Ecological Indicators 18 (2012) 1–10

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

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

Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark

Mogens H. Greve[a], Rania Bou Kheir[a,*], Mette B. Greve[a], Peder K. Bøcher[b]

[a] Department of Agroecology, Faculty of Science and Technology, Aarhus University, Blichers Allé 20, P.O. Box 50, DK-8830 Tjele, Denmark
[b] Ecoinformatics and Biodiversity Group, Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark

A R T I C L E I N F O

Article history:
Received 3 April 2011
Received in revised form 17 October 2011
Accepted 19 October 2011

Keywords:
Soil quantitative relationships
Environmental correlation
GIS regression trees
Soil texture fractions
Terrain analysis

A B S T R A C T

Soil texture is an important soil characteristic that drives crop production and field management, and is the basis for environmental monitoring (including soil quality and sustainability, hydrological and ecological processes, and climate change simulations). The combination of coarse sand, fine sand, silt, and clay in soil determines its textural classification. This study used Geographic Information Systems (GIS) and regression-tree modeling to precisely quantify the relationships between the soil texture fractions and different environmental parameters on a national scale, and to detect the most important parameters that can be used as weighted input data in soil environmental prediction models. Seven primary terrain parameters (elevation, slope gradient, slope aspect, plan curvature, profile curvature, flow direction, flow accumulation) and one compound topographic index (CTI) were generated from a Digital Elevation Model (DEM) acquired using airborne LIDAR (Light Detection and Ranging) systems. They were used along with digital data collected from other sources (existing maps and available pluviometric stations), i.e. parent materials, landscape types, geographic regions, yearly precipitation, seasonal precipitation to statistically explain soil texture fractions field/laboratory measurements (45,224 sampling sites) in the area of interest (Denmark). The developed strongest relationships were associated with clay and silt, variance being equal to 60%, followed by coarse sand (54.5%) and fine sand (52%) as the weakest relationship. This study also showed that parent materials (with a relative importance varying between 47% and 100%), geographic regions (31–100%) and landscape types (68–100%) considerably influenced all soil texture fractions, which is not the case for climate and DEM parameters. Yearly and seasonal precipitation had a significant impact on clay and silt; elevation had higher influence on coarse sand (13%), fine sand (12%) and clay (10%) where; slope gradient influenced silt (11.5%); slope aspect (14%) and CTI (9%) influenced fine sand; and profile/plan curvatures and flow direction/accumulation did not interfere in the building of the soil texture regression trees and associated relationships. The latter can be extrapolated to other areas sharing similar geo-environmental conditions.

1. Introduction

Soil plays a crucial role in the environment and also for human life. A knowledge and understanding of soil and how it is distributed across the landscape is crucial for the effective use and environmental management of this vital resource. Soil texture is one of the most important soil properties, affecting many of the physical and chemical characteristics and behavior of the soil, such as the soil water and nutrient holding capacities, hydraulic conductivity, friability, and resistance to cultivation. Spatial distribution and variability of the various soil texture fractions (coarse sand %, fine sand %, silt % and clay %) is increasingly being required for input into ecological, hydrologic, climatic, and other environmental models, particularly due to ever-rising environmental concerns relating, for example, to the prognosis for agricultural yields and carbon stocks at a global level. In addition, there is a pressing need for precise and accurate predictive quantifications of the relationships between each of the soil texture fractions and key environmental parameters (climate, topography, parent material, etc.) to assist in site-specific, economic and sustainable management of the environment. Such relationships form the basis of environmental prediction modeling and digital soil mapping (DSM) programs, which are widely considered to be the cornerstone of future soil surveys (Grunwald, 2006; Hartemink, 2006; Kværnø et al., 2007; Lagacherie and McBratney, 2007; Gray et al., 2009; He et al., 2010). Traditionally, soil–environment relationships have

* Corresponding author.
E-mail address: Rania.BouKheir@agrsci.dk (R.B. Kheir).

1470-160X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecolind.2011.10.006

been of a qualitative nature rather than a quantitative nature (Gobin et al., 2001; Heuvelink, 2005), dividing soil texture into classes rather than numerically defining specific soil texture fractions. In most studies and soil surveys, soil texture has been applied to particular regions (ranging from site/hillslope to catchment) (Lamsal and Mishra, 2010), and their enunciation is often unclear. For example, it is widely reported that topography influences soil texture (McKenzie et al., 2000; Dobos and Hengl, 2009), but there is little detail on what the precise predictive influence is, particularly in quantitative terms. Moreover, and referring to the fundamental soil equation and the associated state factor model (Clorpt model) of soil formation (Jenny, 1941), the ability to derive quantitative soil texture–environment relationships on a broad scale is inhibited by the complexity of mathematical correlations that are difficult to understand and analyze. The latter can be partially solved through the use of individual pedofunctions (McBratney et al., 2003), but these functions are applied mainly to relatively small locations, and their transferability to wider contexts is questionable.

The advent of digital terrain analysis and allied raster-based Geographic Information Systems (GIS) technologies over the past couple of decades has created an opportunity for the development of more scientifically and statistically based methods and models relating soil properties to digital environmental data [multivariate correspondence analysis (MCA), generalized linear models (GLMs), generalized additive models (GAMs), artificial neural networks (ANNs), etc.] that overcome several of the limitations of conventional soil surveys (McBratney et al., 2003; Scull et al., 2003; Zhai et al., 2006; Kværnø et al., 2007; Hartemink et al., 2008; Zhao et al., 2009; He et al., 2010). However, these methods are heterogeneous in terms of environmental input data, predictive power, ease of use, sensitivity to parsimony, ease of interpretability, handling of mixed data, handling of non-linear relationships, etc. For instance, the multivariate correspondence analysis (MCA) detects the percentages of the total inertias for the uncorrelated principal axes that are linear combinations of the environmental parameters (González et al., 2007). A major limitation of this analysis is the unique combination of axes together defining a unique condition of a given subarea. Generalized linear models (e.g., linear regression, nonlinear and logistic regression, probabilistic regression) are not flexible enough to allow robust integration with a variety of potential data sources [e.g., remote sensing data, topographic surfaces of Digital Elevation Models' (DEMs)] for the investigated soil texture–environment relationships (Fox and Melta, 2005; Selige et al., 2006; Lesch and Corwin, 2008). Linear regression models are limited by their assumed linear relationship between soil and environmental parameters, their assumptions of normally distributed data and their high data requirements. Similar to linear regression models, nonlinear and logistic regression-based models cannot easily look for interactions between soil texture and influencing environmental parameters. An important problem with these models is that we cannot evaluate the contribution to the model of each environmental parameter. Another problem with the Bayesian methods (D'Or and Bogaert, 2003; Grunwald, 2009) is that the identification and weighting of the environmental parameters influencing the various soil texture fractions remain highly subjective, and dependent on the expertise of different soil surveyors, with their overall accuracy and reliability remaining largely unevaluated. Artificial neural networks (ANNs) also have a number of drawbacks for predicting soil texture–environment relationships. Neural nets are unsatisfactory because of the difficulty of interpretation and requirement for specialized skills. They are also criticized for their inability to identify the relative importance of potential predictor environmental parameters (Berk, 2003; Janik et al., 2009). Likewise, generalized additive models (GAMs) are difficult to interpret and might have problems detecting the parsimony (Berk, 2003). Unsupervised machine learning methods like cluster analysis, factor analysis (principal component analysis) treat all environmental parameters influencing soil texture equally without predicting the value of a given parameter (Shukla et al., 2006; Concepción Ramos et al., 2007). These methods tend to look for patterns, groupings or other ways of characterizing the data that may lead to the understanding of the way the data interrelate.

Regression trees are often compared to the previously mentioned numerically oriented techniques and methods (Wilson and Gallant, 2000; Vega et al., 2010), but tree-based models are easier to interpret and discuss when a mix of continuous (e.g., elevation) and nominal (e.g., parent material) environmental parameters are used as predictors. They are scalable to large problems (Wilson and Gallant, 2000), invariant to monotone re-expressions (transformations) of predictor parameters (Scull et al., 2003), and can optimize non-additive and non-linear relationships between inputs (e.g., environmental parameters) and outputs (soil texture fractions) (McKenzie et al., 2000; Grunwald, 2009). In addition, they are non-parametric/probabilistic, and do not require the specification of the form of a function to be fitted to the data, as is necessary for other competing procedures (e.g., non-linear regression) (Breiman, 2001; Lawrence et al., 2004; Henderson et al., 2005). In contrast to ANNs, once regression-tree models have been built, they can be converted to statements that are implemented easily in most computer languages (Razi and Athappilly, 2005). Regression trees also have excellent predictive capabilities (Breiman, 2001; Lawrence et al., 2004; Henderson et al., 2005; Razi and Athappilly, 2005), but they have been criticized for overfitting and for poor performance on small datasets (McKenzie et al., 2000).

In this context, our study aimed to implement systematic regression tree-based models and evaluate their ability to develop broad quantitative relationships between the various soil texture fractions and the environmental soil-forming parameters across large areas (A horizons – depth ranging between 0 and 30 cm) that can be used in quantitative soil mapping and environmental modeling exercises on a national scale. Denmark is used as the case study since national agricultural and environmental authorities have a crucial need for information about the soil textural composition (coarse sand, fine sand, silt and clay expressed in percent) and related digital environmental data. This country has also a massive national database consisting of 45,224 soil samples, which in conjunction with auxiliary environmental parameters can be used to examine the soil texture–environment relationships. The resulting predictive soil texture–environment relationships may provide valuable insights into soil formation/distribution and soil modeling/mapping, and can serve as useful tools for land use management and environmental decision-making.

2. Study area description

The study area (Fig. 1) covers the soils of Denmark, i.e. around 43,000 km^2. The central and eastern part of the country consists of a Last Glacial (Weichselian) morainic landscape (18,330 km^2 or 43% of the total area of Denmark) with mainly loamy soils on calcareous tills. The western part of the country, which was not covered by ice during the Last Glacial, consists of low-relief, glaciofluvial, sandy sediments (5020 km^2 or 12%), emanating from melting glaciers, surrounding slightly protruding 'islands' of the older and strongly eroded landscapes of earlier (Saalian) glacial eras (4731 km^2 or 11%). The northern part of the country consists of a Weichselian glacial core bordered by uplifted marine sediments from early and mid-Holocene. Sand dunes are found in the coastal areas, particularly on the west coast, and as patchy inland deposits. The southwestern coastal region is a salt marsh, dominated by recent fine-textured tidal sediments. Throughout the country, poorly drained basins have been filled with the fine inorganic sediments (gytje) and peat during the Holocene.