PAB
Pesquisa Agropecuária Brasileira

ISSN 1678-3921

Journal homepage: www.embrapa.br/pab

For manuscript submission and journal contents, access: www.scielo.br/pab

Soil Science/ Original Article

Prediction of soil classes in a complex landscape in Southern Brazil

Abstract – The objective of this work was to evaluate the use of covariate selection by expert knowledge on the performance of soil class predictive models in a complex landscape, in order to identify the best predictive model for digital soil mapping in the Southern region of Brazil. A total of 164 points were sampled in the field using the conditioned Latin hypercube, considering the covariates elevation, slope, and aspect. From the digital elevation model, environmental covariates were extracted, composing three sets, made up of: 21 covariates, covariates after the exclusion of the multicollinear ones, and covariates chosen by expert knowledge. Prediction was performed with the following models: decision tree, random forest, multiple logistic regression, and support vector machine. The accuracy of the models was evaluated by the kappa index (K), general accuracy (GA), and class accuracy. The prediction models were sensitive to the disproportionate sampling of soil classes. The best predicted map achieved a GA of 71% and K of 0.59. The use of the covariate set chosen by expert knowledge improves model performance in predicting soil classes in a complex landscape, and random forest is the best model for the spatial prediction of soil classes.

Index terms: digital soil mapping, pedometry, predictive covariates, predictive models, soil-landscape relationship.

Predição de classes de solo em uma paisagem complexa no Sul do Brasil

Resumo – O objetivo deste trabalho foi avaliar o uso da seleção de covariáveis por conhecimento especializado no desempenho de modelos de predição de classes de solos em uma paisagem complexa, para identificar o melhor modelo preditivo para o mapeamento digital de solos na região Sul do Brasil. Um total de 164 pontos foram amostrados em campo, com uso do hipercubo latino condicionado, tendo-se considerado as covariáveis elevação, declividade e aspecto. A partir do modelo digital de elevação, extraíram-se as covariáveis ambientais que compuseram três conjuntos, formados por: 21 covariáveis, covariáveis após exclusão das multicolineares e covariáveis escolhidas por conhecimento especializado. A predição foi realizada com os seguintes modelos: árvore de decisão, floresta aleatória, regressão logística múltipla e máquina de vetor de suporte. A acurácia dos modelos foi avaliada pelo índice kappa (K), pela acurácia geral (AG) e pela acurácia da classe. Os modelos de previsão foram sensíveis à amostragem desproporcional de classes de solo. O melhor mapa predito obteve AG de 71% e K de 0,59. O uso do conjunto de covariáveis escolhido pelo conhecimento especializado melhora o desempenho do modelo em prever as classes de solo em uma paisagem complexa, e floresta aleatória é o melhor modelo para previsão espacial das classes de solo.

Termos para indexação: mapeamento digital de solos, pedometria, covariáveis preditoras, modelos preditivos, relação solo-paisagem.

Jean Michel Moura-Bueno[1 ✉],
Ricardo Simão Diniz Dalmolin[1],
Taciara Zborowski Horst-Heinen[1],
Luciano Campos Cancian[1],
Ricardo Bergamo Schenato[1],
André Carnieletto Dotto[2] and
Carlos Alberto Flores[3]

[1] Universidade Federal de Santa Maria, Centro de Ciências Rurais, Departamento de Solos, Avenida Roraima, nº 1.000, Cidade Universitária, Camobi,
CEP 97105-900 Santa Maria, RS, Brazil.
E-mail: bueno.jean1@gmail.com,
dalmolin@ufsm.br,
tacihorst@gmail.com,
lucianocancian@msn.com,
ribschenato@gmail.com

[2] Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciência do Solo,
Avenida Pádua Dias, no 11, Caixa Postal 9,
CEP 13418-900 Piracicaba, SP, Brazil.
E-mail: andrecdot@gmail.com

[3] Embrapa Clima Temperado, BR-392, Km 78,
9º Distrito, Monte Bonito, Caixa Postal 403,
CEP 96010-971 Pelotas, RS, Brazil.
E-mail: cflores@terra.com.br

✉ Corresponding author

Received
January 6, 2018

Accepted
June 26, 2019

How to cite
MOURA-BUENO, J.M.; DALMOLIN, R.S.D.; HORST-HEINEN, T.Z.; CANCIAN, L.C.; SCHENATO, R.B.; DOTTO, A.C.; FLORES, C.A. Prediction of soil classes in a complex landscape in Southern Brazil. **Pesquisa Agropecuária Brasileira**, v.54, e00420, 2019. DOI: https://doi.org/10.1590/S1678-3921.pab2019.v54.00420.

Introduction

The global demand for agricultural production, together with environmental sustainability and climate change concerns, has led to a growing interest in soil spatial information (Amundson et al., 2015). Soil maps are important sources of information to be used as a support in several areas of environmental science and engineering. In the state of Rio Grande do Sul, Brazil, as well as across the country, more detailed soil information at compatible scales for land use planning at the watershed and farm level is still necessary (Dalmolin & ten Caten, 2015).

In this scenario, new methodologies have been integrated to computational techniques in Soil Science, mainly in the area of soil mapping. Among these methods, stands out the one proposed by McBratney et al. (2003), named digital soil mapping (DSM), based on generating mathematical relationships between covariates and soil classes to predict the latter's spatial distribution. In this approach, the covariates represent the main factors of soil formation, including climate, organisms, relief, parental material, and time. Moreover, the legacy data obtained by traditional surveys and new field samplings can be used to train models for soil class inferences in unmapped areas (Dalmolin & ten Caten, 2015; Bagatini et al., 2016; Pahlavan-Rad et al., 2016; Silva et al., 2016; Meier et al., 2018).

Currently, for inferences on soil types, different calibration methods are being tested, such as decision tree (Teske et al., 2014; Silva et al, 2016), multiple logistic regression (ten Caten et al., 2011b), support vector machine, and random forest (Taghizadeh-Mehrjardi et al., 2015; Dias et al., 2016). The potential of each one is defined by input data, specifically by the degree of correlation between soil classes and environmental covariates (McKenzie & Ryan, 1999; Brungard et al., 2015), which may be high at some points of the landscape but low in others. It should be noted that not all covariates are directly related to a particular soil formation factor, and that some of them have indirect or multiple relationships with different formation factors (Moore et al., 1993; Ma et al., 2019). Time-related covariates, for example, are absent in predictive models unless they are manually incorporated (Noller, 2010). In this case, geomorphological maps can be a useful information source to represent time and parent material in soil genesis (Scull et al., 2005); however, this legacy information is often in a coarse cartographic scale, which is unsuitable for local studies. The predictive potential of the calibration methods is also influenced by the quality and resolution of the covariates extracted from the digital elevation model (DEM) (Moura-Bueno et al., 2016), the window size from which the DEM derivatives are extracted (Samuel-Rosa et al., 2015), and the quality and quantity of the soil data used in modeling (Teske et al., 2014).

The environmental covariates used in DSM are usually selected based on the relationship between soil distribution and landscape, whose main conditioning factors are geology, geomorphology, land use/land cover, climate, and relief. These covariates also have been widely used in predictive models (Bagatini et al., 2016; Pahlavan-Rad et al., 2016; Silva et al., 2016; Meier et al., 2018). Among them, relief is the main soil formation factor taken into account in DSM (McBratney et al., 2003), due to the easy access to the DEM and to the close relationship of the covariate with the soil distribution pattern in the landscape (Moore et al., 1993; McKenzie & Ryan, 1999). Covariate selection can alter classification patterns and directly influence the result of soil class prediction (Brungard et al., 2015; Dias et al., 2016), which is attributed to the level of information on the environmental variability carried by each covariate (Moore et al., 1993; McKenzie & Ryan, 1999; Samuel-Rosa et al., 2015). This makes it improbable that only one predictive model will be useful for all geomorphic surfaces (Grunwald, 2009).

The pedologist's knowledge about the soil-landscape relationship may be an alternative to covariate selection, in order to improve soil class mapping, mainly in complex landscapes. However, this strategy may be biased or even fail in regions where knowledge about the pedogenetic process is insufficient (Moore et al., 1993; Vasques et al., 2012). Consequently, studies that determine sets of covariate predictors to be used in the training of models under complex landscape conditions are fundamental (Moore et al., 1993; McKenzie & Ryan, 1999; Samuel-Rosa et al., 2015), aiming to generate high quality maps to meet the current soil data demand.

There are still challenges for DSM, such as new theories, methods, and applications, especially for highly heterogeneous landscapes (Zhang et al., 2017). Therefore, soil information, combined with different sets of covariates and machine learning, may have a distinct predictive capacity for the DSM of detailed soil maps (≥ 1:20,000) in complex landscapes, which are characterized by the high variability of geology, relief, and land use/land cover.

The objective of this work was to evaluate the use of covariate selection by expert knowledge on the performance of soil class predictive models in a complex landscape, in order to identify the best predictive model for digital soil mapping in the Southern region of Brazil.

Materials and Methods

The study was carried out in an agricultural area of approximately 12 km^2 in the municipality of Santa Maria, in the state of Rio Grande do Sul, Brazil, between the coordinates 29°37'22.94"S and 53°39'45.28"W, 29°40'40.78"S and 53°38'41.92"W. The climate is classified as Cfa, according to the Köppen-Geiger classification system. The area comprises the transition region called "Rebordo do Planalto", between the "Planalto" and "Depressão Central" physiographic regions of the state.

The landscape characteristics of the study area condition a high variability in soil class distribution. According to the Brazilian soil classification system (Santos et al., 2013), the classes predominant in the region are: Neossolos, Argissolos, Cambissolos, and Planossolos, i.e., Entisols, Ultisols, Inceptisols, and Alfisols, respectively. The local relief varies between plain (slope from 0 to 3%) and mountainous (slope from 45 to 100%), and elevations range between 74 and 454 m, with the periodical occurrence of land slips. The geological heterogeneity of the region is attributed to: the transition of acidic and basic igneous rocks, such as rhyolite-rhyodacite and andesite-basalt; consolidated sedimentary rocks, particularly aeolian and fluvial sandstones; and nonconsolidated rocks, including fluvial and colluvial deposits (Sartori, 2009). The geological variability of the area is shown in Figure 1, which was adapted from the geological map of Camobi, in the state of Rio Grande do Sul, at a scale of 1:50,000 (Maciel Filho et al., 1988). Despite the relevance of geology for local soil genesis, the lack of detailed-scale maps (> 1:50,000) for the study area made it impossible to use this information as a predictor covariate. The land use/land cover that predominates in the area is made up of shrubland and native grassland occupying more than half of the area, followed by native semi-deciduous forests, eucalyptus forestry, and annual crop agriculture. Dullius et al. (2018) observed that there is a relationship between vegetation variability and pedology. Native forests are predominant in strong wavy and mountainous relief. In wavy relief, native fields and shrubs are common, whereas flat relief areas are dominated by annual crops, mainly irrigated rice (*Oryza sativa* L.). These characteristics show that "Rebordo do Planalto" is both an environmentally complex region and a constantly changing landscape, where the soil formation factors do not act uniformly, making it difficult to fit soil class prediction models.

A total of 164 points were sampled in the field (Figure 1) using the conditioned Latin hypercube (Minasny & McBratney, 2006), considering the covariates elevation, aspect, and slope; this procedure was performed in the R programming language (R Core Team, 2017), with 10,000 interactions. This sampling method was chosen for taking into account the best geographical distribution of the points, based on the distribution frequency of the covariates in the landscape. At each of the 164 points, sampling was carried out with an auger or by opening trenches, in order to identify and classify the soil to the second categorical level, according to the Brazilian soil classification system (Santos et al., 2013). The points were randomly separated into training (70%, n=115) and validation (30%, n=49) sets, preserving the distribution of classes between sets.

The soil classes identified in the study area are presented in Table 1. Two soil classes, located in very similar parts of the landscape, were joined: Cambissolo Háplico (CX), an Udept; and Neossolo Regolítico (RR), an Orthent. This strategy aimed to reduce prediction errors and facilitate spatialization (Dias et al., 2016). The Neossolo Litólico (RL) class, an Orthent, was more frequently observed in areas of higher elevation and steeper slope; Argissolo Bruno-Acinzentado (PBAC) and Argissolo Vermelho-Amarelo (PVA), both Udults, as well as CX and RR, were found at intermediate elevations; and Planossolo Háplico (SX), an Aqualf, at sites with low elevation and flat relief.

The following terrain covariates were calculated, according to Wilson & Gallant (2000), from a 30-m resolution DEM obtained from Shuttle Radar

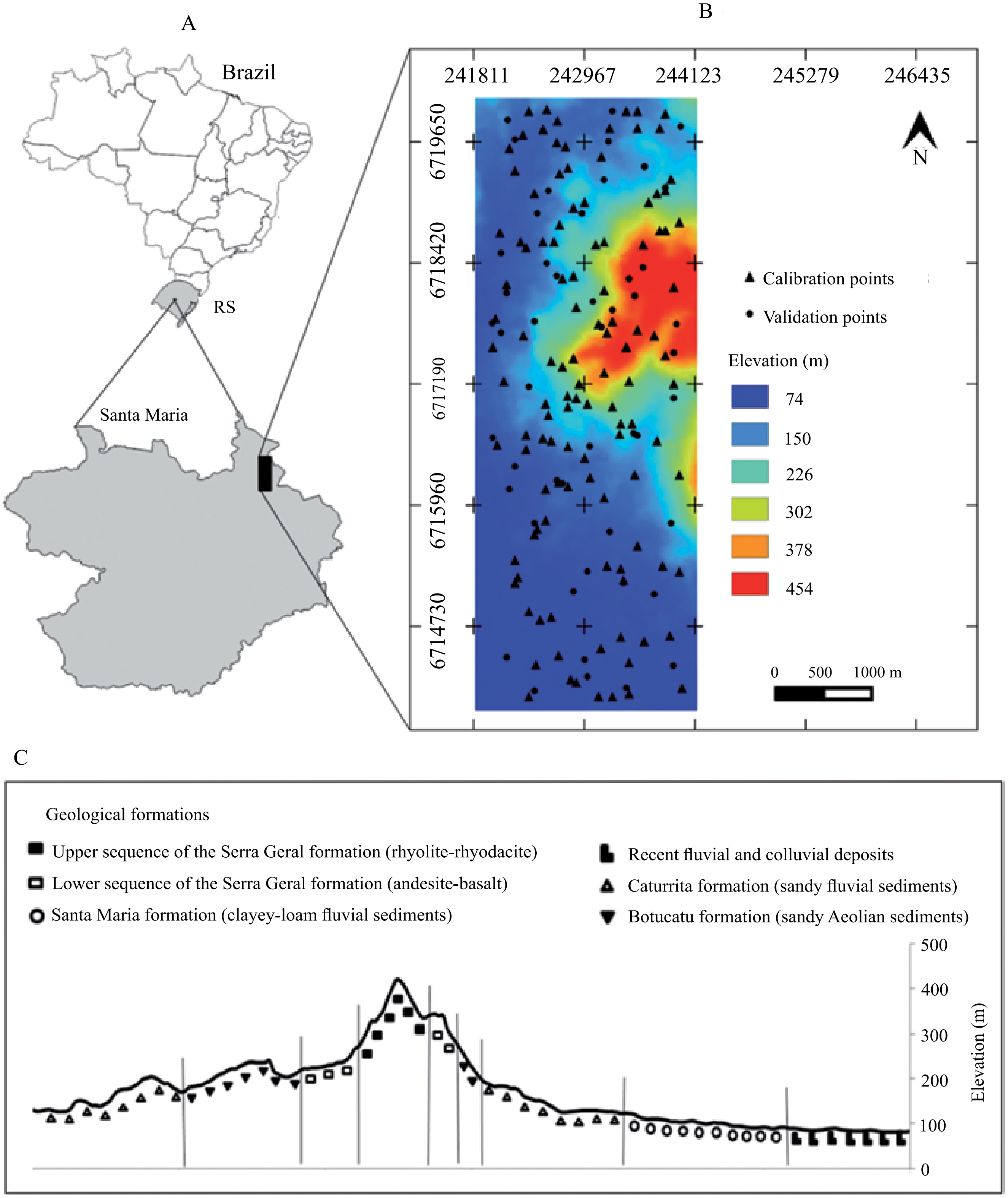
A
Brazil
RS
Santa Maria
B
241811
242967
244123
245279
246435
6719650
6718420
6717190
6715960
6714730
N
▲ Calibration points
● Validation points
Elevation (m)
74
150
226
302
378
454
0 500 1000 m
C
Geological formations
■ Upper sequence of the Serra Geral formation (rhyolite-rhyodacite)
□ Lower sequence of the Serra Geral formation (andesite-basalt)
○ Santa Maria formation (clayey-loam fluvial sediments)
Recent fluvial and colluvial deposits
△ Caturrita formation (sandy fluvial sediments)
▼ Botucatu formation (sandy Aeolian sediments)
500
400
300
200
100
0
Elevation (m)

Figure 1. Location of the study area in the municipality of Santa Maria, in the state of Rio Grande do Sul (RS), Brazil (A); extended study area with elevation representation, indicating soil identification points (training and validation) underlying the prediction of soil classes (B); and elevation profile in the north-south direction of the area, representing the geomorphological sequence (C).

Topography Mission data: elevation, slope, topographic wetness index (TWI), convergence index (CI), profile curvature (ProfC), vertical distance to channel network (VDCN), cumulative flow (CF), analytical hillshading (AH), general curvature (GC), channel network base level (CNBL), transverse curvature (TrC), valley depth (VD), planar curvature (PlanC), longitudinal curvature (LC), terrain ruggedness index (TRI), slope length and steepness factor (LS), aspect, tangential curvature (TangC), topographic position index (TPI), and relative slope position (RSP). In addition, the normalized difference vegetation index (NDVI), obtained from Landsat 8 satellite images from November 2016, was used. This covariate was included to represent the organism factor in soil formation, considering the relationship between land use/land cover and soil class (Samuel-Rosa et al., 2011; Dullius et al., 2018), which was observed during the soil sampling stage in the study area. The covariates were derived in the SAGA-GIS, version 2.1.2, software (Conrad et al., 2015).

Two strategies were used to identify the model most suitable for soil class prediction for the DSM of a complex landscape. The first was selecting the covariates with the best predictive response. For this, three sets were defined, composed of: A, 21 covariates extracted from the DEM; B, covariates applied in set A reduced by the principal component analysis (PCA), as suggested by ten Caten et al. (2011a); and C, covariates chosen after the analysis of their frequency distributions in each soil class, observed in the boxplots, performed by expert knowledge of the soil-landscape relationship in the study area. The covariates selected to best distinguish soil classes were: elevation, slope, TWI, CI, AH, GC, NDVI, CNBL, TrC, and VD in set B; and elevation, slope, TWI, CNBL, NDVI, and aspect in set C.

The second strategy was assessing different models, i.e., decision tree (DT) (algorithm J48), random forest (RF), support vector machine (SVM), and multiple logistic regression (MLoR), considering the A, B, and C covariate predictor sets. For the accuracy assessment of the models, the respective confusion matrices for the training and validation sets were generated, to calculate general accuracy (GA), class accuracy (CA), and the kappa index (K). GA was used to evaluate the proportion of correctly predicted map pixels in relation to the number of total pixels, whereas CA was used to assess the correct pixel ratio of each soil class. All statistical analyses were performed in the R programming language (R Core Team, 2017).

Table 1. Soil classes identified in the study area in the municipality of Santa Maria, in the state of Rio Grande do Sul, Brazil.

Symbol	Brazilian soil classification system[(1)]	Soil taxonomy[(2)]	Sampling points (%)
RL	Neossolo Litólico	Orthent	36
PBAC	Argissolo Bruno-Acinzentado	Udult	26
PVA	Argissolo Vermelho-Amarelo	Udult	14
CX	Cambissolo Háplico	Udept	9
RR	Neossolo Regolítico	Orthent	8
SX	Planossolo Háplico	Aqualf	7
Total			100

[(1)]Santos et al. (2013). [(2)]Soil Survey Staff (2014).

Results and Discussion

The boxplot analysis showed that only some covariates presented direct relationships with a given soil class (Figure 2), indicating that not all of them are directly related to soil class distribution in a complex landscape (McKenzie & Ryan, 1999; Brungard et al., 2015; Ma et al., 2019), where the pedogenetic processes of the soil act differently (Huggett, 1975).

The mean values of CNBL differed for PBAC+PVA, CX, RL, RR, and SX (Figure 2), showing the potential of this covariate for the distinction of these soil classes. Covariates with a different numerical range between soil classes are considered the most relevant because they explain the differences between the sites where each soil class occurs in the landscape (Silva et al., 2016). Among the most used covariates in DSM, elevation and slope (McBratney et al., 2003; Bishop & Minasny, 2006) stand out due to their high relationship with the residual water content in the profile and soil formation processes, followed by CNBL, which is calculated from elevation. Therefore, these covariates are important for soil class prediction, mainly for classes located at higher elevations and in steeper regions, as RL in the present study. The PBAC, PVA, CX, and RR classes are distributed at intermediate elevations below 200 m; furthermore, the last two are poorly represented among the points sampled in the area. Class SX was well characterized in the soil

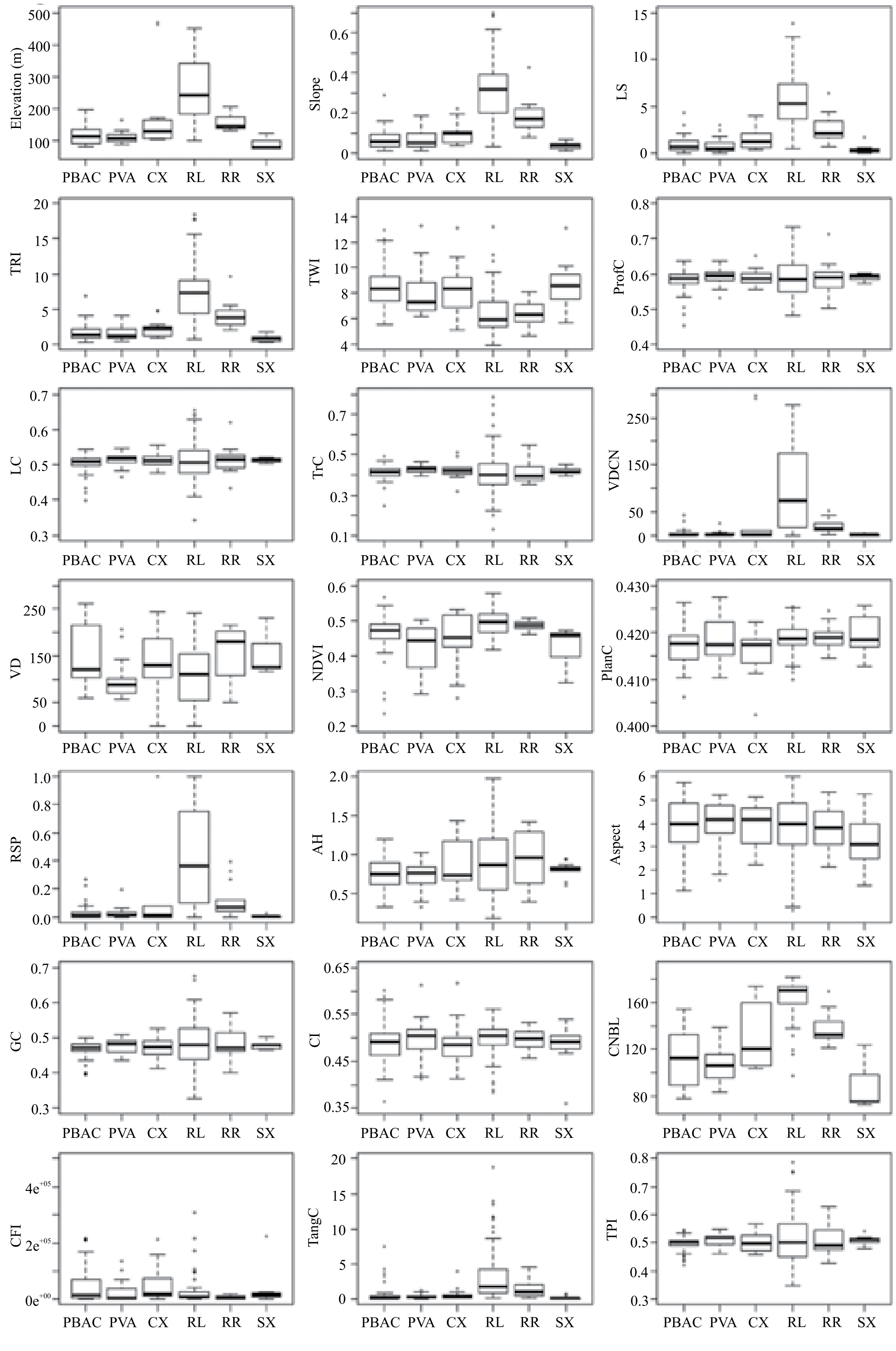

Figure 2. Boxplot representing: range of values; median, minimum, maximum, first, and third quartiles; and interquartile interval of the set of covariates of each soil class. LS, slope length and steepness factor; TRI, terrain ruggedness index; TWI, topographic wetness index; ProfC, profile curvature; LC, longitudinal curvature; TrC, transverse curvature; VDCN, vertical distance to channel network; VD, valley depth; NDVI, normalized difference vegetation index; PlanC, planar curvature; RSP, relative slope position; AH, analytical hillshading; GC, general curvature; CI, convergence index; CNBL, channel network base level; CF, cumulative flow; TangC, tangential curvature; TPI, topographic position index; PBAC, Argissolo Bruno-Acinzentado, an Udult; PVA, Argissolo Vermelho-Amarelo, an Udult; CX, Cambissolo Háplico, an Udept; RL, Neossolo Litólico, an Orthent; RR, Neossolo Regolítico, an Orthent; and SX, Planossolo Háplico, an Aqualf.

survey and clearly identified at lower elevation sites below 100 m and in flat areas, despite its low sampling frequency in the study area.

The VDCN covariate is related to sediment erosion and deposition and organic matter concentration, allowing the prediction of the distribution of different soil classes in the landscape. In the present study, it presented greater potential to distinguish the RL class (Figure 2); however, its range was short for the other soil classes, implying their low distinction power in the construction of the predictive models. Similarly to VDCN, the NDVI has the potential to discriminate the RR and RL classes from the others due to differences in the range of this index. This fact is related to the association of soil classes with incipient pedogenesis and the land use of native forest (Dullius et al., 2018), which predominate in areas of relief varying from strong wavy to mountainous (Samuel-Rosa et al., 2011).

Also based on boxplot analyses, Silva et al. (2016) observed that VDCN was the covariate with the highest power to discriminate soil classes in an area of approximately 4.85 km^2, with flat to undulating relief, in the state of Minas Gerais, Brazil. Moreover, Meier et al (2018) found that precipitation, annual temperature, TWI, slope, VDCN, and bands 1, 7, and 11 of Landsat 8 were the most important covariates for the prediction of soil classes in an area located in Zona da Mata, also in the state of Minas Gerais. The study of Teske et al. (2014) showed that elevation, slope, flow length, and slope orientation were the covariates that best explained soil distribution in a landscape with flat and slightly undulating relief in the "Encosta Inferior" physiographic region, located in the northeast of the state of Rio Grande do Sul. According to these authors, elevation was the most important covariate in the distinction of Rhodudults, soils with a base saturation <35%, a clayey B horizon, and a dark-colored surface horizon. However, for Eutrudepts, soils with an incipient development horizon and a high base saturation in the 25–75-cm layer, the elevation range was higher, and the inclusion of the covariates slope and flow length was required for discrimination between soil classes.

Despite the wide usage of TWI in DSM studies, its potential to discriminate soil classes was not confirmed in the present work, since its median values were close and it varied greatly in the five evaluated soil classes (Figure 2); this result disagrees with that of Meier et al (2018), who found that TWI was the fourth most important covariate for soil discrimination. Although, during the sampling stage, the PBAC and PVA classes occurred in sites with different soil moisture, the boxplot analysis did not allow this distinction based on TWI values. This discrepancy can be explained, in part, by the effect of the parent material on these soil classes; the Santa Maria formation, for example, was related to PBAC and the Caturrita formation to PVA. In addition, data on local geology are not available on the scale needed to allow this distinction, hampering the development of better DSM. This is an indicative that the same covariate and/or set of covariates perform differently in distinguishing soil classes due to the heterogeneity of the covariates and the complex interaction between them in the pedogenesis process. In the landscape, soil formation factors do not act uniformly, which makes it difficult to construct more accurate predictive models (Ma et al., 2019). Therefore, in complex landscapes, as those assessed in the present study, the difficulty in establishing the covariates linked to the most important formation factors involved in soil genesis is evident. A similar behavior was observed by Samuel-Rosa et al. (2015), when fitting models to predict the soil granulometric composition of this same physiographic region of the state of Rio Grande do Sul.

The quality of DEM-derived data is an important factor in DSM, especially in areas with a complex landscape, with great soil variability. Teske et al. (2014) and Moura-Bueno et al. (2016) showed that relief is represented differently by each DEM. Teske et al. (2014) reported a greater variation in elevation values obtained from orbital sensors with a higher spatial resolution of 30 m, reflected in a reduction in the accuracy of the soil-landscape relationship for an area of approximately 68.5 km^2. However, when evaluating the quality of the DEM for DSM, Moura-Bueno et al. (2016) found that a spatial resolution of 30 m may be restrictive in cases of errors in altitude values when detailed DSM is applied in areas of gently undulating relief. Therefore, the DEM resolution used and the accuracy of the covariates may explain the small differences between the values of: TWI for the PBAC, CX, and SX classes; elevation for RR+CX, PBAC, PVA, and SX; and slope for PBAC and PVA (Figure 2). In this case, the detail level of the terrain underlying the DEM was insufficient to discriminate these soil classes in the landscape, resulting in a loss of the predictive capacity of the models due to the difficulty in establishing rules capable of distinguishing classes. However, to discriminate the RL from the other classes and RR from SX, the covariate elevation was important. Furthermore, to distinguish PBAC, PVA, and SX from RR and RL, the TWI median value should differ.

Regarding the sets of predictor covariates, the accuracy of the models trained with all covariates (set A) was higher than that of those trained with the other sets, with a mean GA of 76% and K of 0.66 (Table 2). In validation, the highest values were obtained with set B, with a mean GA of 67% and K of 0.53. The reduction in the number of covariates in set B did not improve the mean accuracy of the training model; however, despite the smaller number of soil data used in validation, the mean accuracy values of set B increased, compared with those of set A. Decreasing the number of covariates facilitates the prediction process, since, when working with many covariates, whether multicollinear or irrelevant, relationships that are not pertinent to soil mapping can be established between soil geomorphology and soil classes, reducing the performance of the model. The use of many covariates may also hamper the establishment of pedological relationships at the moment of interpretation of the soil-landscape relationship (Brungard et al., 2015).

In this sense, set C was made up of covariates with a greater potential to discriminate soil classes based on expert knowledge of the soil-landscape relationship in the study area. For this set, the mean model accuracy was lower than that for sets A and B, both in training and validation. However, each model performed differently regarding accuracy. RF, for example, performed best in set C, with a GA of 71% and K of 0.59 (Figure 3); in set A, it had difficulties in predicting the PBAC and PVA classes, resulting in their generalization, especially of PVA, which showed the lowest CA of 29% (Table 2). With this model, a lower area was also predicted for the RR+CX association, resulting in a CA of 0%. It should be noted that, during the soil sampling stage, it was observed that PBAC and PVA occur in the same position in the landscape; the difference between both is only the color of the B horizon, which is bright and gray for PBAC due to its aquic condition. The covariate TWI, therefore, would have potential in discriminating these classes in the landscape; however, in the field, the parental material is more significant due to its relationship with soil drainage, which conditions different moisture regimes. In the present study, PBAC is predominantly derived from the Santa Maria formation, which is clayey-loamy and provides residual wetness. In spite of the importance of geology, it was not included in the predicted model because of the coarse scale used in the studied area, as aforementioned.

The RF model fitted with set B showed difficulties in accurately predicting the PBAC and RR+CX classes (Table 2), which were confused with RL (Figure 3); the model was more generalist and more uncertain to predict RR+CX, with a CA of 50%. However, when fitted by set C, RF had higher accuracy in the predictions of RR+CX and RL, with a CA of 100% for the RR+CX association. These soil classes occur in the middle third of the slope in the study area, associated with the land uses/land covers shrubland and native grassland, as well as with distant shrubs and pastures of the network of drainage channels. This makes it difficult for pedologists to separate RR+CX from the other soil classes in the landscape based on elevation and slope, for example. However, the selection of covariates, such as NDVI and CNBL, from expert knowledge has proved to be an efficient strategy in the construction of predictive models, especially of RF.

The obtained GA values were higher than those reported by Dias et al. (2016), who found a GA between 48.3 and 58% and K between 0.41 and 0.50, while evaluating strategies to predict soil classes using the RF model. Franco et al. (2015) also obtained a lower GA of 53% and K of 0.34 when reducing covariates, via PCA, for soil mapping in an area with a complex landscape at the “Depressão Central” physiographic region, in the state of Rio Grande do Sul.

The higher precision values obtained by the RF model are an indicative of its potential when the predictor covariates are identified based on expert knowledge of the soil-landscape relationships and on the boxplot analysis. This result is the opposite of that observed by Brungard et al. (2015), who verified lower soil CA prediction when the covariates were selected by expert knowledge. However, Ma et al. (2019) suggested that pedologists should have a better understanding of the factors controlling the variation of the soil-landscape relationship, not only of the tools used for mathematical modelling of data.

In young and unstable geomorphic surfaces, such as those assessed in the present study, there is a greater difficulty in constructing mathematical models that explain pedogenesis (Samuel-Rosa et al., 2015), and the use of environmental covariates selected by expert knowledge may be a strategy in DSM.

Using legacy data to update soil class maps of northern Iran, Pahlavan-Rad et al. (2016) obtained K values of up to 0.34 and concluded that the RF model is more efficient than MLoR in predicting soil classes. According to Hengl et al. (2007), MLoR models depend on a strong correlation between predictor covariates and the soil, besides requiring a minimal representativeness of each soil class in the training phase. In the present study, the highest GA and K values for the validation of set B were obtained with model MLoR; however, for set C, both values were lower than those with RF (Table 2). According to ten

Table 2. Accuracy indicators for the three sets of predictor covariates and four prediction models evaluated.

Covariate set[1]	Prediction model[2]	Class accuracy (%)[3]					General accuracy (%)	Kappa index
		PBAC	PVA	RL	RR+CX	SX		
				Training set				
A	DT	94	88	98	83	100	93	0.91
	RF	65	50	88	17	67	63	0.50
	SVM	54	89	83	60	0	69	0.56
	MLoR	84	56	90	39	100	77	0.68
	Mean	74	71	90	50	67	76	0.66
B	DT	94	88	100	78	100	93	0.91
	RF	61	50	83	11	67	60	0.46
	SVM	54	73	83	50	0	67	0.54
	MLoR	66	62	84	42	80	71	0.61
	Mean	69	68	87	45	62	73	0.63
C	DT	97	93	88	82	100	90	0.87
	RF	63	45	77	13	100	60	0.46
	SVM	56	73	78	50	100	67	0.54
	MLoR	62	54	78	43	89	68	0.56
	Mean	70	66	80	47	97	71	0.61
				Validation set				
A	DT	75	43	80	25	100	65	0.52
	RF	83	29	90	0	100	65	0.50
	SVM	52	0	75	0	0	61	0.43
	MLoR	58	29	85	25	100	61	0.46
	Mean	67	25	83	13	75	63	0.48
B	DT	75	29	90	25	100	67	0.54
	RF	47	75	79	50	100	67	0.53
	SVM	48	33	83	50	0	63	0.47
	MLoR	50	40	86	75	100	69	0.57
	Mean	55	44	85	50	75	67	0.53
C	DT	50	23	72	25	34	51	0.32
	RF	56	40	87	100	100	71	0.59
	SVM	40	0	80	0	0	55	0.34
	MLoR	53	50	87	50	50	67	0.53
	Mean	50	28	82	44	46	61	0.44

[1]A, 21 covariates extracted from the digital elevation model; B, covariates applied in set A reduced by the principal component analysis; and C, covariates chosen after the analysis of their frequency distributions, associated with expert knowledge of the soil-landscape relationship. [2]DT, decision tree; RF, random forest; SVM, support vector machine; and MLoR, multiple logistic regression. [3]PBAC, Argissolo Bruno-Acinzentado, an Udult; PVA, Argissolo Vermelho-Amarelo, an Udult; RL, Neossolo Litólico, an Orthent; RR, Neossolo Regolítico, an Orthent; CX, Cambissolo Háplico, an Udept; and SX, Planossolo Háplico, an Aqualf.

Caten et al. (2011b), the MLoR model is sensitive to prediction errors for soil classes found close to each other in the landscape, as is the case of PBAC and PVA in the present study; MLoR showed the lowest accuracy for the latter. To solve this problem, ten Caten et al. (2011a) suggested the use of a greater number of representative covariate predictors. However, in set A, which has the highest number of covariates, the accuracy for PVA by MLoR was not improved. This may be associated with the negative multicollinearity effect in linear models (Hengl et al., 2007) or with the fact that the covariates are not being representative in distinguishing soil classes in the landscape.

The SVM model presented the poorest performance in predicting the PVA, RR+CX, and SX classes, with zero accuracy for the three sets of covariates (Table 2). Taghizadeh-Mehrjardi et al. (2015) also found that this model performed worse than MLoR, RF, and DT. The authors reported that the accuracy of the SVM model decreased as the level of soil taxonomic detail increased, reaching values of order K = 0.74, subgroup K = 0.68, and family K = 0.60. The highest values were obtained for the model of artificial neural networks (order K = 0.84, subgroup K = 0.75, and family K = 0.69), and intermediate ones for RF (order K = 0.78, subgroup K = 0.73, and family K = 0.65). The generalization by the SVM model, mainly in sets A and C, was only able to discriminate the most representative classes of the area, i.e., 37% RL and 26% PBAC. Similarly, Bagatini et al. (2016), when evaluating the extrapolation of the soil-landscape relationship in two watersheds in Southern Brazil by

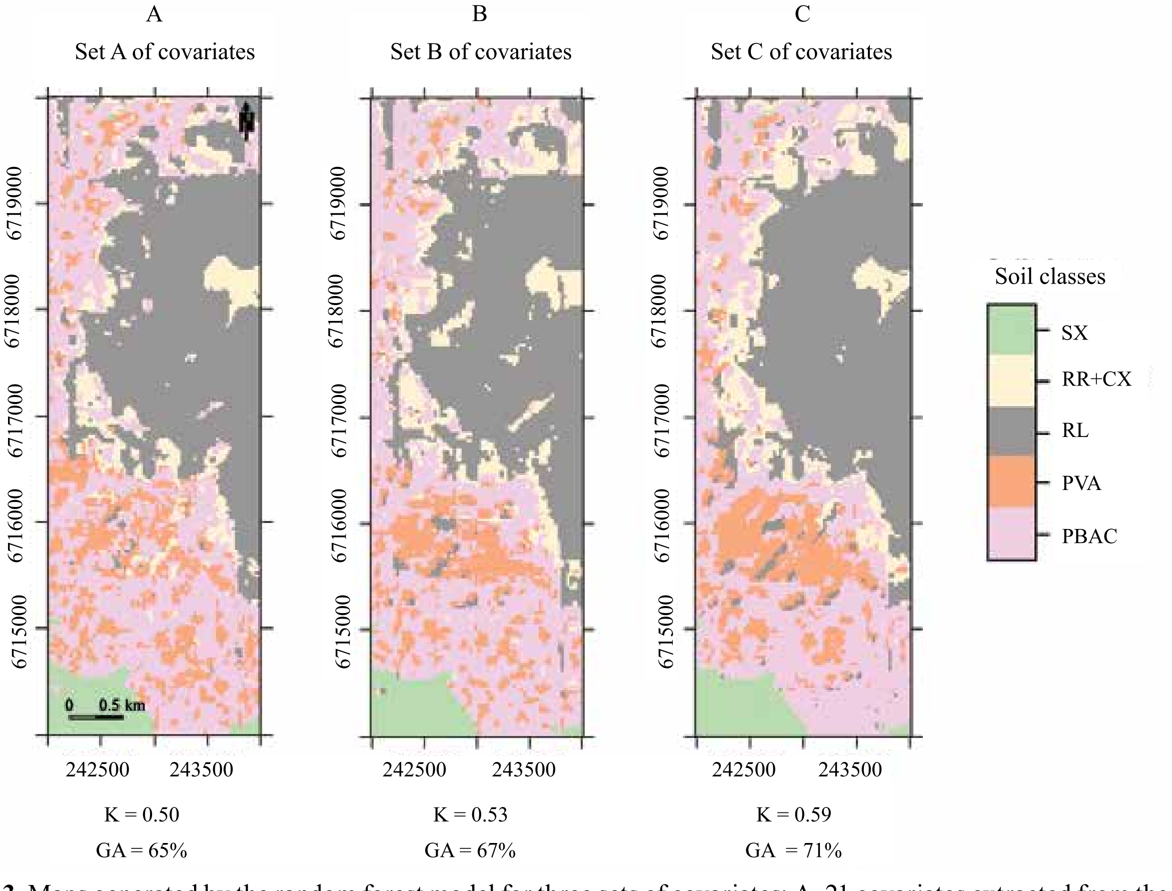

Figure 3. Maps generated by the random forest model for three sets of covariates: A, 21 covariates extracted from the digital elevation model; B, covariates applied in set A reduced by the principal component analysis; and C, covariates chosen after the analysis of their frequency distributions, associated with expert knowledge of the soil-landscape relationship. K, kappa index; GA, general accuracy; SX, Planossolo Háplico, an Aqualf; RR, Neossolo Regolítico, an Orthent; CX, Cambissolo Háplico, an Udept; RL, Neossolo Litólico, an Orthent; PVA, Argissolo Vermelho-Amarelo, an Udult; and PBAC, Argissolo Bruno-Acinzentado, an Udult.

the DT method, concluded that unrepresentative soil classes in the landscape were underestimated in the training and validation of the model. Taghizadeh-Mehrjardi et al. (2015) pointed out that the spatial distribution and number of representative samples per soil class influence the quality of DSM.

In the validation stage, a reduction in accuracy was observed in the three sets of covariates with the DT, MLoR, and SVM models, especially with the DT. The algorithm used by the DT constructed a sophisticated model training solution, presenting the highest GA and K values in the training stage (Table 2), with a minimum error in CA. However, when challenged with a new data set (validation set), classification errors occurred, reducing the robustness of the model, which is attributed to a problem known as overfitting. This problem can be induced by the characteristics of the variables, mainly when the number of independent variables used to construct the models is higher than that of dependent variables. Moreover, each predictor covariate presents soil class information, which increases the possibility of inferring patterns, but decreases the possibility of generalizing such patterns, especially in areas of complex topography and geology, as the region evaluated in the present study. This is more frequent in generalized linear models, where no multicollinearity effects on data are detected in the model fitting phase (Hengl et al., 2007). This problem was also described by Kempen et al. (2009), in the prediction of soil classes by the MLoR model. In the current study, this behavior was mainly observed for set A, in which GA and K decreased in the validation stage. For the B and C sets, the reduction of covariates eliminated the effect of multicollinearity, slightly reducing the accuracy of the MLoR model in validation.

The values of CA varied between the sets of covariates and models (Table 2). The main underlying factors for this were: representativeness of each class in the landscape, covariates of the terrain used as predictors, differentiated potential of each covariate in the discrimination of classes, and capacity of each model in the construction of the classification rules. In the validation of the RF and SVM models, the CA for SX was, respectively, 100 and 0% in all sets, evidencing the low capacity of SVM in predicting SX. However, the DT and MLoR models reached 100% CA for SX in validation for sets A and B, which dropped to 34% for set C. This result is a consequence of the low sampling frequency in class SX, representing 7% of the area, and of the sensitivity of the SVM, DT, and MLoR models in relation to this factor. Despite the low representativeness of SX, when using the RF model, the CA value was higher than that for the most representative classes in the area – RL, PBAC and PVA. The representation of SX is associated with well-defined relief features, i.e., lower and plain parts of the landscape (Figure 3), with a low amplitude regarding the elevation and slope values (Figure 2). In addition, the CA values of the NDVI are associated with the irrigated rice crop, and the ones obtained for the SX class are in agreement with those recorded by Cordeiro et al. (2017) in intensive agricultural areas, during the summer period, at the “Depressão Central” physiographic region of the state of Rio Grande do Sul. A similar performance was also observed for RL, which, although quite representative in the area, is found in higher and steeper sites, constituting landscapes well discriminated by the covariates elevation, slope, NDVI, and VDCN, which showed the highest CA values in all models.

Regarding the variability of soil classes in the study area, the highest CA values in the prediction of RR+CX were obtained by the RF model (Table 2), evidencing its robustness, particularly when using the covariates of set C, especially slope, NDVI and CNBL, which differ in relation to those of the other soil classes (Figure 2). When evaluating different strategies to predict soil classes in areas with no reference data in a sedimentary watershed of the São Francisco River, in the state of Minas Gerais, Dias et al. (2016) found that the use of detailed taxonomic information (subgroup level) resulted in an increase in map fragmentation and in accuracy loss. Moreover, the accuracy of the different sets of predictor covariates and prediction models varied, being the highest for the RF model.

Regional prediction models based on landscape characteristics must be developed with expert knowledge input, according to Grunwald (2009). Therefore, in the present study, the evaluation and selection of covariates by expert knowledge regarding the soil-landscape relationship was an efficient strategy. The RF model was more robust than DT, MLoR, and SVM for soil class prediction in a complex landscape, characterized by heterogeneous relief and geology. The prediction models were sensitive to disproportionate soil class sampling, which is a limitation to obtain accurate soil maps in complex landscapes. Consequently, the classes that are difficult to discriminate by the terrain covariates require more samples for model training and validation.

Conclusions

1. The use of the covariate set chosen by expert knowledge improves the performance of the models – decision tree, random forest, multiple logistic regression, and support vector machine – in predicting soil classes in a complex landscape.

2. Random forest is the best spatial prediction model.

3. The support vector machine model is only able to discriminate the most representative soil classes in a complex landscape.

4. The multiple logistic regression model is negatively affected by the multicollinearity of the covariates.

Acknowledgement

To Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for financial support and for the research productivity scholarship granted to the second author.

References

AMUNDSON, R.; BERHE, A.A.; HOPMANS, J.W.; OLSON, C.; SZTEIN, A.E.; SPARKS, D.L. Soil and human security in the 21st century. **Science**, v.348, p.647-654, 2015. DOI: https://doi.org/10.1126/science.1261071.

BAGATINI, T.; GIASSON, E.; TESKE, R. Expansão de mapas pedológicos para áreas fisiograficamente semelhantes por meio de mapeamento digital de solos. **Pesquisa Agropecuária Brasileira**, v.51, p.1317-1325, 2016. DOI: https://doi.org/10.1590/S0100-204X2016000900031.

BISHOP, T.F.A.; MINASNY, B. Digital soil-terrain modeling: the predictive potential and uncertainty. In: GRUNWALD, S. (Ed.). **Environmental soil-landscape modeling**: geographic information technologies and pedometrics. Boca Raton: CRC Press, 2006. p.185-213. (Books in Soils, Plants, and the Environment).

BRUNGARD, C.W.; BOETTINGER, J.L.; DUNIWAY, M.C.; WILLS, S.A.; EDWARDS JR., T.C. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v.239-240, p.68-83, 2015. DOI: https://doi.org/10.1016/j.geoderma.2014.09.019.

CONRAD, O.; BECHTEL, B.; BOCK, M.; DIETRICH, H.; FISCHER, E.; GERLITZ, L.; WEHBERG, J.; WICHMANN, V.; BÖHNER, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. **Geoscientific Model Development**, v.8, p.1991-2007, 2105. DOI: https://doi.org/10.5194/gmd-8-1991-2015.

CORDEIRO, A.P.A.; BERLATO, M.A.; FONTANA, D.C.; MELO, R.W. de; SHIMABUKURO, Y.E.; FIOR, C.S. Regiões homogêneas de vegetação utilizando a variabilidade do NDVI. **Ciência Florestal**, v.27, p.883-896, 2017. DOI: https://doi.org/10.5902/1980509828638.

DALMOLIN, R.S.D.; TEN CATEN, A. Mapeamento digital: nova abordagem em levantamento de solos. **Investigación Agraria**, v.17, p.77-86, 2015. DOI: https://doi.org/10.18004/investig.agrar.2015.diciembre.77-86.

DIAS, L.M. da S.; COELHO, R.M.; VALLADARES, G.S.; ASSIS, A.C.C. de; FERREIRA, E.P.; SILVA, R.C. da. Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco. **Pesquisa Agropecuária Brasileira**, v.51, p.1396-1404, 2016. DOI: https://doi.org/10.1590/s0100-204x2016000900038.

DULLIUS, M.; DALMOLIN, R.S.D.; PEDRON, F. de A.; LONGHI, S.J.; HORST, T.Z.; MOURA-BUENO, J.M. Influência pedológica e topográfica na distribuição de espécies arbóreas em diferentes estágios de regeneração. **Revista Brasileira de Ciências Agrárias**, v.13, e5589, 2018. DOI: https://doi.org/10.5039/agraria.v13i4a5589.

FRANCO, A.M.P.; DALMOLIN, R.S.D.; RUIZ, L.F.C.; TEN CATEN, A.; SOARES, J.W. Delineamento das unidades de mapeamento de solos utilizando o Google Earth. **Geociências**, v.34, p.861-871, 2015.

GRUNWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. **Geoderma**, v.152, p.195-207, 2009. DOI: https://doi.org/10.1016/j.geoderma.2009.06.003.

HENGL, T.; TOOMANIAN, N.; REUTER, H.I.; MALAKOUTI, M.J. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. **Geoderma**, v.140, p.417-427, 2007. DOI: https://doi.org/10.1016/j.geoderma.2007.04.022.

HUGGETT, R.J. Soil landscape systems: a model of soil genesis. **Geoderma**, v.13, p.1-22, 1975. DOI: https://doi.org/10.1016/0016-7061(75)90035-X.

KEMPEN, B.; BRUS, D.J.; HEUVELINK, G.B.M.; STOORVOGEL, J.J. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. **Geoderma**, v.151, p.311-326, 2009. DOI: https://doi.org/10.1016/j.geoderma.2009.04.023.

MA, Y.; MINASNY, B.; MALONE, B.P.; MCBRATNEY, A.B. Pedology and digital soil mapping (DSM). **European Journal of Soil Science**, v.70, p.216-235, 2019. DOI: https://doi.org/10.1111/ejss.12790.

MACIEL FILHO, C. L.; MEDEIROS E.; VEIGA N.V.L.; SARTORI P.L.; GASPARETO, M. **Mapa Geológico das folhas de Camobi e Santa Maria – RS**. Santa Maria: [s.n.], 1988. Convênio FINEP – UFSM.

MCBRATNEY, A.B.; MENDONÇA SANTOS, M.L.; MINASNY, B. On digital soil mapping. **Geoderma**, v.117, p.3-52, 2003. DOI: https://doi.org/10.1016/S0016-7061(03)00223-4.

MCKENZIE, N.J.; RYAN, P.J. Spatial prediction of soil properties using environmental correlation. **Geoderma**, v.89, p.67-94, 1999. DOI: https://doi.org/10.1016/S0016-7061(98)00137-2.

MEIER, M.; SOUZA, E. de; FRANCELINO, M.R.; FERNANDES FILHO, E.I.; SCHAEFER, C.E.G.R. Digital soil mapping using machine learning algorithms in a tropical mountainous area. **Revista Brasileira de Ciência do Solo**, v.42, e0170421, 2018. DOI: https://doi.org/10.1590/18069657rbcs20170421.

MINASNY, B.; MCBRATNEY, A.B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, v.32, p.1378-1388, 2006. DOI: https://doi.org/10.1016/j.cageo.2005.12.009.

MOORE, I.D.; GESSLER, P.E.; NIELSEN, G.A.; PETERSON, G.A. Soil attribute prediction using terrain analysis. **Soil Science Society of America Journal**, v.57, p.443-452, 1993. DOI: https://doi.org/10.2136/sssaj1993.03615995005700020026x.

MOURA-BUENO, J.M.; DALMOLIN, R.S.D.; TEN CATEN, A.; RUIZ, L.F.C.; RAMOS, P.V.; DOTTO, A.C. Assessment of digital elevation model for digital soil mapping in a watershed with gently undulating topography. **Revista Brasileira de Ciência do Solo**, v.40, e0150022, 2016. DOI: https://doi.org/10.1590/18069657rbcs20150022.

NOLLER, J.S. Applying geochronology in predictive digital mapping of soils. In: BOETTINGER, J.L.; HOWELL, D.W.; MOORE, A.C.; HARTEMINK, A.E.; KIENAST-BROWN, S. (Ed.). **Digital soil mapping**: bridging research, environmental application, and operation. Dordrecht: Springer, 2010. p.43-53. DOI: https://doi.org/10.1007/978-90-481-8863-5_4.

PAHLAVAN-RAD, M.R.; KHORMALI, F.; TOOMANIAN, N.; BRUNGARD, C.W.; KIANI, F.; KOMAKI, C.B.; BOGAERT, P. Legacy soil maps as a covariate in digital soil mapping: a case study from Northern Iran. **Geoderma**, v.279, p.141-148, 2016. DOI: https://doi.org/10.1016/j.geoderma.2016.05.014.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.r-project.org/>. Accessed on: Nov. 13 2017.

SAMUEL-ROSA, A.; HEUVELINK, G.B.M.; VASQUES, G.M.; ANJOS, L.H.C. Do more detailed environmental covariates deliver more accurate soil maps? **Geoderma**, v.243-244, p.214-227, 2015. DOI: https://doi.org/10.1016/j.geoderma.2014.12.017.

SAMUEL-ROSA, A.; MIGUEL, P.; DALMOLIN, R.S.D.; PEDRON, F. de A. Uso da terra no Rebordo do Planalto do Rio Grande do Sul. **Ciência e Natura**, v.33, p.161-173, 2011.

SANTOS, H.G. dos; JACOMINE, P.K.T.; ANJOS, L.H.C. dos; OLIVEIRA, V.A. de; LUMBRERAS, J.F.; COELHO, M.R.; ALMEIDA, J.A. de; CUNHA, T.J.F.; OLIVEIRA, J.B. de. **Sistema brasileiro de classificação de solos**. 3.ed. rev. e ampl. Brasília: Embrapa, 2013. 353p.

SARTORI, P.L.P. Geologia e geomorfologia de Santa Maria. **Ciência e Ambiente**, v.38, p.19-42, 2009.

SCULL, P.; FRANKLIN, J.; CHADWICK, O.A. The application of classification tree analysis to soil type prediction in a desert landscape. **Ecological Modelling**, v.181, p.1-15, 2005. DOI: https://doi.org/10.1016/j.ecolmodel.2004.06.036.

SILVA, S.H.G.; MENEZES, M.D. de; OWENS, P.R.; CURI, N. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. **Geoderma**, v.267, p.65-77, 2016. DOI: https://doi.org/10.1016/j.geoderma.2015.12.025.

SOIL SURVEY STAFF. **Keys to soil taxonomy**. 12th ed. Washington: Usda, 2014. 372p.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; MINASNY, B.; TRIANTAFILIS, J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. **Geoderma**, v.253-254, p.67-77, 2015. DOI: https://doi.org/10.1016/j.geoderma.2015.04.008.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M. de L. Estatística multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo. **Pesquisa Agropecuária Brasileira**, v.46, p.554-562, 2011a. DOI: https://doi.org/10.1590/S0100-204X2011000500014.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F.A.; MENDONÇA-SANTOS, M. de L. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v.35, p.53-62, 2011b. DOI: https://doi.org/10.1590/S0100-06832011000100005.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação do uso de modelos digitais de elevação em mapeamento digital de solos em Dois Irmãos, RS, Brasil. **Revista Brasileira de Ciência do Solo**, v.38, p.1367-1376, 2014. DOI: https://doi.org/10.1590/S0100-06832014000500002.

VASQUES, G.M.; GRUNWALD, S.; MYERS, D.B. Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. **Landscape Ecology**, v.27, p.355-367, 2012. DOI: https://doi.org/10.1007/s10980-011-9702-3.

WILSON, J.P.; GALLANT, J.C. (Ed.). **Terrain analysis**: principles and applications. New York: John Wiley & Sons, 2000. 485p.

ZHANG, G.-L.; LIU, F.; SONG, X.-D. Recent progress and future prospect of digital soil mapping: a review. **Journal of Integrative Agriculture**, v.16, p.2871-2885, 2017. DOI: https://doi.org/10.1016/S2095-3119(17)61762-3.