content significantly affects plant growth and fertile accumulation in this area. The spatial variation of SOM content is impacted by the geographical information, terrain and climatic conditions in this area according to knowledge of local soil experts.

2.2. Soil sample data and environmental covariates

2.2.1. Soil samples

There are 282 soil samples in total in the whole study area, which were collected with different investigation aims in the northern hemisphere summers and autumns of 2010, 2011, 2012 and 2015. Soil samples in this area are collected from two projects using different sampling strategies. One hundred twenty nine samples are collected based on experts' knowledge (hereafter referred to samples 2010) and 153 samples are collected using a stratified random sampling strategy across the whole area (hereafter referred to samples 2015). The spatial distribution of soil samples is shown in Fig. 2.

2.2.2. Environmental covariates

Climate, terrain, vegetation and parent material variables were chosen to characterize the environmental conditions for mapping soil organic (SOM) content of the top soil in Anhui province. Twelve environmental variables were generated including annual average temperature (TEMP), annual average precipitation (PREP), annual average evaporation (EVAP), annual accumulated temperature above 10 °C (ACC10), arid index (ARID), moisture index (IM), slope gradient (SLP), planform curvature (PLAN), profile curvature (PROF), topographic wetness index (TWI), Normalized Difference Vegetation Index (NDVI) and parent material (GEO). Table 1 lists the information of environmental data in detail.

Climate variables were from the Chinese Academy of Agricultural Sciences. The DEM data with a 90 m resolution was download from the Spaceshuttle Radar Topographical Mission (SRTM) website (https://search.earthdata.nasa.gov/). Slope gradient, planform curvature and profile curvature were calculated using the terrain analysis software 3DMapper (www.terriananalytics.com). TWI was calculated using the multiple flow direction strategy MFD-md by Qin et al. (2007); Qin et al. (2011) in SimDTA V1.0. The NDVI data was a Moderate Resolution Imaging Spectroradiometer (MODIS) vegetation index product with a resolution of 250 m downloaded from the website (https://search.earthdata.nasa.gov/) (NASA LP DAAC, 2016) . The NDVI used in this paper is the mean of NDVI products during 2010 to 2015. The parent material data was obtained from the 1:500, 000 geological map database of China (http://www.ngac.org.cn/) (The mapping group of the 1:500, 000 geological map database of China, 1999). There are 16 categories of parent material in the study area, as shown in Fig. 3.

Environmental layers with a 1 km resolution were resampled to 90 m to be consistent with the resolution of DEM using nearest neighbor assignment algorithm. All environmental layers were re-projected to WGS1984_UTM_50N projection system corresponding to the coordinates of sampling points.

2.3. The method to identify representative samples

A representative sample here refers to a sample that is typical of a soil class or a spatial variation type of soil property. The soil-environment relationships of representative samples reflect are typical. The basic idea of the method to discern representative samples was to use clusters of environmental covariates to approximate types of soil variations, and check the occupancy of the existing samples in centroids of environmental clusters. Those samples locating in typical locations or centroids of environmental clusters (Yang et al., 2010) were considered as representative samples.

Before clustering on environmental covariates, the pre-processing of the environmental variables including outliers removing and standardization was conducted (Zhu et al., 2008). Outliers were removed for each environmental covariate except for the parent material layer.

Please cite this article as: An, Y., et al., Identification of representative samples from existing samples for digital soil mapping, Geoderma (2017), http://dx.doi.org/10.1016/j.geoderma.2017.03.014