Catena 156 (2017) 161–175

Contents lists available at ScienceDirect

Catena

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

Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus

Anicet Sindayihebura[a,b,*], Sam Ottoy[b], Stefaan Dondeyne[b], Marc Van Meirvenne[c], Jos Van Orshoven[b]

[a] Department of Earth Sciences, Burundi University, P.O. Box 1550, Bujumbura, Burundi
[b] Department of Earth and Environmental Sciences, University of Leuven, Celestijnenlaan 200E, box 2411, Belgium
[c] Department of Soil Management, Ghent University, Coupure 653, 9000 Gent, Belgium

ARTICLE INFO

Keywords:
Auto-correlation
Choropleth map
Digital elevation model
Prediction
Residual kriging
SCORPAN

ABSTRACT

Whereas contemporary land use and land management planning require specific and quantitative georeferenced soil information, often only general-purpose and qualitative soil maps are available. With a view to fill this gap for topsoil clay and organic carbon content in the central plateaus of Burundi, we tested for a representative 15 km^2 hilly landscape, six types of SCORPAN models. The SCORPAN models were first applied as standalone trend models and next extended with a component accounting for the spatial autocorrelation of the residuals from the trend. Various sets of predictors, including class variables derived from the available soil map and continuous derivatives from a Digital Elevation Model (DEM) and from Landsat-imagery were incorporated. For clay, the best prediction method was a Residual Kriging (RK) using a Generalized Additive Model (GAM) as trend built with only DEM derivatives and spectral normalized difference vegetation index (NDVI). Furthermore, the classical and simplest RK, i.e. using a Least Squares Linear Regression (LR) trend built with only continuous covariates, outperformed all standalone trend models. For organic carbon, residuals from the trend models were not significantly auto-correlated, making RK meaningless. In this case the best model was a GAM combining lithologic units with DEM derivatives and NDVI. Overall, the contribution of soil map–derived predictors to the model performance was rather weak. It was concluded that, for prediction of specific soil characteristics in the study area, a SCORPAN approach is preferred the more as the performance can be boosted by kriging of trend residuals if auto-correlated.

1. Introduction

Current approaches in mapping of soil characteristics explicitly account for spatial variation of Jenny's soil forming factors (Jenny, 1941) and for possible spatial auto-correlation. Such approaches were termed *digital soil mapping* and formalized in so-called "SCORPAN-SSPFe" models (McBratney et al., 2003), where "SCORPAN" stands for the Jenny's soil forming factors supplemented with geographic position (n): soil (s), climate (c), organisms (o), relief (r), parent material (p) and age (a), and "SSPFe" for "Soil Spatial Prediction Function with spatially auto-correlated errors". This family of models is attractive in that it can incorporate various kinds of previous knowledge in the form of trend components, including conventional soil maps, while compensating for their limitation by accounting for spatial auto-correlation of residuals from the trend components using kriging.

Conventional choropleth soil maps divide in categorical soil mapping units (SMUs) what often is a continuum. Moreover, they are, at least partly, constructed based on (expert) tacit knowledge (Carré et al., 2007), which prevents their reproduction, even among expert soil surveyors. The prediction ability of choropleth soil maps has been augmented by: (1) extrapolating soil profile data to non sampled SMUs by a multi-level statistical approach (Ottoy et al., 2015), (2) integrating the accuracy and sharpness of soil map delineations into ordinary kriging (Boucneau et al., 1998), (3) extrapolating punctual profile data to all locations in a SMU based on correlation between target soil attributes and environmental covariates (Meersmans et al., 2008; Moore et al., 1993).

Other researchers have taken another path consisting in segmentation of the landscape into units to be used as alternatives to SMUs of conventional soil maps. For example, MacMillan (2003) and MacMillan et al. (2000) developed so-called LandMapR toolkits to extract a series of contextual topographic variables based on which they defined landforms in terms of relative slope position that they found significant in explaining variations of soil properties and crop yield in Canada.

* Corresponding author at: Department of Earth Sciences, Burundi University, P.O. Box 1550, Bujumbura, Burundi.
E-mail address: anicet.sindayihebura@ub.edu.bi (A. Sindayihebura).

http://dx.doi.org/10.1016/j.catena.2017.04.003
Received 20 November 2015; Received in revised form 16 February 2017; Accepted 1 April 2017
Available online 26 April 2017
0341-8162/

Florinsky et al. (2002) segmented the landscape into so-called accumulation, transit and dissipation zones, and found that this segmentation explained a larger proportion of variance of solum thickness than linear regression with continuous DEM derivatives. Wood (1996, 2009) applied the multi-scale approach to model channels, ridges, pits, peaks, passes, and planar regions from a DEM. Although not initially oriented to soil mapping, this approach is easily repeatable by different users as it relies less on expert knowledge. In the framework of the Global Soil Partnership and the e-SOTER project (FAO and ISRIC, 2012), DEM-derived landforms were presented as proxy for improving existing soil maps, or as alternatives to soil maps in countries with limited soil data (Köthe, 2012). The lack of soil maps and associated data is obvious in countries where the progress of national surveys is limited (Mora-Vallejo et al., 2008).

Choropleth soil maps have also been used as a source of covariates. Many other covariates are increasingly becoming available and multivariate prediction models able to handle different types of covariates have been devised and improved. In this respect, ordinary kriging, also referred to as the best linear unbiased predictor (Matheron, 1963) has been updated to more accurate geostatistical methods that account for secondary information, of which Residual Kriging (RK) was found the most flexible. RK is a "SCORPAN-SSPFe" model where the trend component is usually obtained, either by global linear regression of continuous covariates on the target variable, or by geo-matching (Goovaerts, 1997). However, soil variation is often too complex to be modeled by such simple trend models. For example, using RK based on a trend built with global linear regression of covariates from LUCAS database for mapping topsoil organic carbon of Europe, de Brogniez (2015) obtained unrealistic hot-spots of organic carbon content in northern Europe. She explained this by the difficulty to model the presence of microtopography and related high variability of water regime that influences mineralization rate and organic carbon dynamics. Recently more complex modeling techniques have been proposed. For example, Kumar et al. (2012) used Geographically Weighted Regression (GWR) to spatially adjust the trend component of RK and found this approach to be less biased and more accurate for predicting soil organic carbon stocks than RK based on classical linear regression. For predicting nitrogen oxides levels in South California, Li et al. (2012) found that co-kriging of the residuals of a Generalized Additive Model (GAM) was better than universal kriging, multiple linear regression and GAM alone. Dai et al. (2014) applied kriging to organic carbon residuals from a machine learning technique, Artificial Neural Networks (ANN) and this was more accurate than Inverse Distance Weighting, Universal Kriging and ANN alone. Boosted Regression Trees (BRT) is another powerful machine learning technique (Heremans et al., 2015; Martin et al., 2014; Van Meerbeek et al., 2014). Martin et al. (2014) compared BRT and related RK for the prediction of soil organic carbon stocks in France by varying the number of covariates. With a limited number of covariates, RK significantly improved BRT predictions. But when several covariates were included in the BRT model, the spatial auto-correlation of BRT residuals almost vanished, and RK did not significantly improve the standalone BRT predictions. Vaysse and Lagacherie (2015) predicted several soil characteristics from legacy soil profile data in the Languedoc-Roussillon region (France) using topographic, geologic, and climatic and land use data as covariates and Random Forests as prediction model. They found that Random Forests captured most of the spatially-structured variance of the soil characteristics shown by the available soil data. Aertsen et al. (2012) also obtained non-auto-correlated residuals from GAM estimates of forest site index. In principle, any prediction technique can be cast in RK, provided that related residuals are spatially auto-correlated. This is the reason why we prefer using the term "Residual Kriging", although the same technique has been referred to as "Simple Kriging with varying local means" in Goovaerts (1997) or as "Regression Kriging" e.g. in Hengl et al. (2004).

The lack of spatial auto-correlation in residuals from BRT (Martin et al., 2014), Random Forests (Vaysse and Lagacherie (2015) and from GAM (Aertsen et al., 2012) raises the question whether in such cases these standalone models can also outperform RK based on other trend model types, thereby simplifying the SCORPAN-approach. Furthermore, as exemplified in the above studies, often one soil variable was studied, whereas in general soil users need quantitative information on many soil characteristics, possibly with different dynamics. Lastly, soil maps of many regions and of Burundi in particular are based on limited field observations, and hence SCORPAN methods could be useful for improving soil information in these regions. In this paper the objective was to compare, based on a case study in the Burundi central plateaus, the performance of digital soil mapping methods for the spatial prediction of clay and organic carbon (OC) content. Clay content and OC content are two key soil characteristics strongly related to soil fertility, water retention, soil erodibility, water pollution and soil carbon dynamics and hence to land use planning and land management (Tiessen et al., 1994; Baize, 2000). We adopted a four-level methodological approach, as follows:

a. Derive a trend model for OC and for clay content from the extract of the choropleth soil map by geo-matching point observations with SMUs;
b. Test alternative trend models based on easily available environmental covariates. Six models were tested: (1) geo-matching using land units obtained by spatial overlay of DEM-derived landforms and lithologic units, (2) least squares linear regression (LR), (3) GWR, (4) GAM, (5) BRT and (6) ANN;
c. Test for spatial auto-correlation among residuals from all seven models mentioned in a. and b., and conduct RK if residuals are significantly auto-correlated;
d. Compare the prediction performance of all seven trend models each without and with related RK.

2. Materials and methods

2.1. Study area

The study area (between 3°54′33″–3°56′44″ S and 29°41′30″–29°45′18″ E) is located in the Burundi central plateaus, which cover more than half of the country. It is a 15 km^2 area composed of two contiguous catchments drained by Mutandu River and Nyabuyumpu River, respectively. Mutandu and Nyabuyumpu Rivers are tributaries of the Jiji River and Siguvyaye River, respectively, which in turn flow into Lake Tanganyika (Fig. 1).

The study area is composed of meta-sedimentary and granitic rocks that belong to the Karagwe-Ankolean Belt, with Holocene alluvial deposits in the valley bottoms. The Karagwe-Ankolean Belt is an orogenic belt of middle Proterozoic age that spans Burundi, Rwanda, SW Uganda, NW Tanzania and the Kivu-Maniema region in DR Congo (Tack et al., 2010; Fernandez-Alonso et al., 2012). Both meta-sedimentary and granitic rocks are deformed along a North–South direction. The research was conducted at catchment scale so as to account for landscape complexity at the *catena* level, i.e. from the valley floor to the hilltop or catchment divide. This landscape is hilly with round-shape summits. Hillsides are occupied by small farms separated by grasslands (mainly *Eragrostis*) and patches of woodlands (mainly eucalyptus, occasionally coniferous trees). Subsistence agriculture is dominant and crops are usually intercropped. These are bananas, maize, beans, peas and Irish and sweet potatoes. The same crops are also grown in valleys, except for bananas. Fallows are dominated by *Digitaria abyssinica*.

2.2. Datasets

2.2.1. Response variables

The response variables were topsoil (0–30 cm depth) clay and OC