Geoderma 318 (2018) 137–147

Contents lists available at ScienceDirect

Geoderma

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

Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation

Giulia Bondi[a,*], Rachel Creamer[b], Alessio Ferrari[c], Owen Fenton[a], David Wall[a]

[a] Teagasc Crops, Environment and Land-Use Research Centre, Wexford, Ireland
[b] Soil Biology and Biological Soil Quality, Wageningen University, Wageningen, The Netherlands
[c] Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (CNR-ISTI), Pisa, Italy

ARTICLE INFO

Handling editor: Morgan Cristine L.S.

Keywords:
Soil bulk density
Soil structure
Soil quality
Machine learning

ABSTRACT

Soil structure is a key factor that supports all soil functions. Extracting intact soil cores and horizon specific samples for determination of soil physical parameters (e.g. bulk density (B_d) or particle size distribution) is a common practice for assessing indicators of soil structure. However, these are often difficult to measure, since they require expensive and time consuming laboratory analyses. Our aim was to provide tools, through the use of machine learning techniques, to estimate the value of B_d based solely on soil visual assessment, observed by operators directly in the field. The first tool was a decision tree model, derived through a decision tree learning algorithm, which allows discrimination among three B_d ranges. The second tool was a linear equation model, derived through a linear regression algorithm, which predicts the numerical value of soil B_d. These tools were validated on a dataset of 471 soil horizons, belonging to 201 soil profile pits surveyed in Ireland. Overall, the decision tree model showed an accuracy of ~60%, while the linear equation model has a correlation coefficient of about 0.65 compared to the measured B_d values. For both models, the most relevant property affecting soil structural quality appears to be the humic characteristics of the soil, followed by soil porosity and pedogenic formation. The two tools are parsimonious and can be used by soil surveyors and analysts who need to have an approximate in-situ estimate of the structural quality for various soil functional applications.

1. Introduction

The importance of soil structure in relation to soil quality is well known (Mueller et al., 2009; Karlen, 2004; Kay et al., 2006). A commonly used soil physical measurement to characterize soil structural quality is soil bulk density (B_d) (Armindo and Wendroth, 2016; Dam et al., 2005; Håkansson and Lipiec, 2000; Logsdon and Karlen, 2004; Moncada et al., 2015), which is defined as the oven-dry mass per unit volume of soil (IUSS Working Group, 2006; Mueller et al., 2009). Measurement of soil B_d is useful as it describes both the packing structure of the soil and its permeability (Dexter, 1988), whereby drainage characteristics can be inferred (Reidy et al., 2016). B_d measurement is often used in agronomic studies as it indicates the presence of compacted layers resulting from machinery or animal traffic (Reidy et al., 2016; Saffih-Hdadi et al., 2009), which may affect crop production. It is commonly considered an efficient measurement of soil carbon and nutrient stocks (Ellert and Bettany, 1995; Reidy et al., 2016).

However, the process of measuring B_d is often time consuming and open to human bias in the field and requires accurate laboratory analyses using trained personnel. Furthermore, soil texture has an important influence on the assessment of B_d e.g. in soils with high clay or sand content, or very humic soils, it may be difficult to obtain a representative sample and large variability between replicate samples can represent a problem. Also, in some soils the presence of stones can make sampling almost unmanageable. For such reasons, or constraints of budget or laboratory facilities, B_d measurements are commonly missing from soil databases (Reidy et al., 2016).

The main methods employed for the prediction of B_d are pedotransfer functions (PTF) methods, based on measurable soil attributes, such as organic carbon (OC) and clay content (Kaur et al., 2002; Leonavičiutė, 2000; Reidy et al., 2016). However, many of these methods ignore horizonation and depth variances for soil B_d prediction (Reidy et al., 2016). Furthermore, the nature of these methods, based on chemical/physical or landscape parameters, do not capture the intrinsic nature of the soil structural properties.

Our experience with respect to soil descriptions and classification has shown that the visual observations collected in the field at horizon level are often very important for the evaluation of soil quality (Fenton

* Corresponding author.
E-mail addresses: Giulia.Bondi@teagasc.ie (G. Bondi), rachel.creamer@wur.nl (R. Creamer), alessio.ferrari@isti.cnr.it (A. Ferrari), Owen.Fenton@teagasc.ie (O. Fenton), David.Wall@teagasc.ie (D. Wall).

https://doi.org/10.1016/j.geoderma.2017.11.035
Received 31 May 2017; Received in revised form 15 November 2017; Accepted 23 November 2017

et al., 2015, 2017) and they become essential during the interpretation of the trend of some analytical parameters used as indicators of soil structure status e.g. B_d.

Soil structural quality has been assessed visually for millennia (Batey, 2000) e.g. soil survey manuals used in the field such as the Soils Survey Division Staff Manual (1993) or the WRB for soil resources (FAO, ISRIC and ISSS, 1998) include soil structure visual observations. However, soil scientists, for a long time, have presented repeatable procedures for the examination of soil structural form, stability and resilience (see latest review by Emmet-Booth et al., 2016 with examples from 1940 to present; Ball and Munkholm, 2015).

Taking this into account, in the present work we investigated whether, and to what extent those visual observations, called descriptors, can be used to predict soil B_d, which is considered one of the most efficient indicators in the assessment of soil structure quality (Moncada et al., 2015).

In order to achieve this objective, machine learning techniques were used. The potential of machine learning techniques have been rediscovered in the last few years through various applications in environmental sciences.

Worldwide, decision tree approaches have been used for different purposes: identifying sources of soil pollution (Xue et al., 2015); describing the extension of different forms of soil erosion in Mexico (Geissen et al., 2007); predicting chemical soil properties at national level in Australia (Henderson et al., 2005); classifying the surface soil freeze/thaw status in China (Jin et al., 2009) and even studying soil structure through the prediction of soil hydraulic properties (Pachepsky and Rawls, 2003). However, limited literature has been found on the use of these powerful tools for environmental science in Europe.

The decision tree model output applied in this paper is based on a series of rules generated by the software, which can be visualised as paths starting from the root of the decision tree and ending at one of the leaves (Bhargava et al., 2013; Xue et al., 2013; Xue et al., 2015). Each of those paths corresponds to one or more soil *descriptors*, which are related to an internal *node* (Henderson et al., 2005). The model is able to examine all possible descriptors and then to select the most decisive *splitting attribute* (Xue et al., 2013). This operation occurs several times until all the instances are correctly classified in a set of *rules*. Each descriptor included in the model corresponds to a more defined *level* of classification.

The linear regression model applied in this paper is a classical statistical technique used to predict numerical data. It is based on the modelling of the relationship between a scalar dependent variable and one or more explanatory variables.

With our work we want to:

(i) Provide an operational strategy to estimate a range of B_d values, based on the visual soil parameters by means of a decision tree approach. This model can be used as a field tool to predict a general class of B_d (Low, Medium and High). It is an instrument able to discriminate between macro classes and has to be considered as a descriptive tool for *qualitative* estimation.

(ii) Propose an algorithm that can predict a numerical estimate of B_d. This second model should discriminate better between smaller increments. This instrument has to be considered as a more refined tool for *quantitative* estimation.

2. Methods

2.1. Primary data source and descriptors

Two pedological surveys, where full soil profile descriptions and supporting laboratory analyses, were carried out in Ireland with the aim of defining a coherent and homogeneous way to study soil formation, functions and quality:

1. The Irish Soil Information System (Irish SIS) project was established in 2008. It aimed to conduct a programme of structured research into the national distribution of soil types and construct a soil map, at 1:250,000 scale, able to identify and describe the soils according to a harmonised national legend. Irish SIS included more than 225 sites distributed around Ireland (Creamer et al., 2014).
2. The Soil Quality and Research project (SQUARE) started in 2013. The aim was to establish a baseline of soil quality in Ireland. The SQUARE soil survey included 38 grassland sites distributed within the five major agro climatic regions of Ireland defined by Holden and Brereton (2004) and classified into two drainage classes on the basis of the Irish Soil classification System.

During both (1) and (2) profile pits approximately 1 m deep, were observed and described by different operators. For the present study data from 201 profiles (168 Irish SIS, 33 SQUARE) was extracted from the larger database to cover a wide variety of Irish soil types with a specific focus on mineral soils. This data represents 471 horizons (http://gis.teagasc.ie/soils/map.php).

Although different surveyors worked across the projects mentioned, a systematic procedure was applied to describe the nature of the soil profiles, which included each of the soil horizons. Training was given to field operators. Using knowledge of soil structure and quality, the operators followed a widely understood schema of observation (developed by FAO through the Guidelines for Soil Description in 2006) which was able to investigate and finally characterize soil structure through visual parameters (FAO, 2006; FAO, ISRIC, and ISSS, 1998). Herein we have selected eleven descriptors presented in Table 1 (justifications are provided in Table 1), which may be considered the most important for the qualitative judgment of soil structure. Each descriptor was described and recorded on the basis of a set of pre-defined categories, reported in Table 1 in the Supplementary material.

2.2. Soil analysis

The procedure to determine B_d of intact cores is a version of the ISO 11272:1998 – Soil Quality Part 5: Physical methods Sect. 5.6 – Determination of dry bulk density. The primary difference between the ISO and the applied methodology is that the ISO does not account for stone mass and volume in its core method, whereas the methodology applied in this study includes the following equation to calculate B_d (stone free):

$$B_d\ (\mathrm{g\ cm^{-3}}) = (\mathrm{Md} - \mathrm{Ms})/(\mathrm{V} - \mathrm{Vs}) \tag{1}$$

where; Md: oven dry soil material weight (g), Ms: oven dry stone weight (g), V: volume of soil core ($\mathrm{cm^{-3}}$), Vs: volume of stones (mL). Soil B_d values reported in this paper correspond to the mean of the three values obtained for each horizon sampled.

2.3. Model frameworks

Two models were built by means of the modelling tool WEKA (Waikato Environment for Knowledge Analysis). WEKA 3.8 is open source software for machine learning and data mining under the General public license developed at the University of Waikato in New Zealand (http://www.cs.waikato.ac.nz/ml/weka, Bhargava et al., 2013). This software includes different implementations of several machine learning algorithms. In our context, we used two specific algorithms that are made available by the tool, namely:

- The j48 algorithm, which corresponds to the WEKA's implementation of the C4.5 decision tree learner (Quinlan, 1993; Xue et al., 2015) which was used to build Model (1);
- A linear regression algorithm, used to build the Model (2). The M5 Method was used as attribute selection method for the linear model presented.