Digital Soil Mapping at Parque Estadual da Mata Seca, Minas Gerais state, Brazil: applying Regression Tree to predict soil classes

R.O.Dart[1], M.R. Coelho[1], M.L. Mendonça-Santos[1], J.G. Pares[2] and R.L.L.Berbara[3]

[1]Researcher at EMBRAPA Solos – Brazilian Agricultural Research Corporation / The National Centre of Soil Research. Rua Jardim Botânico 1024, 22.460-000, Rio de Janeiro, RJ, Brazil.
[2]Embrapa's fellowship grant. EMBRAPA Solos –Brazilian Agricultural Research Corporation, The National Centre of Soil Research. Rua Jardim Botânico 1024, 22.460-000, Rio de Janeiro, RJ. Brazil.
[3]Professor at UFRRJ – The Federal Rural University of Rio de Janeiro. BR 465, km 7, 23.890-000, Seropédica, RJ, Brazil.

Abstract

The use of Digital Soil Mapping (DSM) to predict soil classes is an important issue to decrease costs and subjectivity of soil maps. The main objective of this study was to use DSM to produce soil maps of a relatively small area (about 100 km^2) and compare it to a preliminary soil map made by traditional techniques. The study area is located at north of Minas Gerais State, southwest of Brazil. In this study we used decision tree classifier, See5, and 278 soil samples to predict soil class at order level of the Brazilian System of Soil Classification. We also did use ancillary data as Landsat ratios and variables of the topography. DSM didn't show a good performance of soil prediction because basically three factors: (a) taxonomic similarity between Argissolos and Latossolos, (b) great spatial and attributes variability of Cambissolos that occurred in different landscapes types, and (c) low accuracy of soil prediction to Gleissolos, Neossolos and Cambissolos of the river plain domain because its shows great environment complexity. Following works will make a better selection of environmental covariates, predict the soil classes in higher categorical level and assessment of quality of digital soil maps.

Keywords: DSM, Regression Tree, Soil, *s.c.o.r.p.a.n* model.

1. Introduction

In Brazil there is a shortage and rising demand for soil surveys in more detailed than 1:50,000 scale, aiming at the planning of land use for various purposes. However, most of Brazilian territory lacks basic cartographic material (especially aerial photography) on a scale compatible with the publication of the final soil map, which ultimately has serious implications on their quality, time and cost of implementation. The Digital Soil Mapping (DSM) can help meet that need. The integration of both mapping procedures, DSM and traditional soil survey, which is actually already in the works of DSM using the variable "*s*" of *s.c.o.r.p.a.n* model, can generate products suitable for the needs of users in relatively small areas, on time and with good accuracy, which can be measured in a digital environment.

DSM was developed as a substitute for the traditional polygon soil maps (McBratney et al., 2003). A digital map of soil classes or properties allows researchers to use them in digital programs, like GIS, and also to organize, store, analyze and interpret large amounts of data anytime. Because of this, we are seeing an increase interest in use of DSM for soil and properties mapping in the worldwide.

In this study we use DSM procedures to predict soil classes at the order level according to Brazilian System of Soil Classification (Embrapa Solos, 2006) in a small area (about 100 km^2) with absence of adequate cartographical material. For this, we use regression tree classification and *s.c.o.r.p.a.n* variables as the predictive modeling framework. Four different models were built. The best one was compared with a preliminary soil map of the area made by traditional procedure in order to provided qualitative and quantitative information about the quality of digital soil map.

2. Material and Methods

2.1. Study Area

The study area named *Parque Estadual da Mata Seca* (PEMS) is located in Minas Gerais State, Manga county (Fig. 1). PEMS has an area of 10.281ha distributed in 3 landscape types (Tab.1).

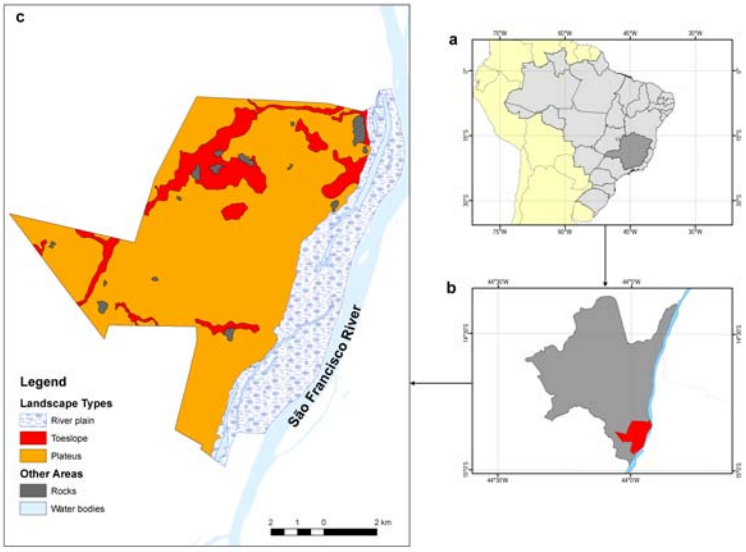

Figure 1. The study area location: (a) Minas Gerais State; (b) PEMS location at Manga county; (c) Landscape types at PEMS.

Table 1. Landscape types and corresponding soil class in PEMS.

Landscape types		Environmental characteristic	Soil class
River plain	Floodplain	Permanentely flooded	Gleissolos
	Alluvial deposits	Permanentely dry	Cambissolos and Neossolos
Toeslope		Temporaly flooded	Cambissolos and Plintossolos
Flat to undulating plateaus		Higher in the landscape	Argissolos, Cambissolos and Latossolos

2.2. Digital and field data and inference models

The soil dataset consist of 278 soil samples which was randomly divide in 209 samples for training and 69 for validation. Fig. 2 shows the distribution of the soil orders according to Brazilian System of Soil Classification (SiBCS; Embrapa Solos, 2006), and its taxonomic correlation with WRB (2007) is presented in Tab. 2.

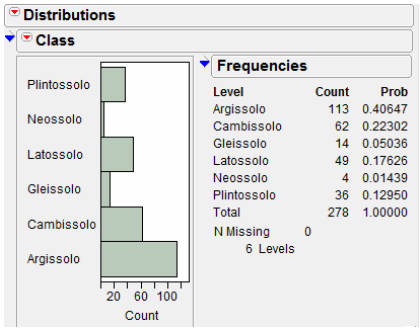

Figure2. Soil Order and frequency of the 278 soil samples used in the modeling process.

Table 2. Soil class in PEMS according SiBCS and WRB.

SiBCS (Embrapa Solos, 2006)	WRB (2007)	%
Latossolo	Ferrasol	18.8
Argissolo	Ferrasol	35.3
Cambissolo	Cambisol	32.7
Plintossolo	Plinthosol	1.0
Neossolo	Fluvisol	3.4
Gleissolo	Gleysols	8.8
Total		100.0

The following covariates were used as predictor variables of *s.c.o.r.p.a.n* model: Landsat ^{ETM} observed in September 2000 (GLCF, 2003); Landsat ratios NDVI, B3/B2 and B5/B7; and the SRTM DEM 30 m (Valeriano, 2008). After 2 passes of mean filters 3x3 SRTM DEM 30m was used as input to LandMapR software (MacMillan, 2003) to obtain DEM derivates to be used in predictive models. (Tab. 3)

Table 3. Environmental covariates used to predict soil class and reference map used to compare the predictive model.

Name	Scale / Resolution	Nature of the soil variable derived	Name	Description	*S.C.O.R.P.A.N* factors	Type
Preliminary Soil Map	1/50000	Reference map	soil	6 classes representing soil-landscape units of PEMS	Sc	Categorical
DEM	30m	Topography	elev	elevation (m)	R	Quantitative
			slope	slope gradient (%)		
			prof	profile curvature		
			plan	plan curvature		
			aspect	downslope direction		
			z2pit	absolute height (Z) above the local pit cell (MacMillan *et al.*, 2000)		
			pctz2top	percent Z relative to top & botton of each watershed (MacMillan *et al.*, 2000)		
			pctz2pit	percent Z relative to local pit & peaks (MacMillan *et al.*, 2000)		
			pimin2max	percent Z relative to min & max elevation for the entire study area (MacMillan *et al.*, 2000)		
		Hydrology	qweti	wetness index (MacMillan *et al.*, 2000)		
			pctz2st	percent Z relative to nearest stream & divede (MacMillan *et al.*, 2000)		
Landsat ^{ETM}	30m	Spectral reflectance	NDVI	Normalized Difference Vegetation Index	O	Quantitative
			B5/B7	acentuate hydroxyl radicals (Boettinger *et al.*, 2008)		
			B3/B2	acentuate carbonate radicals (Boettinger *et al.*, 2008)		

An ancillary dataset representing the whole study area were compiled on a 30m grid, and populated with environmental covariates. The modeling and prediction of soil classes was done by a regression/classification tree, using See5 software (RuleQuest Research, 2003).

Four models were created with the most comonly environmental covariates used in DSM for soil class prediction (Grinand et al., 2008; Mendonça-Santos et al., 2008). The main difference between the models is the set of prediction variables used to build them (tab. 4). Akaike's Information Criterion (AIC) was used in order to compare the performance of the models. The best model is the one that has the smallest AIC (Akaike, 1973).

A randomaly selected, indenpendent validation set was used to assess the accuracy of classification, in order to compare the best model with a existing preliminary soil map of PEMS.

Table 4. Predictive models *s.c.o.r.p.a.n* used to estimate soil classes and results of the prediction.

Models	Predictor variables	Number of parameters	Training data errors %	Validation data errors %	AIC
M1	***O*** (NDVI, B5/B7, B3/B2), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	9	5.7	27.5	1659.6906
M2	***O*** (NDVI, B5/B7, B3/B2), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE, Z2PIT, PCTZ2TOP, PCTZ2ST, PCTZ2PIT, PIMIN2MAX)	14	5.7	33.3	1616.4177
M3	***O*** (NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	7	2.4	27.5	1602.4177
M4	***O*** (NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE, Z2PIT, PCTZ2TOP, PCTZ2ST, PCTZ2PIT, PIMIN2MAX)	12	2.4	27.5	1456.2174

3. Results and Discussion

3.1. Results from the Models

In general, the 4 models (M1...M4) were able to predict all 6 soil classes, showing reasonable and low error values (higher validation data error was 33.3%; Tab. 5). The percentage of correctly predicted soil class in 4 models was 66-72%, similar to findings by Minasny and McBratney (2007). It means that soil classes in the study area could be predicted using environmental covariates that can be easier and cheaper (or even free) to acquire, like Landsat^{ETM} images and the SRTM DEM 30 m.

The M4 was considered the best model (smallest AIC, Tab.3) with a good spatial distribution all over the study area. The soil classes appeared as expected by knowledge expert acquired from field work and preliminary soil map. Fig. 3 illustrates the resulting map of M4.

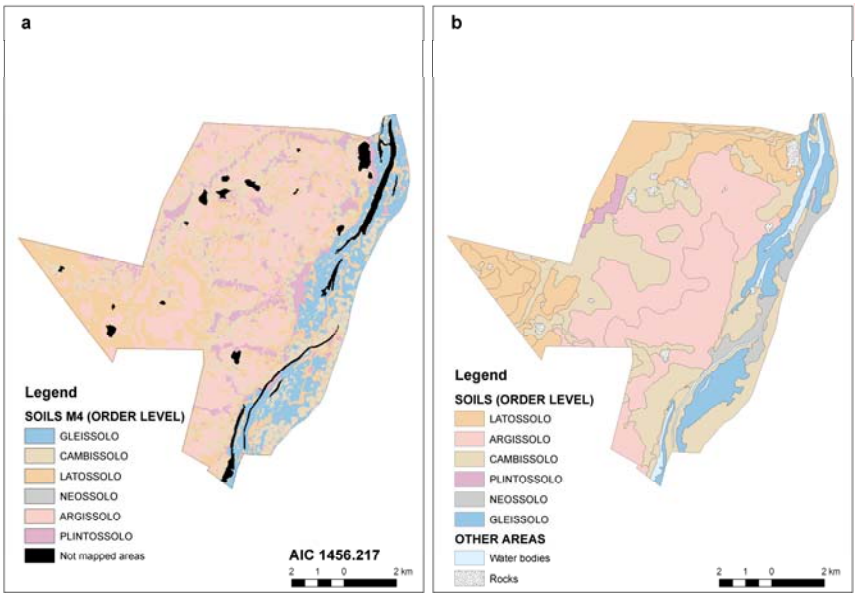

Figure 3. Soil maps of PEMS: (a) resulting from M4; (b) preliminary soil map (1/50000).

3.2. Analysis of the best Model

Indubitally, floodplain and alluvial deposits dominated by Gleissolos, Neossolos and Cambissolos had the best prediction compared with preliminary soil map and expert knowledge. This assertive can be partialy analyzed in the Tab. 5 that compare traditional and predictive soil maps. Gleissolo had the best prediction with about 55% of accuracy (Tab. 5), although it's occuring in the most heterogeneous physiographic domain of PEMS.

Tab. 5 shows that, from the 33 soil samples classified as Gleissolos by the DSM, the traditional mapping classified 12 as Cambissolos and 3 as Neossolos. It must be taken in consideration that part of the samples classified as Cambissolos and all the samples classified as Neossolos were found under alluvial deposits, and this domain naturally have a large spatial variability of soil types.

Table 5. The confusion matrix for regression tree model (M4) of PEMS shows the number and percentages of correctly classified pixels.

	Soil Order (SiBCS)	Soil classes of preliminary soil map (reference)						Total	**User's accuracy (%)**
		G	C	L	R	P	F		
Classified	Gleissolo (G)	**18**	12		3			33	54,5
	Cambissolo (C)	2	**24**	8	4	23	3	64	37,5
	Latossolo (L)	5	22	**26**	1	30		84	31,0
	Neossolo (R)		1					1	0,0
	Argissolo (P)		22	28		**45**		95	47,4
	Plintossolo (F)	2	7	3		9	**2**	23	8,7
	Total	27	88	65	8	107	5	300	
	Producer accuracy (%)	66,7	27,3	40,0	0,0	42,1	40,0		**38,3**

Fig. 4a presents the results of discriminant analysis of the soil classes. Soil classes that are significantly different tend to have non-intersecting circles (Minasny and McBratney, 2007). Plintossolos are well separated from other classes, because of their unique environmental properties. However, the other classes didn't occupied unique position in the discriminant place. Gleissolos e Neossolos appear to be related despite the great uncertainty about the centroid because of the small sample sizes, specially to Neossolos. This result confirms the previous one in terms of the difficulty to separate those soil classes in a varied and complex landscape as floodplain and alluvial deposits.

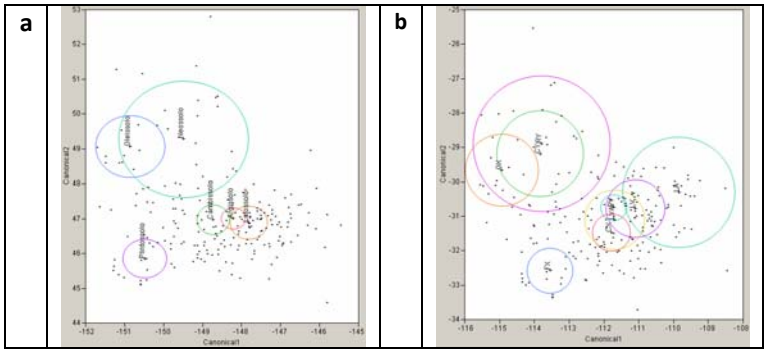

Figure 4. Canonical plot of PEMS using environmental covariates of M4 shows: (a) soil orders; (b) soil sub-orders. Symbols: CX = Cambissolo Háplico; CY = Cambissolo Flúvico; FX = Plintossolo Háplico; GX = Gleissolo Háplico; LA = Latossolo Amarelo; LV = Latossolo Vermelho; LVA = Latossolo Vermelho-Amarelo; PV = Argissolo Vermelho; RY = Neossolo Flúvico.

As seeing in Fig. 4a Argissolos and Latossolos are much related to each other, which agree with the taxonomic parameters used to distinguish both classes in the study area. The classes were morphologic separated just by the presence of clay coatings in Argissolos, which is low expressive or not present in Latossolos. However, many Argissolos and Latossolos samples were found in the same landscape type (flat to undulating plateau); consequently, they present similar environmental covariates. On the other hand, Cambissolos occurs in all landscape types of PEMS and shows great soil

variability attributes. In the Fig. 4a Cambissolos are related to Argissolos because of the large samples density and commom ocorrence in similar landscape types. These results are confirmed when we analyse canonical plot at soil sub-orders (Fig. 4b) and compare traditional and predictive soil maps, considering the landscape type in PEMS (Tab. 6). The loss matrix showed in Tab. 6 emphasizes the significant increase of the overall accuracy (85,7% - Tab. 6 *versus* 38,3% - Tab. 5) when we divide the study area in landscape types. These landscape types, in turn, coincide with those ones resulting from groupings of soil classes with taxonomic similarities (Tab. 1).

Table 6. The confusion matrix for regression tree model (M4) of PEMS grouped in landscape types, shows the number and percentages of correctly classified pixels.

	Landscape types	Soil classes of preliminary soil map (reference)			Total	**User's accuracy (%)**
		River plain	Tooslope	Plateaus		
Classified	River deposits	**47**	1		48	97,9
	Toeslope	4	**14**	16	34	41,2
	Plateaus	7	15	**196**	218	89,9
	Total	58	30	212	300	
	Producer accuracy (%)	81,0	46,7	92,5		**85,7**

4. Conclusions

Four different models (M1...M4) were built and tested using DSM procedures in order to predict soil orders in an area of approximately 100 km^2. The best model, M4, was compared to a preliminary soil map of the area made by traditional survey procedure. Both mapping procedures were efficient to delimitate river plain domains (floodplain and alluvial deposits). However, the 3 different soil orders identified in these landscape type, Cambissolos, Neossolos and Gleissolos, did not show similar predictions in both of mapping procedure, mainly due to the large spatial variability and attributes of Cambisols in the area. Similar trend was observed for the other 3 soil orders identified in the PEMS, Argissolos, Latossolos and Plintossolos. This problem was also reported in Minasny and McBratney (2007).

When we grouped soil orders of PEMS by similarity of environments had a significant increase in overall acuracy (38.3 to 85.7%). However, in case of Argissolos and Latossolos of the park area, SiBCS makes use of subtle and subjective morphological attributes (for example, intensity and quantitative of clay coatings) for discern them, which is not reflected in the diversity of environmental factors between soils classes. Following works will make a better selection of environmental covariates, predict the soil classes in higher categorical level and evaluate the uncertanty of digital soil maps.

7. References

Akaike, H. 1973. Information theory and an extension of maximum likelihood principle. *In*: B.N. Petrov and F. Csaki (Eds.), Second International Symposium on Information Theory. Akademia, Kiado, Budapest, pp. 267–281.

Boettinger, J.L, Ransey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M. and Stum, A.K. 2008. Landsat Spectral Data for Digital Mapping. P. 193-202. In: A.E. Hartemink,

A. McBratney and M.L. Mendonça-Santos (eds.). Digital Soil Mapping with Limited Data. Springer, Amsterdam.

Embrapa Solos. 2006. Sistema Brasileiro de Classificação de Solos. 2. ed. 306p.

Global Land Cover Facility - GLCF. 2003. NASA Landsat Program. Available in: <http://www.landcover.org/> Last verified 25 Sep 2008.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P. 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180–190.

IUSS Working Group WRB. 2007. World Reference Base for Soil Resources 2006, first update 2007. World Soil Resources Reports No. 103. FAO, Rome.

MacMillan, R. A. 2003. LandMapR© Software Toolkit- C++ Version: User manual. LandMapper Environmental Solutions Inc., Edmonton, AB. 110 pp.

MacMillan, R.A., Pettapiece,W.W., Nolan, S.C., Goddard, T.W. 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113, 81–109.

McBratney, A.B., Mendonça-Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Mendonça-Santos, M.L., Santos, H.G., Dart, R.O., Pares. J.G. 2008. Digital Mapping of Soil Classes in Rio de Janeiro State, Brazil: Data, Modelling and Prediction. P. 381-396. In: A.E. Hartemink, A. McBratney and M.L. Mendonça-Santos (eds.). Digital Soil Mapping with Limited Data. Springer, Amsterdam.

Minasny, B., McBratney, A. B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285–293.

RuleQuest Research. 2003. See5/C5.0 version 1.20. RuleQuest Research Pty Ltd., Sydney, Australia. http://www.rulequest.com/see5-info.html.

Valeriano, M. de M. 2008. TOPODATA: Guia para utilização de dados geomorfométricos locais. São José dos Campos: INPE.