AN APPROPRIATE DATA SET SIZE FOR DIGITAL SOIL MAPPING IN ERECHIM, RIO GRANDE DO SUL, BRAZIL[1]

Alexandre ten Caten[2], Ricardo Simão Diniz Dalmolin[3], Fabrício de Araújo Pedron[4], Luis Fernando Chimelo Ruiz[5] & Carlos Antônio da Silva[6]

SUMMARY

Digital information generates the possibility of a high degree of redundancy in the data available for fitting predictive models used for Digital Soil Mapping (DSM). Among these models, the Decision Tree (DT) technique has been increasingly applied due to its capacity of dealing with large datasets. The purpose of this study was to evaluate the impact of the data volume used to generate the DT models on the quality of soil maps. An area of 889.33 km² was chosen in the Northern region of the State of Rio Grande do Sul. The soil-landscape relationship was obtained from reambulation of the studied area and the alignment of the units in the 1:50,000 scale topographic mapping. Six predictive covariates linked to the factors soil formation, relief and organisms, together with data sets of 1, 3, 5, 10, 15, 20 and 25 % of the total data volume, were used to generate the predictive DT models in the data mining program Waikato Environment for Knowledge Analysis (WEKA). In this study, sample densities below 5 % resulted in models with lower power of capturing the complexity of the spatial distribution of the soil in the study area. The relation between the data volume to be handled and the predictive capacity of the models was best for samples between 5 and 15 %. For the models based on these sample densities, the collected field data indicated an accuracy of predictive mapping close to 70 %.

Index terms: decision tree, pedometry, soil survey, mapping unit.

---

[1] Part of the Doctoral Thesis of the first author. Received for publication on August 8, 2012 and approved on February 26, 2013.

[2] Adjunct professor, Universidade Federal de Santa Catarina, Campus Curitibanos - Rod. Ulysses Gaboardi, km 3. Caixa Postal 101. CEP 89520-000 Curitibanos (SC). E-mail: alexandre.ten.caten@ufsc.br

[3] Associate professor, Universidade Federal de Santa Maria - UFSM, Prédio 42. Av. Roraima 1000, Camobi, CEP 97105-900 Santa Maria (RS). E-mail: dalmolin@ufsm.br

[4] Adjunct professor, UFSM. E-mail: fapedron@ymail.com

[5] Undergraduate student, UFSM. E-mail: ruiz.ch@gmail.com

[6] Professor, Universidade Regional Integrada Campus Erechim. Av. Sete de Setembro, 1621. Caixa Postal 743. CEP 99700-000 Erechim (RS). E-mail: scarlos@uri.com.br

RESUMO: *VOLUME DE DADOS ADEQUADO PARA O MAPEAMENTO DIGITAL DE SOLOS NO MUNICÍPIO DE ERECHIM, RIO GRANDE DO SUL, BRASIL*

Informações digitais tornam possível um elevado grau de redundância das informações disponíveis para o ajuste de modelos preditores aplicados ao Mapeamento Digital de Solos (MDS). Entre esses modelos, a técnica de Árvores de Decisão (AD) tem aplicação crescente, em razão da sua potência no tratamento de grandes volumes de dados. Objetivou-se com este trabalho avaliar o impacto do volume de dados utilizados para gerar os modelos por AD, na qualidade dos mapas de solos gerados pela técnica de MDS. Uma área de estudo com 889,33 km² foi escolhida na região do Planalto Médio do Rio Grande do Sul. As relações solo-paisagem foram obtidas a partir de reambulação da área de estudo e delineamento das unidades de mapeamento em cartas topográficas de escala 1:50.000. Seis covariáveis preditoras ligadas aos fatores de formação do solo, relevo e organismos, juntamente com os conjuntos de dados de um, três, cinco, 10, 15, 20 e 25 % do volume total de dados, foram usadas para gerar os modelos preditivos por AD no programa WEKA. Neste estudo, densidades de amostragem menores do que 5 % resultaram em modelos com menor poder de capturar a complexidade da distribuição espacial do solo da área estudada. Amostragens entre cinco e 15 % conduziram a uma melhor relação entre o volume de dados a ser manipulado e a capacidade preditiva dos modelos gerados. Dados coletados no campo indicaram acurácia dos mapas preditos próxima a 70 %, para os modelos oriundos dessas densidades de amostragem.

Termos de indexação: árvore de decisão, pedometria, levantamento de solos, unidade de mapeamento.

INTRODUCTION

Technological advances in areas such as remote sensing, computer processing speed, ability to handle large volumes of data, quantitative methods to describe spatial patterns and three-dimensional visualization have created new opportunities for understanding soil processes and properties (Grunwald, 2009). Digital Soil Mapping (DSM) in soil science is the computer-assisted production of digital maps of soil types and soil properties. DSM involves the creation and population of spatial soil information through the use of field and laboratory observational methods, coupled with spatial and non-spatial soil inference systems (Lagacherie & McBratney, 2007).

The application of digital information to DSM enables a high level of redundancy for fitting the models. On the other hand, the volume of information available will require the handling and processing of large volumes of data (McBratney et al., 2003). The search for an improved relationship between the sampling density in the original data bank and the predictive accuracy of the models has spurred research in this direction.

Among the data mining methods applied to the DSM, the Decision Tree (DT) technique is being increasingly applied, due to its ability of dealing with large data volumes (Witten & Frank, 2005). The DT approach has been used for the advantage of allowing an explicit expression of the soil-landscape relationship (Kheir et al., 2010a). DTs have been used in studies related to soil surface and underground erosion (Geissen et al., 2007), to prediction of soil classes (Giasson et al., 2011) and to spatialization of soil properties (Lemercier et al., 2012).

In addition to allowing grouping and the search for patterns, the DT permits an understanding of how these data are inter-related (Kheir et al., 2010b). The DT does not require an *a priori* specification of the form of the model that will be fitted to the data; moreover, as the DT model is constructed, it can be converted to algorithms that can easily be implemented by programming language (Kheir et al., 2010a).

DT models were used by Scull et al. (2005) with 39,877 samples for prediction of soil types in a 2,590 km² desert area, corresponding to approximately one percent of the total available data. In a study carried out by Giasson et al. (2011), 1,333 samples were used for the prediction of soil classes in an area of approximately 6.1 km², the sample density corresponded to less than one percent of the original data volume. Qi & Zhu (2003) sampled around 12 % of the original data in an area of 4 km² for fitting the models by DT.

In view of the availability of databases with a growing number of observations and the diversity of sampling densities reported in studies, the purpose of this study was to evaluate the effect of the proportion of the original data set used on the quality of soil type mapping generated by the decision tree technique.

MATERIAL AND METHODS

Area of study

The study area was defined by the polygon of the municipality of Erechim, in the Northern region of the State of Rio Grande do Sul (52º 06' - 52º 24'

longitude W, 27º 28' - 27º 47' latitude S, SIRGAS 2000). In the area of 889.33 km², altitudes range from 423 to 900 m (Figure 1). The region was chosen for its lack of soil spatial information with soil maps at recognition and exploratory level only.

With a view to generate DTs, the soil-landscape relationships were obtained from reambulation of the area under study with information obtained in the field by an experienced soil scientist. Soil samples were collected for laboratory analysis in seven profiles, and sets of auger holes were drilled throughout the study area. This information permitted outlining the mapping units on 1:50,000 scale topographic maps and defining soil classes of significant occurrence, according to the Brazilian System of Soil Classification - SiBCS (Embrapa, 2006) (Table 1). This information, together with the environmental covariates was used as data for the DT training.

Environmental covariates

In this study, covariates of the 'scorpan' model (McBratney et al, 2003) were used, related to the soil formation factors of relief (*r*) and organisms (*o*). Other covariates of the model were not available or the spatial resolution was insufficient for this study. The "organisms" factor represented the covariate Normalized Difference Vegetation Index (NDVI) of October (Oct.) (NOUT) and the standard deviation of the NDVI (STAN). The choice of Oct for the NDVI was because the contrast between the different uses of the land in the region is greatest in this month. For the calculation of both covariates, the images used were all from 2004 due to the greater availability of cloudless images in that year. All NDVIs were calculated according to Jensen (2009).

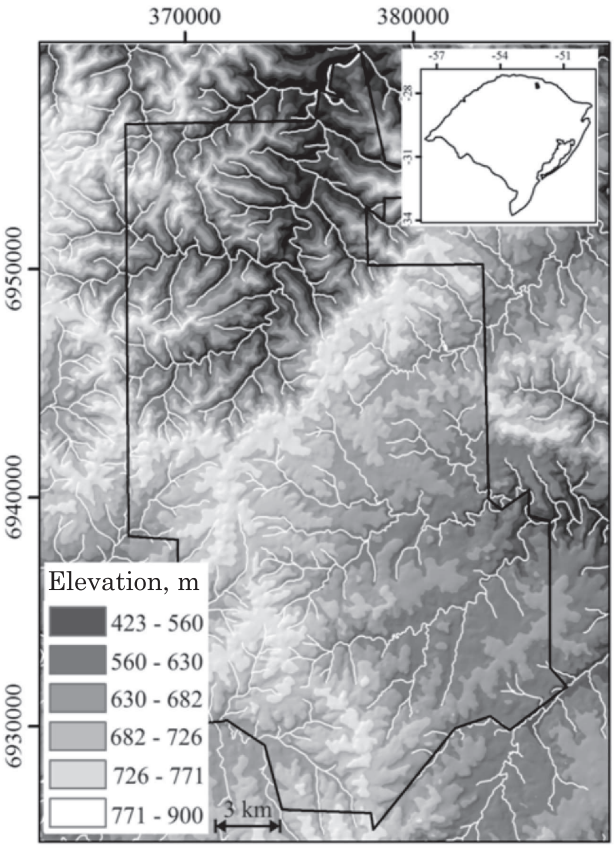

Figure 1. Drainage network and hypsometric map of the study area overlaid by the perimeter of the municipality of Erechim. The inset in the upper right corner shows the study area in the State of Rio Grande do Sul.

The STAN predictor was generated from the data of eight dates in 2004 obtained by the Landsat 5 sensor TM platform with a spatial resolution of 30 m. The NDVI value was calculated individually for each date. Then, the standard deviation of the NDVI value between the eight dates was computed for each pixel in ArcGIS (ESRI, 2008), in the Raster Calculator function of the program. The standard deviation was calculated as follows:

$$STAN = \sqrt{\frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(NDVI_i - \overline{NDVI}\right)^2}$$

Where STAN = standard deviation of NDVI; n = number of different dates used to generate STAN; NDV_i = from 1st to 8th NDVI; and $\overline{NDVI}$ = average value of NDVI.

The relief factor was represented by the elevation (ELEV), vertical distance of the drainage system (DIST), slope (SLOP), topographic wetness index (TWI), sediment transport capacity (STC), planar curvature (PLAN), profile curvature (PROF), and rugosity (RUGO) covariates. The covariates were generated according to Wilson & Gallant (2000) from a digital elevation model (DEM). The DEM with a spatial resolution of 30 m was derived from contour lines on topographic maps on a 1:50,000 scale. The interpolation of the contour lines to the raster format occurred in the *Topo do Raster* tool of the ArcGIS 9.3 program (ESRI, 2008), using the spline technique. In the SAGA-GIS program (Olaya, 2004), the attributes ELEV, DIST, SLOP, PLAN, PROF, and RUGO were generated by the standard terrain analysis tool and the TWI and STC attributes by the grid calculator tool.

Data sets corresponding to the sampling densities of 1, 3, 5, 10, 15, 20 and 25 % of the total area were tabulated for DT modeling. They were extracted by random sampling of the training soil map from the 10 layers of environmental covariates with SAGA-GIS. All soil classes were sampled proportionally to their area, according to the adopted sampling method.

Decision trees

The DTs were developed using the data mining program WEKA (Hall et al., 2009). For data processing, the J48 algorithm was used since it had the best performance in a study carried out by Giasson

Table 1. Soil Mapping Units (MU) surveyed in the study area of the municipality of Erechim (RS)

Symbol	SiBCS (Embrapa, 2006)/Soil Taxonomy	Area	
		ha	%
MU1	Gleissolo Háplico/Typic Endoaquept	31.28	12.52
MU2	Latossolo Vermelho/Rhodic Hapludox	122.99	49.22
MU3	Associação Neossolo Litólico e Cambissolo Hálico/Lithic Udorthent - Typic Dystrudept association	18.27	7.31
MU4	Neossolo Litólico/Lithic Udorthent	26.26	10.51
MU5	Associação Nitossolo Vermelho e Chernossolo Argilúvico fase altitude elevada/Typic Rhodudalf - Typic Argiudoll higher altitude phase association	26.01	10.41
MU6	Associação Nitossolo Vermelho e Chernossolo Argilúvico fase altitude baixa/Typic Rhodudalf - Typic Argiudoll lower altitude phase association	18.03	7.22
MU7	Associação Nitossolo Vermelho e Cambissolo Háplico/Typic Rhodudalf - Typic Dystrudept association	7.05	2.82

et al. (2011). The minimum number of observations (*minNumObj* on WEKA) per leaf for each data set was determined after an analysis of the percentage of misclassified observations. This analysis was performed using the 23 values of 2; 5; 10; 50; 75; 80; 95; 100; 150; 200; 250; 300; 400; 500; 600; 700; 1,000; 1,250; 1,500; 1,750; 2,000; 2,250; and 2,500 for the minimum number of observations per leaf in each of the seven data sets. The pruning method, which mitigated the error in the generated DT, was also selected (*reducedErrorPruning* = *True* on WEKA). During DT generation, an independent data set of 5 % of the total area was used (*Supplied test set* on WEKA) to verify the quality of the generated tree.

Soil map

The DTs with the best ratio between the model complexity and the number of misclassified observations were implemented in the ArcGIS program, in the raster calculator function. The information derived from the tree was converted by the conditional function 'con(condition, if true, if false)' of the program. By this function, the raster files of the environmental covariates can be processed according to the set of rules derived from the DT. Since the intended publication scale of each created soil map was 1:50,000, isolated pixel regions with a minimum mappable area of less than 1 ha were merged with a neighboring polygon.

Quality of models and maps

The quality of the DT models was evaluated based on the percentage value of misclassified observations across the tree. This value is one of the outputs of the WEKA program once the created model is validated using the set of 5 % of the reserved data. To calculate this, the program adds up all misclassified observations and divides them by the total observations of the testing set, multiplying the result by 100 to obtain the percentage value (Hall et al., 2009).

The precision of the fitted model was evaluated by the set of correctly classified soil classes and divided by the sum of the correctly classified and misclassified sets of map units. This value is then multiplied by 100 to obtain a relative value (Hall et al., 2009).

The kappa index was used to verify the quality of information contained in the map (Hengl et al., 2007). The error matrix was established to calculate the kappa index according to Congalton (1991), from 50 observations per map unit obtained from field reambulations for identification of local soils. In all, 350 sites of occurrence of the mapped soil types were identified in this study.

RESULTS AND DISCUSSION

Decision Tree Models

The application of the seven data sample sets in the WEKA program, configured to evaluate the effect of 23 different groups of observations in the terminal nodes (*minNumObj* on WEKA), resulted in the development of 161 different DT models. The DTs in which a large number of observations were gathered in a single terminal node gave rise to more simplified models, but with less predictive power, since observations with a lower degree of similarity were united in the same terminal node. Figure 2 illustrates a DT derived from the data set with a sampling density of 10 %.

To generate the DT of the figure 2, the WEKA program was set to collect up to 1500 observations in the terminal nodes (*minNumObj* = 1500). It was then possible to generate a less complex model with only 12 nodes, which led to 23.83 % incorrectly classified observations. When the same data set was used, but the program set to gather a maximum of 10 observations per terminal node (*minNumObj* = 10), a

DT with 475 nodes was formed, with a percentage of misclassified observations of 13.22 %. This model had greater predictive power but was also more complex.

Among the predictors applied to describe the soil-landscape relationship of the study area, those used to define the main nodes of the trees were: slope, elevation and vertical distance of the drainage system (Figure 2). This is related to the predominant soil-landscape relationship in the study region, where relief plays a significant role. The pattern of predominance of these three predictors was also observed in other generated DT models, confirming the importance of relief for soil formation in the area under study.

The DTs with most nodes overfitted the data and maximized the predictive power (Figure 3). During the development of the models by DT, it is therefore necessary to evaluate the predictive gain of the models to the extent that they become more complex. For the seven sets of samples, it was verified that to the extent that the DT become more complex with a greater number of leaves, the fewer the number of observations misclassified by the models. However, even for very complex trees and higher volumes of data to fit the models, the percentage of misclassified observations remains around 13 % (Figure 3). Possibly, this arises from the impossibility of explaining the complexity of spatial distribution of the soils in the region with only the six covariates utilized. Scull et al. (2005) applied DTs for soil classification at the order level and observed that the model misclassified about 18.52 % of the observations located in the training area and 24.34 % of those located in the validation area.

For the data sets of 1 and 3 %, even with a smaller number of observations per leaf in more complex trees, no results better than 16.5 and 15 %, respectively, were achieved. One of the effects of utilizing a small number of samples for DT generation may be the non-representative nature in relation to the total available data set, reducing the predictive power.

As the DTs are pruned, the number of misclassified observations increases (Figure 3). Nevertheless, above a certain number of leaves, the increase rate in errors changes significantly and begins to increase linearly for DTs with less than 65 leaves (vertical dashed line). This behavior was also observed in the smaller data sets (1 or 3 %), although less pronounced.

To use models with a combination of simplicity and predictive power in the different data sets, the option was made to use a lower limit of 65 leaves in the DTs. With this value, DTs were generated for the data sets of 5, 10, 15, 20 and 25 %. The data sets of 1 and 3 % were discarded for presenting the worst results in terms of number of misclassified observation in the region of the DT with 65 leaves. Grinand et al. (2008) stated that sample volumes below 10 % can lead to a substantial reduction in the model quality.

The results for model development (Table 2) indicate a tendency for a small decrease in the number of misclassified observations as the data sets to generate

```
DIST <= 5.7655
|  ELEV <= 597.5438: MU6
|  ELEV > 597.5438
|  |  ELEV <= 700.0499: MU1
|  |  ELEV > 700.0499: MU2
DIST > 5.7655
|  ELEV <= 659.9101
|  |  SLOP <= 0.211: MU3
|  |  SLOP > 0.211: MU4
|  ELEV > 659.9101
|  |  SLOP <= 0.2366
|  |  |  DIST <= 62.162: MU2
|  |  |  DIST > 62.162
|  |  |  |  ELEV <= 766.5282: MU7
|  |  |  |  ELEV > 766.5282
|  |  |  |  |  ELEV <= 813.9938
|  |  |  |  |  |  DIST <= 98.6077: MU2
|  |  |  |  |  |  DIST > 98.6077: MU5
|  |  |  |  |  ELEV > 813.9938: MU5
|  |  SLOP > 0.2366
|  |  |  ELEV <= 763.2329: MU4
|  |  |  ELEV > 763.2329: MU5
```

Figure 2. Decision tree generated by the WEKA program. Terminal leaves are associated with one of the seven mapping units. Mapping Unit (MU), elevation (ELEV), vertical distance of the drainage system (DIST) and slope (SLOP).

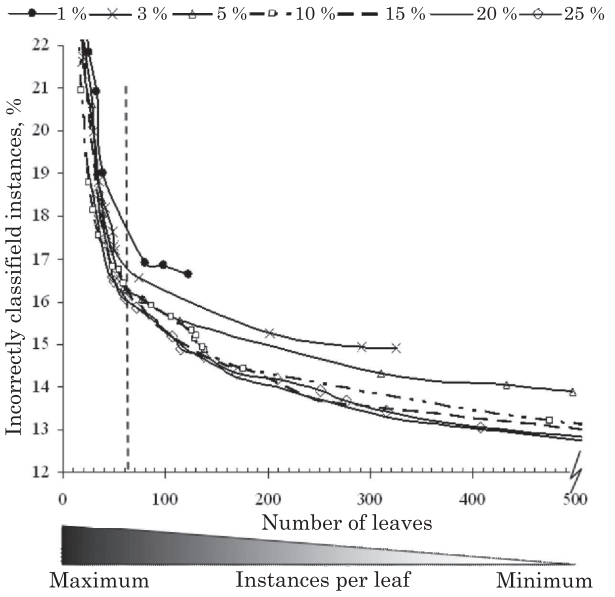

Figure 3. Relationship between the number of leaves in the decision tree and the percentage of misclassified samples. The vertical dashed line indicates the abrupt increase in the number of incorrectly classified instances.

Table 2. Results for model development and testing of the model predictive power

Data set	Model development[1]			Quality of the predictive model[2]				
	Incorrectly classified instance	Kappa	Precision	5	10	15	20	25
5	16.27	76.50	83.00	-	-	-	-	-
10	16.23	76.80	83.20	79.4	-	-	-	-
15	16.14	76.94	83.30	78.4	81.1	-	-	-
20	15.79	77.36	83.70	81.1	84.1	84.1	-	-
25	16.02	77.06	83.00	82.4	85.3	85.6	90.4	-
Training	-	-	-	77.3	77.3	77.1	77.7	77.6
field	-	-	-	69.7	68.5	71.1	71.4	71.1

[1] Output information of the WEKA program after testing the five models in the independent data set (supplied test set). [2] Kappa index between the maps generated with different sampling densities, the training area and the 350 field samples. All values in percentage.

the models increased. The fact that the number of incorrect observations remained around 16 % for all five data sets is possibly due to the quality and quantity of the predictor covariates used in this study, which were incapable of predicting the total complexity of spatial distribution of the soils in the study area. The values of the Kappa index and the precision of the models generated by the WEKA program were also significantly improved as a higher data volume was used to generate the models (Table 2). However, these results indicate that an improvement in the fitting of the models generated by DT is associated with other factors, such as the use of a larger quantity and diversity of predictor covariates, and not only the use of a larger data volume for model fitting.

In a study carried out by Giasson et al. (2011), the authors identified a kappa index of 57.1 % for one DT with 79 leaves, used to map six soil classes. According to these authors, oversimplified DTs implied in a reduction of predictive power of the models. In that study, a DT with only five leaves was evaluated, which resulted in a kappa index of approximately 43.9 % and the prediction of only two soil classes.

Soil maps

Five DTs were used to generate soil maps of the municipality of Erechim (Figure 4). All five soil maps represented the soil-landscape relationship very similarly to those established by pedologists in a field reambulation. Rhodic Hapludox soils were found on slopes at higher altitudes (Figure 4). At these locations, the relief conditions made the formation of these deep, homogeneous and weathered soils possible.

Combinations of Lithic Udorthent and Typic Dystrudept were found at lower positions of the toposequence (Figure 4). Such associations can be observed throughout the upper part of the Rio Dourado Valley. At these locations, slopes make a constant removal of material possible. Consequently, soils tend to be shallow and rocky. Lithic Udorthent soils were situated predominantly on the longer and steeper slopes (Figure 4). The association of Typic Rhodudalf with Typic Argiudoll at low altitude locations occurred in the valleys along the drainage network. At these sites, flat areas made the formation of deeper and more developed soils possible (Figure 4).

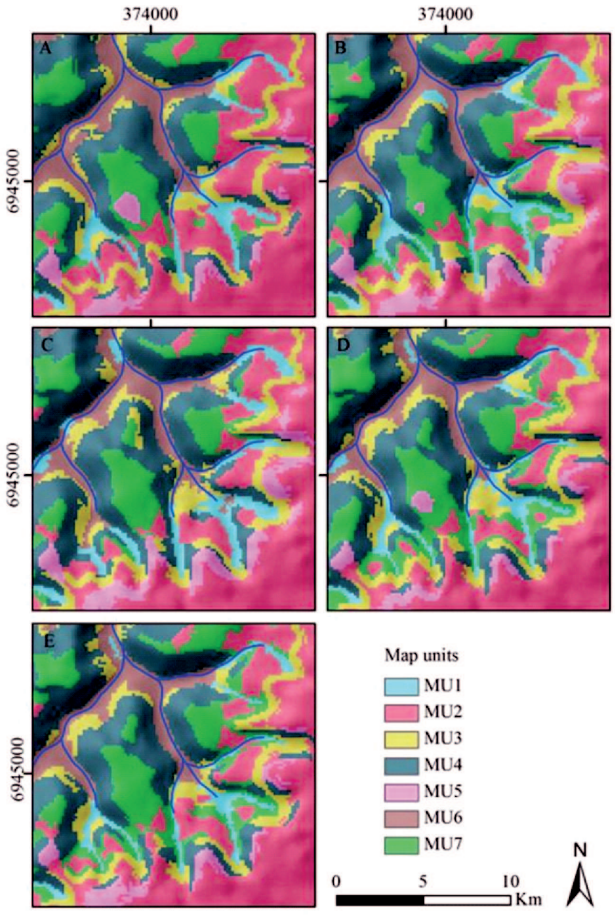

Figure 4. Extracts of the soil maps predicted from five sets of samples: (a) 5 %, (b) 10 %, (c) 15 %, (d) 20 %, (e) 25 %. Coordinate System SIRGAS2000 UTM zone 22.

A visual analysis of the five maps provided insufficient information to determine the most adequate data density. Subsequently, a field evaluation of the predictive capacity of each one of the five DTs generated was carried out.

Quality of the soil maps

From the five soil maps and the field data, it was possible to evaluate the quality of the predictive models (Table 2). First of all, it was decided to evaluate the similarity among the maps by comparison. It was observed that as the data volume increased, the higher the agreement between the generated maps. When the maps with data volumes of 5 and 10 % were compared, the kappa index was lower than the index obtained from the comparison between the maps with data volumes of 20 and 25 %. This result is an indication that larger volumes of data allow a better comprehension of the complexity to be mapped, although the similarities between the maps prepared with a lower data volume in this study were also high.

The reproducibility of the five DT models employed in this study can be visualized in the sixth row of table 2. All models tested reproduced more than 77 % of the soil types correctly, as defined by pedologists in the model training areas. This value is greater than the 46.12 % found by ten Caten et al. (2011), who used multiple logistic regressions as predictive models.

Despite the high reproducibility in this study, the models diverged from the soil types observed in the field in approximately 23 % of the training area. This fact can be associated with the errors in the soil description in the field or in the inability of the models to capture the inherent complexity of local spatial distribution of the soils.

The accuracy of the predictive models is represented in the last line of table 2. All models permitted a mapping accuracy of about 70 % of the soil classes. As expected, the accuracy of the predictive models was slightly lower than the reproducibility of the models mentioned in the paragraphs above. However, both quality indicators in the mapping showed that the five tested models mapped the soil types satisfactorily.

With regard to the effect of the volume of data used to generate the DTs, there is a tendency for better performance of the predictive models when the data volume to train models is greater (Table 2). The best predictive performance was achieved by data volumes of 15, 20 and 25 %. A study of Grinand et al. (2008) tested samples of 10 to 90 % of the original data. The authors verified large variations in accuracy from 10 to 20 % of the original data, with no significant increase in the model quality above 30 % of the total sample volume. According to the results of this study, a data volume of 15 % would combine the features of mapping with a smaller data volume, maintaining reproducibility and accuracy.

On the topographical maps of Dois Córregos (SP), the soil classes were mapped by Crivelenti et al. (2009) with a data volume of 714,000 samples in a total area of 772 km². Although an exact determination of the data volume used to generate the decision trees in that study was not possible, it is believed that more than 80 % of the original data were used to generate the models. The authors reported a kappa index of 43 % in relation to the map used to generate the model.

A study carried out by Bui & Moran (2003) demonstrated that the predictive capacity of the models underwent a significant increase above the utilization of 10 % of the original data. The quality of the predictive models had a maximum value when the samples used represented 25 % of the total data, with a kappa index of 64 % in relation to the original map. However, the authors confirmed that a lower sampling density (of 10 %), would be enough to capture much of the diversity contained in the data for construction of the predictive models by DT.

CONCLUSION

This study demonstrated that unrepresentative data sets (below 5 %) for the generation of Decision Tree models are unable to capture the complexity of the spatial soil distribution. On the other hand, data sets > 20 % could result in excessive redundancy for the generation of predictive models, requiring the handling of unnecessary data that would fail to produce predictive gain by models derived from these sets. Data sets from 5 to 15 % resulted in the best ratio regarding the data volume to be handled and predictive capacity of the models generated. Reproducibility values of around 77 % and accuracy near 70 % were achieved with these sample volumes used to generate the models by decision trees.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) with resources for this research and the competitive research grant to the second author and financial support of the National Council for Scientific and Technological Development (CNPq).

The authors want to thank the anonymous reviewers in recognition for their contribution to this paper.

LITERATURE CITED

BUI, E.N. & MORAN, C.J. A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling basin of Australia. Geoderma, 111:21-44, 2003.

CRIVELENTI, R.C.; COELHO, R.M.; ADAMI, S.F. & OLIVEIRA, S.R.M. Data mining to infer soil-landscape relationships in digital soil mapping. Pesq. Agropec. Bras., 44:1707-1715, 2009.

CONGALTON, R.G. A review of assessing the accuracy of classification of remotely sensed data. Remote Sens. Environ., 37:35-46, 1991.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Centro Nacional de Pesquisa de Solos. Sistema brasileiro de classificação de solos. Brasília, Embrapa Produção de Informação; Rio de Janeiro, Embrapa Solos, 2006. 306p.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE - ESRI. ArcGIS Desktop. Inc. Redlands, 2008.

GIASSON, E.; SARMENTO, E.C.; WEBER, E.; FLORES, C.A. & HASENACK, H. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci. Agríc., 68:167-174, 2011.

GEISSEN, V.; KAMPICHLER, C.; LOPEZ-DE LLERGO-JUAREZ, J.J. & GALINDO-ACANTARA, A. Superficial and subterranean soil erosion in Tabasco, tropical Mexico: Development of a decision tree modeling approach. Geoderma, 139:277-287, 2007.

GRINAND, C.; ARROUAYS, D.; LAROCHE, B. & MARTIN, M. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma, 143:180-190, 2008.

GRUNWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma, 152:195-207, 2009.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. & WITTEN, I.H. The WEKA Data Mining Software: An Update. 2009. (SIGKDD Explorations, 11)

HENGL, T.; TOOMANIAN, N.; REUTER, H.I. & MALAKOUTI, M.J. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma, 140:417-427, 2007.

JENSEN, J.R. Sensoriamento remoto do ambiente: Uma perspectiva em recursos terrestres. São José dos Campos, Parêntese, 2009. 598p.

KHEIR, R.B.; GREVE, M.H.; BØCHER, P.K.; GREVE, M.B.; LARSEN, R. & MCCLOY, K. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. J. Environ. Manage., 91:1150-1160, 2010a.

KHEIR, R.B.; GREVE, M.H.; ABDALLAH, C. & DALGAARD, T. Spatial soil zinc content distribution from terrain parameters: A GIS-based decision-tree model in Lebanon. Environ. Pollut., 158:520-528, 2010b.

LAGACHERIE, P. & McBRATNEY, A.B. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: LAGACHERIE, P.; McBRATNEY, A. & VOLTZ, M., eds. Digital soil mapping: An introductory perspective. Amsterdam, Elsevier, 2007. p.3-22.

LEMERCIER, B.; LACOSTE, M.; LOUM, M. & WALTER, C. Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. Geoderma, 171-172: 75-84, 2012.

McBRATNEY, A.B.; MENDONÇA-SANTOS, M.L. & MINASNY, B. On digital soil mapping. Geoderma, 117:3-52, 2003.

OLAYA, V. A gentle introduction to SAGA GIS. The SAGA user group. Göttingen, Niedersachsen, 2004. 216p.

QI, F. & ZHU, A.X. Knowledge discovery from soil maps using inductive learning. Inter. J. Geogr. Inf. Sci., 17:771-795, 2003.

SCULL, P.; FRANKLIN, J. & CHADWICK, O.A. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model., 181:1-15, 2005.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A. & MENDONÇA-SANTOS, M.L. Multivariate analysis applied to reduce the number of predictors in digital soil mapping. Pesq. Agropec. Bras., 46:554-562, 2011.

WILSON, J.P. & GALLANT, J.C. Digital terrain analysis. In: WILSON, J.P. & GALLANT, J.C., eds. Terrain analysis: Principles and applications. New York, Wiley & Sons, 2000. p.1-27.

WITTEN, I.H. & FRANK, E. Data mining: Practical machine learning tools and techniques. 2.ed. San Francisco, Morgan Kaufmann, 2005. 629p.