Geoderma 192 (2013) 341–353

Contents lists available at SciVerse ScienceDirect

Geoderma

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

A methodology for digital soil mapping in poorly-accessible areas

A.H. Cambule [a,*], D.G. Rossiter [b], J.J. Stoorvogel [c]

[a] Universidade Eduardo Mondlane, Faculdade de Agronomia e Engenharia Florestal, C.P. 257, Maputo, Mozambique
[b] University of Twente, Faculty of Geoinformation Science and Earth Observation (ITC), P.O. Box 217, AA Enschede, The Netherlands
[c] Wageningen University, Land Dynamics Group, P.O. Box 47, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 7 July 2011
Received in revised form 16 May 2012
Accepted 19 August 2012
Available online 17 November 2012

Keywords:
Soil organic carbon
Landscape
Kriging
Limpopo National Park
Mozambique
Mapping inaccessible areas

ABSTRACT

Effective soil management requires knowledge of the spatial patterns of soil variation within the landscape to enable wise land use decisions. This is typically obtained through time-consuming and costly surveys. The aim of this study was to develop a cost-efficient methodology for digital soil mapping in poorly-accessible areas. The methodology uses a spatial model calibrated on the basis of limited soil sampling and explanatory covariables related to soil-forming factors, developed from readily available secondary information from accessible areas. The model is subsequently applied in the poorly-accessible areas. This can only be done if the environmental conditions in the poorly-accessible areas are also found in the accessible areas in which the model is developed. This study illustrates the methodology in an exercise to predict soil organic carbon (SOC) concentration in the Limpopo National Park, Mozambique. Readily-available secondary data was used as explanatory variables representing the soil-forming factors. Conditions in the accessible and poorly-accessible areas corresponded sufficiently to allow the extrapolation of the spatial model into the latter. The spatial variation of SOC in the accessible area was mostly described by the sampling cluster (71.5%) and the landscape unit (46.3%). Therefore ordinary (punctual) kriging (OK) and kriging with external drift (KED) based on the landscape unit were used to predict SOC. A linear regression (LM) model using only landscape stratification was used as control. All models were independently validated with test sets collected in both accessible and poorly-accessible areas. In the former the root mean squared error of prediction (RMSEP) was 0.42–0.50% SOC. The ratio between the RMSEP in the poorly-accessible and accessible areas was 0.67–0.72, showing that the methodology can be applied to predict SOC in poorly-accessible areas as successful as in accessible areas. The methodology is thus recommended for areas with similar access problems, especially for baseline studies and for sample design in two-stage surveys.

1. Introduction

Effective land management depends on knowing the spatial distribution of soil properties. Traditionally this knowledge is represented as soil maps conforming to the discrete model of spatial variation; DMSV (Heuvelink and Webster, 2001), showing polygons within which soils are considered homogeneous and with boundaries where changes in soil properties are considered to be abrupt. However, many soil properties can be better modeled with a continuous model of spatial variation (CMSV), in which properties vary continuously in space. The recent rapid development of information technology along with the availability of new types of secondary data (e.g., digital elevation models and satellite imagery) allow for more quantitative approach to soil survey producing continuous surfaces based on soil forming factors. Furthermore, these methods give spatial estimates of the uncertainty of the predictions. This "predictive" (Scull et al., 2003) or "digital" soil mapping (McBratney et al., 2003) uses relationships between soil properties and auxiliary data at sample points to predict over a study area. in addition, the high sampling costs can be reduced by applying recent developments in the field of diffuse reflectance spectroscopy (e.g. near-infrared spectroscopy), a fast, non-destructive and inexpensive soil analysis method that can enhance or replace traditional laboratory methods (Shepherd and Walsh, 2002; Viscarra-Rossel and McBratney, 2008).

Digital soil mapping (DSM) techniques have been successfully applied in studies at field scale where soil variability is largely due to the effect of topography on soil genesis (e.g., Florinsky et al., 2002) and therefore much of the success is attained by integration of terrain attributes as auxiliary data. To capture the spatial structure of soil variation as well as the soil–environment relations over larger poorly-accessible areas due to poor road networks (such as much of Africa) or difficult terrain (e.g., mountainous regions), a large number of observations following a sound sampling design, covering the feature and geographic space of the predictors (e.g., Minasny and McBratney, 2006) are required, which is impractical or prohibitively expensive. A DSM approach which can concentrate sampling in accessible areas, yet deliver results of sufficient quality, would greatly reduce costs and survey effort.

Our objective was to develop a methodology for DSM for poorly accessible areas. It consists in developing a quantitative predictive

* Corresponding author. Tel.: +258 843285400.
E-mail address: armindo.cambule@uem.mz (A.H. Cambule).

0016-7061/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2012.08.020

model based on limited sampling (mainly in accessible areas) combined with readily-available auxiliary spatial data representing soil forming factors. We hypothesize that if the auxiliary data in accessible and inaccessible areas are sufficiently similar, models built in the former can be applied in the latter, with very few or even no soil samples. It is thus applicable in mapping projects where legacy samples from accessible areas are available.

This paper first explains the proposed method, then introduces a case study where it is tested and discusses the performance of the method in the test area.

2. Proposed method

The proposed method is based on similarities between accessible and poorly-accessible areas in terms of the relation between soil-forming explanatory variables (covariables) and soil properties (target variables). If the areas are similar, the predictive model based on soil samples and explanatory variables from accessible areas can be applied in inaccessible areas. The predictive model uses the conceptual model *scorpan-SSPFe* proposed by McBratney et al. (2003) and widely-applied as a generic method for DSM. *Scorpan* represents the list of soil-forming factors that has been expanded from the original definition by Jenny (1980) representing the initial soil conditions (s), climatic conditions (c), organisms (o) including animals, land cover and human occupation; relief (r), parent material (p), age (a), and the neighborhood (n). The conceptual model uses a soil spatial prediction function with spatially-autocorrelated errors (*SSPFe*) that uses (1) a prediction based on environmental covariables and (2) a prediction based on soil properties measured at a limited set of observation points.

3. Implementation approach

The methodology for mapping poorly accessible areas in the LNP is shown schematically in Fig. 1. We believe the methodology has potential for worldwide application, which we illustrate with the LNP case. Below is described each step of the methodology.

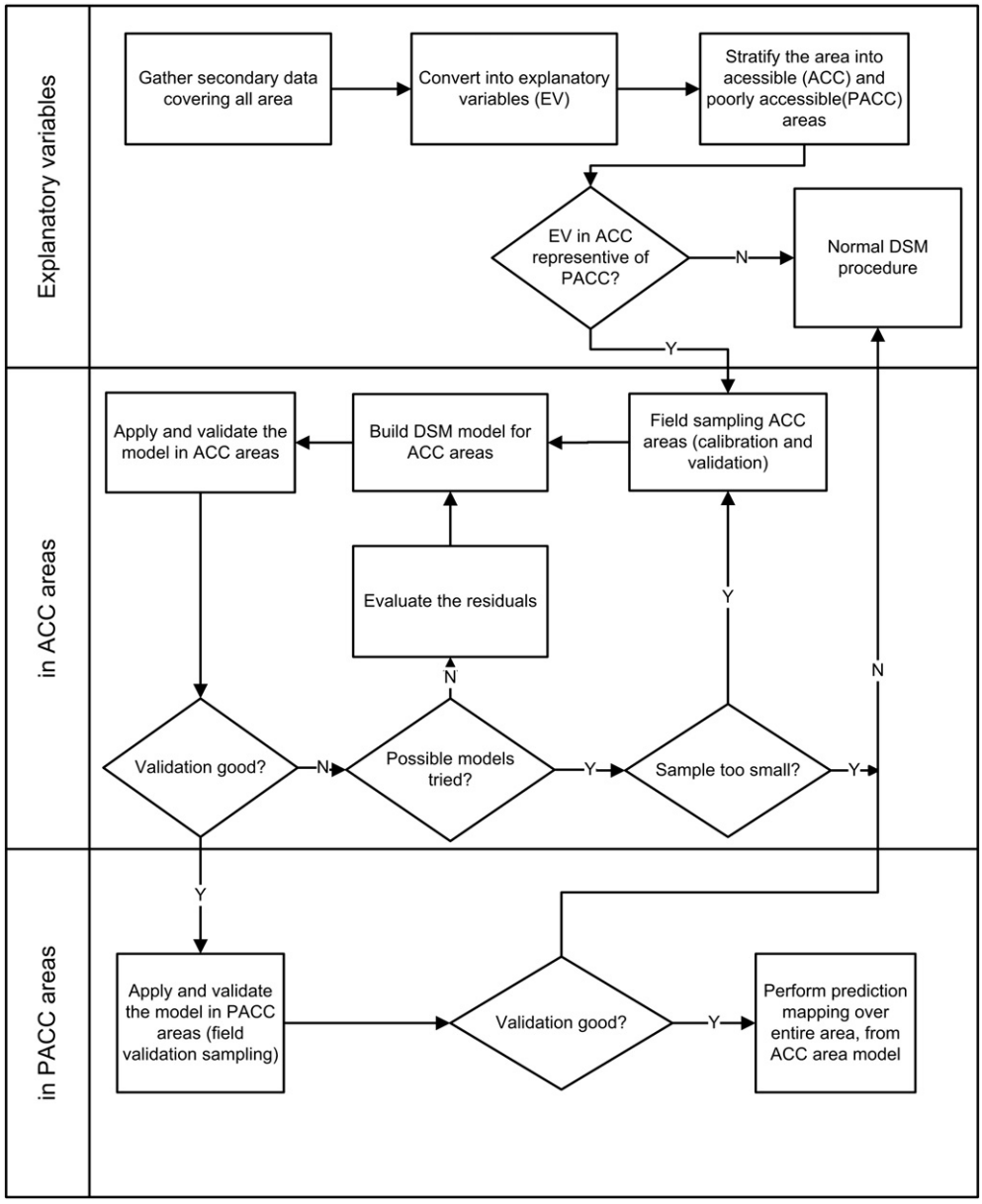


Fig. 1. Flowchart of the proposed methodology for digital soil mapping in poorly accessible areas.

3.1. Gathering of secondary data

Secondary data is the raw material to derive insight in the soil forming factors. The secondary data must cover the area of interest (i.e., have values at all locations). A good example is a digital elevation model that covers the area and provides insight in the soil forming factor "relief" (r).

3.2. Conversion of secondary data into covariables with a direct link to soil formation

The Scorpan approach aims to elucidate quantitative relationships between soil properties and the soil-forming factors. The covariables should provide information on soil formation. If covariables describing relevant soil formation processes are lacking the predictive power of the model will be limited. An example is the conversion of a DEM into terrain derivatives such as a wetness index and potential erosion rates (Gessler et al., 2000; McKenzie et al., 2000).

3.3. Stratification on the base of accessibility

The study area should be divided into accessible (ACC) and poorly-accessible (PACC) areas. The latter are those beyond easy reach by common means due to, for example, poor road infrastructure, difficult navigation, wildlife hazard, or poor security.

3.4. Evaluation of similarities between ACC and PACC areas in terms of covariables

ACC and PACC must be compared to evaluate the degree to which conditions in PACC areas are found in the ACC areas. This determines the potential applicability of the methodology. Similarities can be assessed by comparing, e.g., the histograms, ranges, clusters, class frequencies, or trends of covariables between the two areas, either qualitatively or with formal similarity measures. A decision is taken as to whether PACC areas are sufficiently represented by ACC areas; if not, the method is not applicable and both areas need to be sampled.

3.5. Sampling in accessible area

A sampling strategy must be designed and implemented to gather a representative sample of the target soil properties in ACC areas. The sampling strategy will be based on the available information from the covariables (Minasny and McBratney, 2006), the expected spatial structure (Lark, 2002; Webster et al., 2006), or a combination (Brus and Heuvelink, 2007). If there are legacy samples, and if these can be harmonized with current methods, they can be used to optimize the sampling plan, e.g., by simulated annealing (Brus and Heuvelink, 2007).

3.6. Building of a quantitative calibration model for the accessible areas

The correlation between explanatory variables (i.e., the environmental covariables) and the soil properties of interest must be evaluated, using pedometric modeling approaches (McBratney et al., 2000, 2003) to build a quantitative model for the accessible areas. A separate model must be built for each property. The model may include local spatial correlation, e.g., regression kriging (Hengl et al., 2007), but since the PACC area is by definition not or very sparsely sampled, the local spatial structure cannot be used to explain much of the variability in these areas.

3.7. Application of the model in accessible areas

The calibrated model must be applied to the environmental covariables and measured soil properties to make a prediction map of the soil properties of interest across ACC areas. This should also produce an estimate of the prediction variance as an internal measure of model quality.

3.8. Model validation in ACC areas

An independent field sample must be taken, using the same strategy as the calibration sample; for practical reasons this could be during the original sampling campaign, with a proportion taken out randomly for this validation. The model prediction must then be compared to the true values with measures of quality such as root mean square error of prediction (RMSEP), or bias and gain of modeled vs. actual. If the model quality does not match requirements, one of the following corrections must be undertaken: (1) try another model structure or explanatory variables; (2) make more observations to refine the model; (3) abandon the DSM project if properties cannot be predicted with this approach.

3.9. Application of the model in PACC areas

The calibrated model must be applied to the environmental covariables and field-sampled soil properties (from ACC areas) to make a prediction map of the soil properties of interest across poorly-accessible areas. If there is any local spatial structure represented in the model, the prediction quality will naturally be better nearer to ACC areas.

3.10. Model validation in PACC areas

An independent field sample using the same strategy as the validation set in ACC must be taken; but this will be by definition quite limited (this is the motivation of the methodology), given the difficulty of access. Validation is as in step 8.

3.11. Assessment of the relative performance of the prediction model in PACC areas

Validation results in the two areas must be compared; the ratio between the validation RMSE and other validation statistics should then be used to determine the degree of success of the methodology for poorly-accessible areas. The performance in accessible areas should already have been judged adequate (step 8); if the relative performance in poorly-accessible areas was satisfactory, by deduction so will be the absolute performance. If relative performance is too poor, there is no remedy but to conduct a full (expensive) sampling in the poorly-accessible area, following the same scheme that produced a satisfactory result in the accessible areas.

4. Test case for Limpopo National Park

4.1. Study area

The proposed method arose from a research objective to assess soil organic carbon (SOC) stocks in the Limpopo National Park (LNP), Mozambique, as part of a project to understand how competing claims on natural resources affects land use and livelihoods (Giller et al., 2008). This objective is a good test case because (1) LNP is a conservation area where SOC plays a vital role for natural vegetation growth which is the source of wildlife nutrition; (2) SOC is considered a key indicator of soil quality (Cécillon et al., 2009; Yemefack et al., 2006), playing a vital role in ecosystem function, determining soil fertility, water holding capacity and susceptibility to land degradation (Milne et

al., 2007); (3) SOC has been the focus of many studies, especially in the context of land degradation, climate change and loss of biodiversity (Gisladottir and Stocking, 2005) but also in digital soil mapping both in the early development (Bell et al., 2000; Florinsky et al., 2002; Simbahan et al., 2006) and recently with the emphasis on SOC in public policy (Li, 2010; Miklos et al., 2010; Phachomphon et al., 2010; Ungar et al., 2010; Vasques et al., 2010) so that prior knowledge on possible environmental predictors is available; (4) a single soil property should be simpler to model for evaluation of methodology's potential than a set of properties, a land quality, or a soil function.

The 10.400 km^2 LNP is located in southwest Mozambique, between 22″ 25′ and 24″ 10′ S and 31″ 18′ and 32″ 38′ E. It forms part of the proposed Great Limpopo Transfrontier Park, which also includes Kruger National Park (KNP, South Africa) to the west and Gonarezhou National Park (Zimbabwe) to the north. The LNP is bounded by the Elephant River to the south and the Limpopo River to the north and east; the Singwedzi River flows through the park in a NW–SE direction, joining the Elephant River at the southern border. The area was declared a national park in 2001, after long been used as a hunting zone. LNP has a warm arid climate with dry winters and a mean annual temperature exceeding 18 °C. Maximum temperatures increase northwards to above 40 °C (between November and February). Annual rainfall decrease northwards from above 500 mm around the confluence of Limpopo and Elephant Rivers to about 350 mm at the extreme north (Ministerio do Turismo, 2003; Stalmans et al., 2004). Dominant geological features include (1) the extensive sandy (aeolian sands) cover along the NNW–SSE spine of the park lying to a greater extent between the Limpopo and Singwedzi Rivers, (2) sedimentary rocks (limestones, sandstone) where the sand mantle has been exposed closer to the drainage lines, (3) rhyolite rocks along the western border with KNP, and (4) alluvial deposits along the main drainage lines (Manninen et al., 2008; Rutten et al., 2008). Soils derived from aeolian sands range from shallow to deep, those derived from rhyolite are shallow and clayey, those derived from sedimentary rocks are deep, structured and clayey and those derived from alluvium materials are clayey (Stalmans et al., 2004). LNP is dominated by plant communities with *Colophospermum mopane* classified in as the mopane vegetation of the Sudano–Zambezian region. The mopane vegetation can be found throughout the park except for the aeolian sands and in waterlogged landscapes along the major drainage lines. The LNP was classified into ten landscape units by Stalmans et al. (2004). These landscape units were delineated on the basis of plant communities and environmental characteristics. LNP is covered by (1) (semi-) deciduous open forest composed of broadleaved deciduous and semi-deciduous woodlands, (2) (semi-) evergreen open forest made up of broadleaved evergreen and semi-evergreen woodland, (3) (semi-) deciduous forest which includes the broadleaved deciduous and semi-deciduous trees, (4) (semi-) evergreen forest made up of broadleaved evergreen and semi-evergreen trees, (5) thickets referred to closed shrubland, (6) Open shrublands, and (7) grasslands made up of herbaceous closed to open vegetation (the national land cover and land use map; Cenacarta, 1997).

4.2. Operationalizing the methodology: results and discussion

This section explains how the proposed methodology was carried out for the specific objectives and context of the test case, with the steps in parallel with Section 2. To avoid duplication, the results and discussion are included as well. It illustrates the decisions that must be made, and how they can be justified. All statistical analyses were carried out in the R environment for statistical computing (R Development Core Team, 2011) version 2.12 including geostatistical analyses with the gstat R package (Pebesma, 2004) version 1.0.

Table 1
Summary statistics of the soil-forming explanatory variables in LNP as a whole.

Variable	unit	Min	Max	Range	Mean	SD
Elevation	m	54	531	477	241	99
Flow accumulation	no. of pixels	0	50	50	4	8.2
NDVI wet season	–	−1.0	0.69	1.69	0.35	0.13
NDVI dry season	–	−0.34	0.56	0.91	0.11	0.08
Annual precipitation	mm	362	580	218	461	40

4.2.1. Gathering of secondary data

Selected secondary data for SOC prediction in LNP included (1) mean annual precipitation at a 30 arc-sec resolution grid from the WorldClim database (Hijmans et al., 2011), multispectral Landsat TM satellite imagery at a 30 m resolution for a wet and dry season from the US Geological Survey (www.usgs.gov, row/path: 168/076 from August 2009, preprocessing: L1T level), a digital elevation model at a 3 arc-sec (approximately 90 m) resolution from the shuttle radar topographic mission from the Jet Propulsion Laboratory (www.jpl.nasa.gov, tile: 43_17, preprocessing: research grade), a 1:250,000 lithology map developed by the Geological Survey of Finland (Manninen et al., 2008; Rutten et al., 2008), and a 1:1,000,000 scale landscape map of Stalmans et al. (2004). These latter two are equivalent to, at best, 125 m and 500 m resolution, respectively (Hengl, 2006, Section 2.1), the latter somewhat smaller than the largest cluster dimension, 720 m. However, considering the size of the study area, we decided on a 1 km resolution (thus, about 10,000 pixels) for the final maps.

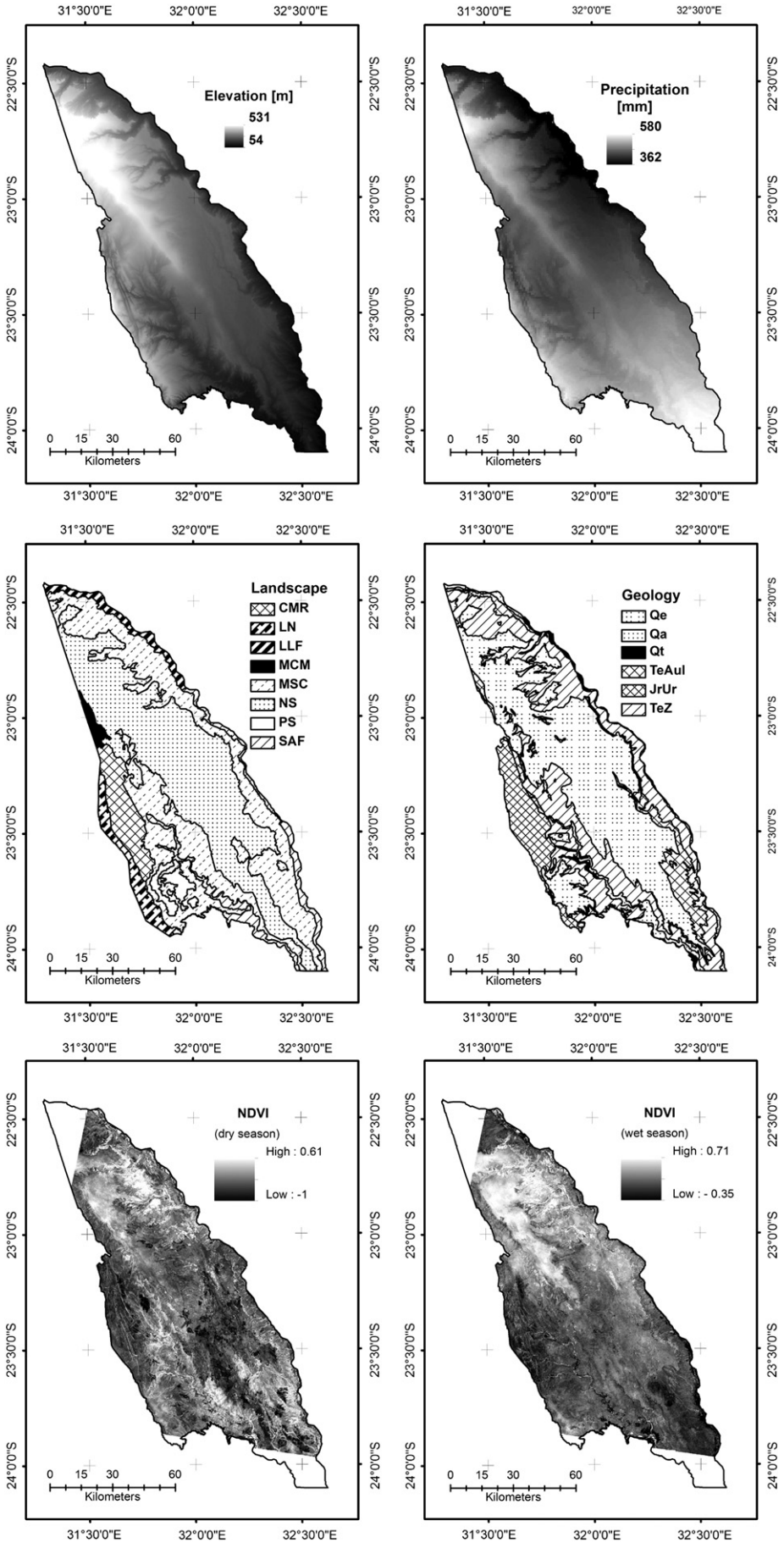
4.2.2. Development of explanatory variables

At this stage coverages were produced from the available secondary maps with potential covariables. The *scorpan-SSPfe* modeling framework was used to organize the coverages by soil-forming factor. Spatial resolution at this stage is kept the same as of the original coverages as it is meant for similarity analysis. We have assumed the time factor as a constant for the present study and therefore no analysis were performed. The summary statistics are presented in Table 1.

Climate (c) influences rates of vegetative growth and turnover of soil organic matter through differences in precipitation, temperature and evaporation (McBratney et al., 2003). The WorldClim database shows a clear rainfall increase to the south with an annual precipitation difference of approximately 220 mm (Fig. 2). The higher grounds in the SW and NW also show precipitation above 500 mm as it is with the SE corner of the study area. Summary statistics (Table 1) show a mean bellow 500 mm, which indicate a rather drier climate. Temperature and evaporation do not vary substantially across the area and therefore were left out.

The most influential organisms (o) for SOC are vegetation and humans (McBratney et al., 2003). The LNP was long used as a hunting zone in colonial times. Later it was declared a conservation area with minimal human influence is minimal, with confined to scattered subsistence farming near the Singuezi River. Wildlife density is low, and therefore vegetation is the principal organism related to SOC. Normalized vegetation index (NDVI) is a surrogate of vegetation biomass (whose decay contributes to SOC) and is calculated as $(NIR-G)/(NIR+G)$ where NIR and G are the reflectances in the near-infrared and green electromagnetic spectrum, respectively. Green NDVI is sensitive to chlorophyll concentrations, adequately measuring the rate of photosynthesis (Gitelson and Merzlyak, 1998; Yoder and Waring, 1994), and can

Fig. 2. Digital elevation model (SRTM), spatial distribution of annual precipitation (Hijmans et al., 2011), the landscapes (Stalmans et al., 2004, modified with permission from Koedoe), the geology and dry and wet season NDVI derived from Landsat TM in the study area.


Elevation [m]
531
54
Precipitation [mm]
580
362
Landscape
CMR
LN
LLF
MCM
MSC
NS
PS
SAF
Geology
Qe
Qa
Qt
TeAul
JrUr
TeZ
NDVI
(dry season)
High : 0.61
Low : -1
NDVI
(wet season)
High : 0.71
Low : - 0.35
0 15 30 60
Kilometers
31°30'0"E
32°0'0"E
32°30'0"E
22°30'0"S
23°0'0"S
23°30'0"S
24°0'0"S

therefore be used as an indicator of vegetation cover. Focal statistics at 3×3 pixels were applied to the NDVI grid in order to match the spatial resolution of NDVI with the support of the sample sites. Dry-season NDVI is an indicator of water availability and hence biological activity in that season. Wet-season NDVI is an indicator of maximum vegetative growth. NDVI for the wet (February) and dry (June) seasons were selected to represent the soil forming factor organism in agreement with the study of Mora-Vallejo et al. (2008) in Kenya. Wet and dry season NDVI (Fig. 2) are derived from Landsat TM scenes that cover most of park. In general higher NDVI values are found along the main drainage lines and at higher grounds of the northern section along the NNW–SSE spine of the park. At this location large patches of distinctive dense 5–10 m high and evergreen *Androstachys johnsonii* forests are located (Stalmans et al., 2004). Summary statistics (Table 1) show the wider range in dry season NDVI. Lower values are found in the southern section with aeolian sands due to the dry conditions and higher values along the drainage lines. Wet season NDVI shows a much wider spatial distribution of higher values, spanning beyond the main drainage lines. This is a result of the vegetation growth during rainy season.

Relief (r) influences water movement and accumulation across the landscape. As a result, relief has indirect consequences on SOC contents through biomass production, erosion, sedimentation and redox conditions. Altitude and flow accumulation (an indirect way of measuring drainage area) were selected as appropriate covariables. Higher elevations are located at the extreme north of the NNW–SSE spine of the park and along the western border with KNP. Lower elevations are found along the major drainage lines. Overall elevation ranges approximately 250 m with a standard deviation is about 20% (Table 1). Flow accumulation was derived from the DEM using ArcGIS 10. Values above 50 pixels are excluded as they correspond to drainage lines. The 50th percentile of flow accumulation was zero (0) indicating that most of the study area has no flow accumulation as a result of the almost flat topography. The summary statistics in Table 1 show a standard deviation twice as higher than the mean, which may indicate the influence of the extreme higher values on the mean and therefore an evidence of the almost flat topography.

Parent material (p) was represented by the lithology map. Six major geological units cover the study area (Fig. 2), three of which are bedrock (sandstone, limestone, and rhyolite) and three surficial sediments (Aeolian sands, fluvial terrace gravel and sand, and alluvium-gravel-and-silt). Small units were merged with neighboring larger ones of similar lithology to avoid a large number of different units.

The spatial factor (n) accounts for spatial trends not revealed by other factors (McBratney et al., 2003). Although in principle any trend should be reflected by the soil forming factors, the selected covariables may not capture all the regional variation. Hence the spatial position was represented by the coordinates.

The soil factor (s) represents soil attributes measured at sampling locations. We have used the SOC concentrations derived from a *partial least square regression* (PLSR) calibration model relating the near-infrared spectral signature of a soil sample to its SOC concentration (%) determined by Walkley–Black method. Field sampling and laboratory analysis details are described below.

Finally, we took advantage of the landscape study of Stalmans et al. (2004) to consider the landscape units as an integrated soil-forming factor, combining elements of lithology, general relief, climate, and soil type into a local eco-region. Stalmans et al. (2004) classified LNP into ten major landscape units (1) *Combretum* spp./*C. mopane* Rugged Veld (CMR), characterized by shallow soils on the hills but deeper in the footslopes and low-lying areas, (2) Limpopo Levubu Floodplains (LLF), subjected to flooding and characterized by sandy alluvial soils, (3) Limpopo north (LN), stoney with loamy to clayey shallow soils derived from rhyolite but also basalts, (4) Mixed *Combretum* spp./*C. mopane* woodland (MCM), made up mainly by rhyolite rock-outcrops, (5) Mopane Shrubveld on Calcrete (MSC), with shallow and calcareous soils derived from sandstones and limestones, (6) Nwambia Sandveld (NS), sandy soils of varying depth derived from the aeolian sands, (7) Pumbe Sandveld (PS), similar to NS but receives more rain and has red sandy soils, (8) *Salvadora angustifolia* Floodplains (SAF), subjected to flooding with black alluvium soils, (9) *Andasonia digitata*/*C. mopane* Rugged Veld (ADR), shallow and calcareous soils with moderate clay concentration, and (10) *C. mopane* Shrubveld on Basalt (CMB), dark soils derived from basalt showing vertic properties. Given the rocky nature of the LN and MCM units, we assumed their SOC contents to be zero. The remaining eight landscape units were reduced to six (Fig. 2) by merging the very small units (<0.1% of total area) ADR into NS and CMB into MCM.

4.2.3. Stratification of the study area based on accessibility

The main road network is comprised of two dirt roads following the N-NW direction, one along the right margin of the Limpopo River, while the other is located about the center of the park, along Singwedzi River (parallel to the Limpopo River). A few other roads connect these main roads. We mapped the road network using a handheld GPS while traversing the entire network in an all-terrain vehicle. Areas within 2.5 km of a road were considered accessible areas. This threshold was considered a practical limit of easy access for field sampling (including carrying tools, water, samples, and a firearm for protection against wildlife) after parking a vehicle along the road. The ACC areas covered 27% of LNP (Fig. 3).

4.2.4. Evaluation of similarity between accessible and poorly-accessible areas

The proposed methodology relies on an adequate similarity between the ACC and the PACC areas. Only if the ecological conditions of the PACC areas are also found in the PACC areas can the prediction model developed for the ACC areas can be applied in the PACC. To evaluate the similarity between the two areas we compared the mean and

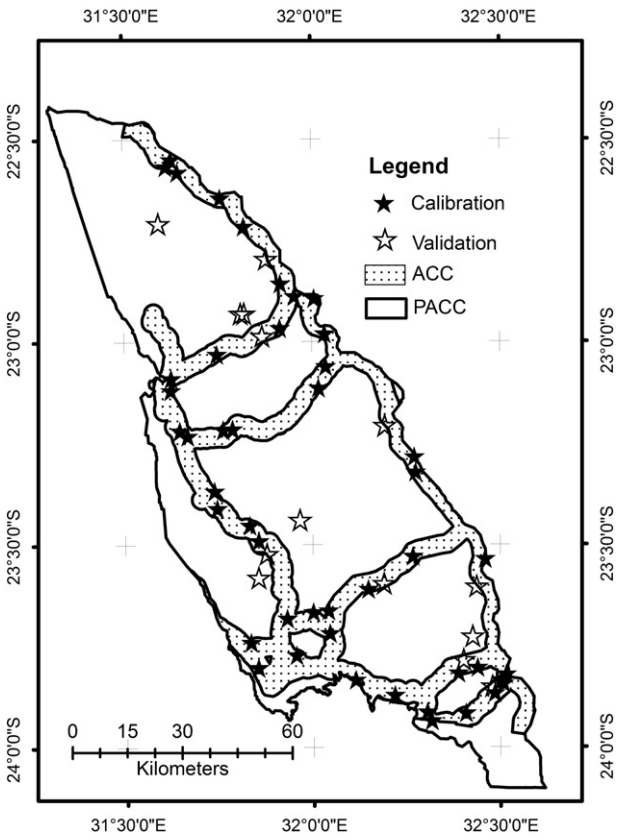


Fig. 3. Accessible and poorly accessible strata and the location of sampling clusters for calibration and validation of spatial prediction models.

the inter quartile range (IQR) for the quantitative covariables and the proportion in which each mapping unit occur in ACC and PACC areas for the categorical covariables. It would have been instructive to do this comparison per-stratum; however, this requires an adequate number of grid cells in each stratum for both ACC and PACC areas. In the present study this was not possible because of the small area of some combinations, e.g., there were only 34 grid cells in the ACC area of the CMR stratum.

Quantitative explanatory variables showed differences in mean between ACC and PACC areas below 10% with exception of elevation (20%). The difference in IQR was less than 6%, with exception of dry season NDVI and precipitation (about 20%) (Table 2). All mapping units of geology (Table 3) and landscape (Table 4) occur in both ACC and PACC areas, however in different proportions. Overall, ACC areas present ecological conditions that do occur in PACC areas. Therefore, we consider the similarity of the ACC and PACC areas to be adequate.

4.2.5. Primary data collection and laboratory analysis

The only legacy soil observations in the LNP are from a 1969 irrigation suitability survey of the extreme SE of the park, with poor georeference and no analytical data. Therefore the sampling design did not take these into account, and started from the "no previous data" situation.

Accessibility and wildlife hazard were major constraints to a random or regular sampling design. Therefore a stratified, clustered random sampling design was applied, which provides a statistical valid sample with high operational efficiency (De Gruijter et al., 2006). The LNP was stratified by landscape units (Stalmans et al., 2004), this being an integrative factor of soil genesis, and so is expected to capture a large part of the SOC variation. The number of clusters per stratum, i.e., landscape unit, was proportional to the stratum size. Sixty clusters were planned, 46 for model calibration in accessible areas and 14 for validation across LNP. Both sets were collected in the same field campaign. Cluster centers were positioned randomly within each stratum. Each cluster was composed out of two orthogonal transects of 720 and 360 m length crossing at their midpoints with a total of 7 sampling points units 180 m apart (Fig. 4). In order to capture the maximum variation the longer transect was oriented along the aspect as determined at the midpoint. At each sample point, five sub-samples from the four corners of a 90×90 m support area plus the center were mixed thoroughly into a composite sample. Sub-samples were from a field-identified A-horizon. The thicknesses of the A-horizons were recorded during the campaign.

In order to minimize the costs of laboratory analysis, all samples were analyzed using NIR spectrometry. A third (104) of the calibration and a quarter (25) of the validation samples were also analyzed using wet-chemical analysis. A PLSR model was developed to relate sample NIR spectra to SOC concentration (Brown et al., 2006; Shepherd and Walsh, 2002). All samples were analyzed in the soil laboratory of the Eduardo Mondlane University, Maputo, following a standard Walkley–Black method for SOC as described by van Reeuwijk (2002). Laboratory quality was assessed by submitting 20% of the samples in duple. All samples were scanned in a NIR spectrometer (Bruker FR-NIR Multi Purpose Analyzer, from Bruker optic GmbH, Ettlingen, Germany).

A PLSR calibration model relating SOC to NIR spectra for the 104 calibration laboratory-analyzed samples was built as described in Cambule et al. (2012). The model was validated by the 25 validation samples. To have a consistent basis for modeling, predicted PLSR SOC was used for further analysis even for those observations with laboratory data. The PLSR predicted SOC was then used as explanatory variable for the "soil" (s) factor. Following the sampling plan, a total of 410 samples from 59 clusters were collected of which 45 calibration and 14 validation (8 in PACC and 6 in ACC).

Laboratory results showed topsoil SOC contents ranging from 0.0% to 2.7% with a mean of 0.9%. The RMSE of the duplicate samples was 0.13% SOC which is in the normal range of variability of the Walkley and Black methodology (Chatterjee et al., 2009). The PLSR model explained 83.7% of the variation in SOC, with a RMSE of 0.32% using cross validation and 0.33% using true validation. The mean of validation residuals is almost zero, i.e., there is no bias, but extremes values are about 0.5% and as high as first quartile of PLSR-predicted SOC. The detailed results are reported separately by Cambule et al. (2012). The calibrated (and validated) model showed it tends to under-predict SOC contents above 1.5–1.8%, but the proportion of under-estimated samples was small and similar in both the wet laboratory sample sets (7%) used to build the model and for the all predicted samples (6%) (Table 5, Fig. 5).

4.2.6. Development of the spatial prediction model

The spatial model was developed on the base of explanatory variables that best explain SOC variation, for which appropriate spatial models were selected. This was then followed by spatial structure (within- and between-cluster) analysis. The main steps are described below:

4.2.6.1. SOC explained variation by explanatory variables. To assess the proportion of SOC variation explained by the continuous explanatory variables, pixel values of each explanatory variable layer at sampling points were extracted and regressed against SOC; the regression model was evaluated by ANOVA of the model compared to a null model, and by visual inspection of regression diagnostic plots (Fox,

Table 2
Summary statistics of the explanatory variables in accessible and poorly accessible area.

Variable	Unit	area	1st Qu.	Mean	3rd Qu.	IQR
Elevation	m	ACC	135	205	272	137
		PACC	188	254	317	129
Flow accumulation	no. of pixels	ACC	0	4	3	3
		PACC	0	4	3	3
NDVI wet season (1% trimmed)	–	ACC	0.28	0.34	0.40	0.12
		PACC	0.31	0.37	0.44	0.13
NDVI dry season, (1% trimmed)	–	ACC	0.07	0.12	0.16	0.09
		PACC	0.06	0.11	0.17	0.11
Annual precipitation	mm	ACC	431	459	494	63
		PACC	438	463	488	50

Table 3
Proportion of each geological unit (%) in accessible and poorly accessible areas.

Geology unit	Code	ACC	PACC
Sandstone	TeZ	39.1	29.8
Limestone	TeAul	9.5	6.8
Fluvial terrace, gravel and sand	Qt	1.5	0.3
Fluvial floodplain, clayey sand	Qps	0.4	0.8
Aeolian sand	Qe	34.7	51.7
Alluvium sand, silt, gravel	Qa	8.3	2.4
Dacite and trachydacite	JrUt	0.1	0.1
Rhyolite	JrUr	5.6	7.0
Basalt	JrSba	0.9	1.0

Table 4
Proportion of each landscape unit (%) in accessible and poorly accessible areas.

Landscape unit	Code	ACC	PACC
Limpopo Levubu Floodplains	LLF	7.0	0.9
Combretum/Mopane Rugged Veld	CMR	5.8	7.0
Nwambia Sandveld	NS	26.6	49.6
Pumbe Sandveld	PS	6.1	1.1
Salvadora angustifolia floodplains	SAF	16.0	2.5
Mopane Shrubveld on Calcrete	MSC	38.5	39.0

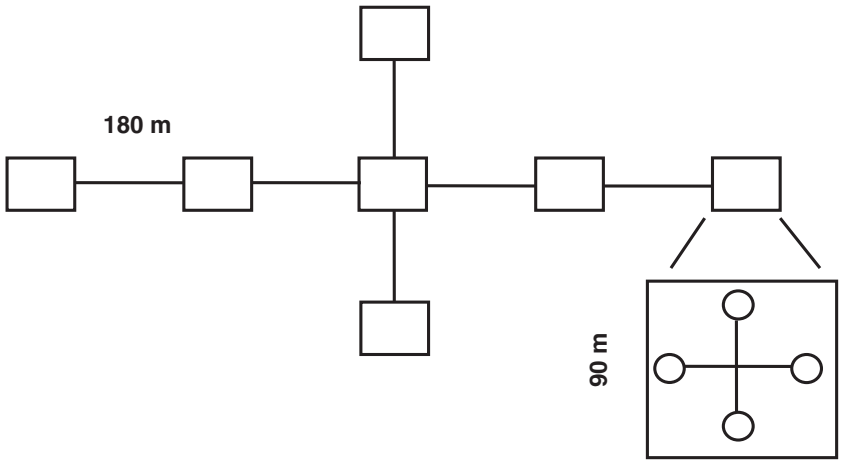


Fig. 4. The cluster (transect) design followed during field sampling, also showing the details of the support area for composite sampling at each sampling sub-station.

1997). The proportion of SOC variation explained by the categorical explanatory variables (clusters, geology and landscape) was evaluated by means of ANOVA of linear models of SOC as a function of each categorical variable. Regression adjusted goodness-of-fit was used to select explanatory covariables for model building.

The SOC variation explained by each of the explanatory variables is shown in Table 6. The soil factor (clusters) explains most SOC variation (71.1%), followed by geology (26.9%). Unfortunately all other single explanatory variables did not explain substantial amount of SOC variation. The landscape, here taken as an integrated explanatory covariable, did explain a substantial amount (39.4%).

Elevation explains little of SOC variation, suggesting height differences (maximum height differences are about 220 m) are not sufficient enough to result in either pronounced temperature differences or elevation differences that could be reflected in steep slopes. This is also corroborated by the consistently low *flow accumulation* across the study area, which indicates that contributing area is too small for water to accumulate.

Similarly, mean annual precipitation did not explain substantial amount of SOC variation (<1%), perhaps because absolute differences in the study area are not large enough to affect SOC. This is also corroborated by the weak regional trend as demonstrated by the explained SOC variation on the coordinates, despite visible variation in greenness, also detected by NDVI. Perhaps the greenness may be explained by below-ground water movement as precipitation easily infiltrates the extensive sand soils.

Wet season NDVI explains a little more than double the SOC variation as the *dry season NDVI*. However the amount explained in both seasons is low. This may be a result of the combined effects from elevation, flow accumulation and mean annual precipitation as all have an effect on water availability across the study area.

Lithology explains about 27% of SOC variation, the best single covariable (Fig. 6). This may be because the soil over most of the area is residual. Rhyolite and aeolian sand have consistently high and low median SOC, respectively; however the rhyolite unit includes only one sampling cluster. Topsoil in this unit was consistently dark and pebble-rich. Other units do not differ substantially.

Table 5
Summary statistics of the PLSR SOC (%) prediction (all samples) and SOC (%) cluster averages.

SOC (%)	Min	1stQ	Med	Mean	3rdQ	Max
PLSR predicted	0.00	0.61	0. 87	0.92	1.19	2.68
cluster mean	0.21	0.61	0.89	0.93	1.10	1.91

The *clusters* predicted SOC, the *soil* factor, explains SOC variation the most (71.1%). Although about 30% of SOC variation is still within the clusters, the clusters' size and the sampling strategy were effective in capturing considerable SOC variation across the LNP.

The *landscape* explained about 40% of SOC variation (Fig. 6); by design it captures both lithology and any vegetation effect. Regression coefficients show CMR landscape unit contributing more to the model. This may be due to its proximity to the Lebombo mountain chain, where rainfall is suspected to be a little higher (Stalmans et al., 2004). This is followed by MSC, SAF and LLF, located along the Singwedzi and Limpopo Rivers under similar surface water regime. The sandvelds (PS and NS) have the least SOC %, perhaps due to sandier soil textures and lower water-holding capacities.

4.2.6.2. Selection of prediction model. Thus there were three possibilities for spatial prediction: (1) ordinary kriging (OK), considering only the known observations (factor *s*); (2) linear regression models (LM) from environmental predictors; (3) kriging with external drift (KED), equivalent to regression kriging (RK) (Hengl et al., 2007), considering the regression model and the spatial correlation of its residuals. In the case where there is demonstrated spatial structure in regression model residuals, the LM method can be replaced by a generalized linear model (GLM). Based on the above, predicted SOC in the clusters and geology represent the *soil* and *parent material* factors in the *scorpan-SSPFe* model, while landscape is an integrated factor, representing all seven *scorpan* factors. Lithology explains less variation in SOC than landscape, which apparently incorporates the lithological information, so it was not used. Separate spatial models were considered, one using the soil factor (OK) and the other using the landscape integrated factor with residuals (KED), as well as the landscape regression model, which has the advantage over kriging methods when spatial structure is weak or have limited range.

4.2.6.3. Variogram analysis. Residuals from selected models show the unexplained variation in SOC. These, as well as the original values of SOC, were examined for local spatial autocorrelation using empirical variograms (Goovaerts, 1999). If structure was evident, models of spatial dependence (both original values and model residuals) were fit to the empirical variogram using weighted least square (WLS) in gstat (Pebesma, 2004). Anisotropy was evaluated visually with a variogram map. In order to minimize irregularities (due to small sampling size and to avoid arbitrary decisions on variogram bin width) and therefore improve the variogram fitting within the range of the variogram model, a residual maximum likelihood (REML) (Marchant and Lark, 2007) was applied directly to the variogram cloud from WLS fit, using gstat. We fitted the ordinary and residual variogram

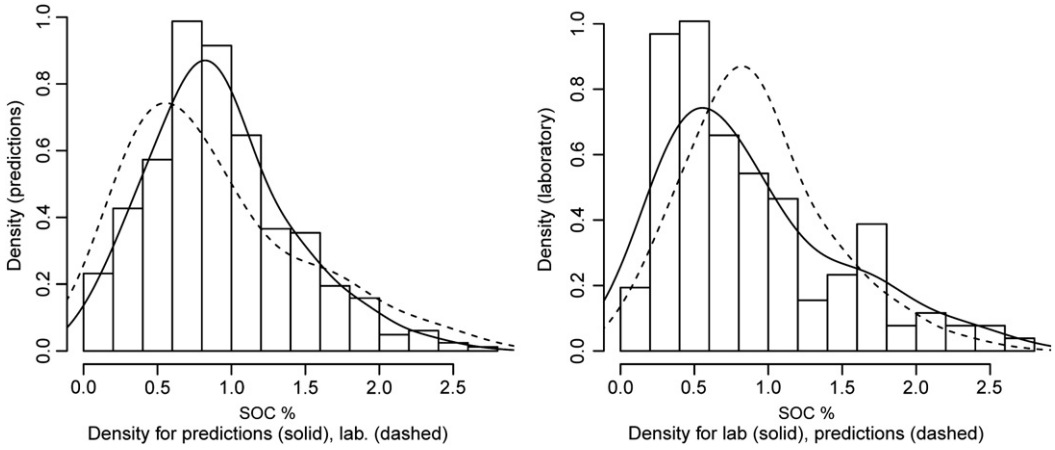


Fig. 5. PLSR predicted SOC concentrations for all samples (left) relative to the laboratory samples only (right).

with spherical models using all calibration samples. The variograms show autocorrelation to ranges of about 16.0 km for SOC and 4.0 km for the residuals from the landscape linear model (Table 7 and Fig. 7). The nugget of REML-fitted variograms is about the same but the residual variogram sill is much lower, about half. The effect of landscape is clear in the shorter range and lower partial sill. This is consistent with the linear regression model with landscape unit as predictor. Both nuggets are higher than RMSE of laboratory analysis on duplicates (about 0.13% squared), so that the laboratory uncertainty is included in the nugget. Despite the relatively higher nuggets, the fitted variograms show the nugget-to-sill ratio of about 22% (ordinary) and 33% (residual), indicating that the short range variability shares some autocorrelation variance, though not by much (Gringarten and Deutch, 2001; Mapa and Kumaragamage, 1996).

While the obtained variogram ranges could be used to design a second-phase sampling, the residual variogram range should enable SOC predictions from ACC through into PACC using explanatory covariables of environmental predictors derived from secondary data. However, in most of the center-southern part of the study area, the obtained residual variogram range is limited relative to the extent of PACC areas, which extend up to about 50 Km away from ACC areas.

4.2.6.4. Within-cluster soc spatial autocorrelation. In order to assess within-cluster spatial autocorrelation, an experimental variogram spanning the cluster range (720 m) was calculated, plotted and visually inspected in order to determine the practical support area, within which SOC variation is controlled by very short-range factors (i.e., within a cluster) and therefore should be ignored when mapping.

The experimental variograms along with a fitted pentaspherical variogram model are shown in Fig. 8. It reveals good spatial structure, with spatial dependence to about 500 m. The spatial dependence at short range was strong: nugget variance was fit to zero, but then raised to the known uncertainty of the laboratory analysis. The originally-modeled zero nugget shows the effect of composite sampling on a 90 m support. Thus most differences in SOC concentration are explained by local factors at scales between cluster range (720 m) and bulk sample range (90 m). The linear model predicting SOC by sampling clusters ($R^2 = 0.71$) has a residual mean square of 0.073%. This is the variance not explained by the clusters and should correspond to the sill of the within-station variograms, which were estimated at about 0.06 (% SOC)^2. This also means that the nugget found in the long-range variogram represents a support of at least a cluster and that the clusters can be represented by their ordinary (unweighted) averages. Therefore spatial models as well as the remainder of the analyses were based on cluster averages. The averaging generally increased the proportion of SOC explained by the different explanatory variables (see Table 6).

Table 6
Explained SOC (%) variation (adjusted R^2) by the explanatory variables.

Variable	All points	Clusters
Elevation	2.6	1.3
Flow accumulation	1.4	1.3
NDVI wet season	8.0	20.9
NDVI dry season	0.1	1.6
Annual precipitation	0.7	−0.7
Geology	26.9	33.2
Landscape	39.4	46.3
SOC (clusters)	71.1	71.5
Coordinates	7.7	6.7

4.2.6.5. Variogram analysis (clusters). Experimental variograms based on calibration cluster averages were difficult to model, due to the low number of point-pairs in each bin. Starting from the parameters of the fitted variograms based on all calibration points, spherical models were fitted (Fig. 9, Table 7), resulting in slightly longer ranges, much lower structural sills and effectively zero nugget. These are all consistent with the averaging effect. The REML fit did not improve the variogram due to the high variance at smaller lag, pulling the REML variogram fit up and introducing an unrealistic nugget. Therefore the WLS fit was retained for mapping. The obtained variogram ranges increased by about 12% (ordinary) and 26% (residual), which potentially improves the ability for predictions in PACC from the ACC areas. This is despite the reduction in the partial sill. Cluster averaging will also be economical in future sampling as the within cluster variation will be ignored.

4.2.7. Application of the model in accessible area

SOC was predicted across ACC areas from the calibration observations, by applying the selected OK, KED and LM spatial models. Internal prediction quality was assessed by kriging prediction standard deviation (Goovaerts, 1999; McBratney et al., 2000). Since the within-cluster analysis showed that SOC in a cluster could be represented by the cluster average, prediction was performed by punctual kriging over 1 × 1 km grid as a support area, assuming that the average of a 1 × 1 km cell would to be similar to that of the 720 × 720 m support area for which spatial structure had a little longer than half the cluster length. The kriging prediction variance is thus realistic: "punctual" in this case means on a cluster-size support.

The summary statistics of OK prediction (Table 8) shows OK with narrower range (1.27%) and the KED with the wider range (1.97%) and LM in between (1.44%). The same is observed for the mean SOC

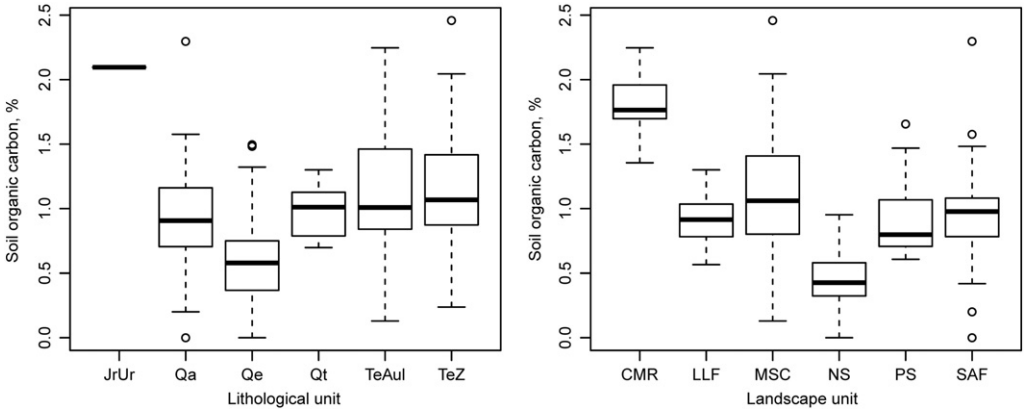


Fig. 6. Boxplot of SOC as a function of geology (left) and landscape (right) as calculated based on calibration clusters.

predictions by the three models. The OK prediction map clearly shows the effect of low sampling density. Areas further away from sampling locations are predicted as a spatially-weighted average (0.93%) as there is no information on spatial variation structure. Predictions by KED much resemble the landscape map (Fig. 10).

Kriging prediction standard deviation (KPSD) is lower (Table 8) for the LM and higher for OK, with KED (Fig. 10) in between the two, all with low IQR, suggesting that kriging prediction SD is a rather precise measure. However, the mean KPSD is about half (OK), a third (KED) and 5% (LM) the median and as high as the minimum predicted SOC (OK, KED) but only about 11% (LM). This suggests prediction quality by internal measure is better for LM and poor for OK and KED.

4.2.8. Model validation in accessible areas

We used (1) the leave-one-out cross-validation (LOOCV) (Goovaerts, 1999; McBratney et al., 2000) as an internal measure of model fitness and (2) true validation with an independent sample set. Since there is no cross-validation concept for the linear regression model, it was not performed for the landscape model. LOOCV RMSE for both OK and KED are low, about 0.02% and mean prediction residuals are about 0.00% which indicate the models are unbiased. However largest residuals (±0.70%, symmetrical) are a little higher than the minimum prediction by the models. OK IQR of residuals is twice that of KED (0.29%). Independent validation results (Table 8) show KED and LM performing similarly (RMSE about 0.43%) and better than OK. Neither method is satisfactory given the fact that RMSE is a substantial proportion (about half) of the median from the model predictions. All methods are also biased (under-predictions). True validation RMSE is about double LOOCV RMSE in both cases, which is consistent with expectations. True validation RMSE and mean kriging SD are almost identical, and therefore kriging standard SD is a reasonable estimate of the actual error. At this point we had to decide if the model was sufficiently accurate to proceed to the next step of the methodology. Given the generally low values of the target variable in the LNP (maximum 2.68%, median 0.87%, see Table 5), and the result that the validation RMSE is about half the median, we are forced to admit that the model is of limited utility. It does show some landscape differences and accounts for spatial structure near the observation points, but even at a 1×1 km block gives predictions that are only about twice as precise as taking the area-weighted average or median observed value over the whole area. Nonetheless, we continue with the method to illustrate the remaining steps.

To put our results in context, we compared them with other studies reported in the literature. Mueller and Pierce (2003) studied the effect of sampling scale on accuracy of SOC predictions of top 20 cm across an area of 12.5 ha in Michigan, USA, and showed that despite the finer grids followed and a wide SOC range (0.2–0.29%), the best RMSEP obtained was 0.28–0.30%, about 30% of the SOC observed mean. Robinson and Metternich (Robinson and Metternicht, 2006) compared the accuracy of OK, lognormal OK, IDW and splines for interpolation of soil proprieties in 60 ha, south west Australia. The best OK RMSEP was 1.43% and about 30% of the average observed OM and 35% of the mean predictions. Chai et al. (2008) compared the performance of empirical best linear unbiased predictor (E-BLUP) with REML with that of RK for prediction of SOM in the presence of different external drifts across an area of 933 km^2 in China. The best RMSEP obtained was 0.38% (RK), which represented about 29% of mean observed data. Grimm et al. (2008) predicted the spatial distribution of SOC following the DSM approach in Panamá for different soil depths in a 1500 ha area. The best RMSEP obtained was 1.72% for the top 10 cm soil depth, corresponding to about 34% of the observed SOC data.

The above results show that the proportion of RMSEP to mean predictions or mean observed SOC in our study is poor relative to other studies.

Comparative studies closer to the study area or in Africa, in general, are few but show different results. For example Stoorvogel et al. (2009) used a classification tree approach combined with existing knowledge from literature and a small data set to map top soil SOC content for a data-poor environment in a 1030 km^2 of the Senegalese peanut basin, with a RMSEP of about 0.17%, representing about 40% of the mean observed SOC.

As another example, Mora-Vallejo et al. (2008) tested whether DSM is suited for exploratory or reconnaissance soil survey of SOC. Their results in a 13,500 km^2 area in southeast Kenya show SOC RMSEP of about 0.2%, corresponding to about 25% of both the mean predictions by regression kriging and mean observed SOC data. While these results are consistent with those from elsewhere, Schloeder et al. (2001) found rather more accurate results when they compared different interpolation methods (OK, IDW and thin-plate with and without tensions) for organic matter (OM) prediction across a 70×20 km area in the Omo basin, south-west Ethiopia. The best MSE was 0.08%, i.e., RMSE = 0.28%, for OK, which represented about 20% of the mean observed data. Regardless of the different results, all are better than the one found in our study.

Table 7
REML fitted variogram parameters.

Variogram type	Nugget [m^2]	Partial sill [m^2]	Range (m)
Ordinary, points	0.065	0.236	15,986
Residual, points	0.057	0.115	3908
Ordinary (within cluster)	0.016	0.069	528
Ordinary, clusters	0.000	0.225	18,126
Residual, clusters	0.008	0.100	5278

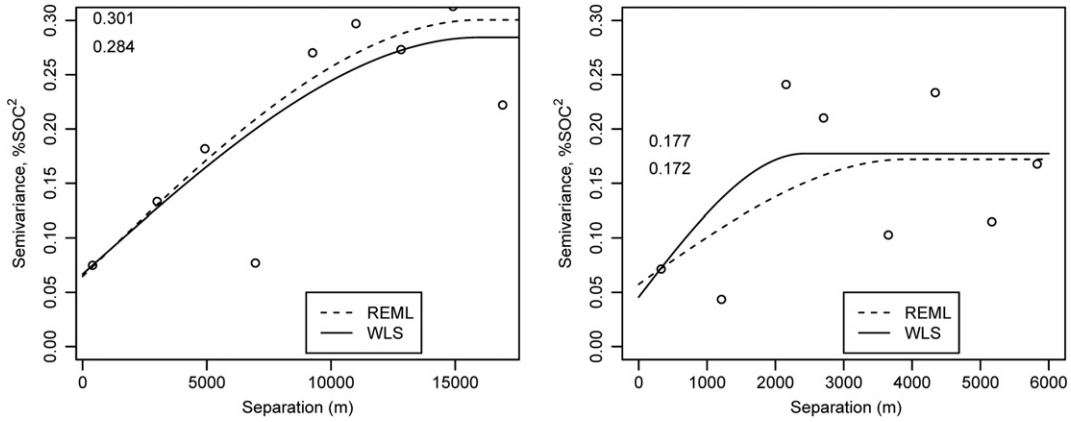


Fig. 7. Ordinary and residual (landscape as covariable) variograms, all calibration points.

4.2.9. Model application in poorly-accessible areas

SOC was predicted here using the same models (OK, KED and LM) for the same support area, also for the same reasons as in the ACC area. We also assessed the internal prediction quality by kriging prediction standard deviation (KPSD) (Goovaerts, 1999; McBratney et al., 2000). The summary statistics of OK prediction (Table 8) shows OK with narrower range (1.22% SOC) and the KED with the wider range (1.77%) and LM in between (1.44%). The same is observed for the mean SOC predictions by the three models. The OK prediction map (Fig. 10) clearly shows the effect of low sampling density. Areas further away from sampling locations are predicted as a spatially-weighted average (0.93%) as there is no information on the structure of spatial variation. KPSD is lower (Table 8) for the LM and higher for OK, with KED in between the two, all with low IQR, suggesting that Kriging prediction SD is a rather precise measure. However, the mean KPSD is a little less than half (OK and KED) and 5% (LM) the median and as high as the minimum predicted SOC (OK and KED) but only about 11% (LM). This suggests prediction quality by internal measure is better for LM and poor for OK and KED.

Examples of the predictions into PACC based on models built from the ACC area, as proposed here, are not available in the literature. However, the obtained prediction results are within the range of those obtained (and discussed) for ACC areas.

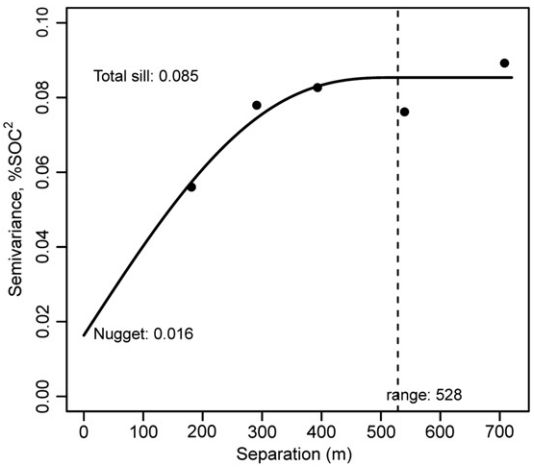


Fig. 8. Ordinary kriging experimental variogram of SOC up to a cut-off of cluster length (720 m), based on all calibration points.

4.2.10. Model validation in poorly-accessible areas

We performed the true validation with an independent sample set as planned. Validation results (Table 8) show, surprisingly, all models with RMSEP lower than the one for ACC areas. Further KED and LM performed similarly (RMSEP about 0.31% SOC) and better than OK. However, all models performed poorly, given the fact that RMSEP are about 4/10 of the SOC prediction median, so effective mapping is not possible with the present sampling density. All models were also biased (under-prediction); with LM similar to KED and both a little better than OK. Mean KPSD was a little higher that validation RMSE (OK and KED) so KPSD is a reasonable estimate of actual error.

Similar to predictions, validation results both the RMSEP (true validation) and KPSD (with exception to LM) found in the present study are about the minimum predictions and as high as double the mean predictions, which confirms our poor results.

4.2.11. Relative performance of prediction model in PACC areas

When comparing validation RMSE between ACC and PACC, the three models performed better in PACC than in ACC areas by about 28% (OK) and 26% (KED) and 31% (LM). This is likely due to the different test set sizes (larger for the PACC). Thus the extrapolation into non-sampled PACC areas seems justified for KED, although predictions are largely determined by landscape away from sampling points in accessible area. LM performed relatively best and does not suffer from the requirement of spatial autocorrelation for interpolation into PACC areas.

Despite poor predictions by both models, the methodology is promising because predictions into PACC areas are close to predictions made in ACC areas. The poor model predictions result from cumulative error effects brought about along the different steps, namely laboratory analysis, PLSR calibration, model building, and spatial predictions. The weak SOC variation explained by most of the explanatory variable here selected may also have contributed to the poor model predictions, although many authors have demonstrated the role of secondary data to improve prediction of SOC (Mueller and Pierce, 2003; Simbahan et al., 2006). Nevertheless, one of the strong points of our results lies on the spatial models' range, which allows interpolation into PACC (about 5 km KED and 18 km for OK). Despite a longer range for OK, the low sampling density is a limiting factor as information on spatial structure is absent in the PACC. By contrast, KED allows mapping based on the covariable but the range of spatial structure is rather limited. LM can take over beyond the OK and KED range, into the PACC areas.

The spatial models building was in this case made possible based on the integrated soil-forming factor (landscape) and the clusters, which explained most SOC variation. The OK results could be used

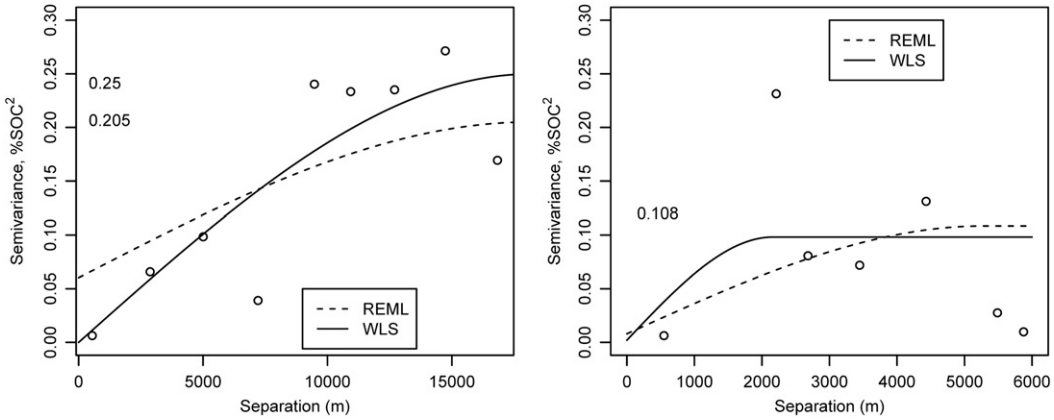


Fig. 9. WLS and REML fitted ordinary (left) and REML fitted residual (right) variograms drawn based on calibration clusters (accessible areas).

Table 8
Summary statistics of SOC (%) spatial prediction, Kriging prediction standard deviation (PSD) and model independent validation.

Model	Prediction				KPSD		Independent validation		
	Min	Median	Mean	Max	Mean	IQR	Mean	RMSE	Bias
OK_ACC	0.42	0.90	0.91	1.69	0.40	0.08	−0.03	0.50	−0.02
OK_PACC	0.46	0.90	0.89	1.68	0.44	0.04	0.09	0.36	0.09
KED_ACC	0.35	1.01	1.22	2.32	0.37	0.03	−0.01	0.42	−0.01
KED_PACC	0.40	0.97	1.06	2.17	0.37	0.03	0.06	0.31	0.06
LM_ACC	0.46	0.93	0.87	1.90	0.05	0.02	−0.03	0.45	−0.03
LM_PACC	0.46	0.92	0.84	1.90	0.05	0.01	0.07	0.31	0.07

to aid future sampling to improve prediction since the within cluster spatial structure is rather weak and could be bulked. Therefore future sampling could be based on the obtained structural range.

5. Conclusions

The chosen test case turned out to be a difficult one. The range of SOC concentrations was narrow, weakly-dependent on covariables, and exhibited most of its spatial structure within the support of a cluster. We conclude that SOC concentration in the study area varies mostly by local factors, probably current and past vegetation and animal activity (including termites), not captured by any covariable. The proposed method did work as planned in the sense that the models did as well in poorly-accessible as in accessible areas. The use of a previous integrative survey (Stalmans et al., 2004) was quite helpful in this case and was able to substitute for a large number of coverages. Such a survey substitutes for multiple factors in the *scorpan-SPPfe* framework.

Despite the somewhat disappointing performance in this test case, we feel that the proposed methodology as such was appropriate, certainly as the first stage in a survey in areas with difficult access. At this point we know the spatial structure and relation of target variable with covariables, and we have evidence that the model structure in poorly-accessible areas is likely to be similar to that in accessible areas. Thus if we are not satisfied with the predictions mostly as landscape spatial averages, we can plan a sampling campaign by optimizing the KED variance to a realistic target (set here by the PLSR precision) as proposed by Brus and Heuvelink (2007). We also know, in this case, to sample on a 1 km support and not try to map variation in smaller areas. All this prepares us for the most efficient approach possible in the difficult circumstances of a survey in poorly-accessible areas.

Acknowledgments

The authors thankfully acknowledge support from The LNP Management, The International Research and Education Fund (INREF) of the Wageningen University through the “Competing Claims on Natural Resources Programme” for funding and from The International Institute for Geo-information Science and

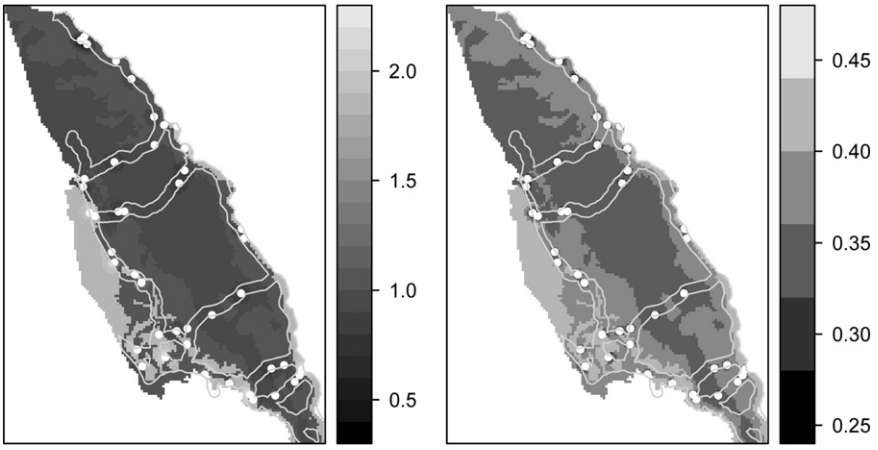


Fig. 10. SOC (%) prediction maps by KED using landscape as a covariable (left) and its kriging prediction standard deviation (right).

Earth Observation of the University of Twente for the scientific support.

References

Bell, J.C., Grigal, D.F., Bates, P.C., 2000. A soil-terrain model for estimating spatial patterns of soil organic carbon. In: Wilson, J.P., Gallant, J. (Eds.), Terrain Analysis: Principles and Applications. Wiley & Sons, New York, pp. 295–310.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132 (3–4), 273–290.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1–2), 86–95.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. Geoderma 183–184, 41–48.

Cécillon, L., Barthès, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., Brun, J.J., 2009. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). European Journal of Soil Science 60 (5), 770–784.

Cenacarta, 1997. Carta de Uso e Cobertura da Terra. Ministério da Agricultura, Maputo.

Chai, X., Shen, C., Yuan, X., Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma 148 (2), 159–166.

Chatterjee, A., Lal, R., Wielopolski, L., Martin, M.Z., Ebinger, M.H., 2009. Evaluation of different soil carbon determination methods. Critical Reviews in Plant Sciences 28 (3), 164–178.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.

Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling and Software 17 (3), 295–311.

Fox, J., 1997. Applied Regression, Linear Models, and Related Methods. Sage, Newbury Park.

Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling soil–landscape and ecosystem properties using terrain attributes. Soil Science Society of America Journal 64, 2046–2056.

Giller, K.E., Leeuwis, C., Andersson, J.A., Andriesse, W., Brouwer, A., Frost, P., Hebinck, P., Heitkonig, I., van Ittersum, M.K., Koning, N., Ruben, R., Slingerland, M., Udo, H., Veldkamp, T., van de Vijver, C., van Wijk, M.T., Windmeijer, P., 2008. Competing claims on natural resources: what role for science? Ecology and Society 13 (2), 18.

Gisladottir, G., Stocking, M., 2005. Land degradation control and its global environmental benefits. Land Degradation & Development 16, 99–112.

Gitelson, A.A., Merzlyak, M.N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. Advances in Space Research 22 (5), 689–692.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89 (1–2), 1–45.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. Geoderma 146 (1–2), 102–113.

Gringarten, E., Deutch, C.V., 2001. Teacher's aid variogram interpretation and modeling. Mathematical Geology 33 (4), 507–534.

Hengl, T., 2006. Finding the right pixel size. Computers & Geosciences 32 (9), 1283–1298.

Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: from equations to case studies. Computers & Geosciences 33 (10), 1301–1315.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100 (3–4), 269–301.

Hijmans, R.J., Cameron, S., Parra, J., 2011. WorldClim — Global Climate Data, Berkeley, CA.

Jenny, H., 1980. The soil resource: origin and behavior. Ecological Studies, 37. Springer-Verlag, New York.

Lark, R.M., 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105 (1–2), 49–80.

Li, Y., 2010. Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? Geoderma 159 (1–2), 63–75.

Manninen, T., Eerola, T., Makitie, H., Vuori, S., Luttinen, A., Senvano, A., Manhica, V., 2008. The Karoo volcanic rocks and related intrusions in southern and central Mozambique. Geological Survey of Finland, Special Paper 48, 211–250.

Mapa, R.B., Kumaragamage, D., 1996. Variability of soil properties in a tropical Alfisol used for shifting cultivation. Soil Technology 9 (3), 187–197.

Marchant, B.P., Lark, R.M., 2007. Robust estimation of the variogram by residual maximum likelihood. Geoderma 140 (1–2), 62–72.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97 (3–4), 293–327.

McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McKenzie, N.J., Gessler, P.E., Ryan, P.J., O'Connell, D.A., 2000. The role of terrain analysis in soil mapping. In: Wilson, J.P., Gallant, J. (Eds.), Terrain Analysis: Principles and Applications. Wiley & Sons, New York, pp. 245–265.

Miklos, M., Short, M.G., McBratney, A.B., Minasny, B., 2010. Mapping and comparing the distribution of soil carbon under cropping and grazing management practices in Narrabri, north-west New South Wales. Australian Journal of Soil Research 48 (3), 248–257.

Milne, E., Adamat, R.A., Batjes, N.H., Bernoux, M., Bhattacharyya, T., Cerri, C.C., Cerri, C.E.P., Coleman, K., Easter, M., Falloon, P., Feller, C., Gicheru, P., Kamoni, P., Killian, K., Pal, D.K., Paustian, K., Powlson, D.S., Rawajfih, Z., Sessay, M., Williams, S., Wokabi, S., 2007. National and sub-national assessments of soil organic carbon stocks and changes: the GEFSOC modelling system. Agriculture, Ecosystems and Environment 122 (1), 3–12.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32 (9), 1378–1388.

Ministerio do Turismo, 2003. Limpopo National Park: Management and Development Plan. Ministério do Turismo, Maputo.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J., Heuvelink, G.B.M., 2008. Small scale digital soil mapping in southeastern Kenya. Catena 76 (1), 44–53.

Mueller, T.G., Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. Soil Science Society of America Journal 67 (1), 258–267.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30 (7), 683–691.

Phachomphon, K., Dlamini, P., Chaplot, V., 2010. Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. Geoderma 155 (3–4), 372–380.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. The R Foundation for Statistical Computing, Vienna.

Robinson, T.P., Metternicht, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Computers and Electronics in Agriculture 50 (2), 97–108.

Rutten, R., Makitie, H., Vuori, S., Marques, J.M., 2008. Sedimentary rocks of the Mapai Formation in the Massingir–Mapai region, Gaza province, Mozambique. Geological Survey of Finland, Special Paper 48, 251–262.

Schloeder, C.A., Zimmerman, N.E., Jacobs, M.J., 2001. Comparison of methods for interpolating soil properties using limited data. Soil Science Society of America Journal 65 (2), 470–479.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27, 171.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66 (3), 988–998.

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., Haddix, M.L., 2006. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. Geoderma 132 (3–4), 471–489.

Stalmans, M., Gertenbach, W.P.D., Carvalho-Serfontein, F., 2004. Plant communities and landscapes of the Parque nacional do Limpopo, Mozambique. Koedoe 47 (2), 61–81.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M., de Bruin, S., 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149 (1–2), 161–170.

Ungar, F., Staffilani, F., Tarocco, P., 2010. Assessing and mapping topsoil organic carbon stock at regional scale: a Scorpan kriging approach conditional on soil map delineations and land use. Land Degradation & Development 21 (6), 565–581.

van Reeuwijk, L.P., 2002. Procedures for Soil Analysis. ISRIC Technical Paper 9.

Vasques, G.M., Grunwald, S., Sickman, J.O., Comerford, N.B., 2010. Upscaling of dynamic soil organic carbon pools in a north-central Florida watershed. Soil Science Society of America Journal 74 (3), 870–879.

Viscarra-Rossel, R.A., McBratney, A.B., 2008. Diffuse reflectance spectroscopy as a tool for digital soil mapping. In: Hartemink, A.A.M.L.M.-S. (Ed.), Digital Soil Mapping with Limited Data. Springer.

Webster, R., Welham, S.J., Potts, J.M., Oliver, M.A., 2006. Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. Computers & Geosciences 32 (9), 1320–1333.

Yemefack, M., Jetten, V.G., Rossiter, D.G., 2006. Developing a minimum data set for characterizing soil dynamics in shifting cultivation systems. Soil and Tillage Research 86 (1), 84–98.

Yoder, B.J., Waring, R.H., 1994. The normalized difference vegetation index of small Douglas-fir canopies with varying chlorophyll concentrations. Remote Sensing of Environment 49 (1), 81–91.