DOI: http://dx.doi.org/10.1590/1678-992X-2017-0171

Scientia Agricola

ISSN 1678-992X

Preprocessing procedures and supervised classification applied to a database of systematic soil survey

Alan Pessoa Valadares[1], Ricardo Marques Coelho[1]*, Stanley Robson de Medeiros Oliveira[2]

[1]Instituto Agronômico de Campinas/Centro de Solos e Recursos Ambientais, Av. Dr. Theodureto de Almeida Camargo, 1500 – 13075-630 – Campinas, SP – Brasil.
[2]Embrapa Informática Agropecuária, Av. André Tosello, 209 – 13083-886 – Campinas, SP – Brasil.
*Corresponding author <rmcoelho@iac.sp.gov.br>

Edited by: Thomas Kumke

Received May 21, 2017
Accepted May 07, 2018

ABSTRACT:Data Mining techniques play an important role in the prediction of soil spatial distribution in systematic soil surveying, though existing methodologies still lack standardization and a full understanding of their capabilities. The aim of this work was to evaluate the performance of preprocessing procedures and supervised classification approaches for predicting map units from 1:100,000-scale conventional semi-detailed soil surveys. Sheets of the Brazilian National Cartographic System on the 1:50,000 scale, "Dois Córregos" ("Brotas" 1:100,000-scale sheet), "São Pedro" and "Laras" ("Piracicaba" 1:100,000-scale sheet) were used for developing models. Soil map information and predictive environmental covariates for the dataset were obtained from the semi-detailed soil survey of the state of São Paulo, from the Brazilian Institute of Geography and Statistics (IBGE) 1:50,000-scale topographic sheets and from the 1:750,000-scale geological map of the state of São Paulo. The target variable was a soil map unit of four types: local "soil unit" name and soil class at three hierarchical levels of the Brazilian System of Soil Classification (SiBCS). Different data preprocessing treatments and four algorithms all having different approaches were also tested. Results showed that composite soil map units were not adequate for the machine learning process. Class balance did not contribute to improving the performance of classifiers. Accuracy values of 78 % and a Kappa index of 0.67 were obtained after preprocessing procedures with Random Forest, the algorithm that performed best. Information from conventional map units of semi-detailed (4^{th} order) 1:100,000 soil survey generated models with values for accuracy, precision, sensitivity, specificity and Kappa indexes that support their use in programs for systematic soil surveying.

Keywords: machine learning algorithms, random forest, tacit soil-landscape relationships, digital soil mapping

Introduction

One of the challenges of modeling soil classes for digital soil mapping has been to reproduce soil-landscape relationships through tacit information on conventional soil maps (Hudson, 1992). A possible strategy for overcoming this is to make the assumption that conventional soil survey databases implicitly carry information on soil-landscape relationships. Databases derived from soil surveys and those from soil predictive covariates, such as relief and parent material covariates (McBratney et al., 2003), can then be analyzed to produce patterns of soil spatial variation with techniques that belong to the field conceived as Knowledge Discovery in Databases or KDD, of which data preprocessing and data mining are essential steps in the entire process (Fayyad et al., 1996).

Optimal routines for the application of data mining techniques are far from reaching a consensus on digital soil surveys (Bagatini et al., 2016; Behrens and Scholten, 2007), but they can accelerate the generation of information on spatial distribution of soil classes. Classification algorithms such as Artificial Neural Networks (ANN) and Decision Trees (DT) have been widely used for soil survey modeling (Behrens et al., 2005; Silva et al., 2013). ANN simulates biological neural networks; its basic component, a neuron, receives input signals that are aggregated and compared to a threshold or bias of the neuron. If the aggregated signal is greater than the bias, the neuron will be activated and the output signal generated by an activation function (Zhou, 2012). Neurons are linked by weighted connections to form a network. DT uses the divide and conquer process, based on the values of information gained, to create classification rules visually similar to trees (Witten et al., 2016). Algorithms with integrated approaches are also being tested for pedological modeling. Bayesian Neural Networks, which integrate the maximization of probability estimation by Bayes' theorem with ANN (Zhou, 2012) and Random Forest, an Ensemble Method that uses the strategy of Bootstrap Aggregating to create a stronger classifier, based on random DT (Zhou, 2012; Breiman, 2001), are expected to produce very robust models (Hastie et al., 2009; Chagas et al., 2017).

The aim of this research was to evaluate the performance of data preprocessing procedures and supervised classification approaches applied to conventional map units and environmental covariates as reference data sources for predicting soil map units.

Materials and Methods

Studied settings

The research was carried out in the Geographic Information System (GIS) environment with map polygons and legend from three 1:50,000-scale sheets of the 1:100,000-scale soil survey maps of the Brotas and Piracicaba quadrangles, in the state of São Paulo, Brazil (Figure 1).

The studied region has its largest extension located on the Peripheral Depression, but also has part of it on the Basaltic *Cuestas*, both being geomorphological provinces of the state of São Paulo, elevation ranging from 453 to 1069 m. On these landscapes, relief classes range from nearly level to very steep and lithology is mostly of sedimentary rocks but also, in the province of *Cuestas*, of basaltic rocks. Köppen's climate are Cwa and Aw (Alvares et al., 2013).

Databases

A 30-m resolution digital elevation model generated from the Brazilian Institute of Geography and Statistics (IBGE) 1:50,000-scale toposheets provided seven predictive relief attributes: Elevation, Slope Gradient, Relief Class, Profile Curvature, Plane Curvature, Distance to Drainage, and Topographic Wetness Index (TWI). Geological Formation or Lithology, as on the geological map of the state of São Paulo (1:750,000-scale) (Perrota et al., 2005), was the 8th predictive variable.

The target variable was either locally named Soil Units or soil classes in the 2nd (suborder), 3rd (great group) or 4th (subgroup) level of the Brazilian System the Brazilian System of Soil Classification (SiBCS) (Santos et al., 2013) extracted from the Brotas (Almeida et al., 1981) and Piracicaba (Oliveira and Prado, 1989) 1:100,000-sheet soil surveys. Equivalence among map Soil Units, SiBCS subgroups, and U.S. Soil Taxonomy subgroups (Soil Survey Staff, 2014) is shown on Table 1. Mapping concepts used for Soil Units were those from the original 1:100.000-sheet soil surveys (e.g. Oliveira, 1999). Nomenclature of soil classes in the three categorical levels of SiBCS plus Soil Unit names applied to a database of soil map units enabled the structuring of four matrices of predictive attributes plus soil classes.

In order to favor machine learning of soil classes with a low number of instances, minority classes were merged, producing a larger number of instances per class. This was the case for the classes of Hydromorphic soils (Glei), Orthents (Litólicos), certain Alfisols with abrupt textural changes (Diamante), and Spodosols (Podzóis).

Data mining

Preprocessing

Efficiency of the procedures for variable discretization, data selection, under- and oversampling, class balancing, and variable selection (Table 2) was evaluated using the Weka software, version 3.8.0 by "Hold-Out" (Supplied Test Set - 2/3 for training, 1/3 for test).

The predictive variable "relief class" was obtained by discretization of the slope gradient into the following classes: 0-4, 4-8, 8-20, 20-45, 45-75, > 75 %. These variables (slope gradient and relief class) were used simultaneously in both the continuous and discrete forms in data matrices. Discretization of continuous predictive variables in arbitrary classes was also carried out for profile curvature and plane curvature. The discrete classes described above can better represent the hydrodynamic behavior of landscapes than continuous variables. Discretization was tested for the variables TWI and distance to drainage using intervals of equal ranges and equal frequencies. The adopted interval for defining

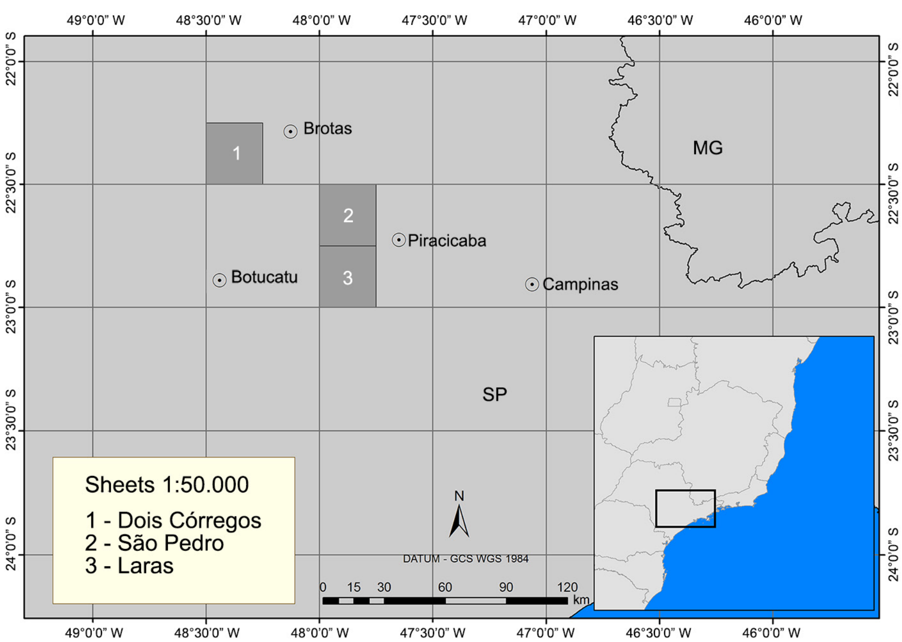

Figure 1 – Sheets on 1:50,000 scale from study area (São Paulo, Brazil).

Table 1 – Equivalence of mapped Soil Units to soil classes of the Brazilian System of Soil Classification (SiBCS) and U.S. Soil Taxonomy.

Soil Units	Soil classes at the 4th level of the SiBCS[a]	U.S. Soil Taxonomy
Alva	PVAd e PVAe abrúptico, A moderado, textura arenosa/média	Sandy over Fine-loamy, Arenic and Typic Paleudult
Areia Quartzosa	RQo típico, A moderado	Typic Quartzipsamment
Baguari	PVAd e PVAe típico e abrúptico, A moderado, textura média e média/argilosa	Fine-loamy, Typic Kandiudult
Barão Geraldo	LVdf típico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Rhodic Hapludox
Campestre	PVe nitossólico e NVe típico, A moderado, textura argilosa/muito argilosa	Fine, Rhodic Kandiudult and Kandiudalf
Canela	PVd e PVAd típico, A moderado, textura média e média/argilosa	Fine-loamy over Fine, Typic Kandiudult
Coqueiro	LVAd psamítico e típico, A moderado e fraco, textura média	Coarse-loamy, Typic Hapludox
Diamante	SXe e SXd típico e vertissólico, A moderado, textura média/argilosa	Fine-loamy over Fine, Vertic, Albaquic and Typic Hapludalf
Engenho	MTf e MTo típico, textura argilosa	Very Fine and Fine, Typic Paleudoll
Estruturada	NVef e NVdf típico, A moderado, textura argilosa e muito argilosa	Very Fine and Fine, Kandiudalfic Eutrudox
Hidromórficos	GXvd, GXve, GXbd e GXbe típico, A moderado e proeminente, textura argilosa	Fine, Aquept, Aquent, Aquox, Aquult, Aqualf
Hortolândia	LVd típico, A moderado, textura média	Fine Loamy, Rhodic Hapludox
Itaguaçu	NVdf latossólico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Kandiudalfic Eutrudox and Rhodic Kandiudox
Laranja Azeda	LVAd típico, A moderado, textura média	Fine-loamy, Typic Hapludox
Limeira	LVd típico, A moderado, textura argilosa e muito argilosa	Very Fine and Fine, Rhodic Hapludox
Litólicos	RLe e RLm típicos, A moderado e chernozêmico, textura média	Loamy, Lithic Udorthent
Monte Cristo	PVAd e PVAe abrúptico e arênico abrúptico, A moderado, textura arenosa/média e média/argilosa	Sandy over Fine-loamy and Sandy over Fine, Arenic Kandiudult and Arenic Kandiudalf
Olaria	NXd típico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Typic and Rhodic Kandiudult
Podzóis	ESKo típico, textura arenosa/média	Sandy over Coarse-loamy, Humod
Ribeirão Preto	LVef típico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Rhodic Eutrudox
Santa Cruz	PVAd e PVAe abrúptico, A moderado, textura média/argilosa media/muito argilosa e argilosa/muito argilosa	Fine-loamy over Fine, Typic Kandiudult and Typic Kandiudalf
Santana	NXe chernossólico, textura média/argilosa	Fine-loamy over Fine, Typic Paleudoll
São Lucas	LAd e LVAd psamítico, A moderado, textura média	Coarse-loamy, Typic Hapludox and Kandiudox
Serrinha	PVAd, PVAe, PAd e PAe arênico abrúptico, A moderado e fraco, textura arenosa/média	Sandy over Fine-loamy, Arenic Paleudult, Grossarenic Paleudult, Arenic Paleudalf and Grossarenic Paleudalf
Sete Lagoas	CYbd e CYbe típico, A moderado e proeminente, textura argilosa e média	Fine and Fine-loamy, Fluventic and Typic Dystrudept
Taquaraxim	CXbd e CXbe típico, A moderado e proeminente, textura média e argilosa	Fine-loamy, Typic Dystrudept
Três Barras	LAd úmbrico, textura media	Coarse- and Fine-loamy, Xanthic and Typic Hapludox

[a]Abreviations as in the Brazilian System of Soil Classification (SiBCS) (Santos et al., 2013).

Table 2 – Summary of preprocessing procedures.

Procedures	Importance /Application
Stratified sampling	Stratified data sampling separating training and testing datasets in the reference area.
Data Selection	Identification and exclusion of inconsistent information.
Discretization	Transformation of continuous quantitative variables into categorical ones.
Undersampling	Resampling by gradual elimination of information from the majority classes in the training of unbalanced classes.
Oversampling	Replication sampling of minority classes in training unbalanced classes.
Class Balancing	Resampling with standardization of the distribution (frequency) of prediction classes.
Selection of Variables	Evaluation of the predictive power of each explanatory variable and elimination of those detrimental to machine learning.

criteria was the one that best improved the performance of the classifier, as evaluated by the "Hold-Out" method.

Resampling procedures were evaluated in order to improve the predictive power of the models in the less representative soil classes in the reference area. Undersampling, oversampling, and class balancing procedures were applied to 2/3 of the map unit database. Resampling was applied at three levels, zero (0.0), representing the original distribution of the data, one (1.0), the balanced distribution of soil classes, and 0.5, a distribution involving undersampling of majority classes and oversampling of minority classes. The test database for these procedures was 1/3 of the instances using the "Hold-Out" method (Supplied Test Set).

Ranking predictive variables by importance was carried out by chi-square (χ^2) and information gain, two commonly used feature-selection methods.

Soil class prediction

Preprocessing procedures were evaluated using four algorithms, Random Forest, J48, MLP and Bayes Net to explore their capabilities to predict soil map units

Table 3 – Algorithms used for supervised classification.

Algorithm (classifier)	Reference	Type of approach
J.48 (C4.5)	Quinlan (1993)	DecisionTree (divide and conquer process based on data information gain)
Random Forest	Breiman (2001)	Ensemble (bootstrap aggregating based on random decision trees)
Multi-Layer Perceptron	Si et al. (2003)	Artificial Neural Networks (transfer functions based on input signal, connections weight and neuron bias)
Bayes Net	Hall et al. (2009)	Bayesian Classifiers (integrates Bayesian probability function to ANN)

(Table 3). Performance of classifiers for each prediction class was evaluated by accuracy, true and false positive rates (TPR and FPR), and by the area under the curve (Bradley, 1997). To evaluate the global performance of classifiers, we used global accuracy (Weiss and Zhang, 2003), weighted mean precision, weighted average of the true positive (weighted mean sensitivity) and the false positive (one minus weighted mean specificity), and the Kappa index (Cohen, 1960).

Results

Preprocessing procedures

Discretization

Profile and plane curvature performed better as discrete variables, whereas elevation, distance to drainage and topographic wetness index (TWI) had better performance as continuous variables. Slope had better performance used in conjunction with its discrete form (relief class). Small differences in performance were considered for pre-selecting the type of predictive attribute. The best adjustment results in these cases are shown in Table 4.

Accuracy values were around 50 % and Kappa indices between 0.35 and 0.45, meaning fair and good agreement (Table 4). The weighted averages of true positive rates (average sensitivity) for the best models were between 42 % and 50 % (Table 4). The specificity of the rules generated by the models was high, indicated by the low mean of false positive rates, with values from 5 % to 9 % (Table 4).

Data selection

Data selection was fundamental to the acceptance of models generated by all evaluated algorithms. Accuracy ranged from 65 % to 78 %, and the Kappa index from 0.50 to 0.67, representing good and very good agreement (Table 5). Composite map units (soil associations and soil complexes) from conventional soil maps showed inconsistency (reduction of predictive performance) and were therefore removed.

Weighted average of true positive rates (TPR) (mean sensitivity) for models generated by conventional map units were between 65 % and 68 % (Table 5). Weighted average of false positive rates (FPR) ranged from 6 to 16 %.

Resampling procedures (subsampling, class balancing and oversampling) did not favor model performance. There was a considerable reduction in global

Table 4 – Best algorithm performances after discretization procedures. Classification in soil unit. Accuracy = Global Accuracy; TPR = weighted average of true positive rates; FPR = weighted average of false positive rates.

Algorithms[a]	Accuracy	Error	Precision	TPR	FPR	Kappa
	%					
J48	50.19	49.81	48.10	50.20	5.50	44.70
MLP	47.87	52.13	44.90	47.90	8.80	39.76
Bayes Net	42.52	57.48	37.30	42.50	6.50	35.78

[a]Algorithm Random Forest could not be used with this dataset due to computational limitations.

Table 5 – Performance of the algorithms after discretization and data selection. Classification in soil units. Accuracy = Global Accuracy; Precision = weighted average precision; TPR = weighted average of true positive rates; FPR = weighted average of false positive rates.

Algorithms	Accuracy	Error	Precision	TPR	FPR	Kappa
	%					
Random Forest	78.13	21.87	77.70	78.10	12.80	67.0
J48	76.09	23.91	75.30	76.10	13.50	64.0
MLP	71.64	28.36	71.00	71.60	15.70	57.0
Bayes Net	65.75	34.25	65.00	65.70	15.30	50.0

accuracy and class precision as soil unit distribution approached the fully balanced distribution (1.0) (Table 6). Databases always produced models with better performance when they were not submitted to resampling procedures.

Results for variable selection showed that all the predictive variables were important for the generated models, with classifier performance reduction as any predictive variable was removed from databases. Evaluation of variables by chi-square (χ^2) and information gain methods showed attributes in the following descending order of predictive power: elevation, geology, distance to drainage, slope gradient, relief class, topographic wetness index (TWI), profile curvature and plane curvature (Table 7).

Algorithms

Global model evaluation used the following metrics: accuracy (overall accuracy), precision, weighted mean of true positive rates, weighted mean of false positive rates and Kappa index (Table 5). The algorithm with the best overall performance was Random Forest, with an Ensemble Method for Prediction approach that generated 20 decision trees for the creation of models.

Table 6 – Performance of the algorithms for holdout accuracy (2/3 training and 1/3 test) after subsampling (0.5), class balancing (1.0) and resampling keeping the original data distribution (0.0).

	Random Forest			J48			Bayes Net		
	1.0	0.5	0.0	1.0	0.5	0.0	1.0	0.5	0.0
	Weighted average precision (%)								
	72.10	75.80	76.50	68.80	71.40	71.70	40.68	61.46	65.00
Soil units	Precision per Soil Unit								
Alva	0.131	0.206	0.394	0.098	0.128	0.372	0.041	0.159	0.756
Areia Quartzosa	0.730	0.803	0.805	0.699	0.758	0.755	0.696	0.738	0.736
Baguari	0.433	0.701	0.754	0.360	0.564	0.617	0.161	0.347	0.427
Barão Geraldo	0.563	0.663	0.717	0.517	0.597	0.631	0.155	0.179	0.195
Campestre	0.282	0.495	0.648	0.211	0.346	0.453	0.068	0.123	0.291
Canela	0.761	0.803	0.782	0.709	0.740	0.731	0.501	0.558	0.589
Coqueiro	0.205	0.455	0.746	0.147	0.260	0.378	0.036	0.061	0.077
Diamante	0.500	0.429	0.000	0.231	0.292	0.000	0.000	0.000	0.000
Engenho	0.071	0.127	0.231	0.042	0.079	0.072	0.041	0.071	0.067
Estruturada	0.441	0.565	0.639	0.327	0.434	0.506	0.082	0.111	0.127
Hidromórficos	0.423	0.587	0.646	0.406	0.523	0.570	0.269	0.330	0.356
Hortolândia	0.666	0.680	0.699	0.577	0.585	0.584	0.289	0.320	0.380
Itaguaçu	0.732	0.688	0.700	0.673	0.660	0.560	0.087	0.313	0.167
Laranja Azeda	0.370	0.584	0.718	0.291	0.457	0.590	0.160	0.326	0.574
Limeira	0.743	0.773	0.747	0.695	0.728	0.718	0.510	0.573	0.575
Litólicos	0.371	0.614	0.653	0.300	0.481	0.525	0.187	0.368	0.467
Monte Cristo	0.720	0.735	0.748	0.632	0.642	0.668	0.238	0.285	0.376
Olaria	0.350	0.553	0.655	0.259	0.398	0.529	0.098	0.194	0.277
Podzóis	0.053	0.121	0.250	0.038	0.045	0.083	0.027	0.056	0.000
Ribeirão Preto	0.371	0.394	0.381	0.241	0.263	0.212	0.096	0.129	0.089
Santa Cruz	0.461	0.680	0.699	0.420	0.581	0.589	0.321	0.463	0.486
Santana	0.676	0.800	0.867	0.430	0.630	0.565	0.108	0.350	0.579
São Lucas	0.144	0.341	0.526	0.113	0.199	0.277	0.049	0.063	0.054
Serrinha	0.862	0.806	0.792	0.840	0.794	0.780	0.762	0.742	0.744
Sete Lagoas	0.410	0.463	0.532	0.436	0.493	0.592	0.236	0.323	0.393
Taquaraxim	0.587	0.666	0.774	0.516	0.585	0.674	0.363	0.418	0.449
Três Barras	0.902	0.893	0.904	0.868	0.860	0.868	0.698	0.688	0.728

Table 7 – Assessment of covariates by their prediction power based on information gain and chi-square (χ^2).

Covariates	Information Gain	χ^2
Elevation	0.885	2,141,866
Geology	0.774	1,258,271
Distance to Drainage	0.219	270,550
Slope Gradient	0.108	168,237
Relief Class	0.074	74,552
TWI	0.054	58,413
Profile Curvature	0.049	45,768
Plane Curvature	0.028	23,355

Algorithm J48 (Decision Tree) presented results close to those of Random Forest, but always inferior, whereas the algorithms MLP (Artificial Neural Networks) and Bayes Net (Bayesian Neural Networks) showed worse performance, despite Kappa indexes of 0.57 and 0.50, respectively (Table 5).

Precision per class was evaluated in order to differentiate the performance of the models of each algorithm per predicted class. Results showed that precision per class followed the results for global model evaluation (Table 8). In general, models with better overall performance showed better accuracy performance per prediction class. Exceptions occurred for Campestre and Baguari soil units, where the accuracy of MLP algorithm was greater than that obtained by J48. For Sete Lagoas, Engenho, Alva and Diamante soil units, the model generated by the J48 algorithm showed better performance than that developed by Random Forest. Precision for Itaguaçu, Santana and Alva units was greater in the model generated by Bayes Net algorithm than by MLP, in which they obtained zero precision. The São Lucas unit had better precision in the model generated by MLP algorithm (Table 8).

Assessment of categorical levels of SiBCS

As for the evaluated categorical levels of SiBCS (Suborder, Great Group and Subgroup), there was little difference in accuracy and Kappa between the categories. A small decrease in performance was associated with an increase in the detail of the categorical level, the algorithm MLP (Artificial Neural Networks) was an exception, and had better performance with classification at the 3rd categorical level of SiBCS (Great Group) (Table 9). Predictions of conventional map units classified by the Brazilian System of Soil Classification (SiBCS) (Table 9) were very similar to classification by soil units (Table 5), even though slightly better.

Results for the weighted average of true positive rates in each class by the algorithm of best performance with classification at the 4th level (Subgroup) of SiBCS was 78 %, indicating good average sensitivity (average chance of the classifier to hit a particular class) (Table 10). The weighted mean false positive rates were 13 %, indicating that the classification rules also showed good average specificity (average chance of the classifier failing in a given class) due to the low number of false positive occurrences (Table 10).

Table 8 – Precision of algorithms evaluated by Hold-Out (2/3 training and 1/3 test) in each soil map unit and classification in Soil Units.

Soil Unit	Random Forest	J48	MLP	Bayes Net
Alva	0.411	0.550	0.000	0.806
Areia Quartzosa	0.826	0.809	0.760	0.736
Baguari	0.778	0.708	0.709	0.441
Barão Geraldo	0.726	0.670	0.510	0.205
Campestre	0.678	0.552	0.644	0.276
Canela	0.800	0.749	0.624	0.591
Coqueiro	0.744	0.612	0.000	0.000
Diamante	0.333	0.600	0.000	0.000
Engenho	0.000	0.083	0.000	0.000
Estruturada	0.687	0.586	0.487	0.120
Hidromórficos	0.656	0.625	0.419	0.356
Hortolândia	0.727	0.645	0.582	0.376
Itaguaçu	0.721	0.596	0.000	0.154
Laranja Azeda	0.728	0.640	0.689	0.578
Limeira	0.758	0.751	0.587	0.586
Litólicos	0.700	0.640	0.559	0.477
Monte Cristo	0.783	0.694	0.621	0.375
Olaria	0.693	0.585	0.425	0.322
Podzóis	0.667	0.200	0.000	0.000
Ribeirão Preto	0.494	0.338	0.000	0.075
Santa Cruz	0.718	0.695	0.643	0.489
Santana	0.947	0.735	0.000	0.481
São Lucas	0.574	0.409	0.804	0.067
Serrinha	0.792	0.781	0.753	0.742
Sete Lagoas	0.566	0.635	0.624	0.412
Taquaraxim	0.790	0.714	0.602	0.453
Três Barras	0.911	0.889	0.781	0.733

Table 9 – Best performance of the algorithms evaluated by Hold-Out (2/3 training and 1/3 test), with classification in the 2nd, 3rd and 4th categorical levels of SiBCS.

Algorithms	SiBCS hierarchical levels			SiBCS hierarchical levels		
	2nd	3rd	4th	2nd	3rd	4th
	Accuracy			Kappa		
	%					
Random Forest	78.69	78.61	78.18	67.90	67.81	67.42
J48	76.77	76.62	76.25	65.00	64.89	64.57
MLP	71.74	72.22	71.45	57.42	57.78	57.42
Bayes Net	66.71	66.37	65.84	51.28	50.86	50.42

The lowest sensitivity values were for "Planossolos" (Alfisols with abrupt textural changes), "Espodossolos" (Spodosols), "Chernossolos" (Mollisols), "Latossolos Amarelos and Vermelho-Amarelos psamíticos" (coarse- and fine-loamy, Xanthic and Typic Oxisols), and "Latossolos Vermelhos Eutroférricos" (Rhodic Eutrudox). Rules created for the remaining classes showed good sensitivity, the best results being obtained for "Argissolos Vermelho-Amarelos" (Arenic and Grossarenic Paleudults and Paleudalfs) and "Latossolos Amarelos Distróficos úmbricos, textura media" (coarse- and fine-loamy Typic Hapludox), with sensitivity values close to 0.9 (Table 10).

The low false positive rates indicate that the rules created by the Random Forest algorithm (committee of 20 decision trees) showed good specificity with the conventional map units and classification at the 4th level (Subgroup) of the SiBCS (Table 10). Values obtained for the area under the curve were quite satisfactory, ranging from 0.7 to 1.0 (Table 10).

Discussion

Preprocessing procedures were extremely important to improving the performance of the models generated by the algorithms. When evaluating data selection, certain information showed inconsistent for machine learning (Han et al., 2011), as they drastically reduced the performance of the evaluated models in supervised classification. This was observed for map units of soil associations and soil complexes. Composite map units (soil associations or complexes) are supposed to carry greater complexity of soil forming factors than those present in soil consociations. Thus, as soil forming factors relief and parent material were used for deriving covariates for soil prediction, this greater complexity could affect the results. Exclusion of composite map units from the training set does not preclude to map areas with features associated with these map units since soil complex and soil associations, composed of two or more soil consociations, are represented by the single unit of the main consociation in the training set. Reducing complexity of predictive covariates has been a successful strategy for improving the prediction of soil map units (Ten Caten et al., 2012).

To deal with the substantial amount of information extracted from the training areas (1,013,329 instances after preprocessing) we used the Hold-Out method, 2/3 for training and 1/3 for model testing, increasing the amount of information for training and testing the models, optimizing the analysis procedure in relation to computational capacity or processing time.

Table 10 – Performance per class of the Random Forest algorithm. Classification in the 4th level (Subgroup) of SiBCS. TP = true positive rate; FP = false positive rate; AUC = area under th e curve.

TP	FP	Precision	AUC	Soil class[a]	U.S. Soil Taxonomy
0.701	0.001	0.811	0.972	CXbd e CXbe típicos, A moderado e proeminente, textura média e argilosa	Fine-loamy, Typic Dystrudept
0.683	0.006	0.578	0.990	CYbd e CYbe típicos, A moderado e proeminente, textura argilosa e média	Fine and Fine-loamy, Fluventic and Typic Dystrudept
0.000	0.000	0.000	0.707	ESKo típico, textura arenosa/média	Sandy over Coarse-loamy, Humod
0.601	0.006	0.663	0.973	GXvd, GXve, GXbd e GXbe típicos, A moderado e proeminente, textura argilosa	Fine, Aquept, Aquent, Aquox, Aquult, Aqualf
0.180	0.002	0.564	0.860	LAd e LVAd psamíticos, A moderado	Coarse-loamy, Typic Hapludox and Kandiudox
0.902	0.001	0.922	0.997	LAd úmbrico, textura média	Coarse- and Fine-loamy, Xanthic and Typic Hapludox
0.300	0.000	0.770	0.903	LVAd psamítico e típico, A moderado e fraco, textura média	Coarse-loamy, Typic Hapludox
0.617	0.001	0.739	0.965	LVAd típico, A moderado, textura média	Fine-loamy, Typic Hapludox
0.850	0.005	0.759	0.994	LVd típico, A moderado, textura argilosa e muito argilosa	Very Fine and Fine, Rhodic Hapludox
0.617	0.002	0.737	0.986	LVd típico, A moderado, textura média	Fine Loamy, Rhodic Hapludox
0.682	0.002	0.738	0.987	LVdf típico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Rhodic Hapludox
0.212	0.000	0.607	0.913	LVef típico, A moderado, textura argilosa ou muito argilosa	Fine and Very Fine, Rhodic Eutrudox
0.010	0.000	0.143	0.743	MTf e MTo típicos, textura argilosa	Very Fine and Fine, Typic Paleudoll
0.611	0.000	0.733	1.000	NVdf latossólico, A moderado, textura argilosa ou muito argilosa	Fine and Very Fine, Kandiudalfic Eutrudox and Rhodic Kandiudox
0.546	0.001	0.655	0.964	NVef e NVdf típicos, A moderado, textura argilosa e muito argilosa	Very Fine and Fine, Kandiudalfic Eutrudox
0.527	0.001	0.669	0.943	NXd típico, A moderado, textura argilosa e muito argilosa	Fine and Very Fine, Typic and Rhodic Kandiudult
0.759	0.000	0.837	0.991	NXe chernossólico, textura média/argilosa	Fine-loamy over Fine, Typic Paleudoll
0.707	0.002	0.763	0.988	PVAd e PVAe abrúpticos e arênico abrúpticos, A moderado, textura arenosa/média e média/argilosa	Sandy over Fine-loamy and Sandy over Fine, Arenic Kandiudult and Arenic Kandiudalf
0.572	0.000	0.490	0.998	PVAd e PVAe abrúpticos, A moderado, textura arenosa/média	Sandy over Fine-loamy, Arenic and Typic Paleudult
0.571	0.021	0.714	0.931	PVAd e PVAe abrúpticos, A moderado, textura média/argilosa, média/muito argilosa e argilosa/muito argilosa	Fine-loamy over Fine, Typic Kandiudult and Typic Kandiudalf
0.604	0.006	0.769	0.943	PVAd e PVAe típico e abrúpticos, A moderado, textura média e média/argilosa	Fine-loamy, Typic Kandiudult
0.902	0.235	0.793	0.907	PVAd, PVAe, PAd e PAe arênicos abrúpticos, A moderado e fraco, textura arenosa/média	Sandy over Fine-loamy, Arenic Paleudult, Grossarenic Paleudult, Arenic Paleudalf and Grossarenic Paleudalf
0.836	0.003	0.788	0.994	PVd e PVAd típicos, A moderado, textura média e média/argilosa	Fine-loamy over Fine, Typic Kandiudult
0.412	0.001	0.677	0.947	PVe nitossólico e NVe típico, A moderado, textura argilosa/muito argilosa	Fine, Rhodic Kandiudult and Kandiudalf
0.518	0.011	0.702	0.916	RLe e RLm típicos, A moderado e chernozemico, textura média	Loamy, Lithic Udorthent
0.757	0.040	0.829	0.956	RQo típico, A moderado	Typic Quartzipsamment
0.000	0.000	0.000	0.964	SXe e SXd típicos e vertissólicos, A moderado, textura média/argilosa	Fine-loamy over Fine, Vertic, Albaquic and Typic Hapludalf
0.782	0.128	0.777	0.929	Weighted average	

[a]Abbreviations as in the Brazilian System of Soil Classification (SiBCS) (Santos et al., 2013).

Results from class balancing followed the pattern found by Crivelenti et al. (2009), with a decrease in performance of classifiers after class balancing (training in equal frequency classes). This indicates that, in this case, undersampling majority classes was detrimental to the machine learning process, probably due to the failure of classifiers to learn important relationships. Therefore, even though an improvement in the prediction of minority classes after balancing was expected, class balancing was detrimental to the overall performance of the models.

Random Forest, a supervised classification method with ensemble approach, produced models with the best performances, similar to the findings of Dias et al. (2016), and Chagas et al. (2017), surpassing J.48, a decision tree algorithm. This opposes the findings of Ten Caten et al. (2013), in terms of accuracy and kappa indexes above 70 % when using decision trees in smaller datasets than those studied here. The use of Bootstrap Aggregating (Bagging) in the Random Forest algorithm shows advantages due to the combination of classifiers (Zhou, 2012).

The high performance of the model generated by the Random Forest algorithm at the 4th level (Subgroup) of SiBCS (accuracy above 78 % and kappa index above 67 %) indicates that the approach has great potential for producing digital pedological maps compatible to medium and high intensity reconnaissance (4th order) soil surveys.

Some of the minority classes were better predicted by models with lower global performance. However, in all cases the information gain with these individual classes was not significant enough to improve the overall performance of the models. This fact may be due, in the main, to the prevalence of certain classes in the training area, which resulted in assigning great weight to small decreases in the performance of these majority classes. The high accuracy level in most of the predicted classes is indicative of the high predictive power of the models tested.

Conclusions

Composite soil map units (soil complex and soil associations) proved to be inadequate for the machine learning process, since their exclusion from the training dataset improved overall prediction.

When modeling soil map units for pedological mapping, training on unbalanced databases outperformed training on balanced databases, showing no need for class balancing for machine learning on the studied dataset.

The Random Forest algorithm had good performance for soil class prediction and, though requiring preprocessing procedures, outscored algorithms of different approaches, such as single Decision Trees, Artificial Neural Networks and Bayesian Classification.

The applied predictive variables (terrain attributes and geology) trained on 1:100,000 soil survey maps showed excellent performance for predicting soil map unit distribution, and can be used to create digital pedological maps consistent with high intensity reconnaissance soil survey (4th order) maps.

Authors' Contributions

Conceptualization: Coelho, R.M.; Valadares, A.P. Data acquisition: Valadares, A.P.; Coelho, R.M.; Oliveira, S.R.M. Data analysis: Valadares, A.P.; Oliveira, S.R.M. Design of methodology: Coelho, R.M.; Valadares, A.P.; Oliveira, S.R.M. Software development: Valadares, A.P.; Oliveira, S.R.M. Writing and Editing: Valadares, A.P.; Coelho, R.M.; Oliveira, S.R.M.

References

Almeida, C.L.F.; Oliveira, J.B.; Prado, H. 1981. Semi detailed Soil Survey of the State of São Paulo: Brotas Sheet (SF-22-Z-B-III) Map, 1:100,000 scale. = Levantamento Pedológico Semidetalhado do Estado de São Paulo: Quadrícula de Brotas (SF-22-Z-B-III). Instituto Agronômico, Campinas, SP, Brazil (in Portuguese).

Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Gonçalves, J.L.M.; Sparovek, G. 2013. Koppen's climate classification map for Brazil. Meteorologische Zeitschrift 22: 711-728. DOI 10.1127/0941-2948/2013/0507.

Bagatini, T.; Giasson, E.; Teske, R. 2016. Expansion of pedological maps for physiographically similar areas by digital soil mapping. Pesquisa Agropecuária Brasileira 51: 1317-1325 (in Portuguese, with abstract in English).

Behrens, T.; Förster, H.; Scholten, T.; Steinrücken, U.; Spies, E.D.; Goldschmitt, M. 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168: 21-33.

Behrens, T.; Scholten, T. 2007. A comparison of data mining techniques in predictive soil mapping. p. 353-365. In: Lagacherie, P.; McBratney, A.; Voltz, M., eds. Digital soil mapping: an introductory perspective. Elsevier, Amsterdam, The Netherlands. (Developments in Soil Science, 31).

Breiman, L. 2001. Random forests. Journal of Machine Learning Research 45: 5-32.

Chagas, C.S.; Pinheiro, H.S.K.; Junior, W.C.; Anjos, L.H.C.; Pereira, N.R.; Bhering, S.B. 2017. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. Geoderma Regional 9: 47-55.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37-46.

Crivelenti, R.C.; Coelho, R.M.; Adami, S.F.; Oliveira, S.R.M. 2009. Data mining for soil-landscape relationships inference in digital soil mapping. Pesquisa Agropecuária Brasileira 44: 1707-1715 (in Portuguese, with Abstract in English).

Dias, L.M.S.; Coelho, R.M.; Valladares, G.S.; Assis, A.C.C.; Ferreira, E.P.; Silva, R.C. 2016. Soil class prediction by data mining in an area of the São Francisco sedimentary basin. Pesquisa Agropecuária Brasileira 51: 1396-1404 (in Portuguese, with Abstract in English).

Fayyad, U.M.; Shapiro, G.P.; Smyth, P. 1996. From data mining to knowledge discovery: an overview. p. 1-34. In: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., eds. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Han, J.; Kamber, M.; Pei, J. 2011. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA, USA.

Hastie, T.; Tibshirani, R.; Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Stanford, CA, USA. (Springer Series in Statistics).

Hudson, B.D. 1992. The soil survey as paradigm-based science. Soil Science Society of America Journal 56: 836-841.

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

Oliveira, J.B. 1999. Soils of the Piracicaba Sheet = Solos da Folha de Piracicaba. Instituto Agronômico, Campinas, SP, Brazil (Boletim Científico, 48) (in Portuguese).

Oliveira, J.B.; Prado, H. 1989. Semi detailed Soil Survey of the State of São Paulo: Piracicaba sheet (SF-23-Y-A-IV) Map, 1:100,000-scale. = Carta Pedológica Semidetalhada do Estado de São Paulo: Piracicaba (SF-23-Y-A-IV). Instituto Agronômico, Campinas, SP, Brazil (in Portuguese).

Perrota, M.M.; Salvador, E.D.; Sachs, L.L.B. 2005. Geological Map of the State of São Paulo. 1:750,000-scale. = Mapa Geológico do Estado de São Paulo. CPRM-Serviço Geológico do Brasil, Brasília, DF, Brazil (in Portuguese).

Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, USA.

Santos, H.G.; Jacomine, P.K.T.; Anjos, L.H.C.; Oliveira, V.A.; Lumbreras, J.F.; Coelho, M.R.; Almeida, J.A.; Cunha, T.J.F.; Oliveira, J.B. 2013. Brazilian System of Soil Classification. = Sistema Brasileiro de Classificação de Solos. 3ed. Embrapa, Brasília, DF, Brazil (in Portuguese).

Si, J.; Nelson, B.J.; Runger, G.C. 2003. Artificial neural network 410 models for data mining. p. 41-66. In: Ye, N., ed. The handbook of data mining. Lawrence Erlbaum, Mahwah, NJ, USA.

Silva, C.C.; Coelho, R.M.; Oliveira, S.R.M.; Adami, S.F. 2013. Digital soil mapping of the Botucatu sheet (SF-22-Z-B-VI-3): data training on conventional maps and field validation. Brazilian Journal of Soil Science 37: 846-857 (in Portuguese, with abstract in English).

Soil Survey Staff. 2014. Keys to Soil Taxonomy. 12ed. USDA-Natural Resources Conservation Service, Washington, DC, USA.

Ten Caten, A.; Dalmolin, R.S.D.; Ruiz, L.F.C. 2012. Digital soil mapping: strategy for data pre-processing. Revista Brasileira de Ciência do Solo 36: 1083-1091.

Ten Caten, A.; Dalmolin, R.S.D.; Pedron, F.A.; Ruiz, L.F.C.; Silva, C.A. 2013. An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul, Brazil. Revista Brasileira de Ciência do Solo 37: 359-366.

Weiss, S.M.; Zhang, T. 2003. Performance analysis and evaluation. p. 425-440. In: Nong, Y., ed. The handbook of data mining. Lawrence Erlbaum, Mahwah, NJ, USA.

Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA, USA.

Zhou, Z.H. 2012. Ensemble Methods: Foundations and Algorithms. Chapman & Hall, Cambridge, UK.