Geoderma 243–244 (2015) 214–227

Contents lists available at ScienceDirect

Geoderma

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

Do more detailed environmental covariates deliver more accurate soil maps?

A. Samuel-Rosa [a,b,c,*], G.B.M. Heuvelink [c], G.M. Vasques [d], L.H.C. Anjos [b]

[a] CAPES Grant Holder — Process BEX 11677/13-9, CAPES Foundation, Ministry of Education of Brazil, Brasília, DF 70040-020, Brazil
[b] Graduate School in Agronomy-Soil Science, UFRRJ — Institute of Agronomy, BR-465, km 7, Seropédica, RJ 23890-000, Brazil
[c] ISRIC — World Soil Information, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands
[d] Embrapa Soils, Rua Jardim Botâânico 1024, Jardim Botânico, Rio de Janeiro, RJ 22460-000, Brazil

ARTICLE INFO

Article history:
Received 7 August 2014
Received in revised form 18 November 2014
Accepted 19 December 2014
Available online 21 January 2015

Keywords:
Digital soil mapping
Linear mixed model
Auxiliary information
Variable selection
Model accuracy
Soil mapping cost

ABSTRACT

In this study we evaluated whether investing in more spatially detailed environmental covariates improves the accuracy of digital soil maps. We used a case study from Southern Brazil to map clay content (CLAY), organic carbon content (SOC), and effective cation exchange capacity (ECEC) of the topsoil for a ~ 2000 ha area located on the edge of the plateau of the Paraná Sedimentary Basin. Five covariates, each with two levels of spatial detail were used: area-class soil maps, digital elevation models (DEM), geologic maps, land use maps, and satellite images. Thirty-two multiple linear regression models were calibrated for each soil property using all spatial detail combinations of the covariates. For each combination, stepwise regression was used to select predictor variables incorporated in the model. Model evaluation was done using the adjusted R-square of the regression. The baseline model, calibrated with the less detailed version of each covariate, and the best performing model were used to calibrate two linear mixed models for each soil property. Model parameters were estimated using restricted maximum likelihood. Spatial prediction was performed using the empirical best linear unbiased predictor. Validation of baseline and best performing linear multiple regression and linear mixed models was done using cross-validation. Results show that for CLAY the prediction accuracy did not considerably improve by using more detailed covariates. The amount of variance explained increased only ~2 percentage points (pp), less than that obtained by including the kriging step, which explained 4 pp. On the other hand, prediction of SOC and ECEC improved by ~13 pp when the baseline model was replaced by the best performing model. Overall, the increase in prediction performance was modest and may not outweigh the extra costs of using more detailed covariates. It may be more efficient to spend extra resources on collecting more soil observations, or increasing the detail of only those covariates that have the strongest improvement effect. In our case study, the latter would only work for SOC and ECEC, by investing in a more detailed land use map and possibly also a more detailed geologic map and DEM.

1. Introduction

Digital soil mapping relies on the use of statistical models to produce digital representations of spatial soil distribution using point soil observations and spatially exhaustive environmental covariates (McBratney et al., 2003; Scull et al., 2003; Florinsky, 2012). Three important weaknesses in the statistical soil distribution modelling approach can be pointed out. First, it requires sufficient and appropriately distributed point soil data within the area being mapped (Carré et al., 2007). Second, the model structure explores only the empirical relationship among environmental conditions and soil properties, being less comprehensive than soil-landscape process models (Grunwald, 2009). Last, the covariates are only approximations of the true environmental conditions that helped shape the soil. They serve only as proxies (surrogates) of the current environmental conditions, which in many cases are different from the past conditions under which pedogenesis took place (Heuvelink and Webster, 2001). In spite of these weaknesses, digital soil mapping has proven very successful in the past decades in producing soil property maps that capture the main patterns of soil spatial variation (Moore et al., 1993; McBratney et al., 2000; Grunwald, 2009).

More recently, there has been growing interest in understanding how the characteristics of the environmental covariates influence the success of digital soil mapping — this study contributes to this effort. It is commonly accepted that the more resources are spent on the construction of a covariate and the more spatial information it has, the more accurately it describes the environmental conditions (Hupy

* Corresponding author at: CAPES Grant Holder — Process BEX 11677/13-9, CAPES Foundation, Ministry of Education of Brazil, Brasília, DF 70040-020, Brazil.
E-mail addresses: alessandrosamuelrosa@gmail.com (A. Samuel-Rosa), gerard.heuvelink@wur.nl (G.B.M. Heuvelink), gustavo.vasques@embrapa.br (G.M. Vasques), lanjos@ufrrj.br (L.H.C. Anjos).

http://dx.doi.org/10.1016/j.geoderma.2014.12.017
0016-7061/

et al., 2004; Hengl et al., 2013). It is also generally believed that such *more detailed* covariates will be more valuable for digital soil mapping and lead to more accurate soil property predictions (Cavazzi et al., 2013; Maynard and Johnson, 2014). If these more detailed covariates convey more information and represent more adequately the environmental conditions — the drivers of soil forming processes —, then it is fair to expect that they improve the accuracy of the resulting soil maps. However, some studies have shown the contrary (Thompson et al., 2001; Eldeiry and Garcia, 2008; Kim et al., 2014). For example, the window size at which DEM derivatives are calculated can be more important than the spatial resolution of the DEM (Wood, 1996; Zhu et al., 2008; Behrens et al., 2010). The uncertainty about the added value of using more detailed covariates is of concern for those seeking to use resources efficiently, because using more detailed covariates generally increases soil mapping costs (Shi et al., 2012).

The objective of this study was to evaluate whether investing in more detailed environmental covariates improves the accuracy of digital soil maps. The main difference of our study to previous ones is that we use a rigorous statistical approach to assess the added value of using five more detailed covariates simultaneously. We used a case study in Brazil to compare the accuracy of digital maps of the clay content, organic carbon content and effective cation exchange capacity of the topsoil as obtained from regression kriging on the five covariates, whereby each covariate was evaluated on two levels of spatial detail.

2. Material and methods

2.1. Study area and soil data

The study area constitutes a small catchment (~2000 ha) located on the southern edge of the plateau of the Paraná Sedimentary Basin, Rio Grande do Sul, Brazil (Fig. 1). The climate is classified as Cfa (Köppen — subtropical humid without a dry season) with a mean annual temperature of 19.3 °C, and mean annual precipitation of 1708 mm, well distributed throughout the year (Maluf, 2000). Relief varies between plain (slope between 0 and 3%) and mountainous (slope between 45 and 100%), and elevations range between 140 and 475 m. Geology consists of basic, intermediate and acid igneous rocks (rhyolite-rhyodacite and andesite-basalt) of the Cretaceous period, consolidated sedimentary rocks (aeolian and fluvial sandstones) of the Triassic and Jurassic periods, and non-consolidated (fluvial and colluvial deposits) of the Quaternary period (Gasparetto et al., 1988; Maciel Filho, 1990; Sartori, 2009). Native semi-deciduous forests occupy more than half of the area, followed by native grassland used for animal husbandry, semi-deciduous shrubland, annual crop agriculture, forestry (Eucalyptus), urban areas, and artificial water bodies (Samuel-Rosa et al., 2011).

A dataset containing $n = 350$ point soil observations collected between 2004 and 2011 (Pedron et al., 2006; Samuel-Rosa et al., 2011; Miguel et al., 2011; Samuel- Rosa et al., 2013b) was used in this study (available at http://soil-scientist.net). Sampling locations were selected purposively and by convenience (Samuel- Rosa et al., 2014a). Three soil pits were opened within an area of about 100 m^2 at most sampling locations to obtain composite samples of the topsoil for laboratory analysis. Soil was collected to a depth of 20 cm or less when soil depth was smaller than 20 cm. A few observations ($n = 10$) correspond to individual samples collected up to 30 cm. Sampling depth ranges from 2 to 30 cm, with a mean of 17.3 cm. We assumed that the vertical, horizontal and temporal support differences between soil samples are negligible for the purpose of this study.

Three soil properties (fine earth fraction, <2 mm) were explored: clay content (CLAY, g kg^{-1}), organic carbon content (SOC, g kg^{-1}), and effective cation exchange capacity (ECEC, mmol kg^{-1}). CLAY was determined by the pipette method. SOC was determined using wet digestion. ECEC was calculated as the sum of exchangeable bases plus exchangeable acidity. The soil properties selected were expected to present different patterns of spatial variation and correlation with the most dominant factors of soil formation (Jenny, 1994) in the area: organisms (O), relief (R), and parent material (P). CLAY was presumed to have a stronger relation with P, while SOC was expected to be more correlated with O. Because the soils of the study area were strongly eroded due to intense agriculture in the 20th century, both CLAY and SOC were also expected to be closely related with R. Finally, ECEC was expected to be strongly correlated with P and O, which is supported by its natural relationship with both CLAY and SOC.

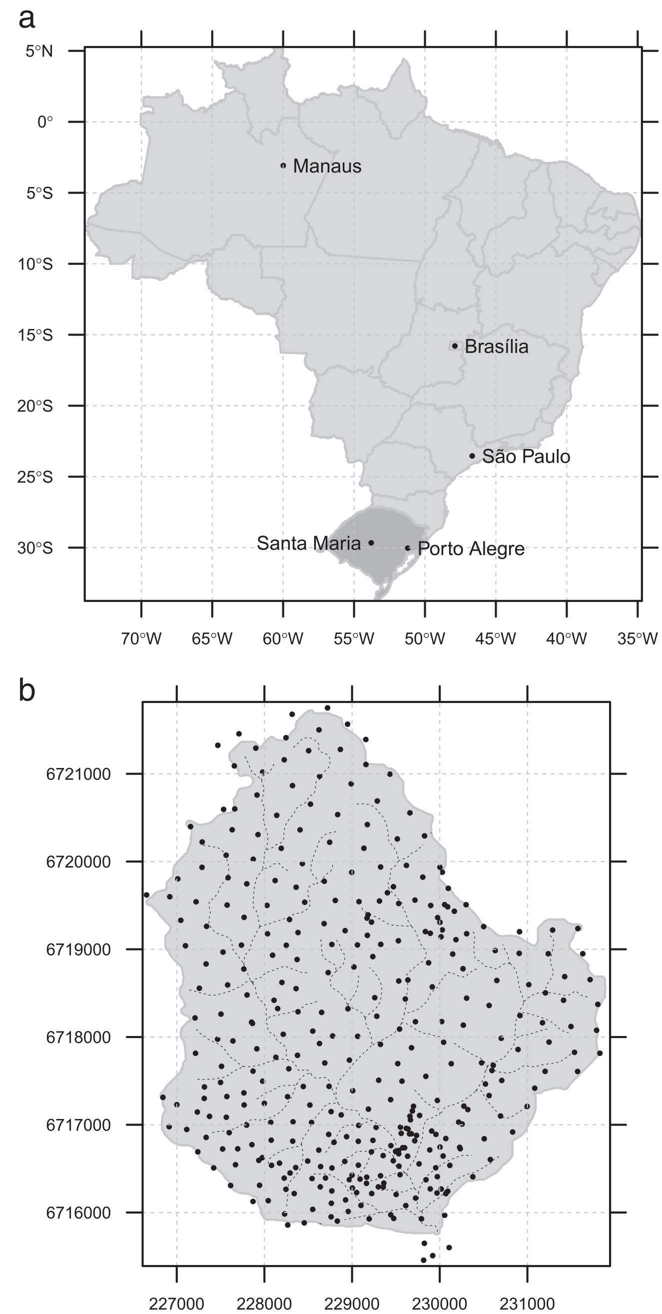

Fig. 1. Location of the study area in Santa Maria (a) and spatial distribution of the point soil observations and drainage network (b).

Point soil data, here denoted by $Z(s)$, where Z is the soil property and s geographic location, showed a positive skew (Fig. 2) and were normalized, $Z'(s)$, using the Box–Cox family of power transformations, where $Z'(s) = (Z(s)^\lambda - 1) / \lambda$, if $\lambda > 0$, and $Z'(s) = log(Z(s))$, if $\lambda = 0$ (Diggle and Ribeiro Jr., 2007). Lambda (λ) values were selected empirically