Soil Advances 3 (2025) 100024

Contents lists available at ScienceDirect

Soil Advances

journal homepage: www.sciencedirect.com/journal/soil-advances

# Integrating proximal geophysical sensing and machine learning for digital soil mapping: Spatial prediction and model evaluation using a small dataset

Danilo César de Mello [a], Gustavo Vieira Veloso [a], Murilo Ferre de Mello [a], Marcos Guedes de Lana [b], Isabelle de Angeli Oliveira [a], Fellipe Alcantara de Oliveira Mello [b], Rafael Gomes Siqueira [a], Lucas Carvalho Gomes [a], Elpídio Inácio Fernandes-Filho [a], Carlos Ernesto Gonçalves Reynaud Schaefer [a], Márcio Rocha Francelino [a], Emilson Pereira Leite [c], Tiago Osório Ferreira [b], José Alexandre Melo Demattê [b,1,*]

[a] Department of Soil Science, Federal University of Viçosa, Av. Peter Henry Rolfs s/n – Campus Universitário, Viçosa, MG CEP: 36570-900, Brazil
[b] Department of Soil Science, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, CP 9, Piracicaba, SP 13418-900, Brazil
[c] Department of Geology and Natural Resources, Institute of Geosciences, University of Campinas, Rua Carlos Gomes, 250, Cidade Universitária, Campinas, SP CEP 13083-855, Brazil

ARTICLE INFO

*Keywords:*
Digital soil mapping
Gamma ray spectrometry
Machine learning algorithms
Magnetic susceptibility
Soil electrical conductivity

ABSTRACT

Geophysical methods support soil security by providing non-invasive tools to assess soil properties, monitor degradation, and guide sustainable management strategies. However, studies focusing the spatial prediction of geophysical data remain limited. In this research, we aimed to model and predict the spatial distribution of soil geophysical properties using parent material and terrain attributes with machine learning algorithms. In addition, we tested the nested leave-one-out cross validation (nested-LOOCV) method to deal with datasets with limited size. We performed a geophysical survey using three types of sensors (radiometric, magnetic and electric methods). The random forest (RF) and support vector machine (SVM) algorithms presented the best results, with RF showing higher performance for $K^{40}$ and magnetic susceptibility, and SVM had higher performance for eU, eTh and apparent electrical conductivity. Parent materials and digital elevation model were the most significant variables for the modelling. The nested-LOOCV method proved to be adequate for small soil dataset. Machine learning techniques are potential tools for modelling soil geophysical variables. The combination with computational techniques shows the great relevance of geophysical measurements for the estimation of soil properties related to fertility and soil genesis.

## 1. Introduction

Geophysical methods contribute to soil security by providing non-invasive and efficient tools for assessing and monitoring soil properties at various scales. Techniques like electromagnetic induction, gamma ray spectrometry and magnetic susceptibility can measure and map soil physical and chemical properties, including moisture content, salinity, and organic carbon. These methods enable the identification of soil degradation, compaction, erosion risks, and nutrient distribution, supporting informed soil management and conservation strategies. By offering a deeper understanding of soil variability, geophysical approaches enhance soil health monitoring, promote sustainable land use, and ultimately contribute to maintaining soil's capacity to deliver essential ecosystem services (Schuler et al., 2011; Beamish, 2013; McFadden and Scott, 2013; Sarmast et al., 2017; Reinhardt and Herrmann, 2019).

The gamma-ray spectrometry measures the natural gamma radiation emissions from radionuclides such as potassium-40 ($K^{40}$); the daughter radionuclides of uranium-238 ($U^{238}$) and thorium-232 ($Th^{232}$) in soils, sediments, and rocks (Minty, 1988). This technique provides information on pedogenesis (Reinhardt and Herrmann, 2019), soil texture, mineralogy, pH and organic carbon (Wong and Harper, 1999; Taylor et al., 2002; Wilford and Minty, 2006; Barbuena et al., 2013; Priori et al., 2016).

* Corresponding author.
*E-mail address:* jamdemat@usp.br (J.A.M. Demattê).
[1] https://esalqgeocis.wixsite.com/english

https://doi.org/10.1016/j.soilad.2024.100024
Received 10 July 2024; Received in revised form 1 October 2024; Accepted 3 October 2024
Available online 10 October 2024

The intensity to which soil can be magnetised comprises soil magnetic susceptibility (κ) (Rochette et al., 1992). This property is related to soil mineralogy, parent material and the formation of magnetite and maghemite (ferrimagnetic minerals) (Ayoubi et al., 2018) and, less commonly, ferrihydrite and hematite (Valaee et al., 2016). Soil κ has been used in geological studies (Shenggao, 2000; Correia et al., 2010), soil granulometry and organic carbon determination (Camargo et al., 2014; Jiménez et al., 2017), soil survey (Grimley et al., 2004) and the study of soil-forming processes (Viana et al., 2006; Sarmast et al., 2017; Mello et al., 2020).

The ability of soil to conduct an electrical current comprises the apparent electrical conductivity (ECa). This property can be applied in pedology, indicating the existence/quantity of solutes in a soil solution (Richards, 1954). As a geophysical method, the ECa is able to identify soil's properties and their spatial variability, which can affect land use and management (Corwin et al., 2003). ECa is a function of soil salinity, clay mineralogy, clay content, cation exchange capacity, porosity, moisture and temperature (Mcneill, 1992; Rhoades et al., 1999; Bai et al., 2013; Cardoso and Dias, 2017).

Machine learning techniques have been applied in digital soil mapping to spatialize the above-mentioned soil geophysical attributes, besides modelling the variability of other attributes through the application of geophysical data. Among the main machine learning algorithms used, we can cite the random forests (RF) (Viscarra Rossel et al., 2014; Sousa et al., 2020; Siqueira et al., 2024), support vector machine (SVM) (Heggemann et al., 2017; Li et al., 2017; Zare et al., 2020), K - nearest neighbors (knn), artificial neural networks (ANN) (Dragovic and Onjia, 2007) and the Cubist tree model (Wilford and Thomas, 2012; Azizi et al., 2023). However, spatial predictions of soil properties based on small datasets of geophysical data (gamma-ray spectrometry, κ and ECa) are still underdeveloped.

Mapping of geophysical properties field sensors is usually conducted either remotely (via aerial platforms) or proximally (on-the-ground) (Wilford, 2012; Moonjun et al., 2017). Proximal geophysical surveys using ground vehicles can collect high-density data, but they require manual surveying, and accessing sites with complex terrain can be challenging (Parshin et al., 2018). In this situation, machine learning algorithms can be a useful tool for making the spatial prediction of geophysical attributes based on fewer samples, from association with environmental and topographic covariates that express the relationship of these attributes with the landscape.

Traditional methods of machine learning require a reasonable number of samples for calibrating (or training) the models and to obtain optimal spatial prediction of soil attributes. At the same time, one of the greater benefits of using digital soil mapping with machine learning is the possibility of obtaining predictions with known accuracy (McBratney et al., 2003). For machine learning tasks involving relatively large datasets, approximately 70–80 percent of the data is used for training, and 20–30 percent for testing (Moquedace et al., 2024; Siqueira et al., 2024).

However, in soil science, the number of samples available may be too small to create reasonable subsets of training and test, due to the difficulties of sampling. These difficulties are even greater for data obtained from geophysical techniques. In these cases, using small datasets-adapted evaluation methods has posed a great alternative, as the case of the nested leave one out cross-validation (nested-LOOCV). The nested-LOOCV method is recommended for small soil datasets (Mello et al., 2022a), for which other testing methods, such as holdout validation and cross-validation, would not be viable due to their low robustness with reduced number of samples (Ferreira et al., 2021; Paes et al., 2022).

Our previous study (Mello et al., 2022) used geophysical sensors and machine learning algorithms to model soil attributes, demonstrating that the integration of gamma-ray spectrometry and magnetic susceptibility data, combined with terrain and parent material information, can effectively predict soil properties. In this study, we explore ways to interpolate geophysical data using limited number of geophysical measurements.

This study had the following objectives: *i)* predict the spatial distribution of soil geophysical attributes (ECa, κ, eU, eTh and K (gamma-ray emission from $K^{40}$); *ii)* test the nested-LOOCV method and five machine learning algorithms (RF, Cubist, SVM, generalised linear models [LM] and adaptive multivariate regression) in a small dataset of soil and geophysical attributes; and *iii)* select the best algorithm for spatial prediction of each geophysical attribute, and to relate the attributes to pedogenesis.

## 2. Materials and methods

### 2.1. Study area and soil sampling

The study area is a 184-hectare farm recently cultivated with sugarcane, located in Southeast Brazil, between 23°00'31.37" and 22°58'53.97" S latitude and 53°39'47.81" and 53°37'25.65" W longitude (**Fig. 1**). It was described in Mello et al. (2022).The climate is subtropical mesothermal (Cwa) according to the Köppen classification system (Alvares et al., 2013). The mean temperature varies from 18°C in July (winter) to 22°C in February (summer), while the mean annual precipitation lies between 1100 and 1700 mm (Nanni and Demattê, 2006, Bazaglia Filho et al., 2013a).

In terms of geomorphology, the area is in the Paulista Peripheric Depression and is mainly composed of sedimentary rocks. The lithological composition of the area is: siltstone, metamorphosed siltstone, diabase and fluvial sediments (Bazaglia Filho et al., 2013a) (Fig. 2**a**).

The study area is composed of Cambisols, Phaeozems, Nitisols, Acrisols and Lixisols (Fig. 2**b**), reflecting the heterogeneity of the parent materials and relief. Pedologists have previously conducted soil surveys in the area (Nanni and Demattê, 2006; Bazaglia Filho et al., 2013b).

A total of 75 locations distributed throughout the study area was chosen. At each site, geophysical readings were performed on the soil surface (0–20 cm). Considering the complexity of the terrain and dense sugarcane cultivation, the readings with the geophysical sensors were performed in the most accessible parts, while simultaneously ensuring the representativeness of the area.

### 2.2. Geophysical data collection

#### 2.2.1. Magnetic method: soil magnetic susceptibility (κ)

Soil magnetic susceptibility (κ) values were collected via the geophysical sensor Terraplus KT10 model (Fig. 1**A**). This sensor measures soil magnetic susceptibility values up to 2 cm below the soil surface (precision of $10^{-6}$ SI units in $m^3\ kg^{-1}$). For κ readings, the sensor was first calibrated following the recommendations of the manufacturer (Sales and C., 2021). The soil κ readings were performed with the sensor in scanner mode. Due to small variations in sensor readings (noise effect), three readings were performed around each collection point, and the mean values of these readings were used for analysis.

#### 2.2.2. Radiometric method: soil radionuclides via gamma-ray spectrometry

Potassium, uranium, and thorium and radionuclides values ($K^{40}$, eU and eTh) were obtained using the near-gamma-ray spectrometer (GM -Radiation Solution RS 230 (Radiation Solution INC, Ontario, Canada) (Fig. 1**B**). The sensor measures radionuclides with an average depth about 30 cm below the soil surface. For radionuclides readings, the sensor was automatically stabilized. Then, the readings were taken with the sensor in direct contact with the soil surface for 2 minutes, in the "essay" mode (which provides better accuracy) (Solutions, 2009). Potassium ($K^{40}$) values were reported in % while uranium (eU) and thorium (eTh) values were given in $mg\ kg^{-1}$ due to the higher and lower proportion of these elements in the environment, respectively.

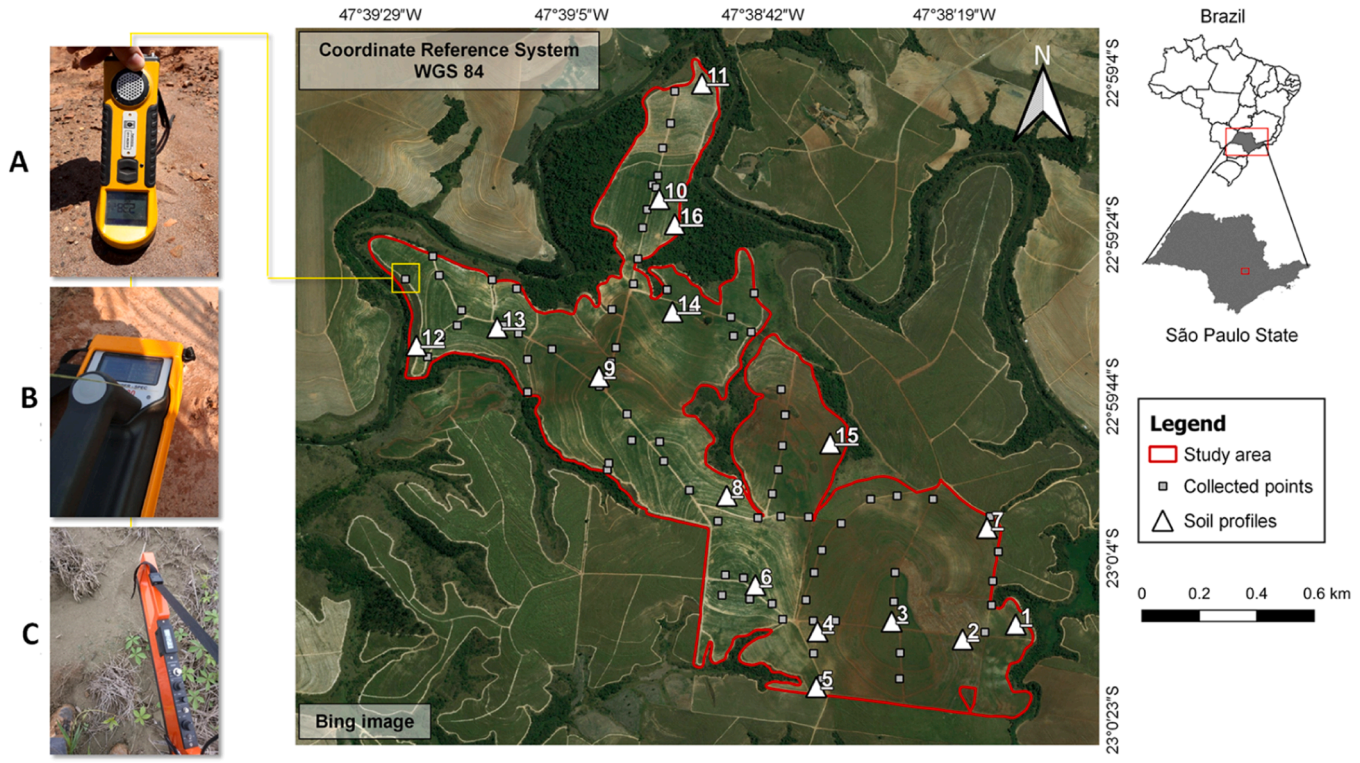


**Fig. 1.** Study area, collected points and geophysical sensors. A - Susceptibilimeter (KT-10 Terraplus); B - Gamma-ray spectrometer (Radiation Solution - RS 230); C - Geonics Ground Conductivity Meter (EM 38).

#### 2.2.3. *Electric method: soil apparent electrical conductivity*

The readings of the soil apparent electrical conductivity (ECa) were undertaken via the geophysical sensor Geonics EM38 (Geonics Ltd., Mississauga, Ontario, Canada) (McNeill, 1986) (Fig. 1**C**). The sensor was previously calibrated following the recommendations of the manufacturer's instruction manual (Heil and Schmidhalter, 2019). The EM38 sensor was positioned vertically in contact with the soil for measurement. This reading is an integrated soil ECa values down to a depth of 1.5 m in $mSm^{-1}$. The ECa readings were taken during the dry period (winter) and during the same period of the day to ensure that humidity was a constant variable and metal objects were removed from the sensor's proximity to avoid interference.

### 2.3. *Digital elevation model and covariates*

A digital elevation model (DEM) was generated using a topographic map with 5-meter contour intervals at a 1:10,000 scale, sourced from the Campinas Geographic Institute. The contour lines were interpolated into a DEM using the Topo to Raster function in ESRI ArcGIS 10.4, and the final DEM was exported at a spatial resolution of 30 m. Based on the DEM, 32 additional terrain attributes were calculated (Table 1) using the *R* software version 4.1.0 (RC Team, 2021), through the "*Rsaga*" (Brenning, 2008) and "*raster*" (Hijmans and Van Etten, 2016) packages.

We also utilized the lithology map layer, created by an expert pedologist at a 1:10,000 scale (Nanni and Dematte, 2000), as a covariate representing the parent material. as a covariate representative of the parent material. Considering the necessity of integrating a categorical variable into the modelling process, the lithology variable was transformed into four new dummy covariates, each for a class of the original layer (siltstone, metamorphosed siltstone, diabase and fluvial sediments).

### 2.4. *Model processing*

The detailed description of the methodological framework is presented in Fig. 3 and comprises four main steps (1) selection of environmental covariates; (2) training process with different algorithms; (3) evaluation of model's performance in the testing process and (4) spatial prediction and uncertainty analysis using the best fitted model.

#### 2.4.1. *Selection of covariates*

Several potential covariates can be used to predict the spatial distribution of soil and geophysical attributes. However, using many covariates, requires computational effort and generates complex final models that are difficult to explain. To overcome this problem, we applied three steps: (1) removal of variables with low variance (near zero), (2) removal of covariates that were highly correlated and (3) selection of important covariates for prediction.

Step 1 removed covariates that have variance near zero applying the nearZeroVar function, using the *caret* package (Kuhn et al., 2020). Then, the remaining covariates were subjected to the phase 2.

Step 2 comprised removing covariates with high correlation. This phase was performed because highly correlated covariates provide redundant information and contribute little to the modelling process. In this phase, Pearson's correlation coefficients were calculated for all covariates, separating those with a $> 95$ % linear correlation value. This processes was calculated using the "*find correlation*" with *caret* package (Kuhn et al., 2020). Then, training and testing samples partition was done, using the nested-LOOCV method, which will be described later. Only training samples were used in the next step.

Step 3 involved removing covariates that did not contribute significantly to the modeling process using recursive feature elimination (RFE) from the caret package (Kuhn et al., 2020). RFE is a backward selection method that iteratively reduces the number of predictors (Kohavi and John, 1997). It ranks covariates by their importance, groups them into subsets, and evaluates these subsets using simpler models based on their

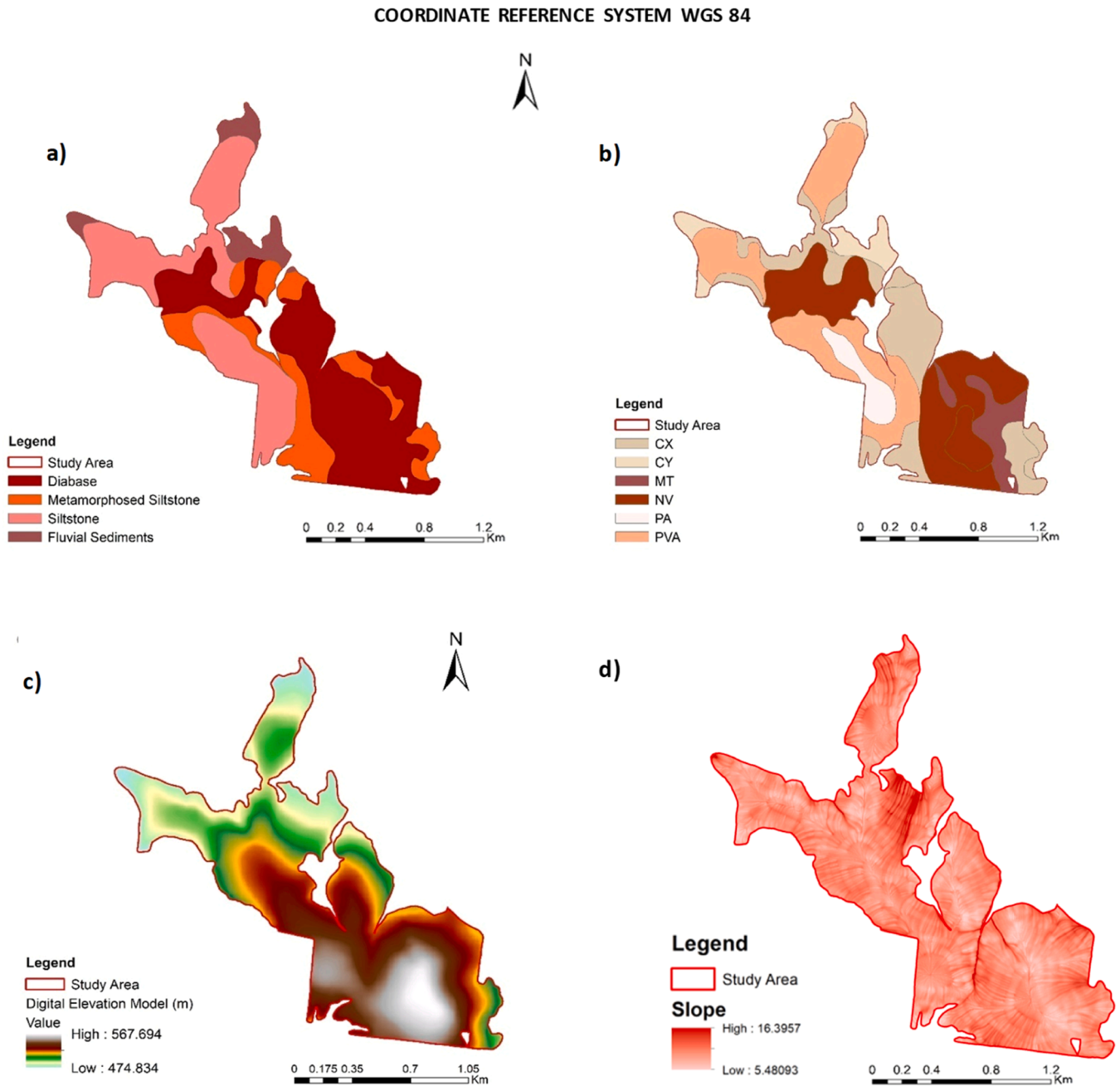


**Fig. 2.** a) Geological compartments of landscape. b) Soil classes: CX: Haplic Cambisols, CY: Fluvic Cambisols, MT: Luvic Phaozem, NV: Rhodic Nitisol: PA: Xanthic Acrisol, PVA: Rhodic Lixisol. The geological and Soil classes maps were adapted from Bazaglia Filho et. al. (2012). c) Digital Elevation Model: d) Slope.

performance. The subset with the best performance is then selected for final model training, representing the most influential covariates for predicting the target phenomenon. In this study, the remaining covariates after correlation removal were grouped into sixteen subsets with varying numbers of covariates (ranging from 5 to 25), including the full set, and each subset was tested iteratively.

RFE considers the base algorithm (e.g., RF, SVM, lm, etc.) (Kuhn and Johnson, 2013), which means that each algorithm had a specific RFE model. For the RF algorithm, *rfFuncs* from the *caret* package was used. For other algorithms, the *caretFuncs* support function were used (Kuhn, 2012).

### 2.4.2. *Nested leave one out cross-validation ("nested-LOOCV")*

The covariates selected by the RFE were associated with soil geophysical data and were used in the training and test processes (Fig. 3). Considering the small number of samples (n = 75), we applied the nested-LOOCV method, which is indicated for modelling databases with a limited number of samples (small number ≤ 100 samples).

The methods used for separating the training and test data as well as the Nested-LOOCV will be detailed next:

- ✓ Firstly, our total dataset contains 75 samples. For the modelling process, we divided the dataset into two subsets: subset1 (training) and subset2 (test).
- ✓ *subset 1* contains 74 samples, (all 75 samples minus one (75−1)).
- ✓ *subset 2* contains only 1 sample, corresponding to the sample that was removed from subset 1.
- ✓ *subset 1* and *2* were further separated using the nested-LOOCV method (Honeyborne et al., 2016; Mello et al., 2022a; Paes et al., 2022; Rytky et al., 2020).

**Table 1**
Terrain attributes generated from the digital elevation model.

| Terrain attributes | Abbreviations | Brief description |
|---|---|---|
| **Convergence index** | CI | Convergence/divergence index in relation to runoff |
| **Cross sectional curvature** | CSC | Measures the curvature perpendicular to the down slope direction |
| **Flow line curvature** | FLC | Represents the projection of a gradient line to a horizontal plane |
| **General curvature** | GC | The combination of both plan and profile curvatures |
| **Hill** | HI | Analytical hill shading |
| **Hill index** | HIINDEX | Analytical index hill shading |
| **Longitudinal curvature** | LC | Measures the curvature in the down slope direction |
| **Mass balance index** | MBI | Balance index between erosion and deposition |
| **Maximal curvature** | MAXC | Maximum curvature in local normal section |
| **Mid-slope position** | MSP | Represents the distance from the top to the valley, ranging from 0 to 1 |
| **Minimal curvature** | MINC | Minimum curvature for local normal section |
| **Multiresolution index of ridge top flatness** | MRRTF | Indicates flat positions in high altitude areas |
| **Multiresolution index of valley bottom flatness** | MRVBF | Indicates flat surfaces at bottom of valley |
| **Normalized height** | NH | Vertical distance between base and ridge of normalized slope |
| **Plan curvature** | PLANC | Described as the curvature of the hypothetical contour line passing through a specific cell |
| **Profile curvature** | PROC | Describes surface curvature in the direction of the steepest incline |
| **Slope** | S | Represents local angular slope |
| **Slope height** | SH | Vertical distance between base and ridge of slope |
| **Standardized height** | STANH | Vertical distance between base and standardized slope index |
| **Surface specific points** | SSP | Indicates differences between specific surface shift points |
| **Tangential curvature** | TANC | Measured in the normal plane in a direction perpendicular to the gradient |
| **Terrain ruggedness index** | TRI | Quantitative index of topography heterogeneity |
| **Terrain surface convexity** | TSC | Ratio of the number of cells that have positive curvature to the number of all valid cells within a specified search radius |
| **Terrain surface texture** | TST | Splits surface texture into 8, 12, or 16 classes |
| **Total curvature** | TC | General measure of surface curvature |
| **Topographic position index** | TPI | Difference between a point elevation with surrounding elevation |
| **Valley depth** | VD | Calculation of vertical distance at drainage base level |
| **Valley** | VA | Calculation fuzzy valley using the Top Hat approach |
| **Valley Index** | VAI | Calculation fuzzy valley index using the Top Hat approach |
| **Topographic wetness index** | TWI | Describes the tendency of each cell to accumulate water as a function of relief |

✓ The nested-LOOCV is made up of an inner and outer loop (Fig. 4).

The inner loop is performed with the samples from *subset 1*. In this process the samples are re-divided into training (***subset A***) and training validation (***subset B***) samples. *Subset A* is resampled with the removal of a sample (74−1), consisting of 73 samples. *Subset B* is composed of the sample that was removed, consisting of 1 sample. *Subset A* is used for a "internal training" and *subset B* for a "internal testing". The inner loop is run 74 times (corresponding to the number of samples in *subset 1*), and at each round the removed sample (*subset B*) is changed subsequently. The entire process consists of the LOOCV method (Kuhn and Johnson, 2013). At the end of this process, the performance of the training is calculated considering the prediction over the subset B at each round.

The outer loop is performed with the samples from *subset 2* (test, with 1 sample), which do not participate in the training process. As the outer loop is run 75 times (one for each sample of the total dataset), the sample of subset 2 is swapped 75 times. In this process, the sample that had been removed returns to *subset 1* and another sample is relocated to *subset 2*, which makes the subsets being alternated every round. At the end of 75 rounds of the outer loop, it results in 75 pairs of predicted and observed values, which are used to calculate the test parameters of model performance. This full round of loops is the principle of Nested-LOOCV.

### 2.4.3. Machine learning algorithms, training and covariates importance

In this study, we tested five algorithms: random forests (RF), Cubist model (C), support vector machines (SVM) with Radial Basis Function Kernel, adaptive multivariate regression (Earth) and generalized linear models (LM). We selected these models to explore different families and linear as well as non-linear algorithms that have been used widely in digital soil mapping (DSM) studies (Hengl et al., 2017; Gomes et al., 2019; Khaledian and Miller, 2020). Algorithms from different 'families' have specific characteristics for processing and optimization. For example, the RF and C are decision tree algorithms; the Earth and LM are linear models; and the SVM is a kernel-based model.

The training was performed using the group of covariates selected by RFE for each algorithm and using the LOOCV method of the inner loop. For training, the hyperparameters of each algorithm were optimized using 5 possible values in the argument *tuneLength* of the *train* function. The optimized hyperparameters used for each algorithm tested are demonstrated in Table 2. The process of hyperparameters optimization is described in chapter 6 of the caret package manual (available at https://topepo.github.io/caret/train-models-by-tag.html.).

Additionally, the importance of covariates for training the models was obtained with the *varImp* function of the caret package, with results averaged over the 75 loops. With this function, values of importance are normalized for a scale of 0–100, where the most important predictor is at 100 and the least important at 0 (Kuhn, 2012).

### 2.4.4. Model's performance

After the training, the model's performance metrics were obtained with the mean of $n$ rounds ($n = 75$ in this study; hence, the parameters were calculated based on the mean of 75 rounds). To evaluate the model's performance, we applied the fitted model to the test data and the accuracy was expressed by the following statistical indexes: R-squared ($R^2$) (Eq. (1)), root mean squared error (RMSE) (Eq. (2)) and mean absolute error (MAE) (Eq. (3)). These indices are very used in the digital soil mapping and are considered very robust to evaluate comparatively the performance of different machine learning models (Gomes et al., 2019; Siqueira et al., 2023, Moquedace et al., 2024). The $R^2$ indicates the proportion of the variance in the target variable that the model explains. In turn, the MAE and RMSE are error metrics related to the models' residuals. They describe the absolute accuracy of the models, which means how close the predicted values are to the actual values. The best model for the spatial prediction and maps creation was the one that presented greater $R^2$ and smaller values of RMSE and MAE.

$$R^2 = \frac{[\sum(\text{Qpred} - \overline{\text{Qpred}}) \times (\text{Qobs} - \overline{\text{Qobs}})]^2}{\left[\sum(\text{Qpred} - \overline{\text{Qpred}})^2\right] \times \left[\sum(\text{Qobs} - \overline{\text{Qobs}})^2\right]} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \times \sum(\text{Qobs} - \text{Qpred})^2} \quad (2)$$

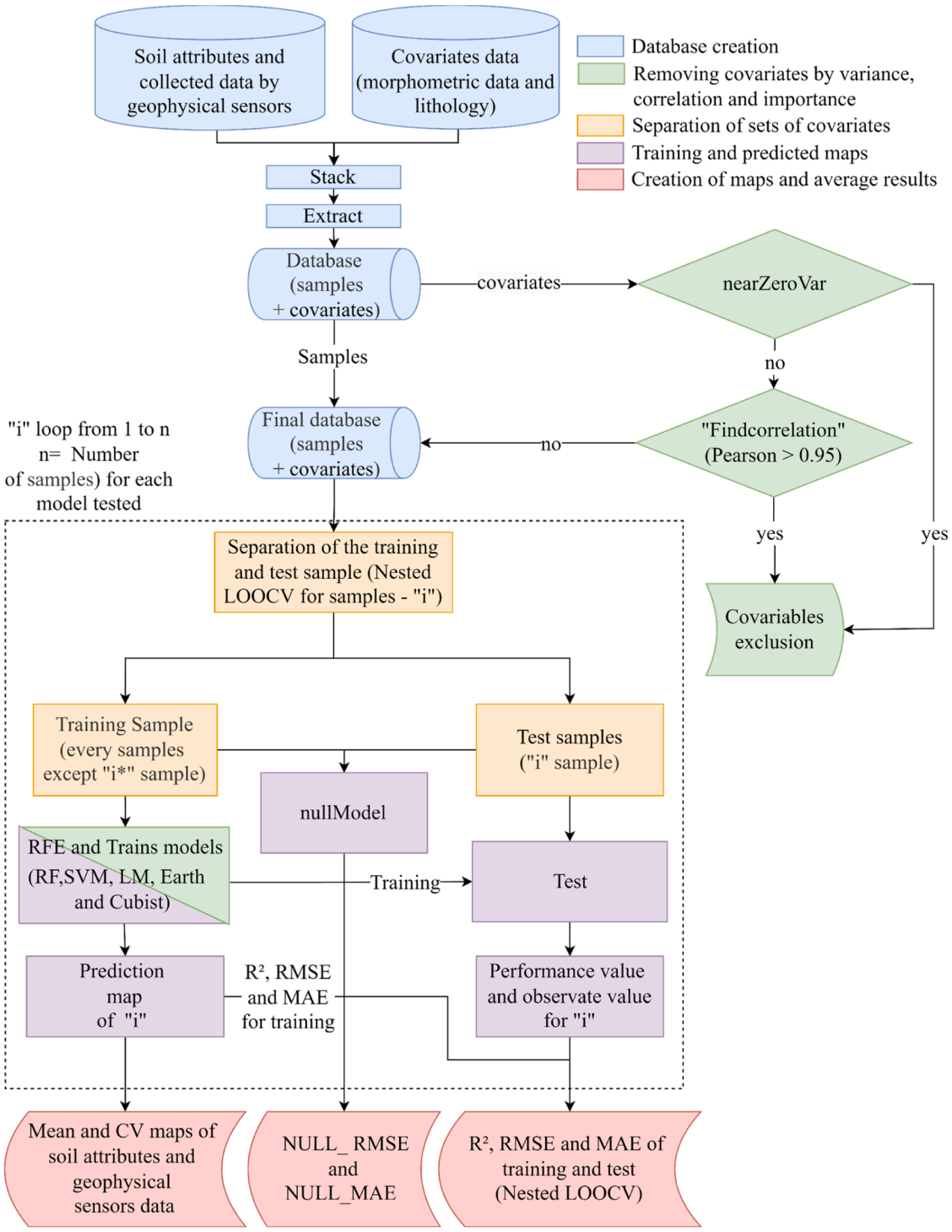


**Fig. 3.** Methodological flowchart showing the sequence of methodologies applied for soil and geophysical attributes prediction. The most accurate model between Cubist(CUB), Random Forests (RF), Support Vector Machines (SVM), Earth and Linear Models (LM) was selected to model and map the geophysical and soil attributes maps.

$$MAE = \frac{1}{n} \times \sum |Qpred - Qobs| \quad (3)$$

Qpred = predicted samples
Qobs = observed samples
n = the number of samples

As an additional validation, we also calculated the RMSE and MAE for the null model (NULL_RMSE and NULL_MAE) (Eqs. (4 and 5)). The null model is considered as the simplest model with predicted values represented by the mean value of the observations. In this way, the NULL_RMSE and NULL_MAE can be used as a baseline to compare the trained models. Any model that presents a lower RMSE and MAE relative to the 'null versions' should not be discarded, since the performance is superior to the simple mean. The null models were estimated using the nullMode function in the *caret* package (Kuhn et al., 2020).

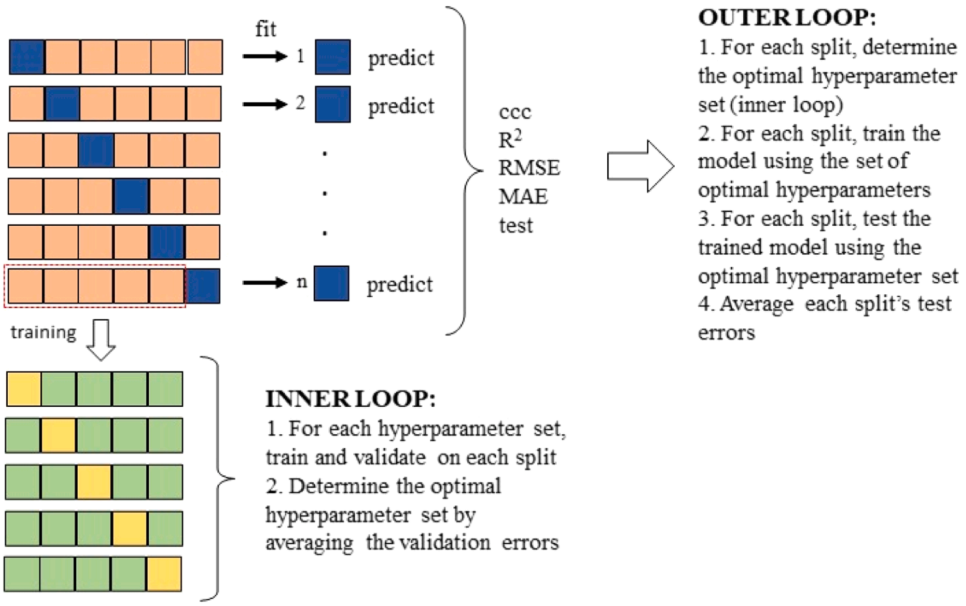


**Fig. 4. :** Methodological flowchart for *Nested Leave One Out Cross-Validation* (nested-LOOCV) method. ccc: Lin's Concordance Correlation Coefficient.

**Table 2**
The optimized hyperparameters used for each algorithm tested.

| Algorithms | hyperparameters |
|---|---|
| ***RF*** | *mtry* |
| ***Cubist*** | *committees; neighbors* |
| ***SVM*** | *sigma; C(cost)* |
| ***LM*** | *intercept* |
| ***Earth*** | *nprune; degree* |

***RF:*** random forests (RF); ***SVM:*** Support Vector Machines with Radial Basis Function Kernel; ***Earth:*** Adaptive Multivariate Regression; ***LM:*** Generalized Linear Models.

$$NULL\_RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(\overline{Qtrain_i} - \quad Qobs_i)^2\right]^{\frac{1}{2}} \tag{4}$$

$$NULL\_MAE = \frac{1}{n} \times \sum|\overline{Qtrain_i} - Qobs_i| \tag{5}$$

$Qtrain$ = the mean of the training samples
$Qobsi$ = the validation sample
$N$ = the number of samples (loop).

#### 2.4.5. Spatial prediction and uncertainty

The predicted maps were generated from the 75 loops of RFE/training using the best model for each soil geophysical attribute. The final maps were generated from the mean of the 75 predicted maps. We also map of the coefficient of variation of prediction (CV% = standard deviation / mean). Values with a lower CV show a more consistent results or less uncertain.

## 3. Results

### 3.1. Covariate's importance

The models selected unique sets of covariates for each geophysical attribute (Fig. 5). Considering the best model for each attribute, the RFE selected the largest number of covariates for the eTh model, which were trained with 29 variables. On the other hand, magnetic susceptibility had the lowest number of covariates, a total of 12.

The most important covariates for the prediction of $K^{40}$ were the metamorphosed siltstone (MST) and siltstone (ST). Among the terrain attributes, the most important were the minimal curvature (MINC) and normalized height (NH). For the eU, the most important variable was the parent material diabase (D), and terrain attributes, mainly the digital elevation model (DEM) and standardized height (STANH). For the eTh, the most important covariates were the maximal curvature (MAXC) and multiresolution index of ridge top flatness (MRRTF). Although the most significant of the parent material variable, the dummy variables of the lithology classes presented very low importance, which separated the eTh from the other gamma ray attributes (Fig. 5).

The magnetic susceptibility (κ) presented the diabase and DEM as the most relevant covariates with importance of more than 50 % to the κ model. At last, the most important covariates for the ECa were the SH and parent material, with 100 % of importance, followed by the topographic position index (TPI), general curvature (GC) and standardized height (STANH). The diabase was the more relevant lithological class, with importance of more than 50 % (Fig. 6).

### 3.2. Model's performance and uncertainty

With few exceptions, the models that presented the greatest performances for all geophysical attributes were the RF and SVM. For all machine learning algorithms, the κ and $K^{40}$ models presented the best performance (Table 3), which evidences the greatest correlation of these geophysical attributes with the parent material and terrain covariates used. The RF presented the best performance for the $K^{40}$ attribute, with $R^2$ of 0.36 presenting the largest difference for the second position (SVM with $R^2 = 0.22$), besides the lowest RMSE (0.26) and MAE (0.18). RF and SVM shared the greatest $R^2$ values for κ (0.49). However, considering the RMSE and MAE, RF presented the greatest performance. The SVM models presented the greatest performance for eU, eTh and ECa, although the $R^2$ values were considerably lower than the previous geophysical attributes. The SVM presented $R^2$ of 0.11, 0.10 and 0.09 for eU, eTh and ECa, respectively (Table 3).

The best performances of RF and SVM are associated with the better generalization capability of these models, besides the greatest capacity to handle the non-linear relationships between the geophysical attributes and the covariates. In turn, the worst models were the LM and Earth models. The Earth presented the lowest $R^2$ values for the gamma spectrometric variables, whereas the LM presented the lowest values for the ECa and κ (Table 3). This is associated with the fact that these models

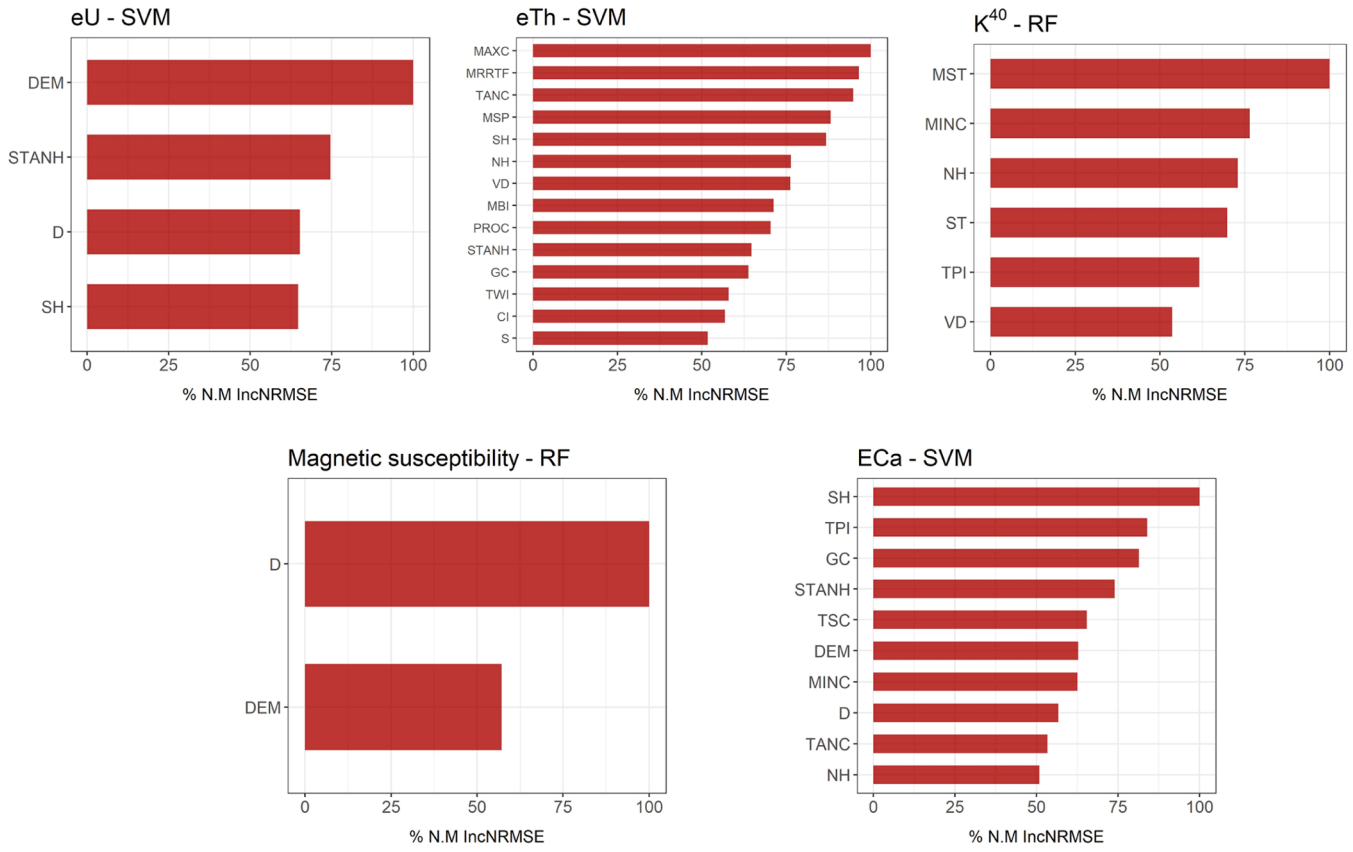


**Fig. 5.** Importance of predictor variables (parent material and terrain attributes) for geophysical attributes. Equivalent uranium (eU), equivalent thorium (eTh), potassium ($K^{40}$), magnetic susceptibility (κ) and apparent electrical conductivity (ECa). Only the variables with importance higher 50 % are represented.

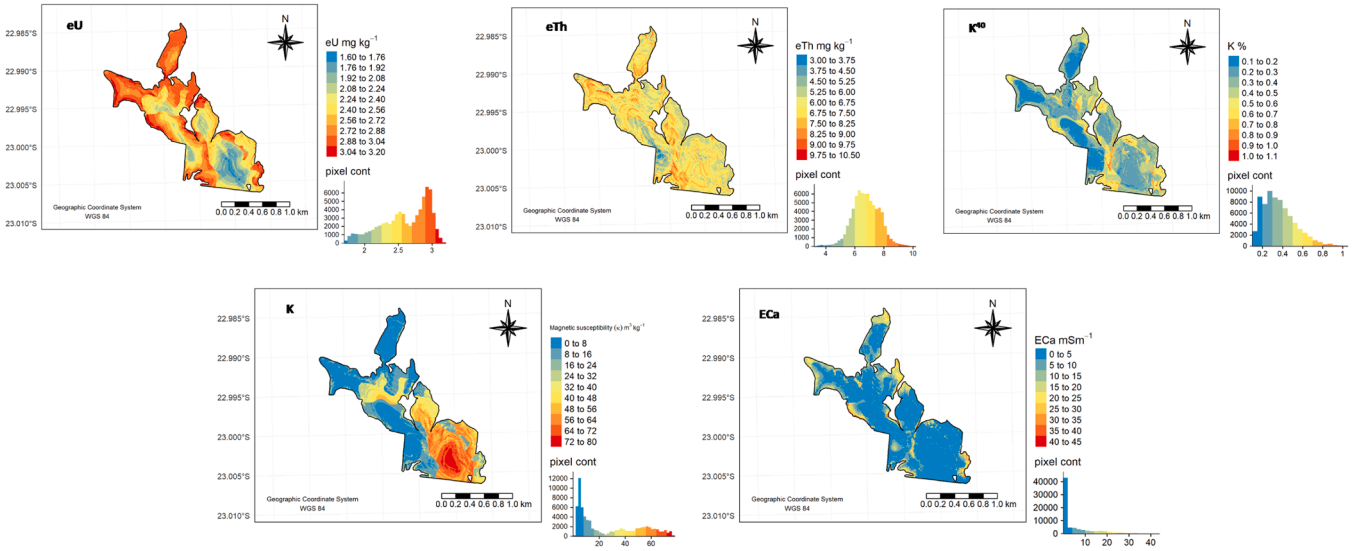


**Fig. 6.** Spatialized geophysical attribute data. Equivalent uranium (eU), equivalent thorium (eTh), potassium ($K^{40}$), magnetic susceptibility (κ) and apparent electrical conductivity (ECa).

are not able to work with the non-linearity of soil data, which limits considerably their performance.

The RMSE and MAE values of the best models for each geophysical attribute presented values always lower than the NULL_RMSE and NULL_MAE (Table 3), whereas for many of the other models this did not happen. In this way, the RF algorithm was selected for the spatial prediction and distribution of $K^{40}$ and κ, while the SVM algorithm was selected for eU, eTh and ECa. Additionally, the κ was the only attribute where even the low-performance models (LM and Earth) presented RMSE and MAE values higher than those of the null models, which corroborate this geophysical attribute was the most appropriate for estimation by machine learning.

**Table 3**
*Outer loop* table. Models' performance for the geophysical attributes, based on $R^2$, RMSE, MAE and NULL_RMSE.

| ***Geophysical attributes*** | ***$R^2$*** | | | | | |
|---|---|---|---|---|---|---|
| | ***Random Forest*** | ***Cubist*** | ***SVM*** | ***LM*** | ***Earth*** | |
| $K^{40}$ | **0.362** | 0.216 | 0.229 | 0.204 | 0.002 | |
| eU | 0.107 | 0.045 | **0.109** | 0.062 | 0.001 | |
| eTh | 0.017 | 0.001 | **0.095** | 0.017 | 0.003 | |
| ECa | 0.024 | 0.000 | **0.080** | 0.002 | 0.046 | |
| κ | **0.490** | **0.490** | 0.444 | 0.340 | 0.447 | |
| ***Geophysical attributes*** | ***RMSE*** | | | | | |
| | ***Random Forest*** | ***Cubist*** | ***SVM*** | ***LM*** | ***Earth*** | ***NULL_RMSE*** |
| $K^{40}$ | **0.261** | 0.295 | 0.294 | 0.313 | 1.430 | 0.331 |
| eU | 0.720 | 0.854 | **0.716** | 0.818 | 2.788 | 0.762 |
| eTh | 2.678 | 2.575 | **2.330** | 2.810 | 4.900 | 2.478 |
| ECa | 33.640 | 36.740 | **31.880** | 36.600 | 53.150 | 33.053 |
| κ | 25.070 | **23.876** | 24.520 | 28.257 | 24.160 | 32.832 |
| ***Geophysical attributes*** | ***MAE*** | | | | | |
| | ***Random Forest*** | ***Cubist*** | ***SVM*** | ***LM*** | ***Earth*** | ***NULL_MAE*** |
| $K^{40}$ | **0.182** | 0.196 | 0.186 | 0.225 | 0.485 | 0.239 |
| eU | **0.566** | 0.605 | **0.566** | 0.650 | 1.020 | 0.593 |
| eTh | 1.911 | 1.924 | **1.620** | 2.095 | 2.770 | 1.805 |
| Eca | 22.630 | 26.700 | **21.890** | 25.720 | 32.670 | 23.875 |
| κ | 16.600 | **15.694** | 15.970 | 19.623 | 16.890 | 26.131 |

$K^{40}$ in %, eU and eTh in mg $kg^{-1}$; AEC in dSm $^{-1}$; κ in $m^3$ $kg^{-1}$. Abbreviations: $K^{40}$: (potassium by gamma-ray spectrometer); eU: (equivalent uranium by gamma-ray spectrometer); eTh: (equivalent thorium by gamma-ray spectrometer); ECa: Apparent electrical conductivity by Geonics EM38 (geophysical sensor); κ: magnetic susceptibility by KT10-Terraplus (geophysical sensor). ***SVM:*** Support Vector Machines with Radial Basis Function Kernel; ***LM:*** Generalized Linear Models; ***Earth:*** Adaptive Multivariate Regression.

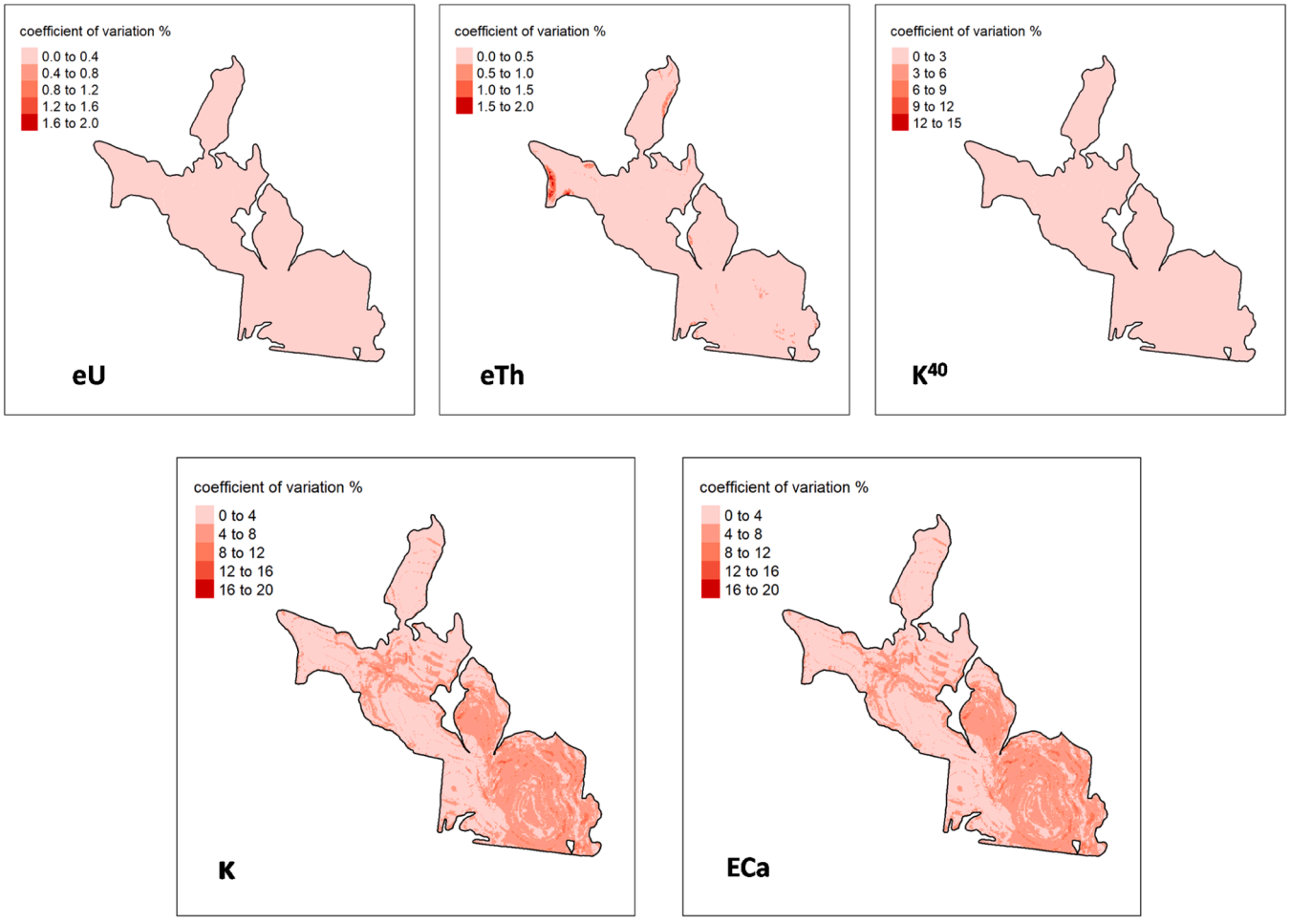


**Fig. 7.** Coefficient of variation for geophysical attributes. Equivalent uranium (eU), equivalent thorium (eTh), potassium ($K^{40}$), magnetic susceptibility (κ) and apparent electrical conductivity (ECa).

### 3.3. Spatial prediction and uncertainty

The final map of $K^{40}$ presented minimum and maximum values of 0.1 % and 1.1 %, respectively. The $K^{40}$ presented a spatial variability clearly marked by the parent material, with lower contents over siltstones and greater contents over metamorphosed siltstones. For the eU, the values ranged from 1.6 to 3.2 and presented a spatial pattern marked by the influence of lithology and topography. The lowest contents of eU were spatially associated with metamorphosed siltstones and areas of lower altitude, the last ones mainly in the western part of the study area. At the same time, the greatest contents were located over the highest parts with dominance of diabase (Fig. 6).

The eTh values ranged from 3.0 to 10.50 mg $kg^{-1}$, and showed a more complex spatial pattern, associated with the great number of covariates used for prediction. The ECa values varied 0 to little values above 30 %. Overall, the ECa was below 5 % at most of the study area, with the exception of higher values in regions of lower altitude, mainly over the plains with fluvial sediments. At last, the κ predicted contents with RF varied from 0 to 80 kg $m^3$ $kg^{-1}$ and were strongly related to the parent material spatial distribution, with the greatest and lowest values found over the diabase and siltstone rocks, respectively (Fig. 6).

Regarding the uncertainty, most maps presented values of coefficient of variation predominantly below 5 %, which indicates a low uncertainty. The eU and eTh predictions presented the best results, with uncertainty not exceeding more than 2 % in the entire study area. In turn, the greatest uncertainties were observed for the κ and ACE, which reached maximum values of coefficient of variation of 20 % (Fig. 7).

## 4. Discussion

### 4.1. Model performance evaluation

The methodological framework using the nested-LOOCV optimised the prediction of soil geophysical variables using a small number of samples, besides providing reliable performance results. Regarding the gamma-ray spectrometric variables, the RF algorithm best predicted $K^{40}$ while the SVM algorithm best predicted eU and eTh. Our results are similar to those found by Viscarra Rossel et al., (2014), who also used the RF algorithm to predict $K^{40}$ in soils from Tasmania, reporting an $R^2$ of 0.43. On the other hand, Cracknell and Reading (2014) stated that the RF algorithm was the best model for multiclass inference using widely available, high-dimensional multisource remotely sensed geophysical variables.

The high importance of the parent material predictor aligns with findings from Dickson and Scott, (1997), Wilford et al. (1997), and Wilford and Minty, (2006), who demonstrated a strong relationship between $K^{40}$ and soil parent materials. Mello et al. (2023a), (2023b) also observed that greater soil weathering, influenced by relief and water movement, results in lower $K^{40}$ values due to increased K leaching. However, our spatial prediction indicated that parent material had a greater influence than weathering and topography in explaining the distribution of $K^{40}$. The lowest $K^{40}$ values were found in areas with siltstones, rocks naturally poor in K due to the pre-weathering. In turn, the highest $K^{40}$ values were found for metamorphosed siltstones, indicating that the thermal metamorphism that affected parts of the siltstone promoted enrichment in K, probably from hydrothermalism (Soares et al., 2004; Rosales et al., 2019).

The low $R^2$ of the RF model for eU and eTh disagrees with results obtained by Anic and Dragovic (2005) who found an $R^2 > 0.90$ using a different approach (neural network algorithm), and Sousa et al. (2020), who used an RF algorithm that efficiently predicted eU and eTh ($R^2 > 0.90$). However, it is important to highlight that these authors used data from an airborne gamma sensor to estimate radionuclide values, which correlations are high, justifying the high $R^2$ values found by the authors.

There are several possible explanations for the low performance of eU and eTh. The studied area is very heterogeneous regarding the parent material (four lithological classes) and soil types (six soil classes in 184 ha) (Mello et al., 2020, 2021). *In-situ* evaluation brings several uncontrolled factors such as rocks or fragments, soil mixture with plant residue, fertilizers and different moisture conditions that could impact the prediction and reduce the $R^2$. Other possible explanations for the low $R^2$ are related to the various effects of field sampling data and larger variations in mineralogy, soil type, pedogenesis and their interactions (Mello et al., 2021).

The lithological covariate was also very important to the eU and eTh predictions. The eU contents were clearly lower in the diabase area where Mello et al. (2023a), (2023b) identified the greatest weathering and leaching rates. Uranium is considered high mobility under oxidizing conditions (Modena et al., 2016). Its increased remotion with leaching from the higher parts of the study area, where the greater weathered and drained soils are located, corroborates the great importance of DEM as a predictor. The importance altitude is also revealed with the greater contents of uranium in the lower parts of the study area, indicating that the accumulation of uranium leached from the higher parts. Among the parent materials, the diabase was the most important class, and the natural lowest contents of uranium of this lithology can be pointed out as another factor to explain the lowest eU contents, since mafic rocks (with the lowest contents of silica) tend to present lower contents of uranium (Modena et al., 2016).

Although lithology was also relevant for this attribute, the eTh contents did not present significant spatial differentiations related to the lithology. By studying the relationship between gamma ray spectrometric attributes and rock weathering in Southern Brazil, Modena et al. (2016) also did not find significant differences between distinct lithologies. Different from uranium, thorium is a relatively immobile element that tends to accumulate in soil according to weathering progress. The spots of lower thorium contents observed in this study had a strong spatial correlation with the zones of less weathered and developed soils as found by Mello et al. (2023a), (2023b). The authors associated the lower weathering in this part of the study area with the great slope, which limits the pedogenesis through erosive processes. This idea is corroborated by the importance of curvature covariates in our study. The erosion in steep slopes is also responsible for the remotion of the immobilized thorium, which is removed with the clastic sediments.

In relation to κ, the RF and SVM models presented satisfactory performance, with $R^2$ of 0.5, indicating 50 % of the total data variance explained by the models, which produces more reliable predictions. The greatest performance of the κ models is related to the strong relationship between soil κ and the parent material, clay and total iron content (Mello et al., 2020; Siqueira et al., 2010), variables that are directly or indirectly related to the covariates used in this work. According to Blundell et al. (2009) a range of 36 % - 46 % of the variances in the magnetic parameters can be explained by the parent material and drainage network.

The nature of mafic parent rock drives the content of magnetic minerals (Ayoubi et al., 2019; Karimi et al., 2017; Teixeira et al., 2018) and the formation of secondary ferrimagnetic minerals (Jordanova, 2016; Mullins, 1977) responsible for the soil κ (Dearing, 1999). This explains the greatest κ contents found over diabase in the study area, dominated by highly weathered Nitisols with clayey B horizons rich in iron oxides (Mello et al., 2023a, 2023b). In turn, the great importance of variables such as DEM and STANH highlights the greater contribution of the topography to the formation of more weathered soils in the higher parts of the study area. According to Mello et al. (2022b) ferralitization dominated in this area. The desilication and concentration of iron oxides in chemically strongly leached soils, tend to contribute to the greater values of soil κ.

Finally, the prediction model of the ECa was the worst among all geophysical attributes. The lowest $R^2$ for the ECa could be related to the small number of collected samples that do not represent the different soil salinity spatial patterns, as reported by Johnston et al. (1997) and Lesch et al. (1992). Soil ECa is usually strongly related to the soil texture,

mainly the clay content (Brus et al., 1992; Weller et al., 2007), which is also related to the parent material. Although parent material was an important covariate for the ECa prediction, the spatial variability of the final ECa maps does not follow the lithology trend, but the relief. The largest ECa were found on areas of lower altitude, mainly associated with the fluvial sediments of the edges of the study area. This reveals the important function of the topography in redistributing salinity from the higher to lower zones in tropical lands, independent of the parent material. The incorporation of salinity-related soil attributes, such as clay contents, is a strategy to improve the ECa prediction (Brus et al., 1992; Harvey and Morgan, 2009).

## 5. Conclusions

Machine learning algorithms produced reliable results for spatially predicting soil geophysical attributes, outperforming null models based on mean values. The null model serves as an effective benchmark for evaluating machine learning outcomes, demonstrating the potential of these models even when $R^2$ values are low. The use of the nested-LOOCV method proved suitable for soil and geophysical datasets with limited samples, providing a robust approach for model performance evaluation and optimizing algorithm training and testing. Moreover, maps generated with nested-LOOCV showed greater spatial consistency, supported by improved results interpretation.

The RF and SVM algorithms presented the best results. Parent material and DEM were the covariates that most contributed to the prediction and controlled the estimated spatial variability of the geophysical variables ($K^{40}$, eU, eTh, ECa and κ).

Machine learning techniques are valuable tools for modeling soil geophysical variables, especially in scenarios with limited observations. The integration of proximal sensing with computational methods highlights the importance of geophysical measurements for estimating other labor-intensive soil attributes (e.g., chemistry, texture, mineralogy), providing critical insights into soil genesis and fertility. This demonstrates the strong potential of geophysical techniques in advancing soil science.

## Open research (availability statement)

All analyses and codes used in this research were developed in R software version 4.0.3 (R Core Team, 2015; Kuhn et al., 2013). The entire database and codes used in the analyses are available at: https://zenodo.org/record/6302012#.Yl6uiujMKUk; DOI: 10.5281/zenodo.6302012 (Veloso et al., 2021).

## CRediT authorship contribution statement

**Rafael Gomes Siqueira:** Writing – review & editing, Visualization, Formal analysis. **Lucas Carvalho Gomes:** Writing – review & editing. **Elpídio Inácio Fernandes-Filho:** Writing – review & editing, Validation, Software, Methodology. **Carlos Ernesto Gonçalves Reynaud Schaefer:** Writing – review & editing. **Márcio Rocha Francelino:** Writing – review & editing. **Emilson Pereira Leite:** Writing – review & editing, Resources. **Danilo César de Mello:** Writing – original draft, Validation, Investigation, Formal analysis, Conceptualization. **Tiago Osório Ferreira:** Writing – review & editing. **Gustavo Vieira Veloso:** Validation, Software, Methodology, Data curation, Conceptualization. **Jose Alexandre Melo Demattê:** Writing – review & editing, Supervision, Resources, Project administration, Formal analysis. **Murilo Ferre Mello:** Writing – review & editing. **Marcos Guedes de Lana:** Software, Formal analysis. **Isabelle de Angeli Oliveira:** Writing – review & editing. **Fellipe Alcantara de Oliveira Mello:** Writing – review & editing.

## Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work, the authors utilized GPT-4 to correct the grammar and structure of the English language, ensuring clarity in the sentences for the reader. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank the following funding agencies and institutions that contributed and/or supported the development of this research: National Council for Scientific and Technological Development (CNPq) (grant No. 134608/2015–1); the São Paulo Research Foundation (FAPESP) (grant No. 2014–22262–0); Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) (Finance code 001); the Geotechnologies in Soil Science group (GeoSS – website http://esalqgeocis.wixsite.com/english) and LabGeo – UFV - 'Post-graduate Program in Soils and Plant Nutrition – PGSNP' of the Soil Department of the Federal University of Viçosa; the Institute of Geosciences at Campinas State University.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.soilad.2024.100024.

## Data availability

Data will be made available on request.

## References

Alvares, C.A., Stape, J.L., Sentelhas, P.C., De Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorol. Z. 22, 711–728. https://doi.org/10.1127/0941-2948/2013/0507.

Anic, I., Dragovic, S., 2005. Artificial neural network modelling of uncertainty in gamma-ray spectrometry 540, 455–463. ⟨https://doi.org/10.1016/j.nima.2004.11.045⟩.

Ayoubi, S., Abazari, P., Zeraatpisheh, M., 2018. Soil great groups discrimination using magnetic susceptibility technique in a semi-arid region, central Iran. Arab. J. Geosci. 11. https://doi.org/10.1007/s12517-018-3941-4.

Ayoubi, S., Adman, V., Yousefifard, M., 2019. Use of magnetic susceptibility to assess metals concentration in soils developed on a range of parent materials. Ecotoxicol. Environ. Saf. 168, 138–145. https://doi.org/10.1016/j.ecoenv.2018.10.024.

Bai, W., Kong, L., Guo, A., 2013. Effects of physical properties on electrical conductivity of compacted lateritic soil. J. Rock. Mech. Geotech. Eng. 5, 406–411. https://doi.org/10.1016/j.jrmge.2013.07.003.

Barbuena, D., de Souza Filho, C.R., Leite, E.P., Miguel Jr, E., de Assis, R.R., Xavier, R.P., Ferreira, F.J.F., Paes de Barros, A.J., 2013. Airborne geophysical data analysis applied to geological interpretation in the Alta Floresta Gold Province. Mt. Rev. Bras. GeofíSci.

Bazaglia Filho, O., Rizzo, R., Lepsch, I.F., Prado, H. do, Gomes, F.H., Mazza, J.A., Demattê, J.A.M., 2013a. Comparison between detailed digital and conventional soil maps of an area with complex geology. Rev. Bras. Ciência do Solo 37, 1136–1148. https://doi.org/10.1590/s0100-06832013000500003.

Bazaglia Filho, O., Rizzo, R., Lepsch, I.F., Prado, H. do, Gomes, F.H., Mazza, J.A., Demattê, J.A.M., 2013b. Comparison between detailed digital and conventional soil maps of an area with complex geology. Rev. Bras. Ciência do Solo 37, 1136–1148. https://doi.org/10.1590/s0100-06832013000500003.

Beamish, D., 2013. Gamma ray attenuation in the soils of Northern Ireland, with special reference to peat. J. Environ. Radioact. 115, 13–27. https://doi.org/10.1016/j.jenvrad.2012.05.031.

Blundell, A., Dearing, J.A., Boyle, J.F., Hannam, J.A., 2009. Controlling factors for the spatial variability of soil magnetic susceptibility across England and Wales. Earth-Sci. Rev. 95, 158–188. https://doi.org/10.1016/j.earscirev.2009.05.001.

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: the example of landslide susceptibility analysis with generalized additive models. Hamburg. Beiträge zur Phys. Geogr. und Landsch. ökologie 19, 410.
Brus, D.J., Knotters, M., Van Dooremolen, W.A., Van Kernebeek, P., Van Seeters, R.J.M., 1992. The use of electromagnetic measurements of apparent soil electrical conductivity to predict the boulder clay depth. Geoderma 55, 79–93.
Camargo, L.A., Marques Júnior, J., Pereira, G.T., Bahia, A.S.R. de S., 2014. Clay mineralogy and magnetic susceptibility of Oxisols in geomorphic surfaces. Sci. Agric. 71, 244–256. https://doi.org/10.1590/S0103-90162014000300010.
Cardoso, R., Dias, A.S., 2017. Study of the electrical resistivity of compacted kaolin based on water potential. Eng. Geol. 226, 1–11. https://doi.org/10.1016/j.enggeo.2017.04.007.
Correia, M.G., Leite, E.P., de Souza Filho, C.R., 2010. Comparação de métodos de estimativa de profundidades de fontes magnéticas utilizando dados aeromagnéticos da província mineral de Carajás, Pará. Braz. J. Geophys 28, 411–426.
Corwin, D.L., Lesch, S.M., Shouse, P.J., Soppe, R., Ayars, J.E., 2003. Identifying Soil Properties that Influence Cotton Yield Using Soil Sampling Directed by Apparent Soil Electrical Conductivity 352–364.
Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008.
Dearing, J.A., 1999. Environmental Magnetic Susceptibility. Using the Bartington MS2 system, Second. ed. Chi Publishing, Kenilworth, UK.
Dickson, B.L., Scott, K.M., 1997. Interpretation of aerial gamma-ray surveys - adding the geochemical factors. AGSO J. Aust. Geol. Geophys. 17, 187–200.
Dragovic, S., Onjia, A., 2007. Classification of soil samples according to geographic origin using gamma-ray spectrometry and pattern recognition methods. Appl. Radiat. Isot. 65, 218–224. https://doi.org/10.1016/j.apradiso.2006.07.005.
Gomes, L.C., Faria, R.M., Souza, E., De, Veloso, G.V., Ernesto, C., Schaefer, G.R., Inácio, E., Filho, F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007.
Grimley, D.A., Arruda, N.K., Bramstedt, M.W., 2004. Using magnetic susceptibility to facilitate more rapid, reproducible and precise delineation of hydric soils in the midwestern USA. Catena 58, 183–213. https://doi.org/10.1016/j.catena.2004.03.001.
Harvey, O.R., Morgan, C.L.S., 2009. Predicting regional-scale soil variability using a single calibrated apparent soil electrical conductivity model. Soil Sci. Soc. Am. J. 73, 164–169.
Heil, K., Schmidhalter, U., 2019. Theory and Guidelines for the Application of the Geophysical Sensor EM38 38.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748.
Hijmans, R.J., Van Etten, J., 2016. raster: Geographic Data Analysis and Modeling. R package version 2.5-8.
Honeyborne, I., McHugh, T.D., Kuittinen, I., Cichonska, A., Evangelopoulos, D., Ronacher, K., van Helden, P.D., Gillespie, S.H., Fernandez-Reyes, D., Walzl, G., Rousu, J., Butcher, P.D., Waddell, S.J., 2016. Profiling persistent tubercule bacilli from patient sputa during therapy predicts early drug efficacy. BMC Med 14, 1–13. https://doi.org/10.1186/s12916-016-0609-3.
Jiménez, C., Benavides, J., Ospina-Salazar, D.I., Zúñiga, O., Ochoa, O., Mosquera, C., 2017. Relationship between physical properties and the magnetic susceptibility in two soils of Valle del Cauca Relación entre propiedades físicas y la susceptibilidad magnética en dos suelos del Valle del Cauca. Cauca. Rev. Cienc. Agric. 34, 33–45. https://doi.org/10.22267/rcia.173402.70.
Johnston, M.A., Savage, M.J., Moolman, J.H., du Plessis, H.M., 1997. Evaluation of calibration methods for interpreting soil salinity from electromagnetic induction measurements. Soil Sci. Soc. Am. J. 61, 1627–1633. https://doi.org/10.2136/sssaj1997.03615995006100060013x.
Jordanova, N., 2016. Soil Magnetism: Applications in Pedology, Environmental Science and Agriculture. Academic Press.
Karimi, A., Haghnia, G.H., Ayoubi, S., Safari, T., 2017. Impacts of geology and land use on magnetic susceptibility and selected heavy metals in surface soils of Mashhad plain, northeastern Iran. J. Appl. Geophys. 138, 127–134. https://doi.org/10.1016/j.jappgeo.2017.01.022.
Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81, 401–418.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97, 273–324.
Kuhn, M., 2012. Variable selection using the caret package. URL ⟨http//cran.cermin.lipi.go.id/web/packages/caret/vignettes/caretSelection.pdf⟩.
Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., 2020. Package 'caret.' R J.
Lesch, S.M., Rhoades, J.D., Lund, L.J., Corwin, D.L., 1992. Mapping soil salinity using calibrated electromagnetic measurements. Soil Sci. Soc. Am. J. 56, 540–548.
Li, H., Wang, J., Wang, Q., Tian, C., Qian, X., Leng, X., 2017. Magnetic properties as a proxy for predicting fine-particle-bound heavy metals in a support vector machine approach. Environ. Sci. Technol. 51, 6927–6935. https://doi.org/10.1021/acs.est.7b00729.
McFadden, M., Scott, W.R., 2013. Broadband soil susceptibility measurements for EMI applications. J. Appl. Geophys. 90, 119–125. https://doi.org/10.1016/j.jappgeo.2013.01.009.
Mcneill, J.D., 1992. Rapid, accurate mapping of soil salinity by electromagnetic ground conductivity meters 2–3.
McNeill, J.D., 1986. Geonics EM38 ground conductivity meter. Tech. Note TN-21. Geonics Ltd., Mississauga, Ontario, Canada.
Mello, D.C. de, Alexandre Melo Demattê, J., Alcantara de Oliveira Mello, F., Roberto Poppiel, R., ElizabetQuiñonez Silvero, N., Lucas Safanelli, J., Barros e Souza, A., Augusto Di Loreto Di Raimo, L., Rizzo, R., Eduarda Bispo Resende, M., Ernesto Gonçalves Reynaud Schaefer, C., 2021. Applied gamma-ray spectrometry for evaluating tropical soil processes and attributes. Geoderma 381. https://doi.org/10.1016/j.geoderma.2020.114736.
Mello, D., Demattê, J.A.M., Silvero, N.E.Q., Di Raimo, L.A.D.L., Poppiel, R.R., Mello, F.A. O., Souza, A.B., Safanelli, J.L., Resende, M.E.B., Rizzo, R., 2020. Soil magnetic susceptibility and its relationship with naturally occurring processes and soil attributes in pedosphere, in a tropical environment. Geoderma 372, 114364. https://doi.org/10.1016/j.geoderma.2020.114364.
Mello, D.C. de, Ferreira, T.O., Veloso, G.V., de Lana, M.G., de Oliveira Mello, F.A., Di Raimo, L.A.D.L., Cabrero, D.R.O., de Souza, J.J.L.L., Fernandes-Filho, E.I., Francelino, M.R., 2023a. Digital mapping of soil weathering using field geophysical sensor data coupled with covariates and machine learning. J. South Am. Earth Sci., 104449
Mello, D.C. de, Veloso, G.V., Lana, M.G. de, Mello, F.A. de O., Poppiel, R.R., Cabrero, D. R.O., Di Raimo, L.A.D.L., Schaefer, C.E.G.R., Leite, E.P., Demattê, J.A.M., 2022. A new methodological framework for geophysical sensor combinations associated with machine learning algorithms to understand soil attributes. Geosci. Model Dev. 15, 1219–1246.
Mello, D.C. de, Vieira, G., Marques, C., Angeli, I.De, Soares, F., Oliveira, D., Demattê, M., 2023b. Chemical weathering detection in the periglacial landscapes of Maritime Antarctica: New approach using geophysical sensors, topographic variables and machine learning algorithms. Geoderma 438. https://doi.org/10.1016/j.geoderma.2023.116615.
Minty, B.R.S., 1988. A Review of Airborne Gamma-Ray Spectrometric Data-Processing Techniques. Aust. Gov. Publ. Serv.
Mullins, C.E., 1977. Magnetic susceptibility of the soil and its significance in soil science–a review. J. Soil Sci. 28, 223–246.
Nanni, M.R., Demattе, J.A.M.P.P.-P., 2000. Dados radiométricos obtidos em laboratório e no nível orbital na caracterização e mapeamento de solos.
Nanni, M.R., Demattê, J.A.M., 2006. Spectral reflectance methodology in comparison to traditional soil analysis. Soil Sci. Soc. Am. J. 70, 393–407. https://doi.org/10.2136/sssaj2003.0285.
Paes, É. de C., Veloso, G.V., Fonseca, A.A. da, Fernandes-Filho, E.I., Fontes, M.P.F., Soares, E.M.B., 2022. Predictive modeling of contents of potentially toxic elements using morphometric data, proximal sensing, and chemical and physical properties of soils under mining influence. Sci. Total Environ. 817, 152972. https://doi.org/10.1016/j.scitotenv.2022.152972.
Parshin, A.V., Morozov, V.A., Blinov, A.V., Kosterev, A.N., Budyak, A.E., 2018. Low-altitude geophysical magnetic prospecting based on multirotor UAV as a promising replacement for traditional ground survey. Geo-Spat. Inf. Sci. 21, 67–74.
Priori, S., Fantappiè, M., Bianconi, N., Ferrigno, G., Pellegrini, S., Costantini, E.A.C., 2016. Field-scale mapping of soil carbon stock with limited sampling by coupling gamma-ray and vis-NIR spectroscopy. Soil Sci. Soc. Am. J. 80, 954–964. https://doi.org/10.2136/sssaj2016.01.0018.
RC Team, 2021. R: A language and environment for statistical computing.(Version 4.1.0).
Reinhardt, N., Herrmann, L., 2019. Gamma-ray spectrometry as versatile tool in soil science: a critical review. J. Plant Nutr. Soil Sci. 182, 9–27. https://doi.org/10.1002/jpln.201700447.
Rhoades, J.D., Chanduvi, F., Lesch, S.M., 1999. Soil salinity assessment: methods and interpretation of electrical conductivity measurements. Food Agric. Org.
Richards, L.A., 1954. Diagnosis and improvement of saline and alkali soils. LWW.
Rochette, P., Jackson, M., Aubourg, C., 1992. Rock magnetism andn the interpretation of magnetic susceptibility. Rev. Geophys. 30, 209–226.
Rytky, S.J.O., Tiulpin, A., Frondelius, T., Finnilä, M.A.J., Karhula, S.S., Leino, J., Pritzker, K.P.H., Valkealahti, M., Lehenkari, P., Joukainen, A., Kröger, H., Nieminen, H.J., Saarakkala, S., 2020. Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography. Osteoarthr. Cartil. 28, 1133–1144. https://doi.org/10.1016/j.joca.2020.05.002.
Sales, S. and C, 2021. Terraplus KT-10 v2 User Manual.
Sarmast, M., Farpoor, M.H., Esfandiarpour Boroujeni, I., 2017. Magnetic susceptibility of soils along a lithotoposequence in southeast Iran. Catena 156, 252–262. https://doi.org/10.1016/j.catena.2017.04.019.
Schuler, U., Erbe, P., Zarei, M., Rangubpit, W., Surinkum, A., Stahr, K., Herrmann, L., 2011. A gamma-ray spectrometry approach to field separation of illuviation-type WRB reference soil groups in northern Thailand. J. Plant Nutr. Soil Sci. 174, 536–544. https://doi.org/10.1002/jpln.200800323.
Shenggao, L., 2000. Lithological factors affecting magnetic susceptibility of subtropical soils, Zhejiang Province, China. Catena 40, 359–373. https://doi.org/10.1016/S0341-8162(00)00092-8.
Siqueira, D.S., Marques, J., Matias, S.S.R., Barrón, V., Torrent, J., Baffa, O., Oliveira, L.C., 2010. Correlation of properties of Brazilian Haplustalfs with magnetic susceptibility measurements. Soil Use Manag 26, 425–431. https://doi.org/10.1111/j.1475-2743.2010.00294.x.
Solutions, R., 2009. Spectrum stabilization and calibration for the RSI RS-125 and RS-230 handheld spectrometers.
Sousa, I., Costa, L., Cavalcanti, I., Oliveira, C.De, Tavares, F.M., José, H., Polo, D.O., Sousa, I., Costa, L., Cavalcanti, I., Oliveira, C.De, 2020. Uranium anomalies detection

through Random Forest regression Uranium anomalies detection through Random Forest regression. ⟨https://doi.org/10.1080/08123985.2020.1725387⟩.
Siqueira, R.G., Moquedace, C.M., Fernandes-Filho, E.I., Schaefer, C.E.G.R., Francelino, M.R., Sacramento, I.F., Michel, R.F.M., 2024. Modelling and prediction of major soil chemical properties with Random Forest: Machine learning as tool to understand soil-environment relationships in Antarctica. CATENA 235, 107677.
Taylor, M.J., Smettem, K., Pracilio, G., Verboom, W., 2002. Relationships between soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia. Explor. Geophys. 33, 95–102. https://doi.org/10.1071/EG02095.
Teixeira, D.D.B., Marques, J., Siqueira, D.S., Vasconcelos, V., Carvalho, O.A., Martins, É. S., Pereira, G.T., 2018. Mapping units based on spatial uncertainty of magnetic susceptibility and clay content. Catena 164, 79–87. https://doi.org/10.1016/j.catena.2017.12.038.
Valaee, M., Ayoubi, S., Khormali, F., Lu, S.G., Karimzadeh, H.R., 2016. Using magnetic susceptibility to discriminate between soil moisture regimes in selected loess and loess-like soils in northern Iran. J. Appl. Geophys. 127, 23–30. https://doi.org/10.1016/j.jappgeo.2016.02.006.
Viana, J.H.M., Couceiro, P.R.C., Pereira, M.C., Fabris, J.D., Fernandes Filho, E.I., Schaefer, C., Rechenberg, H.R., Abrahão, W.A.P., Mantovani, E.C., 2006. Occurrence of magnetite in the sand fraction of an Oxisol in the Brazilian savanna ecosystem, developed from a magnetite-free lithology. Soil Res 44, 71–83.
Viscarra Rossel, R.A., Webster, R., Kidd, D., 2014. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf. Process. Landf. 39, 735–748. https://doi.org/10.1002/esp.3476.
Weller, U., Zipprich, M., Sommer, M., Castell, W.Z., Wehrhan, M., 2007. Mapping clay content across boundaries at the landscape scale with electromagnetic induction. Soil Sci. Soc. Am. J. 71, 1740–1747.
Wilford, P.N., Bierwirth, J.R., Craig, M.A., 1997. Application of airborne gamma-ray spectrometry in soiVregolith mapping and Applied Geomorphology 17.
Wilford, J., Minty, B., 2006. Chapter 16 the use of airborne gamma-ray imagery for mapping soils and understanding landscape processes. Dev. Soil Sci. 31. https://doi.org/10.1016/S0166-2481(06)31016-1.
Wilford, J., Thomas, M., 2012. Modelling soil-regolith thickness in complex weathered landscapes of the central Mt Lofty Ranges, South Australia.
Wong, M.T.F., Harper, R.J., 1999. Use of on-ground gamma-ray spectrometry to measure plant-available potassium and other topsoil attributes. Aust. J. Soil Res 37, 267–277. https://doi.org/10.1071/S98038.
Zare, E., Li, N., Khongnawang, T., Farzamian, M., 2020. Identifying Potential Leakage Zones in an Irrigation Supply Channel by Mapping Soil Properties Using Electromagnetic Induction, Inversion Modelling and a Support Vector Machine.