Geostatistics or machine learning for mapping soil attributes and agricultural practices

Wanderson de Sousa Mendes[1], *José Alexandre Melo Demattê*[2*], *Arnaldo Sousa e Barros*[1], *Diego Fernando Urbina Salazar*[1], *Merilyn Taynara Accorsi Amorim*[2]

10.1590/0034-737X202067040010

ABSTRACT

Applying the upcoming technologies in agriculture has been a major economic, environmental and social challenge for scientists and farmers. In order to overcome such challenge, this study evaluated the advantages and limitations of using geostatistics and machine learning for soil mapping in agricultural practices and soil surveys. The study occurred in Tocantins State, Brazil, and consisted into seven areas with a total extension of 17.24 km^2, 222 meters regular gridded resulting in one-point sampling per 0.0493 km^2 of five randomly sampled cores within a 1 m circle radius. It was collected 332 georeferenced soil samples at 0-20 cm depth using an auger and then, soil laboratory analyses performed. Afterward, liming rate maps were originated from the predicted soil attributes clay, cation exchange capacity and base saturation comparing four methods: ordinary kriging, random forest, cubist, support vector machine and the best model results of each soil attribute. Evaluating the methods, the Pearson's index presented strong results for soil attributes predicted by random forest and ordinary kriging. Machine learning methods can be successfully applied for soil mapping in agricultural practices and soil surveys using less soil samples rather than geostatistical framework.

Keyword: machine learning, digital soil mapping, kriging, remote sensing, decision tree

INTRODUCTION

Applying the upcoming technologies in agriculture has been an enjoyable challenge for scientists and agricultural entrepreneurs. However, the challenge is to reach economic, environmental and social sustainability by using these technologies in the field of agricultural practices. Techniques such as geostatistics and machine learning have been performed in land management (Rodrigo-Comino *et al*., 2018), crop yield prediction (Adamchuk *et al.*, 2017), soil type mapping (Demattê *et al*., 2015) and zone management (Castro-Franco *et al.*, 2018). The geostatistics began in the field of geology by Krige (1951). Since then, these application of random theory functions are extremely used for spatialisation of georeferenced data and as it has been performed for agricultural purposes (Dowd, 1991; Shannon *et al.*, 2018; Wackernagel, 2014). The machine learning approach estimates soil spatial arrangements using ancillary variables such as digital elevation models (DEM) and its covariates, and remote sensing data such as satellite images from Landsat 5 Thematic Mapper (McBratney *et al.*, 2003; Fongaro *et al.*, 2018; Gallo *et al.*, 2018; Castro-Franco *et al.*, 2018).

The foremost difference between the geostatistical and machine learning methods is how each one deals with spatial statistics and sampling data. The sampling density for geostatistics has to be spatially dependent, otherwise it is classical statistics and no prediction can be performed by using this method. Meanwhile, machine learning relies on georeferenced sampling data only. Both methods depend on georeferenced dataset; however, there is no need of spatial dependency on machine learning. Ordinary kriging is a stationary random function of geostatistics, which means to calculate an average of the radium of each point (Wackernagel, 2014). Other three random spatially functions are Random Forest (Flaxman *et al.*, 2011) and

Submitted on June 19th, 2019 and accepted on June 28th, 2020.

[1] Escola Superior de Agricultura "Luiz de Queiroz", Programa de Pós-Graduação em Solos e Nutrição de Plantas, Piracicaba, São Paulo, Brazil.
[2] Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciência do Solo, Piracicaba, São Paulo, Brazil.
* Corresponding author: jamdemat@usp.br

Cubist (Quinlan & Ross, 1993), which are a decision tree normally used in regressions and classifications. The third one is the Support Vector Machine that uses a kernel function to generalise non-linear models (Vapnik, 2000). The three methods are classified as machine learning.

Each method has advantages and limitations when applied to agriculture. The main farmers' complain about geostatistics are the sampling density because the ideal grid for predicting chemical and physical soil attributes is less than one point per hectare (Cherubin *et al.*, 2015; Nanni *et al.*, 2011) and one point per 7.2 hectares, respectively (Nanni *et al.*, 2011). Thus, the geostatistical method can be non-economic sustainable. In order to solve this issue, this study evaluates the advantages and limitations of using geostatistics or machine learning for soil mapping in agricultural practices and soil surveys. We create a liming rate map originated from the predicted soil attributes cation exchange capacity and base saturation comparing four methods: ordinary kriging, random forest, cubist, support vector machine and the best model results of each soil attribute.

MATERIAL AND METHODS

Study area and sampling data

The study area is located in the municipality of Barra do Ouro, State of Tocantins, Brazil (Figure 1). The site was subdivided into seven areas with total extension of 17.24 km^2 (1,724 hectares), 222 meters regular gridded (222 x 222 m) resulting in one-point sampling per 0.0493 km^2 (4.93 ha) of five randomly sampled cores within a 1 m circle radius. It was collected 332 georeferenced soil samples at 0-20 cm depth using an auger. The soil samples were dried (45 ºC for 24 h), grounded and sieved (2-mm mesh). Afterward, the soil chemical and physical analyses were performed as described in Donagemma *et al.* (2011) and Camargo *et al*. (2009).

Prediction of soil attributes

The ordinary kriging (OrdKrig), support vector machine (SVM), cubist (Cub) and random forest (RF) methods were applied to predict the soil attributes: clay (g kg^{-1}), cation exchange capacity (CEC, cmol$_c$ dm^{-3}) and base saturation (V%). We only predicted these three attributes because they are required in the liming rate calculation. The ancillary variables used in the machine learning process were (i) the Digital Elevation Model (DEM) retrieved from the Earth Resource Observation and Science Center that distributes the Shuttle Radar Topography Mission data (SRTM v.3, 30 m), and (ii) the Synthetic Soil Image (SYSI) generated from 27-year time-series (1985-2011) of Landsat 5 TM by the methodology described in Demattê *et al.* (2018). The SYSI represents bare soil pixels along time. The characterisation of the parameters used in the study area

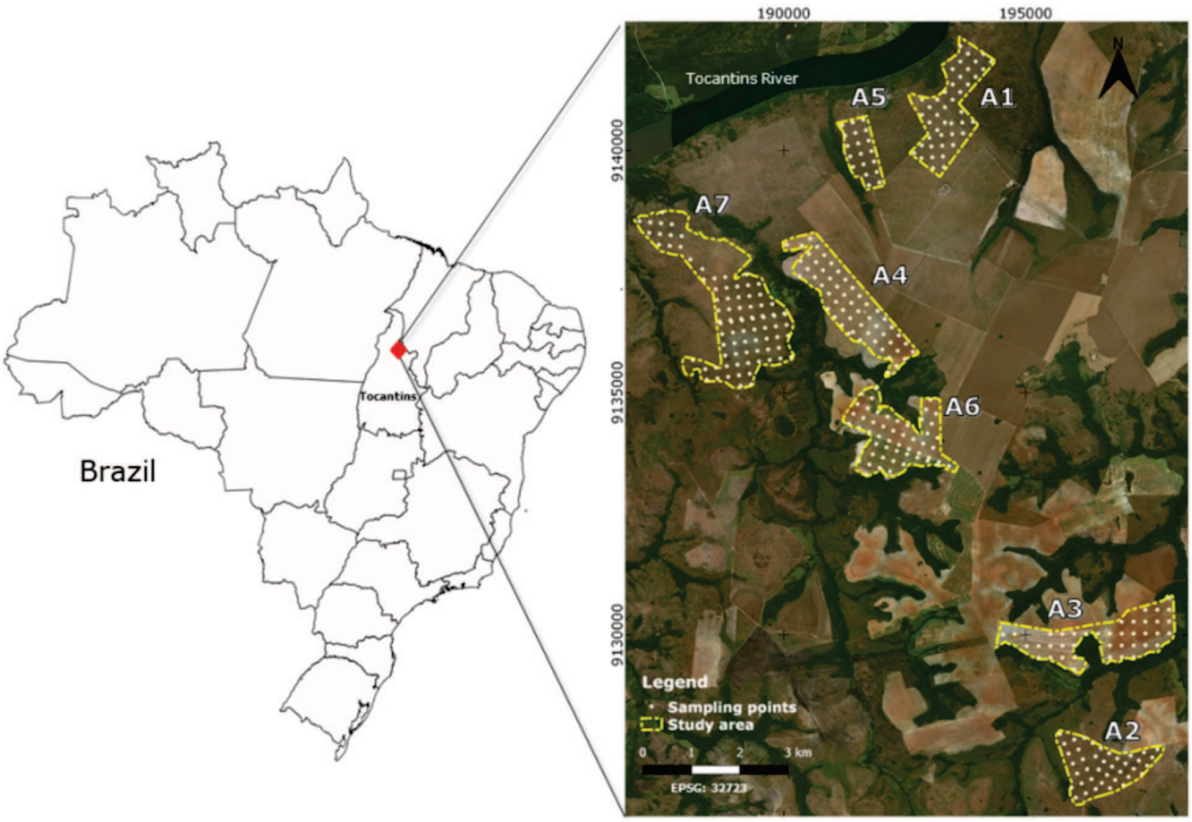

Figure 1: Study area retrieved using the Bing aerial map as background.

Table 1: Descriptive statistics of the parameters used to map the soil attributes

Parameters	Description	Unit	Min	Median	Max	[1]SD
Clay	-	$g\,kg^{-1}$	9.00	130.00	340.00	84.89
V	Base saturation	%	10.13	61.08	86.48	16.04
CEC	Cation Exchange Capacity	$cmol_c\,dm^{-3}$	1.57	3.56	6.90	1.02
[2]SYSI_1	Band n. 1 from the SYSI of the Landsat 5 TM (0.45-0.52 µm)	Reflectance factor	0.03	0.06	0.09	0.01
[2]SYSI_2	Band n. 2 from the SYSI of the Landsat 5 TM (0.52-0.60 µm)	Reflectance factor	0.05	0.10	0.15	0.02
[2]SYSI_3	Band n. 3 from the SYSI of the Landsat 5 TM (0.63-0.69 µm)	Reflectance factor	0.06	0.13	0.21	0.02
[2]SYSI_4	Band n. 4 from the SYSI of the Landsat 5 TM (0.76-0.90 µm)	Reflectance factor	0.10	0.20	0.29	0.03
[2]SYSI_5	Band n. 5 from the SYSI of the Landsat 5 TM (1.55-1.75 µm)	Reflectance factor	0.15	0.28	0.40	0.04
[2]SYSI_7	Band n. 7 from the SYSI of the Landsat 5 TM (2.08-2.35 µm)	Reflectance factor	0.13	0.24	0.36	0.04
DEM	Digital Elevation Model from the SRTM	Meters	188.00	210.00	239.00	10.28

[1]SD - standard deviation; [2]SYSI - Synthetic Soil Image.

is summarized in the Table 1. The methods were performed using the software R (R Development Core Team, 2019) by specific packages as it follows: OrdKrig from "automap" package (Hiemstra *et al.*, 2009), SVM from "e1071" package (Dimitriadou *et al.*, 2011), Cub and RF from "caret" package (Kuhn, 2008). The fitted semivariogram was generated for the three soil properties before of interpolating them using OrdKrig. For machine learning methods (SVM, Cub and RF), the data were initially analysed and no value fields excluded. Then, the 307 sampling points left were randomly divided into 80% for training (247 samples) and 20% for validation purposes (60 samples). Finally, the 80% training data were cross-validated (3 fold, repeated 3 times) for each model (Heung *et al.*, 2014).

Model evaluation

The ordinary kriging method was evaluated by its standard deviation and variance results (Mueller *et al.*, 2004). The machine learning algorithms were evaluated analysing the predicted and observed values of each soil attribute by accessing the Root Mean Squared Error (RMSE), the Pearson correlation coefficient (r) and the Index Of Agreement (IOA) (Willmott *et al.*, 2012).

Liming rate calculation

The liming rate calculation is based on the following classical formula: $LR = [(V\%_2 - V\%_1) * CEC]/TNP$ (Molin *et al.*, 2015; Raij, 1983). LR is the Liming Requirement in ton per hectare; $V\%_2$ is the base saturation that is considered between 60-80% for most crops; $V\%_1$ is the base saturation obtained in laboratory and predicted map; CEC is the cation exchange capacity in cmol dm^{-3} from laboratory and predicted map; and TNP is the total neutralizing power in %. For the LR calculation, we adopted base saturation of 80% for crop and TNP of 100%. We used the raster calculation of QGIS (QGIS Development Team, 2018) to achieve the LR. Basically, this tool retrieved from the V% and CEC maps the value of each pixel at the same location and then, the formula of liming rate was calculated generating the final map with a pixel resolution of 30 m.

RESULTS AND DISCUSSION

The Pearson's correlation coefficient (r) of the dataset presented moderate linearity between the soil attributes and the satellite bands of the SYSI, which represents the bare soil pixels (Figure 2). However, the DEM had low *r* values for all attributes (Figure 1). The linearity among them is a fundamental principle to identify the normal distribution needed before of mapping the soil attributes, and the evaluation of the kriging process in geostatistics is analysing its variance and standard variation.

The variance and standard deviation increases as the spatial dependency decreases (Figure 3). Thus, whether the distance among the collected samples are higher than one sample per hectare, the soil chemical attributes have no fair prediction. For example, the recommended grid size for phosphorus and potassium prediction in Oxisol is less or equal to 1 sample per 100x100 m (Cherubin *et al.*, 2015). Soil grids larger than 100x100 m display high variance and to some degree no spatial dependency, which reallocate them in the conventional statistics. For physical attributes,

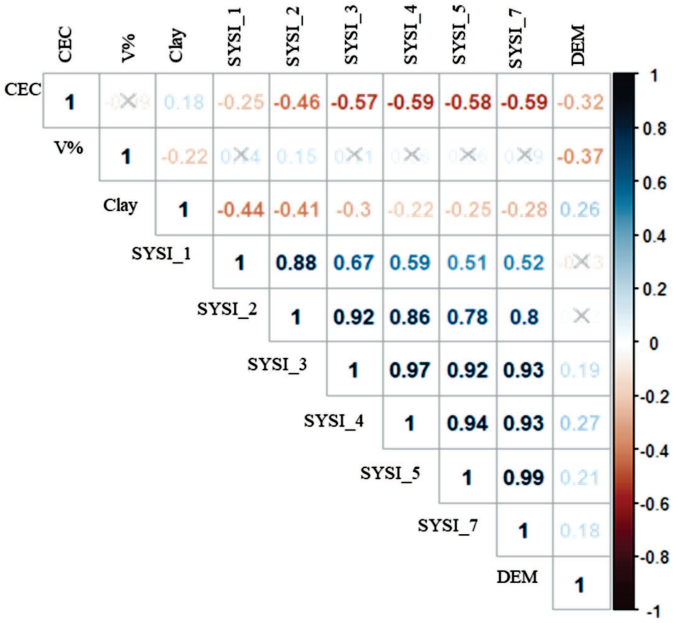

Figure 2: Pearson's correlation coefficient ($p < 0.01$) for all attributes used to map the area. SYSI is the Synthetic Soil Image and the numbers are the satellite's bands; DEM is the digital elevation model; V% is the base saturation; and CEC is the cation exchange capacity. Checkbox values no significant.

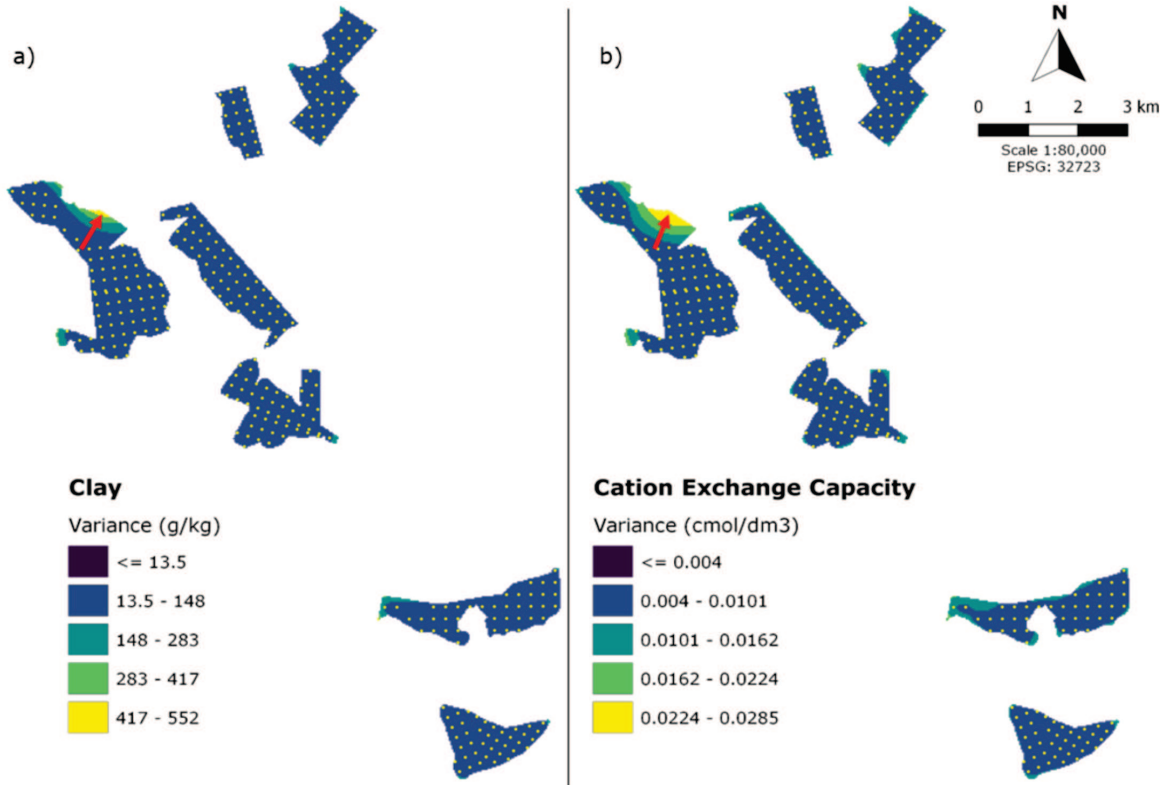

Figure 3: Maps of variance resulted from the ordinary kriging for clay (a) and cation exchange capacity (b). Red arrow indicates increasing variance.

it is acceptable grids up to one sample per 7.2 hectares (Nanni *et al.*, 2011). These authors found moderate correlation for clay prediction ($R^2 \geq 0.53$).

The machine learning methods are evaluated by other meanings, which are the RMSE, r and IOA. The Index Of Agreement (IOA) and r values close to one means that the model prediction fitted well (Table 2). Taking this into account, the RF was the best algorithm to predict clay (IOA = 0.689 and r = 0.83) (Beguin *et al.*, 2017; Fongaro *et al.*, 2018; Vasava *et al.*, 2019) and had close correlation to SVM predicting CEC. For mapping base saturation, the cubist model had better performance rather than the other two methods (Nussbaum *et al.*, 2017). We basically performed the three methods in order to predict the soil attributes. Subsequently, we calculate the liming rate based on the results from RF, Cub, SVM, and the best predicted attribute independent of the method. This last one named Best Model (BM). The ordinary kriging was kept a part from the machine learning methods because our intention is to prove that machine learning methods can be a reasonable framework economically, socially and environmentally sustainable for agriculture. Furthermore, it was calculated the liming rate for an extent of the study area showing the practicability of machine learning

Table 2: Evaluation of the machine learning methods performed for three soil attributes

Soil attributes	Models		
	RMSE	**r**	**[1]IOA**
	Random Forest		
Clay	73.66	0.52	0.59
CEC	1.33	0.69	0.34
V%	9.57	0.83	0.68
	Cubist		
Clay	82.15	0.39	0.55
CEC	1.31	0.64	0.37
V%	8.56	0.86	0.72
	Support Vector Machine		
Clay	76.35	0.49	0.57
CEC	1.27	0.68	0.38
V%	10.54	0.76	0.68

[1]IOA – Index of Agreement (Willmott *et al.*, 2012).

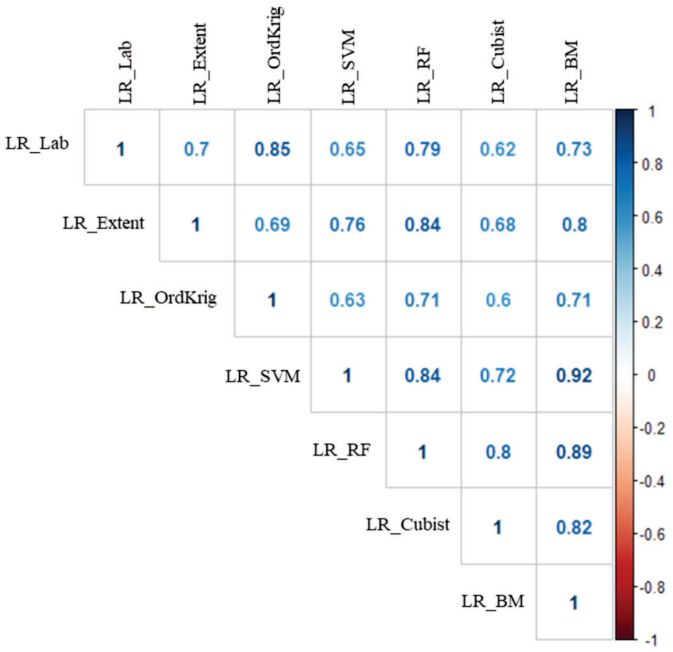

Figure 4: Pearson's correlation coefficient ($p < 0.01$) for liming rate predicted by Ordinary Kriging (LR_OrdKrig), Support Vector Machine (LR_SVM), Random Forest (LR_RF), the best model (LR_BM), the extent of the study area (LR_Extent), and cubist (LR_Cubist) methods compared with the calculation from soil laboratory analyses (LR_Lab).

methods because they can spatialize attributes or response variables without needs of grid size dependency. Checking the reliability of the calculation of all methods, the Pearson's index presented strong results for soil attributes predicted by RF and OrdKrig. These methods had better performance among others (Figure 4). The advantage of using machine learning methods to achieve agricultural field application relies on the absence of spatial dependency among soil sample observations, which is a requirement in geostatistical framework. Another advantage, it is to map large areas (Figure 5) with close performance to geostatistical approaches (r = 0.7).

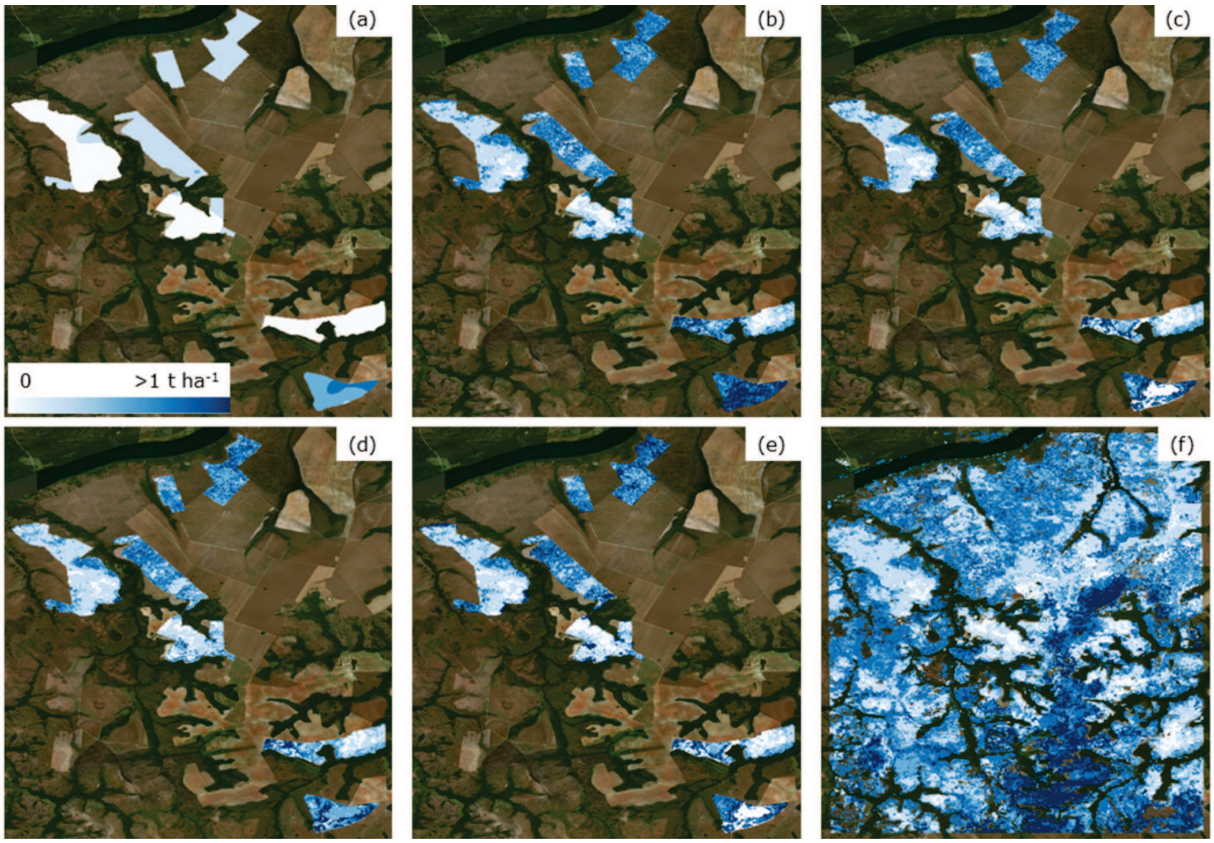

Figure 5: Liming rate maps generated using soil attributes predicted from Ordinary Kriging (a), Random Forest (b), Best Model (c), Support Vector Machine (d), Cubist (e), and Extent area (f). The Bing aerial map is used as background.

CONCLUSIONS

Machine learning methods can be successfully applied for soil mapping in agricultural practices and soil surveys using less samples rather than the geostatistical approaches. This conclusion is a key point to farmers that want to apply optimized methods in their agricultural day life. The machine learning frameworks, mainly Random Forest, proved to be economically and environmentally sustainable for agriculture.

ACKNOWLEDGEMENTS, FINANCIAL SUPPORT AND FULL DISCLOSURE

We acknowledge the financial support of the São Paulo Research Foundation (FAPESP grant numbers 2014/22262-0 and 2016/26124-6). We also thank members of the Geotechnologies in Soil Science Group (GeoCIS/GeoSS) (http:// esalqgeocis.wixsite.com/geocis).

REFERENCES

Adamchuk V, Lacroix R, Shinde S, Tremblay N & Huang H (2017) An uncertainty-based comprehensive decision support system for site-specific crop management. Advances in Animal Biosciences, 8:625-629.

Beguin J, Fuglstad G, Mansuy N & Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma, 306:195-205.

Camargo AO, Moniz AC, Jorge JA & Valadares JMAS (2009) Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agronômico de Campinas. Campinas, IAC. 77p.

Castro-Franco M, Córdoba MA, Balzarini MG & Costa JL (2018) A pedometric technique to delimitate soil-specific zones at field scale. Geoderma, 322:101-111.

Cherubin MR, Santi AL, Eitelwein MT, Amado TJC, Simon DH & Damian JM (2015) Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. Pesquisa Agropecuária Brasileira, 50:168-177.

Demattê J, Rizzo R & Botteon VW (2015) Pedological mapping through integration of digital terrain models spectral sensing and photopedology. Revista Ciência Agronômica, 46:669-678.

Demattê J, Fongaro CT, Rizzo R & Safanelli JL (2018) Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sensing of Environment, 212:161-175.

Dimitriadou E, Hornik K, Leisch F, Meyer D & Maintainer AW (2011) Package "e1071". Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.191.317&rep=rep1&type=pdf. Accessed on: June 1st, 2019.

Donagemma G, Campos DVB, Calderano SB, Teixeira WG & Viana JHM (2011) Manual de Métodos de Análise de Solo. 2ª ed. Rio de Janeiro, Embrapa Solos. 230p.

Dowd P (1991) A review of recent developments in geostatistics. Computers & Geosciences, 17:1481-1500.

Flaxman AD, Vahdatpour A, Green S, James SL & Murray CJ (2011) Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Population Health Metrics, 9:01-29.

Fongaro C, Demattê J, Rizzo R, Safanelli J, Mendes W, Dotto A & Ustin S (2018) Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images. Remote Sensing, 10:01-21.

Gallo B, Demattê J, Rizzo R, Safanelli J, Mendes W, Lepsch I & Lacerda MPC (2018) Multi-Temporal Satellite Images on Topsoil Attribute Quantification and the Relationship with Soil Classes and Geology. Remote Sensing, 10:1571.

Heung B, Bulmer CE & Schmidt MG (2014) Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma, 214-215:141-154.

Hiemstra PH, Pebesma EJ, Twenhöfel CJW & Heuvelink GBM (2009) Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers & Geosciences, 35:1711-1721.

Krige DG (1951) A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical Metallurgical and Mining Society of South Africa, 52:119-139.

Kuhn M (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28:01-26.

McBratney AB, Mendonça Santos ML & Minasny B (2003) On digital soil mapping.Geoderma, 117:03-52.

Molin JP, Amaral LR & Colaço AF (2015) Agricultura de precisão. São Paulo, Oficina de Textos. 224p.

Mueller TG, Pusuluri NB, Mathias K, Cornelius PL, Barnhisel RI & Shearer SA (2004) Map Quality for Ordinary Kriging and Inverse Distance Weighted Interpolation. Soil Science Society of America Journal, 68:2042.

Nanni MR, Povh FP, Demattê J, Oliveira RB, Chicati ML & Cezar E (2011) Optimum size in grid soil sampling for variable rate application in site-specific management. Scientia Agricola, 68:386-392.

Nussbaum M, Walthert L, Fraefel M, Greiner L & Papritz A (2017) Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. Soil, 3:191-210.

QGIS Development Team (2018) QGIS geographic information system. Open source geospatial foundation project. Available at: http://www.qgis.org/. Accessed on: November 20th, 2018.

Quinlan JR, John R & Ross J (1993) C4.5: programs for machine learning. San Francisco, Morgan Kaufmann Publishers. 302p.

R Development Core Team (2019) Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available at: https://www.r-project.org/. Accessed on: June 1st, 2019.

Raij B (1983) Avaliação da fertilidade do solo. 2ª ed. Piracicaba, Instituto da Potassa & Fosfato. 142p.

Rodrigo-Comino J, Martinéz-Hernández C, Iserloh T & Cerdà A (2018) Contrasted Impact of Land Abandonment on Soil Erosion in Mediterranean Agriculture Fields. Pedosphere, 28:617-631.

Shannon DK, Clay DE & Sudduth KA (2018) An Introduction to Precision Agriculture. In: Shannon DK, Clay DE & Kitchen NR (Eds.) Precision Agriculture Basics. Madison, ASA / CSSA / SSSA. p.01-12.

Vapnik VN (2000) The Nature of Statistical Learning Theory. New York, Springer New York. 201p.

Vasava HB, Gupta A, Arora R & Das BS (2019) Assessment of soil texture from spectral reflectance data of bulk soil samples and their dry-sieved aggregate size fractions. Geoderma, 337:914-926.

Wackernagel H (2014) Geostatistics. In: Wiley StatsRef: Statistics Reference Online. Chichester, John Wiley & Sons. p.01-10.

Willmott CJ, Robeson SM & Matsuura K (2012) A refined index of model performance. International Journal of Climatology, 32:2088-2094.