Geoderma 263 (2016) 226–233

Contents lists available at ScienceDirect

Geoderma

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

Deriving World Reference Base Reference Soil Groups from the prospective Global Soil Map product — A case study on major soil types of Africa

Vince Láng, Márta Fuchs *, Tamás Szegi, Ádám Csorba, Erika Michéli

Szent Istvan University, Pater Karoly Street 1, Godollo 2100, Hungary

ARTICLE INFO

Article history:
Received 14 January 2015
Received in revised form 7 July 2015
Accepted 12 July 2015
Available online 5 August 2015

Keywords:
World Reference Base
Soil classification
Global Soil Map
Random Forest
Taxonomic distance

ABSTRACT

The main objectives of the Global Soil Map (GSM) project are the development of high resolution maps for defined soil properties and new digital soil mapping tools. Although many current land use, land capability models and many soil related applications use soil type information, a new global soil type map is not among the objectives of the GSM project. In this presented research legacy data was used to derive centroid values for World Reference Base Reference Soil Groups based on the Global Soil Map project defined soil properties in the specified depth intervals and test the applicability of numerical approaches for classifying soils, based on these specifications. The used variables were strictly following the project definitions such as: organic carbon content, pH, electric conductivity, sand, silt, clay and gravel content, bulk density and effective cation exchange capacity, all calculated for the GSM project defined depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, 100–200 cm). No environmental covariates were used in this study. A taxonomic distance and a Random Forest algorithm were tested to derive soil classes with the use of training and validation profiles for accuracy assessment. Results were studied in overall and for each Reference Soil Group. Reasons of misclassifications were identified, along with soil properties, which could increase the accuracy of the tested methods. Although this study did not derive a soil class map, promising results show a possibility to derive a World Reference Base based soil type map from the Global Soil Map product with numerical approaches as long as reliable archive dataset is available as a training population. The Random forest algorithm performed better with 68% classification success compared to the taxonomic distance based results with 47% success for the total number of validation profiles. Classification success differed by Reference Soil Group, those with strict, well defined and easily measureable and/or observable criteria on a certain property (like Arenosols) were classified better with both methods compared to other soil groups with broader definitions (like Cambisols). The study also unfolded completeness and accuracy problems in archive datasets, and the varying accuracy by soil type also revealed soil groups with narrow, precise definitions and groups which may need better definitions in a future global classification system for the better performance of numerical methods.

1. Introduction

The recognition of the importance of high quality and consistent soil information in land use planning food and water security or climate change and further global environmental issues resulted in the foundation of a global consortium called the GlobalSoilMap project. The original aim of the project was to produce a global high resolution spatial soil information for limited number of surface soil properties (Sanchez et al., 2009). As the project developed, wider range of properties, extended depth specifications, and state-of-the-art digital soil mapping approaches were defined by the GSM expert group (Arrouays et al., 2014). Deriving classes and producing a new global soil type map however have not become a goal of the GSM Project.

The only global coverage of soil maps is the 1:5,000,000 scale FAO-UNESCO "Soil Map of the World" compiled in 1974 (FAO-UNESCO, 1974). The last update and improvement was performed by Food and Agriculture Organization of the United Nations (FAO), the International Institute for Applied Systems Analysis (IIASA) and the EU JRC (Joint Research Center of the EU) using recent national soil information and applying digital soil mapping tool. The new comprehensive Harmonized World Soil Database (HWSD). The Harmonized World Soil Database is a 30 arc-second raster database with over 15,000 different soil mapping units that combines existing regional and national updates of soil information worldwide (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2009).

Recently Hengl et al. (2014) developed the automated soil mapping tool of "SoilGrids1km" aiming to provide consistent and coherent global soil information including spatial predictions for a selection of soil properties, and classification information as World Reference Base Reference Soil Groups (RSGs), and USDA Soil Taxonomy suborders. They

* Corresponding author.
E-mail address: fuchs.marta@mkk.szie.hu (M. Fuchs).

http://dx.doi.org/10.1016/j.geoderma.2015.07.005
0016-7061/

concluded that the major limitations of the model for global predictions are based on the highly variable quality, quantity and spatial detail of the input data.

In this study centroid-based taxonomic distance calculations approach and a Random Forest algorithm was performed to derive soil classes from GSM specified properties.

The idea of using calculated taxonomic distances to express the level of similarity and dissimilarity between different soil taxonomic units was first applied in the 1960s (Hole and Hironaka, 1960; Bidwell and Hole, 1964a,b; Sarkar et al., 1966) but only with local data and limited scope. The rediscovery of taxonomic distance calculations has been started in the 21st century by Minasny and McBratney (2007) who incorporated taxonomic distances into spatial prediction and digital mapping of soil classes. Minasny et al. (2009) derived taxonomic distances for the WRB Reference Soil Groups (RSGs) based on the presence and absence of key properties.

Fuchs et al. (2011) studied the taxonomic relationship of Hungarian soil types based on their dominant soil forming processes to provide numerical support to the improvement of the criteria of the taxonomic units. Soil taxonomic distance calculations were also applied to study the correlation of different national soil classification systems to the WRB (Láng et al., 2013; van Huyssteen et al., 2014).

Random Forest algorithm was designed and testified as an efficient tool in prediction by Breiman (2001). Application of the algorithm in soil science mainly focuses on digital mapping of certain continuous soil properties, using large number of covariates. Application to map or predict soil categorical data like soil types was lacking until recent publications. Heung et al. (2014) successfully applied the Random Forest algorithm to map categorical data, when predicting soil parent material classes, concluding that the training data has large influence on the prediction results. Nauman and Thompson (2014) investigated the algorithm on the United States Soil Survey Geographic (SSURGO) dataset to disaggregate soil survey maps. Reza Pahlavan Rad et al. (2014) used the tool to update soil survey maps combined with Latin hypercube sampling in Iran. Brungard et al. (2015) tested several machine learning techniques to predict low level soil classes of the United States Soil Taxonomy and with the use of high number of environmental covariates and concluded that Random Forest models were constantly the most accurate or was among the most accurate models to derive soil classes.

2. Materials

The World Reference Base for Soil Resources (WRB) (IUSS Working Group WRB, 2006) is an IUSS endorsed international soil correlation system since 1998. The WRB is based on diagnostic approach, hence the classification of soils is based on strict definitions and quantitative criteria defined in terms of diagnostic horizons, properties and materials. At the first level of classification 32 Reference Soil Groups (RSGs) are defined by a key based on the presence, sequence or exclusion of diagnostics, while at the second level qualifiers are added to the names of the RSGs to express the presence of specific soil characteristics.

The selection of the 11 dominant WRB Reference Soil Groups in Africa (Acrisols, Arenosols, Calcisols, Cambisols, Ferralsols, Leptosols, Luvisols, Lixisols, Nitisols, Plinthosols and Vertisols) applied in this study was based on the data published in the “Soil Atlas of Africa” (Jones et al., 2013).

The GlobalSoilMap.net product details are described in the Specifications Version 1 GlobalSoilMap.net products document (GlobalSoilMap.net, 2011). This study was conducted according to the Release 2.1 version of the document, which defines the vertical dimensions (Table 1.) and the soil properties (Table 2.) mapped in these dimensions. Vertical dimensions are defined in 6 depth intervals to a depth of 200 cm and 2 additional properties referred to soil depth described in Table 2.

Table 1
Vertical dimensions defined in the Global Soil Map specifications.

No.	Depth interval
1	0–5 cm
2	5–15 cm
3	15–30 cm
4	30–60 cm
5	60–100 cm
6	100–200 cm

The legacy data was derived from two sources:

The ISRIC-WISE v3.1 dataset is a compilation of soil profile data, collected from 149 countries worldwide. The profiles have been harmonized with respect to the FAO-UNESCO, (1974) and FAO-UNESCO, (1988) of FAO-UNESCO Soil Map of the World (Batjes, 2008, 2009).

The WRB Working Group (WRB database, IUSS WRB Working Group, 2010, unpublished, provided by Otto Spaargaren, 11.26.12.) dataset is a compilation of soil profiles from publications and was compiled to test the WRB. Profiles were quality checked and classified manually.

3. Methods

Prediction possibilities were studied with two approaches:

1. A taxonomic distance calculation based method, where centroid values of Reference Soil Groups, derived from legacy datasets were used as reference to classify unclassified soil profiles.
2. Random Forest algorithm based prediction was selected and tested due to good results in recent publications to derive categorical soil information.
3. Classification results were studied with validation profiles, where all studied soil properties and reliable soil classification information was available.

3.1. Available datasets and data preparation

Although many environmental covariates are available on global scale, derived from remotely sensed data and digital elevation models, the aim of this study was to test only the soil properties and depth increments defined in the GSM specifications, thus variables used in this study were defined by these specifications. Data availability in harmonized global datasets further limited the number of variables. Available water capacity, plant exploitable depth, and depth to bedrock were excluded from this study due to the lack of information or the lack of reliable data in the used datasets. The WISE dataset and the WRB Working Group dataset were combined and profiles of the 11 dominant Reference Soil Groups were selected for the geographical extent of Africa. Variations of laboratory methods were also taken into account. Profiles with properties measured with methods different from the ones accepted by the GSM project were either excluded from the further studies or analytical data was correlated according to the GSM specifications (GlobalSoilMap.net, 2011). This resulted in a total of 2136 soil profiles available for the calculations (Table 3.).

Brungard et al. (2015) concluded that overall prediction accuracy of soil classes with numerical approaches is influenced by the number of observations, soil classes and the frequency distribution of observations within classes. To address this bias 90% of the available pedons of the Cambisol Reference Soil Group were excluded from the further calculations. Pedons included in the calculations for the Cambisol RSG were selected on a random basis to reduce the uneven distribution of profiles available for each RSG. This resulted in 115 Cambisol profiles included in the further studies.

For most of the profiles within each Reference Soil Group missing values occurred for some properties or some depth intervals. For