Geoderma 285 (2017) 35–49

Contents lists available at ScienceDirect

Geoderma

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

A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization

Corrado Camera [a,*], Zomenia Zomeni [b], Jay S. Noller [c], Andreas M. Zissimos [b], Irene C. Christoforou [b], Adriana Bruggeman [a]

[a] Energy, Environment and Water Research Centre, The Cyprus Institute, Aglantzia, Lefkosia 2121, Cyprus
[b] Cyprus Geological Survey, Ministry of Agriculture, Rural Development and the Environment, Strovolos, Lefkosia 2064, Cyprus
[c] Department of Crop and Soil Science, Oregon State University, Corvallis, OR 97331, United States

A R T I C L E I N F O

Article history:
Received 23 February 2016
Received in revised form 13 September 2016
Accepted 21 September 2016
Available online 29 September 2016

Keywords:
Cyprus
Digital soil mapping
Model optimization
Random Forest
Soil landscape model
World Reference Base

A B S T R A C T

Fine-resolution soil maps constitute important data for many different environmental studies. Digital soil mapping techniques represent a cost-effective method to obtain detailed information about soil types and soil properties over large areas. The main objective of the study was to extend predictions from 1:25,000 legacy soil surveys (including WRB soil groups, soil depth and soil texture classes) to the larger area of Cyprus. A multiple-trees classification technique, namely Random Forest (RF), was applied. Specific objectives were: (i) to analyze the role and importance of a large data set of environmental predictors, (ii) to investigate the effect of the number of training points, forest size (*ntree*), the numbers of predictors sampled per node (*mtry*) and tree size (*nodesize*) in RF; (iii) to compare RF-derived maps with maps derived with a multinomial logistic regression model, in terms of validation error (test set and independent profiles) and map uncertainty, using the confusion index and a newly developed reliability index. The optimized RF model was run using half of the input points available (over a million) and with *ntree* equal to 350. The *mtry* parameter was set to 5 (close to half the number of the environmental variables used) for both soil series and soil properties. The *nodesize* calibration showed no relevant performance increase and was kept at its default value (1). In terms of environmental variables, the model used 10 predictors, covering all the soil formation factors considered in the scorpan formula, to derive the three maps. Soil properties, derived from geochemistry data, showed a high importance in deriving soil groups, depths and texture. Random Forest constructed a better predictive model than multinomial logistic regression, showing comparable predictive uncertainty but much lower validation error. The RF-derived maps show very low out of bag (OOB) errors (around 10% for both soil groups and soil properties) but relatively high validation error from independent profiles (45% for soil depth, 51% for soil texture). The resulting reliability index was low in the main mountainous area of Cyprus, where predictions were extrapolations as indicated by the multivariate environmental similarity surface, but medium to high in the main agricultural areas of the country.

1. Introduction

Environmental factors such as climate, organisms, relief, parent material, and time (*clorpt*) drive soil genesis (Jenny, 1941). Following this hypothesis, traditional soil survey maps are developed based on an empirical model (the soil-landscape model) derived from inductive reasoning from field and laboratory data, which represent the interchangeable relationships between soils and environmental factors. Digital soil mapping (DSM) techniques (McBratney et al., 2003) are based on the same hypotheses and aim at predicting soil types and properties linking field soil observations to environmental predictors. In DSM, inductive reasoning for developing the relationships among the soil-landscape model factors is replaced by different machine learning techniques (i.e., decision trees, fuzzy logic, neural networks etc.) (Lagacherie et al., 1995; Scull et al., 2003; Lagacherie et al., 2007; Grinand et al., 2008; Heung et al., 2016). However, pedological expert knowledge remains a key factor in model building to ensure both statistically and pedologically sound outputs (Kempen et al., 2009).

Digital soil mapping as a discipline has experienced a continuous expansion in the last two decades, mainly due to its increased efficiency in comparison to conventional field soil mapping techniques (Kempen et al., 2012). Reasons are the ever growing computational capacities coupled with the development of data-mining algorithms and GIS tools, and the increased availability of spatial remote-sensing data (Minasny and McBratney, 2016). Due to their numerical nature, digital soil maps also allow handling continuous spatial variations of soils, for example through class membership values, as presented by Burrough

* Corresponding author.
E-mail address: c.camera@cyi.ac.cy (C. Camera).

http://dx.doi.org/10.1016/j.geoderma.2016.09.019
0016-7061/

et al. (1997). This overcomes the problem of soil spatial patterns being traditionally captured and displayed as choropleth maps with discrete lines representing the boundaries between soil map units, which implies homogeneity within map units (Burrough, 1986; Bolstad et al., 1990). Digital soil mapping expands the notion of the soil-forming equation to that of a soil-mapping equation, the *scorpan* equation, which adds preexisting soil information and spatial location into Jenny's *clorpt* equation.

Random Forest (RF) is a fairly recent data mining algorithm (Breiman, 2001) that has been increasingly used for digital soil mapping applications in recent years. Its success is related to several advantages over other statistical (e.g. linear regression or generalized linear models), geo-statistical (e.g. regression or co-kriging), and machine learning (e.g. neural networks, logistic regression, support vector machines, classification trees) techniques. These advantages have been summarized by Grimm et al. (2008): ability of modelling high dimensional non-linear relationships; simultaneous handling of categorical and continuous predictors; robustness against over-fitting; measures of error rate and variable importance; requirement of only three user-defined input parameters; and relatively low sensitivity to parameter values. In particular, the measure of variable importance has proved to be, in many circumstances, a useful tool for enlightening soil-environment relationships to allow authors to infer the effects of possible future environmental changes on soil characteristics (e.g. Barthold et al., 2013).

Among its many applications, RF has been used for predicting the spatial distribution of various soil properties, such as soil organic and/or inorganic carbon (e.g. Grimm et al., 2008; Wiesmeier et al., 2011; Poggio et al., 2013; Akpa et al., 2016; Sreenivas et al., 2016); soil texture and cation-exchange capacity (e.g. Lagacherie et al., 2013; Chagas et al., 2016); and soil taxonomic units in unmapped areas (Stum et al., 2010; Barthold et al., 2013; Pahlavan Rad et al., 2014; Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015; Heung et al., 2016; Lang et al., 2016). Heung et al. (2014) pointed out that few studies have applied RF for mapping categorical soil properties, specifically referring to soil taxonomic units. In the last 3–4 years this gap has started to get filled.

Recent soil classification studies mainly deal with the comparison of the performance of many different algorithms (including RF) and sampling techniques (e.g. Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015; Heung et al., 2016), partially disregarding model building and model optimization. Models often include continuous variables representing topography, climate, vegetation or land use from remote sensing products (e.g. Pahlavan Rad et al., 2014; Brungard et al., 2015; Heung et al., 2016), sometimes categorical variables representing parent material (e.g. Barthold et al., 2013; Taghizadeh-Mehrjardi et al., 2015), but only Taghizadeh-Mehrjardi et al. (2015) and Lang et al. (2016) also consider soil information and properties. However, none of these authors clearly quantifies the role and the importance of these different predictors in the model. An optimization of the number of trees in the forest and of the number of variables to be used to split branches is quite typical for RF (e.g. Grimm et al., 2008; Barthold et al., 2013; Heung et al., 2016). Conversely, the investigation of the effect of tree pruning, which is common for classification tree and boosted classification tree modelling approaches (Scull et al., 2005; Schmidt et al., 2008; Lemercier et al., 2012), has not been extensively reported in RF soil mapping applications. Finally, Barthold et al. (2013) and Taghizadeh-Mehrjardi et al. (2015) are the only authors explicitly relating soil groups to major soil properties such as soil depth and soil texture, although they do not derive each of them independently. These properties are of major importance for application studies such as agricultural crop modelling and soil erosion (Bird et al., 2016; Djuma et al., 2016). There is also a paucity of studies dealing with soil prediction in complex topographical and pedological environments and in the Eastern Mediterranean region.

The main aim of this study is to develop digital soil maps of the soil groups, depth and texture classes of a topographical and pedological complex area of the eastern Mediterranean, namely the island of Cyprus, based on extensive soil legacy data. Specific objectives are: (i) to analyze the role and importance of a large data set of environmental predictors, covering all the soil formation factors considered in the *scorpan* formula, both for single and groups of predictors; (ii) to investigate the effect of number of training points, forest size (number of trees), number of predictors sampled at each node, and tree size (terminal node) in RF; (iii) to compare RF-derived maps with maps derived with a multinomial logistic regression model, in terms of validation error and map uncertainty, using the confusion index and a newly developed reliability index.

2. Materials and methods

2.1. Study area

Cyprus is the third largest island in the Mediterranean and is located between 34–36°N and 32–35°E. The main physical characteristics of the island are represented by the two mountain chains, the Troodos, located in the central-west part with the highest peak at Mount Olympus (1951 m a.s.l.), and the Pentadaktylos Range along the north coast with its highest peak at Mount Kyparissovouno (1024 m a.s.l.). The main agricultural area of the country is the Mesaoria Plain, which lies in between the two mountain ranges and the coastal lowlands.

Soils on Cyprus are exceptional due to the geological complexity of the island, the Mediterranean climate and the long presence of man on the landscape. The Troodos Ophiolite, a fragment of fully developed oceanic crust, consisting of Turonian plutonic, intrusive and volcanic rocks and chemical sediments dominates the central topographic high of the island. Older allochthonous rocks are juxtaposed in the southwest (Mamonia Terrane, Middle Triassic – Middle Cretaceous) and the long east-west Pentadaktylos range in the north coast (Keryneia Terrane, Carboniferous – Middle Miocene). Autochthonous carbonate sediments cover the slopes and plains. Quaternary deposits are predominately of gravity and fluvial origins inland and of marine and aeolian origins on the coast. The soils on Cyprus vary between leptosols, regosols, solonchaks, solonetz, vertisols, luvisols, fluvisols, and cambisols based on the World Reference Base of the FAO (Food and Agriculture Organization of the United Nations) soil classification system (IUSS, 2015). They are generally poor in organic matter (Koudounas and Makin, 1978; Grivas, 1988) and closely associated to parent material and landscape position (Zomeni, 2012; Zomeni and Bruggeman, 2013). Thin (leptic) and stony (lithic) soils dominate the mountainous areas developing mostly as residuum. Other soils form on transported materials such as alluvial deposits (alluvial fans, fluvial terraces and deltas), colluvial deposits, aeolian deposits, marine deposits (sands and gravels) and lake and estuarine deposits (hydromorphic silts and clays).

The geochemistry of the island also reflects the geological complexity and the impact of humans. A recent high sampling density (5350 sites on a 1 km^2 grid), multi-element (60 elements) and multi-method analysis soil geochemical survey has resulted in the compilation of the Geochemical Atlas of Cyprus (Cohen et al., 2011, 2012a). The survey was carried out at two depths. Surface soil samples were collected at a depth of 0–20 cm and bottom samples at a depth of 50–70 cm. The survey has demonstrated that chemical processes and element concentrations are dominated by parent lithology. Other processes such as the physical concentration of heavy minerals (Ren et al., 2015), ocean influences along the coastal plains, and human activities also affect the spatial geochemical patterns of the soils on the island (Cohen et al., 2012b; Zissimos et al., 2014). For the purpose of this study we have calculated geochemical parameters using data from surface soil samples.

The present study covers the areas under the effective control of the government of the Republic of Cyprus, with the addition of the UN buffer zone (tot. 5979 km^2), where data are available (Fig. 1).