Revista Brasileira de Ciência do Solo
ISSN: 0100-0683
revista@sbcs.org.br
Sociedade Brasileira de Ciência do Solo
Brasil

Comissão 1.2 - Levantamento e classificação do solo

DIGITAL SOIL MAPPING: STRATEGY FOR DATA PRE-PROCESSING[1]

Alexandre ten Caten[2], Ricardo Simão Diniz Dalmolin[3] & Luis Fernando Chimelo Ruiz[4]

SUMMARY

The region of greatest variability on soil maps is along the edge of their polygons, causing disagreement among pedologists about the appropriate description of soil classes at these locations. The objective of this work was to propose a strategy for data pre-processing applied to digital soil mapping (DSM). Soil polygons on a training map were shrunk by 100 and 160 m. This strategy prevented the use of covariates located near the edge of the soil classes for the Decision Tree (DT) models. Three DT models derived from eight predictive covariates, related to relief and organism factors sampled on the original polygons of a soil map and on polygons shrunk by 100 and 160 m were used to predict soil classes. The DT model derived from observations 160 m away from the edge of the polygons on the original map is less complex and has a better predictive performance.

Index terms: choropleth map, pedometrics, soil survey, decision tree.

RESUMO: *MAPA DIGITAL DE SOLOS: ESTRATÉGIAS PARA PROCESSAMENTO DE DADOS*

Mapas de solos têm na borda dos polígonos a região de maior variabilidade, o que leva os pedólogos a divergir quanto ao delineamento das classes de solos nesses locais. O objetivo deste estudo foi propor uma estratégia de pré-processamento de dados aplicada ao mapeamento digital de solos. Polígonos de solos em um mapa de treinamento foram deslocados para seu interior em 100 e 160 m. Essa estratégia possibilitou que covariáveis localizadas próximas à borda das classes de solos não fossem utilizadas na geração dos modelos de Árvore de Decisão

[1] Received for publication in June 20, 2011 and approved in May 10, 2012.

[2] Adjunct Professor at Universidade Federal de Santa Catarina campus Curitibanos, Rod. Ulysses Gaboardi, Km 3, Caixa Postal 101, Curitibanos (SC), CEP 89.520-000. E-mail: alexandre.ten.caten@ufsc.br

[3] Associate Professor of the Soil Deparment, Universidade Federal de Santa Maria - UFSM. Av. Roraima 1000, Camobi, CEP. 97105-900 Santa Maria (RS). E-mail: dalmolin@pq.cnpq.br

[4] Student of Geoprocessing Technology, Colégio Politécnico da UFSM, Av. Roraima 1000, Camobi, CEP. 97105-900 Santa Maria (RS). E-mail: ruiz.ch@gmail.com

(AD). Três ADs geradas a partir de oito covariáveis preditoras, ligadas aos fatores relevo e organismos, amostradas nos polígonos originais de um mapa de solos e em polígonos deslocados em 100 e 160 m para o seu interior, foram utilizadas para predizer classes de solos. O modelo de AD a partir de observações distantes 160 m da borda dos polígonos no mapa original é menos complexo e tem melhor desempenho preditivo.

Termos de indexação: Mapa coroplético, pedometria, levantamento de solos, árvore de decisão.

INTRODUCTION

Soil maps can be predicted from pre-existing soil information. The 'scorpan' model (McBratney et al., 2003) assumes that existing information about soil classes and properties can assist in the prediction and digital soil mapping (DSM) in areas where spatial information of soils is missing or unavailable at the scale required. According to Qi & Zhu (2003), the basic idea of formalization of the soil-landscape relationships contained in choropleth soil maps consists of recovering pedological knowledge contained in the map by application of data mining techniques. The information of the soil polygons on choropleth maps has proven to be applicable to DSM (Crivelenti et al., 2009; Giasson et al., 2011; ten Caten et al., 2011a).

In conventional soil mapping, the map units are delineated by visual interpretation of stereoscopic aerial photo pairs at a scale compatible with the objective (Dalmolin et al., 2004). The spatial position of the soil polygons implies knowledge of relationships among different soil classes and the environmental conditions present in the landscape. When delineating the boundaries of the polygons, pedologist are guided by their implicit knowledge of the relationships among the multiple information layers that reflect the local soil genesis, such geology, relief and land use (Qi & Zhu, 2003).

A certain degree of subjectivity is intrinsic to the conventional method, whether due to the scale of the basic material used, or due to the nature of spatial variation of the soil itself, where transitions between classes are not abrupt. For these reasons, tiny or gradual variations in environmental conditions are difficult to locate through the conventional method (Zhu et al., 2001), which raises an uncertainty with regard to the real location of limits between soil classes in the landscape. The use of these polygons as a reference for training of predictive models will imply the addition of deviant information. They, for their part, will have different effects on the predictive quality of the models (Qi, 2004).

In an attempt to improve the quality of the database used to predict soil classes, Qi & Zhu (2003) used only the information that was near the mode of each predictive covariate. Through the construction of a histogram for each covariate, the authors discarded the data outside the mode of the data set. The accuracy of the models developed by the decision tree in the set of filtered data reached 86 %, whereas models using the total data set obtained a mean value of 75 %. For these authors, the filtering method was effective as a strategy of pretreatment of data applied to the DSM. Nevertheless, Schmidt et al. (2008) warned that the histogram method has the disadvantage of requiring the construction of a histogram for each predictive covariate, creating difficulties in studies with a high number of covariates and observations.

Methods for selection of observations are used when a large number of samples are available. Strategies appropriate for selection of the observations lead to better results if compared to models which are adjusted with the total number of observations available (Schmidt et al.; 2008). The challenge is in "doing more with less" (Liu & Motoda, 2002), in other words, extracting part of the data set which should be representative of the original data and could be more easily handled by the algorithms. In this case, there may be a gain in accuracy and speed in data processing.

Studies in DSM seek to evaluate the quality of the covariates and of the observations aiming into achieving a higher accuracy of the predictions (Schmidt et al., 2008). Among the studies that consider the observations used in the DSM are those that seek to define the best sampling density (Moran & Bui, 2002; Scull et al., 2003; Grinand et al., 2008; Schmidt et al., 2008), although studies that are concerned with patterns of the observations within each soil class to be predicted are rare (Qi & Zhu, 2003; Qi, 2004).

Evaluating the application of multiple logistic regressions for prediction of soil classes, ten Caten et al. (2011b) observed the failure of the models to distinguish among nearby classes in the landscape. It was observed that the greatest percentages of prediction confounding occurred among the four distinct suborders of Haplic Acrisol, since these soil classes occupy very similar positions in the landscape. For the authors, this difficulty of the models may arise from the map design itself that served for training, since in the landscape there are no abrupt limits between the soil classes as represented on the original map (with choropleth polygons), or, due to very tiny differences among the terrain attributes (environmental covariates), which may present no type of gradient at all at the boundary of the soil class polygons.

The influence of the transition areas among different soil classes was also registered in a study of

Carvalho et al. (2009). According to these authors, due to the practice of cartography based on polygons (Boolean) adopted in the conventional soil mapping method, the application of fuzzy logic for digital soil mapping will imply in the appearance of areas that would be related to transition zones among two or more soil units, which would lead to the emergence of non-existent delineations on the original map used for model training.

Definition of the exact position of the boundaries of the polygons is a controversial matter among pedologists. Legros (2005) evaluated the quality of delineation of mapping units performed by 20 different pedologists in the same area. It was observed that the polygons tend to overlap. However, small deviations in delineation are induced by the perception of each pedologist of the information contained in the aerial photos. These deviations of delineation allow definition of a region of uncertainty with regard to the most adequate location for the boundary of the polygons. In this region of uncertainty, the contribution of the predictive covariates to the quality of the predictive models may be doubtful.

The objective of this work was to propose a strategy for pre-processing of data applied to the DSM, evaluating the effect of non-use of information derived from predictive covariates present at the boundaries of the soil polygons in decision tree models for prediction of soil classes. The sampling method proposed excludes deviant samples and selects observations that contain only the characteristics of the predictive covariates in each soil class.

MATERIAL AND METHODS

Soil map used

The semi-detailed soil survey at a scale of 1:50,000 by Klamt et al. (2001) of the municipality of São Pedro do Sul in the central region of the state of Rio Grande do Sul, with an extension of 874 km², was used as test map for this study. The soil classes at the suborder level contained on the test map were vectorized using ArcGIS 9.3 (ESRI, 2008).

The proposal of this study consists of displacing each soil class polygan present on the original map inwardly to the polygon defined by the pedologist. Therefore, the regions of greatest uncertainty with regard to the real position of the edges of the soil class polygons are not sampled for use of the predictive models. To check the effect of the method on the data set to be generated, the distances of 100 and 160 m were used in relation to the original position of each soil class (Figure 1). As all soil polygons are displaced inwardly, the width of the strip of effectively discarded data was 200 and 320 m on the edges of neighboring polygons. The values of 100 and 160 m were defined based on an analysis of gradients of values occurring in the predictive covariates near the edge by the polygons. The new polygons were created by the Buffer function of the program ArcGIS 9.3. The proposed sampling method did not prevent any soil class from being sampled by the reduction of the total area covered by each soil class.

Predictive covariates

In this study, covariates of the 'scorpan' model (McBratney et al., 2003) were used related to the soil relief (r) and organism (o) formation factors. The organism factor was represented by the standard deviation covariate of the Normalized Difference Vegetation Index (NDVI), hereinafter called STAN. The STAN covariate was used rather than the NDVI due to the fact that the latter presents different values throughout the year as a result of different agricultural uses.

To calculate the STAN covariate, images taken between February 2004 and January 2005 were used. This period was selected due to greater availability of images with absence of clouds in that year. The STAN predictor was generated using data of eight different dates of the period obtained by the platform Landsat 5 Thematic Mapper with a spatial resolution of 30 m. All NDVI calculations were performed according to

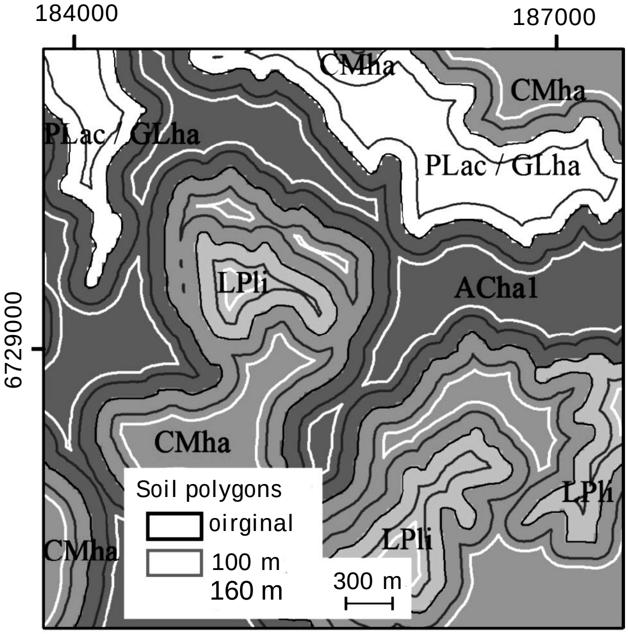

Figure 1. Sampling strategies of the predictive covariates adopted in the study using the original soil map, with a displacement of 100 and 160 m from the edge of the polygons. Acric Planosol (Albic, Epiarenic)/Haplic Gleysol (Dystric, Siltic) [PLac/GLha], Haplic Cambisol (Dystric, Chromic)[CMha], Lithic Leptosol (Humic, Eutric) [LPli] and Haplic Acrisol (Profondic, Chromic) [ACha1]. Coordinate system SIRGAS2000/UTM zone 22.

Jensen (2009). The NDVI value of each date was calculated individually and then the standard deviation of the NDVI value was computed among the eight dates for each pixel on the ArcGIS 9.3 program in the Raster Calculator function.

The relief factor was represented by the covariates: elevation (ELEV), slope (SLOP), topographic wetness index (TWI), sediment transport capacity (STC), plane curvature (PLAN), profile curvature (PROF), and terrain roughness index (TRI). The covariates were generated according to Wilson & Gallant (2000), based on a digital elevation model (DEM). The DEM (spatial resolution 30 m) was derived from topographic map contour lines at a 1:50,000 scale. The contour lines for the matrix format were interpolated by the Topo to Raster tool of the ArcGIS 9.3 program, using the spline technique (Wahba, 1990). In the SAGA-GIS program, the attributes ELEV, SLOP, PLAN, PROF, and TRI were derived by the standard terrain analysis tool, and TWI and STC with the grid calculator tool.

These attributes were tabulated from the soil polygons, creating three separate data sets to construct the decision tree models: original, 100 and 160 m.

Decision tree (DT)

Digital soil mapping uses the DT technique to split the set of original data into smaller blocks containing increasingly homogeneous data. The data partitioning occurs in the form of a tree where a criterion of segregation among the data is tested at each node of the tree, creating new data sub-blocks. A no longer partitioned data set is a leaf of the tree (Giasson et al., 2011).

In the three data sets used to develop the models, all soil classes were sampled proportionally to their area. The DTs were developed on the data mining program WEKA 3.6.3 (Hall et al., 2009). For data processing, algorithm J48 was used, which presented the best results in a study of Giasson et al. (2011). The minimum number of observations per leaf (*minNumObj* - WEKA) for each data set was determined after analysis of the percentage of incorrectly classified observations. This analysis was performed with a series of values for the minimum number of observations in each one of the three data sets. The pruning method that mitigated the error on the derived tree was also selected (*reducedErrorPruning* = *True* - WEKA). During the phase of creating the tree, each data set was partitioned in 70 % for creation of the model and 30 % for validation of the tree (*Percentage split* - WEKA).

Soil map

After analysis of the complexity of the DT created and of the number of wrongly classified observations, the DTs were implemented in ArcGIS 9.3 with 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. This function allows the raster files of the environmental covariates to be processed according to the set of rules derived from the DT. Since the intended publication scale is 1:50,000, in each soil map created, isolated pixel regions with a minimum mappable area of less than one hectare were merged with a neighbor class.

Quality of models and maps

The quality of the DT models was evaluated based on the percentage of wrongly classified observations in the tree. This value is one of the outputs of the WEKA program after the model created was validated in the set of 30 % of the reserved data. For its calculation, the program adds up all wrongly classified observations and divides them by the total observations of the test set, multiplying the result by 100 to indicate the percentage value (Hall et al., 2009). The soil maps created based on the three distinct sets of data were compared to each other by calculating the Kappa coefficient. The Kappa coefficient is used to attest the quality of predictive mappings (Giasson et al., 2011). The error matrix for calculation of the Kappa coefficient was established according to Congalton (1991).

RESULTS AND DISCUSSION

The proposed pre-processing method resulted in a reduction of the total number of observations available for creating the models by a decision tree. In the data set derived from the original map, 100 % of the information used in this study is present. The data set created from displacement of the polygons by 100 m retained 60 % of the original data. In the event of displacement of 160 m of the soil polygons, 43 % of the original data were tabulated to create the decision tree. These data percentages were greater than the 30 % used by Grinand et al. (2008), and the 25 % used by Moran & Bui (2002) to adjust the decision trees applied to digital soil mapping.

The descriptive measurements of the data sets derived from Lithic Leptosol polygons displaced by 100 and 160 m, used as example here, were different from the original data set (Figure 2). Among the observations most distant from the central characteristics of each predictive covariate are the outliers. With the exception of the elevation, the covariates contained outliers in the data sampled from the original polygon. With the application of the shifting of the soil polygons, observations near the polygon edges were discarded and there were no distant values in the covariates. Pre-processing led to similar changes in the data pattern originated in the other soil classes (not represented here).

Visual analysis of the graphs indicates that the pre-processing did not change the data distribution. The position of the mean value in relation to the set of

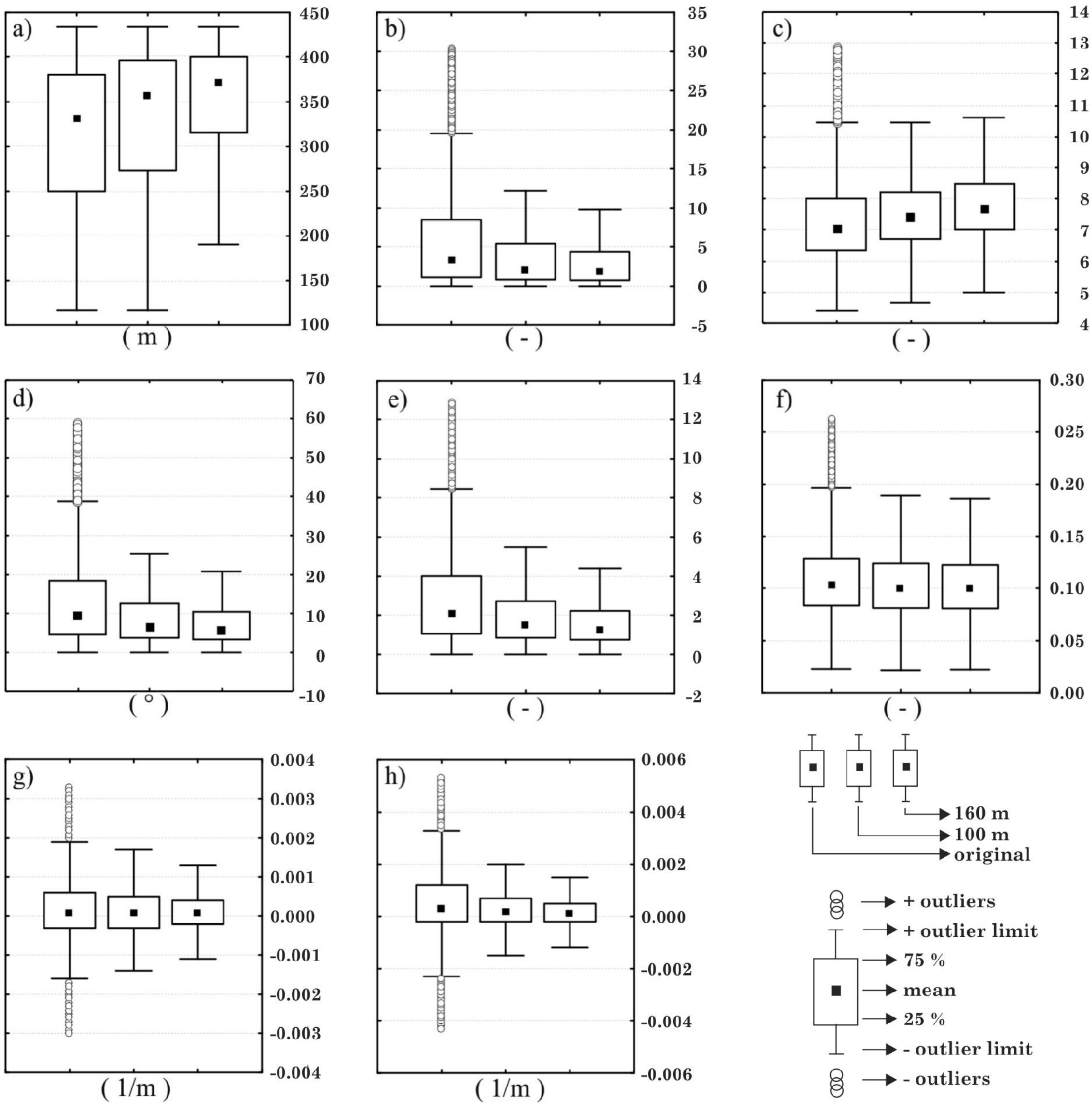

Figure 2. Boxplot of the predictive covariates created from the data sampled for the Lithic Leptosol class in the three situations of polygon positions. a) elevation (m), b) sediment transport capacity, c) topographic wetness index, d) slope (°), e) terrain roughness index, f) standard deviation of the NDVI, g) plane curvature (1/m) and h) profile curvature (1/m). Units of each covariate in parentheses. The outlier limit was calculated from the height of the central box of the boxplot multiplied by 1.5. Outliers were defined as the values beyond the outlier limit.

50 % of the total of data indicated that the data did not have normal distribution. This pattern was repeated for the data derived from the 100 and 160 m shifts, although data dispersion decreased; the distance of the data from the mean value was reduced. This performance is less noticeable in the data coming from the standard deviation covariate of the NDVI (Figure 2f). This may be attributed to the fact that in the study area, the Lithic Leptosol areas are situated on slopes covered by native forest. From the position of the polygons on the original map, even the polygons shifted by 160 m indicated no notable changes in land use at the location.

Decision Trees (DTs) obtained by the three sets of data have distinct performance with regard to the number of incorrect classified observations (Figure 3). The DT created from the total data set resulted in the classification of the observations with a minimum error of 38.5 %; in other words, the DT created from the totality of the samples of the test area allowed that in 61.5 % of the cases, the soil class present in the training set was attributed to the predictive map. The other data sets resulted in minimum errors of 32 and 28 % for displacements of 100 and 160 m, respectively. This indicates a benefit of data pre-processing for a better adjustment of the DT model. Nevertheless, even both trees adjusted to the data with less deviant observations misclassified around one third of the data. This may be due to the complexity of spatial distribution of soils at the location, not represented by the eight predictive covariates chosen for this study, or, moreover, it may be related to characteristics of the conventional map used for training of the DT, such as the scale and the arrangement of the polygons, which may have led to a generalization of the soil information.

In the three data sets, the error percentage was lowest with trees with a greater number of nodes and terminal leaves. The number of wrongly classified observations remained practically unaltered insofar as the program was allowed to group a greater number of observations in the terminal leaves. Nevertheless, near the number of 30 leaves, the three data sets showed an increasing tendency for the number of misclassified observations. It is believed that as of this limit, the tree is overly simplified and becomes incapable of predicting the complexity present in the data.

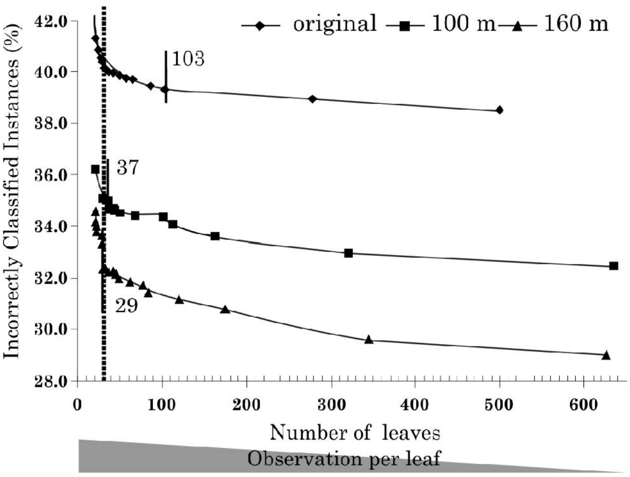

Figure 3. Relationship between the number of leaves and wrongly classified observations per decision tree. Sets of data without displacement (original) and with displacement of 100 and 160 m from the edge of the polygons. The numbers near the vertical bars indicate the minimum number of leaves to ensure that all soil classes would be predicted with that data set. The dashed vertical line indicates that the number of 30 leaves is the limit above which there is a sudden increase in the number of incorrectly classified observations in the three data sets.

The predictive power of the DTs created from the data distant from the edges of the soil polygons was greater. With regard to the minimum number of leaves so that the decision trees would still map the seven soil classes of the study area, it was observed that for the data derived from the 100 m shift, the minimum number of leaves was 37. When the displacement performed was 160 m, the minimum number of leaves was 29. On the other hand, with the use of the totality of observations of the study area, only with trees of greater complexity (above 103 leaves), it was possible to predict all soil classes. These results indicate that the presence of deviant observations demands more complex trees to capture the data pattern. In a study of Giasson et al. (2011), based on 1,333 randomly distributed observations, the authors observed that among the trees tested, only the more complex were capable of predicting all soil classes.

The results do not allow conclusions on the adequate displacement distance from the edges of the soil polygons. This distance will be a function of various factors, such as scale, complexity of the elements that control soil distribution, and the experience and ability of the pedologist that mapped the region used to create the models. Possibly in a single map, different distances could be adopted since the transitions of the soil classes present a range of complexity. Nevertheless, an analysis of our results (Figure 3) shows that the pre-processing by displacement of the soil polygons, avoiding use of the observations located in the regions of greatest uncertainty on the map, allows more precise and less complex DTs.

Soil maps were created by DT with a minimum number of 30 leaves based on the three sets of data (Figure 4). This number was adopted for being the value resulting in the best cost-benefit ratio between complexity and predictive power of the models (shown in Figure 3). Nevertheless, as previously indicated, this minimum number of leaves means that the trees created from the original set and shifted by 100 m do not predict all soil classes.

Visually, the Acric Planosol (Albic, Epiarenic)/ Haplic Gleysol (Dystric, Siltic) [PLac/GLha] were spatialized similarly to the training map. In the three predictive maps, the Haplic Acrisol (Sombric, Abruptic, Profondic) [ACha3] class was spatialized in an intermediate position on the landscape between the floodplain areas and the more elevated hills where the Haplic Acrisol (Profondic, Chromic) [ACha1] soils are located. This is in disagreement with the extract of the original map (Figure 4a), possibly due to the fact that in the region the characteristics are nearly indistinguishable for the perception of the pedologists, but nevertheless identified by the predictive model as being adequate for the formation of ACha3 at those locations. In this case, the predictive model is consistent with the known soil landscape relation for the study area, where the soils of the ACha3 class occupy a transition strip between the floodplain areas and the hills where the ACha1 class is found.

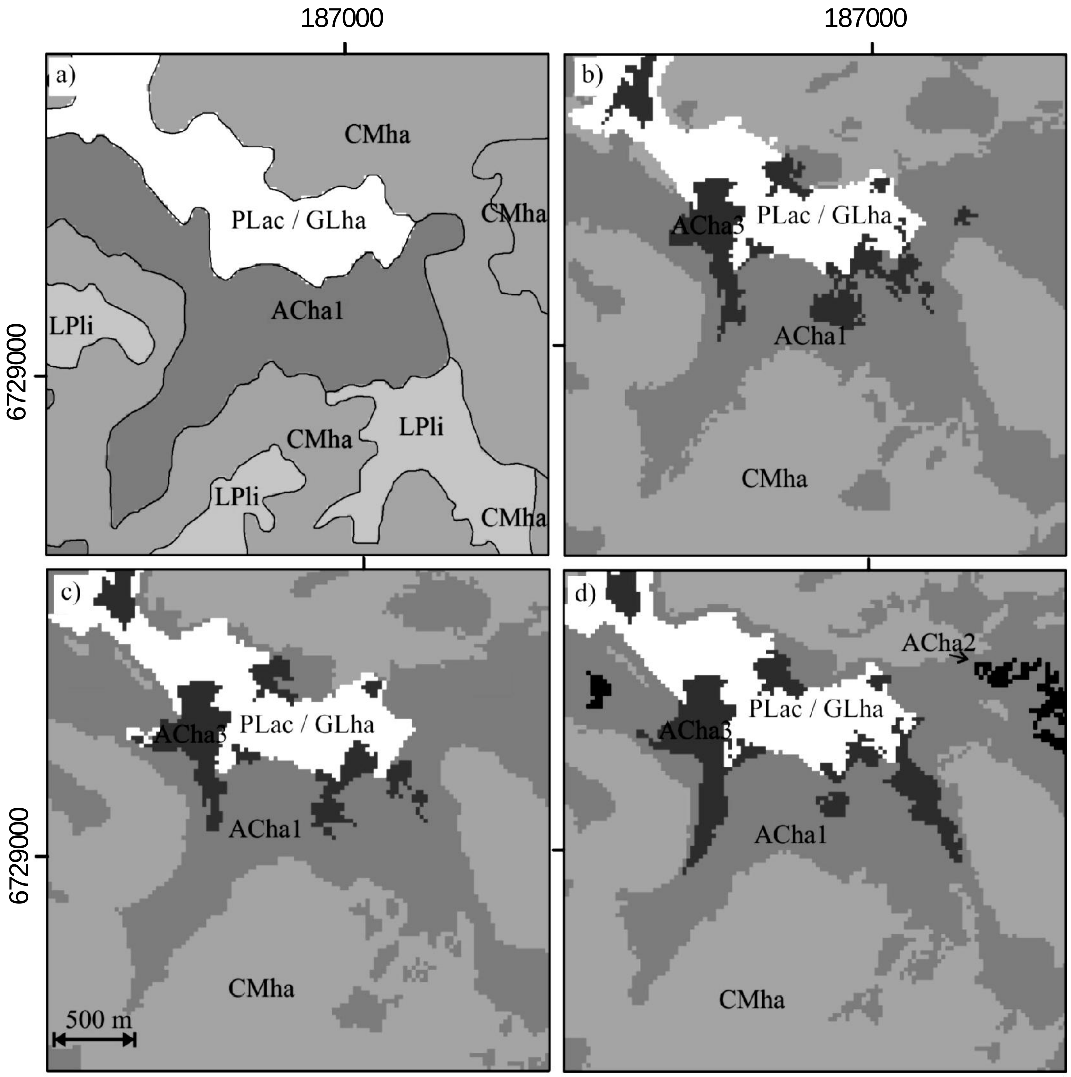

Figure 4. a) Extract of the soil map used for training; b) Predicted based on the totality of the data; c) Predicted based on displacement of 100 m in the polygons; d) Predicted based on displacement of 160 m in the polygons. Acric Planosol (Albic, Epiarenic)/Haplic Gleysol (Dystric, Siltic) [PLac/GLha], Haplic Cambisol (Dystric, Chromic)[CMha], Lithic Leptosol (Humic, Eutric) [LPli], Haplic Acrisol (Profondic, Chromic) [ACha1], Haplic Acrisol (Profondic, greyic)[ACha2] and Haplic Acrisol (Sombric, Abruptic, Profondic)[ACha3].Coordinate system SIRGAS2000/UTM zone 22.

The model created by the data set based on the shifting of 160 m allowed a spatialization of the Haplic Acrisol (Profondic, greyic) [ACha2] class (Figure 4d), although this class, present in other regions of the original map, is not located in the area of the extract delineated by the pedologist and used to show the results (Figure 4a). In the same way as for the ACha3 class, the use of an automated method for delimitation of the soil classes may have captured very small variations of the local environmental conditions, and, in accordance with the model created, those more favorable for the occurrence of the ACha2 class.

Comparisons of the three soil maps created and the original map allowed the observation of a degree of agreement among them. In approximately 60 % of the points (pixels) on the landscape, the same class existing on the original map was attributed by the models to the predictive maps (Table 1). This value is greater than the values found in the literature of 43 % (Crivelenti et al., 2009), 38.6 % (Coelho & Giasson, 2010) and 51.8 % (Giasson et al., 2011). In spite of a greater prediction quality obtained in this study in relation to the studies found in the literature, it may be observed that with the displacement of 100 m, the class ACha2 was not predicted. This probably occurred due to the fact that the limitation of a maximum number of 30 leaves is a very strong restriction in view of the complexity of the information in the data arising from positions on the landscape where ACha2 occurs. In this case, there is the need that, for the most disperse sets of data, more complex trees should be created for prediction of all soil classes.

Comparisons of the three soil maps indicated a high degree of similarity among them (Table 1). The three predicted maps were identical to each other in around 90 % of the points of the landscape, although it must be considered that one of the soil classes (ACha2) was not mapped by the DT, created by two of the data sets (original and 160 m). The Kappa coefficient of 59.34 % in relation to the original map was reached with an amount of data to be handled that was approximately 60 % less than the original volume, and with all classes present in the original map being predicted. If we consider that the pre-processing of the data resulted in a reduction of the volume of observations, the data set derived from displacement of 160 m at the edge of the polygons made the proposal of doing more with less viable (Liu & Motoda, 2002).

The Lithic Leptosol (Humic, Eutric) [LPli] class was not spatialized by the predictive models as it occurs in the extract area of the Figure 4a of the although it was predicted in other regions of the study area. The predictive potential of the models is related to the capacity of the environmental covariates used in representing the complexity that governs the spatial distribution of the soils in the landscape. In this study, the three models created did not adjust to 40, 35 and 33 % of the test set data used in the creation phase of the predictive models (Figure 3). This leads to a lack of prediction of determined classes where they possibly occur, or to their prediction in areas where they are not naturally present, as I the case of the ACha1 class (Figures 4b, c, d). Improvement of the predictive potential of the digital soil mapping technique will require research seeking predictive covariates by which a maximum of the soil formation factors that determine soil spatial distribution can be captured.

Pre-processing of the data presented here differs from the method used by Qi & Zhu (2003), in which the histogram and mode concept are used in two main aspects. The first is linked with the fact that the selection of the observations based on their relationship with the mode of a covariate may represent the exclusion of relevant information. Situations may occur in which a covariate presents multimodal characteristics in a single soil class such that it is difficult to determine how many and which are the modes present in the data. The other aspect is related to the volume of processing required; shifting the soil polygons is an automatic procedure, while building up and analyzing the histograms will be more time-consuming.

Table 1. Kappa coefficient indicating the degree of agreement among the maps created from three data sets of this study and the training map

	Training map	Original*	100 m*
Original*	60.64	-	-
100 m*	60.16	90.5	-
160 m	59.34	87.6	88.85

* The Haplic Acrisol (Profondic, greyic) [ACha2] soil class was not predicted.

CONCLUSIONS

1. Pre-processing the observations reduces the data volume to be handled in DSM.

2. Observations from the edges of soil polygons increase the number of misclassified observations by the decision tree.

3. Decision trees created from observations distant from the edges of the soil polygons are less complex and have greater predictive power in DSM; although in this study it was not possible to specify the minimum distance, results indicate that for environmental covariates with a pixel size of 30 m, a shift of 160 m from the polygon edges should be observed.

4. The soil map obtained by decision trees based on unbiased observations has greater similarity to the training map.

5. The method presented here is easily implemented and is more easily applied to DSM than the observation selection method based on histograms.

ACKNOWLEDGEMENTS

The authors are indebted to the National Council for Scientific and Technological Development (CNPq) for the award for Excellence in Research of the second author and for the financial support of this study.

LITERATURE CITED

CARVALHO, C.C.N.; FRANCA-ROCHA, W. & UCHA, J.M. Mapa digital de solos: Uma proposta metodológica usando inferência fuzzy. R. Bras. Eng. Agríc. Amb., 13:46-55, 2009.

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

COELHO, F.F. & GIASSON, E. Comparação de métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Ci. Rural, 40:2099-2106, 2010.

CRIVELENTI, R.C.; COELHO, R.M.; ADAMI, S.F. & OLIVEIRA, S.R.M. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. R. Agropec. Bras., 44:1707-1715, 2009.

DALMOLIN, R.S.D.; KLAMT, E.; PEDRON, F.A. & AZEVEDO, A.C. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. Ci. Rural, 34:1479-1486, 2004.

ESRI. Environmental Systems Research Institute. 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.

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.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. & WITTEN, I.H. The WEKA Data Mining Software: An Update. SIGKDD Explorations Newsletter, 11:10-18, 2009.

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

KLAMT, E.; FLORES, C.A. & CABRAL, D.R. Solos do Município de São Pedro do Sul. Santa Maria, Departamento de Solos/ CCR/UFSM, 2001. 96p.

LEGROS, J.P. Mapping of the soil. Enfield, Science Publisher, 2005. 411p.

LIU, H. & MOTODA, H. On issues of instance selection. Data Min. Knowl. Disc., 6:115-130, 2002.

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

MORAN, C.J. & BUI, E.N. Spatial data mining for enhanced soil map modeling. Inter. J. Geogr. Inf. Sci., 16:533-549, 2002.

QI, F. Knowledge discovery from area-class resource maps: Data preprocessing for noise reduction. Trans. GIS, 8:297-308, 2004.

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

SCHMIDT, K.; BEHRENS, T. & SCHOLTEN, T. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma, 146:138-146, 2008.

SCULL, P.; FRANKLIN, J.; CHADWICK, O.A. & McARTHUR, D. Predictive soil mapping: A review. Progr. Phys. Geog., 27:171-197, 2003.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A. & MENDONÇA-SANTOS, M.L. Extrapolação das relações solo-paisagem a partir de uma área de referência. Ci. Rural., 41:812-816, 2011a.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A. & MENDONÇA-SANTOS, M.L. Regressões logísticas múltiplas: Fatores que influenciam sua aplicação na predição de classes de solos. R. Bras. Ci. Solo, 35:53-62, 2011b.

ZHU, A.X.; HUDSON, B.; BURT, J.; LUBICH, K. & SIMONSON, D. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J., 65:1463-1472, 2001.

WAHBA, G. Spline models for observational data. In: CBMS-NSF. Regional Conference Series in Applied Mathematics. Philadelphia, Soc. Ind. Appl. Maths., 1990. v.59. 169p.

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.