Applying Conditional Latin Hypercube (cLHS) for selecting soil sampling location for Digital Soil Mapping at Parque Estadual da Mata Seca, MG, Brazil

M.L. Mendonça-Santos[1], R.O. Dart[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, CEP 22.460-000, Rio de Janeiro, RJ, Brazil. e-mail: loumendonca@cnps.embrapa.br
[2] Fellow at Embrapa Solos. e-mail: rdart81@yahoo.com.br
[3] Professor at UFRRJ - Universidade Federal Rural do Rio de Janeiro, BR 465 Km47, CEP23890-000, Seropédica, RJ, Brazil. e-mail: berbara@ufrrj.br

Abstract

The use of the Conditional Latin Hypercube (cLHS) to select soil samples location to be used in the prediction of soil properties as soil organic carbon seems to be an important tool to decrease costs and subjectivity of sampling schemes. The main objective of this work was to test this method in a no sampled area, aiming to evaluate its efficiency in Digital Soil Mapping (DSM) procedures. The results show that this method was able to significantly increase the reliability in the spatial distribution of the sampled points in Parque Estadual da Mata Seca (PEMS). The study area is located in the North of Minas Gerais State, Brazil, in Tropical Dry Forest (TDF) ecosystems mainly. In this study, the cLHS algorithm developed by Minasny & McBratney (2006) was used, in addition to some ancillary data as Land Use/Land Cover and Normalized Difference Vegetation Index (NDVI), in order to select 60 soil sampling location to the study of soil organic carbon and others soil properties. The cLHS will be further analyzed in its performance to select representative points to be sampled and how this method could be helpful for DSM.

1 Introduction

In traditional soil mapping, the location of each soil survey is generally determined by empirical procedure, based on air-photo interpretation and the mental “mapping” strategies of the surveyor, that undertake field prospecting correlating soil with underlying geology, landforms, and vegetation. There are no statistical criteria for traditional soil sampling which may lead to bias in the areas being sampled, in addition to the fact that soil sampling is always limited by financial and human resources availability, as well as time. Thus an efficient sampling strategy is welcome in order to optimize soil survey.

Several sampling strategies have been developed in the earth sciences, as related by Minasny & McBratney, 2006. Latin hypercube sampling (LHS) has been proposed as a sampling design for digital soil mapping when there is no prior soil samples (Minasny and McBratney, 2006; Carré et al., 2007).

In digital soil mapping (DSM), the prediction of soil properties and soil classes is based on forming relationships between observed soil attributes and ancillary soil and environmental variables (McBratney et al., 2003). Based on these relationships, ancillary data can be used

in LHS to help on deciding the location of soil sampling, since these data can be cheaper over large areas.

The LHS is a constrained Monte Carlo sampling scheme which uses a stratified random procedure that provides an efficient way of sampling variables from their multivariate distributions (McKay et al., 1979). According to Minasny and McBratney (2006), a Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions (*K*) whereby each sampling unit is the only one in each axis-aligned hyperplane containing it. LHS involves sampling *n* values from the prescribed distribution of each of the variables. The cumulative distribution for each variable is divided into *n* equiprobable intervals, and a value is selected randomly from each interval. The *n* values obtained for each variable are then paired with the other variables. This method ensures a full coverage of the range of each variable by maximally stratifying the marginal distribution.

In this work the methodology proposed by Minasny and McBratney (2006) is applied to a non-sampled area in the Manga District of Minas Gerais State, Brazil, in order to locate 60 soil samples based on ancillary data (NDVI and LULC map), aiming to maximise the efficiency of the sampling scheme while ensuring that the variability within the sampling area is adequately considered.

2 Materials and Methods

2.1 Study Area

Manga district is located north of Minas Gerais State, Brazil. It belongs to the Sao Francisco River Basin. The area has 10,281 Ha (Figure 1). Land cover is predominantly composed of tropical dry forest reminiscent expressing high diversity and is situated in the transition between Cerrado and Caatinga Biomes (IEF, 2007). According to the Soil Map (Embrapa, 2006a) the soil is predominantly Oxisol or *Latossolo*, according to the Brazilian Soil Classification (EMBRAPA SOLOS, 2006b). The dynamics of this Biomes transition will be studied in this work, by remote sensing and digital soil mapping techniques, in order to predict soil attribute variation, with a special focus on organic carbon stock.

2.2 Digital and Field Data

In this application the following covariates were used as predictor variables to the location of soil samples: Land Use/Land Cover (LULC) and Normalized Difference Vegetation Index (NDVI). The LULC and NDVI derived from Quickbird image with 2.5m spatial resolution and four spectral band widths, from 04/06/2004. The LULC was done with *eCognition* 2.1, which were selected 12 representative classes (Table 1) of LULC from PEMS and was performed a supervised classification with standard nearest neighbor.

We did resample the spatial resolution of the predictor's variables to 10 meters, in order to increase the performance of the processing. Thus, we exclude some classes of LULC (roads, rocks, constructions, water) that don't need to be considered to select soil samples we prepare a mask to not consider them in cLHS analysis.

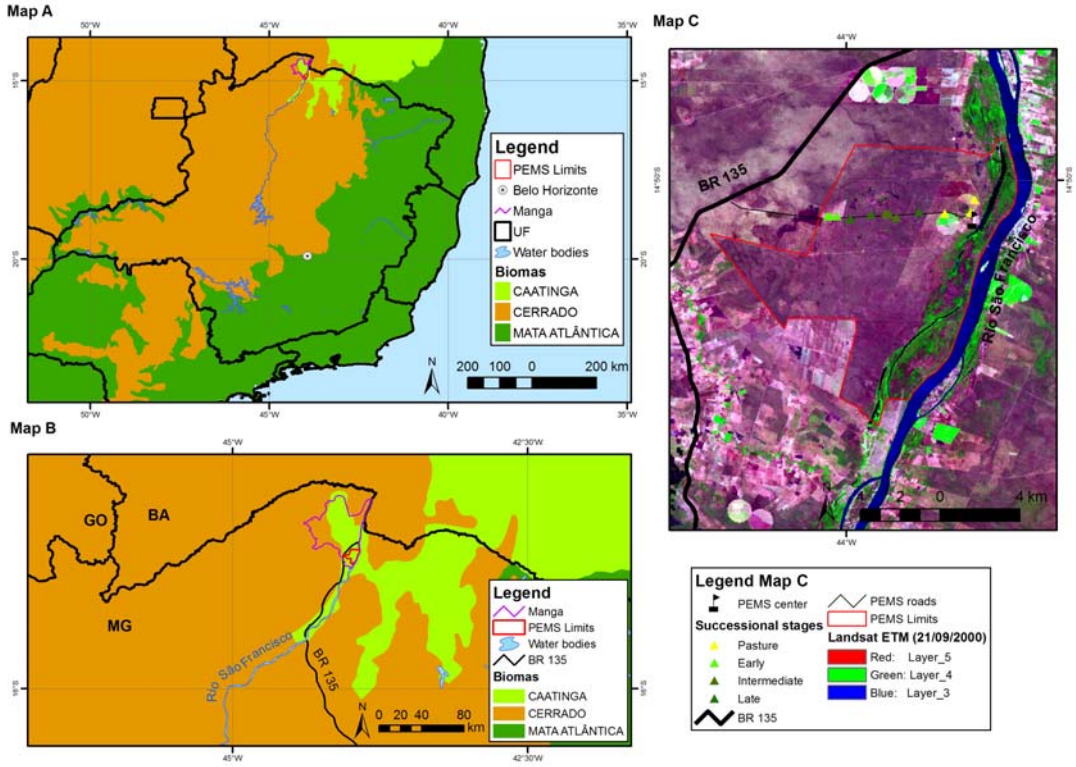

Figure 1. Map A – Location of Manga District and the Parque Estadual da Mata Seca (PEMS) at the North of Minas Gerais State, Brazil; Map B – Detail of Map A; Map C – Details of the PEMS (Satellite image Landsat 7 ETM^{+} , RGB 543, from September 17, 2000).

2.3 Inference Models

To predict the location of the soil samples, the cLHS algorithm developed by Minasny & McBratney (2006) was applied to some ancillary data as LULC and NDVI. In order to use cLHS code we prepared the database to run the algorithm. The predictor's files (LULC and NDVI) were converted into text files. Then they were linked as the ancillary data (LULC+NDVI). In order to spatialize the results we made some changes on the cLHS code to generate a data table (dbf. file) and then to execute cLHS. We run the algorithm to provide 60 points in the study area.

3 Results

The resulted data table (dbf. file) was converted to shapefile in ArcGis 9.3, in order to be a reference to our next soil sample survey in PEMS. Location of the 60 points were soil will be sampled is presented in Figure 2. In order to check if these samples are representative of the soil variability, field work is necessary.

Table 1 Statistical distribution of ancillary variables from original and sampled data

	No. of points	25% quartile	Median	75% quartile	Mean	Std. dev.
NDVI						
Original data	985252	0.254	0.297	0.393	0.330	0.108
cLHS	60	0.258	0.295	0.433	0.335	0.116

Table 2. Details of predictor variables

Predictor Variables	Details	Spatial Resolution (meters)
LULC	12 Classes: three successions of tropical dry forest – early stage, intermediate stage and late stage, pasture, savanna, bare soil, roads, agriculture, rocks, constructions, water and shadow of trees	2.5
NDVI	$\frac{IR - R}{IR + R}$	2.5

Abbreviations: IR = infrared; R = red.

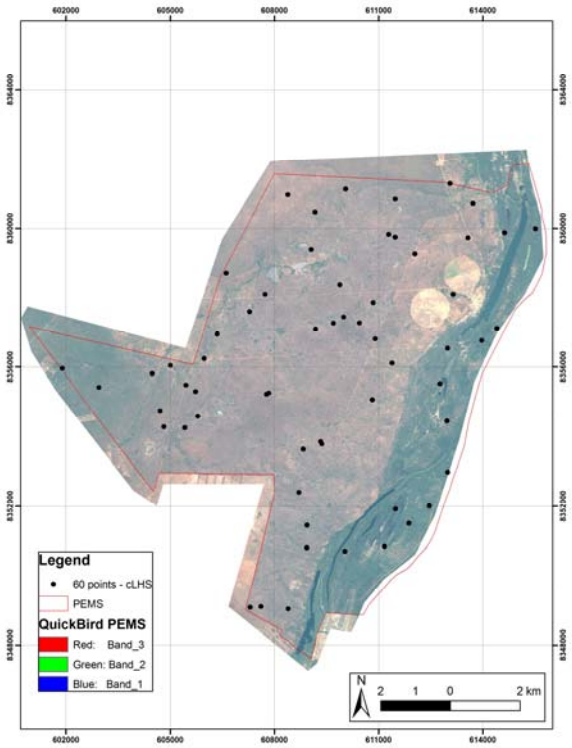

Figure 2. Result of the application of cLHS with 60 points

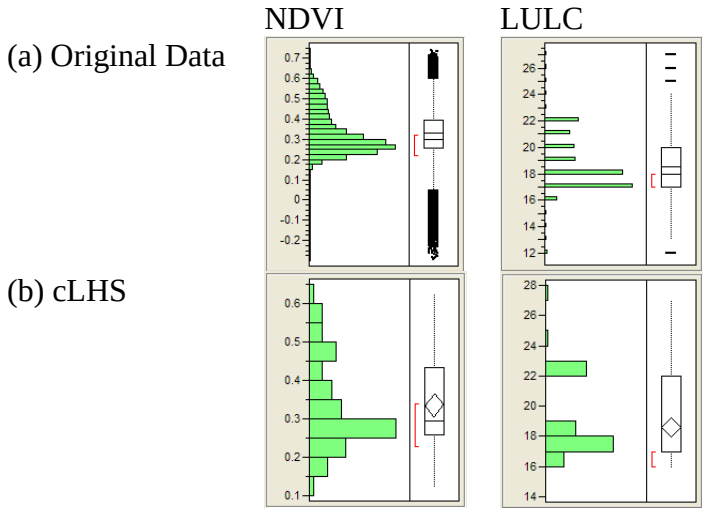

Figure 3. Histogram and box plot of variables (a) original data, (b) cLHS. For box plot, ends of box are 25th and 75th quantiles, line across middle of box identifies median sample value and means diamond indicates sample mean and 95% confidence interval. Bracket along edge of box identifies shortest half, which indicates most dense 50% of observations. (minasny, 2006). LULC class: 12 (pasture), 13 (bare soil), 16 (TDF early stage), 17 (TDF intermediate stage), 18 (savanna), 20 (agriculture), 22 (TDF late stage), 24 (shadow of trees).

4 Conclusions

This work is still in development. Field work is planned, in order to sample soil for digital mapping of soil properties as organic carbon and others. Further work must be done in order to analyse in what measure this kind of previous determination of sampling points can help to optimize the procedures of soil survey and mapping, as well as to compare this methodology with others.

5 References

Carré, F., McBratney, A. B. and Minasny, B. 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma, 141: 1-14,

EMBRAPA SOLOS (Empresa Brasileira de Pesquisa Agropecuária - Solos). 2006a. Mapa de solos do Estado de Minas Gerais. Rio de Janeiro. http://200.20.158.13/website/pub/MG_Solos/ (last verified 02 September 2008).

EMBRAPA SOLOS, 2006b. Sistema Brasileiro de Classificação de Solos. 2. ed. 306p.

IEF (Instituto Estadual de Florestas). 2000. Parque Estadual da Mata Seca. Belo Horizonte, 2000. http://www.ief.mg.gov.br/index.php?option=com_content&task=view&id=204&Itemid=37 (last verified 23 May 2007).

Jenny, H., 1941. Factors of soil formation. McGraw-Hill, New York.

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

McKay, M.D., Beckman, R.J. and Conover, W.J. 1979. A comparison of three methods for selecting values of input variables in the análisis of output from a computer code. Technometrics, vol.1, No.2, pp. 239-245.

Minasny, B., McBratney, A.B. 2006. A conditional Latin hypercube method for sampling in the presence of ancillary information. Computer & Geosciences 32:1378-1388.