Geoderma 237–238 (2015) 237–245

Contents lists available at ScienceDirect

Geoderma

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

Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps

Romina Lorenzetti *, Roberto Barbetti, Maria Fantappiè, Giovanni L'Abate, Edoardo A.C. Costantini

Consiglio per la ricerca e la sperimentazione in agricoltura, CRA-ABP Agrobiology and Pedology Research Center, Florence, Italy

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form 18 July 2014
Accepted 4 September 2014
Available online 19 September 2014

Keywords:
Learning machine
Bayesian predictivity
Soil classification
Geomatic
Italy

ABSTRACT

The assessment of class frequency in soil map legends is affected by uncertainty, especially at small scales where generalization is greater. The aim of this study was to test the hypothesis that data mining techniques provide better estimation of class frequency than traditional deterministic pedology in a national soil map.
In the 1:5,000,000 map of Italian soil regions, the soil classes are the WRB reference soil groups (RSGs). Different data mining techniques, namely neural networks, random forests, boosted tree, classification and regression tree, and supported vector machine (SVM), were tested and the last one gave the best RSG predictions using selected auxiliary variables and 22,015 classified soil profiles. The five most frequent RSGs resulting from the two approaches were compared. The outcomes were validated with a Bayesian approach applied to a subset of 10% of geographically representative profiles, which were kept out before data processing. The validation provided the values of both positive and negative prediction abilities.
The most frequent classes were equally predicted by the two methods, which differed however from the forecast of the other classes. The Bayesian validation indicated that the SVM method was more reliable than the deterministic pedological approach and that both approaches were more confident in predicting the absence rather than the presence of a soil type.

1. Introduction

Soils vary in a rather complex manner across the landscape. Therefore, attribution of soil types to a map unit is always affected by some degree of uncertainty. The level of complexity and uncertainty of soil information within map units often increase as the map scale decreases, and this can bias or lower the confidence in interpretations (Kros et al., 1999). Considerable information may be lost due to map generalization and the results can only be considered rough estimates (Ibañez et al., 1998). Nevertheless, small scale maps are very important since they play a specific role in synthesizing soil information and representing a first partition of the soilscape, allowing comparisons on a broad scale and between disciplines (Ibañez and Feoli, 2013).

The traditional deterministic approach (DA) considers soil cover as partitioned into discontinuous patches of discrete classes (Soil Survey Division Staff, 1993). The mapper delineates soilscapes with relatively homogeneous expressions of soil-forming processes at the reference scale (Jenny, 1941, 1980). According to the guidelines for soil mapping provided by different systems, such as the Soil Survey Manual (Soil Survey Division Staff, 1993) and the WRB (IUSS working group WRB, 2014), soil types are classified as "dominant", "co-dominant" and "associated" depending on their estimated coverage in the polygon. The estimate can be based on expert judgment, field surveys, image analysis of aerial photographs or satellite images, by attribution of similar soils to similar soilscapes. Thus the spatial inference model of DA is the well-known "soil paradigm" (Hudson, 1992) based on knowledge on the deterministic effect of the factors of pedogenesis on the nature and distribution of soil in a studied environment. This knowledge, which can be more or less precise, is usually reported in the map report and can also be retrieved using a set of artificial intelligence techniques to improve both the spatial detail and the degree of attribute accuracy (Zhu et al., 2001). At smaller scales, the synthesis of landscapes drawn at more detailed scales is derived from a taxonomic and geographical generalization (Soil Survey Staff, 1999).

In contrast, digital soil mapping approaches can replace or better integrate the soil scientist's field survey and image analysis with different types of spatial inference models, such as geostatistical techniques and data mining, to link soil types or characteristics to the factors of pedogenesis and to assess the soil spatial distribution (Carré et al., 2007). McBratney et al. (2003) conceptualized this linkage in the so-called scorpan model:

$$Sc = f(s,\ c,\ o,\ r,\ p,\ a, n) + e$$

where: Sc is the soil class or attribute to be modeled, s refers to existing soil information, c is the climatic condition at the site, o is organisms, r is local relief, p is parent materials, a is soil age, n is space (perceived as spatial topology or spatial relationship) and e is the spatially correlated

* Corresponding author.

http://dx.doi.org/10.1016/j.geoderma.2014.09.006
0016-7061/

error. An advantage of statistical methods is the possibility to have a quantitative comparison of uncertainty in the prediction made by inference models (Lagacherie, 2008).

Several data mining techniques have been used to predict categorical soil properties at all scales. Moonjun et al. (2010) produced a predictive soil map showing soil taxonomy subgroups by applying neural network and decision tree algorithms to 57 soil observations on a surface of about 20 km^2. The two methods gave similar results when similar predictors were used. At a broader scale, Mendonça-Santos et al. (2008) adopted a decision tree algorithm to predict soil classes from environmental variables and a few soil profiles (only 431 in 44,000 km^2). They obtained the lowest error in the validation dataset by adding a legacy soil class map to the other predictors. Similarly, but at the national scale, Adhikari et al. (2013) constructed a soil class map of Denmark based on the FAO legend by applying a decision tree algorithm to 1170 soil profiles and 17 environmental variables. They found that clay content was the most important predictor, followed by geology. Instead, neural network was preferred by Oleg et al. (2003), who mapped the occurrence of seven WRB reference soil groups (RSGs, IUSS Working Group WRB, 1998) in forest soils of Croatia and obtained 63% correct correspondence. Hahn and Gloaguen (2008) demonstrated that SVM was more robust than linear classifiers in the prediction of soil type in a surface area of about 28 km^2 with 3218 training soils. The performance improved when coordinates were included among the input variables, since classes were well represented within the training datasets and showed a uniform distribution over the whole study area. Another data mining technique, namely boosted classification trees, was used by Lemercier et al. (2012) to predict natural soil drainage classes.

The reliability of the different mapping techniques were assessed by different approaches but never with the Bayesian methodology. However the Bayesian approach appears to be particularly suitable for comparing soil class maps, since it can produce a dual quantitative expression of uncertainty, i.e. positive and negative predictions of the occurrence of an event. Hence the Bayesian approach can allow the comparison of different maps in their estimate of a soil class being either present or absent in a polygon.

In the legend of the 1:5,000,000 maps of the soil regions of Europe (BGR, 2011) and of Italy (Costantini et al., 2012), the named soils were attributed to WRB classes through a deterministic approach. Our general goal was to devise a system able to follow the rules proposed in the WRB for creating map legends (IUSS working group WRB, 2014),while at the same time adding a confidence index of the class occurrence in the map legend. As data mining is capable of assessing the frequency of classes in the legend of soil maps, the specific aim of the present study was to use a Bayesian approach to test the hypothesis that data mining techniques can improve the reliability of the frequency order of WRB classes in the legend of the Italian 1:5,000,000 soil region map.

2. Materials and methods

2.1. Materials

2.1.1. The map of Italian soil regions

The 1:5,000,000 map of Italian soil regions produced by the deterministic approach (Costantini et al., 2012) (Fig. 1) is a component of Italy's Soil Information System (SISI, www.soilmaps.it, Costantini et al., 2013). SISI is a spatial data infrastructure that stores geographical and semantic information about soils and soil-forming factors, e.g. climate, geology, relief and land-use, at different scales (Table 1). The currently available soil geodatabases for all of Italy are those of the soil regions (1:5,000,000), subregions (1:1,000,000) and systems (1:500,000), while the soil subsystem geodatabase (1:250,000) is available for most of Italy. Thus far, soil unit (1: 50,000) and element (1:25,000) databases cover only a limited portion of Italy. Soil regions are intended to describe the soil geography at the European level (Finke et al., 1998), while subregions, systems and subsystems indicate national and regional dimensions, and soil units and elements are mainly of local interest. Italian soil regions, in particular, were created as a result of a comprehensive national effort, with a multi-author approach, which also involved regional soil services. The soilscapes resulted from a careful work of correlation and harmonization of both soils and landscapes carried out between regions and between them and the national level.

At present, the soil system geodatabase is the main completed soilscape level of SISI. Polygons are composed of areas that are homogeneous with regard to relief, lithology, drainage network and land cover at the reference scale. There are up to seven land components (LCs) in each polygon of a soil system. A "land component" of the soil system is a specific combination of morphological class, lithology and land cover. A linkage between geography and soil was created by allocating one or more soil typological units (STUs) to a LC, attributing a percentage of coverage. LCs were not delineated but their incidence in the polygon was quantified. The database stores 1413 STUs that summarize the information from the pedological observations. There were about 44,000 pedological observations in the database, including 26,942 classified and georeferenced soil profiles. The soil profiles of the database were collected from several national, regional and local investigations over many years, mainly since the 1980s. However, the information concerning environmental characteristics and chemical and physical properties were all standardized according to the manual of the database (Costantini, 2007). Moreover, the laboratory data were comparable since they were all obtained by means of the national official analysis methods (MIPAF, 2000).

The dataset used for the analysis consisted of all the sites with a WRB classification. Because of the origin of the database, they resulted clustered according to the specific study areas. However, this clustering of the data always occurs when a study is carried out without a dedicated sampling activity. Hence this case study can be considered representative of many national or other large datasets.

The geography and semantics of soil regions (SRs) were created through a generalization of the soil system database. The geography was generalized by considering the main climatic and lithological factors of pedogenesis, as reported in the European manual and adapted for Italy (Finke et al., 1998; Righini et al., 2001). The map legend was created by generalizing the information on the soil systems (SSs). Since a SR polygon contains several SS polygons, a SR has a set of STUs with an occurrence averaged on the LC percentage in the SSs. STU classification was generalized by considering the RSGs of WRB, and a list of the five most frequent RSGs was reported in the map legend.

2.1.2. Auxiliary variables ('scorpan factors') and software used

The influence on pedogenesis of the soil parent material was derived from a lithological map of Italy (Servizio geologico d'Italia, 1978, reference scale 1:500,000), the organisms from land-use (Corine land cover with a grid spacing of 100 m; De Jacher, 2012) and local topography from elevation and slope obtained from the digital elevation model (DEM) with 100 m spacing. Climatic and pedoclimatic conditions were based on the soil aridity index (SAI, Costantini and L'Abate, 2009). SAI expresses the mean annual number of days when the first 50 cm of soil is dry. Since SAI was obtained by multiple linear regression from the mean annual temperature, annual rainfall and available water capacity, it was used as a unique proxy for both pedoclimatic and climatic conditions. The spatial relationships among soils were determined from the geographical affiliation to a specific soilscape. The considered soilscapes were the soil regions and the soil systems, both extracted from SISI. Other soil information consisted of the map of carbon stock (Fantappiè et al., 2010), the map of soil inorganic carbon (Barbetti et al., 2012), and the mean clay and sand values of SSs. The considered soil variables were processed by taking into account all the analyzed soil observations of SISI, which were many more than the classified soils (Table 2).