Revista Brasileira de Ciência do Solo

Division – Soil in Space and Time | Commission – Pedometrics

Selection of Environmental Covariates for Classifier Training Applied in Digital Soil Mapping

Alcinei Ribeiro Campos[(1)*], Elvio Giasson[(2)], José Janderson Ferreira Costa[(1)], Israel Rosa Machado[(1)], Elisângela Benedet da Silva[(3)] and Benito Roberto Bonfatti[(4)]

[(1)] Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Ciência do Solo, Porto Alegre, Rio Grande do Sul, Brasil.
[(2)] Universidade Federal do Rio Grande do Sul, Departamento de Solos, Porto Alegre, Rio Grande do Sul, Brasil.
[(3)] Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina - Centro de Informações de Recursos Ambientais e de Hidrometeorologia de Santa Catarina, Florianópolis, Santa Catarina, Brasil.
[(4)] Universidade do Estado de Santa Catarina, Centro de Ciências Agroveterinárias, Departamento de Engenharia Ambiental e Sanitária, Lages, Santa Catarina, Brasil.

ABSTRACT: A large number of predictor variables can be used in digital soil mapping; however, the presence of irrelevant covariables may compromise the prediction of soil types. Thus, algorithms can be applied to select the most relevant predictors. This study aimed to compare three covariable selection systems (two filter algorithms and one wrapper algorithm) and assess their impacts on the predictive model. The study area was the Lajeado River Watershed in the state of Rio Grande do Sul, Brazil. We used forty predictor covariables, derived from a digital elevation model with 30 m resolution, in which the three selection models were applied and separated into subsets. These subsets were used to assess performance by applying four prediction algorithms. The wrapper method obtained the best performance values for the predictive model in all the algorithms evaluated. The three selection methods applied reduced the number of covariables in the predictive models by 70 % and enabled prediction of the 14 soil mapping units.

Keywords: data mining, geomorphometric variables, soil prediction.

*** Corresponding author:**
E-mail: alcineicampos@gmail.com

Received: December 27, 2017
Approved: May 21, 2018

How to cite: Campos AR, Giasson E, Costa JJF, Machado IR, Silva EB, Bonfatti BR. Selection of environmental covariates for classifier training applied in digital soil mapping. Rev Bras Cienc Solo. 2018;42:e0170414.
https://doi.org/10.1590/18069657rbcs20170414

INTRODUCTION

The primary predictor covariables used in digital soil mapping (DSM) are derived from digital elevation models (DEMs), which can be used to characterize the environment on a number of detailed scales (ten Caten et al., 2012). Digital elevation models can be used to obtain covariables that characterize the relief and have a direct or indirect relationship with other pedogenetic factors from the Scorpan model (McBratney et al., 2003).

The level of agreement obtained on the maps predicted by DSM remain low even when a large number of predictor variables are used. Average overall accuracy values are around 65 %, and it is not always possible to predict all the soil classes that make up the reference maps (Behrens et al., 2010; Coelho and Giasson, 2010; ten Caten et al., 2012). These accuracy values may be attributed to the low correlation between predictor covariables and soil classes, but there is still no consensus on which covariables should be used, as observed in the range of combinations used in DSM studies (Campos et al., 2010; ten Caten et al., 2012; Höfig et al., 2014; Teske et al., 2015; Dias et al., 2016).

An awareness lack of the relevance of predictor covariables in discriminating soil classes has led to the use of sets with an insufficient number of covariables to make a prediction or excessively large sets with redundant covariables. Large sets of covariables together with the presence of redundant variables increase the complexity of the predictive model and hinder prediction for some soil classes, reducing map accuracy and precluding understanding of the relationships between predictors and soil types (Guyon and Elisseeff, 2003; Behrens et al., 2010; ten Caten et al., 2012; Brungard et al., 2015).

Selection algorithms can be applied to attenuate the complexity and low performance of predictive models and reduce the number of predictor covariables (Guyon and Elisseeff, 2003; Hall and Holmes, 2003; Coelho and Giasson, 2010; Giasson et al., 2013). The main selection algorithms applied to data mining can be grouped into wrapper (*envelope*), filter, and embedded methods (integrated) (Guyon and Elisseeff, 2003; Hall and Holmes, 2003).

Wrapper algorithms select the predictor variables by assessing their relevance via induction of a predictive model. These models select by subtracting or adding covariables to the set and estimating performance indices of the respective predictive model, until achieving the smallest subset of predictors with performance greater than or equal to the set composed of all the predictor covariables under study (Guyon and Elisseeff, 2003; Hall and Holmes, 2003). This type of selection exhibits maximum performance in the classifier used to select the predictor covariables.

Filter-type algorithms are applied independently of the predictive model, and the selection criteria are parameters, such as correlation, distance, information gained, and covariable consistency (Hall, 1999; Dash et al., 2000; Guyon and Elisseeff, 2003). This type of selection is commonly used to select variables (Giasson et al., 2013; Paes et al., 2013; Subburayalu and Slater, 2013; Subburayalu et al., 2014; Taghizadeh-mehrjardi et al., 2016; Vasu and Lee, 2016) and the results can be applied to any prediction algorithm (Dash et al., 2000; Guyon and Elisseeff, 2003; Hall and Holmes, 2003). Embedded methods are integrated with learning models and are specific to these classifiers (Guyon and Elisseeff, 2003; Paes et al., 2013).

Applying each selection method results in different subsets of covariables and, consequently, affects the predictive ability of the model. Studies have demonstrated that the wrapper algorithm obtains the best result when the predictive model is hierarchical (Hall and Holmes, 2003; Brungard et al., 2015). Although hierarchical methods are mainly used in DSM as the classification method, the literature reports a predominance of filter-type algorithms for selecting predictor covariables since the procedures are faster and do not depend on the predictive model (Coelho and Giasson, 2010; Giasson et al., 2013;

Subburayalu and Slater, 2013; Subburayalu et al., 2014; Taghizadeh-mehrjardi et al., 2016). Applying filter-type models may result in sets of covariables weakly correlated with soil types, explaining the low accuracy values observed in the literature.

Thus, the wrapper algorithm is an alternative that could result in the selection of sets of covariables more associated with soil types, in turn, more accurate predictive models (Brungard et al., 2015). The present study aimed to compare three systems for selection of predictor covariables (two filter-type algorithms and one wrapper selection system) and assess their impacts on the predictive model.

MATERIALS AND METHODS

The study area was the Lajeado Grande River Watershed (Figure 1). The watershed encompasses an area of approximately 532 km^2 in the extreme northwest of the state of Rio Grande do Sul - Brazil, in the Alto Uruguai Hydrographic Region (Freitas et al., 2012).

The climate of the region is humid subtropical (Cfa according to the Köppen climate classification system), with average annual rainfall of 1,778 mm and temperature of 18.5 °C. The geology corresponds to the Parana Province, characterized primarily by basalt flows from the Serra Geral formation (Bagatini et al., 2015). The soil map of the area (Figure 1) has a scale of 1:50,000 (Kämpf et al., 2004), composed of 14 soil map units (MUs) (Table 1).

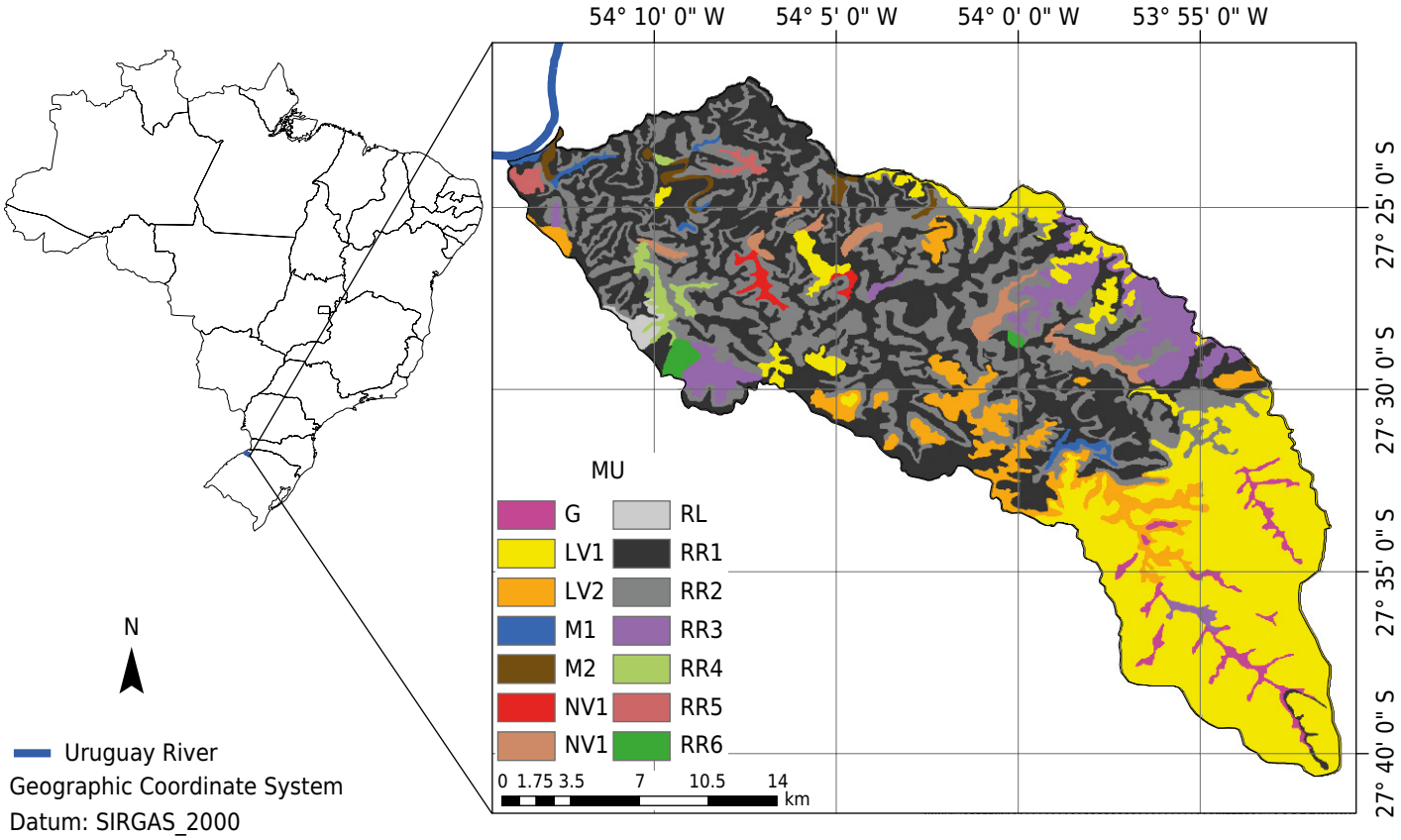

Figure 1. Location and soil map of Lajeado Grande River Watershed. MU = soil map units. G = *Gleissolos Melânico eutrófico típico*; LV1 = *Latossolo Vermelho distroférrico típico*; LV2 = *Latossolo Vermelho distroférrico típico + Neossolo Regolítico eutrófico típico*; M1 = *Chernossolo Argilúvico férrico típico*; NV1 = *Nitossolo Vermelho distroférrico típico*; NV2 = *Nitossolo Vermelho distroférrico típic + Neossolo Regolítico eutrófico típico*; RL = *Neossolo Litólico eutrófico típico + Neossolo Regolítico eutrófico típico*; RR1 = *Neossolo Regolítico eutrófico típico*; RR2 = *Neossolo Regolítico eutrófico típico + Neossolo Litólico eutrófico típico*; RR3 = *Neossolo Regolítico eutrófico típico + Latossolo Vermelho distroférrico típico*; RR4 = *Neossolo Regolítico eutrófico típic + Chernossolo Argilúvico férrico típico*; RR5 = *Neossolo Regolítico eutrófico típico + Cambissolo Háplico Ta eutrófico típico + Nitossolo Vermelho distroférrico típico*; RR6 = *Neossolo Regolítico eutrófico típico* + rock beds.

Table 1. Soil map units (MUs) in the Lajeado Grande River Watershed

MU	Composition of soil map units	Area	Area
		km^2	%
G	*Gleissolos Melânico eutrófico típico*[(1)]	8.57	1.61
LV1	*Latossolo Vermelho distroférrico típico*[(2)]	144.85	27.21
LV2*	*Latossolo Vermelho distroférrico típico*[(2)] + *Neossolo Regolítico eutrófico típico*[(4)]	31.09	5.84
M1	*Chernossolo Argilúvico férrico típico*[(3)]	3.57	0.67
M2*	*Chernossolo Argilúvico férrico típico*[(3)] + *Neossolo Regolítico eutrófico típico*[(4)]	3.89	0.73
NV1	*Nitossolo Vermelho distroférrico típico*[(2)]	2.40	0.45
NV2*	*Nitossolo Vermelho distroférrico típico*[(2)] + *Neossolo Regolítico eutrófico típico*[(4)]	8.94	1.68
RL*	*Neossolo Litólico eutrófico típico*[(4)] + *Neossolo Regolítico eutrófico típico*[(4)]	2.18	0.41
RR1	*Neossolo Regolítico eutrófico típico*[(4)]	161.19	30.28
RR2*	*Neossolo Regolítico eutrófico típico*[(4)] + *Neossolo Litólico eutrófico típico*[(5)]	127.28	23.91
RR3*	*Neossolo Regolítico eutrófico típico*[(4)] + *Latossolo Vermelho distroférrico típico*[(2)]	28.00	5.26
RR4*	*Neossolo Regolítico eutrófico típico*[(4)] + *Chernossolo Argilúvico férrico típico*[(3)]	4.58	0.86
RR5*	*Neossolo Regolítico eutrófico típico*[(4)] + *Cambissolo Háplico Ta eutrófico típico*[(5)] + *Nitossolo Vermelho distroférrico típico*[(2)]	2.77	0.52
RR6*	*Neossolo Regolítico eutrófico típico*[(4)] + rock beds	2.61	0.49

*Associations of soil. [(1)] Endoaquent, [(2)] Hapludox, [(3)] Hapludol, [(4)] Udorthents, [(5)] Dystrudept (Soil Survey Staff, 2014).

Forty predictor covariables (Table 2) were derived from a digital elevation model (DEM) that was produced based on ASTER/GDEM v2 sensor data (Global Digital Elevation Models), with spatial resolution of 30 m, dated October 17, 2011, and obtained from the American Geological Service (Tachikawa et al., 2011). The geoform covariable was derived from the LandMapR tool package (MacMillan, 2003). The drainage density index (km km^{-2}) was obtained with the Raster Calculator extension of ArcGIS 9.2, following the methodology proposed by Cardoso et al. (2006). The other covariables were derived using RSAGA v 2.2.2 (Brenning et al., 2018), integrated into R 3.3.1 software (R Development Core Team, 2018).

Next, all the predictor covariables were sampled together with the response variable (soil map units), following a stratified sampling scheme, with approximately 30,000 points (ten Caten et al., 2013; Bagatini et al., 2015). The samples were stratified based on the number and size of the polygons of each soil map unit. To that end, simulations were made with different numbers of points in each MU, and a minimum of 300 samples was established for the smaller MUs. Thus, 3,000 points were randomly distributed in the nine units with an area of less than 10 km^2 (Table 1), and the remaining points were randomly stratified, obtaining approximately 300 samples in the smaller area and 6,000 in the larger.

Three selection methods were applied to the dataset containing 40 predictor covariables (CJ40), and the soil map units and the following four subsets were separated:

Subset 1 = selected to apply the Correlation-based Feature Selection (CFS) algorithm (Hall and Smith, 1999); this algorithm performs a heuristic assessment based on a correlation in order to find subsets containing covariables highly correlated with the class and not correlated with one another, while covariables with strong intercorrelation are considered redundant and excluded;

Table 2. Covariables extracted from the digital elevation model

Predictor covariables	
Negative Openness (NEGOP)	Analytical Hillshading (HILLSAHD)
Positive Openness (POSOP)	Convergence Index (CONVINDEX)
Total Catchment Area (CATAREA)	Drainage Density Index (DDI)
Catchment Slope (CATSLOPE)	Stream Power Index (SPI)
Convexity (CONVEXITY)	Topographic Position Index (TPI)
Curvature (CURVATURE)	Topographic Wetness Index (TWI)
Profile Curvature (PROCURV)	Terrain Ruggedness Index (TRI)
Longitudinal Curvature (LOGCURV)	Diffuse Insolation (DIFINSOL)
Maximum Curvature (CURVMAX)	Direct Insolation (DIRINSOL)
Minimum Curvature (MINCURV)	Total Insolation (TOTINSOL)
Slope (DECL)	Generalized Surface (MDEGENER)
Slope Height (SLHEIGHT)	Multiresolution Ridge Top Flatness Index (MRRTF)
Channel Network Distance (CND)	Multiresolution Index of Valley Bottom Flatness (MRVBF)
Euclidean distance of the rivers (DISTRIVERS)	Channel network base level (CNBL)
Elevation (MDE)	Aspect (ASPECT)
Standardized Height (MDEHEIGTH)	Mid-Slope Position (MSP)
Normalized Height (NORHEIGTH)	Relative Slope Position (RSP)
Flow Accumulation (FLOWACU)	Valley Depth (VALLEYDEP)
Morphometric Features (MORFFEAT)	Direct to Diffuse Ratio (DDR)
Landforms (LANDFORMS)	Cross-Sectional Curvature (CSC)

Subset 2 = selected to apply the Consistency Subset Eval (CSE) (Liu and Setiono, 1996). The CSE uses the class consistency rate to select the properties that divide the original dataset into subsets containing most of the classes, and uses the consistency assessor proposed by Liu and Setiono (1996). The CFS and CSE aimed to compare filter selection by correlation and consistency of the data. To that end, the Best First internal method (D1-N5) was applied in the respective algorithms, using Weka 3.8.0 software (Hall et al., 2009);

Subset 3 = this subset was selected following the wrapper selection principles (Hall and Smith, 1999). An R programming language script composed of predictive models with the J48 algorithm (J48 -C 0.25 -M 2) was written and cross validated by five blocks. The recursive command “while” was used, whereby the predictive model was repeated and covariables eliminated one by one until reaching a minimum number of covariables in the model that exhibited overall accuracy greater than or equal to the set with all the covariables (CJ40);

Subset 4 = composed of covariables selected simultaneously in subsets 1, 2, and 3.

The set of predictor covariables (subsets 1, 2, 3, 4, and CJ40) was used to assess performance by applying four prediction algorithms (J48, REPTree, BFTree, and the Multilayer Perceptron) used in earlier studies (Coelho and Giasson, 2010; Giasson et al., 2011; Sarmento et al., 2012; Arruda et al., 2013; Giasson et al., 2013; ten Caten et al., 2013; Calderano Filho et al., 2014; Dias et al., 2016). These algorithms were selected to compare covariable performance in classifiers with different architecture, the first three with decision tree architecture and the last one with artificial neural networks (ANNs).

For the decision trees, a minimum of five instances per leaf were used. All the algorithms were cross-validated by five blocks, applying Weka 3.8.0 software (Hall et al., 2009).

The results were assessed by an error matrix using Kappa's coefficient, mapping accuracy (MA), overall accuracy (OA) (Congalton, 1991), and the following evaluators: area under the precision-recall curve (PRC), Matthews correlation coefficient (MCC) (Saito and Rehmsmeier, 2015), mean absolute error (MAE), and root mean square error (RMSE), the last two calculated based on the likelihood of the occurrence of the data observed and estimated for each class (Shi, 2007).

RESULTS AND DISCUSSION

Of the 40 covariables extracted from the DEM, only 20 were able to predict soil type, having been selected to make up some of the subsets (Table 3). It means that half the covariables under study exhibited a weak correlation with the spatial distribution of soils or were redundant, therefore they were discarded.

Subset 3, selected by applying the wrapper algorithm, resulted in a predictive model with higher overall accuracy and Kappa coefficient than the values obtained in subsets 1 and 2, selected by CFS and CSE filters (Figure 2). This behavior was observed for the four algorithms, indicating no interaction between the selection methods and respective prediction algorithms.

The smallest predictive model, containing 21 layers, was obtained by the neural network algorithm and the largest (2142 leaves) by J48 (Table 4). This difference is largely due to the architecture adopted by these classifiers. Neural networks are composed of neurons stacked in layers that make the connection between the layers and the neurons themselves, whereas the other algorithms display decision tree architecture consisting of nodes that, when combined, give rise to the leaves, which leads to a larger model than that obtained by the neural networks (Witten et al., 2011). The architecture and size of the predictive models are directly linked to their complexity, ideal models exhibiting high predictive power and complexity that allows them to be interpreted (Ruiz et al., 2014). The apparently smaller ANNs show greater complexity and prevent complete understanding of the nature of the data under analysis (ten Caten et al., 2012).

None of the selection methods assessed showed a significant decrease in the size of the predictive model compared to the model obtained with all the covariables. For the BFTree and J48 algorithms, the smallest predictive models were obtained from subset 3, selected by the wrapper method, whereas for the REPTree, the smallest model was

Table 3. Predictor covariables selected to make up the respective subsets

Sets	Predictor variables selected
Subset 1	Drainage density index (DDI); Channel Network Base Level (CNBL); Aspect (ASPECT); Channel Network Distance (CND); Multiresolution Index of Valley Bottom Flatness (MRVBF); Slope Catchment (CATAREA); Diffuse Insolation (DIFINSOL); Generalized Surface (MDEGENER)
Subset 2	Drainage Density Index (DDI); Channel Network Base Level (CNBL); Aspect (ASPECT); Channel Network Distance (CND); Multiresolution Index of Valley Bottom Flatness (MRVBF); Slope Catchment slope (CATAREA); Curvature (CURV); Longitudinal Curvature (LOGCURV); Elevation (MDE); Topographic Wetness Index (TWI); Valley Depth (VALLEYDEP); Direct to Diffuse Ratio (DDR)
Subset 3	Drainage density index (DDI); Channel Network Base Level (CNBL); Aspect (ASPECT); Diffuse Insolation (DIFINSOL); Generalized Surface (MDEGENER); Positive Openness (POSOP); Convexity (CONVEXITY); Slope (Slope); Euclidean Distance of the Rivers (DISTRIVERS); Convergence Index (INDCONER); Mid-Slope Positon (MSP)
Subset 4	Drainage Density Index (DDI); Channel Network Base Level (CNBL); Aspect (ASPECT)

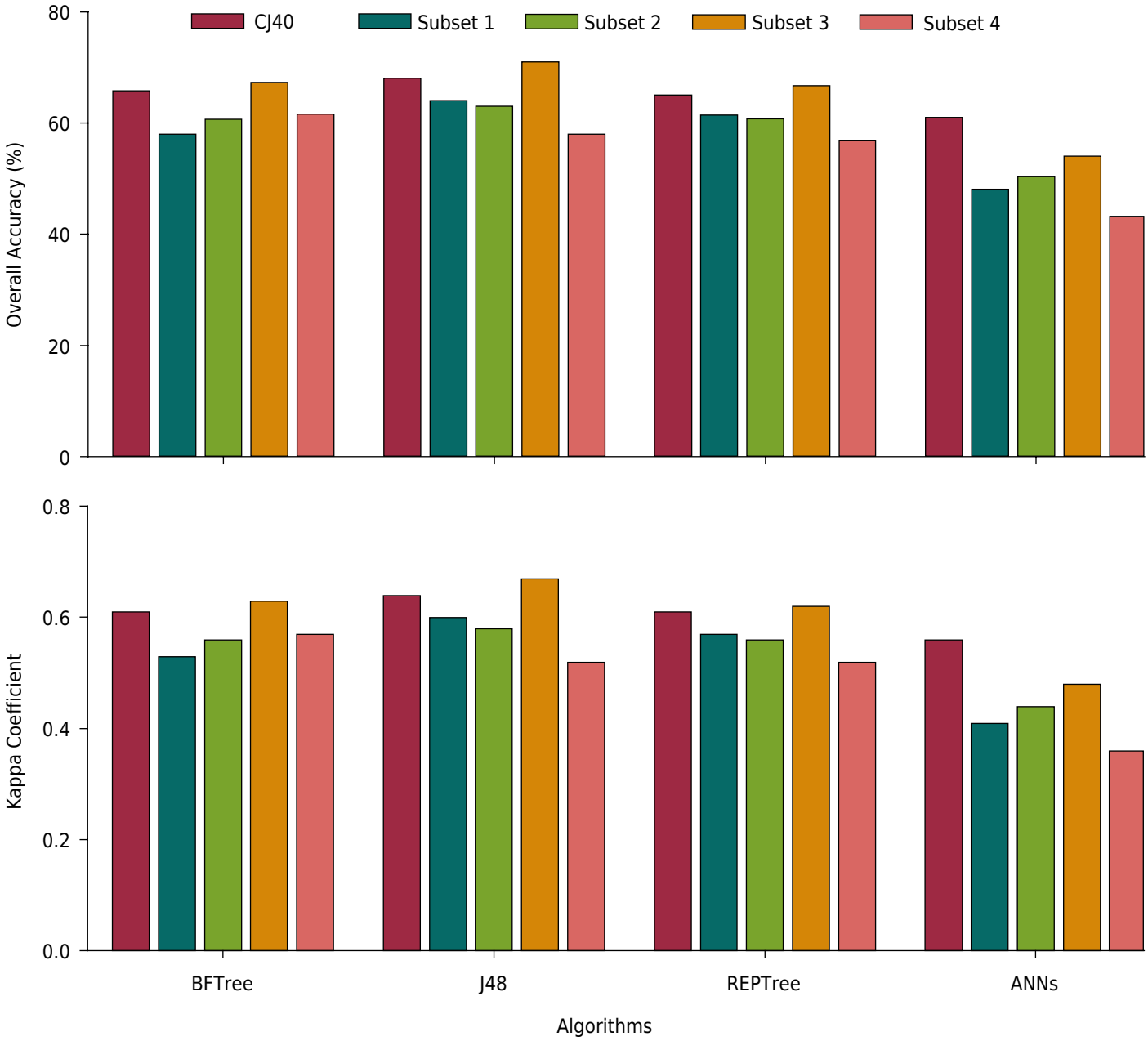

Figure 2. Performance of covariable sets using four prediction algorithms (ANNs = Artificial neural networks).

Table 4. Size of predictive models for the sets of predictor variables

Set	Number of variables	Algorithm			
		BFTree[(1)]	J48[(1)]	REPTree[(1)]	RNA[(2)]
CJ40	40	1048	1969	1579	40
Subset 1	7	1376	1979	1487	23
Subset 2	12	1243	2142	1647	26
Subset 3	11	1162	1923	1601	25
Subset 4	3	1269	1891	1693	21

[(1)] Number of leaves. [(2)] Number of layers.

obtained from subset 1, selected by application of the CFS filter. However, it is important to underscore that only in J48 did a decline in the size of the predictive model not result in lower overall accuracy and Kappa coefficient values.

Of the predictor covariables tested, only slope orientations (ASPECT), channel network base level (CNBL), and drainage density index (DDI) were selected simultaneously by the three selection algorithms (subset 4). The simultaneous selection of these covariables is due to their strong association with the spatial distribution of soils. Slope orientation (Figure 3a) shows a direct effect on the microclimate, changing water availability and biological activity in pedogenesis, thereby correlating with the spatial distribution of soils in the landscape (Schaetzl and Anderson, 2005).

The covariable channel network base level (Figure 3b) is an intermediate variable used to calculate the vertical distance from the drainage network. This covariable is obtained by relating the vertical distance to the channel network base level, and is used to indicate the depth of soils influenced by underground waters (Bock and Köthe, 2008). The CNBL exhibits the lowest values in enclosed valleys, indicating that in these areas the soil surface occurs closer to the base level of the drainage network, helping to separate soils that occur in the enclosed valleys of the soils that occur in the upland areas (Bock and Köthe, 2008). The study area contains MUs formed by associations with predominant *Latossolos* (Hapludox) in upland areas and *Neossolos Regolíticos* (Udorthents) on valley slopes. These two groups of soils account for approximately 70 % of the total study area, making the CNBL covariable important in differentiating these soils.

The drainage density index (Figure 3c) shows the relationship with the properties that influence water infiltration into the soil, such as depth, texture, and drainage, and is used to differentiate soils in terms of drainage. In areas with a high DDI, there is theoretically less infiltration and, consequently, shallower soils or soils with limited percolation in the B horizon (Demattê and Demétrio, 1998; Dobos et al., 2000).

The three selection methods resulted in different combinations of predictor covariables. This occurred due to the criteria used by each method for selecting the covariables most important for predicting soil type. Application of the CFS filter eliminated the largest number of covariables. Of the 40 covariables under study, only seven were selected to form subset 1; thus, CFS led to an 80 % decline in the number of variables. The correlation matrix of subset 1 (Figure 4) shows that fewer than 30 % of correlations exhibit a magnitude greater than 0.4, which are classified as moderate or strong (Dancey and Reidy, 2006). The selection criterion of the CFS filter is the weak correlation between covariables and high correlation with the response variables, which contributes to the selection of poorly intercorrelated covariables (Hall and Holmes, 2003).

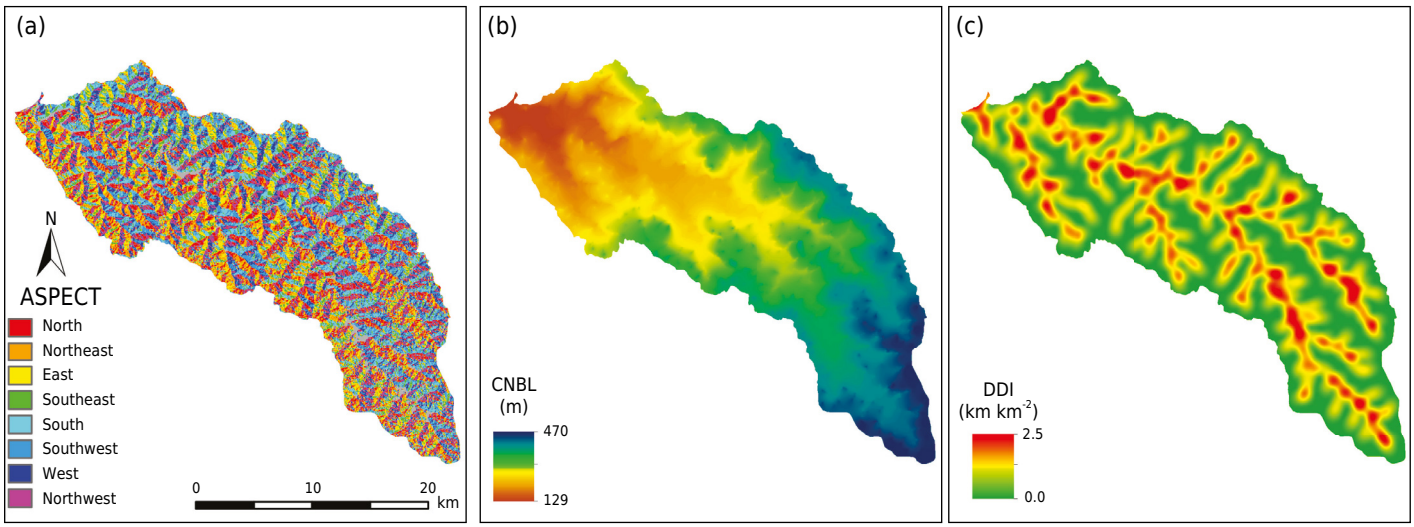

Figure 3. Covariables selected in the three methods tested. Slope orientation (a); channel network base level (b); drainage density index (c).

The CFS was highly efficient in reducing the number of predictors; however, it decreased the performance of the predictive model. The highest correlation (0.96) occurred between the CNBL and MDEGENER covariables, due to the way the former is calculated, using the horizontal distance from the drainage network and the vertical distance (elevation) from the base of the drainage network, resulting in a strong correlation between them (Figure 4).

Applying the CSE filter resulted in subset 2, with 12 covariables selected, which represented a 70 % decline in the number of covariables. In this subset, fewer than 20 % of the correlations are greater than 0.4 (Figure 5a). The VALLEYDEP and MDEPAD covariables, collinearity of 0.63, and CND and MDEPAD covariables, collinearity of -0.71, were the strongest correlation values observed in subset 2. The selection criterion of the CSE algorithm is the consistency of the subsets in relation to the response variable, which allowed higher correlation values between the selected covariables. Nevertheless, applying these filters eliminated the highest number of covariables with strong intercorrelation (Hall and Holmes, 2003). However, as in the CFS, the CSE also reduced the performance of the predictive model.

Subset 3, selected by applying the wrapper approach, consisted of 11 covariables. In this subset, 35 % of the correlations are greater than 0.4 (Figure 5b), this method being the least efficient in eliminating strongly correlated covariables. The highest collinearities occur between covariables CNBL and MDEGENER (0.95) and PRDECL and CONVINDEX (0.85). Only the wrapper selection did not decrease the performance of the predictive model in relation to the set with all 40 covariables. Moreover, a slight decline in the size of the predictive model was also observed in algorithms BFTree, REPTree, and J48. This result corroborates the study by Hall and Holmes (2003), who also obtained a better prediction result when the wrapper method was applied to select the predictor variables.

The presence of strongly intercorrelated covariables in all the subsets indicates that they may exhibit different degrees of importance for the soil classes, justifying maintaining them in the subsets. This result shows that the correlation between covariables alone is insufficient to select the predictors most relevant for predicting soil types.

	ASPECT	CNBL	CND	DDI	DIFINSOL	POSOP	MDEGENER
ASPECT	1	0.09	−0.1	0.06	−0.03	0.02	0.06
CNBL	0.09	1	−0.08	−0.29	−0.68	0.39	0.96
CND	−0.1	−0.08	1	−0.54	−0.29	0.32	0.21
DDI	0.06	−0.29	−0.54	1	0.39	−0.22	−0.44
DIFINSOL	−0.03	−0.68	−0.29	0.39	1	0.01	−0.75
POSOP	0.02	0.39	0.32	−0.22	0.01	1	0.46
MDEGENER	0.06	0.96	0.21	−0.44	−0.75	0.46	1

Figure 4. Correlation matrix of covariables that form subset 1.

(a)

	CATSLOPE	ASPECT	CNBL	CND	CURVATURE	LOGCURV	DDR	DDI	MRVBF	NORHEIGTH	TWI	VALLEYDEP
CATSLOPE	1	−0.03	−0.31	0	−0.19	−0.31	−0.23	0.02	−0.45	−0.25	−0.01	0.1
ASPECT	−0.03	1	0.09	−0.1	0	−0.03	0.03	0.06	0.02	0.01	0.01	0.01
CNBL	−0.31	0.09	1	−0.08	−0.01	0.14	0.25	−0.29	0.04	0.56	−0.03	−0.38
CND	0	−0.1	−0.08	1	0.15	0.62	0.06	−0.54	−0.44	0.63	−0.2	−0.56
CURVATURE	−0.19	0	−0.01	0.15	1	0.19	0.02	0.04	−0.09	0.2	−0.1	−0.11
LOGCURV	−0.31	−0.03	0.14	0.62	0.19	1	0.08	−0.32	−0.25	0.6	−0.17	−0.56
DDR	−0.23	0.03	0.25	0.06	0.02	0.08	1	−0.14	0.03	0.18	−0.01	−0.16
DDI	0.02	0.06	−0.29	−0.54	0.04	−0.32	−0.14	1	0.35	−0.51	0.08	0.46
MRVBF	−0.45	0.02	0.04	−0.44	−0.09	−0.25	0.03	0.35	1	−0.37	0.16	0.52
NORHEIGTH	−0.25	0.01	0.56	0.63	0.2	0.6	0.18	−0.51	−0.37	1	−0.2	−0.71
TWI	−0.01	0.01	−0.03	−0.2	−0.1	−0.17	−0.01	0.08	0.16	−0.2	1	0.21
VALLEYDEP	0.1	0.01	−0.38	−0.56	−0.11	−0.56	−0.16	0.46	0.52	−0.71	0.21	1

(b)

	ASPECT	CNBL	CONVEXITY	DISTRIVERS	DDI	CONVINDEX	DIFINSOL	POSOP	RSP	SLOPE	MDEGENER
ASPECT	1	0.09	−0.05	0.01	0.06	−0.05	−0.03	0.02	−0.08	−0.05	0.06
CNBL	0.09	1	0.15	0.49	−0.29	0.12	−0.68	0.39	0.15	−0.27	0.96
CONVEXITY	−0.05	0.15	1	0.37	−0.62	0.75	−0.4	0.23	0.75	0.27	0.35
DISTRIVERS	0.01	0.49	0.37	1	−0.51	0.28	−0.42	0.25	0.36	−0.06	0.54
DDI	0.06	−0.29	−0.62	−0.51	1	−0.56	0.39	−0.22	−0.67	−0.09	−0.44
CONVINDEX	−0.05	0.12	0.75	0.28	−0.56	1	−0.35	0.34	0.85	0.14	0.35
DIFINSOL	−0.03	−0.68	−0.4	−0.42	0.39	−0.35	1	0.01	−0.4	−0.32	−0.75
POSOP	0.02	0.39	0.23	0.25	−0.22	0.34	0.01	1	0.46	−0.56	0.46
RSP	−0.08	0.15	0.75	0.36	−0.67	0.85	−0.4	0.46	1	0.13	0.4
SLOPE	−0.05	−0.27	0.27	−0.06	−0.09	0.14	−0.32	−0.56	0.13	1	−0.19
MDEGENER	0.06	0.96	0.35	0.54	−0.44	0.35	−0.75	0.46	0.4	−0.19	1

Figure 5. Correlation matrix for subsets 2 (a) and 3 (b).

In all the subsets of variables and algorithms tested, it was possible to predict the 14 map units. Mapping Accuracy (MA) behaved identically to that of overall accuracy (OA), with the highest values observed in subset 3 associated with algorithm J48 (Table 5). The lowest MA values (<0.3) occurred in the BFTree and ANN algorithms associated with subsets 1 and 2. The greatest variation between the minimum and maximum MA value was observed in these two subsets. The greater variation between minimum and maximum MA in subsets 1 and 2 indicates that the prediction error was concentrated in only a few map units.

The opposite behavior can be observed for algorithm J48, which displayed the lowest variation between minimum (0.50) and maximum (0.90) MA values in CJ40 and subset 3, suggesting greater error distribution among the MUs and, in turn, better prediction of the 14 soil map units. Only map units LV2, NV2, and RL did not reach maximum MA in J48 associated with subset 3; however, the MA values observed for this combination between the prediction algorithm and subset were higher than 0.67, indicating good prediction for these three soil map units also.

The map units with the greatest difficulty for prediction were NV2, LV2, RR1, RR2, RR3, and RR6, which showed the lowest MA values. These map units (NV2, LV2, RR1, RR2, RR3, and RR6) consist of Typic Udorthents (*Neossolos Regolíticos*) or these classes in association with other classes, such as Lithic Udorthents (*Neossolos Litólicos*) and Typic Haplustox (*Latossolos Vermelhos*). Thus, these MUs may occur adjacently or in similar environments, hindering their prediction by classifiers. Map units occupying very similar positions in the landscape may cause more discrimination difficulties for predictive models (Höfig et al., 2014).

The performance of the subsets and algorithms is also supported by the values obtained in the other indices evaluated (Table 6). Except for the ANN algorithm, the values of the area under the PRC curve, Kappa coefficient, and OA exhibit a variation of less than 5 % in the predictive models, indicating low randomness in the classifications obtained by these models, showing good consistency in the covariables used to predict the occurrence of the MUs. The results for subset 4 demonstrate that only the three covariables selected simultaneously in the three methods tested account for more than 56 % of the agreement

obtained in the decision tree models and 43 % in the ANN model, reinforcing the relevance of these covariables in predicting soil type.

The mean absolute error (MAE) and root mean square error (RMSE) values were relatively low, the highest being 0.10 and 0.22, respectively, confirming the good performance of the predictive models (Witten et al., 2011). The observed values of MCC indicate good correlation between the predicted data and the reference map, with the highest value (0.67) in algorithm J48 associated with subset 3. For the MCC, only the ANN algorithm obtained values below 0.5; however, it is important to underscore that the MCC is better used when working with a balanced number of samples from the response variable. In the present study, since the number of samples per class range from 300 to 6,000 between the smaller and larger soil classes (RL and RR1), the values observed may have been underestimated (Saito and Rehmsmeier, 2015).

The three selection methods resulted in different combinations of the predictor variables and, in turn, different validation performances; the wrapper type exhibited slightly better performance than that of the other methods in all the algorithms tested, corroborating the studies by Hall and Smith (1999) and Brungard et al. (2015). All the selection methods led to fewer covariables, reinforcing the need for data pre-processing in order to maximize the performance of the classifiers used in digital soil maps, which is in agreement with

Table 5. Mapping Accuracy for predictive models fit to the five sets of predictor covariables.

Algorithm	Soil map unit													
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6
CJ40														
BFTree	0.75	0.72	0.74	0.70	0.69	0.73	0.32	0.65	0.45	0.59	0.45	0.65	0.76	0.31
J48	0.79	0.79	0.64	0.79	0.82	0.83	0.65	0.89	0.50	0.55	0.65	0.75	0.85	0.79
REPTree	0.79	0.77	0.61	0.75	0.76	0.84	0.56	0.89	0.42	0.56	0.63	0.73	0.80	0.77
RNA	0.80	0.71	0.51	0.57	0.56	0.69	0.49	0.91	0.45	0.55	0.55	0.71	0.84	0.68
Subset 1 (CFS filter)														
BFTree	0.69	0.74	0.79	0.73	0.66	0.78	0.43	0.64	0.24	0.56	0.25	0.63	0.67	0.51
J48	0.79	0.76	0.56	0.78	0.79	0.82	0.64	0.86	0.43	0.55	0.57	0.68	0.81	0.77
REPTree	0.80	0.75	0.56	0.68	0.78	0.83	0.59	0.85	0.34	0.55	0.53	0.68	0.79	0.72
RNA	0.63	0.65	0.39	0.21	0.44	0.70	0.20	0.59	0.38	0.56	0.40	0.46	0.44	0.20
Subset 2 (CSE filter)														
BFTree	0.68	0.76	0.72	0.67	0.63	0.76	0.47	0.65	0.33	0.50	0.18	0.64	0.55	0.51
J48	0.77	0.76	0.54	0.77	0.77	0.80	0.59	0.83	0.43	0.53	0.56	0.69	0.77	0.78
REPTree	0.82	0.75	0.55	0.67	0.79	0.80	0.56	0.85	0.36	0.53	0.51	0.69	0.76	0.72
RNA	0.70	0.66	0.40	0.29	0.46	0.63	0.28	0.64	0.38	0.59	0.43	0.54	0.51	0.22
Subset 3 (wrapper selection)														
BFTree	0.79	0.74	0.73	0.73	0.67	0.72	0.29	0.68	0.46	0.58	0.46	0.67	0.75	0.34
J48	0.83	0.80	0.68	0.84	0.84	0.83	0.69	0.90	0.52	0.59	0.67	0.75	0.88	0.84
REPTree	0.81	0.78	0.64	0.76	0.80	0.84	0.64	0.90	0.43	0.56	0.65	0.74	0.80	0.82
RNA	0.75	0.69	0.48	0.33	0.41	0.62	0.34	0.77	0.40	0.59	0.40	0.57	0.71	0.43
Subset 4 (1∩2∩3)														
BFTree	0.71	0.72	0.48	0.76	0.67	0.75	0.52	0.85	0.36	0.43	0.58	0.66	0.72	0.68
J48	0.65	0.73	0.47	0.75	0.75	0.73	0.62	0.82	0.38	0.41	0.51	0.67	0.73	0.76
REPTree	0.71	0.74	0.47	0.77	0.73	0.74	0.51	0.86	0.37	0.35	0.54	0.68	0.70	0.69
RNA	0.46	0.68	0.31	0.22	0.43	0.56	0.26	0.62	0.37	0.25	0.32	0.51	0.63	0.31

Table 6. Performance of the sets of predictor variables in each prediction algorithm

Set	Algorithm	MAE	RMSE	Area PRC	MCC	Kappa	OA
							%
CJ40	BFTree	0.06	0.19	0.63	0.61	0.61	65.76
	J48	0.05	0.19	0.64	0.63	0.64	68.00
	REPTree	0.06	0.19	0.66	0.60	0.61	65.00
	RNA	0.07	0.20	0.61	0.56	0.56	60.97
Subset 1	BFTree	0.07	0.20	0.58	0.52	0.53	57.95
	J48	0.07	0.21	0.56	0.52	0.60	64.00
	REPTree	0.07	0.20	0.62	0.56	0.57	61.41
	RNA	0.09	0.22	0.49	0.41	0.41	48.06
Subset 2	BFTree	0.07	0.20	0.59	0.55	0.56	60.63
	J48	0.06	0.20	0.60	0.57	0.58	63.00
	REPTree	0.07	0.20	0.62	0.55	0.56	60.77
	RNA	0.09	0.21	0.51	0.44	0.44	50.33
Subset 3	BFTree	0.06	0.19	0.65	0.63	0.63	67.32
	J48	0.05	0.18	0.68	0.67	0.67	71.00
	REPTree	0.06	0.19	0.68	0.62	0.62	66.66
	RNA	0.08	0.21	0.54	0.48	0.48	54.05
Subset 4	BFTree	0.07	0.20	0.60	0.56	0.57	61.61
	J48	0.06	0.20	0.62	0.60	0.52	58.00
	REPTree	0.08	0.20	0.58	0.51	0.52	56.90
	RNA	0.10	0.22	0.42	0.36	0.36	43.25

MAE = mean absolute error; RMSE = root mean square error; Area PRC = area under the precision/recall curve; MCC = Matthews correlation coefficient; OA = overall accuracy.

other studies (Giasson et al., 2013; Paes et al., 2013; Subburayalu and Slater, 2013; Subburayalu et al., 2014; Taghizadeh-mehrjardi et al., 2016; Vasu and Lee, 2016). Wrapper selection assesses a model that predicts the relevance of each covariable independently, which made it possible to select variables highly correlated with the response variable and achieve maximum performance in the algorithms used for classification.

CONCLUSIONS

The wrapper selection method produced the best performance for the predictive model in all the algorithms.

Applying the three selection methods reduced the number of covariables by approximately 70 % in the predictive models and made it possible to predict the 14 map units.

Applying the Correlation-based Feature Selection (CFS) and Consistency Subset Eval (CSE) filter algorithms decreased accuracy and Kappa coefficient values in relation to the set of all the covariables.

Only the covariables of slope orientation, channel network base level, and drainage density index were selected simultaneously by the three methods tested.

REFERENCES

Arruda GP, Demattê JAM, Chagas CS. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. Rev Bras Cienc Solo. 2013;37:327-38. https://doi.org/10.1590/S0100-06832013000200004

Bagatini T, Giasson E, Teske R. Seleção de densidade de amostragem com base em dados de áreas já mapeadas para treinamento de modelos de árvore de decisão no mapeamento digital de solos. Rev Bras Cienc do Solo. 2015;39:960-7. https://doi.org/10.1590/01000683rbcs20140289

Behrens T, Zhu A-X, Schmidt K, Scholten T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma. 2010;155:175-85. https://doi.org/10.1016/j.geoderma.2009.07.010

Bock M, Köthe R. Predicting the depth of hydromorphic soil characteristics influenced by ground water. In: SAGA-seconds Out. Hamburg: Universität Hamburg, Institut für Geographie; 2008. v. 2. p. 13-22.

Brenning A, Bangs D, Becker M. RSAGA: SAGA geoprocessing and terrain analysis. R package; 2018. Available from: https://CRAN.R-project.org/package=RSAGA

Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards Jr TC. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma. 2015;239-240:68-83. https://doi.org/10.1016/j.geoderma.2014.09.019

Calderano Filho SB, Polivanov H, Chagas CS, Carvalho Júnior W, Barroso EV, Guerra AJT, Calderano SB. Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar. Rev Bras Cienc Solo. 2014;38:1681-93. https://doi.org/10.1590/S0100-06832014000600003

Campos PM, Lacerda MPC, Silva CL, Sá MAC, Sousa DMG. Drenagem interna como fator de diferenciação de Latossolos do Distrito Federal. Pesq Agropec Bras. 2010;45:306-14. https://doi.org/10.1590/S0100-204X2010000300011

Cardoso CA, Dias HCT, Soares CPB, Martins SV. Caracterização morfométrica da bacia hidrográfica do rio Debossan, Nova Friburgo, RJ. Rev Arvore. 2006;30:241-8. https://doi.org/10.1590/S0100-67622006000200011

Coelho FF, Giasson E. Métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Cienc Rural. 2010;40:2099-106. https://doi.org/10.1590/S0103-84782010005000156

Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ. 1991;37:35-46. https://doi.org/10.1016/0034-4257(91)90048-B

Dancey CP, Reidy J. Estatística sem matemática para psicologia. 3. ed. Porto Alegre: Artmed; 2006.

Dash M, Liu H, Motoda H. Consistency based feature selection. In: Terano T, Liu H, Chen ALP, editors. Knowledge discovery and data mining. current issues and new applications. Berlin: Springer; 2000. p. 98-109.

Demattê JAM, Demétrio VA. Caracterização de solos por padrões de drenagem e sua relação com índices de intemperismo. Pesq Agropec Bras. 1998;33:87-95.

Dias LMS, Coelho RM, Valladares GS, Assis ACC, Ferreira EP, Silva RC. Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco. Pesq Agropec Bras. 2016;51:1396-404. https://doi.org/10.1590/s0100-204x2016000900038

Dobos E, Micheli E, Baumgardner MF, Biehl L, Helt T. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma. 2000;97:367-91. https://doi.org/10.1016/S0016-7061(00)00046-X

Freitas MA, Binotto RB, Nanni AS, Rodrigues ALM, Bortoli CR. Avaliação do potencial hidrogeológico, vulnerabilidade intrínseca e hidroquímica do sistema aquifero Serra Geral no noroeste do estado do Rio Grande do Sul. Revista Brasileira de Recursos Hídricos. 2012;17:31-41. https://doi.org/10.21168/rbrh.v17n2.p31-41

Giasson E, Hartemink AE, Tornquist CG, Teske R, Bagatini T. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil. Cienc Rural. 2013;43:1967-73. https://doi.org/10.1590/S0103-84782013001100008

Giasson E, Sarmento EC, Weber E, Flores CA, Hasenack H. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci Agr. 2011;68:167-74. https://doi.org/10.1590/S0103-90162011000200006

Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-82.

Hall MA, Smith LA. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: Proceedings of the twelfth International Florida Artificial Intelligence Research Society Conference; May 1999; Orlando. Menlo Park: AAAI Press; 1999. p. 235-9.

Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations. 2009;11:10-8. https://doi.org/10.1145/1656274.1656278

Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE T Knowl Data En. 2003;15:1437-47. https://doi.org/10.1109/TKDE.2003.1245283

Hall MA. Correlation-based feature selection for machine learning [thesis]. Hamilton: University of Waikato; 1999.

Höfig P, Giasson E, Vendrame PRS. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. Pesq Agropec Bras. 2014;49:958-66. https://doi.org/10.1590/S0100-204X2014001200006

Kämpf N, Giasson E, Streck EV. Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da microbacia do Lajeado Grande. Porto Alegre: Secretaria do Meio Ambiente do Rio Grande do Sul; 2004.

Liu H, Setiono R. A probabilistic approach to feature selection - a filter solution. In: Proceedings of the thirteenth International Conference on International Conference on Machine Learning; July 1996; Bari. San Francisco: Morgan Kaufmann Publishers Inc.; 1996. p. 319-27.

MacMillan RA. LandMapR© Software Toolkit- C++ Version: users manual. Alberta: LandMapper Environmental Solutions; 2003.

McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117:3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

Paes BC, Plastino A, Freitas AA. Selection of attributes applied to hierarchical classification. In: Proceedings of the first Symposium on Knowledge Discovery, Mining and Learning; July 2013; São Carlos. São Carlos: Universidade de São Paulo; 2013. p. 1-8.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2018 [cited 2018 Jun 23]. Available from: http://www.R-project.org/.

Ruiz LFC, ten Caten A, Dalmolin RSD. Árvore de decisão e a densidade mínima de amostras no mapeamento da cobertura da terra. Cienc Rural. 2014;44:1001-7. https://doi.org/10.1590/S0103-84782014000600008

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. Plos One. 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432

Sarmento EC, Giasson E, Weber E, Flores CA, Hasenack H. Predição de ordens de solos com alta resolução espacial: resposta de diferentes classificadores à densidade de amostragem. Pesq Agropec Bras. 2012;47:1395-403. https://doi.org/10.1590/S0100-204X2012000900025

Schaetzl RJ, Anderson S. Soils: genesis and geomorphology. New York: Cambridge University Press; 2005.

Shi H. Best-first decision tree learning [thesis]. Hamilton: University of Waikato; 2007.

Soil Survey Staff. Keys to soil taxonomy. 12th ed. Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service; 2014.

Subburayalu SK, Jenhani I, Slater BK. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma. 2014;213:334-45. https://doi.org/10.1016/j.geoderma.2013.08.018

Subburayalu SK, Slater BK. Soil series mapping by knowledge discovery from an Ohio County soil map. Soil Sci Soc Am J. 2013;77:1254-68. https://doi.org/10.2136/sssaj2012.0321

Tachikawa T, Kaku M, Iwasaki A, Gesch D, Oimoen M, Zhang Z, Danielson JJ, Krieger T, Curtis B, Haase J, Abrams M, Carabajal RCC. ASTER Global digital elevation model version 2 - summary of validation results [internet]. NASA; 2011. [cited 2018 Jun 13]. Available from: https://ssl.jspacesystems.or.jp/ersdac/GDEM/ver2Validation/Summary_GDEM2_validation_report_final.pdf

Taghizadeh-Mehrjardi R, Toomanian N, Khavaninzadeh AR, Jafari A, Triantafilis J. Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran. Eur J Soil Sci. 2016;67:707-25. https://doi.org/10.1111/ejss.12382

ten Caten A, Dalmolin RSD, Mendonça-Santos ML, Giasson E. Mapeamento digital de classes de solos: características da abordagem brasileira. Cienc Rural. 2012;42:1989-97. https://doi.org/10.1590/S0103-84782012001100013

ten Caten A, Simão R, Dalmolin D, Pedron FA. An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul. Rev Bras Cienc do Solo. 2013;37:359-66. https://doi.org/10.1590/S0100-06832013000200007

Teske R, Giasson E, Bagatini T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos. Rev Bras Cienc do Solo. 2015;39:14-20. https://doi.org/10.1590/01000683rbcs20150344

Vasu NN, Lee S-R. A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea. Geomorphology. 2016;263:50-70. https://doi.org/10.1016/j.geomorph.2016.03.023

Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington: Morgan Kaufmann; 2011.