Digital soil mapping for soil class prediction in a dry forest of Minas Gerais, Brazil[1]

Ricardo de Oliveira Dart[2]; Gustavo M Vasques[3]; Maurício Rizzato Coelho[3]; Nelson Ferreira Fernandes[4]

[1] Trabalho executado com recursos da Embrapa (Projeto MP3 número 03.10.06.013.00.00)
[2] Analista; Embrapa Solos; Rio de Janeiro, RJ; ricardo.dart@embrapa.br; [3] Pesquisador; Embrapa Solos; [4] Professor; Universidade Federal do Rio de Janeiro.

Abstract: Investment on soil survey has become scarce over the past decades. Digital Soil Mapping (DSM) techniques emerged as an economic alternative to produce soil maps. We applied a classification tree algorithm to predict soil suborders in a tropical dry forest area with 102 km^2 in the north of Minas Gerais state, Brazil. We tested environmental covariates with different spatial resolutions as predictors, and used 361 observations to train the model and 64 independent observations to validate the map. Prediction models included three decision trees and one logistic regression model. The results showed that freely available environmental covariates with coarser spatial resolution can produce as good or better suborder predictions than more expensive covariates with finer resolution.

Keywords: Environmental covariates, classification tree, spatial resolution.

INTRODUCTION

Brazil has less than 3% of the territory with soil mapped in scales of 1/50.000 or more (Santos et al., 2013), although soil maps are very important for land planning and management (Silva et al., 2014). Digital Soil Mapping (DSM) has been proposed as an alternative to represent continuous soil variation in space, as opposed to discrete maps produced by traditional soil survey. Moreover, DSM has potential to reduce the time of soil surveying (Zijl et al., 2014).

Approaches for soil class prediction based on point support (Brungard et al., 2015) or by disaggregating legacy soil maps (Collard et al., 2014) have been tested in different places, but to our knowledge soil suborder-environmental correlations in tropical dry forests have not been studied using DSM methods.

Our objective were to: 1) predict soil classes at the suborder level according to the Brazilian System of Soil Classification (Embrapa, 2006) in an area of tropical dry forest using environmental covariates with different spatial resolutions; and 2) validate the results using independent validation samples.

METHODS

Study area

The *Parque Estadual da Mata Seca* (PEMS; Dry Forest State Park) spans across 102 km^2 in the county of Manga in the north of Minas Gerais state, Brazil **(Figure 1)**. The relief in the park is flat (64%) and undulating (31%).

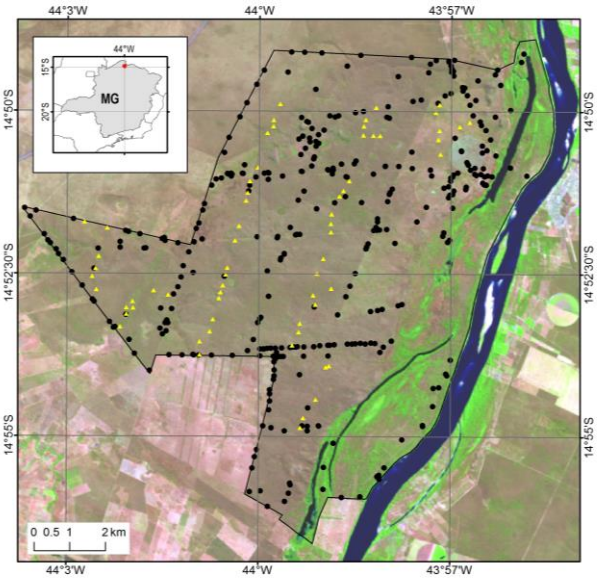

Figure 1 – *Parque Estadual da Mata Seca* (PEMS) in north of Minas Gerais (MG) state; and PEMS limits with the training (black dots) and validation (yellow dots) observations, Landsat 8 Operational Land Imager is shown in false color (RGB = 6,5,4).

The main soils that occur in the PEMS according to the polygon soil map of Coelho et al., (2013) are *Gleissolo Háplico* (GX), *Neossolo Flúvico* (RY) and *Cambissolo Flúvico* (CY) in the floodplain and terrace of the *São Francisco* River, presence of riparian forest. Under *Carrasco* vegetation, small vegetation that occurs in arid highlands, there is a presence of *Latossolo Amarelo* (LA) and *Latossolo Vermelho-Amarelo* (LVA). In the great area of Dense Arboreal *Caatinga* between the *Carrasco* and the floodplain, *Latossolo Vermelho* (LV) and *Cambissolo Háplico* (CX) dominate, followed by *Chernossolo Háplico* (MX) and *Vertissolo Háplico* (VX) (**Figure 2**;). The soil map of PEMS was classify according the Brazilian Soil System of Classification (Embrapa, 2006).

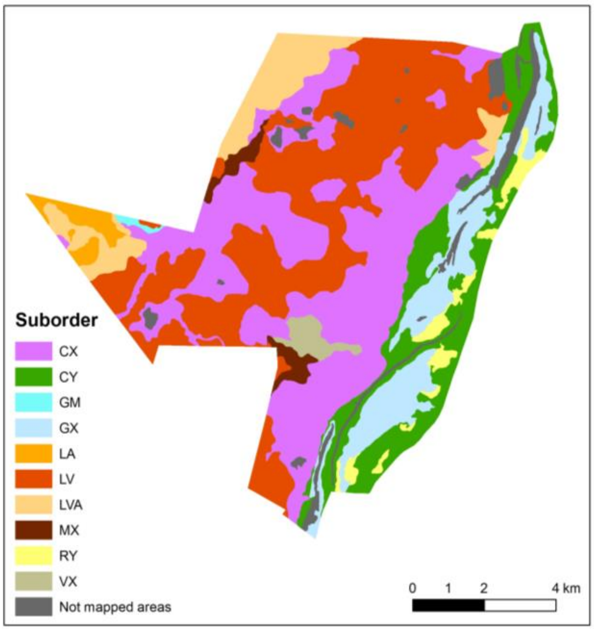

Figure 2 – Polygon soil map of PEMS at suborder categorical level (Coelho et al., 2013).

Soil sampling

In the first field work 361 training sites were visited, where soils were sampled and classified at the suborder level (Embrapa, 2006). They were located using a combination of purposive (261 sites) and conditioned Latin Hipercube (Minasny & McBratney, 2006) samples (100 sites). In the last field campaign the 64 validation sites were visited, which were located using stratified sampling, with strata defined as the combination of density of training sites and environmental heterogeneity. Suborder class *Gleissolo Melânico* (GM) with only 2 observations were grouped with the most similar suborder class Chernossolo Háplico (MX).

Environmental covariates

Two sets of environmental covariates were prepared with different spatial resolutions, namely: *detailed* (10 m), and *regional* (30 m). The detailed set was derived from Ikonos and RapidEye imagery, whereas the regional set was derived from SRTM (30 m) and Landsat 8 (L8) Operational Land Imager (OLI) imagery, respectively.

In the detailed set, a Digital Elevation Model (DEM) was obtained from a 1-m Ikonos stereo pair, which was resampled to 10-m resolution and then corrected by filling spurious depressions (Planchon & Darboux, 2002). Sixteen terrain derivatives were derived from the DEM: Slope (SLO), Profile Curvature (PFCV), Plan Curvature (PLCV), Aspect (ASP), LS-Factor (LSF), Valley Depth (VDP), Relative Slope Position (RSP), Multiresolution Index of Valley Flatness (MRVBF), Multiresolution Ridgetop Flatness Index (MRRTF), Topographic Position Index (TPI), Terrain Ruggness Index (TRI), Terrain Surface Texture (TST), Slope Length (SLG), Slope Height (SHT), Mid Slope Position (MSP), and Topographic Wetness Index (TWI). Two RapidEye (RE) images taken in the wet (REWet; May, 2013) and dry (REDry; August’ 2012) periods were orthorectified, atmospherically corrected, resampled from 5- to 10-m resolution, and then used to derive: Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI), and Soil Adjusted Vegetation Index (SAVI) for both periods.

The regional set was initially composed of the same set of covariates of the detailed set, but they were derived from coarser imagery with 30-m resolution, and included: a SRTM DEM and 16 terrain derivatives, and two L8 OLI images from the wet (L8Wet; March, 2014) and dry (L8Dry; July, 2014) periods with the same three vegetation indices from both periods. The same preprocessing steps were applied to the regional imagery. In addition, three indices related to parental material were added to the regional set: Normalized Difference Ratio Carbonate (NDRC), Ratio Carbonate Index (RCI), and Ratio Hydroxyl Index (RHI) (Boettinger et al., 2008).

Modeling and validation

Four models were created considering the soils classified in suborder categorical level (Embrapa, 2006): M1 using the detail set, M2 using the regional set, and M3 and M4 using both detail and regional covariates. Models 1 through 3 were derived using the C5.0 decision tree algorithm (QUINLAN, 1993), and 10-fold cross-validation. Model 4 was derived using multinomial logistic regression with stepwise selection with p-value threshold of 0.25 to enter variables and 0.10 to remove them.

The overall prediction errors calculated from the confusion matrices were used for model comparison. The best model was the one with the lowest prediction error of external validation.

Visually evaluate the resulted maps from models M1 to M4 against polygon soil map (Coelho et al., 2013) into a geographical information system (GIS).

RESULTS AND DISCUSSION

In general, only model M2 was not able to predict all the nine-suborder soil class, however presented good results (Table 1). According to the external validation errors, the best models were M2 and M3, with a validation error of 42% **(Table 1)**. Compared to M1 and M2, M3 had all predictors with different resolutions to choose from, resulting in a map that is a combination of fine and coarse spatial patterns observed from the detailed and regional covariates, respectively **(Figure 3c)**. From M3, some classes were underestimated (MX and VX) and some were overestimated (LV and CX), compared to Coelho et al. (2013; **Figure 2**).

The equally accurate M2 was more parsimonious than M3, however it did not predict the VX suborder in the study area **(Figure 3b)**. Due to the coarser spatial resolution of the covariates (30 m), M2 found it difficult to correctly predict MX and VX, which occur in small areas in the park. On the other hand, M2 produced a simpler map that shows smoother soil variation, without loss of prediction quality. This can be appealing to users.

Model 1 produced the suborder map with most variation at the short scale **(Figure 3a)**, because it only used detailed covariates with 10-m resolution. However, this was the worst map according to external validation, even though these covariates were more detailed. Thus, whether to invest in covariates that are more detailed for DSM should be decided carefully and on a case-by-case basis (Samuel-Rosa et al., 2015).

Finally, M4 was the most parsimonious model, with 24 predictors. Most predictors (14 out of 24) were selected from the detailed set, and three covariates were selected from both sets with different resolutions: TWI, VDP, and infra-red band from the wet images (L8WetB5 and REWetB5).

Table 1 – Summary of model results.

Model	Number of covariates	Training error (%)	Validation error (%)
M1	33	50.7	51.5
M2	43	48.5	42.2
M3	76	51.0	42.2
M4	24	44.3	48.4

CONCLUSIONS

Soil suborder variations in the Brazilian dry forest relate to relief and vegetation patterns at different spatial resolutions.

In DSM, using more detailed environmental predictors, which usually cost more, does not necessarily mean achieving better predictions.

Along the same lines, combining covariates with different spatial resolutions may or not improve model quality.

REFERENCES

BOETTINGER, J. L.; RAMSEY. R. D.; BODILY, J. M. et al. Landsat spectral data for digital soil mapping. In: HARTEMINK, A. E.; MCBRATNEY, A. B.; MENDONÇA-SANTOS, M. L. (Eds.). **Digital soil mapping with limited data.** Amsterdam: Springer, 2008. p. 193-202.

BRUNGARD, C. W.; BOETTINGER, J. L.; DUNIWAY, M. C. et al. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma,** 239–240:68-83, 2015.

COELHO, M. R.; DART, R. O.; VASQUES, G. M. et al. Levantamento pedológico semidetalhado (1:30.000) do Parque Estadual da Mata Seca, município de Manga - MG. **Boletim de Pesquisa e Desenvolvimento,** Rio de Janeiro, n. 217, 264 p: Embrapa Solos, 2013.

COLLARD, F.; KEMPEN, B.; HEUVELINK, G. B. M. et al. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). **Geoderma Regional**, 1:21-30, 2014.

EMBRAPA. **Sistema Brasileiro de Classificação de Solos**. 2. ed. Rio de Janeiro: Embrapa Solos, 2006. 306p.

MINASNY, B. & MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences,** 32:1378-1388, 2006.

PLANCHON, O. & DARBOUX, F. A fast, simple and versatile algorithm to fill the depressions of digital elevation models. **Catena,** 46:159-176, 2002.

QUINLAN, J. R. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann Publishers, 1993. 302p.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M. et al. Do more detailed environmental covariates deliver more accurate soil maps? **Geoderma,** 243-244:214-227, 2015.

SANTOS, H. G.; ÁGLIO, M. L. D.; DART, R. O. et al. Distribuição espacial dos níveis de levantamento de solos no Brasil. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 34., Florianópolis, 2013. Anais. Florianópolis: Sociedade Brasileira de Ciência do Solo, 2013.

SILVA, A. F. PEREIRA, M. J.; CARNEIRO, J. D. et al. A new approach to soil classification mapping based on the spatial distribution of soil properties. **Geoderma,** 219-220: 106-116, 2014.

ZIJL, G. M. V.; BOWVER, D.; TOL, J. J. V. et al. Functional digital soil mapping: A case study from Namarroi, Mozambique. **Geoderma,** 219-220:155-161, 2014.

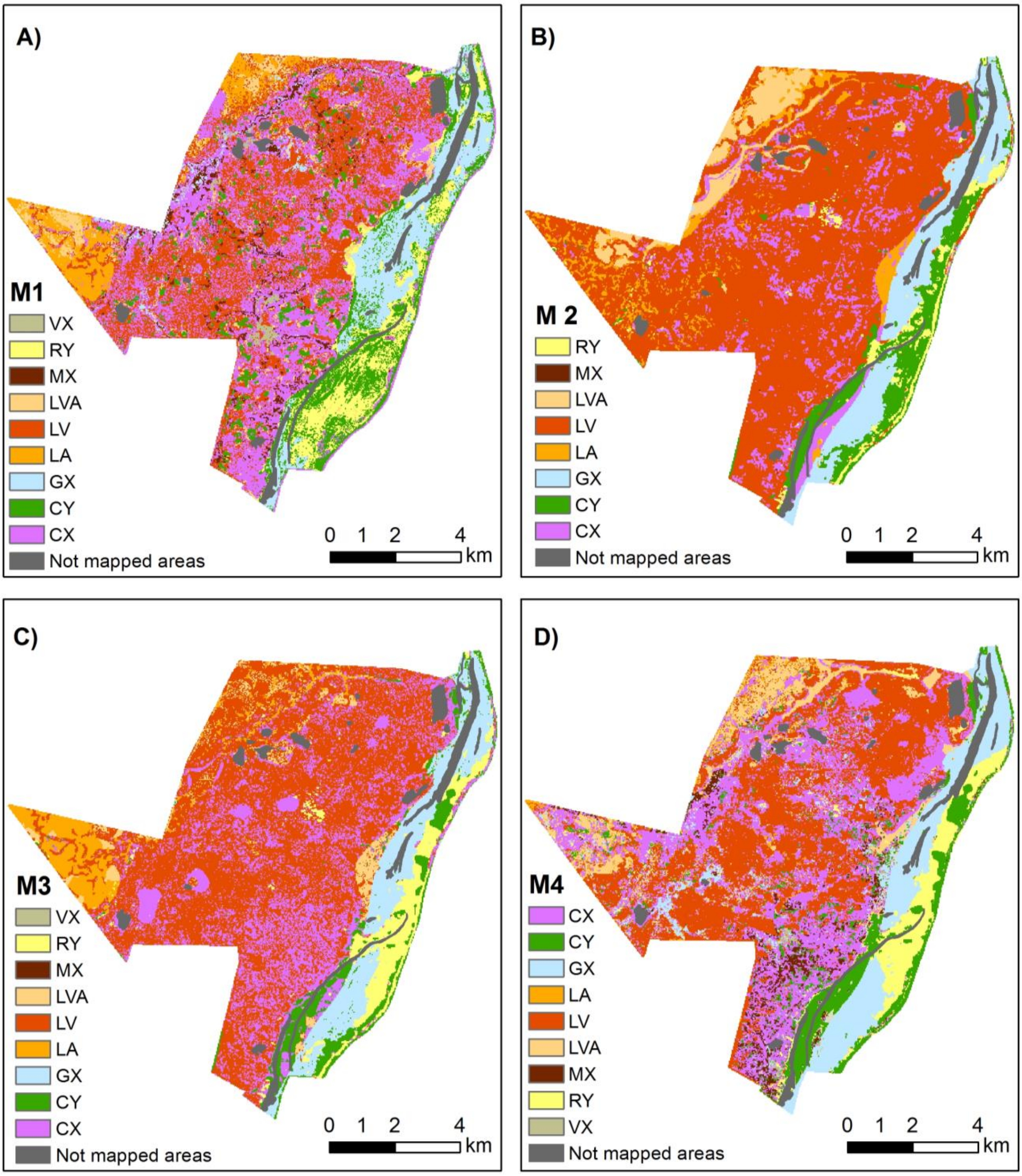

Figure 3 – Soil suborder maps at the *Parque Estadual da Mata Seca* produced by the different models: A) M1; B) M2; C) M3; and D) M4.