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

Corrado Camera[1], Zomenia Zomeni[2], Jay S. Noller[3], Andreas M. Zissimos[2], Irene C. Christoforou[2], Adriana Bruggeman[1]

[1] *Energy, Environment and Water Research Centre, The Cyprus Institute, Aglantzia, Lefkosia, 2121, Cyprus*

[2] *Cyprus Geological Survey, Ministry of Agriculture, Rural Development and the Environment, Strovolos, Lefkosia, 2064, Cyprus*

[3] *Department of Crop and Soil Science, Oregon State University, Corvallis, Oregon, 97331, United States*

Corresponding author: Corrado Camera

Email: c.camera@cyi.ac.cy

Telephone: +357 22208691

Fax: +357 22208625

Highlights

- We created a digital soil map of Cyprus (1:25,000) using Random Forest
- We developed a reliability index to qualitatively show prediction uncertainties
- The model showed greater sensitivity to the number of input points than trees grown
- Soil geochemistry predictors had a prevalent role in identifying soil properties

Accepted manuscript version of the paper published in Geoderma

(https://doi.org/10.1016/j.geoderma.2016.09.019)

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

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 tree, fuzzy logic, neural network 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 algorithm 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 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 modeling 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; Láng 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. Berthold et al., 2013; Taghizadeh-Mehrjardi et al., 2015), but only Taghizadeh-Mehrjardi et al. (2015) and Láng 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., under review). 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 34-36ºN and 32-35ºE. 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 (1,024 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 (5,350 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. 5,979 km^2), where data are available (Fig. 1).

Fig. 1. The island of Cyprus with its main physical characteristics and the location of the study area.

2.2 Soil data

The most detailed soil references on the island are ten 1:25,000 scale soil sheets prepared between 1967 and 1985 by the Soil Section of the Department of Agriculture, using traditional survey methods (Soteriades and Georgiades, 1967, Soteriades and Grivas, 1968, Soteriades et al. 1968, Soteriades and Markides, 1969, Grivas and Georgiades, 1972, Markides, 1975, Koumis, 1980a, Koumis, 1980b, Koumis, 1980c, Markides, 1985a). The soils were mapped and classified based on their development stage, origin and parent material. These sheets are always accompanied by an agricultural land suitability map and two of them - the Pafos sheet (Soteriades and Koudounas, 1968) and the Polemi sheet (Markides, 1973) - have an extensive soil memoir.

The 10 sheets cover around 1,600 km^2, classified in 369 soil sub-series (52 soil series with local soil names). Based on soil-profile descriptions, provided in the legend of the map sheets or in the available soil memoirs, we independently harmonized soil (sub)series, soil depth, and soil texture. The harmonization of the soil series was based on the World Reference Base (WRB) for soil resources (IUSS, 2015) and led to recognize 31 soil classes (soil groups accompanied by one or two qualifiers) and two miscellaneous classes, namely Water Bodies and No Data (e.g. quarries and excavated areas). Soil depth was harmonized in four classes, based on the depth ranges available in the original maps (Table 1). Soil texture (Table 2) was

harmonized in four classes (None, Coarse, Medium, Fine) following the guidelines provided by FAO (2008). As only differences, we grouped all sandy loams as coarse textured and all clay loams as medium textured. In FAO (2008) these two classes were split between coarse-medium and medium-fine, based on the relative abundance of sand and clay, which information is rarely available in our legacy data. The soil profiles usually included: depth class; texture class; color; lithic properties and their quantity and composition; bedrock geology; horizon descriptions; and chemical properties (the latter two available only in the two soil memoirs). The soil classes, soil depth and soil texture maps were digitised and converted to raster format with cell size of 25x25 m^2 (2,561,849 cells in total) and used as training data for DSM purposes. The training areas, shown in Fig. 2, cover 17% of the island and 27% of the study area.

Additional mapping activities were carried for different sheets (Morphou, Polis-Tylliria, Krasochoria, Pissouri-Paramali, Limassol, and Larnaka) of the Land Suitability Map of Cyprus (Grivas, 1969; Markides, 1969; 1985b; Koumis, 1970a; 1970b; 1970c). These studies provided 126 profiles, described in terms of both soil depth and soil texture, outside the training areas that we used as independent validation observations (see Section 2.4.4 - Map extrapolation, prediction uncertainty, map reliability, and independent profile evaluation). Additional 199 validation profiles, reporting data only on soil depth, were derived from Robins (2004). The author assessed the soil depth variability of the Troodos Mountains at road cuts in three valleys on the north slope of the mountain chain. The location of the validation profiles are presented in Fig. 2.

Fig. 2. Location of the study area, the existing 1:25,000 scale soil sheets used for training, the 126 soil profiles derived from Land Suitability Maps (LSM), and the 199 soil profiles from Robins (2004) used to evaluate the digital soil map of Cyprus.

Since Water Bodies and No Data are not proper soils and represent mainly human modification of the environment, we superimposed these data layers (of known location) on the map resulting from the applied computational methods (see section 2.4). For both Water Bodies and No Data, soil depth and soil texture were considered not applicable (N/A).

Table 1. Soil depth classes identified from the 1:25,000 scale soil sheets and reclassified in four consistent classes.

Table 2. Soil texture classes identified from the 1:25,000 scale soil sheets and reclassified in four consistent classes. The harmonization has been performed using the guidelines provided in FAO (2008).

2.3 Environmental covariates

A general summary of the methodology applied in this study is presented in this and the following sections (Fig. 3). For the present study, predictors are environmental covariates that have been selected based on the *scorpan* formula (McBratney et al., 2003). The selected 20 variables (17 continuous, 3 categorical) are described in the following paragraphs. A correlation analysis was performed on the 17 continuous predictors to reduce the dimensionality of the input dataset by removing redundant features. Building a parsimonious model brings two different advantages, as explained by Behrens et al. (2010). First, it helps recognize soil formation processes more clearly, and second it usually yields more reliable predictions.

Fig. 3. Summary of the methods used in the study. OOB is the Out Of Bag error, MESS is the Multivariate Environmental Similarity Surface, and CI is the Confusion Index.

2.3.1 Soil chemistry and organism covariates

To describe soil chemistry we selected Electrical Conductivity (EC), pH, Total Carbon (TotC), Organic Carbon (OrgC), Loss on Ignition (LOI), Mafic Index of Alteration (MIA), Chemical Index of Alteration (CIA), and Vogt Ratio (VR). All these covariates were obtained from the Geochemical Atlas of Cyprus (Cohen et al., 2011; 2012a) by using the raw data, which were collected on a regular 1 km^2 grid, and re-interpolating them by Inverse Distance Weighting (IDW) over the 25 x 25 m^2 grid. Both TotC and OrgC were measured using an Eltra CS-800 automatic Carbon–Sulfur analyser, while LOI was measured on soil sub samples at 1000° C. The Mafic Index of Alteration (MIA) was selected to account for the chemical changes during weathering of mafic lithologies (Babechuk et al., 2014), while the CIA and the VR are ratios that quantify the loss of mobile major elements relative to immobile elements during weathering. The three indices were calculated as follows:

$$MIA = 100 \cdot \left[\left(Al_2O_3 \right) / \left(Al_2O_3 + Fe_2O_3 + MgO + CaO + Na_2O + K_2O \right) \right] \quad (1)$$

$$CIA = Al_2O_3 / (Al_2O_3 + Na_2O + K_2O + CaO) \quad (2)$$

$$VR = (Al_2O_3 + K_2O) / (MgO + CaO + Na_2O) \quad (3)$$

For further information on the analytical methods used to obtain the geochemical data as well as the spatial variability of the used soil chemistry covariates see Cohen et al. (2011, 2012a) and Zissimos et al. (2014).

OrgC data can be considered a proxy for the presence of soil biota like micro-organisms and plants (Yeomans 1988; Soon 1991; Panagos et al 2008), In addition, we considered land use as an organism covariate. We derived this parameter from the CORINE land cover data set of 2006, which has a 1:250,000 scale (Büttner and Kosztra, 2007). We use categories from the Level 1 of the database except for "Forest and semi-natural areas", which have been differentiated to Level 2.

2.3.2 Climate covariates

As climate covariates, we included minimum, maximum and average temperature (T_{min}, T_{max}, T_{ave}). We calculated these parameters over each 25 x 25 m^2 cell from the daily gridded temperature data set developed by Camera et al. (2013). The data set has a horizontal resolution of 1 km and covers the period 1980-2010. Therefore, the smaller cells of the present study were assigned the value of the coarser cell by nearest neighbour resampling. T_{min} and T_{max} were calculated as the mean of the daily minimum and maximum temperatures recorded in January and July, respectively. They represent the influence of the hottest and coldest temperatures on soil genesis (Scull et al., 2005). T_{ave} is the mean annual average temperature.

2.3.3 Relief covariates

Relief covariates included: elevation from a Digital Elevation Model (DEM); aspect; curvature; planar curvature; profile curvature; slope; and landscape units. The DEM has a horizontal resolution of 25 x 25 m^2 and was created from digitised contour and point data of published 1:50,000 scale topographical maps using the Topogrid command in ArcGIS®, which is based on the ANUDEM gridding application by Hutchinson (1993). Landscape units were derived on the basis of the DEM, by generating the Shary classes (Shary, 2008) of topographic curvature and form, wetness index, topographic complexity index, in order to

classify the landscape in 21,794 polygons and 13 categorical landscape units based only on surface morphology. We calculated all the other covariates from this digital elevation model by also using ArcGIS® functions.

2.3.4 Parent material and age covariates

In this group of covariates we selected to use a map of geological formation grouped in 10 classes based on lithology, depositional environment and relative age. This map was prepared specifically for this study as a summary of two different layers: bedrock geology, as derived from a number of geological maps of the Cyprus Geological Survey, of various scales, published and unpublished, digitised, merged and harmonised; and quaternary geology, as derived from recent field, aerial images and satellite remote mapping (Noller, 2009).

2.4 Computational methods

The soil groups and soil property maps were derived for the study area using RF, as implemented in R (www.r-project.org) within the randomForest package (Liaw and Wiener, 2014) and multinomial logistic regression, as implemented in package nnet (Venables and Ripley, 2002). All the random subsets used as model input were created using the caret package (Khun, 2008), preserving the original class distribution of the soil classes. All computations were performed on a single node of the CyTera HPC facility (http://cytera.cyi.ac.cy/) consisting of 12 cores with 4 GB RAM each.

2.4.1 The Random Forest algorithm

Random Forest is a multiple tree classification and regression algorithm (Breiman, 2001) that can be used to predict a target variable at locations where it is unknown, on the basis of previously defined relationships between the target variable itself and identified predictors. In this study, we applied RF as a classification tool. A clear overview of the method's functioning is presented by Diaz-Uriate and de Anders (2006) and Boulesteix et al. (2012).

To construct the relationships between the target variable and the predictors, many classification trees are grown. Besides the number of trees (*ntree*), only two other parameters can be defined by the user before running a RF classification: the minimum size of terminal nodes (*nodesize*), and the number of variables randomly sampled as candidates at each split (*mtry*). Each tree is a standard classification tree. At each node the code randomly samples a number (*mtry*) of predictors and among these it picks the predictor that ensures the best split. Each target point is then classified by aggregating the trees and picking the class that received the maximum number of votes.

Bootstrap samples from the original data set are used to build the trees; this means that some observational values are not used to construct the trees and can be grouped to form the out-of-bag (OOB) sample. This sample can be used for validation purposes by comparing it to the model outputs and calculating the corresponding relative error (OOB error). An additional feature of RF is the capacity to rank the relative importance of the variables in the prediction. In particular, we used the Mean Decrease of Accuracy (MDA) value. It is calculated as follows: first, the original OOB sample is run down a tree and the number of votes for the correct target class is kept. Second, the values of a variable are randomly permuted in the OOB sample, which is then run down a tree, saving the number of votes of the correct class. Finally, the difference between the two vote numbers is calculated, and the results from the different trees are averaged (Breiman, 2001).

2.4.2 Random Forest model parameterization and optimization

The RF classification model optimization was carried out in four steps. In the **first step**, we identified the best combination of number of trees to be grown and number of training points (raster cells) to be used, considering the evolution of the average OOB error and the computational resources available. The number of training points was expressed in terms of percentage of the total number of observation points from the 10 available maps.

As a **second step**, we quantified the role of each predictor and group of predictors. Firstly, we evaluated the importance matrix (MDA) calculated by the RF algorithm run using all the covariates not excluded in the correlation analysis. For this purpose we did not scale the variable importance (as default in RF), since this can affect the interpretability of the MDA investigation, as suggested in Strobl et al. (2007). Secondly, we calculated the modification in the OBB error when removing predictors one by one, and group by group (e.g. relief, climate, geochemistry predictors etc.), from the complete set. In addition, similar to Xiong et al. (2014) and Brungard et al. (2015), we used recursive co-variable elimination (based on importance and OOB error modification) to derive the subset of covariates leading to the lowest OOB error. The *nodesize* and *mtry* parameters were kept to their default values for the first two steps.

In the **third step**, we calibrated the *mtry* parameter leaving again *nodesize* to its default value. We attributed to *mtry* all the odd values between 1 and the maximum number of predictors coming from the optimized subset of covariates and selected the value giving the minimum OOB error.

In the **fourth step**, we investigated the effect of tree pruning by modifying the value of the *nodesize* parameter setting it to 1 (default), 4, 12, and 20. We selected the best value based on the minimum OOB error, minimum average uncertainty value, and minimum independent validation error. The computation of uncertainty value and validation error is explained in section 2.4.4 (Map extrapolation, prediction uncertainty, map reliability, and independent profile evaluation).

2.4.3 Multinomial Logistic Regression model

To determine the quality of the maps derived with RF, we compared its outputs with those of an alternative, more standard technique from the generalized linear model family. We

selected a multinomial logistic regression (MLR) model, since it has been proven a reliable method by many authors (Debella-Gilo and Etzelmüller, 2009; Kempen et al., 2009; Jafari et al., 2012; Collard et al., 2014).

MLR is the generalization of the binomial logistic regression, where multiple possible outcomes are derived from a set of predictors to which coefficients are applied. In the binomial case, two possible outcomes (A_1, A_2) are given and the probability of occurrence (π) of the second outcome is $\pi_2 = 1 - \pi_1$. Logistic regression relates probability π_1 to a set of predictors using the logit link function:

$$\log it(\pi_1) = \ln\left(\frac{\pi_1}{\pi_2}\right) = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \mathbf{x}'\boldsymbol{\beta} \quad (4)$$

where $\mathbf{x}$ is the vector of predictors and $\boldsymbol{\beta}$ is the vector of model coefficients. In case of categorical variables, every class level is fitted with a different coefficient. Model coefficients are usually estimated through maximum likelihood. From eq. 4 it is possible to obtain:

$$\pi_1 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} \quad (5)$$

which can be generalized, for a multinomial case with K target classes, to:

$$\pi_k = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\sum_{k=2}^{K}\exp(\mathbf{x}'\boldsymbol{\beta}_k)}. \quad (6)$$

One class is selected as reference and logits are then calculated that compare other classes to it. The constraint is that the summation of the probabilities of the K classes must add up to 1.

In this study, for MLR model building we used the same training points and set of predictors optimized for RF. To compare the two methods we analyzed prediction error in comparison to original soil map data, prediction error based on independent soil profiles, and prediction uncertainty. Map errors were calculated based on a test set of points not used for model generation for MLR and based on OOB errors for RF, since OOB errors can be considered a

valid alternative to independent validating data sets (Grimm et al., 2008). The computation of the predictive uncertainty and independent validation of both maps are described in the next section.

2.4.4 Map extrapolation, prediction uncertainty, map reliability, and independent profile evaluation

We analysed different indices to evaluate and discuss the quality of the obtained soil maps. Firstly, we calculated the Multivariate Environmental Similarity Surface (MESS), as presented by Elith et al. (2010), based on all continuous predictors and using the dismo R package (Hijmans, 2013). MESS identifies and quantifies the areas where model predictions are extrapolations by measuring the similarity of any prediction point to the used training points, with respect to the chosen predictor variables. The higher the score, the more common the environment of the point is. Negative values indicate a novel environment, i.e. the presence of at least one variable outside its range in the training points. Categorical predictors are not handled by the MESS algorithm. However, they would not have affected the resulting map, because we had reclassified all categorical variables to have all classes represented in the training areas. Secondly, we evaluated the uncertainty of the prediction with the confusion index (CI) as presented in Burrough et al. (1997):

$$CI = \left[1 - \left(\mu_{\max i} - \mu_{(\max-1)i}\right)\right] \quad (7)$$

where $\mu_{\max}$ represent the probability value of the class with the maximum odds at cell i and $\mu_{\max\text{-}1}$ is the second largest probability at the same cell. The higher the CI is, the higher is the confusion, i.e. increasing similar probabilities for the two classes. For RF the probabilities were calculated from the number of tree votes, while for MLR they came from a probability prediction.

We developed a reliability index for the created soil maps (groups, depth, texture) with the aim of qualitatively evaluating a combination of model error and extent of extrapolation limited to original data density. The reliability index we derived is based on the ratio of the number of training points and the number of points in the final map per class (unknown areas potential extrapolation error), and the error in the training map areas not used for model construction (OOB error for RF and test set error for MLR). We implemented a two-step approach to derive three reliability classes (High, Medium, Low). We calculated a reliability index for each class of the three maps produced, according to the following equation:

$$R_{ij} = \left(1 - MAP_{ij}\right) \cdot 2 \frac{P_{Tij}}{P_{Mij}} \quad (8)$$

where R is the reliability index, MAP is the error in the training map areas not used for model construction, P_T is the number of training points, P_M is the number of predicted cells in the derived maps, i is one of the three maps (soil groups, soil depth, soil texture), and j is a class of one of the three maps. The coefficient 2 is introduced because only half the points in the training map areas were used to train the models. A perfect reliability of 1 would therefore be achieved with a MAP error equal to zero and a P_{Mij} value double P_{Tij}, meaning that all the points inside the training area were well predicted, and the specific class is not extrapolated to unknown areas. We decided to limit the value of Rij between 0 (minimum reliability) and 1 (maximum reliability), meaning that also classes under-represented by the prediction in comparison with the training areas have a maximum reliability of 1. For mapping purposes, we then classified the R values into three classes: Low ($R < 0.25$), Medium ($0.25 \leq R \leq 0.50$), High ($R > 0.50$).

The created soil depth and soil texture maps were also evaluated through 325 and 126 independent soil profiles (Fig. 2), respectively. The profiles are those described in Section 2.2 (Soil data). For each profile from the Land Suitability Map, we identified the maximum depth

and then we attributed it a soil depth class according to Table 1. In these profiles, clay, silt, fine sand, and coarse sand percentages were specified for three or four depth intervals. We classified the texture of each depth according to the soil textural triangle (FAO, 2006; IUSS, 2015), and then we attributed texture classes, according to Table 2. Robins (2004) provides depth ranges for each road cut analysed. In case the range fell in two or more depth classes as presented in Table 1, we assigned a range of classes to the validation point. Depth and texture of the profile were than compared with those of the predicted maps, at the exact location. For soil depth from Land Suitability map profiles, we evaluate the point as a perfect match (1) or a complete miss (0). For texture we evaluated the point as a perfect match (1) if the prevailing texture was correctly mapped, a partial match (0.5) if the predicted texture matched the texture of at least one of the depth intervals, and a complete miss (0) in case of no match at all. Similarly, for soil depth over the profiles of Robins (2004) we evaluated a perfect match (1) if the average depth of the range was falling in the predicted depth class, a partial match (0.5) if the mapped depth class covered, at least partially, the depth range, and a complete miss (0), if the predicted class was out of range. An average hit ratio (HR) was then derived summing the contribution of each profile and dividing by their total number.

3 Results and discussion

3.1 Covariates correlation analysis

An assessment of a 17 x 16 matrix of correlation scatter plots and Pearson correlation coefficients led to the discard of seven out of the 17 continuous covariates presented in Section 2.3. We removed single covariates out of couples presenting a correlation coefficient higher than 0.6. Loss on Ignition, Total Carbon, the Chemical Index of Alteration, the Vogt Ratio and the Mafic Index of Alteration (MIA) are all highly correlated with each other. We

preferred keeping only the MIA because it allowed a broader differentiation of the main lithologies of the study area. We removed planar curvature and profile curvature because they were both highly correlated with curvature. Finally, we removed average temperature because it was highly correlated with minimum temperature and showing a larger correlation coefficient with T_{max} (0.56) than T_{min} (0.24). Fig. 4 shows six out of the seven pairs of correlations just described. The relation between TotC and MIA is not shown because very similar to that of LOI and MIA, also in terms of correlation coefficient (-0.88).

Fig. 4. Scatter plots, drawn with a 0.05% subset of the original input dataset, showing highly correlated covariates. The R value shown in each graph is the Pearson correlation coefficient calculated using the full input dataset. The covariates on the x-axis were removed from the data set used as input for the digital mapping techniques. CIA is the chemical index of alteration; MIA is the Mafic Index of Alteration; VR is the Vogt Ratio; and LOI is the Loss on Ignition.

3.2 Model parameterization and optimization

The results of the **first optimization step** (number of target points and number of trees) are presented in Fig. 5. The results concern the soil groups model. On the one hand, the graph shows how the error depends mainly on the number of target points rather than on the number of trees, especially for forests larger than 200-250 trees. On the other hand, using too many training points (>50%) reduces the maximum number of trees that we can calculate without a memory failure of the system to less than 300. For this reason, we performed our analysis using 50% of the training points and building forests of 350 trees. This guaranteed a satisfactory low average OOB error (around 10%) and a certain robustness of the classification model.

Fig. 5. Average OOB error for increasing number of trees in the Random Forest classification model and different number (% of the total) of training points as derived from the rasterized 1:25,000 scale soil maps.

The results of the **second step**, for the systematic removal of single covariates are reported in Table 3. Table 3 shows that there is generally a good agreement between the MDA value calculated by the RF algorithm for each variable and the decrease/increase of the average OOB error when removing a certain predictor. The most illustrative case is that of the elevation (DEM) covariate. In fact, it appears as the most important covariate in terms of MDA, and the predictor causing the largest increase of OOB error when removed for all the three maps. However, a high decrease/increase in the OOB error does not always correspond to a low/high value of MDA. The most enlightening case (soil groups prediction) is the one of the pH, which ranked seventh based on the MDA results, while being the variable causing the third highest increase of OOB average error when removed. The opposite can be seen for Maximum Temperature, which was for all the three maps among the four most important variables for MDA. However, if removed it caused a decrease in the OOB error.

Table 3. Average OOB error for forests (350 trees, 50% data) built removing single variables, and importance value (Mean Decrease of Accuracy - MDA) calculated by removing the variable from the complete model. OOB values are presented as differences from the model derived with the complete set of variables. Between parenthesis the rank of each variable.

Mean Decrease of Accuracy has a great advantage over the analysis of the OOB error variations due to variable removal; it can cover the impact of each predictor also in terms of multivariate interactions with the other variables (Strobl et al., 2008). However, the same

authors point out how often the MDA of correlated variables is overestimated due to a preference for correlated predictors in the early splits. This can explain why Maximum Temperature, which from the previously performed correlation analysis showed a Pearson's correlation coefficient with elevation of -0.46, presents such differences in behaviour in Table 3. When using the original RF algorithm as implemented in the randomForest package in R, it is therefore very useful combining an analysis of the MDA values with a correlation analysis of the predictors involved and an analysis of OOB errors following single variable removal, to identify potential anomalies.

Table 4 shows the importance of the group of soil chemical covariates. The removal of the five selected soil chemical properties from the complete model caused an increase in the error of more than 9% for soil groups, 12% for soil depth, and 9% for soil texture. The removal of the other four groups of covariates leads to maximum error increases of 3.5%, 3.6%, and 2.9% for soil groups, soil depth and soil texture, respectively. However, it is worth noting how the second most important group is always relief and the rank of the groups of variables is consistent for the three maps. In addition, the removal of the climate covariates (T_{min} and T_{max}) leads to a decrease in the OOB error, for all the maps, of around 0.5%. This is probably due to an overfitting of the model, when temperatures are included, which can be related to correlation with other covariates (e.g. elevation).

The relatively high importance of relief (terrain) attributes was not a surprise. Many authors conducting DSM studies using different algorithms defined them, or demonstrated them to be useful predictors (e.g. Lemercier et al., 2012; Barthold et al., 2013; Taghizadeh-Mehrjardi et al., 2015). Additionally, the impact of their scale and resolution has been extensively studied (Behrens et al., 2010). The high impact of the soil geochemical properties is more surprising. Lawley and Smith (2008) discussed how geochemical surveys can help to describe some soil and weathered-zone characteristics and so be used to improve geological maps, which was

their final aim. The results of our study confirm that differences in soil characteristics can be spotted and highlighted by these variables. Láng et al. (2016) used similar variables (organic carbon content, pH, electric conductivity, sand, silt, clay and gravel content, bulk density, and cation exchange capacity) to derive major soil types (WRG soil groups) in Africa. These variables allowed a RF prediction with 68% classification success, thus confirming again the reliability of this type of variables for soil type prediction. However, their study does not use any other environmental predictor and therefore it is not possible to define a relative importance for soil properties predictors. Taghizadeh-Mehrjardi et al. (2015) used variables such as clay index, carbonate index, gypsum index and salinity ratio to predict USDA-family soil groups in Iran. These variables were derived from Landsat 8 products along with other indices. They conclude that Landsat spectral data and terrain parameters are the most useful auxiliary variables, although no specific reference is made to the soil property indices. Considering our results and those of these recent studies, we suggest further investigation of the role of soil properties (geochemical in particular) in the prediction of soil groups, in comparison with other categories of variables, and across different environments.

Table 4. Average OOB error for forests (350 trees, 50% data) built removing groups of variables according to the *scorpan* formula. Values are presented as differences from the model derived with the complete set of variables. Between parenthesis the rank of each group.

Based on the recursive elimination analysis, we removed Curvature, Aspect, and Slope for the soil groups map. The resulting 10-variable classification model had an average OOB error of 8.7%. Thus, the three covariates with the lowest MDA - and three out of the five covariates that gave a decrease of the OOB average error when singularly removed - were excluded (Table 3). For soil depth and soil texture recursive elimination led to a classification model

with 10 covariates each. The removed variables were the same for both maps: slope, curvature, and minimum temperature. Using only the remaining predictors the OOB error was reduced to 10.7% and 9.0% for soil depth and soil texture, respectively. These results demonstrate that a quick analysis of the model errors derived from single variable removal can provide more insight for the construction of the most parsimonious model than a simple analysis of MDA. In fact, a removal of variables based only on the MDA values of covariates could lead to keeping uninformative predictors in the model (due to preference for correlated variables, as discussed before) or removing layers with useful information.

In the **third step** of the model optimization, we defined the best *mtry* parameter value. In Table 5, we present the OOB errors calculated for the different classification models built with varying *mtry* values, 50% of the target points, and 350 trees. Besides an *mtry* value equal to 1, which gives fairly poor results for all the maps, all other *mtry* values give very similar OOB error. Poor performances with low *mtry* values are likely to be expected due to reduced probabilities of a relevant variable to be selected at each split (Hastie et al., 2008). As discussed in Breiman (2001), the forest error rate decreases with decreasing correlation between trees and increasing strength of each individual tree. Both correlation and strength depend on the *mtry* parameter and increase with it. The optimum range of the *mtry* parameter can be quite wide (Breiman 2001) as is also demonstrated by the results in Table 6 and previous literature (e.g., Grimm et al., 2008; Barthold et al., 2013; Heung et al., 2014). For the prediction of all maps, an *mtry* value of 5 was selected, being the one giving the lowest OOB error and ensuring a good balance between correlation among trees and strength of the trees.

Table 5. Average OOB errors calculated using different *mtry* values. The forest is made up of 350 trees.

In the **fourth step** the effect of tree pruning, through the setting of the *nodesize* parameter was investigated. The results are presented in Table 6. An increase in *nodesize* led to an increase in the OOB error for all the three maps, while it almost did not affect map uncertainty, expressed in terms of CI, and validation HR calculated using independent profiles. Hastie et al. (2009), in a general introduction of the RF algorithm, discussed that fully grown trees (*nodesize* equal to 1) can lead to model unnecessary variance (i.e. to overfitting), although in their experience no relevant performance increase is obtained from *nodesize* tuning. They showed an example in which the tuning of the *nodesize* parameter slightly reduced the model error, having used RF in its regression form. Our results confirm those of Hastie et al. (2009) and add an example for RF used as a classifier. For our final RF prediction, we kept *nodesize* to the default value (1).

Table 6. Average OOB (for Random Forest, RF) or MAP (for Multinomial Logistic Regression, MLR, from 50% unused points in 1:25,000 soil maps) errors, validation hit rate (HR) from profiles (not applicable to soil groups for lack of profile data) and confusion index (CI) calculated using different nodesize values (n). The forest is made up of 350 trees and the mtry parameter is fixed to 5.

The OOB errors of the optimized models are 8.6%, 10.5% and 8.8% for soil groups, soil depth, and soil texture respectively. Compared with other studies, these can be considered very low values: Stum et al. (2010) obtained an OOB error of 55.2% in predicting soil series; Barthold et al. (2013) reported an OOB error of 51.6%; Pahlavan Rad et al. (2014) calculated OOB errors of 48.5%, 51.5% and 56.5% for great group, subgroup and series levels, respectively; Taghizadeh-Mehrjardi et al. (2015) obtained a OOB error of 78% for family soil groups. OOB error values closer to the ones obtained by our study were obtained in other research fields. As an example Peters et al. (2008) achieved a minimum OOB error of 19% while modelling wetland vegetation distribution.

3.3 Comparison with Multinomial Logistic Regression

Consistent with many other studies (e.g., Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015; Heung et al., 2016), RF provided a better predictive model than MLR for all three maps, based on average scores. The validation error calculated for MLR using the 50% of the 1:25,000 map points are much larger than the comparable OOB errors obtained from the final RF models, as presented in Table 6. Coversely, the HR from independent validation of soil depth and soil texture from profiles is comparable with RF, which performs a little better for soil depth and a little worse for soil texture (Table 6).

A spatial comparison of CI results is presented in Fig. 6. The average CI values are usually lower for RF than MLR, with the exception of soil groups where MLR performs comparably. However, average values in this case do not explain all, especially for soil depth and soil texture. The CIs calculated from the RF output for these two maps show generally low CIs in the training areas (< 0.4) and higher values (> 0.4) in the rest of the study area. Zooming in, we can see large areas with low CIs bordered by narrow bands with higher CI values, perfectly representing diffuse boundaries among classes. Since this level of detail is not kept in all the prediction areas, it can be interpreted as an overfitting of the RF model, although no benefit appeared from tree pruning. CI maps derived from the RF built with *nodesize* equal to 20 (not presented here) show exactly the same pattern. Conversely, The CIs calculated from MLR are more homogeneous throughout the study area. Additional visual comparison between RF and MLR shows, in any case, similar trends of CI over the areas not included in training maps, with rather low values over the Troodos Mountains, in the centre of the island, and higher value along the borders of the study area. Considering these results, and above all the large difference in the validation from independent map values, we decided to present as final maps those predicted using RF.

Fig. 6. Confusion Index (CI) maps calculated for soil groups, soil depth and soil texture prediction from both Random Forest (left) and Multinomial Logistic Regression (right).

3.4 Final maps, map extrapolation, map reliability, and evaluation

Fig. 7, shows the final map of soil groups, whereas Fig. 8 shows the soil depth and the soil texture maps. Three main areas can be recognized: the very shallow (< 10 cm), lithic, and poorly developed soils of the Troodos Mountains; the young, (weakly) developed, shallow to moderately deep (between 10 cm and 100 cm), calcaric soils of the southern flank of the Troodos Mountains; and the well-developed, deep to very deep (> 100 cm), fertile soils of the western Mesaoria Plain and coastal plains in the southeast. Among the latter, it is worth mentioning the characteristic red soils here classified as Chromic Luvisols (with additional qualifiers according to the substrate). The good detail of the predicted maps is demonstrated along the southern coast, where differences between valley bottoms (Calcaric Cambisols) and slopes (Calcaric Regosols) can be recognized, and the most fertile soils of the country, located in the area of Pafos, can be spotted as Calcaric and Eutric Fluvisols. However, small errors, especially over the Troodos Mountains, can be found. Here, small lenses of Gypsiric and Calcaric Regosols have been predicted although, considering the igneous, mafic and ultramafic nature of the bedrock, these can be considered impossible soil types for the region.

Fig. 7. Digital soil map of Cyprus. The map is presented with WRB soil group names accompanied by one or two qualifiers as predicted using Random Forest.

Fig. 8. Digital soil depth and soil texture maps of Cyprus as derived from Random Forest.

The MESS and the reliability of the produced maps are presented in Fig. 9 and Fig. 10, respectively. Although slightly different set of predictors were used for soil groups, soil depth and soil texture, only the MESS related to soil groups is presented, since the differences between these maps were minimal. It is worth noting how both in terms of MESS and reliability the study area can be roughly split into two main regions. On one hand, the MESS and the reliability of the predicted soils in the main agricultural areas of the country (Mesaoria plain, southeastern coast, and mountain foothills) show medium to high values, confirming that the RF prediction is legitimate (not an extrapolation) and its quality is good. On the other hand, the predominant soils of the Troodos Mountains show a low reliability and are located over areas with a highly negative MESS. This low reliability is driven by the large size of the extrapolated area rather than a high OOB error. An important limiting factor for the prediction of soils over the Troodos region is the lack of a training area covering the highest elevations of the mountains. Limited training areas are only available along the foothills of the mountains, where the typical mountain soils start to form, but none of these areas cover regions characterized by steep valleys, as this terrain is found at higher elevations only. In addition, some of these valleys are rather peculiar, because they have been modified by human activities since the Bronze Age, especially by the construction of dry stone terraces to favour agricultural practices (Fall et al., 2012). Thus, while based on the topographic, geological, geomorphological, and climatic conditions, the predicted prevalent soil depth (0-10 cm) and soil texture (none) of the Troodos region (Fig. 8) can be considered reasonable, some of the natural variability and the human modifications of the terrain are not captured. The produced soil depth and soil texture maps generally fail to represent the peculiarities of the mountainous agricultural land, which could be also of interest for applied environmental studies (e.g. water needs for agriculture, soil erosion).

Fig. 9. Multivariate Environmental Similarity Surface (MESS) showing areas where model predictions are extrapolations in comparison to the training data set (values lower than zero).

Fig. 10. Reliability maps derived for the soil groups, soil depth, and soil texture maps of Cyprus predicted with Random Forest.

Table 7 summarizes the validation HRs based on independent profiles for each soil depth and soil texture class and for areas where the MESS values is larger or lower than zero (i.e.non extrapolation or extrapolation). The soil profiles cover all soil depth and soil texture classes, although with a non-uniform distribution. For soil depth, the moderate to deep (50-100 cm) class is highly under-represented in comparison with the other three, while the class shallow (10-50 cm) has the largest number of profiles. For soil texture, the two classes Medium and Fine are similarly represented and comprise 83% of all the profiles.

At the profile locations, the calculated overall validation HR for soil depth and soil texture are 55% and 49%, respectively. Heung et al. (2016) obtained a 54% HR predicting soil great groups with RF using the same method for training data sampling. Barthold et al. (2013) got much better results mapping WRB soil groups in the Mongolian grasslands, having a misclassification error of 29%. Comparably with the latter authors, Lang et al. (2016) obtained a 32% classification error for WRB soil groups in Africa. Considering that the error rate is a little higher than typical literature value, it is worth some discussion.

Table 7 shows that there is not much difference in the results of points predicted in the known or the extrapolated areas, but that there are large differences in the error between different classes, which have different reliabilities. In terms of soil depth, we can state that the model

is masking short distance heterogeneity, which is represented by the profiles. For example, the whole area of the Troodos Mountains is represented as a zero depth soil. This is due to the fact that 90% of the limited training areas presenting the main environmental characteristic of the Troodos is mapped as zero depth soil. Variability in this region exists, as testified by the soil-profile analyses of Robins (2004), who found a soil depth range between 0 and 0.96 m, with a median of 0.15 m and an average value of 0.18 m. Our model has not been trained to match Robins' soil depth data, as this would require larger training areas in the Troodos Region and a significantly finer resolved terrain model. Similarly, very-deep soils are more common than moderate-to-deep soils in the training areas and are better represented in the prediction, with less error. For texture, the smallest HR is associated with the class Coarse, which is the least represented in the training data set (6%). The texture class Medium, which is the most represented with 45% of the training points, is also the one with the best HR. The model therefore shows a tendency to predict the classes that occur more frequently in the training areas.

Table 7. Number of independent soil profiles (N. prof.) and validation hit rate (HR), calculated based on positive or negative Multivariate Environmental Surface Similarity (MESS), for soil depth classes and soil texture classes.

4 Conclusions

This study produced a digital soil map of Cyprus (including soil groups, soil depth classes and soil texture classes) with RF using training areas from ten 1:25,000 published soil maps (27% of the study area). The study also focused on the optimization of the RF model in terms of environmental variables used as predictors and of the model's adjustable parameters

(number of training points, number of trees in the forest, and number of variables sampled at each branch split). The study proved how the average OOB error, calculated by the RF model, increases at a higher rate with the reduction of the number of training points than with the reduction of the number of trees grown in the forest. This can be especially noticed for forests larger than 200-250 trees (when an asymptotic behaviour of the OOB error is reached). Considering this asymptotic behaviour, in this study forests of 350 trees were calculated.

For the selection of the optimal set of environmental variables to run the model (lowest error and lowest number of predictors), the recursive feature elimination based on OOB error was found to be generally superior to elimination based on the lowest Mean Decrease of Accuracy (MDA) value per predictor, as calculated by the RF algorithm. This is mainly due to MDA being influenced by variable correlation. In the optimum models derived for the prediction of soil groups, soil depth and soil texture, a prevalent role in keeping a low model error is played by the variables linked with the geochemistry of the soil. They appeared as the most important group of variables in our models. The importance of these variables should be further investigated in different geographic and pedological areas to confirm their quality as soil group predictors.

The calibration of the *mtry* parameter (number of variables sampled at each branch split) revealed a wide range of optimum values, distributed around half the number of environmental variables used to drive the prediction. The calibration of the *nodesize* parameter showed no relevant performance increase and was kept at its default value (1).

The produced soil groups, soil depth and soil texture maps showed very low OOB error: 8.6%, 10.5%, and 8.8%, respectively. However, when evaluated for soil depth and soil texture outside the training areas by means of independent soil profiles, the average error shows values equals to 45% and 51%, respectively. In particular, the maps show a medium to

high reliability over the major agricultural areas of the country (plain and mountain foothills) but a low reliability over the mountainous region. This low reliability is mainly due to a lack of training data over the mountains, whose soils are largely extrapolated from few training areas on the mountain foothills. Despite low OOB errors, peculiar characteristics of steep mountain environments, like medium to deep soils typical of terraced agricultural land, are missed by the model. Also, the model showed a tendency to predict more commonly the classes that occur more frequently in the training areas, therefore masking local variance in soil properties.

Maps derived with MLR had comparable prediction uncertainty and higher validation error than RF, indicating the better performance of the latter for DSM in topographical and pedological complex regions. If properly applied, RF was found to be a reliable instrument for DSM activities. Future research will focus on the development of a specific digital soil map for the Troodos Mountains of Cyprus, including both field and modelling studies.

5 Acknowledgements

This study is part of the AGWATER project (ΑΕΙΦΟΡΙΑ/ΓΕΩΡΓΟ/0311(ΒΙΕ)/06), co-financed by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation. All the simulations have been run at the Cy-Tera High Performance Computing (HPC) facility of the Cyprus Institute, whose staff, Thekla Loizou in particular, we would like to thank for the support to our study. Special thanks also to Hakan Djuma and Marinos Eliades for the digitization of the data of the soil profiles analysed in the 1970s and 1980s.

6 References

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., Amapu, I.Y., 2016. otal soil organic carbon and carbon sequestration potential in Nigeria. Geoderma 271, 202-215. doi:10.1016/j.geoderma.2016.02.021.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop Evapotranpiration: Guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper No 56. Food and Agriculture Organisation, Land and Water. Rome, Italy.

Babechuk, M.G., Widdowsonc, M., Kamber, B.S., 2014. Quantifying chemical weathering intensity and trace element release from two contrasting basalt profiles, Deccan Traps, India. Chemical Geology 363, 56–75. doi:10.1016/j.chemgeo.2013.10.027.

Barthold, F.K., Wiesmeier, M., Breuer, L., Frede, H.-G., Wu, J., Blank, F.B., 2013. Landuse and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. J. Arid Environ. 88, 194–205. doi:10.1016/j.jaridenv.2012.08.004.

Behrens, T., Zhu, A., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175-185. doi:10.1016/j.geoderma.2009.07.010.

Bird, D.N., Benabdallah, S., Gouda, N., Hummel, F., Koeberl, J., La Jeunesse, I., Meyer, S., Prettenthaler, F., Soddu, A., Woess-Gallasch, S., 2016. Modelling climate change impacts on and adaptation strategies for agriculture in Sardinia and Tunisia using AquaCrop and value-at-risk. Sci. Tot. Environ. 543, 1019-1027. doi:10.1016/j.scitotenv.2015.07.035.

Bolstad, P.V., Gessler, P., Lillesand, T.M., 1990. Positional uncertainty in manually digitized map data. Intl. Jr. Geog. Info. 4 399-412. doi: 10.1080/02693799008941555.

Boulesteix, A.L., Janitza, S., Kruppa, J., and König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining Knowl. Discov. 2, 493–507. doi:10.1002/widm.1072.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil series in three semi-arid landscapes. Geoderma 239–240, 68–63.doi: 10.1016/j.geoderma.2014.09.019.

Burrough, P.A., 1986. Principles of geographical information systems for land resources assessment. Clarendon Press, Oxford, UK.

Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115-135. doi:10.1016/S0016-7061(97)00018-9.

Büttner, G., Kosztra, B., 2007. CORINE Land Cover 2006 Technical guidelines. Technical Report No. 17/2007.EEA, 2007. Available at: http://www.eea.europa.eu/publications/technical_report_2007_17.

Camera, C., Bruggeman, A., Hadjinicolaou, P., Pashiardis, S., Lange, M.A., 2013. High resolution gridded datasets for meteorological variables: Cyprus, 1980-2010 and 2020-2050. AGWATER Scientific Report 5, The Cyprus Institute, Nicosia, 70 pp.

Chagas, C.D., de Carvalho, W., Bhering, S.B., Calderano, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139, 232-240. doi:10.1016/j.catena.2016.01.001.

Cohen, D.R., Rutherford, N.F., Morisseau, E., and Zissimos, A.M., 2011. Geochemical Atlas of Cyprus. Sydney: UNSW Press.

Cohen, D.R., Rutherford, N.F., Morisseau, E., Zissimos, A.M., 2012a. Geochemical patterns in the soils of Cyprus. Sci. Total Environ. 420, 250–262. doi:10.1016/j.scitotenv.2012.01.036.

Cohen, D.R., Rutherford, N.F., Morisseau, E., Christoforou, E., Zissimos, A.M., 2012b. Anthropogenic versus lithological influences on soil geochemical patterns in Cyprus. Geochem. Explor. Environ. Anal. 12, 349–360. doi:10.1144/geochem2011-111.

Collard, F., Kempen, B., Heuvelink, G.B.M., Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Reg. 1, 21–30.

Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77, 8–18.

Diaz-Uriate, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. Bmc Bioinformatics 7. doi:10.1186/1471-2105-7-3.

Djuma, H., Bruggeman, A., Camera, C., Zoumides, C., under review. Combining qualitative and quantitative methods for soil erosion assessments: an application in a sloping Mediterranean watershed, Cyprus. Land Degrad. Dev.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1: 330-342. doi:10.1111/j.2041-210X.2010.00036.x.

Fall, P.L., Falconer, S.E., Galletti, C.S., Shirmang, T., Ridder, E., Klinge J., 2012. Long-term agrarian landscapes in the Troodos foothills, Cyprus. J. Archaeol. Sci. 39, 2335–2347. doi: 10.1016/j.jas.2012.02.010.

FAO, 2006. Guidelines for soil description, Fourth edition. Food and Agriculture Organization of the United Nations, Rome.

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2008. Harmonized World Soil Database (version 1.0). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008.Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146, 102-113. doi:10.1016/j.geoderma.2008.05.008.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 180–190. doi:10.1016/j.geoderma.2007.11.004.

Grivas, G.C., 1969. Report on the soils of the Morphou Watershed. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 53 pp.

Grivas, G.C., and Georgiades, M., 1972. 1:25,000 Sheet 30 Lakatamia. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Grivas, G., 1988. Development of land resources in Cyprus. In: Zomenis, S.L. et al. (ed.) Proceedings - Workshop on conservation and development of natural resources in Cyprus - case studies - soils - groundwater - mineral resources. Ministry of Agriculture and Natural Resources, Cyprus and Federal Institute for Geosciences and Natural Resources, W. Germany, 13-18 Oct 1986, Nicosia. p.7-16.

Hastie T., Tibshirani R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction, Second Edition. Springer Series in Statistics, Springer, 763 pp.

Hijmans, R.J., 2013. Dismo, Species Distribution Modeling. R package version 1.0-15 [Available at https://cran.r-project.org/web/packages/dismo/dismo.pdf]

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214-215, 141-154. doi: 10.1016/j.geoderma.2013.09.016.

Hueng, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77. doi:10.1016/j.geoderma.2015.11.014.

Hutchinson, M. F., 1993. Development of a continent-wide DEM with applications to terrain and climate analysis. In Environmental Modeling with GIS, ed. M. F. Goodchild et al., 392–399. New York: Oxford University Press

IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.

Jafari, A., Finke, P.A., Van deWauw, J., Ayoubi, S., Khademi,H., 2012. Spatial prediction of USDA-great soil groups in the Arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63, 284–309.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

Kempen B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311-326. doi:10.1016/j.geoderma.2009.04.023.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., De Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76 (6), 2097–2115. doi:10.2136/sssaj2011.0424.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Koudounas, C., Makin, J., 1978. A study of representative soil profiles from the Limassol Paphos districts, published report, Department of Agriculture, Nicosia, Cyprus, 61 pp.

Koumis, C.I., 1970a.Report of the soils of the Limassol Watershed. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 229pp.

Koumis, C.I., 1970b.Report of the soils of the Larnaka Watershed. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 119pp.

Koumis, C.I., 1970c.Report of the soils of the Pissouri-Paramali Watershed. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 59pp.

Koumis, C.I., 1980a. 1:25,000 Sheet 53 Ypsonas, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Koumis, C.I., 1980b. 1:25,000 Sheet 54 Limassol, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Koumis, C.I., 1980c. 1:25,000 Sheet 58&59 Akrotiri, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Lagacherie, P., Legros, J.P., Burfough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65 283–301. doi:10.1016/0016-7061(94)00040-H

Lagacherie, P., McBratney, A.B., Voltz, M., 2007. Digital soil mapping: An introductory perspective. Elsevier, Amsterdam, 658 pp.

Lagacherie, P., Sneep, A.R., Gomez, C., Bacha, S., Coulouma, G., Hamrouni, M.H., Mekki, I., 2013. Combining vis–NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). Geoderma 209–210, 168–176. doi:10.1016/j.geoderma.2013.06.005

Lang, V., Fuchs, M., Szegi, T., Csorba, A., Micheli, E., 2016. Deriving World Reference Base Reference Soil Groups from the prospective Global Soil Map product - A case study on major soil types of Africa. Geoderma 263, 226-233. doi:10.1016/j.geoderma.2015.07.005.

Lawley, R., Smith, B., 2008. Digital soil mapping at a national scale: A knowledge and GIS based approach to improving parent material and property information. In: Hartemink, A.E., McBratney, A., Mendoça-Santos, M., Eds, Digital Soil Mapping with Limited Data, Springer, 173-182.

Lemercier, B., Lacoste, M., Loum, M., Walter, C., 2012. Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171-172 75–84. doi:10.1016/j.geoderma.2011.03.010.

Liaw, A., Wiener, M., 2014. randomForest: Breiman and Cutler's random forests for classification and regression. R package version 4.6-10. [Available at http://cran.r-project.org/web/packages/randomForest/randomForest.pdf]

Markides, L., 1973. Soils Memoirs of Polemi, Sheet no. 44 & 45, includes 1:25,000 map, Ministry of Agriculture and Natural Resources, Department of Agriculture, 138 p.

Markides, L., 1969. Report on the soils of the Polis-Tylliria Watershed. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 55 pp.

Markides, L., 1975. 1:25,000 Sheet 41 Ormidhia. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Markides, L., 1985a. 1:25,000 Sheet 50&56 Kiti. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Markides, L., 1985b. Report on the soils of the Krasochoria integrated rural development project area. Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus, 35 pp.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma117, 3–52. doi:10.1016/S0016-7061(03)00223-4.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. doi:10.1016/j.geoderma.2015.07.017.

Noller, J., 2009. The Geomorphology of Cyprus. Cyprus Geological Survey, Open File Report, 269 pp.

Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232–234, 97–106. doi:10.1016/j.geoderma.2014.04.036.

Panagos, P., Van Liedekerke, M., Montanarella, L., Jones, R.A, 2008. Soil organic carbon content indicators and web mapping applications. Environmental Modelling and Software 23, 1207–1209. doi:10.1016/j.envsoft.2008.02.010.

Peters, J., Verhoest, N.E.C., Samson, R., Boeckx, P., De Baets, B., 2008. Wetland vegetation distribution modelling for the identification of constraining environmental variables. Landsc. Ecol. 23, 1049-1065.

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209–210, 1–14. doi:10.1016/j.geoderma.2013.05.029.

Ren, L., Cohen, D.R., Rutherford, N.F., Zissimos, A.M., Morisseau, E.G., 2015. Reflections of the geological characteristics of Cyprus in soil rare earth element patterns. Applied Geochemistry 56, 80–93. doi:10.1016/j.apgeochem.2015.02.011.

Robins, C.R., 2004. Spatial Analysis of soil depth variability and pedogenesis along toposequences in the Troodos Mountains, Cyprus. Master of Science Thesis, Oregon State University, Corvallis, 119 pp.

Saxton, K.E., Rawls, W.J., 2006. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci. Soc. Am. J. 70, 1569–1578. doi:10.2136/sssaj2005.0117.

Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146, 138-146. doi:10.1016/j.geoderma.2008.05.010.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress Phys. Geogr. 27, 171–197. doi:10.1191/0309133303pp366ra.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model. 181, 1–15. doi:10.1016/j.ecolmodel.2004.06.036.

Shary, P.A., 2008. Models of topography. In: Zhou, Q., Lees, B., Tang G.A. (eds.), Advances in digital terrain analysis, Berlin, Springer, 29-57.

Soon, Y. K., Abboud, S., 1991. A comparison of some methods for soil organic carbon determination. Communications in Soil Science and Plant Analysis 22, 943–954. doi:10.1080/00103629109368465.

Soteriades, C., Georgiades, M., 1967. 1:25,000 Sheet 22 Kythrea, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Soteriades, C., Grivas, G., 1968. 1:25,000 Sheet 20 Kokkinotrimithia, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Soteriades, C., Koudounas, C, Markides, L., 1968. 1:25,000 Sheet 51 Paphos, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Soteriades, C., Koudounas C., 1968. Soils Memoirs of Pafos, Sheet no. 51, includes 1:25,000 map, Ministry of Agriculture and Natural Resources, Department of Agriculture, 96 p.

Soteriades, C., Markides, L., 1969. 1:25,000 Sheet 44&45 Polemi, Soil Section, Department of Agriculture, Ministry of Agriculture and Natural Resources, Cyprus.

Sreenivas, K., Dadhwal, V.K., Kumar, S., Harsha, G.S., Mitran, T., Sujatha, G., Suresh, G.J.R., Fyzee, M.A., Ravisankar, T., 2016. Digital mapping of soil organic and inorganic carbon status in India. Geoderma 269, 160-173. doi:10.1016/j.geoderma.2016.02.002.

Strobl, C., Boulesteix, A. L., Zeileis, A., Hothorn, T., 2008. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8, 25. doi:10.1186/1471-2105-8-25.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinformatics 9, 307. doi:10.1186/1471-2105-9-307.

Stum, A.K., Boettinger, J.L., White, M.A., Ramsey, R.D., 2010. Random forests applied as a soil spatial predictive model in Arid Utah. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, Dordrecht, pp. 179–190.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253, 67-77. doi: 10.1016/j.geoderma.2015.04.008.

Venables, W.N.,Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth ed. Springer, New York.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24.

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil–landscape modeling of soil organic carbon. Environ. Model. Softw. 57, 202–215. doi:10.1016/j.envsoft.2014.03.004.

1 Yeomans, J.C., Bremner, J.M., 1988. A rapid and precise method for routine determination of organic
2 carbon in soil. Communications in Soil Science and Plant Analysis 19, 1467–1476.
3 doi:10.1080/00103628809368027.

4 Zissimos, A.M., Christoforou, I.C., Morisseau, E., Cohen, D.R., Rutherford, N.F., 2014. Distribution
5 of water-soluble inorganic ions in the soils of Cyprus. Journal of Geochemical Exploration 146, 1–8.
6 doi:10.1016/j.gexplo.2014.07.004

7 Zomeni, Z., 2012. Quaternary marine terraces on Cyprus: constraints on uplift and pedogenesis, and
8 the geoarchaeology of Palaipafos, PhD Dissertation, Oregon State University.

9 Zomeni, Z., Bruggeman, A., 2013. Chapter III, Soil resources of Cyprus, in Soil Resources of
10 Mediterranean and Caucasus Countries, edited by Yigini, Y., Panagos, P., Montanarella, L.,
11 Luxembourg: Publications Office of the European Union, EUR25988EN Scientific and Technical
12 Research series, pp. 37-59.

13

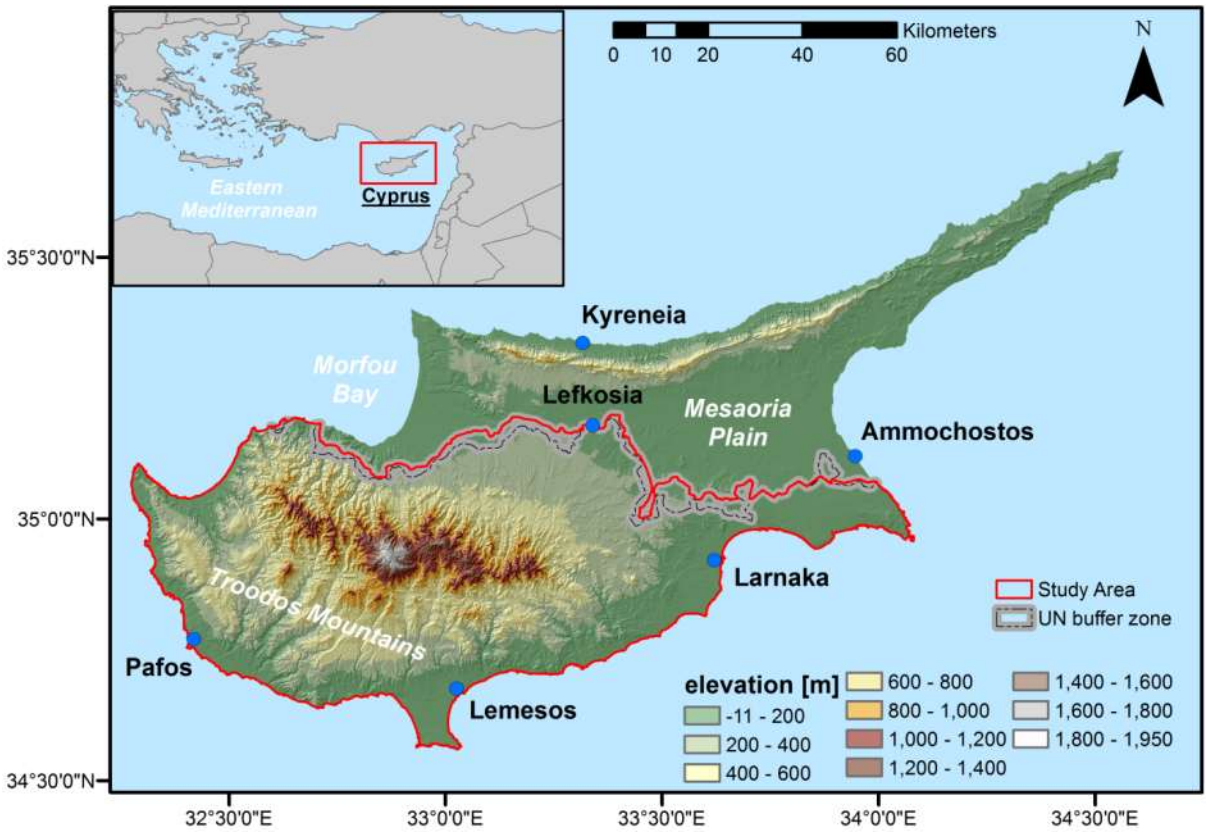

Fig. 1.The island of Cyprus with its main physical characteristics and the location of the study area.

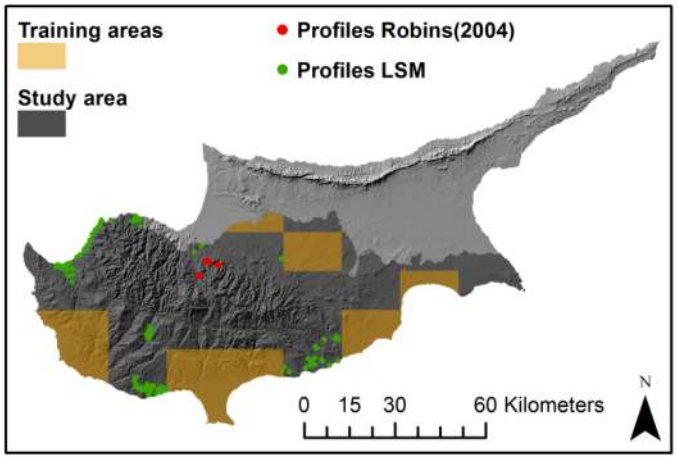

Fig. 2. Location of the study area, the existing 1:25,000 scale soil sheets used for training, the 126 soil profiles derived from Land Suitability Maps (LSM), and the 199 soil profiles from Robins (2004) used to evaluate the digital soil map of Cyprus.

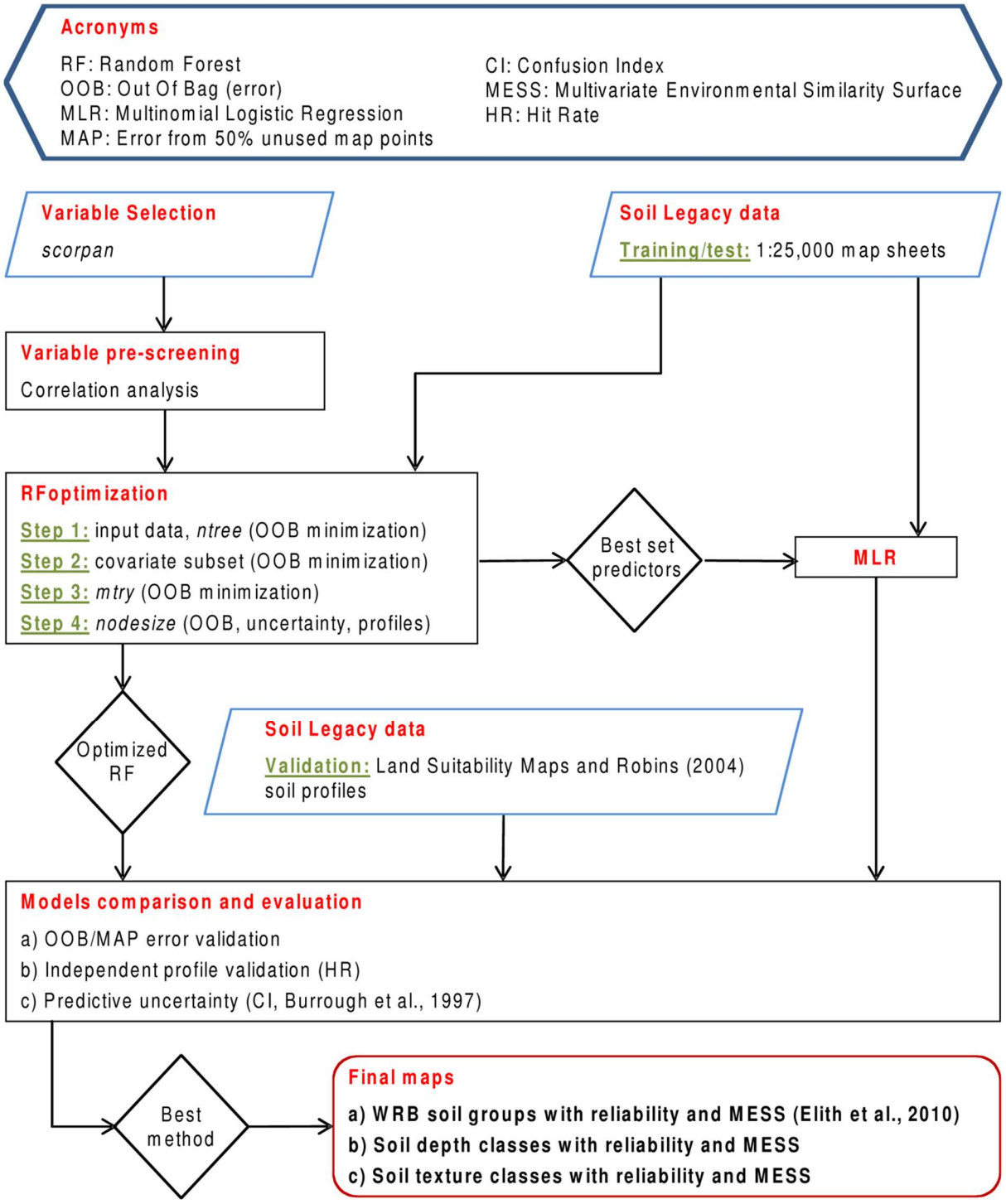

Fig. 3. Summary of the methods used in the study. OOB is the Out Of Bag error, MESS is the Multivariate Environmental Similarity Surface, and CI is the Confusion Index.

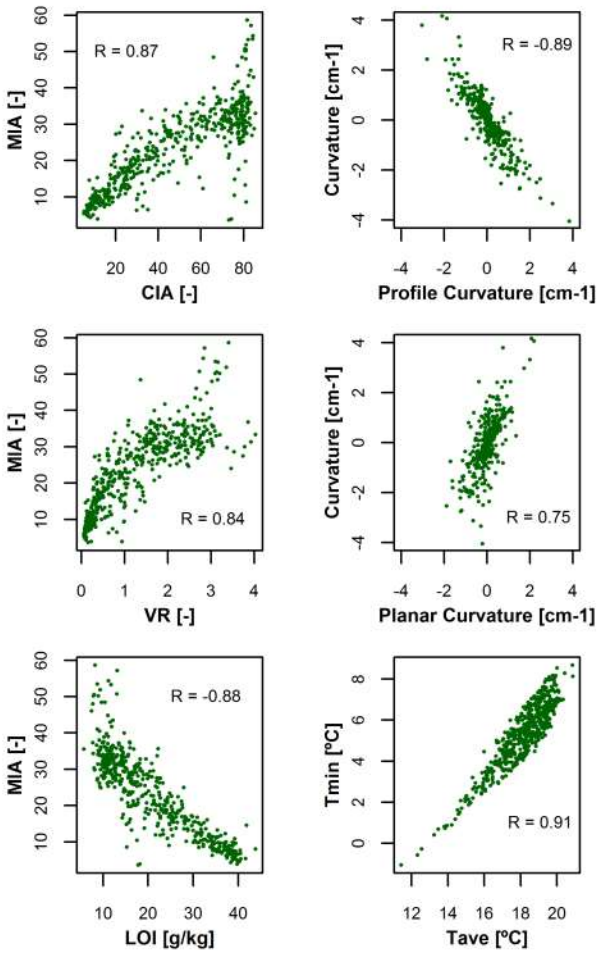

Fig. 4. Scatter plots, drawn with a 0.05% subset of the original input dataset, showing highly correlated covariates. The R value shown in each graph is the Pearson correlation coefficient calculated using the full input dataset. The covariates on the x-axis were removed from the data set used as input for the digital mapping techniques. CIA is the chemical index of alteration; MIA is the Mafic Index of Alteration; VR is the Vogt Ratio; and LOI is the Loss on Ignition.

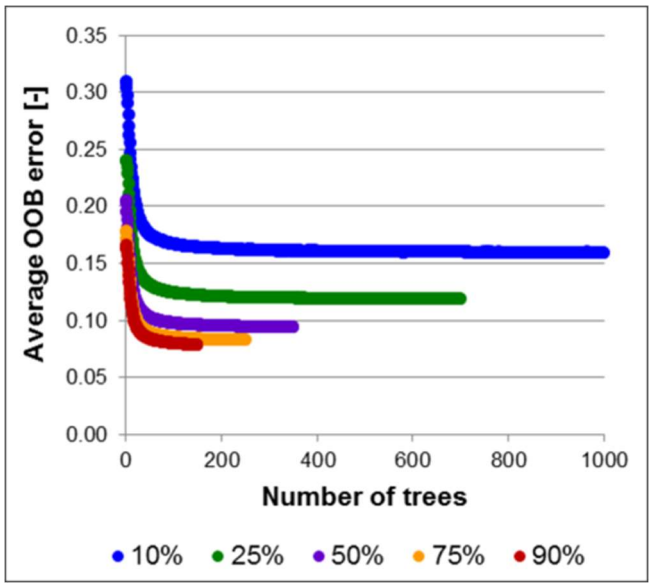

Fig. 5. Average OOB error for increasing number of trees in the Random Forest classification model and different number (% of the total) of training points as derived from the rasterized 1:25,000 scale soil maps.

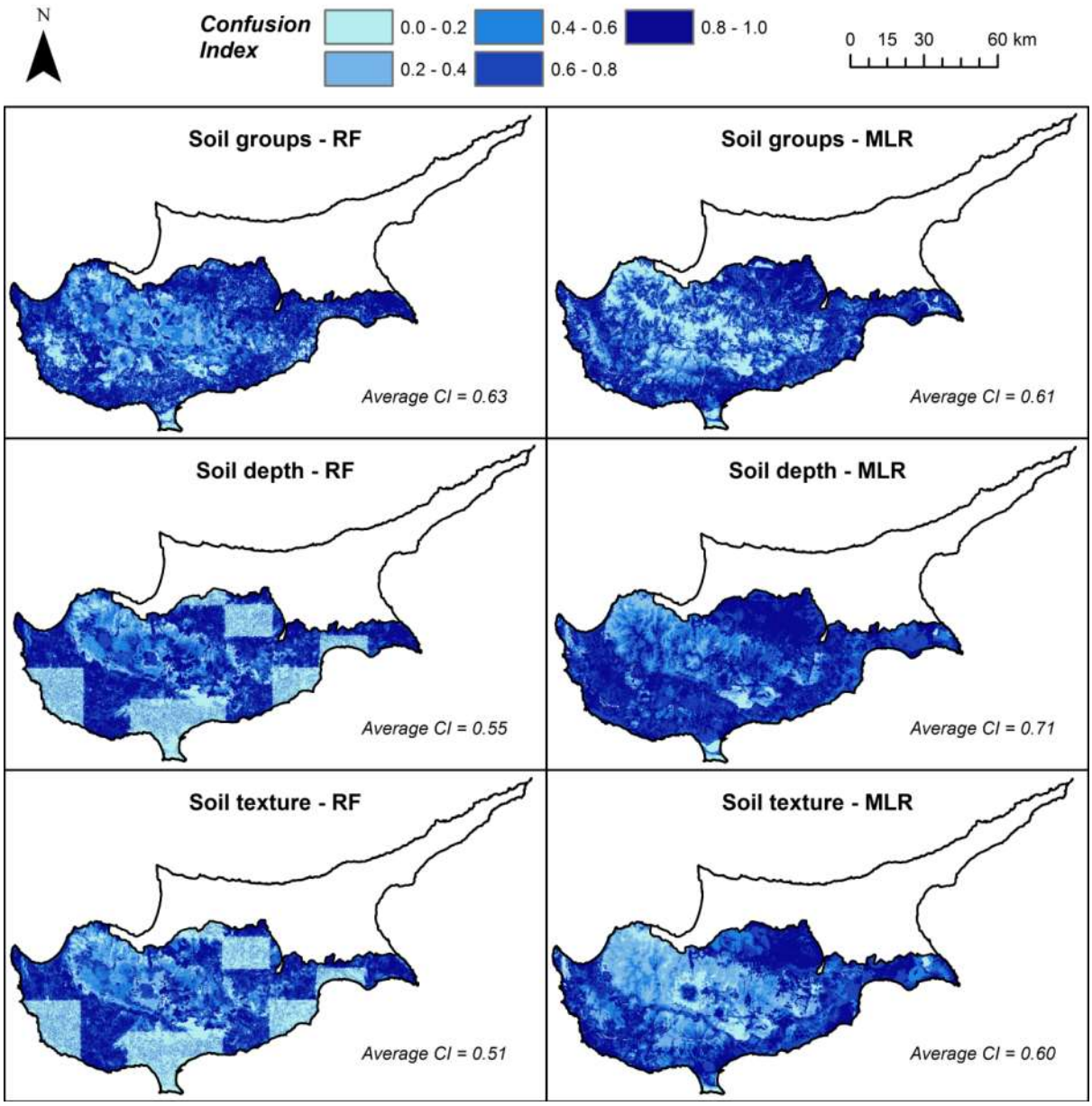

Fig. 6. Confusion Index (CI) maps calculated for soil groups, soil depth and soil texture prediction from both Random Forest (left) and Multinomial Logistic Regression (right).

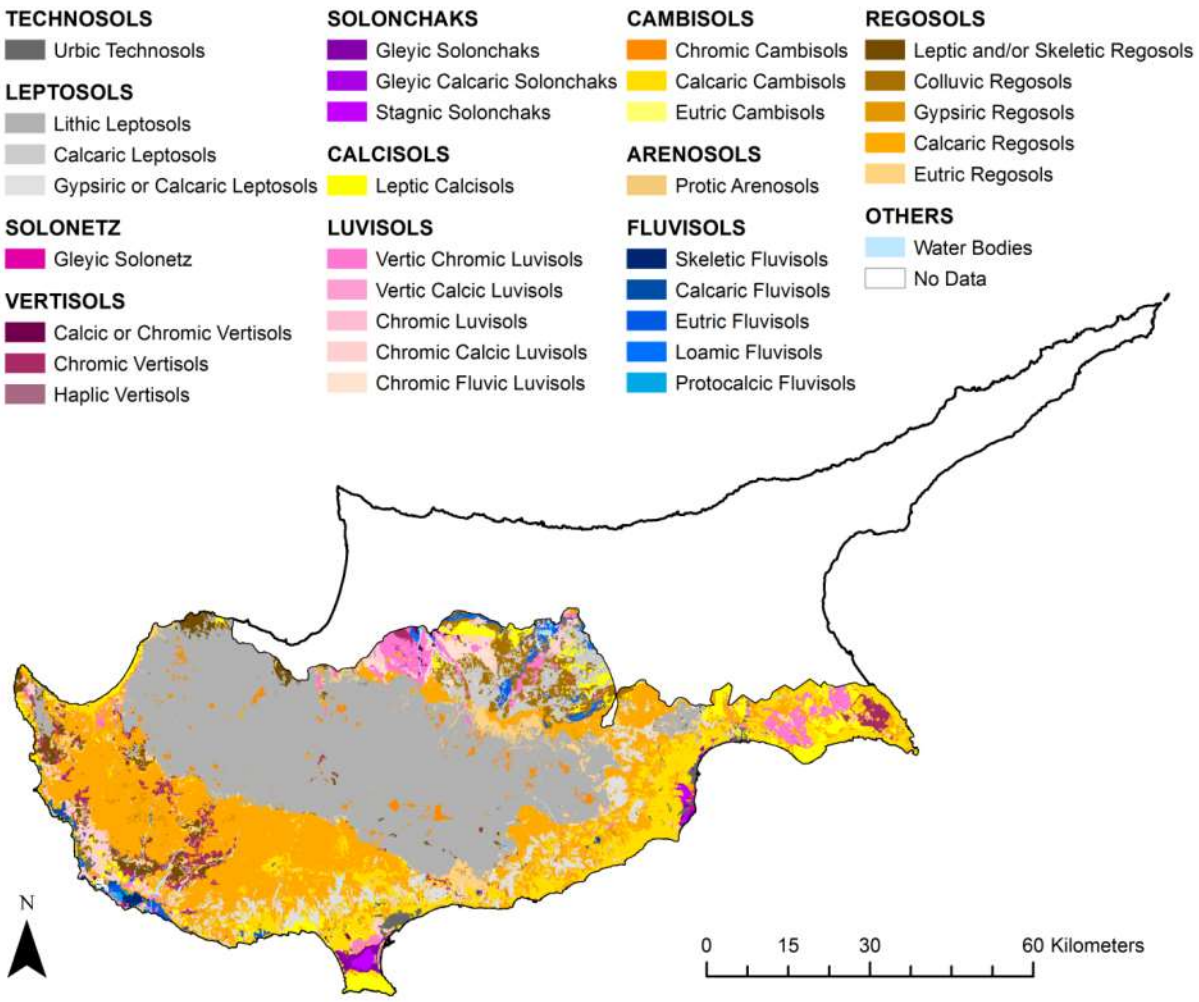

Fig. 7. Digital soil map of Cyprus. The map is presented with WRB soil group names accompanied by one or two qualifiers, as predicted using Random Forest.

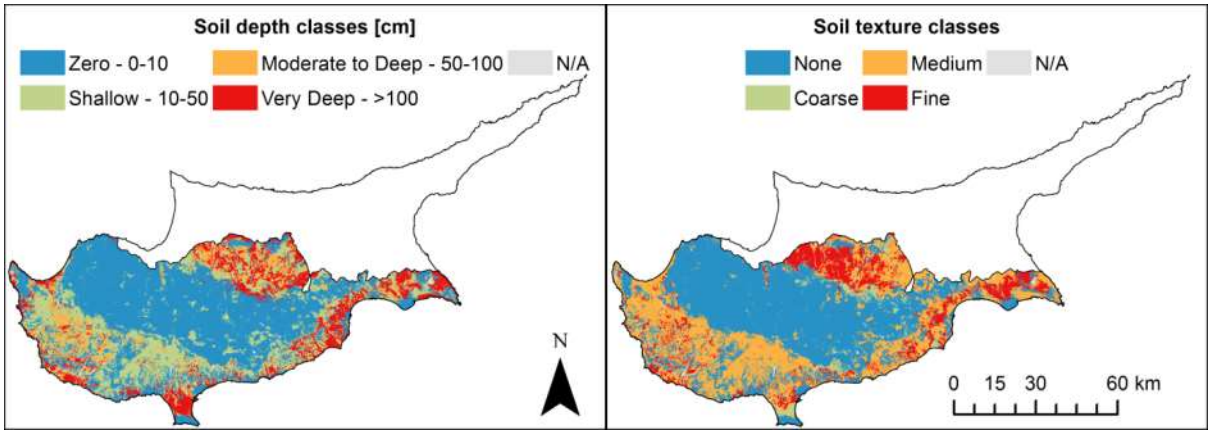

Fig. 8. Digital soil depth and soil texture maps of Cyprus, as derived using Random Forest.

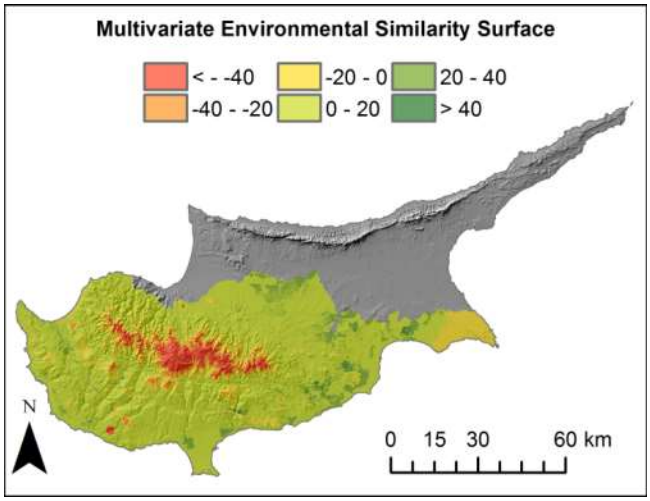

Fig. 9. Multivariate Environmental Similarity Surface (MESS) showing areas where model predictions are extrapolations in comparison to the training data set (values lower than zero).

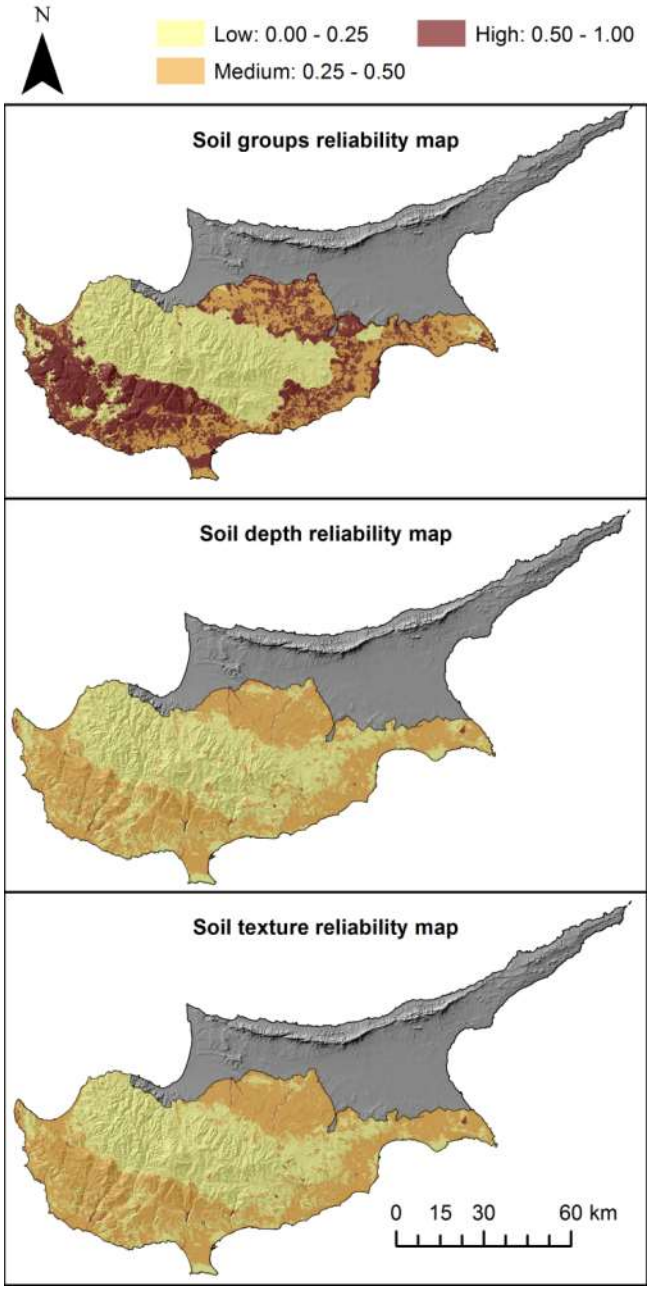

Fig. 10. Reliability maps derived for the soil groups, soil depth, and soil texture maps of Cyprus predicted with Random Forest.

Table 1. Soil depth classes identified from the 1:25,000 scale soil sheets and reclassified in four consistent classes..

Depth class (from soil maps)	Depth [cm] (from soil maps)	Depth [cm] (new class)
Zero	0-10	0-10 Zero
Very shallow	10-25	10-50
Very shallow to shallow	10-50	
Shallow	25-50	Shallow
Shallow to moderate	25-75	50-100
Moderately deep	50-75	
Deep	75-100	Moderate to Deep
Very deep	> 100	> 100 Very Deep

Table 2. Soil texture classes identified from the 1:25,000 scale soil sheets and reclassified in four consistent classes. The harmonization has been performed using the guidelines provided in FAO (2008).

Texture classes (from soil maps)	Texture 4 classes (reclassified FAO, 2008)	AWC[1] [%]
Rock	**None**	
Gravel	Stony and gravelly	
Sand		
Light to medium		
Coarse to medium		
Coarse	**Coarse**	5-15
Light	Sand and loamy sand, and sandy loam	
Sandy loam gravelly		
Sandy loam		
Clay loam	**Medium**	
Medium	Sandy clay loam, loam, clay loam,	10-20
Medium to fine	silty clay loam, silt loam, silt	
Moderately heavy		
Medium heavy		
Heavy	**Fine**	11-20
Fine	Clay, sandy clay, silty clay	
Clayee		

[1]AWC: Available water capacity, range of values taken from Saxton and Rawls (2006) and Allen et al. (1998).

Table 3. Average OOB error for forests (350 trees, 50% data) built removing single variables, and importance value (Mean Decrease of Accuracy - MDA) calculated by removing the variable from the complete model. OOB values are presented as differences from the model derived with the complete set of variables. Between parenthesis the rank of each variable.

	Soil groups		Soil depth		Soil texture	
Removed variable	**OOB error [%]**	**MDA**	**OOB error [%]**	**MDA**	**OOB error [%]**	**MDA**
None	9.4	N/A	12.0	N/A	9.9	N/A
Electrical conductivity	+0.5 (5)	0.22 (6)	+0.6 (5)	0.19 (5)	+0.5 (5)	0.18 (5)
pH	+0.7 (3)	0.20 (7)	+0.8 (3)	0.17 (7)	+0.7 (3)	0.17 (7)
Organic Carbon	+0.9 (2)	0.25 (5)	+1.1 (2)	0.22 (2)	+1.0 (2)	0.21 (3)
Mafic Index of Alteration	+0.7 (3)	0.31 (3)	+0.8 (3)	0.22 (2)	+0.7 (3)	0.24 (2)
Maximum Temperature	-0.3 (12)	0.33 (2)	-0.2 (10)	0.20 (4)	-0.1 (10)	0.21 (3)
Minimum Temperature	-0.2 (10)	0.18 (8)	-0.2 (10)	0.16 (10)	-0.1 (10)	0.15 (9)
Land Use	+0.3 (7)	0.18 (8)	+1.2 (7)	0.17 (7)	+0.2 (7)	0.17 (7)
Elevation (DEM)	+1.9 (1)	0.38 (1)	+2.0 (1)	0.29 (1)	+1.6 (1)	0.26 (1)
Aspect	+0.0 (8)	0.07 (12)	+0.1 (8)	0.07 (12)	+0.1 (8)	0.06 (12)
Curvature	-0.2 (10)	0.01 (13)	-0.2 (10)	0.02 (13)	-0.1 (10)	0.01 (13)
Slope	-0.6 (13)	0.17 (11)	-0.7 (13)	0.17 (7)	-0.4 (13)	0.12 (11)
Landscape Units	+0.4 (6)	0.18 (8)	+0.4 (6)	0.16 (10)	+0.3 (6)	0.14 (10)
Geology	+0.0 (8)	0.27 (4)	+0.1 (8)	0.19 (5)	+0.1 (8)	0.18 (5)

Table 4. Average OOB error for forests (350 trees, 50% data) built removing groups of variables according to the scorpan formula. Values are presented as differences from the model derived with the complete set of variables. Between parenthesis the rank of each group.

Removed group	Removed variables	Soil groups OOB error [%]	Soil depth OOB error [%]	Soil texture OOB error [%]
None	None	9.4	12.0	9.9
Soil chemistry	EC, pH, OrgC, MIA	+13.6 (1)	+13.0 (1)	+11.6 (1)
Climate	Tmin, Tmax	-0.5 (5)	-0.4 (5)	-0.4 (5)
Organisms	Land use, OrgC	+1.3 (3)	+1.5 (3)	+1.3 (3)
Relief	Elevation, aspect, curvature, slope, landscape units	+3.5 (2)	+3.6 (2)	+2.9 (2)
Parent material (Age)	Geology	+0.0 (4)	+0.1 (4)	+0.1 (4)

Table 5. Average OOB errors calculated using different *mtry* values. The forest is made up of 350 trees.

mtry	Soil groups OOB error [%]	Soil depth OOB error [%]	Soil texture OOB error [%]
1	22.5	26.2	23.0
3	8.7	10.7	9.0
5	8.6	10.5	8.8
7	8.7	10.6	8.9
9	8.8	10.7	9.1

Table 6. Average OOB (for Random Forest, RF) or MAP (for Multinomial Logistic Regression, MLR, from 50% unused points in 1:25,000 soil maps) errors, validation hit rate (HR) from profiles (not applicable to soil groups for lack of profile data) and confusion index (CI) calculated using different *nodesize* values (n). The forest is made up of 350 trees and the *mtry* parameter is fixed to 5.

	Soil groups			Soil depth			Soil texture		
Method	**OOB/MAP [%]**	**CI [-]**	**HR [%]**	**OOB/MAP [%]**	**CI [-]**	**HR [%]**	**OOB/MAP [%]**	**CI [-]**	**HR [%]**
RF n1	8.6	0.63	N/A	10.5	0.55	55	8.8	0.51	49
RF n4	8.7	0.64	N/A	10.7	0.56	55	9.0	0.51	49
RF n12	9.7	0.63	N/A	11.8	0.57	53	9.9	0.52	46
RF n20	10.6	0.63	N/A	12.8	0.57	54	10.6	0.53	53
MLR	48	0.61	N/A	52	0.71	58	44	0.60	46

Table 7. Number of independent soil profiles (N. prof.) and validation hit rate (HR), calculated based on positive or negative Multivariate Environmental Surface Similarity (MESS), for soil depth classes and soil texture classes.

Soil depth					Soil texture				
Class [cm]	**N. prof. MESS+**	**N. prof. MESS-**	**HR MESS+**	**HR MESS-**	**Class**	**N. prof. MESS+**	**N. prof. MESS-**	**HR MESS+**	**HR MESS-**
0-10	80	2	0.97	1.00	None	5	0	1.00	N/A
10-50	129	1	0.43	0.50	Coarse	11	5	0.09	0.20
50-100	17	1	0.03	0.00	Medium	47	8	0.74	0.50
> 100	76	19	0.43	0.42	Fine	42	8	0.31	0.37