Available online at www.sciencedirect.com
ScienceDirect
Geoderma 140 (2007) 417 – 427
GEODERMA
www.elsevier.com/locate/geoderma

Methods to interpolate soil categorical variables from profile observations: Lessons from Iran

Tomislav Hengl [a,*], Norair Toomanian [b], Hannes I. Reuter [a], Mohammad J. Malakouti [c]

[a] European Commission, Directorate General JRC, Institute for Environment and Sustainability, TP 280, Via E. Fermi 1, I-21020 Ispra (VA), Italy
[b] College of Agriculture, Isfahan University of Technology, 84154 Isfahan, Iran
[c] Soil and Water Research Institute, PO Box: 114155-6185, Tehran, Iran

Available online 11 May 2007

Abstract

The paper compares semi-automated interpolation methods to produce soil-class maps from profile observations and by using multiple auxiliary predictors such as terrain parameters, remote sensing indices and similar. The Soil Profile Database of Iran, consisting of 4250 profiles, was used to test different soil-class interpolators. The target variables were soil texture classes and World Reference Base soil groups. The predictors were 6 terrain parameters, 11 MODIS EVI images and 17 physiographic regions (polygon map) of Iran. Four techniques were considered: (a) supervised classification using maximum likelihoods; (b) multinominal logistic regression; (c) regression-kriging on memberships; and (d) classification of taxonomic distances. The predictive capabilities were assessed using a control subset of 30% profiles and the kappa statistics as criterion. Supervised classification and multinominal logistic regression can lead to poor results if soil-classes overlap in the feature space, or if the correlation between the soil-classes and predictors is low. The two other methods have better predictive capabilities, although both are computationally more demanding. For both mapping of texture classes and soil types, the best prediction was achieved using regression-kriging of indicators/memberships (κ=45%, κ=54%). In all cases kappa was smaller than 60%, which can be explained by the preferential sampling plan, the poor definition of soil-classes and the high variability of soils. Steps to improve interpolation of soil-class data, by taking into account the fuzziness of classes directly on the field are further discussed.
Published by Elsevier B.V.

Keywords: Categorical variables; Terrain parameters; MODIS; Multinominal logistic regression; Fuzzy κ-means; Regression-kriging

1. Introduction

Soil-class data is a generic name for various categorical (multistate) soil variables commonly recorded by soil profile descriptions. A soil surveyor, or a soil expert, assigns classes directly in field or in office by following some mental or empirical model. Such field designations are in a way irreplaceable by a mechanical system, because they depend on the human understanding of soils, soil forming factors and soil use systems in all their complexity. Other soil-class designations can probably be replaced by a mechanical system, but these are commonly too expensive, too slow or too sensitive to field conditions (McBratney et al., 2003) The most common field designations are the taxonomic soil-classes, *e.g.* soil types or diagnostic horizons. Other groups of soil-class data are the *in situ* field-descriptive data such as texture classes, type of soil structure, drainage status, erosion groups *etc.*

The major problem with field designations is that they are rarely used to actually produce maps. Although soil-class field designations are often of unknown reliability, soil mappers would still like to use it for spatial prediction purposes. There are several good reasons for this: (a) it is less expensive to estimate the texture class by hand than by laboratory analysis; (b) due to high local soil variability, it is better to visit more locations than to allocate most of the budget on measuring a smaller number of highly precise analytical variables; and (c) much of the soil data available in the world today are of this type. Furthermore, the identification of some soil-related feature by an experienced surveyor conveys a lot of useful information

* Corresponding author. Tel.: +39 0332 785535; fax: +39 0332 786394.
E-mail addresses: tomislav.hengl@jrc.it (T. Hengl), norair@sepahan.iut.ac.ir (N. Toomanian), hannes.reuter@jrc.it (H.I. Reuter), SWRI@Sinasoft.net (M.J. Malakouti).

0016-7061/$ - see front matter. Published by Elsevier B.V.
doi:10.1016/j.geoderma.2007.04.022

on local conditions that simply cannot be replaced by any mechanistic system (Lark, 2005).

As with quantitative soil variables, digital soil mappers are more and more interested to interpolate categorical variables using semi-automated procedures. A standard geostatistical technique used to interpolate soil-class type variables is (multiple) indicator kriging (Goovaerts, 1997, p.301). For this reason, geostatistical analysis of categorical variables is by many referred to as the *indicator geostatistics* (Bierkens and Burrough, 1993). In practice, indicator kriging leads to many computational problems, which probably explains why there are not many operational applications of mapping soil-classes using geostatistics in the world (Hession et al., 2006). For example, it will typically be difficult to fit variogram for less frequent classes that occur at isolated locations (Fig. 1).

With the increasing availability of new sources of terrain and remote sensing based predictors, there is an increasing interest to include such exhaustively sampled auxiliary data in the interpolation of soil-class data. In recent years, pedometricians developed several novel concepts to produce maps of soil categories from profile observations by utilizing the auxiliary data. At least five distinct groups can be distinguished:

- *Pure classification techniques* — This is the equivalent classification of remote sensing images/bands for land cover classification or similar. Examples of classification of auxiliary maps to predict soil or landform types are given by Dobos et al. (2000) and Hengl et al. (2004). Note that the classification techniques are not interpolators as such, but lead to similar outputs as the regression techniques.
- *Pure regression techniques* — The most common statistical technique to predict soil categories is the multinomial logistic regression, which is an extension of (binomial) logistic regression to a case where there are more than two classes. Examples are given by Bailey et al. (2003) and Antonić et al. (2003).
- *Pure geostatistical techniques* — The simple multi-indicator kriging can also be extended to a case where several covariates are used to improve the predictions. This technique is known by the name *indicator (soft) co-kriging* (Journel, 1986). Although the mathematical theory is well explained (Bierkens and Burrough, 1993; Goovaerts, 1997; Pardo-Iguzquiza and Dowd, 2005), the application is cumbersome because of the need to fit a very large number of cross-covariance functions.
- *Hybrid statistical/geostatistical approaches* — One approach to interpolate soil categorical variables is to first assign memberships to point observations and then to interpolate each membership separately. This approach was first elaborated by De Gruijter et al. (1997) and then applied by Bragato (2004) and Triantafilis et al. (2001). An alternative is to first map cheap, yet descriptive, diagnostic distances and then classify these per pixel in a GIS. Examples are given by Carré and Girard (2002) and Hengl, (2003, §7). Another hybrid interpolator is the Bayesian Maximum Entropy (BME) approach by D'Or and Bogaert (2005). Li et al. (2004, 2005) further introduced the use of Markov-chain algorithms to enhance spatial simulation and modelling of soil categories. Although use of the BME and Markov-chain type of algorithms is a promising development, its computational complexity makes it still far from use in operational survey.
- *Expert systems* — Most experienced soil mappers do not believe that automated statistical techniques can be applied to map soil types (or similar) without mappers control of the system. Instead, the prediction of soil-classes should be based on empirical rules that are created and iteratively adjusted by the mapper. Zhu et al. (2001) demonstrated how auxiliary data can be combined with experts' knowledge to derive soil-class maps. MacMillan et al. (2003) used whole sets of semi-automated DEM classification rules to map soil series over large areas, even without doing any new field sampling. Although such techniques are not as computationally demanding, they rely heavily on the interpreters' knowledge of the soil-landscape combinations. In many cases, they also require extensive work on data preparation, *e.g.* to design classification rules and to adjust the final outputs.

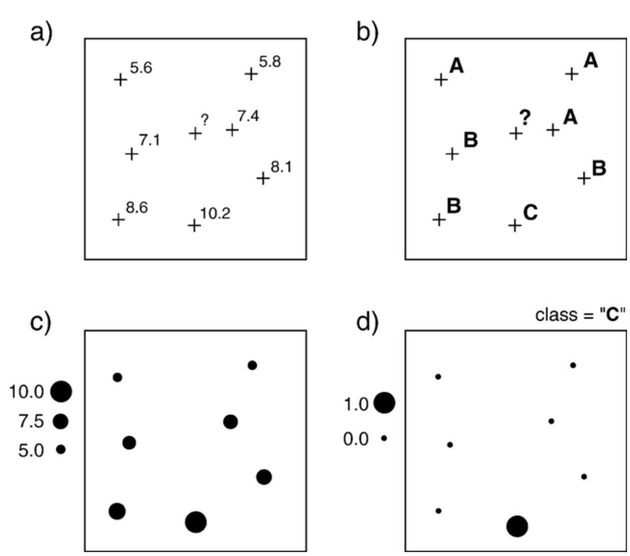

Fig. 1. Difficulties of interpolating soil-class data from point observations (b), as compared to quantitative variables (a), is that the soil-class interpolators are typically more complex and computationally more time-consuming. In addition, it is commonly easier to model spatial dependence structure for quantitative variables (c). In the case of indicators, modelling of the spatial dependence structure is especially difficult for isolated classes (d).

This research aimed to compare four existing soil-class interpolators using soil profile data from Iran as a model study. We were asked to recommend a methodology that can be implemented in other national and regional scale soil surveys to improve semantic and spatial detail of soil-class maps without repeating expensive and extensive field data collection.

Our objectives ranged from pure research to pure application. First, we were interested in evaluating the predictive capabilities of different soil-class interpolators. Second, we wanted to evaluate practical limitations of running data processing on an extensive national data-sets.