Environmental Modelling & Software xxx (2014) 1–14

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Holistic environmental soil-landscape modeling of soil organic carbon

Xiong Xiong [a], Sabine Grunwald [a,*], D. Brenton Myers [b], Jongsung Kim [a], Willie G. Harris [a], Nicolas B. Comerford [c]

[a] Department of Soil and Water Science, University of Florida, Gainesville, FL 32611, USA
[b] Division of Plant Science, University of Missouri, Columbia, MO 65211, USA
[c] North Florida Research and Education Center, University of Florida, Quincy, FL 32351, USA

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form
11 March 2014
Accepted 12 March 2014
Available online xxx

Keywords:
Environmental soil-landscape modeling
STEP-AWBH model
Soil organic carbon
Variable selection
Categorical variables
Data reduction

ABSTRACT

In environmental soil-landscape modeling (ESLM), the selection of predictive variables is commonly contingent on the researchers' domain expertise on soil–environment processes. This variable selection strategy may suffer bias or even fail in regions where the process knowledge is insufficient. To overcome this problem, this study demonstrates a holistic ESLM framework which consists of five components: model conceptualization, data compilation, process identification, parsimonious model calibration, and model validation. Based on the STEP-AWBH conceptual model, a comprehensive pool of 210 potential environmental variables that exhaustively cover pedogenic and environmental factors was constructed. This was followed by strategic variable selection and development of parsimonious prediction models using machine learning techniques. The all-relevant variable selection successfully identified the major and minor factors relevant to the SOC variation, showing that the major factors important for explaining SOC variation in Florida were vegetation and soil water gradient. Topography and climate showed moderate effects on SOC variation. Parsimonious SOC models developed using four minimal-optimal variable selection techniques and simulated annealing yielded optimal predictive performance with minimal model complexity. The holistic ESLM framework not only provides a new view of selecting and utilizing variables for predicting soil properties but can also assist in identifying the underlying processes of soil-environment systems of interest. Due to the flexibility of the framework to incorporate various types of variable selection and modeling techniques, the holistic environmental modeling strategy can be generalized to other environmental modeling domains for both prediction and process identification.

1. Introduction

Environmental soil-landscape modeling (ESLM) is a useful tool for predicting soil properties and classes and understanding the relationships between soils and the environmental factors (Grunwald, 2005). The ESLM lays its foundation on work by Jenny (1941) and V.V. Dokuchaev (Glinka, 1927) who conceptualized the soil formation as a function of five factors, i.e., CLimate, Organism, Relief, Parent material, and Time (CLORPT model). It has been undergoing additional development over the past decades. McBratney et al. (2003) first encapsulated the conceptual model into a quantitative framework with the SCORPAN model which describes the relationships between soil and environmental factors for the purpose of spatial prediction of soils. For the past century, human activities have been influencing the environment, exerting critical impact on the pedosphere in terms of soil formation, change, and degradation (Richter et al., 2011). In response, the STEP-AWBH model (S: soil, T: topography, E: ecology, P: parent material, A: Atmosphere, W: Water, B: Biota, H, Human) was devised to explicitly model the effects induced by human activity on the soil system (Grunwald et al., 2011; Thompson et al., 2012).

It is a common practice in ESLM that the environmental factors are selected based on the researchers' domain knowledge of the soil–environment processes in the study area (Florinsky et al., 2002; Grunwald, 2009). This variable selection strategy heavily relies on the legitimacy of the researchers' knowledge. In some cases when the process knowledge is not comprehensive, a limited set of predictor variables could lead to biased and suboptimal model performance (Grunwald, 2009). Therefore, it is necessary to adopt a more unbiased strategy that allows models to access a broad set of environmental variables that represent a spectrum of possible soil-forming processes operating in a given landscape. The more exhaustive such a set of predictive variables is, the higher the

* Corresponding author. Tel.: +1 352 294 3145.
E-mail address: sabgru@ufl.edu (S. Grunwald).

http://dx.doi.org/10.1016/j.envsoft.2014.03.004
1364-8152/

potential is to unravel complex soil–environmental interactions and identify an unbiased, optimal model to predict a soil property of interest. Just as Jakeman et al. (2006) argued, a good practice in the development of environmental models should embrace alternative model families and structures to allow model comparison that avoids biased or even false conclusions drawn from a certain model favored by the model developer(s). Poggio et al. (2013) used stepwise methods to select the most predictive variables from a large pool of satellite-derived variables in a regional application of mapping properties, however their covariate scope was confined in Digital Elevation Model (DEM) and vegetation indices derived from Moderate-Resolution Imaging Spectroradiometer (MODIS) products while other important factors (e.g., soil, atmosphere, water) were not considered.

With the advance of Geographic Information Systems (GIS), Global Positioning System (GPS), and remote and proximal sensing technologies, it is feasible to build a comprehensive pool of spatially exhaustive environmental variables to characterize a full spectrum of environmental properties. These spatially explicit environmental datasets are available in much more abundance and finer spatial resolutions when compared with more sparsely sampled soil pedon data. In that sense, digital environmental covariates serve as critical predictors to infer on soil properties, although it is usually not known which combination of the environmental predictors has the highest predictive power in a given geographic region due to their scale dependent behavior (Vasques et al., 2012). It should be noted that collecting a large set of predictive variables for models can potentially be problematic as well. Some key issues are redundancy and collinearity between the variables, and the deleterious effects of noisy or non-informative variables. Strategic variable selection is required to identify the major ecosystem processes and identify parsimonious predictive models (Guyon and Elisseeff, 2003). In addition, variable selection can reduce model development and application time, increase model interpretability, and reduce overfitting (Belanche-Muñoz and Blanch, 2008; May et al., 2008). Variable selection has been an important research topic in machine learning. It involves two problems – minimal-optimal and all-relevant. The former is aimed at searching for the minimum set of predictor variables yielding the best prediction accuracy (Guyon and Elisseeff, 2003; Nilsson et al., 2007), while the latter is focused on finding all-relevant variables to the target property. Therefore, the minimal-optimal set is of special interest for developing predictive models, while the all-relevant set has great value in understanding the mechanisms underlying the soil–environment relationship. Nilsson et al. (2007) gave an in-depth discussion about the relationships between the two problems and showed that the minimal-optimal set is a subset of the all-relevant set when the data conforms to the strictly positive distribution which is the case for most data encountered in practical applications (Fig. 1).

Environmental variables that represent soil-landscape processes may come as different data types, generally continuous and categorical (including ordinal and nominal). Categorical variables (e.g., land use and geology type) discretize observations (samples) into unbalanced groups and may impose problems for model validation and predictions. A common approach in model validation is data-division, in which the observation data are split into calibration and validation sets (Bennett et al., 2013). The split of data may result in some classes of a categorical predictor under-represented or not represented by the calibration set, which can lead to poor predictions or failed predictions for the underrepresented or non-represented classes. The same issues related to modeling using categorical predictors can occur in validation mode. The occurrence of this problem increase exponentially as the number of categorical variables included in a model increases.

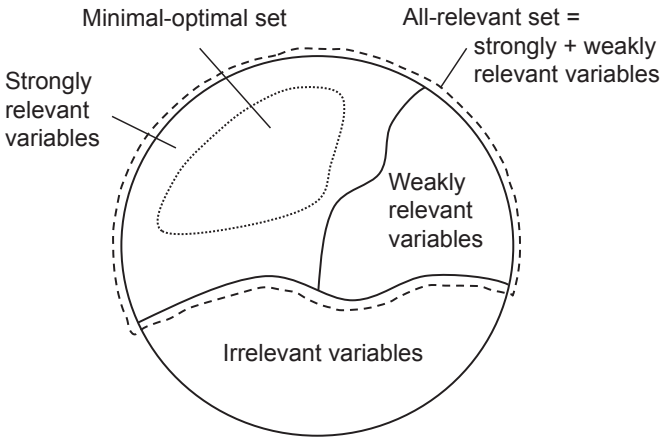

Fig. 1. Topological representation of variable sets. The circle denotes all variables, the dashed line all-relevant set and the dotted line the minimal-optimal set. The figure was redrawn based on Nilsson et al. (2007).

Therefore, it worthwhile to pay special attention to the categorical variables in ESLM and build models that strike the balance between model performance and the number of categorical variables used.

Soil organic carbon (SOC) is a key property that not only indicates soil quality but also has profound significance to the global climate system (Trumbore et al., 1996). Thus, the focus of this study is to model SOC in Florida, USA. The aim of the study was to demonstrate a new holistic ESLM strategy based on a comprehensive environmental variable pool using variable selection techniques that serve two purposes – revealing the underlying processes and making predictions of SOC. It involves five steps – model conceptualization, data compilation, process identification, predictive model calibration and model validation (Fig. 2). The specific objectives are threefold: 1) from a comprehensive set of environmental variables, identify an all-relevant set of variables of topsoil SOC in order to reveal the underlying SOC processes; 2) from the all-relevant set, identify the minimal-optimal sets that simplify models and optimize model performance for prediction; 3) explore the possibility of reducing the use of categorical variables in predictive models.

2. Materials and methods

2.1. Study area

The study area is the state of Florida, located in the southeastern region of the United States, with latitudes from 24°27′ N to 31°00′ N and longitudes from 80°02′ W to 87°38′ W. Florida covers approximately 150,000 km^2 (United States Census Bureau, 2000). The climate is humid and subtropical in northern and central Florida and is humid and tropical in southern Florida. The mean annual precipitation of Florida is 1373 mm and the mean annual temperature is 22.3 °C (National Climatic Data Center, 2008). Overall, soils in Florida are sandy in texture. Dominant soil orders of Florida are: Spodosols (32%), Entisols (22%), Ultisols (19%), Alfisols (13%), and Histosols (11%). Most frequent soil subgroups are: Aeric Alaquods, Ultic Alaquods, Lamellic Quartzipsamments, Typic Quartzipsamments, and Arenic Glossaqualfs (Natural Resources Conservation Service, 2009). Land use and land cover consists mainly of wetland (28%), pinelands (18%), urban and barren lands (15%), agriculture (9%), rangelands (9%), and improved pasture (8%) (Florida Fish and Wildlife Conservation Commission, 2003). Florida's topography is muted with gentle slopes varying from 0 to 5% in almost the whole State (Fig. 3) (United States Geological Survey, 1999).

2.2. Soil organic carbon data

A total of 1080 soil samples in the topsoil (0–20 cm) across Florida (Fig. 3) were collected between 2008 and 2010 based on a random sampling design stratified by the combination of soil suborder and reclassified LULC (Table 1). The number of samples designated to each stratum is proportional to the area of the strata. The reclassification of LULC is based on the data produced by Florida Fish and Wildlife Conservation Commission (2003). Essentially, the original LULC classes with similar soil moisture regime were generalized into a broader class. The purpose of

Please cite this article in press as: Xiong, X., et al., Holistic environmental soil-landscape modeling of soil organic carbon, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.03.004