

# Mapping soil drainage classes: Comparing expert knowledge and machine learning strategies

Danilo César de Mello<sup>a</sup>, Nélida E.Q. Silvero<sup>a</sup>, Bradley A. Miller<sup>b</sup>, Nicolas Augusto Rosin<sup>a</sup>, Jorge Tadeu Fim Rosas<sup>a</sup>, Bruno dos Anjos Bartsch<sup>a</sup>, Gustavo Vieira Veloso<sup>a</sup>, Jean Jesus Macedo Novais<sup>a</sup>, Renan Falcioni<sup>c</sup>, Marcos Rafael Nanni<sup>c</sup>, Marcelo Rodrigo Alves<sup>d</sup>, Elpídio Inácio Fernandes-Filho<sup>e</sup>, Uemeson José dos Santos<sup>f</sup>, José Alexandre Melo Demattê<sup>a,\*</sup>

<sup>a</sup> Department of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, Piracicaba, São Paulo 13418-260, Brazil

<sup>b</sup> Department of Agronomy, Iowa State University, Ames, IA, USA

<sup>c</sup> Department of Agronomy, State University of Maringá, Av. Colombo, 5790, Maringá, Paraná 87020-900, Brazil

<sup>d</sup> Department of Soil Science, University of Western Paulista, Rod. Raposo Tavares, km 572 - Limoeiro, Presidente Prudente, São Paulo, Brazil

<sup>e</sup> Department of Soil and Plant Nutrition, Federal University of Viçosa, campus UFV, Viçosa 36570-900, Brazil

<sup>f</sup> Federal Institute of Education, Science, and Technology of Pará, Campus Santarém, Av. Mal. Castelo Branco, 621, Interventoria, Santarém, PA 68020-820, Brazil

## ARTICLE INFO

### Keywords:

Environmental covariates  
Ferralsol  
Predictive modeling  
Remote sensing  
Soil color

## ABSTRACT

Soil drainage is an essential factor that influences plant growth and various biophysical processes, such as nutrient cycling and greenhouse gas fluxes. Therefore, soil drainage maps are fundamental tools for managing crops, forests, and the environment. This study compared two approaches to mapping soil drainage classes in the state of São Paulo, Brazil, using geographic information systems (GIS). The first approach employed expert knowledge (EK) to develop a simple model based on soil color and texture, while the second used machine learning (ML) with an extensive set of covariates and a decision tree algorithm. To evaluate the full, operational implementation of soil mapping, this study assessed the two approaches in terms of accuracy, labor efficiency, transferability, interpretability, and agreement/disagreement statistical methods. In terms of accuracy, the ML-based strategy showed greater agreement with the reference map (53 %) compared to the EK approach (50 %). However, the EK strategy was more time- and resource-efficient, as well as being more transferable and interpretable due to the simplicity of its rules based on soil properties. Given its higher interpretability and ease of application, the EK approach was recommended as the most suitable for operational soil drainage mapping in tropical environments.

## 1. Introduction

Demand for detailed soil information has increased over the years as a means of supporting not only land management for agriculture but also the roles of soil in ecosystems (Brevik et al., 2016). The gap between this demand and the state of legacy soil maps is becoming even greater in developing countries such as Brazil. Here, most available soil maps are incomplete or insufficient to support public policies and other applications involving soil resources (Santos et al., 2015).

An important characteristic lacking in most Brazilian maps is the soil drainage class, defined as the frequency and duration of wet periods or

degree and frequency the soil matrix is free of water saturation (Soil Science Division Staff, 2017). A soil drainage class is used as an indicator of the general conditions of water movement in the soil, which is primarily influenced by climate and regulated by soil texture, structure, and topography (Troeh, 1964; Gerardin and Duerue, 1990; Fausey, 2005). Information on soil drainage becomes essential because it controls several processes in the soil, such as the decomposition of soil organic matter (Wickland et al., 2010), distribution of soil organic carbon (Raymond et al., 2012), and microbial adaptations (Graça et al., 2021). Subsequently, soil drainage informs management decisions, such as irrigation rate, cow stocking rates, dry matter production (Fitzgerald

\* Corresponding author.

E-mail addresses: [n.silvero@leeds.ac.uk](mailto:n.silvero@leeds.ac.uk) (N.E.Q. Silvero), [millerba@iastate.edu](mailto:millerba@iastate.edu) (B.A. Miller), [narosin@usp.br](mailto:narosin@usp.br) (N.A. Rosin), [jorge.fimrosas@usp.br](mailto:jorge.fimrosas@usp.br) (J.T.F. Rosas), [brunobartsch@usp.br](mailto:brunobartsch@usp.br) (B.A. Bartsch), [jeannovais@usp.br](mailto:jeannovais@usp.br) (J.J.M. Novais), [mrnanni@uem.br](mailto:mrnanni@uem.br) (M.R. Nanni), [marceloalves@unoeste.br](mailto:marceloalves@unoeste.br) (M.R. Alves), [elpidio@ufv.br](mailto:elpidio@ufv.br) (E.I. Fernandes-Filho), [uemeson.santos@ifpa.edu.br](mailto:uemeson.santos@ifpa.edu.br) (U.J. Santos), [jamdemat@usp.br](mailto:jamdemat@usp.br) (J.A.M. Demattê).

<https://doi.org/10.1016/j.soilad.2024.100028>

Received 1 November 2024; Received in revised form 14 December 2024; Accepted 17 December 2024

Available online 22 December 2024

2950-2896/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

et al., 2008) and which plant species are more suitable for a specific area (Levine et al., 1994).

The description of soil drainage is usually performed in terms of seven discrete units or classes varying from very well to poorly drained, depending on the proportion and rate of water infiltration. Rather than directly observing the long-term hydrologic conditions of a site, drainage class tends to be interpreted from soil morphological features such as soil color, texture class, and the presence of redoximorphic features that are observed when soil profiles are described in the field. Soil color and redoximorphic features reflect processes that take place in the soil due to the duration of water saturation, specifically oxidation and reduction of iron. Well aerated and drained soils usually present reddish, brownish, and yellowish colors, which are characteristic of the presence of iron in the oxidized form. In these soils, water is removed rapidly, especially if coarse-textured soils are present. In poorly drained soils, however, gray and dark colors reflect the presence of organic matter and silicate minerals, and altered forms of iron that have been reduced or removed from the system due to the presence of water during extended periods (Franzmeier et al., 2001).

On the other hand, the depth to the water table can be more useful than morphological characteristics for describing soil drainage classes (Mackintosh and Van Dern Hulst, 1979). This concept relates to soil sequences along a slope, known as catena, and is grounded in soil-landscape relationships. Hillslope positions—summit, shoulder, backslope, footslope, and toeslope—are crucial in governing water movement (Schaetzl et al., 2022). Well-drained soils with deep water tables are typically on uplands and upper backslopes, while poorly drained soils with higher water tables are usually at the slope's bottom.

Evaluating landscapes and allocating land areas into drainage classes is a complex task that relies on spatial predictions of observable surface variables. Traditionally, creating soil drainage class maps involves two steps: 1) field observations and allocation of drainage classes and 2) applying knowledge from observed soil profiles to areas where direct observation is impractical. In the second step, pedologists use their mental models, that is, their perceptions and knowledge acquired over time from prior experiences, scientific understanding, and technical intuition, in addition to the understanding of soil-landscape relationships. Previous studies have relied on expert knowledge, morphological characteristics, and the experiences of pedologists (Hudson, 1992). For example, Gerardin and Duerue (1990) proposed a method for assessing natural soil drainage in forest soils to assist non-specialists. Their approach, which involved measuring various geomorphological, topographical, edaphic, and vegetation variables, showed promise but was site-specific, limiting its broader applicability.

Traditional techniques have produced valuable soil maps for over a century; however, reliance on tacit knowledge can lead to inconsistencies among experts (Bouma, 1973; Kerebel and Holden, 2013). To address this issue, quantitative frameworks for mapping soil drainage classes have emerged, utilizing mathematical calculations and statistical tools to describe the landscape's relationship with drainage classes. For instance, Troeh (1964) developed 3D equations to derive landscape parameters related to soil drainage, establishing a strong correlation. Bell et al. (1992) proposed a "soil-landscape model" using multivariate discriminant analysis to identify variables that optimally separate soil drainage classes, achieving a 74 % agreement with field survey data.

Currently, digital soil mapping is the primary method for spatially predicting soil drainage classes. Common covariates include optical and radar remote sensing images (Cialella et al., 1997; Peng et al., 2003; Niang et al., 2012), terrain derivatives (Zhao et al., 2013), electrical conductivity (Kravchenko et al., 2002), magnetic susceptibility (Grimley et al., 2008; Asgari et al., 2018) and soil color (Malone et al., 2018). Despite its advances, field descriptions of soil drainage classes are still essential before employing digital mapping techniques. Most studies relied on field descriptions or legacy soil maps. An exception is Malone et al. (2018), who defined soil drainage classes based on soil color groups. They predicted these classes using terrain attributes and

machine learning, marking the first quantitative attempt to define soil drainage classes with information on soil color from legacy data.

In this research, we compared two strategies for mapping soil drainage classes in Brazil. The first was based on expert knowledge (EK), in which maps of soil color and texture were used to predict the soil drainage class. The second strategy was based on machine learning (ML), in which soil sampling points with the labeled soil drainage class were used in conjunction with environmental covariates to spatially predict the distribution of soil drainage classes in the study area. An existing soil drainage class map was used as a reference. Finally, we compared and selected the most appropriate strategy between EK and ML based on their best performance, taking into account four criteria (accuracy, labor efficiency, transferability, and interpretability), and agreement and disagreement accuracy statistical method.

## 2. Materials and methods

### 2.1. Study site description

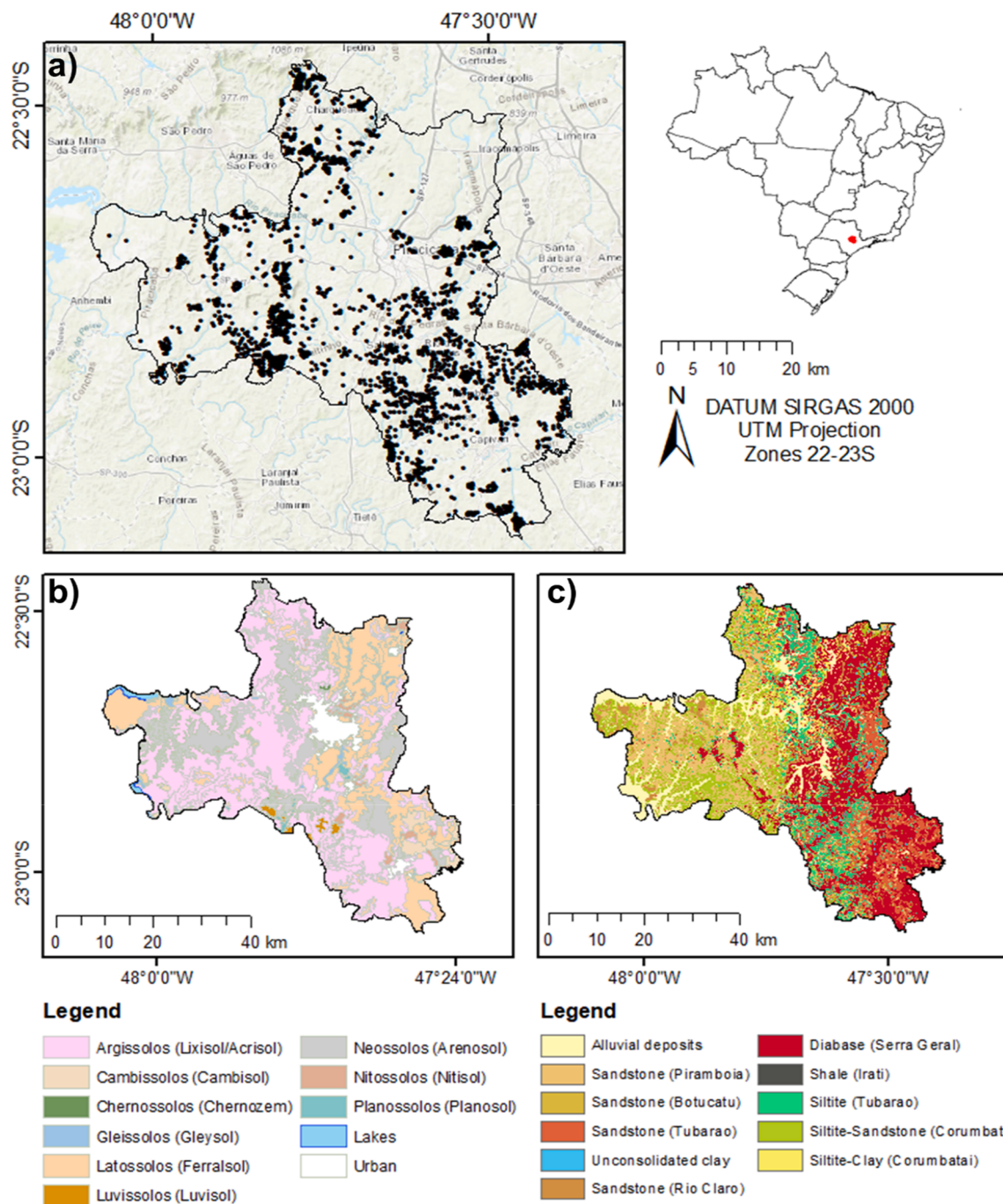
The study area is in southeastern Brazil, in the state of São Paulo, covering an area of approximately 2574 km<sup>2</sup> (Fig. 1). The study area is part of the Rio Paraná Basin, where 2309 soil samples were collected from various collection points (Fig. 1a). It is characterized by gentle slopes, undulating hills, and rolling uplands, with elevation varying between 450 and 950 m.a.s.l. The region's climate is classified as Cwa in Köppen's climatic classification, which is characterized by dry winters and hot summers with mean annual precipitation of 1200 mm and a mean annual temperature of 24 °C (Alvares et al., 2013). In the study area, 83 % of the land is used for agriculture and pasture, 9 % for forest, savanna, and forest plantations, and the remaining corresponds to urban areas and water. More than 40 % of the agricultural land is cultivated with sugarcane, almost 15 % with pasture, 22 % with a mixture of agriculture and pasture, and 3 % with soybean and other perennial and temporary crops (Souza et al., 2020).

The main soil classes in the area according World Reference Base for Soil Resources (WRB, 2014), with the corresponding Brazilian Soil Classification System in parenthesis, are Acrisols (Argissolos), Ferralsol (Latossolos), Arenosols (Neossolos), Cambisol (Cambissolos) and Nitisol (Nitossolos) (Fig. 1b) (Rossi, 2017). In the study area, three Acrisols (PA, PVA, and PV) are found, the main differences of which are related to soil color. PA (Yellow Acrisol) has a hue of 7.5YR in the first 100 cm of the B horizon, PV (Red Acrisol) has a hue of 2.5YR, and PVA (Red Yellow Acrisol) represents a gradient between PA and PV with a hue varying from red to yellow. Four Ferralsols can be found in the study area: LH (Humic Ferralsol), LA (Yellow Ferralsol), LV (Red Ferralsol), and LVA (Red Yellow Acrisol Ferralsol), whose differences in classification are also due to soil color, except for LH, which has higher organic matter content than the others. Two suborders of Arenosols are found in the study area: RQ (Quartzarenic Arenosols), a sandy soil, and RL (Lithic Arenosols), which has a lithic contact in the first 50 cm. Finally, the Nitisol, being, Red Nitisol (NVf) with high presence of iron oxides are common and the Haplic Nitisol (NX), which is a soil class that does not fit in the other categories of this soil order.

The lithological formation is complex and consists of diverse sedimentary rocks, including sandstone, siltstone, shale, unconsolidated clay, and alluvial deposits (Fig. 1c). The spatial arrangement of these lithological formations is also depicted in Fig. 1c. For a detailed description of the area's geology, the readers are referred to Bonfatti et al. (2020).

### 2.2. Strategies for mapping soil drainage classes

The strategies evaluated in this study were based on expert knowledge (EK) and machine learning (ML). We used an existing soil drainage class map as a reference for accessing accuracy. The more appropriate mapping strategy was selected based on a set of criteria related to the



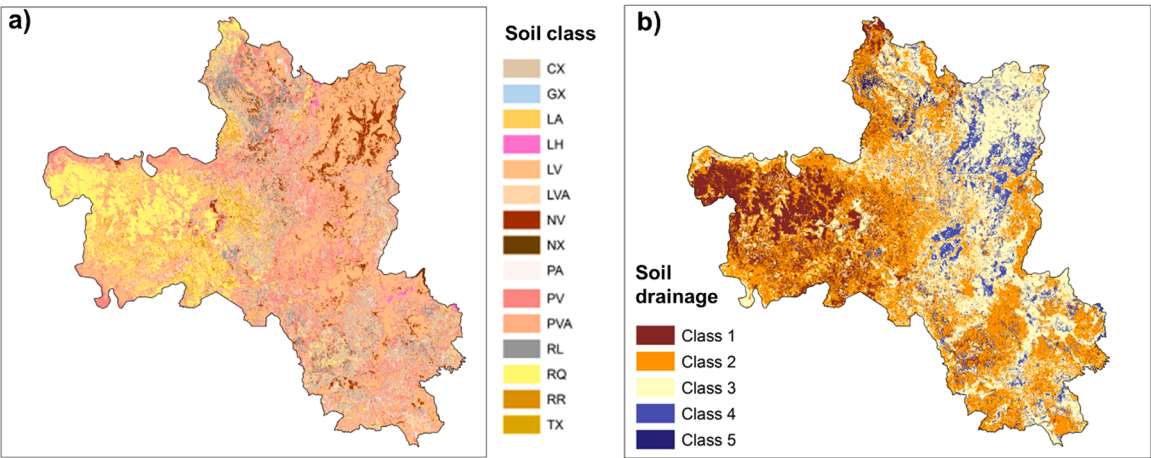
**Fig. 1.** Study area with a) sampling point locations, b) legacy soil class map at 1:250,000 obtained from Rossi (2017), with soil classes named according to the Brazilian Soil Classification System and the corresponding WRB in parenthesis, and c) Lithological map with a spatial resolution of 30 m, obtained from Bonfatti et al. (2020). The map unit names in parenthesis correspond to the main geological formations according to the Brazilian classification (IGG, 1996). Neossolos in this map represents the Neossolos Quartzarênicos.

implementation and results of the mapping procedure. A thorough description of each step is given in the following sections.

The reference drainage map used for accuracy assessment was generated based on soil classes previously mapped by Mello et al. (2021) (Fig. 2). A conventional soil class map was created at a scale of 1:20,000, incorporating environmental variables such as drainage and relief attributes, drainage classes, and satellite images to extend the soil map over a larger area. A point grid with a resolution of  $30 \times 30$  m was established to refine the map further to extract soil and variable information. Subsequently, these data were used to calibrate a random forest model, with cross-validation employed to fine-tune model selection and

optimize performance. Each soil class in the map by Mello et al. (2021) has a corresponding drainage class (Table S1), assigned by an expert using descriptions from other soil maps available for the country and knowledge of soil-landscape relationships for each soil class. For simplicity, we used only five classes to represent drainage, ranging from very well drained (Class 1) to poorly drained (Class 5). The reference soil drainage map (Fig. 2b) shows that the study area is predominantly (39 %) Class 2 (well drained), followed by Class 3 (moderately drained), which accounts for 37 % of the area. Drainage classes 1, 4, and 5 represent 13 %, 9 %, and 2 % of the area, respectively.

It is important to realize that the soil class map in Fig. 1 is based on



**Fig. 2.** Soil class map (a) and the resulting soil drainage class map (b) used as reference, with drainage classes varying from well drained (Class 1) to poorly drained (Class 5). Cambisols (CX), Gleisols (GX), Ferralsols (LH, LA, LV, LVA), Nitisols (NV, NX), Acrisol/Lixisol (PA, PV, PVA), Arenosol/Regosol (RL, RQ, RR), and Planosol (TX).

broader lithological and soil type data, designed to give an overview of the area’s soil types and their formation factors, particularly lithological substrates that also influence drainage intensity. On the other hand, the soil map in Fig. 2 is derived from a more detailed dataset specifically tailored to analyze the relationship between soil type and drainage class.

2.2.1. Soil drainage mapping by expert knowledge (EK)

The strategy based on EK used soil color and textural classes as base maps and applied an expert’s mental model to determine how those covariates indicated drainage class. Soil color variables (hue, value, and

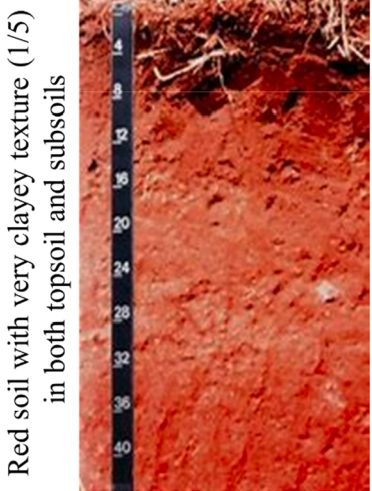
chroma) along with texture class maps were obtained from Mendes et al. (2021), who predicted these properties for two layers (0–20 cm and 80–100 cm) (Figure S1). These maps were created based on data from 1500 observations acquired through various soil surveys conducted by the Geotechnologies by Soil Science Group (GEOCIS, <https://esalqgeocis.wixsite.com/english>) between 2010 and 2019.

The 0–20 cm layer map was used as the topsoil and the 80–100 layers were averaged to represent the subsoil. We addressed the two primary soil layers that significantly influence soil composition, which we have designated as the ‘surface’ layer (0–20 cm) and the ‘undersurface’ layer

| a) | Color  | Code | Textural class                                        | Code | Drainage description          | Class   |
|----|--------|------|-------------------------------------------------------|------|-------------------------------|---------|
|    | Red    | 1    | Very sandy (Clay content <150 g kg <sup>-1</sup> )    | 1    | Very well drained             | Class 1 |
|    | Brown  | 2    | Sandy (Clay content <250 g kg <sup>-1</sup> )         | 2    | Well drained                  | Class 2 |
|    | Yellow | 3    | Loam (250 > Clay content <350 g kg <sup>-1</sup> )    | 3    | Moderately drained            | Class 3 |
|    | Gray   | 4    | Clayey (350 > Clay content <=600 g kg <sup>-1</sup> ) | 4    | Imperfectly to poorly drained | Class 4 |
|    | Black  | 5    | Very clayey (Clay content > 600 g kg <sup>-1</sup> )  | 5    | Very poorly drained           | Class 5 |

| b) | C/T topsoil | C/T subsoil | Drainage Topsoil | Drainage Subsoil | Final drainage |
|----|-------------|-------------|------------------|------------------|----------------|
|    | 1/5         | 1/5         | Class 3          | Class 3          | Class 3        |

C/T: Color/texture code  
Code 1/5: red color and very clayey soil  
Class 3 (drainage): moderately drained



**Fig. 3.** Measurement in soil classes using a) A table for assigning codes based on soil color and texture class, which were then interpreted by expert knowledge (EK) to predict drainage class. The final drainage class was determined by an expert after evaluating the relationships between the topsoil and subsoil; and b) An example of assigning drainage class based on the color and texture of the topsoil and subsoil. The soil drainage class was assigned based on the combination of color/texture (CT) codes from the topsoil and subsoil layers.

(80–100 cm). The soil color maps were classified into five groups (red, brown, yellow, gray, and black) to represent the sequence of five soil drainage classes, varying from very well to poorly drained, similar to the approach presented in [Malone et al. \(2018\)](#). The data for the mentioned attributes can be viewed in [Figure S1](#) and [Table S1](#) in the [supplementary material](#).

The classified soil color and textural maps were combined in ArcGIS Pro using the “combine” tool to join the maps and define a numerical code to represent the color and textural groups ([Fig. 3](#)). For example, red soil with a very clayey texture had the code 1/5 while a gray soil with a very clayey texture had the code 4/5. With this procedure, 24 color/texture codes for each layer were obtained (topsoil and subsoil, [Table S3](#) in the [supplementary material](#)). After this procedure, a drainage class was assigned to each color/texture code by an expert pedologist for each layer. Both layers were then considered to define the final drainage class. The theory and experience guiding the assignment of drainage class by an expert pedologist to each color/texture code, and the final drainage class through the consideration of the two layers, were based on a threefold approach: i) a comprehensive examination of the relevant literature concerning the relationship between soil color and drainage, ii) three decades of hands-on experience in pedology involving soil profile investigations and field surveys, and iii) developing an adaptation of the international system of drainage classes (drainage classes recognized by the Natural Resource Conservation Service - NRCS) ([Schaeztl, 2013](#)) tailored to Brazilian conditions and soils, while taking into account the nuances of the Brazilian Soil Classification System (SIBICS, 2018). More details can be found in [Tables S2 and S3](#) of the [supplementary materials](#).

## 2.2.2. Soil drainage mapping by machine learning (ML)

The ML strategy used 2309 locations observed by the Geotechnologies in Soil Science Group between 2000 and 2020. Each sample point had its respective soil drainage class confirmed in the field. The stack of covariates supporting the ML was derived from elevation, remotely sensed, and previous soil maps. These spatially exhaustive variables and sampling points labeled for soil drainage class were analyzed by a decision tree algorithm (C5.0) to build a model for predicting soil drainage classes at unseen locations. Unlike EK, the ML-based strategy does not rely on using only two soil variables (color and texture) but instead a set of variables that seek to represent processes related to the distribution of soil drainage classes in the landscape.

**2.2.2.1. Environmental covariates.** A total of 603 environmental covariates were considered in a digital soil mapping framework. All terrain derivatives were calculated using SAGA GIS, GRASS GIS, and ArcGIS Pro. The majority (552) of these derivatives were obtained from a digital elevation model (DEM) with a resolution of 5 m. These derivatives are scale-dependent and were calculated in a GIS using a moving window, typically  $3 \times 3$  cells; however, we considered analysis scales ranging from 15 to 5070 m. Eight terrain derivatives, including slope gradient, northness, mid-slope position, and curvatures, resulted in a total of 69 raster layers for each derivative. Additionally, derivatives that do not use moving windows were included, such as vertical distance to the channel network (VDCN) and the topographic wetness index (TWI). Remote sensing products from Sentinel-2A (S2A) and Landsat were also incorporated. For the S2A, images from the dry season (April to October) and the moist season (October to April) from 2015 to 2020 were processed on the Google Earth Engine platform. The Normalized Difference Vegetation Index (NDVI) and the Synthetic Soil Image (SYSI) were also obtained following the methodology proposed by [Demattê et al. \(2018\)](#).

Not all environmental covariates were used for spatial modelling. A Spearman rank correlation analysis was performed to select those that had the highest correlation with the drainage class. Considering the raster layers that were obtained at different analysis scales, only the one

with the highest correlation was selected. For example, among the 69 raster layers of slope gradient at different analysis scales, the slope gradient at 115 m (slp\_115m, window size of  $15 \times 15$  cells) was chosen. The same was performed on all-terrain derivatives with multiple analysis scales, satellite images, and the SYSI. The final covariate stack consisted of 23 raster layers. The environmental covariates that were selected and used in the modelling process are shown in [Table 1](#).

**2.2.2.2. Building and validating ML model.** The sample points were randomly divided into two sets: 70 % of the data (1616 samples) was utilized for model building, while the remaining 30 % (693 samples) served for independent validation. The predictive model was implemented in the R software using the C5.0, caret, and raster packages ([Hijmans; Van Etten, 2016](#); [Kuhn, 2020](#); [Su et al., 2021](#)) a decision tree approach derived from the C4.5 and ID3 algorithms created by [Quinlan \(1986\)](#). This supervised machine learning algorithm builds a classification model based on training data with known classes to identify the best rules for splitting the data. Validation involved predicting drainage classes for each sample using a test dataset. Model accuracy, defined as the number of correct predictions divided by the total predictions, was assessed using a confusion matrix. Additionally, the kappa index, user's accuracy (UA), and producer's accuracy (PA) were calculated. PA reflects the map maker's accuracy, while UA indicates the reliability from the map user's perspective. The Shapley value statistic was employed to

**Table 1**

Environmental covariates used for spatially modelling soil drainage classes with their corresponding abbreviation. The analysis scale corresponds to the different moving windows used to represent some terrain derivatives at different spatial scales. The last column ('Selected') represents the environmental covariates that were selected after analyzing their relationship with soil drainage.

| Environmental covariates                           | nLayers | Abbreviation            | Analysis scales | Selected        |
|----------------------------------------------------|---------|-------------------------|-----------------|-----------------|
| Slope gradient                                     | 69      | slp                     | 15–5070 m       | slp_075m        |
| Profile curvature                                  | 69      | prc                     | 15–5070 m       | prc_115m        |
| Plan curvature                                     | 69      | plc                     | 15–5070 m       | plc_1110m       |
| Northness                                          | 69      | nnes                    | 15–5070 m       | nnes_3630m      |
| Eastness                                           | 69      | enes                    | 15–5070 m       | enes_3630m      |
| Cross-sectional curvature                          | 69      | ccurv                   | 15–5070 m       | ccurv_115m      |
| Longitudinal curvature                             | 69      | lcurv                   | 15–5070 m       | lcurv_115m      |
| Relative elevation                                 | 69      | rel                     | 15–5070 m       | rel_3090m       |
| Topographic wetness index                          | 1       | TWI                     | NA              | YES             |
| Mid-slope position                                 | 1       | mslp                    | NA              | YES             |
| Vertical distance to channel network               | 1       | VDCN                    | NA              | YES             |
| Multi-resolution valley bottom flatness            | 1       | MRVBF                   | NA              | YES             |
| Soil color subsoil and topsoil                     | 6       | (Hue, value and chroma) | NA              | All             |
| Sentinel–2A bands - dry season                     | 9       | S2A dry                 | NA              | S2A_dry_swir1   |
| Sentinel–2A bands - moist season                   | 9       | S2A moist               | NA              | S2A_moist_swir1 |
| Sentinel–2A Normalized Difference Vegetation Index | 2       | S2A NDVI dry            | NA              | S2A_NDVI_moist  |
| Sentinel–2A Normalized Difference Vegetation Index | 2       | S2A NDVI moist          | NA              | S2A_NDVI_dry    |
| Sentinel–2A Normalized Difference Red-Edge Index   | 2       | S2A NDRE dry            | NA              | S2A_NDRE_moist  |
| Sentinel–2A Normalized Difference Red-Edge Index   | 2       | S2A NDRE moist          | NA              | S2A_NDRE_dry    |
| Sentinel–2A Synthetic Soil Image                   | 6       | SYSI                    | NA              | SYSI_swir2      |

assess the contributions of environmental covariates to drainage class predictions.

### 2.3. Comparison between reference soil drainage map and those from EK and ML strategies

The soil drainage maps obtained from the EK and ML strategies were compared with the reference map in two ways. First, a cell-by-cell comparison was conducted by subtracting the EK and ML-based maps from the reference map to assess their agreement. The resulting difference maps showed positive, negative, and zero values, indicating the discrepancies between the EK and ML drainage maps relative to the reference data.

Next, the soil drainage class at each sampling location shown in Fig. 1 was extracted from the three maps, and two confusion matrices were created. During this process, it was noted that Class 5 was not present among the values extracted from the ML-based soil drainage map, leading to its exclusion from the comparison and resulting in only four classes being presented in the confusion matrices. Accuracy, kappa values, along with user accuracy (UA) and producer accuracy (PA), were calculated from the confusion matrices.

Additionally, we employed the agreement accuracy and disagreement accuracy methods to evaluate the predictive performance of the ML and EK strategies (Pontius Junior and Millones, 2011). The R software (R Core Team, 2023) was used to calculate these performance parameters. The terms "agreement accuracy" and "disagreement accuracy" are commonly used in predictive model evaluation and are part of model evaluation metrics. Agreement accuracy measures the percentage of agreement between predicted and observed maps, while disagreement accuracy gauges the percentage of correct predictions when the model disagrees with the observed map (Pontius Junior and Millones, 2011).

### 2.4. Selecting the most appropriate strategy for mapping soil drainage classes

Based on the appropriate operational strategy for mapping soil drainage classes in tropical environments, each map was evaluated using four key criteria for the overall soil mapping process (accuracy, labor-efficiency, transferability, and interpretability) and a statistical method for agreement and disagreement accuracy. Scores of 1 (good), 2 (moderate), or 3 (poor) were assigned to each criterion, considering our experience with the respective strategies based on expert knowledge (EK) and machine learning (ML) for generating soil drainage class maps.

For the first criterion, accuracy was evaluated by model performance expressed by the kappa value of the confusion matrices obtained between the strategies and the reference map. Accuracies higher than 0.5 received a value of 1 while accuracies between 0.3 and 0.5 and lower than 0.3 received the values of 2 and 3, respectively. This division was clearly arbitrary and involved the adaptation of the table from Landis and Koch (1977). While these divisions are clearly arbitrary, they provide useful 'benchmarks' for discussing the interpretation of results.

Labor-efficiency was considered as the ease or difficulty in obtaining the data for each strategy. In other words, if the base maps or environmental covariates are easier to calculate or to obtain from other sources. A higher value was given to the strategy that required fewer resources in terms of the number of covariates or base maps.

Method transferability, in turn, was evaluated analogously as the labor-efficiency criterion. This means the ranking would be higher if a strategy was easily reimplemented in any other region with similar characteristics to our study area.

The interpretability criterion considered how understandable the results were for how predictions were made, especially for an audience with no statistical background. A high value was given to strategies that presented simple classification prediction rules or employed statistical analyses to summarize model function in a manner that was easy to

grasp. Finally, the criterion ratings were averaged per strategy, and the strategy that received the lowest value was selected as the most appropriate for mapping soil drainage classes.

Agreement accuracy refers to the percentage of cases in which the EK and ML provide results that match our reference map. On the other hand, disagreement accuracy represents the percentage of cases which the EK and ML do not match with our reference map, implying the accuracy of identifying cases where the methods differ or provide conflicting results. These statistical parameters help assess the consistency and discrepancy between the two methods and their alignment with the reference map in the context of the study. As the agreement accuracy values increase, so does the percentage accuracy of the predicted map relative to the reference map. Conversely, lower disagreement accuracy values correspond to decreased prediction errors.

## 3. Results

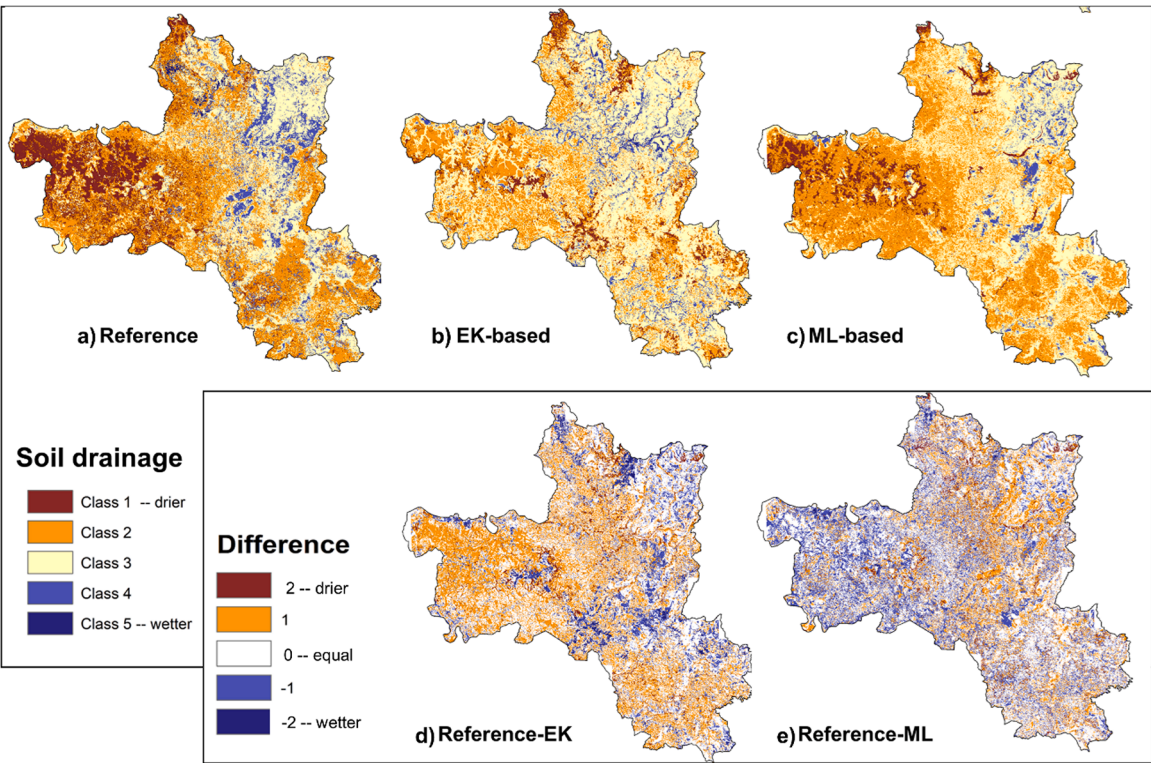
### 3.1. Predicting soil drainage classes by EK and ML strategies and comparison with reference data

The strategy based on EK used topsoil and subsoil color and textural classes for building numerical codes that were used to define drainage classes. The combination of these colors and textural classes produced a soil drainage map (Fig. 4b) in which 58 % of the area had Class 3 (moderately drained), which was mostly placed on areas with loam to clayey texture. The second most common drainage was Class 2 (well drained), which was characterized as having brownish color and sandy to loam texture. This soil drainage class covered approximately 29 % of the area. Class 1 and Class 4 were in less than 10 % of the area. Class 5, which represented poorly drained soils with low chroma or gleyed colors, was predicted to cover even less of the map area and was mostly restricted to areas near rivers.

The strategy based on ML used a set of environmental covariates to obtain a prediction model suitable for predicting soil drainage classes. The ML-based strategy produced a soil drainage map (Fig. 4c) in which Class 1 (very well drained) was the most common, with 33 % of the area, followed by Class 2, Class 3 and Class 4, with 24 %, 22 % and 20 % of the area, respectively. Class 5 (poorly drained), in turn, was predicted in only 0.11 % of the area. Overall, the ML-based map represented a much more even distribution across the drainage classes than the EK-based map. Instead of a single drainage class (moderately drained) dominating most of the map, the dominant soil drainage class in ML-based map (very well drained) was only a third of the map areas, with most of the other drainage classes sharing a similar portion of the map area. Additional results from the ML model are presented in Table S4 and Figure S2 of the supplementary material, in which model validation and most important covariates are presented.

Difference maps and confusion matrices between both strategies and the reference soil drainage map were used to investigate which strategy developed to map soil drainage classes was most similar to the reference map. Fig. 4d and e show the difference maps while Table 2 depicts the confusion matrix of each strategy with the reference data. The greater the negative or positive values in the difference maps, the higher the difference between the respective strategy and the reference data. A negative value meant the drainage class obtained either by EK or ML strategies was higher (i.e., wetter) than the reference map. Meanwhile, a positive value represented a lower (drier) drainage class. For example, suppose at a specific cell, the value of the difference map was -2. In that case, it means that the strategy for mapping soil drainage class predicted a Class 3 (moderately drained) soil drainage, where the reference had a Class 1 (very well drained). The reverse applied for positive values. Zero values, in turn, represent areas where the predicted maps correspond to the soil drainage class described in the reference soil map.

The first difference map (Fig. 4d) represents the values obtained after subtracting the reference map from the predicted EK-based map. The EK-based strategy had an agreement of 50 % with the reference map but



**Fig. 4.** Reference (a), EK- (b), and ML-based soil drainage maps with their corresponding difference maps (d-e). In the difference maps, positive values indicate that the reference map indicated wetter drainage classes than the EK- or ML-based soil drainage maps, while negative values represent areas where the reference map was drier than the EK- or ML-based soil drainage maps. Zero values indicate areas where either EK- or ML-based soil drainage classes are equal with the reference map.

**Table 2**  
Cell-by-cell confusion matrix between the reference drainage class map and the drainage class maps produced by the EK and ML strategies. PA%: Producer's accuracy, UA%: User's accuracy.

|                   |                | Reference soil drainage map |        |        |        |       | UA (%) |
|-------------------|----------------|-----------------------------|--------|--------|--------|-------|--------|
|                   |                | Class1                      | Class2 | Class3 | Class4 | Total |        |
| EK-based drainage | Class1         | 21                          | 28     | 11     | 13     | 73    | 29     |
|                   | Class2         | 233                         | 587    | 138    | 42     | 1000  | 59     |
|                   | Class3         | 37                          | 312    | 578    | 146    | 1073  | 54     |
|                   | Class4         | 0                           | 3      | 11     | 86     | 100   | 86     |
|                   | Total          | 291                         | 930    | 738    | 287    | 2246  |        |
|                   | PA (%)         | 7 %                         | 63 %   | 78 %   | 30 %   |       |        |
|                   | Accuracy kappa | 0.57                        |        |        |        |       | 0.33   |
| ML-based drainage | Class1         | 109                         | 22     | 5      | 0      | 136   | 80     |
|                   | Class2         | 159                         | 753    | 194    | 41     | 1147  | 66     |
|                   | Class3         | 23                          | 154    | 578    | 99     | 774   | 64     |
|                   | Class4         | 0                           | 1      | 41     | 147    | 189   | 78     |
|                   | Total          | 291                         | 930    | 738    | 287    | 2246  |        |
|                   | PA%            | 38 %                        | 81 %   | 68 %   | 51 %   |       |        |
|                   | Accuracy kappa | 0.67                        |        |        |        |       | 0.51   |

was prone to underestimating the drainage classes, especially in sandy soils. We found that 43 % of the predicted drainage classes were between  $\pm 1$  class from the drainage class observed in the reference map (values 1 and  $-1$  in Fig. 4d). The reference-ML difference map (Fig. 4e) had an agreement of 53 % with the reference data, being 41 % between  $\pm 1$  Class (values 1 and  $-1$  values in Fig. 4e).

Table 2 shows the confusion matrix between the strategies (rows) and the reference data (columns). Highlighted cells represent the number of samples that were correctly classified by both strategies.

Because Class 5 was not present in any of the sampling points, it was excluded from the confusion matrix, reducing the total number of samples from 2309 to 2246. Between the reference sampling points and the ML-based strategy, an accuracy of 0.67 and kappa of 0.50 were found. For the EK-based strategy, accuracy and kappa were lower (0.57 and 0.33, respectively).

The PA% and UA% indices were applied to assess the accuracy of individual drainage classifications on maps generated by ML- and EK-based strategies. Compared to reference data, the UA% for the ML strategy ranged from 64 % to 80 %, while the EK strategy showed lower UA% values, between 29 % and 86 %. Regarding PA%, the ML-generated drainage map showed values ranging from 38 % to 81 %, while the EK-based map displayed greater variability, with PA% values ranging from 7 % to 78 % (Table 2).

The confusion matrix revealed that both ML and EK strategies tend to overestimate well and very well-drained areas, assigning them to wetter classes, particularly in Classes 1 and 2, where the ML strategy achieved more correct classifications. In contrast, the EK strategy struggled with Class 1, exhibiting low UA% and PA% indices. For Classes 3 and 4, both strategies underestimated drainage levels to the next lower class, although EK was more accurate in predicting Class 3, while ML performed better for Class 4. The statistical agreement and disagreement method effectively assessed strategy performance, affirming ML's superior predictive accuracy for soil drainage classification.

3.2. Selecting the most appropriate strategy for mapping soil drainage classes

Table 2 depicts the scoring values for each criterion analyzed to select the most appropriate strategy for mapping soil drainage classes. The first qualitative criterion, accuracy, is represented by kappa values of each strategy depicted in Table 2. The EK-based strategy received a value of 2 (kappa values between 0.3 and 0.5) while the ML-based

strategy received the higher score as it presented a kappa value of 0.51.

The second criterion, labor efficiency, considers the data needed to build each strategy and the effort to obtain this data. A value of 1 was assigned for the EK-based strategy, whereas a value of 2 was attributed to the ML-based strategy.

Regarding the method transferability of both strategies, considering the easiness of whether to transfer or to repeat the strategy to develop these maps, EK received a value of 2 while a value of 3 was given to the ML-based strategy, since it requires more data and machine learning models are sometimes more difficult to replicate.

The interpretability criterion considered the easiness or not of understanding the rules for each strategy. The highest value was assigned for the ML-based strategy because ML models are still considered black boxes. The EK-based strategy received the lowest value because it is based on a set of rules that can be considered easier to understand and interpret.

In addition, we statically evaluate the predicted and observed maps using the agreement and disagreement method (Table 3). The machine learning (ML) strategy demonstrated superior performance, with the highest overall agreement (0.53) and lowest overall disagreement (0.47) compared to the Empirical Kriging (EK)-based strategy (0.45 and 0.55, respectively) (Table 3).

## 4. Discussion

### 4.1. Agreement of soil drainage maps with reference data

In Brazil, due to tropical climate conditions and deep and permeable soils, the existence of poor drainage classes is minimal compared to temperate regions. This spatial pattern may compromise the representation of all drainage classes and hamper further improvement of the map accuracy. Conversely, the almost equal agreement observed between the strategies and the reference map in the cell-by-cell approach can be explained by the fact that a high number of samples are available for each class, covering all the spatial domain of our study area. However, it is worth mentioning that the reference data is not free of error, as it was conceived from a soil class map with an overall accuracy of 82 % (Mendes et al., 2021). This needs to be considered when evaluating both strategies, which will be discussed later.

The ML-based strategy had the highest agreement and accuracy in mapping soil drainage classes in tropical environments. This approach uses a set of covariates as explanatory variables of the spatial distribution of drainage classes in the study area, which was not the case with the EK-based strategy, in which only soil color and textural maps were used. However, an interesting result found in the ML strategy was that soil color variables (hue and value in the subsoil) were second and third most important covariates that explained the spatial patterns of soil drainage classes, after the SWIR2 band of Landsat image (see supplementary material). Several researchers have demonstrated the importance of soil color in predicting soil drainage classes as it can reveal insights into the local hydrological regime (Schoonover; Crim, 2015), where color such as red, brown, and yellow are usually associated with

iron oxidation in soil environments where oxygen is abundant, indicating well-draining conditions. In other words, the soil is not permanently saturated or flooded, and infiltration conditions are optimal. On the other hand, low chroma values or gleying color are indicators of poor drainage and/or low iron content. This coloration is caused by water staying longer in the soil, favoring the reduction of iron.

The work of Malone et al. (2018) is an example, where color descriptions found in legacy soil maps were used to describe and spatially predict a soil drainage index in New South Wales, Australia. They used fuzzy set theory to allocate soil colors in pre-defined five color groups (red, brown, yellow, gray and black), which represented drainage classes ranging from 1 to 5 (very well to poorly drained), similar to the rules built for this work. The authors argued that their soil drainage index incorporated tacit knowledge, and it can be used in other areas since the soil color/soil drainage relationships were well defined by an expert. The SWIR2 band, in turn, is related to the variability of clay content, whose absorption feature due to minerals is centred at 2200 nm (Gomez et al., 2018). Environmental variables that were considered to have a direct relationship with water movement, such as topographic wetness index (TWI) and multi-resolution valley flatness bottom (MRVFB) were less important.

Some examples were reported using prediction models to predict soil drainage classes. One of the earlier works using prediction models was the work of Bell et al. (1992) who related landscape attributes to soil drainage classes to obtain the probability of occurrence of each class in the field. In a similar research, Levine et al. (1994) studied the relationship of environmental covariates with soil drainage classes and found a positive correlation with the values of NDVI, which are the highest values associated with well-drained soils. This is an indication of the conditions of the site for plant growth as poorly drained soils do not provide a suitable rhizosphere environment for most roots, but they are more fertile due to the accumulation of organic matter. Zhao et al. (2013) used topographic and hydrological variables and artificial neural networks to build a model for soil drainage classification. They used data from coarse-resolution maps and aimed at improving the resolution of the soil drainage map by using machine learning and digital soil mapping frameworks. In their results, the most efficient model utilized 12 linear equations corresponding to 12 landform types, which improved soil drainage class predictions with accuracy gains ranging from 7.5 % to 21.3 %. This approach proved to be effective in enhancing accuracy across large and complex areas.

In recent works, machine learning and deep learning have been more common. Beucher et al. (2019), for example, used artificial neural networks (ANN) to predict soil drainage classes in Denmark and compared their performance to that obtained by a decision tree classification (DTC) model. They used 31 covariates, which included topographical, hydrological, and soil variables. They found that the ANN outperformed the DTC, but the difference was just 2 %, reporting an overall accuracy of 54 % and 52 %, respectively. Considering the contribution of the environmental covariates, they found that the clay content in depth (100–200 cm) was the one that contributed the most. Among the ten most important variables, they also reported slope to channel network, geology, vertical and horizontal distance to channel network, wetness index and depth to groundwater. It is worth mentioning that the results obtained by Beucher et al. (2019) with the DTC model were similar to those reported here, with an overall accuracy of 52 % on the validation dataset, using all environmental variables (31 in total) and differential misclassification costs, this performance was comparable to the artificial neural network (ANN) model, which achieved a slightly higher overall accuracy of 54 % after variable selection. Cialella et al. (1997) also predicted soil drainage classes by using remote sensing data and topographical variables in a decision tree algorithm. They achieved an average accuracy of 78 % and topography being the most correlated variable with the drainage classes followed by NDVI.

The ML approach is the most common framework used nowadays for mapping soil drainage classes, but its main disadvantage is that before

**Table 3**

Scoring values for each criterion considered for selecting the most appropriate strategy and based on statistical "agreement accuracy" and "disagreement accuracy".

| Criteria                          | EK   | ML   |
|-----------------------------------|------|------|
| Accuracy                          | 2    | 1    |
| Labor-efficiency                  | 1    | 2    |
| Transferability                   | 2    | 3    |
| Interpretability                  | 1    | 3    |
| Mean                              | 1.5  | 2.25 |
| Statistical performance parameter | EK   | ML   |
| Overall Agreement                 | 0.45 | 0.53 |
| Overall Disagreement              | 0.55 | 0.47 |

modelling, the description of the soil drainage class in the field or from soil reports is needed. In turn, approaches based on expert knowledge are only used to first describe soil profiles in the field and then use ML approaches to predict drainage classes at unobserved locations spatially. Strategies similar to that used in this work with two soil properties are not common.

The ML approach is the most common framework used nowadays for mapping soil drainage classes, but its main disadvantage is that before modelling, the description of the soil drainage class in the field or from soil reports is needed. In turn, approaches based on expert knowledge are only used to first describe soil profiles in the field and then use ML approaches to predict drainage classes at unobserved locations spatially. Strategies similar to that used in this work with two soil properties are not common. However, there were attempts to build strategies based on expert knowledge. This was the case of [Geradin and Duerue \(1990\)](#), who aimed to build an objective approach for mapping soil drainage classes in forest soils, easily accessible by non-specialists. They used topographical variables (hillslope position, terrain shape, effective upper slope position and declivity), abundance and mottling depth, solum depth and humus thickness to build rules and allocate soil drainage classes considering the variability of these parameters. Although particularly useful, this approach was not used very much because it appeared overly site-specific and, at some point, difficult to replicate. Therefore, considering the accuracy and agreement with reference data, ML based approaches are still the better strategy for mapping soil drainage classes.

#### 4.2. The most appropriate strategy for mapping soil drainage class based on operational criteria

Although the ML-based strategy presented the best results in terms of accuracy and agreement with the reference data, we evaluated three other criteria to select the most appropriate strategy for mapping soil drainage classes. The first criterion was the accuracy, which, as mentioned above, was the more appropriate for the ML-based strategy, as it had the highest kappa value and the highest agreement in the cell-by-cell comparison (although this latter was not used to evaluate accuracy).

For the second criterion, labor efficiency, the EK-based strategy received the lower value (1), because it required fewer resources when compared to the ML-based strategy (2), which required 23 environmental covariates (initial number of covariates: 603) to build prediction models to represent the spatial variability of soil drainage classes. Besides that, the environmental covariates used required a set of pre-processing steps that demanded considerable time and knowledge of statistics and mathematics. The EK-based strategy used only two soil properties (soil color and texture maps with accuracy > 0.6) that now are easily available on web platforms such as SoilGrids, for example [Poggio et al. \(2021\)](#).

Although the resolution of the soil property maps available on the SoilGrids platform might not be sufficient for a specific area, it is a straightforward way to have an idea of the variability of soil texture and color. Another option is the availability of this information in legacy soil maps or soil reports. This is especially important when it is not an option to perform field surveys to describe the soil drainage of an area, which also was mentioned as one of the limitations of the ML-based strategy. Of course, legacy soil data and descriptions available in legacy reports might not be sufficient to spatially represent soil drainage classes, but they can also be an easier and less resource-requiring strategy for obtaining information on soil drainage classes.

In the third criterion, transferability, the EK-based strategy also received the best score (2), as it is based on a set of numerical codes that used soil color and textural classes that are easy to apply to other areas. The EK-based strategy did not receive the highest best value (1) because the rules used here may apply only to tropical areas, which are not 100 % transferable to other regions where the variability of soil color

and texture might differ. Another limitation of the transferability of the EK-based strategy can be related to the use of soil color, if not a proper separation of groups representing the sequence of drainage classes exists. The rules obtained from the ML strategy, in turn, were built automatically by decision trees based specifically on the covariates used for this work and might not be transferable to any other region, unless equal conditions in terms of climate and soil are characteristic. Therefore, the value assigned to this criterion for the ML strategy was the highest, considering that the transferability of ML models is limited by the feature space ([Meyer and Pebesma, 2021](#)).

The last criterion, interpretability, is related to the ability to grasp the rules used in each strategy. The EK-based strategy again received the highest score, as it is based on a set of rules that consider the variability of soil color from red (well drained) to black (poorly drained) and the textural classes from sandy to very clayey. The example presented in [Fig. 3](#) provides a good understanding of these rules. We decided to give the EK-based strategy the highest score as we considered the facility to understand the numerical code presented. An example of the use of soil color for classifying drainage classes was presented in [Malone et al. \(2018\)](#) and was the basis for this work, which can be seen as an example of the facility to interpret and apply rules based on soil color. Conversely, although the ML strategy is based on a simple decision tree to simplify the process, the selection of covariates that would have more importance in the model and how this process happens inside the model is still not well understood. Besides that, as the number of covariates increases, the decision tree's complexity also increases, making it impossible to follow all the rules and have a unique rule for a specific soil drainage class.

#### 4.3. Expert knowledge and machine learning drainage maps for agricultural decision-making and limitations

The machine learning (ML)-based soil drainage class mapping strategy demonstrated greater efficiency than the expert knowledge (EK) strategy in identifying soil classes with drainage restrictions, particularly in classes 4 and 5 ([Fig. 4b](#) and [c](#)) and areas with negative values ([Fig. 4e](#)). This superiority stems from the ML approach's ability to incorporate a broader range of environmental covariates related to drainage deficiencies, compared to the EK strategy, which relies solely on soil color and texture. Consequently, the ML-derived soil drainage maps achieved higher producer's accuracy for classes 4 and 5 than those generated using the EK method. In this sense, the ability of the ML strategy to effectively identify rare critical conditions, as reflected in classes 4 and 5, highlights its value for decision-making.

This distinction is significant, as drainage restrictions impose critical limitations on land use and management for both agricultural and environmental purposes ([Ramalho Filho and Beek, 1995](#)). Poor drainage affects oxygen availability and flow to crop roots, narrows the optimal soil consistency range for preparation and management, and alters nutrient dynamics, particularly for elements like nitrogen and reducible nutrients such as iron, manganese, and sulfur ([Vymazal and Kröpfelová, 2008](#); [Lamers et al., 2012](#)). Furthermore, poorly drained soils are vital considerations in agronomic planning (e.g., crop selection and drainage improvements) and ecosystem services (e.g., wetland conservation).

Conversely, the EK-based maps underrepresent well-drained soils (class 1). In practical terms, this misclassification can lead to misguided decision-making regarding the management of well-drained soils, particularly when these areas are intended for aquifer recharge purposes. Furthermore, such inaccuracies could have broader implications for other environmental and agricultural applications such as land-use management and planning. For instance, failing to accurately identify drainage-restricted soils may result in management strategies that are inconsistent with actual field conditions, leading to inefficiencies in assessing the agricultural suitability of the land and environmental degradation.

When assessing the transferability of the proposed methodology, it is

essential to consider the specific limitations of both EK-based and ML-based strategies. While EK-based approaches may face challenges such as the dependency on bare soil conditions and the resolution of texture maps, ML approaches require extensive training datasets and computational resources. Both strategies should be evaluated in the context of the region's data infrastructure, environmental conditions, and technical expertise to ensure realistic applicability.

## 5. Conclusions

This study compared two strategies for mapping soil drainage classes in Brazil: one based on expert knowledge (EK) and the other on machine learning (ML). Both strategies achieved an agreement of approximately 50 % with the reference map in cell-by-cell comparisons, with the ML-based map showing slightly higher accuracy. However, the EK-based strategy demonstrated significant advantages in terms of labor efficiency, transferability, and interpretability.

The EK approach required fewer covariates, relying mainly on soil color and texture, making it more resource-efficient and easier to apply in other regions with minimal adjustments. Its rules, derived from pedological knowledge, were also more interpretable. Conversely, the ML strategy showed better performance in point-by-point evaluations and provided deeper insights into spatial variability, benefiting from the integration of multiple environmental covariates.

Our article highlights that, while ML is more suitable for detailed analyses of variability and uncovering relationships between influencing factors, EK offers a practical and efficient solution for mapping soil drainage classes, especially in data-scarce environments. However, both strategies showed weak performance in estimating drainage classes, as the results suggest an equal likelihood of correct or incorrect predictions with either approach. Future research should focus on improving surface-to-subsurface predictions to enhance the applicability of these methods.

## Ethics approval

All authors have read, understood, and complied with all requirements regarding the ethical responsibility of the journal and are aware that, with minor exceptions, no changes may be made in authorship after submission of the article.

## Author's contributions

All authors participated in all stages of conducting the research and writing the manuscript.

## Consent to publish

The authors consent to the case manuscript being submitted to and published in the journal.

## Funding

This research was funded by the National Council for Scientific and Technological Development (CNPq): Programa de Apoio à Fixação de Jovens Doutores no Brasil, grant number 168180/2022 –7; Fundação Araucária: CP 19/2022—Jovens Doutores; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), grant number 001; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant number 2021/05129–8, for sensor financial support; and by CEA-GRE—Centro de Excelência em Agricultura Exponencial for financial support.

## CRedit authorship contribution statement

**Jose Alexandre Melo Demattê:** Visualization, Supervision, Project

administration, Conceptualization. **Bradley A. Miller:** Visualization, Methodology, Investigation. **Uemeson José dos Santos:** Writing – review & editing, Writing – original draft. **Nélida Elizabet Quiñonez Silvero:** Investigation, Formal analysis, Data curation. **Elpidio Inácio Fernandes-Filho:** Writing – review & editing, Writing – original draft. **Daniilo César de Mello:** Formal analysis, Conceptualization. **Marcelo Rodrigo Alves:** Writing – original draft, Visualization. **Marcos Rafael Nanni:** Writing – original draft, Visualization. **Renan Falcioni:** Visualization, Supervision. **Gustavo Vieira Veloso:** Writing – original draft, Visualization, Data curation, Conceptualization. **Jean Jesus Macedo Novais:** Visualization, Software. **Bruno dos Anjos Bartsch:** Methodology, Investigation. **Jorge Tadeu Fim Rosas:** Software, Resources. **Nicolas Augusto Rosin:** Software, Formal analysis.

## 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.

## Acknowledgments

We would like to thank to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the scholarship awarded to the first author (project 2024/06285–1), which enabled the completion of this manuscript (ESALQ/USP). We also extend our gratitude to the Programa Nacional de Becas de Postgrado en el exterior “Don Carlos Antonio López” (BECAL) of the Government of Paraguay for granting the scholarship to the second author, the São Paulo Research Foundation (FAPESP) for financial support (Project grant n. 2014/22262–0). The authors would like to thank Nelida Silvero, co-author of this work, for her valuable contributions and for allowing this article to be derived from her thesis. Her dedication and efforts have been fundamental to the development and success of this research and the Geotechnologies on Soil Science group – GeoCIS ([esalqgeocis.wixsite.com/english](http://esalqgeocis.wixsite.com/english)), the Geospatial Laboratory for Soil Informatics at Iowa State University and to everyone who directly or indirectly contributed to this study.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.soilad.2024.100028](https://doi.org/10.1016/j.soilad.2024.100028).

## Data availability

No data was used for the research described in the article.

## References

- Alvares, C.A., Stape, J.L., Sentelhas, P.C., Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. *Meteorol. Z.* 22 (6), 711–728. <https://doi.org/10.1127/0941-2948/2013/0507>.
- Asgari, N., Ayoubi, S., Demattê, J.A.M., 2018. Soil drainage assessment by magnetic susceptibility measures in western Iran. *Geoderma Reg.* 13, 35–42. <https://doi.org/10.1016/j.geodrs.2018.03.003>.
- Bell, J.C., Cunningham, R.L., Havens, M.W., 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. *Soil Sci. Soc. Am. J.* 56 (6), 1860–1866. <https://doi.org/10.2136/sssaj1992.03615995005600060035x>.

- Beucher, A., Møller, A.B., Greve, M.H., 2019. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. *Geoderma* 352, 351–359. <https://doi.org/10.1016/j.geoderma.2017.11.004>.
- Bonfatti, B.R., Demattè, J.A.M., Marques, K.P.P., Poppiel, R.R., Rizzo, R., Mendes, W.S., Silvero, N.E.Q., Safanelli, J.L., 2020. Digital mapping of soil parent material in a heterogeneous tropical area. *Geomorphology*, 107305. <https://doi.org/10.1016/j.geomorph.2020.107305>.
- Bouma, J., 1973. Use of physical methods to expand soil survey interpretations of soil drainage conditions. *Soil Sci. Soc. Am. J.* 37 (3), 413–421. <https://doi.org/10.2136/sssaj1973.03615995003700030030x>.
- Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma* 264, 256–274. <https://doi.org/10.1016/j.geoderma.2015.05.017>.
- Cialella, A.T., Dubayah, R., Lawrence, W., Levine, E., 1997. Predicting soil drainage class using remotely sensed and digital elevation data. *Photogramm. Eng. Rem. S.* 63 (2), 171–178.
- Demattè, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial Soil Sensing System (GEOS3): a powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sens. Environ.* 212, 161–175. <https://doi.org/10.1016/j.rse.2018.04.047>.
- FAO. (2014). World Reference Base for Soil Resources 2014: International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. Rome: Food and Agriculture Organization of the United Nations.
- Faussey, N.R., 2005. Drainage, surface and subsurface. In: Hille, D. (Ed.), *Encyclopedia of Soils in the Environment*. Elsevier, pp. 409–413.
- Fitzgerald, J.B., Brereton, A.J., Holden, N.M., 2008. Simulation of the influence of poor soil drainage on grass-based dairy production systems in Ireland. *Grass Forage Sci.* 2008 (63), 380–389. <https://doi.org/10.1111/j.1365-2494.2008.00637.x>.
- Franzmeier, D.P., Klavivko, E.J., 2001. Drain. Wet. Soil Manag. Drain. Wet. Soil Manag. *Wet. Soils Indiana Wet. Soils Indiana*.
- Gerardin, V., Duerue, J.P., 1990. An objective approach to evaluating natural drainage of forest mineral soils for non-specialists. *Vegetatio* 87 (2), 127–133. <https://doi.org/10.1007/BF00042949>.
- Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P., Roger, J.M., Briottet, X., 2018. Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. *Remote Sens. Environ.* 204, 18–30. <https://doi.org/10.1016/j.rse.2017.10.047>.
- Graça, J., Daly, K., Bondi, G., Ikoyi, I., Crispie, F., Cabrera-Rubio, R., Cotter, P.D., Schmalenberger, A., 2021. Drainage class and soil phosphorus availability shape microbial communities in Irish grasslands. *Eur. J. Soil Biol.* 104, 103297. <https://doi.org/10.1016/j.ejsobi.2021.103297>.
- Grimley, D.A., Wang, J.S., Liebert, D.A., Dawson, J.O., 2008. Soil magnetic susceptibility: a quantitative proxy of soil drainage for use in ecological restoration. *Restor. Ecol.* 16 (4), 657–667. <https://doi.org/10.1111/j.1526-100X.2008.00479.x>.
- Hijmans, R.J., Van Etten, J., 2016. Raster: geographic data analysis and modeling. R. Package Version 2, 5–8.
- Hudson, H.D., 1992. The soil survey as paradigm-based science. *Soil Sci. Soc. Am. J.* 56, 836–841.
- IGG, 1966. Folha Geológica de Piracicaba. Escala 1:100.000.
- Kerebel, A., Holden, N.M., 2013. Allocation of grass fields to hybrid soil moisture deficit model drainage classes using indicators. *Soil Till. Res.* 127, 45–59. <https://doi.org/10.1016/j.still.2012.04.004>.
- Kravchenko, A.N., Bollero, G.A., Omonode, R.A., Bullock, D.G., 2002. Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. *Soil Sci. Soc. Am. J.* 66, 235–243. <https://doi.org/10.2136/sssaj2002.2350>.
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, S., Kenkel, B., Team, R.C., 2020. Package 'caret'. R. J.
- Lamers, L.P., Van Diggelen, J.M., Op den Camp, H.J., Visser, E.J., Lucassen, E.C., Vile, M. A., Roelofs, J.G., 2012. Microbial transformations of nitrogen, sulfur, and iron dictate vegetation composition in wetlands: a review. *Front. Microbiol.* 3, 156. <https://doi.org/10.3389/fmicb.2012.00156>.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*. 159–174.
- Levine, E.R., Knox, R.G., Lawrence, W.T., 1994. Relationships between soil properties and vegetation at the Northern Experimental Forest, Howland, Maine. *Remote Sens. Environ.* 47 (2), 231–241. [https://doi.org/10.1016/0034-4257\(94\)90158-9](https://doi.org/10.1016/0034-4257(94)90158-9).
- Mackintosh, E.E., Van, Dern, Hulst, J., 1979. Soil drainage classes and soil water tables relations in medium and coarse textured soils In Southern Ontario. *Can. J. Soil Sci.* 58, 287–288.
- Malone, B.P., Mcbratney, A.B., Minasny, B., 2018. Description and spatial inference of soil drainage using matrix soil colours in the Lower Hunter Valley, New South Wales, Australia. *PeerJ* 6, e4659. <https://doi.org/10.7717/peerj.4659>.
- Mello, F.A.O., Demattè, J.A.M., Rizzo, R., Dotto, A.C., Poppiel, R.R., Mendes, W.S., Guimarães, C.C.B., 2021. Expert-based maps and highly detailed surface drainage models to support digital soil mapping. *Geoderma* 384, 114779.
- Mendes, W.S., Demattè, J.A.M., Silvero, N.E.Q., Rabelo Campos, L., 2021. Integration of multispectral and hyperspectral data to map magnetic susceptibility and soil attributes at depth: a novel framework. *Geoderma* 385, 114885. <https://doi.org/10.1016/j.geoderma.2020.114885>.
- Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* 12 (9), 1620–1633. <https://doi.org/10.1111/2041-210x.13650>.
- Niang, M.A., Nolin, M., Bernier, M., Perron, I., 2012. Digital mapping of soil drainage classes using multitemporal RADARSAT-1 and ASTER images and soil survey data. *Appl. Environ. Soil Sci.* 17. <https://doi.org/10.1155/2012/430347>.
- Peng, W., Wheeler, D.B., Bell, J.C., Krusemark, M.G., 2003. Delineating patterns of soil drainage class on bare soils using remote sensing analyses. *Geoderma* 115, 261–279. [https://doi.org/10.1016/S0016-7061\(03\)00066-1](https://doi.org/10.1016/S0016-7061(03)00066-1).
- Poggio, L.S., Batjes, L.M., Heuvelink, N.H., Kempen, G.B.M., Ribeiro, B.E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* 7 (1), 217–240. <https://doi.org/10.5194/soil-7-217-2021>.
- Pontius, Jr, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* 32, 4407–4429.
- Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn.* 1, 81–106.
- R Core Team, 2023. R: A language and environment for statistical computing. Ramalho Filho, A., Beek, K.J., 1995. Sistema de avaliação da aptidão agrícola das terras. In: Rio de Janeiro, 1995. EMBRAPA-CNPS.
- Raymond, J.E., Fernandez, I.J., Ohno, T., Simon, K., 2012. Soil drainage class influences on soil carbon in a New England Forested Watershed. *Soil Sci. Soc. Am. J.* 77, 307–317. <https://doi.org/10.2136/sssaj2012.0129>.
- Rossi, M., 2017. Map pedológico do Estado de São Paulo: Revisado e ampliado. Instituto Florestal de São Paulo.
- Santos, M.L.M., Ten Caten, A., 2015. Mapeamento Digital de Solos (MDS): Avanços e desafios. *Bol. Inf. Da SBSC* 39–42.
- Schaetzl, R., 2013. Catenas and soils. In: Shroder, J. (Ed.), *Treatise on Geomorphology*. Academic Press, pp. 145–158.
- Schaetzl, R.J., Miller, B.A., Baish, C.J., 2022. Catenas and Soils. In: Shroder, J.F. (Ed.), *Treatise on Geomorphology*, Second edition. Academic Press, pp. 221–235. <https://doi.org/10.1016/B978-0-12-818234-5.00214-5>.
- Schoonover, J.E., Crim, J.F., 2015. An introduction to soil concepts and the role of soils in watershed management. *J. Contemp. Water Res. Educ.* 154 (1), 21–47. <https://doi.org/10.1111/J.1936-704X.2015.03186.X>.
- Sistema Brasileiro de Classificação de Solos. 2018, 5.ed. Brasília, DF: Embrapa. 355 p.
- Soil Science Division Staff, 2017. Soil survey manual. In: Ditzler, C., Scheffe, K., Monger, H.C. (Eds.), *USDA Handbook 18*. Government Printing Office, Washington, D.C.
- Souza, C.M., Shimbo, J.Z., Rosa, M.R., Parente, L.L., Alencar, A.A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., Ferreira, L.G., Souza-Filho, P.W.M., Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vêlez-Martin, E., Azevedo, T., 2020. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat archive and Earth Engine. *Remote Sens.* 12 (17), 2735. <https://doi.org/10.3390/RS12172735>.
- Su, Q., Tao, W., Mei, S., Zhang, X., Li, K., Su, X., Guo, J., Yang, Y., 2021. Landslide susceptibility zoning using C5.0 decision tree, random forest, support vector machine and comparison of their performance in a coal mine area. *Front. Earth Sci.* 9, 781472.
- Troeh, F.R., 1964. Landform parameters correlated to soil drainage. *Soil Sci. Soc. Am. J.* 28 (6), 808–812. <https://doi.org/10.2136/sssaj1964.03615995002800060035x>.
- Vymazal, J., Kröpfelová, L., 2008. Transformation mechanisms of major nutrients and metals in wetlands. *Wastewater Treat. Constr. Wetl. Horiz. sub-Surf. Flow., Environ. Pollut.* 14, 11–91. <https://doi.org/10.1007/978-1-4020-8580-2.2>.
- Wickland, K.P., Neff, J.C., Harden, J.W., 2010. The role of soil drainage class in carbon dioxide exchange and decomposition in boreal black spruce (*Picea mariana*) forest stands. *Can. J. For. Res.* 40 (11), 2123–2134. <https://doi.org/10.1139/X10-163>.
- Zhao, Z., Ashraf, M.I., Meng, F., 2013. Model prediction of soil drainage classes over a large area using a limited number of field samples: a case study in the province of Nova Scotia, Canada. *Can. J. Soil Sci.* 93, 73–83. <https://doi.org/10.4141/CJSS2011-095>.