MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

RÔMULO JOSÉ ALENCAR SOBRINHO

AI-DRIVEN DIGITAL SOIL MAPPING OF THE EASTERN AMAZON: INTEGRATING PEDOLOGICAL KNOWLEDGE AND PEDOMETRICS

BELÉM
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

RÔMULO JOSÉ ALENCAR SOBRINHO

AI-DRIVEN DIGITAL SOIL MAPPING OF THE EASTERN AMAZON: INTEGRATING PEDOLOGICAL KNOWLEDGE AND PEDOMETRICS

A thesis presented to the Agronomy Graduate Program at the Universidade Federal Rural da Amazônia as part of the requirements for the Doctor degree in Agronomy.

Field of Concentration: Environmental Resource Management and Conservation.

Advisor: Dr. João Fernandes da Silva Júnior

BELÉM
2026

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas da Universidade Federal Rural da Amazônia
Gerada automaticamente mediante os dados fornecidos pelo(a) autor(a)

S677a Sobrinho, Rômulo José Alencar
AI-Driven Digital Soil Mapping of the Eastern Amazon : Integrating Pedological Knowledge and Pedometrics / Rômulo José Alencar Sobrinho. - 2026.
152 f. : il. color.

Tese (Doutorado) - Programa de Pós-Graduação em Agronomia (PPGAGRO), Campus Universitário de Belém, Universidade Federal Rural Da Amazônia, Belém, 2026.
Orientador: Prof. Dr. João Fernandes da Silva Júnior

1. Soil mapping. 2. Remote sensing. 3. Artificial intelligence. 4. Soil-landscape relationships. 5. hydromorphic soils. I. Silva Júnior, João Fernandes da , *orient.* II. Título

CDD 631.4

RÔMULO JOSÉ ALENCAR SOBRINHO

AI-DRIVEN DIGITAL SOIL MAPPING OF THE EASTERN AMAZON: INTEGRATING PEDOLOGICAL KNOWLEDGE AND PEDOMETRICS

A thesis presented to the Agronomy Graduate Program at the Universidade Federal Rural da Amazônia as part of the requirements for the Doctor degree in Agronomy.

Field of Concentration: Environmental Resource Management and Conservation.

Advisor: Dr. João Fernandes da Silva Júnior

Approval date: April 30, 2026

Committee members:

Documento assinado digitalmente
JOAO FERNANDES DA SILVA JUNIOR
Data: 05/05/2026 19:27:21-0300
Verifique em https://validar.iti.gov.br

Dr. João Fernandes da Silva Junior - Committee Chair
Universidade Federal Rural da Amazônia - UFRA

Documento assinado digitalmente
ELVIO GIASSON
Data: 05/05/2026 09:17:35-0300
Verifique em https://validar.iti.gov.br

Dr. Elvio Giasson – 1st Examiner
Universidade Federal do Rio Grande do Sul – UFRGS

Documento assinado digitalmente
MICHELE DUARTE DE MENEZES
Data: 05/05/2026 18:55:06-0300
Verifique em https://validar.iti.gov.br

Dra. Michele Duarte de Menezes– 2nd Examiner
Universidade Federal de Lavras – UFLA

Documento assinado digitalmente
FABRICIO DA SILVA TERRA
Data: 05/05/2026 08:14:15-0300
Verifique em https://validar.iti.gov.br

Dr. Fabrício da Silva Terra – 3rd Examiner
Universidade Federal dos Vales do Jequitinhonha e Muruci – UFVJM

Documento assinado digitalmente
WALDIR DE CARVALHO JUNIOR
Data: 04/05/2026 13:29:49-0300
Verifique em https://validar.iti.gov.br

Dr. Waldir de Carvalho Junior – 4th Examiner
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA

Acknowledgments

À Deus.

À minha família, minha esposa Sueny, minha filha Alícia, meu pai Romildo, minha mãe Eugênia, minha irmã Rayanne, meus sogros, meus sobrinhos e cunhados, pela compreensão, pela motivação para superar os desafios e pele amor a mim dedicados.

À Universidade Federal Rural da Amazônia (UFRA) pela oportunidade de realizar este doutorado, através do Programa de Pós-Graduação em Agronomia (PGAGRO) e aos docentes. Ao Campus da UFRA Capanema por ceder a estrutura para a realização de parte dos estudos. À Coordenação de Aperfeiçoamento e Pessoal de Nível Superior (CAPES) pela bolsa de estudo concedida e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento do projeto de pesquisa

Ao meu orientador e amigo, Prof. Dr. João Fernandes da Silva Júnior, pela dedicação em compartilhar o seu conhecimento e ao Prof. Dr. Daniel Pereira Pinheiro por todo o apoio.

Aos meus amigos do PGAGRO pelo companheirismo durante a execução das disciplinas. À Secretária do PGAGRO, Nena, pelo auxílio no cumprimento dos prazos e das diretrizes estabelecidas pelo PGAGRO. Aos amigos Yan Nunes e Flávio Henrique, pelo grande apoio na realização das análises no Laboratório de Elementos Traços no Ambiente (LETAM) da Universidade Federal Rural da Amazônia.

Aos membros da banca avaliadora de qualificação e defesa, pela contribuição.

Rômulo José Alencar Sobrinho

“Jesus, eu confio em vós”

RESUMO

A lacuna de informações mais detalhadas da distribuição de solos na Amazônia é um impedimento para a efetiva governança de solos nessa região. Portanto, o objetivo geral do presente estudo foi avaliar performance e a incerteza das abordagens de assembleia de algoritmos, mapa legado, covariáveis ambientais e Área de Referência. Para tanto, a presente tese foi dividida em dois capítulos, cujos títulos são: I) Integração de conhecimento tácito e IA para mapeamento digital de solos na Amazônia oriental: assembleia de algoritmos, desempenho do modelo e incorporação de incertezas; e II) Transferência das relações solo-paisagem em regiões da Amazônia com dados limitados: uma abordagem de mapeamento digital guiada pelo conhecimento e baseada em área de referência. Dados de saída de ambos os capítulos seguiram os passos comuns a seguir. Foram usadas 15 covariáveis ambientais baseadas no modelo SCORPAN. Para gerar os modelos preditivos nós usamos os algoritmos e quatro conjuntos de dados de entrada Randon Forest, Ranger, Xgboost, C 5.0 e Ensemble learning. Avaliamos também dois níveis de detalhamento das unidades de mapeamento do mapa legado: nível taxonômico de ordem e de Grande grupo. Covariáveis de terreno foram derivadas de dois modelos de elevação distintos: modelo digital de superfície e modelo digital de terreno. Foram analisados ao todo 20 modelos preditivos. Como destaques das conclusões do Capítulo 2: a abordagem assembleia de algoritmos usando modelo digital de superfície e unidade de mapeamento ao nível taxonômico de Grande grupo obtiveram performance estatísticamente superior a partir de precisão e Kappa em relação aos modelos individuais, ao nível taxonômico de Grande grupo. Covariáveis ambientais como: Temperatura Média Anual de 22,000 anos atrás, Nível de Base da Rede de Canais, Altitude e Temperatura da Superfície Terrestre apresentaram maior importância na modelagem preditiva; e como destaques do Capítulo 3: na validação externa na área de extrapolação os modelos preditivos baseados em assembleia de algoritmos usando dados derivados do modelo digital de superfície, na unidade de mapeamento ao nível de Ordem, assembleia de algoritmos usando dados derivados do modelo digital de terreno, usando unidades de mapeamento ao nível de Grande grupo e assembleia de algoritmos usando dados derivados do modelo digital de superfície, usando unidades de mapeamento ao nível taxonômico de Grande grupo e a previsão das unidades de mapeamento GLEISSOLOS e GLEISSOLOS SÁLICOS Sódicos se destacaram com as mais precisas na área de extrapolação. Os mapas gerados a partir do mapa legado e para a área de extrapolação, utilizando área de referência, mapearam com sucesso unidades de mapeamento relacionadas ao ambiente hidromórfico. O pioneirismo deste estudo na Amazônia revelou o potencial da abordagem de assembleia de algoritmos do mapeamento de classes de solos sob ambientes hidromorficos, ambientes que ocupam grande área na Amazônia e são de relevante importância.

Palavras-chave: mapeamento do solo, sensoriamento remoto, inteligência artificial, relação solo-paisagem, solos hidromórficos, SCORPAN.

ABSTRACT

The lack of more detailed information on soil distribution in the Amazon is an impediment to effective soil governance in this region. Therefore, the overall objective of this study was to evaluate the performance and uncertainty of assembly algorithm approaches, legacy mapping, environmental covariates, and reference area. To this end, this thesis was divided into two chapters, titled: I) Integrating Tacit Knowledge and AI for Digital Soil Mapping in Eastern Amazonia: Ensemble Learning, Model Performance, and Uncertainty Incorporation; and II) Transferring soil–landscape relationships in data-limited regions of the Amazon: a knowledge-guided digital soil mapping approach based on reference areas. Output data from both chapters followed the common steps below. Fifteen environmental covariates based on the SCORPAN model were used. To generate the predictive models, we used the algorithms and four input datasets: Random Forest, Ranger, Xgboost, C 5.0, and Ensemble Learning. We also evaluated two levels of detail for the Mapping Units of the legacy map: taxonomic order level and large group level. Terrain covariates were derived from two distinct elevation models: digital surface model and digital terrain model. A total of 20 predictive models were analyzed. Key findings from Chapter 2: the algorithm assembly approach using a digital surface model and unit mapping at the Large Group taxonomic level achieved statistically superior performance in terms of accuracy and Kappa compared to individual models at the Large Group taxonomic level. Environmental covariates such as: 22,000-year-old Mean Annual Temperature, Channel Network Base Level, Altitude, and Land Surface Temperature showed greater importance in predictive modeling; And as highlights of Chapter 3: in the external validation in the extrapolation area, the ensemble learning predictive models using data derived from the digital surface model, in mapping units at the Oreder taxonomic level, ensemble learning using data derived from the digital terrain model, in mapping units at the Order taxonomic level and ensemble learning using data derived from the digital surface model, in mapping units at the Great group taxonomic level, and the prediction of the mapping units GLEISSOLOS and GLEISSOLOS SÁLICOS Sódicos stood out as the most accurate in the extrapolation area. The maps generated from the legacy map and for the extrapolation area, using reference area, successfully mapped mapping units related to the hydromorphic environment. The pioneering nature of this study in the Amazon revealed the potential of the ensemble learning approach for mapping soil classes under hydromorphic environments, environments that occupy a large area in the Amazon and are of significant importance.

Keywords: soil mapping, remote sensing, artificial intelligence, soil-landscape relationships, hydromorphic soils, SCORPAN.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

MU Mapping Unit

LD Level of Detail

RF Random Forest algorithm

Xgb Xgboost algorithm

C5 C 5.0 algorithm

EL Ensemble Learning (EL)

DSM Digital Surface Model

DTM Digital Terrain Model

AAP Average Annual Precipitation

PR_22k Annual Precipitation 22,000 years ago

Tmean_22k Annual Mean Temperature 22,000 years ago

SAVI Soil-Adjusted Vegetation Index

LST Land Surface Temperature

CNBL Channel Network Base Level

DD Drainage Density

PDSM Predictive Digital Soil Mapping

MRRTF Multiresolution Index of Ridge Top Flatness

MRVBF Multiresolution Index of Valley Bottom Flatness

PFC Profile Curvature

RSP Relative Slope Position

IOR Iron Oxide Ratio

GEOM Geomorphons

RFS1 RF algorithm using data derived from DSM, in LD1 of the MU

RangerS1 Ranger algorithm using data derived from DSM, in LD1 of the MU

XgbS1 Xgboost algorithm using data derived from DSM, in LD1 of the MU

C5S1 C5 algorithm using data derived from DSM, in LD1 of the MU

ELS1 EL algorithm using data derived from DSM, in LD1 of the MU

RFT1 RF algorithm using data derived from DTM, in LD1 of the MU

RangerT1 Ranger algorithm using data derived from DTM, in LD1 of the MU

XgbT1 Xgboost algorithm using data derived from DTM, in LD1 of the MU

C5T1 C5 algorithm using data derived from DTM, in LD1 of the MU

ELT1 EL algorithm using data derived from DTM, in LD1 of the MU

RFS2	RF algorithm using data derived from DSM, in LD2 of the MU
RangerS2	Ranger algorithm using data derived from DSM, in LD2 of the MU
XgbS2	Xgboost algorithm using data derived from DSM, in LD2 of the MU
C5S2	C5 algorithm using data derived from DSM, in LD2 of the MU
ELS2	EL algorithm using data derived from DSM, in LD2 of the MU
RFT2	RF algorithm using data derived from DTM, in LD2 of the MU
RangerT2	Ranger algorithm using data derived from DTM, in LD2 of the MU
XgbT2	Xgboost algorithm using data derived from DTM, in LD2 of the MU
C5T2	C5 algorithm using data derived from DTM, in LD2 of the MU
ELT2	EL algorithm using data derived from DTM, in LD2 of the MU
UA	User’s Accuracy
PA	Producer’s Accuracy
AO	Overall Accuracy
IBGE	Instituto Brasileiro de Geografia e Estatística

SUMMARY

CHAPTER 1

1.1 Introdution

Conventional soil surveys are the primary source of information on soil spatial variability for the rational and sustainable planning of this natural resource (Giasson et al., 2006). This dataset is essential for assessing the potential and/or limitations of a given soil class or mapping unit (Oliveira Junior et al., 1999). However, the level of detail in a survey is largely related to the minimum mapping area (MMA), which corresponds to the smallest area that can be delineated on a map or chart (Cavararo, 2015).

Currently, only 5% of Brazil is covered by soil maps at a scale of 1:100,000 or larger—that is, surveys ranging from medium-intensity reconnaissance to ultra-detailed—whose practical application depends on the objectives of the survey (Polidoro et al., 2016; Santos et al., 1995). In general, most of the Brazilian territory is covered by pedological surveys at a scale of 1:250,000.

Among Brazilian biomes, the Amazon biome presents one of the largest gaps in soil information in terms of level of detail, with the 1:250,000 scale predominating across most of the region. This limitation contributes to inadequate land management, potentially accelerating soil weathering processes and leading to soil degradation and low productivity (Giasson et al., 2006).

Given this gap, the vast territorial extent of the Amazon biome and the difficulty of access—especially in hydromorphic environments—make conventional soil mapping in this region costly and time-consuming (McBratney, Mendonça Santos, and Minasny, 2003). In this context, digital soil mapping has emerged as a feasible alternative.

In recent decades, a large number of studies have explored the use of supervised classification through machine learning techniques to improve existing soil maps (Kempen et al., 2009; Lamichhane, Kumar, and Adhikari, 2021), as well as the reference-area approach proposed by Favrot (1989), aiming to extrapolate known soil classes distribution to other locations within a region. This approach enables modeling capable of predicting the spatial variability of potential soil classes or mapping units (Arruda et al., 2016; Rodrigues et al., 2025; Wolski et al., 2017).

1.1.1 General hypothesis

The central hypothesis is that soil knowledge is spatially transferable within the Amazonian environment to data-scarce areas through ensemble learning approaches that incorporate regional environmental covariates—based on the SCORPAN framework—to capture complex soil–landscape relationships in hydromorphic environments.

By employing machine learning techniques, the SCORPAN model, environmental covariates, and performance and uncertainty metrics—including external validation of predictive models—the following chapters aim to improve legacy soil maps. The integration of tacit knowledge (intrinsic pedological expertise) with pedometric techniques within the framework of digital soil mapping can serve as a valuable resource for planning and managing sustainable soil use in the Amazon.

1.1.2 Objectives

The main objectives of this thesis were: (i) to evaluate the internal validation performance of five machine learning algorithms—Random Forest, Ranger, XGBoost, C5.0, and an ensemble learning approach—in digital soil mapping of the municipality of Tracuateua, using a legacy map and environmental covariates related to local soil-forming factors; (ii) to assess the extrapolation of Mapping Units (MUs) to other locations within the region through external validation; and (iii) to evaluate the potential of the ensemble learning approach for predicting MUs under hydromorphic conditions. These objectives resulted in two articles: (a) the Integration of tacit knowledge and artificial intelligence for digital soil mapping in the Eastern Amazonia, focusing on ensemble learning, model performance, and the incorporation of uncertainty; and (b) the Transferring soil–landscape relationships in data-limited regions of the amazon: a knowledge-guided digital soil mapping approach based on reference areas.

1.1.3 Expected impact on society

The results of this thesis may contribute significantly to soil sustainability in the Amazon, through agricultural planning and land management. A more detailed understanding of the spatial variability of soil and the composition of mapping units enables more appropriate land-use management and water conservation. The integration of legacy maps with predictive modeling, combined with the reference-area approach, has the potential to reduce gaps in detailed soil information across the Amazon.

Finally, the results obtained in this study can support regional soil surveys, particularly by helping to reduce the disparity between soil surveys and cartographic updates conducted in the South and Southeast regions and those in North Brazil, such as the Amazon region.

1.2 Review

1.2.1 Conventional soil mapping

Conventional soil mapping consists of empirical, intuitive, and deterministic mental models developed by soil scientists to represent the geographic distribution of the complex natural soil system across the landscape (Dijkerman, 1974). This classical approach relies heavily on the surveyor's tacit knowledge regarding soil–landscape relationships and is formally grounded in the interaction of the five soil-forming factors established by Dokuchaev (1949) and formalized by Jenny (1941): climate, organisms, relief, parent material, and time.

The fundamental importance of conventional soil surveys lies in their capacity to identify, interpret, and separate complex soil systems into distinct, homogeneous mapping units that delineate both the potentials and environmental limitations of the land. By doing so, these traditional surveys provide the essential spatial and taxonomic baseline required to guide the sustainable use, management, and conservation of soil as a vital natural resource.

The conventional mapping workflow is applied through a structured sequence divided into three main phases: i) Office Work (Pre-field): The initial compilation and synthesis of all available cartographic information for the target area to be mapped; ii) Fieldwork: The direct, empirical observation of landscape features—such as relief, topography, vegetation types, and land use—to identify soil boundaries and the opening of soil pits and collection of physical and chemical samples for laboratory analysis; e iii) Interpretation and Synthesis (Post-field): The integration of field and laboratory data to interpret the predominant taxonomic units that compose each mapping unit, culminating in the formal preparation of the final soil map, descriptive legends, and technical reports (Santos et al., 1995).

1.2.2 Tacit knowledge

In the context of predictive digital soil mapping, tacit knowledge refers to the empirical, implicit, and specialized expertise accumulated by soil scientists over years of conventional field surveys and mapping. It is formally described as a qualitative "mental model" of soil–landscape relationships, through which experts inherently decode and delineate the occurrence of specific soil classes by interpreting subtle visual and spatial cues from topography, vegetation, and geology (Hudson, 1992).

The integration of this human expertise is of paramount importance because modern pedometrics, despite its heavy reliance on mathematical algorithms and quantitative environmental covariates, cannot operate in a pedological vacuum. Excluding tacit knowledge from data structuring can result in models that are statistically validated yet pedologically incoherent, completely failing to capture the true geographic logic and genetic factors of soil distribution across the landscape. Bridging this gap is fundamental to converting subjective empirical models into an objective, quantitative soil–landscape framework endowed with both spatial rigor and mathematical reproducibility (Rentschler and Scholten, 2025; Wadoux et al., 2019).

The major challenge in contemporary digital soil mapping lies in translating, formalizing, and codifying this implicit knowledge into explicit digital rules that can be successfully assimilated by machine learning algorithms. In practical workflows, this digitalization of human expertise is applied through three main technical bridges: i) stratified sampling design: Structuring field sampling strategies based on the conditioning boundaries of legacy soil maps (McBratney et al., 2003); e ii) covariates selection: the careful and deliberate selection of terrain and environmental attributes tailored specifically to the local pedogenetic context (Lima et al., 2013).

1.2.3 Pedometrics

Pedometrics is a fundamental and interdisciplinary branch of soil science formally structured upon the foundational contributions of Webster (1994), McBratney et al. (2003), and Hengl (2003). Its core foundation consists in the mathematical and statistical decoding of soil genesis, spatial distribution, and landscape variability, representing a paradigm shift from the qualitative mental models of classical pedology (Bockheim et al., 2005) toward a quantitative and fully reproducible numerical approach.

The scientific importance of pedometrics lies in its capacity to substitute empirical subjectivity with statistical rigor. By providing a structured, data-driven alternative to traditional approaches, it allows for the objective quantification of soil patterns (Lima et al., 2013). It establishes the mathematical validation necessary to understand how soils vary continuously across space and time, thereby reducing mapping uncertainties and ensuring that soil science operates as a highly verifiable and reproducible discipline (McBratney et al., 2019; Wadoux et al., 2021).

Pedometrics is applied through the rigorous integration of computational modeling, applied statistics, and geoinformation sciences. It serves as the driving theoretical and methodological engine that directly underpins digital soil mapping (Hengl, 2003). In practice, its techniques are deployed to process high-dimensional environmental covariates, run spatial interpolation algorithms, optimize sampling designs, and compute predictive models that translate field observations into continuous digital soil maps (Lima et al., 2013).

1.2.4 Digital Soil Mapping

Considered an evolution of conventional soil survey methods, Digital Soil Mapping integrates quantitative techniques (mathematical and statistical) and geoinformation to accurately characterize and evaluate soil classes or attributes (Lima et al., 2013; Minasny and McBratney, 2016). The foundational framework of digital soil mapping is the SCORPAN model, proposed by McBratney et al. (2003), which mathematically relates soil properties and classes to environmental covariates. The acronym SCORPAN represents the main soil-forming factors and spatial predictors: soil (S), climate (C), organisms (O), relief (R), parent material (P), age or time (A), and spatial position (N).

The core importance of Digital Soil Mapping lies in its capacity to modernize and update the subjective, conceptual models of conventional soil mapping by leveraging contemporary quantitative techniques and the full range of available digital environmental data (McBratney et al., 2003). Supported by advances in Geographic Information Systems (GIS), remote sensing, digital environmental layers, and geostatistical modeling, Digital Soil Mapping significantly increases the effectiveness, reproducibility, and accuracy of spatial soil predictions across diverse scales.

The Predictive Digital Soil Mapping (PDSM) are deployed through two main practical approaches: a) legacy data upgrading: enhancing the spatial and taxonomic resolution of existing legacy soil maps; e b) spatial extrapolation with reference area: Predicting soil classes or attributes in data-scarce regions where no previous pedological surveys are available (Arruda et al., 2016; Lamichhane et al., 2021). Originally proposed by Favrot (1989), this method consists in extrapolating identified soil-landscape relationships from a detailed reference area to physiographically similar regions (Lagacherie et al., 1595).

In the literature, the integration of PDSM with this reference-area approach is executed through various predictive modeling workflows, including classification trees (Grinand et al., 2008), artificial neural networks (Arruda et al., 2016), and advanced semi-supervised learning frameworks (Taghizadeh-Mehrjardi et al., 2022).

1.2.5 Artificial Intelligence (AI) and Machine Learning (ML)

Artificial Intelligence (AI) serves as a broad computational framework designed to simulate human cognitive processes, within which Machine Learning (ML) acts as a specialized subset focused on pattern recognition and statistical learning directly from data. Grounded in mathematical optimization, supervised ML algorithms excel in Earth sciences by establishing complex, non-linear functions that link spatial environmental predictors (SCORPAN covariates) to field soil observations.

Within this framework, the algorithms commonly applied to supervised soil classification operate under distinct mathematical principles:

a) Random Forest (RF) and Ranger: Shared tree-based ensembles grounded in bagging (Bootstrap Aggregating) and feature bagging, where independent decision trees are trained on random data subsets with replacement, and final predictions are obtained by majority voting (Breiman, 2001; Wright and Ziegler, 2017);

b) Extreme Gradient Boosting (XGBoost): An architecture based on gradient-boosted decision trees constructed iteratively and sequentially, where each new tree is designed specifically to correct the residual errors of previous ones based on the difference between predicted and observed values (Chen and Guestrin, 2016); e

c) C5.0: A decision tree algorithm designed to generate classifiers focused on reducing misclassification costs rather than simply minimizing the overall error rate, capable of handling noisy and incomplete data (Quinlan, 2014; Arif, 2018).

In PDSM, predictive modeling relies on these data-driven approaches to identify and represent the relationships between soil classes and environmental covariates (McBratnet et al., 2003). However, the selection and execution of these algorithms depend on specific technical trade-offs, advantages, and drawbacks:

RF and Ranger: Both perform exceptionally well with high-dimensional datasets without requiring complex covariate preprocessing. While RF is implemented via the "randomForest" R package, Ranger operates through the "ranger" package as a fast implementation optimized for multi-core parallel computing, making it highly scalable for massive raster matrices. However, they demand high memory costs as the forest grows, offer reduced direct visual interpretability, and possess an inherent mathematical limitation in extrapolating values outside the boundaries of the training domain (Seyedrahimi-Niaraq; Mahdiyanfar and Olyaee, 2025).

XGBoost: It offers native handling of missing values that bypasses the need for mandatory imputation routines and supports parallel and distributed computing. Nonetheless, its sequential tree-building nature prevents full parallelization at the node-construction level, increasing computational demands. It is also highly sensitive to hyperparameter tuning (e.g., learning rate and maximum depth), exhibiting a high risk of overfitting if poorly configured or applied to heavily imbalanced classes (Natekin and Knoll, 2013; Seyedrahimi-Niaraq; Mahdiyanfar and Olyaee, 2025).

C5.0: It operates with impressive processing speed and low memory consumption. A distinct advantage for pedological studies is its unique ability to extract explicit, human-readable decision rules directly from the tree structure, which directly aids in interpreting soil–landscape relationships. However, it is strictly focused on categorical classification, lacks robust native support for continuous regression, and suffers from structural instability, meaning minor variations in the training data can drastically alter branch configurations unless its internal boosting feature is activated (Wadoux; Minasny; McBratney, 2020).

The predictive performance of digital soil models varies significantly according to factors such as landscape complexity, sampling strategy, algorithm architecture, and the quality and representativeness of covariates (Arrouays et al., 2020; Farias et al., 2026). Therefore, selecting appropriate machine learning algorithms is a critical step for improving prediction accuracy and reducing spatial uncertainty.

1.2.5.1 Ensemble Learning

In Predictive Digital Soil Mapping (PDSM), while models are traditionally implemented using a single machine learning algorithm, ensemble learning constitutes an advanced framework that integrates diverse mathematical logics to generate a combined prediction. It encompasses three main algorithmic architectures: bagging, boosting, and stacking (Adeniyi et al., 2023). The core mechanism relies on mitigating the individual weaknesses of standalone models by combining their outputs through aggregation methods, such as majority vote (hard voting) or probability averaging (soft voting) (Odegua, 2019).

Operationally, the framework is applied by concurrently deploying distinct algorithmic approaches—such as bagging and boosting simultaneously—and combining them via a final meta-regressor. However, its implementation demands a high computational cost, requiring separate training phases for all base models alongside strict cross-validation workflows to prevent data leakage. Furthermore, because of this multi-layered architectural complexity, the approach suffers from an amplified "black-box" effect, which severely complicates the direct physical interpretation of soil covariates across the study area.

Despite its operational and interpretative challenges, ensemble learning is highly critical for digital soil mapping because it significantly enhances predictive stability by reducing both model bias and variance (Qiu et al., 2026). This mathematical synergy makes the approach exceptionally proficient at managing class imbalances and ensuring continuous spatial representation across complex landscapes, successfully overcoming the structural and predictive limitations inherent to standalone algorithms (Tajik; Ayoubi; Zeraatpisheh, 2020).

1.2.6 Reference Area

The concept of the reference area, originally introduced by Favrot (1989), defines a small natural region where dominant soil classes and their respective soil–environment relationships are thoroughly known and structured as cartographic rules. This approach is based on the premise that spatial distribution patterns of soils across the landscape are recurrent and identifiable, allowing the assumption that a purposely and carefully chosen representative area is fully capable of capturing the environmental signatures required to identify soil classes and determine their spatial relationships over a much larger geographic extent (Lagacherie et al., 1995).

In practice, the application of this method consists in transferring pedological knowledge and predictive models calibrated within this reference zone to unsurveyed sectors sharing the same physiographic domain. This spatial extrapolation is widely documented in the literature through the integration of reference areas with data mining tools and machine learning algorithms, such as classification trees, artificial neural networks, Random Forest, and fuzzy logic expert systems (Arruda et al., 2016; Grinand et al., 2008). The strategic importance of this approach lies in its capacity to optimize, cost-reduce, and accelerate the execution of soil surveys in data-scarce regions, overcoming the historical reliance on the surveyor's exclusive tacit knowledge by digitally formalizing human expertise for continuous soil prediction across the landscape.

1.2.7 Error, Accuracy and Uncertainty

Measures of error, accuracy, and uncertainty assess the reliability of predictive models. The data validation and map quality are defined by three distinct core parameters: i) Error: Fundamentally defined as the difference between the predicted soil class and its true value observed directly in the field; ii) Accuracy: measures the agreement between the predicted soil property and the actual observed value; e iii) Uncertainty: Provides a probabilistic assessment of the reliability of spatial predictions across the entire surveyed area (Lilburne et al., 2024).

The transition from a subjective, qualitative mental model to a predictive, quantitative framework is crucial because it introduces objectivity into soil surveys, offering the distinct advantage of quantifying spatial uncertainty (Braga, 2013; Lepsch, 2011). When model precision is high, it indicates a robust capacity to capture the underlying spatial structure of soil–landscape relationships (Carvalho Monteiro et al., 2023). Furthermore, understanding and constraining model uncertainty is essential for determining map reliability, where the integration of expert pedological knowledge—or tacit knowledge—serves as a powerful tool to bound and mitigate prediction variances across complex environments (Schmidinger and Heuvelink, 2023).

Operationally, the evaluation of predictive modeling performance and map quality in this study was executed through a comprehensive set of spatial metrics: Performance and Accuracy Assessment: evaluated based on the derivation of a validation confusion matrix, the Kappa index, and overall accuracy metrics (Poppiel et al., 2019); e Spatial Uncertainty and Quality Mapping: assessed and mapped continuously across the landscape using the confusion index (Esfandiarpour-Boroujeni et al., 2020), Shannon entropy (Rossiter and Poggio, 2025), and map purity spatial metrics (Kempen et al., 2012).

REFERENCES

ADENIYI, Odunayo David et al. Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. **Land**, v. 12, n. 2, 15 fev. 2023.

ARIF, Muhammad. Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review. **International Journal of Database Theory and Application**, v. 11, n. 1, p. 1–8, 31 mar. 2018.

ARROUAYS, Dominique et al. Impressions of digital soil maps: The good, the not so good, and making them ever better. **Geoderma Regional**, v. 20, p. e00255, 1 mar. 2020.

ARRUDA, Gustavo Pais De et al. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, v. 73, n. 3, p. 266–273, jun. 2016.

ARRUDA, Gustavo Pais De; DEMATTÊ, José Alexandre M.; CHAGAS, César Da Silva. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. **Revista Brasileira de Ciência do Solo**, v. 37, n. 2, p. 327–338, abr. 2013.

BOCKHEIM, J. G. et al. Historical development of key concepts in pedology. **Geoderma**, v. 124, n. 1, p. 23–36, 1 jan. 2005.

BRAGA, A. **Mapeamento de solos do tradicional ao digital - Doc 316**. Embrapa Cerrados, , 2013. Disponível em: <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1005933/1/doc316.pdf>. Acesso em: 15 abr. 2026

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001.

CARVALHO MONTEIRO, Maria Eduarda et al. Digital mapping of soil classes in Southeast Brazil: environmental covariate selection, accuracy, and uncertainty. **Journal of South American Earth Sciences**, v. 132, p. 104640, 1 dez. 2023.

CAVARARO, R. **Manual técnico de pedologia**. 3. ed. Rio de Janeiro: IBGE, Coordenação de Recursos Naturais e Estudos Ambientais, 2015.

CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: A Scalable Tree Boosting System**. In: 13 ago. 2016. Disponível em: <http://arxiv.org/abs/1603.02754>. Acesso em: 14 mar. 2026

DIJKERMAN, J. C. Pedology as a science: The role of data, models and theories in the study of natural soil systems. **Geoderma**, v. 11, n. 2, p. 73–93, 1 mar. 1974.

DOKUCHAEV, V. **Russian chernozem: complete set of works**. Moscow. 1883.

ESFANDIARPOUR-BOROUJENI, I. et al. Assessment of different digital soil mapping methods for prediction of soil classes in the Shahrekord plain, Central Iran. **Catena**, v. 193, p. 104648, 1 out. 2020.

FARIAS, Fabrício do Carmo et al. Digital soil mapping in the Eastern Amazon: algorithm performance and classification uncertainty in hydromorphic and well-drained landscapes. **Soil Research**, v. 64, n. 4, p. SR25209, 11 maio 2026.

FAVROT, J. **Une stratégie d'inventaire cartographique à grande échelle: la méthode des secteurs de référence**. In: 1989. Disponível em: <https://www.semanticscholar.org/paper/Une-strat%C3%A9gie-d'inventaire-cartographique-%C3%A0-grande-Favrot/ff344adb072f3afcbdb5c1d319cb43abe02d4037>. Acesso em: 14 abr. 2026

GIASSON, Elvio et al. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, v. 63, p. 262–268, 2006.

GRINAND, Clovis et al. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, v. 143, n. 1, p. 180–190, 15 jan. 2008.

HENGL, T. **Pedometric mapping: bringing the gaps between conventional and pedometric approaches**. Enschede: International Institute for Geo-Information Science and Earth Observation (ITC), 2003. 214 p.

HUDSON, Berman D. The Soil Survey as Paradigm-based Science. **Soil Science Society of America Journal**, v. 56, n. 3, p. 836–841, 1992.

JENNY, H. **Factors of soil formation**. New York: 281 p. 1941.

KEMPEN, B. et al. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. **Soil Science Society of America Journal**, v. 76, n. 6, p. 2097–2115, 2012.

KEMPEN, Bas et al. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. **Geoderma**, 2009.

LAGACHERIE, P.; LEGROS, J. P.; BURFOUGH, P. A. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. **Geoderma**, v. 65, n. 3, p. 283–301, 1 mar. 1995.

LAMICHHANE, Sushil; KUMAR, Lalit; ADHIKARI, Kabindra. Updating the national soil map of Nepal through digital soil mapping. **Geoderma**, 2021.

LEPSCH, Igo Fernando. **19 Lições de pedologia**. São Paulo: Oficina de Textos, 2011.

LILBURNE, Linda et al. Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids. **Geoderma**, v. 450, p. 117052, 1 out. 2024.

LIMA, L. A. de S. et al. **Mapeamento de solos: do tradicional ao digital**. Planaltina, DF: Embrapa Cerrados, 2013. 41 p. (Documentos, 316). Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1005933/1/doc316.pdf. Acesso em: 24 maio 2026.

MACHADO, Diego Fernandes Terra et al. Soil type spatial prediction from Random Forest: different training datasets, transferability, accuracy and uncertainty assessment. **Scientia Agricola**, v. 76, p. 243–254, 2019.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1, p. 3–52, 1 nov. 2003.

MCBRATNEY, Alex; DE GRUIJTER, Jaap; BRYCE, Alisa. Pedometrics timeline. **Geoderma**, v. 338, p. 568–575, 15 mar. 2019.

MELLO, Fellipe et al. Expert-based maps and highly detailed surface drainage models to support digital soil mapping. **Geoderma**, v. 384, p. 114779, 15 fev. 2021.

MINASNY, Budiman; MCBRATNEY, Alex. B. Digital soil mapping: A brief history and some lessons. **Geoderma**, Soil mapping, classification, and modelling: history and future directions. v. 264, p. 301–311, 15 fev. 2016.

NATEKIN, Alexey; KNOLL, Alois. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, 2013.

ODEGUA, R. **An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking)**. : IndabaXAt. Proceedings of the conference on deep learning. Anais... . , 2019.

OLIVEIRA JUNIOR, R. C. de et al. **Zoneamento agroecológico do município de Tracuateua**, Estado do Pará. 1999.

POLIDORO, J. C. et al. **Programa Nacional de Solos do Brasil (PronaSolos)**. 2016.

POPPIEL, Raúl R. et al. Pedology and soil class mapping from proximal and remote sensed data. **Geoderma**, v. 348, p. 189–206, 15 ago. 2019.

QUINLAN, J. Ross. C4.5: Programs for Machine Learning. [S.l.]: **Elsevier**, 2014.

RENTSCHLER, Tobias; SCHOLTEN, Thomas. A Note on Spurious Correlations and Explainable Machine Learning in Digital Soil Mapping. **European Journal of Soil Science**, v. 76, n. 4, p. e70172, 2025.

RIZZO, Rodnei et al. Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in southeastern Brazil. **Geoderma**, 2020.

RODRIGUES, Hugo et al. autoRA: An Algorithm to Automatically Delineate Reference Areas—A Case Study to Map Soil Classes in Bahia, Brazil. **Land**, v. 14, n. 3, 12 mar. 2025.

ROSSITER, David G.; POGGIO, Laura. Representing soil landscapes from digital soil mapping products – helping the map to speak for itself. **SOIL**, v. 11, n. 2, p. 849–881, 21 out. 2025.

SANTOS, H. G. dos et al. **Procedimentos normativos de levantamentos pedológicos**. [S.l.]: Brasília, DF: EMBRAPA-SPI; Rio de Janeiro: EMBRAPA-CNPS, 1995., 1995.

SCHMIDINGER, Jonas; HEUVELINK, Gerard B. M. Validation of uncertainty predictions in digital soil mapping. **Geoderma**, v. 437, p. 116585, 1 set. 2023.

SEYEDRAHIMI-NIARAQ, Mirmahdi; MAHDIYANFAR, Hossein; OLYAEE, Mohammad hossein. Comparison of the performance of gradient boost, linear regression, decision tree, and voting algorithms to separate geochemical anomalies areas in the fractal environment. **Artificial Intelligence in Geosciences**, v. 6, n. 2, p. 100156, 1 dez. 2025.

SILVA, Cristiano Cassiano da et al. Mapeamento pedológico digital da folha Botucatu (SF-22-Z-B-VI-3): treinamento de dados em mapa tradicional e validação de campo. **Revista Brasileira de Ciência do Solo**, v. 37, p. 846–857, 2013.

SILVA, Sérgio Henrique Godinho et al. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. **Geoderma**, v. 267, p. 65–77, 1 abr. 2016.

TAGHIZADEH-MEHRJARDI, Ruhollah et al. Semi-supervised learning for the spatial extrapolation of soil information. **Geoderma**, v. 426, 2022.

TAJIK, Samaneh; AYOUBI, Shamsollah; ZERAATPISHEH, Mojtaba. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. **Geoderma Regional**, v. 20, p. e00256, 1 mar. 2020.

TEN CATEN, Alexandre et al. Regressões Logísticas Múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v. 35, n. 1, p. 53–62, fev. 2011.

VILLELA, André Luis Oliveira. **Mapeamento digital de solos da formação solimões sob floresta tropical amazônica**. Rio de Janeiro: UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO, ago. 2023.

WADOUX, Alexandre M. J. C. et al. A note on knowledge discovery and machine learning in digital soil mapping. **European Journal of Soil Science**, v. 71, n. 2, 2019.

WADOUX, Alexandre M. J. C.; MINASNY, Budiman; MCBRATNEY, Alex B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, p. 103359, nov. 2020.

WADOUX, Alexandre M. J. C. et al. Ten challenges for the future of pedometrics. **Geoderma**, v. 401, p. 115155, 1 nov. 2021.

WEBSTER, R. The development of pedometrics. **Geoderma**, v. 62, n. 1-3, p. 1-15, 1994.

WOLSKI, Mario Sergio et al. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 633–642, ago. 2017.

WRIGHT, Marvin N.; ZIEGLER, Andreas. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 77, n. 1, 2017.

CHAPTER 2

2 INTEGRATING TACIT KNOWLEDGE AND AI FOR DIGITAL SOIL MAPPING IN EASTERN AMAZONIA: ENSEMBLE LEARNING, MODEL PERFORMANCE, AND UNCERTAINTY INCORPORATION

Published at: Soil Systems

Available at: https://doi.org/10.3390/soilsystems10030041

2.1 Abstract

Predictive Digital Soil Mapping (PDSM) in Eastern Amazonia faces challenges due to its environmental complexity, difficult access, and scarce legacy data. While legacy soil maps contain valuable tacit knowledge, updating them requires methods that can handle uncertainty. This study evaluates the integration of old soil maps with machine learning to update soil information in Tracuateua, Pará, with a specific focus on the performance of ensemble learning and the explicit incorporation of uncertainty metrics in soil mapping units under hydromorphic influence, which, in addition to being difficult to access, are influenced by complex pedogenetic processes. We combined 270 sampling points, equivalent to the total pixels that captured the variability of soil mapping units, with environmental covariates and historical data. Several algorithms were tested, including an ensemble approach, to predict mapping units and quantify uncertainty through entropy and confusion indices. The ensemble model demonstrated improved stability and reduced classification uncertainty compared to single models, particularly in challenging hydromorphic environments. Although accuracy gains were modest, the models captured soil–environment relationships, with climate as: Annual Mean Temperature 22,000 years ago (Tmean_22k), relief: Channel Network Base Level (CNBL and altitude) and organism variables: Land Surface Temperature (LST) emerging as the main predictors. Spatialized uncertainty estimates, expressed through entropy and the confusion index, provide a practical decision-support tool for guiding field surveys and identifying areas of low mapping reliability. By explicitly transferring the pedologist's mental model—encoded as tacit knowledge in legacy soil maps—into ensemble learning, this study presents a robust and transferable framework for updating soil maps in data-scarce tropical regions, balancing predictive performance, spatial consistency, and uncertainty-aware interpretation.

Keywords: machine learning, SCORPAN, soil survey, hydromorphic pedoenvironments, knowledge-discover, tropical soils

2.2 Introduction

Knowledge of the spatial distribution of soils is of fundamental importance for supporting the management of environmental resources such as the atmosphere, river basins, and soil itself (Lamichhane; Kumar; Adhikari, 2021; Moquedace et al., 2024). In tropical regions such as the Amazon biome, soil surveys can provide essential information for the development of projects and the planning of soil use, management, and conservation (Pavão et al., 2024). However, maps with sufficient detail to guide decision-making in the Amazon are still scarce, and the lack of high-resolution information remains one of the main limitations for sustainable agricultural planning and the implementation of conservation policies in this biome.

In Brazil, soil surveys at a scale of 1:100,000 or larger cover less than 6% of the entire territory, according to information from the National Soil Program (Pronasolos). Furthermore, in the Brazilian Amazon, which covers approximately 5 million km2, the vast majority of available surveys are at a smaller, exploratory, or low-intensity reconnaissance scale (1:1,000,000 to 1:250,000). Among the challenges hindering the slow progress of soil surveys in this biome since the late 1980s, when the RADAMBRASIL project was completed, are the vast territory and poor accessibility (except via major waterways) (Fernandes Filho et al., 2023), which demands time and considerable availability of human and financial resources (Polidoro et al., 2016).

In this context, many studies have addressed the possibility of not only increasing the level of detail in legacy maps using PDSM techniques (Carvalho Monteiro et al., 2023), but also incorporating uncertainty into these estimates (Lamichhane; Kumar; Adhikari, 2021). The formalization of the PDSM, based on the SCORPAN approach (McBratney; Mendonça Santos; Minasny, 2003), is performed by adding the spatial position (n) to the soil formation factors formulated by (Jenny, 1941). It allows for the prediction of soil distribution patterns or attributes in the landscape based on their relationship with environmental covariates through mathematical and statistical methods, generating a spatial model (Carvalho Monteiro et al., 2023).

Thus, PDSM consists of developing an empirical predictive model using statistical and mathematical methods, which combines, in short, environmental information and the tacit knowledge of soil scientist, the latter materialized in conventional soil survey maps (Lagacherie; McBratney, 2006). This predictive model, however, is also capable of learning the implicit uncertainty of soil classes from legacy maps (Wadoux et al., 2020). Considering that the uncertainty of predictions may be related, among other causes, to the quality of the database and the variability of the soil, in the pixel-by-pixel conversion of a soil mapping unit, in which the target is classified according to its similarity to the object, it is possible to evaluate the degree of certainty in a pixel classification, which is associated with a specific class or attribute of the soil (Qi; Zhu, 2011).

The possibility of improving mapping quality based on legacy maps has been addressed in a literature review of 90 publications related to digital land mapping and modeling (Grunwald, 2009) and in a study (Lin et al., 2025) exploring the potential of machine learning to accelerate the mapping of parental soil material in the northern region of Europe. In conventional mapping, among soil formation factors, relief is the primary parameter used by soil scientists to mentally disaggregate the landscape. Consequently, the establishment of soil–landscape relationships strongly depends on the expert's tacit knowledge and subjective interpretation of morphogenetic criteria (Wadoux; Heuvelink, 2023; Zhu et al., 2024). Mukumbuta et al. (2022) also highlights how protocols such as the Cornell method, along with other systematic evaluation approaches, reveal the subjectivity and variability in the quality of historical information contained in conventional soil survey maps, whose cartographic generalization results in greater spatial uncertainty (Bazaglia Filho et al., 2013).

When combined field pedological knowledge and important pedogenetic predictive covariates (Wadoux et al., 2021), PDSM becomes a more efficient, reproducible, and economical approach to reduce soil information gaps, like in the Amazon biome, where detailed scale soil surveys are widely scarce and economically challenging in the context of the territorial dimension of the Brazilian Amazon. This approach helps to overcome the limitations associated with the subjectivity of tacit knowledge applied in conventional soil surveys (Kempen et al., 2012) and has shown particular promise in addressing the challenges posed by conventional maps, whose spatial accuracy often does not correspond to their nominal publication scale, thus limiting their usefulness for detailed assessments (Mukumbuta et al., 2022).

The use of machine learning-based modeling techniques—integrating field samples, pedological knowledge, and interrelated environmental variables—constitutes the conceptual core of PDSM (Poggio et al., 2021). Among the algorithms most commonly applied in pedometric studies are Random Forest, Ranger, XGBoost, and C5.0, all based on decision tree structures. These methods have become fundamental for PDSM due to their ability to model complex nonlinear relationships, handle high-dimensional predictor spaces, and capture interactions between multiple environmental covariates (Esfandiarpour-Boroujeni et al., 2020; Hengl et al., 2018; Neyestani et al., 2021). Furthermore, ensemble learning approaches have gained increasing prominence by combining predictions from multiple algorithms, resulting in more robust models with lower variance and superior predictive performance compared to individual learning (Padarian; Minasny; McBratney, 2020; Taghizadeh-Mehrjardi et al., 2021; Wadoux, 2025).

This study addresses the improvement of a soil map published in 1998 for the municipality of Tracuateua, with a scale of 1:100,000, in northeastern Pará, by integrating PDSM, environmental covariates, and multiple machine learning algorithms, including ensemble modeling. Its innovation lies in improving the detail of soil maps for the Amazon biome, focusing on soils in hydromorphic environments, known for their high pedogenetic complexity and difficulty of access. It explicitly examines the implications of using legacy maps as training labels and the spatial dependence of the models in the context of evaluating the uncertainty of the predicted maps.

Given this, our hypothesis is that ensemble modeling increases the stability of predictions and reduces the spatial uncertainty in predicting soils under hydromorphic environments compared to individual models. Therefore, the objectives of this study were (i) to compare the predictive performance of different machine learning algorithms, (ii) to analyze the effect of ensemble modeling on the stability of predictions, and (iii) to quantify and interpret the spatial uncertainty of classification, with an emphasis on soils under critical areas for field surveys such as in hydromorphic environments.

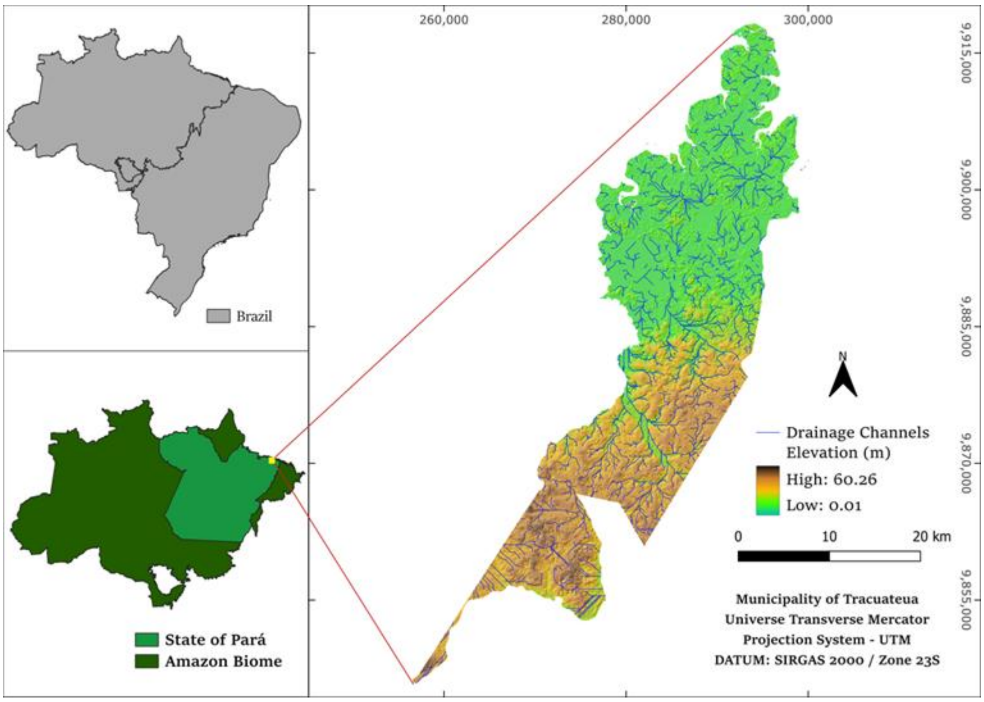
2.3 Materials and methods

2.3.1 Study area

The study area encompasses the municipality of Tracuateua, covering 828.025 km2 and located within the Amazon biome in the northeastern Pará mesoregion, between 01°02′ S latitude and 46°56′ W longitude (Figure 1). The regional climate is classified as Am - tropical monsoon - according to the Köppen system, with a mean annual precipitation of 2581 mm and a mean annual temperature of 27.8 °C (Alvares et al., 2013). This tropical climatic regime strongly influences soil weathering processes, acting in conjunction with the parent material and the variability of geomorphological positions that structure the local landscape (Pavão et al., 2024).

The study area is characterized by the predominance of sediments from two geological periods. (i) The Quaternary period is represented by recent alluvial deposits along the banks of large rivers such as the Japerica River, as well as beaches and mangroves. In these hydromorphic environments, GLEISSOLOS HÁPLICOS, NEOSSOLOS QUARTZARÊNICOS Hidromórficos, and PLINTOSSOLOS, according to the Brazilian Soil Classification System (SiBCS), represent common soil classes in Amazonian hydromorphic landscapes (floodplains, lowlands, and transition areas); and (ii) The Tertiary period is represented by the Barreiras Formation, which occupies approximately 70% of the study area and forms a lower Amazonian plateau characterized by LATOSSOLOS and ARGISSOLOS, predominate under flat to gently undulating relief. These highly weathered soils typically exhibit yellow to red hues associated with the presence of iron oxides, which exert strong control over their spectral behavior. Soil reflectance in the visible (VIS) and near-infrared (NIR) regions is primarily governed by soil color and mineralogical composition, particularly iron oxides, clay minerals, and organic matter. The dominance of hematite and goethite in these tropical soils results in characteristic spectral patterns and reflectance variations in the VIS–NIR range. In contrast, soils formed under hydromorphic conditions - where prolonged water saturation promotes iron reduction and organic matter accumulation - generally exhibit lower reflectance and weaker spectral features, often associated with grayish or bluish colors typical of gleyed environments (Baumgardner et al., 1985; Costa; Almeida, 1998; Santos et al., 2025).

Figure 1 - Location of the study area in the State of Pará, Eastern Amazon.



Source: The author (2026).

2.3.1 Data Sources

The database is composed of mapping units (MUs) defined in the conventional soil map of the municipality of Tracuateua, originally produced by (Oliveira Junior et al., 1999) at a scale of 1:100,000. This legacy soil map was published at the Great Group level, ac-cording to the SiBCS, and grouped into 9 MUs (Table 1).

To enhance the predictive performance of the PDSM models, the original MUs were reorganized into two levels of detail:

(i) LD1, the least detailed level, comprising five mapping units composed of simple taxonomic units and associations at the Soil Order level derived from the legacy map; and (ii) LD2, the most detailed level, comprising nine mapping units composed of simple taxonomic units and associations at the Great Group level present in the original legacy soil map (Figure 2a, b).

This hierarchical structuring aimed to evaluate model sensitivity under different taxonomic resolutions and to explore the effects of class aggregation on classification ac-curacy and uncertainty.

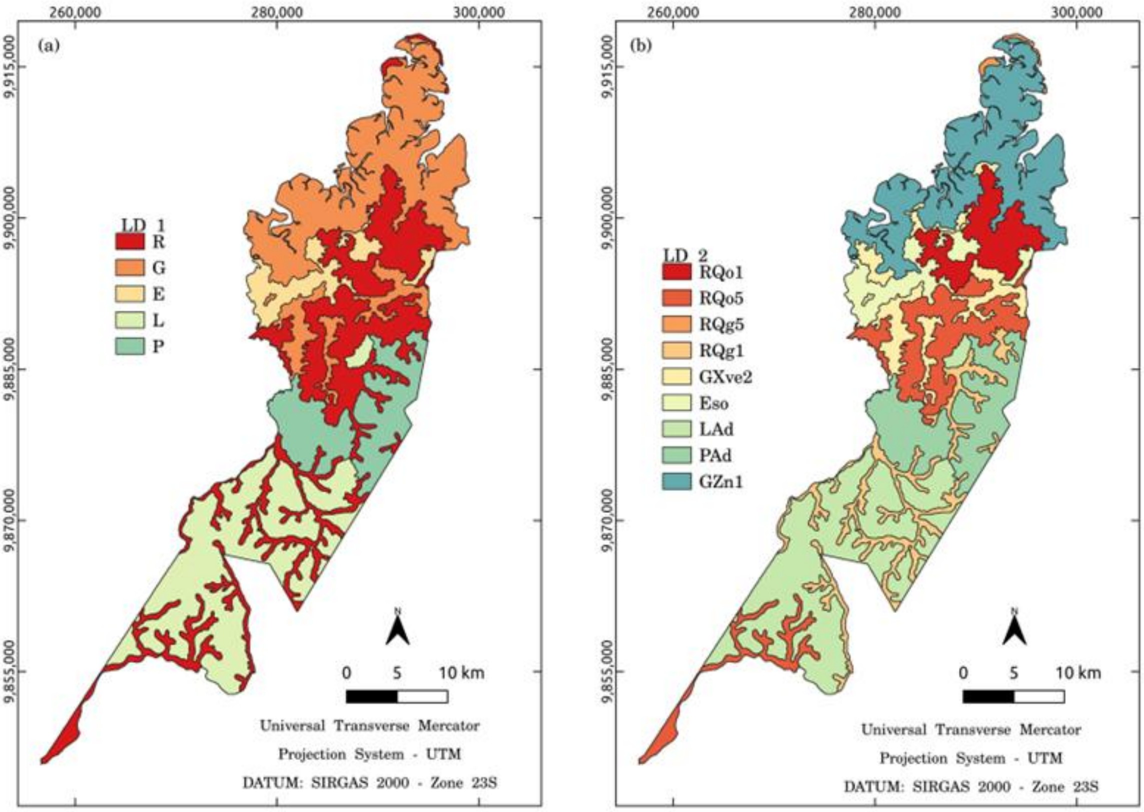
Table 1 - Mapping units (MU) from soil classes derived from the legacy map levels of detail (LD1 and LD2) and their correspondence to the WRB/FAO classification.

MU[1]	WRB/FAO	Mapping Units Adapted	Level of Detail		Area (km^2)
			LD1	LD2	
NEOSSOLOS QUARTZARÊNICOS Órticos	Arenosols	NEOSSOLOS	R	RQo1	58.18
NEOSSOLOS QUARTZARÊNICOS Órticos + ARGISSOLOS VERMELHO-AMARELOS Distróficos	Arenosols + Acrisols			RQo5	107.79
NEOSSOLOS QUARTZARÊNICOS Órticos + NEOSSOLOS QUARTZARÊNICOS Hidromórficos + GLEISSOLOS HÁPLICOS Tb Distróficos	Arenosols + Gleysols			RQg5	3.27
NEOSSOLOS QUARTZARÊNICOS Hidromórficos	Fluvisols			RQg1	74.17
GLEISSOLOS HÁPLICOS Ta Eutróficos	Gleysols	GLEISSOLOS	G	GXve2	56.18
GLEISSOLOS SÁLICOS Sódicos	Solonchaks			GZn1	177.33
ESPODOSSOLOS FERRILÚVICOS Órticos	Podzols	ESPODOSSOLOS	E	ESo	45.59
LATOSSOLOS AMARELOS Distróficos	Ferralsols	LATOSSOLOS	L	LAd	212.26
ARGISSOLOS AMARELOS Distróficos	Acrisols	ARGISSOLOS	P	PAd	102.94

Source: Oliveira Junior *et. al* (1999)

[1] SiBCS – Sistema Brasileiro de Classificação do Solo

Figure 2 - Legacy conventional soil map of the municipality of Tracuateua, State of Pará, Brazil. (a) LD1 classification scheme and (b) LD2 classification scheme.



Source: adapted from Oliveira Junior *et. al* (1999).

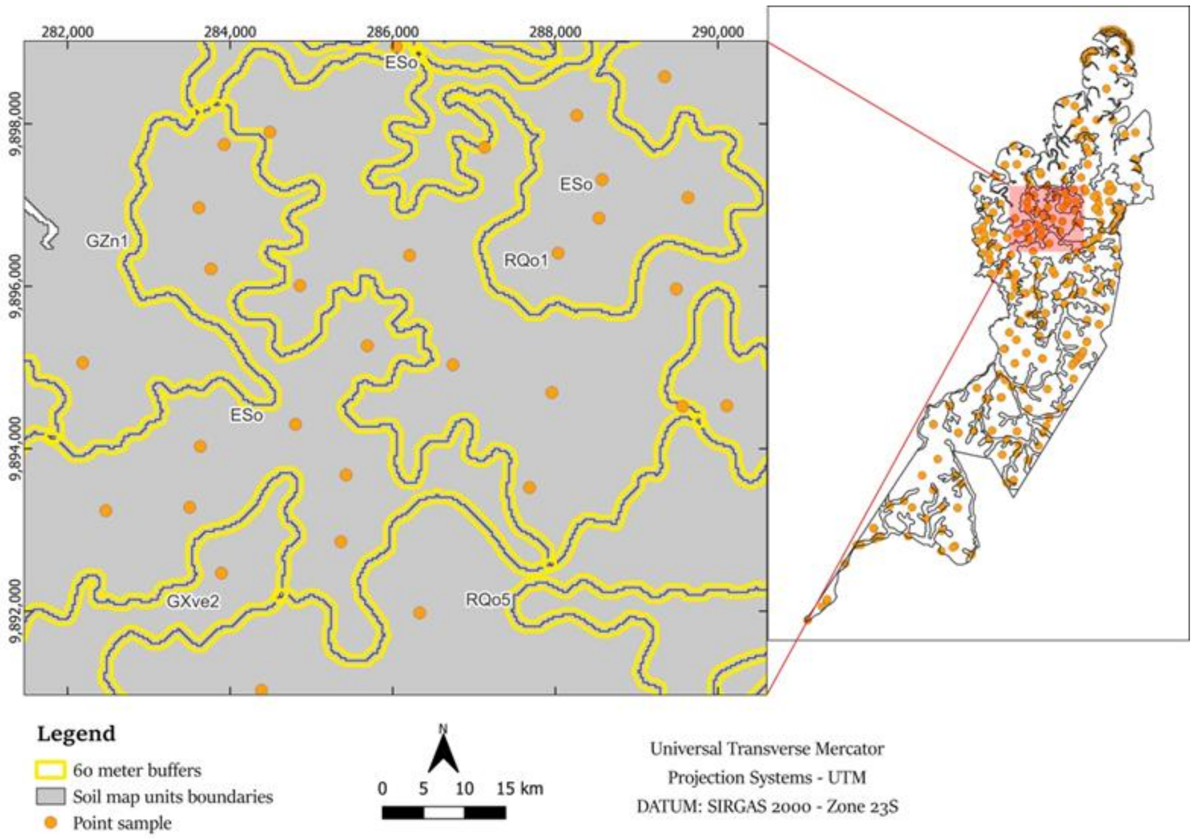
For model training, 270 pixels of 30 m × 30 m each (900 m2) were selected, aiming to guarantee the spatial variability of the pedons at both detail levels LD1 and LD2, thus covering the entire study area (828 km2). Field surveys using an auger, at a depth of 0 to 1.50 m, were carried out in the area of the aforementioned 270 samples with the objective of, in addition to confirming the MUs described in the legacy soil map (Table 2), avoiding transition zones between the MUs (Figures 3 and A1 in Appendix A), thus establishing a robust protocol to guarantee the independence between the MUs training. This method sought to overcome the challenges of sparse regional inference, which is due to the logistical challenge of soil surveying in the Amazonian environment. This approach allows for the identification of areas where the uncertainty of the prediction, such as the entropy and confusion index, may require additional field investigations.

Table 2 - Distribution of digital samples across LD1 and LD2.

LD1	No. of Digital Samples
R	109
G	62
E	33
L	36
P	30
Total	270
LD2	
RQo1	28
RQo5	36
RQg5	23
RQg1	22
GXve2	30
Eso	33
LAd	36
PAd	30
GZn1	32
Total	270

Source: The author (2026).

Figure 3 - Representation of selected pixels (point sample) of the boundaries of the soil mapping units and the 60 m strip indicating the transition zone between the soil mapping units.



Souce: The author (2026).

2.3.1.1 SCORPAN Covariates

Fifteen predictor covariates were pre-selected based on their pedological relevance to the MUs in the study area (Table 3), aiming to align machine learning techniques with knowledge discovery approaches (Wadoux et al., 2020). These covariates represent soil-forming factors following the SCORPAN framework (McBratney; Mendonça Santos; Minasny, 2003): S—soil properties or classes; C—climate; O—organisms; R—relief; P—parent material; and A—time (age). To ensure spatial consistency across all datasets, these covariates were resampled using the nearest neighbor method and standardized to a uniform spatial resolution of 30 x 30 m, matching the pixel size of the primary digital elevation model.

The climate factor was represented by annual average precipitation (AAP) and paleoclimatic variables PR_22K and Tmean_22K, derived from data estimated 22,000 years ago (Carvalho Monteiro et al., 2023). Climate is a key driver of soil formation, and tropical conditions promote the development of highly weathered soils. This data was obtained from Google Earth Engine (GEE) and WorldClim at 1–6 km resolution (Kampf; Curi, 2012; Soil Science Division Staff, 2017).

Table 3 - SCORPAN covariates used for predictive modeling based on pedological relevance.

SCORPAN Factors	Abbreviation	Description	Spatial Resolution/Scale	Source	References
Climate (C)	AAP	Average Annual Precipitation	6 km	Google Earth Engine. ImageCollection ('UCSB-CHG/CHIRPS/DAILY'). Time series from 1981 to 2023	(Taylor; Stouffer; Meehl, 2012)
	PR_22K	Annual Precipitation 22,000 Years Ago	1 km	https://geodata.ucdavis.edu/climate (Accessed on 31 July 2024)	
	Tmean_22K	Annual Mean Temperature 22,000 Years Ago	5 km	https://geodata.ucdavis.edu/climate/cmip5/2_5m/ (Accessed on 31 July 2024)	
Organism (O)	SAVI	Soil-Adjusted Vegetation	30 m	Multispectral data. LANDSAT 8—sensor OLI (U.S. Geological Survey). Equation: (NIR-RED)/(NIR + RED + L)×(1 + L)	(Huete, 1988)
	LST	Ground Surface Temperature—during the day	1 km	Google Earth Engine. ImageCollection ("MODIS/061/MOD11A1"). Time series from 2000 at 2024	(Kuenzer; Dech, 2013)
Relief (R)	CNBL	Channel Network Base Level	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Conrad et al., 2015)
	DD	Drainage Density		Copernicus GLO-30 (DSM) and ANADEM (DTM)	(Christo-foletti, 1980)
	MRRTF	Multiresolution Index of Ridge Top Flatness	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Gallant; Dowling, 2003)
	MRVBF	Multiresolution Index of Valley Bottom Flatness	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Gallant; Dowling, 2003)
	PFC	Profile Curvature	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Florinsky, 2012)
	RSP	Relative Slope Position	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Florinsky, 2012)
	Altitude	Level	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM)	(Florinsky, 2012)
Parent material (P)	Geology		1:250,000	IBGE (2023)	(Florinsky, 2012)
	IOR	Iron Oxide Ratio	30 m	Multispectral data. LANDSAT 8—sensor OLI (U.S. Geological Survey). Equation: (Band 4-Red/Band 2-Blue)	(Singer, 1981)
Age (A)	GEOM	Geomorphons	30 m	Copernicus GLO-30 (DSM) and ANADEM (DTM). SAGA GIS 9.3.1	(Jasiewicz; Stepinski, 2013)

Source: The author (2026).

For the organism factor, the Soil-Adjusted Vegetation Index (SAVI) (Mirchooli et al., 2020) and daytime Land Surface Temperature (LST) were used (Sayão; Demattê, 2018). These were obtained from LANDSAT-8 OLI multispectral imagery and the MODIS/061/MOD11A1 daily time series (2000–2024) via GEE (Gorelick et al., 2017). In the Amazon biome, organic matter inputs and high temperatures strongly influence particle aggregation, darkening of the surface horizon, water infiltration, soil erodibility, nutrient dynamics, and forest biomass (Kampf; Curi, 2012).

Terrain attributes were derived using two elevation models: the Digital Surface Model (DSM) and Digital Terrain Model (DTM). The DSM was obtained from the Copernicus DEM GLO-30 (1 Arc Second), which provides global coverage at 30 m spatial resolution, selected for its free availability, high quality, and proven accuracy in tropical environments (Bielski et al., 2024).

The DTM was derived from the South America Digital Terrain Model (ANADEM) (Laipelt et al., 2024), also at 30 m resolution. The development of ANADEM involved (i) calculating the Enhanced Vegetation Index (EVI), Normalized Difference Moisture Index (NDMI), Modified Simple Ratio (MSR), and Modified Soil-Adjusted Vegetation Index (MSAVI) from multispectral imagery acquired by the LANDSAT- 8 Operational Land Imager (OLI) and Sentinel-2 MultiSpectral Instrument (MSI); (ii) incorporating vegetation height data from the Global Ecosystem Dynamics Investigation (GEDI), collection L2A; (iii) using the DSM from Copernicus GLO-30; (iv) applying the TreeBoost machine learning algorithm to estimate bias induced by vegetation; and (v) obtaining the DTM by subtracting the vegetation-induced bias from the Copernicus DSM, accounting for correction based on the vegetation fractional cover.

Terrain covariates were derived from the DSM and DTM using the Basic Terrain Analysis module in SAGA GIS v. 9.3.1 (Olaya; Conrad, 2009), resampled to the UTM SIRGAS 2000 Zone 23S projection at 30 m resolution. These include CNBL, DD, MRRTF, MRVBF, PFC, RSP, and GEOM. Linear drainage density (DD) was calculated in ArcGIS 10.5 as the ratio of total channel length to basin area (Strahler, 1952), according to (Mello et al., 2021), who reported its significant contribution to model performance and soil map extrapolation. CNBL, DD, MRRTF, MRVBF, PFC, and RSP represent terrain surface configuration and relate to the spatial distribution of soils across landscape scales (Kampf; Curi, 2012). GEOM captures terrain form patterns, allowing for intuitive classification of topographic features and providing useful estimates of surface age within the SCORPAN framework (Jasiewicz; Stepinski, 2013; McBratney; Mendonça Santos; Minasny, 2003).

Information regarding geology and Iron Oxide Ratio (IOR) was sourced from Instituto Brasileiro de Geografia e Estatística (IBGE, 2023) and LANDSAT-8 multispectral imagery, respectively. Parent material influences key soil properties, including texture and chemical composition. The accumulation of iron oxides and hydroxides such as hematite and goethite occur in highly weathered soils and provides an estimate of the degree of soil weathering (Kampf; Curi, 2012).

2.3.1.2 Machine Learning Models

Machine learning models are capable of modeling nonlinear relationships between SCORPAN covariates, including soil classes as targets and terrain attributes, location, and temporal series as predictor variables, which can not only improve mapping accuracy but also quantify uncertainty. In this study, four tree-based algorithms were used—Random Forest (RF), Fast implementation of Random Forests (Ranger), XGBoost (Xgb), and C5.0 (C5)—as well as an ensemble learning (EL) approach combining these algorithms (Table 3). Testing different algorithms is essential for identifying the most suitable method for local conditions (Moquedace et al., 2024).

The RF (Breiman, 1996) and Ranger (Wright; Ziegler, 2017) algorithms share the following fundamental principles: (i) bagging (Bootstrap Aggregating), where each decision tree is trained on a randomly selected sample with replacement from the training set; and (ii) feature bagging, in which only a subset of features is selected for each decision tree in the forest. Final predictions are made by majority vote, as in the current classification study. Differences between the algorithms include computational efficiency, with Ranger being faster, and the use of different R packages—“randomForest” and “ranger” for RF and Ranger, respectively.

These models have three hyperparameters for optimization: (i) “ntree” (1000) for RF and “num.trees” (1000) for Ranger; (ii) “nodesize” (5); and (iii) “mtry”. Among these, “mtry” was optimized using the “caret” package, testing 10 different values ac-cording to the number of covariates, with the value yielding the best performance selected, in accordance with (Moquedace et al., 2024).

XGBoost (Xgb), developed by (Chen; Guestrin, 2016), is based on gradient-boosted decision trees. In this model, new trees are created to correct the errors of the previous tree, and model performance is assessed based on the difference between predicted and observed values. Xgb offers advantages such as improved predictive performance via gradient descent, enhanced model accuracy, computational efficiency, parallel and distributed computing, and the ability to handle missing values, making it well-suited for complex prediction problems. The algorithm was implemented using the “xgbTree” package, with the hyperparameter “nrounds” set to 1000, while other hyperparameters—“max_depth,” “eta,” “gamma,” “colsample_bytree,” “min_child_weight,” and “subsample”—were optimized using the “caret” package.

C5.0, proposed by (Quinlan, 1993), is based on a broad decision tree capable of handling noise and missing data. The algorithm generates classifiers that help reduce the cost of misclassification rather than simply minimizing the error rate (Arif, 2018). The C5.0 model includes three hyperparameters: (i) “trials” (100), (ii) “model” (tree), and (iii) “win-now” (false). The final model was implemented using the “C50” package.

In PDSM, a single machine learning algorithm is often used to implement the model. However, ensemble learning, which has the advantage of combining the strengths of different algorithms, remains underutilized. Ensemble learning encompasses three main techniques: bagging, boosting, and stacking. In this study, the bagging technique proposed by (Breiman, 1996) was employed, aiming to combine similar models (e.g., tree-based) to reduce variance and increase model robustness. Final predictions can be aggregated by majority vote (hard voting) or probability averaging (soft voting), with the latter approach applied in the present study (Odegua, 2019).

To enhance model generalizability by mitigating overfitting and underfitting, a two-step covariate selection process was adopted (Moquedace et al., 2024). First, Spearman's rank correlation coefficient was calculated, applying an absolute cutoff threshold of 0.95 via the 'findCorrelation' function from the 'caret' package (Kuhn, 2021). When pairs of covariates exhibited a correlation exceeding this threshold, the variable with the higher average absolute correlation was removed.

As the target, the dataset of 270 samples for both LD1 and LD2 (representing the soil taxonomic levels of Order and Great Group, respectively) was split into 80% for training with 5-fold cross-validation and 20% for testing. Using the “caret” package, this split was performed with the “createDataPartition” function, which executes more representative random sampling (Moquedace et al., 2024). Subsequently, to identify the optimal subset of predictor covariates for model performance, Recursive Feature Elimination (RFE) was applied as a second method for covariate selection. With the adjusted parameters, 100 iterations were run, with each generated map resulting from an independent sampling iteration. The best model was selected based on the highest performance according to the Kappa index and the importance of the respective predictor covariates in predicting the targets (Figure 4). To generate a single map, we use the mode among the 100 maps.

The parameters pre-configured before the start of the process are called hyperparameters. Their selection can drastically change the model's performance, and is essential for the reproducibility of the modeling. Therefore, in this study, we used the configurations shown in Table 4.

Table 4 - Hyperparameters used in modeling according to the algorithms, at different levels details different (LD1 and 2) using DSM and DTM.

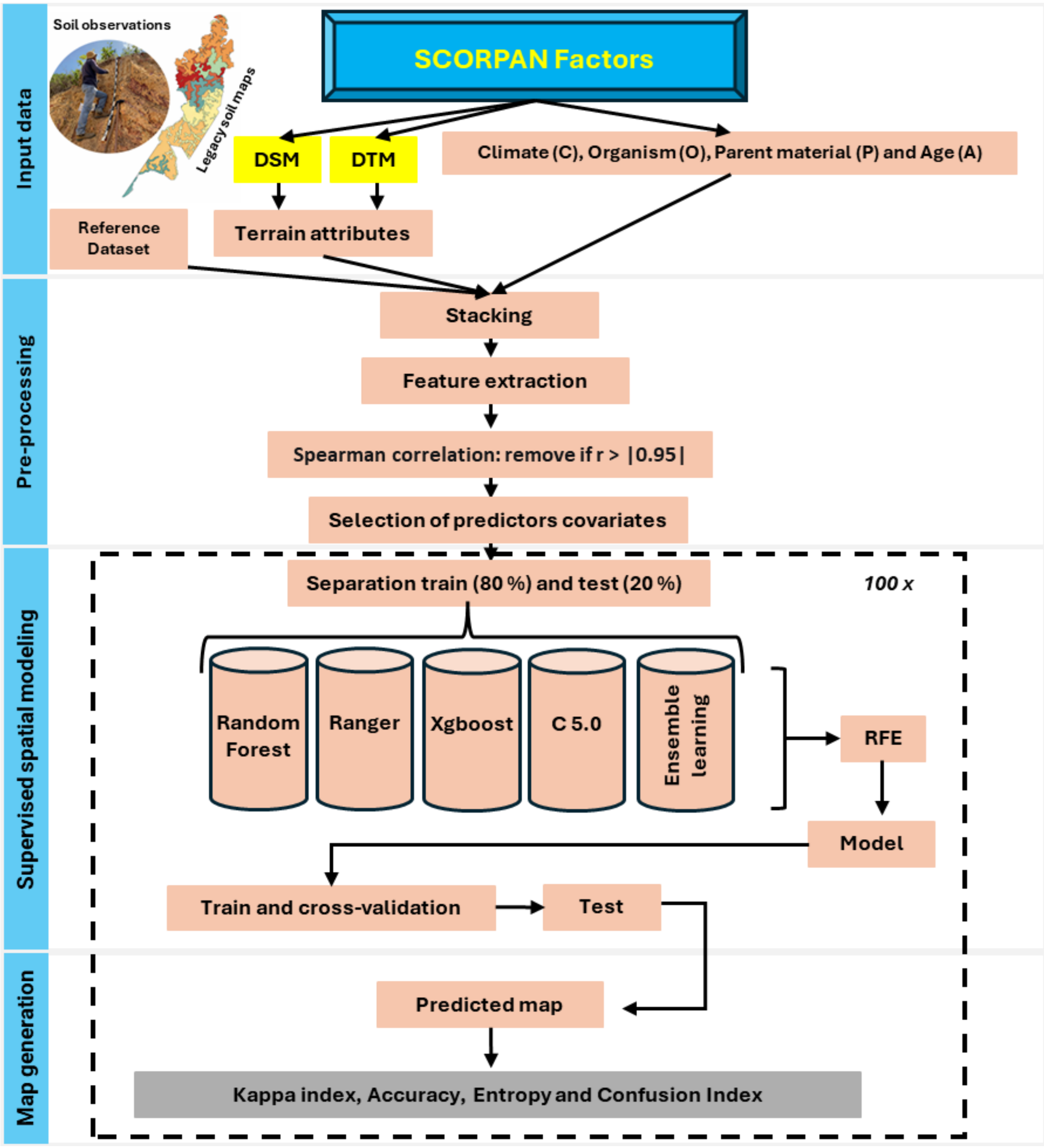
Algorithms	LD	Hyperparameters	
		DSM	**DTM**
RF	LD1	mtry = 2; ntree = 1000	mtry = 2; ntree = 1000
	LD2	mtry = 4, ntree = 1000	mtry = 6; ntree = 1000
Ranger	LD1	mtry = 10; num.trees = 1000	mtry = 4; num.trees = 1000
	LD2	mtry = 2; num.trees = 1000	mtry = 2; num.trees = 1000
Xgb	LD1	nrounds = 1000; max_deph = 6; eta = 0.01; gamma = 0; colsample_bytree = 0.8; min_child_weight = 1; subsample = 0.8	
	LD2	nrounds = 1000; max_deph = 3; eta = 0.01; gamma = 0; colsample_bytree = 0.8; min_child_weight = 1; subsample = 0.8	
C5	LD1 LD2	trials = 100. Multiple model combination (10 models C 5.0)	
EL	LD1 LD2	bestTune	

Source: The author (2026).

To represent the models used, we employed the following structure: abbreviation of the algorithm, e.g., RF, plus the origin of the terrain attributes used as environmental co-variates, either DSM (S) or DTM (T), and the level of detail of the taxonomic units and as-sociations, either LD1 or LD2. Thus, the nomenclature of the XgbS2 model, for example, means that the Xgboost algorithm was used, with terrain covariates derived from DSM, at LD2.

In this study we used Google Gemini (Free Version) to assist in refining the English translation and performing grammar checks of the manuscript text and generating the graphical abstract figure. The results from the generative artificial intelligence (GenAI) tool were used to assist only and were carefully reviewed, edited and validated by the authors. All methodological decisions, data processing, statistical analyses, interpretation of results, and scientific conclusions were performed exclusively by the authors, who take full responsibility for the content of the manuscript.

Figure 4 - Flowchart of this study.



Source: The author (2026).

2.3.2 Spatial Prediction Performance on the Test Dataset

The performance of the algorithms was evaluated using the mean Kappa index (k) and overall accuracy. The Kappa index (Equation (1)) indicates the level of agreement in predicting the mapping unit (Cohen, 1960).

$$k = \frac{n \sum_{i=1}^{C} n_{ii} - \sum_{i=1}^{C} n_{i+} + n_{+i}}{n^2 - \sum_{i=1}^{C} n_{i+} + n_{+i}} \quad \text{(Eq. 1)}$$

Where n_{ii} is the value in row i and column i; n_{i+} is the sum of row i and n_{+i} is the sum of column i in the confusion matrix; n is the total number of samples; and n is the total number of MUs.

The k was interpreted using the scale proposed by (Landis; Koch, 1977), which defines the following ranges: Zero (0)—"Terrible," meaning no agreement beyond chance; "Bad" (0.01–0.20); "Fair" (0.21–0.40); "Moderate" (0.41–0.60); "Very Good" (0.61–0.80); and "Excellent" (0.81–1), indicating almost perfect agreement between the models on the test data.

The k results from the n = 100 maps generated for each algorithm were compared using Tukey's Honest Significant Difference (HSD) test at a 1% confidence level to assess statistically significant differences among the algorithms studied, including the ensemble learning approach (Equation (2)).

$$HSD = q\sqrt{\frac{MSE}{n_c}} \qquad \text{(Eq. 2)}$$

Where HSD is the minimum difference between two group means required for statistical significance; q is the critical value from the table; MSE is the mean squared error; and n_c is the sample size for each group.

Overall accuracy, defined as the degree of agreement between model predictions and observed values, was calculated as the number of correctly classified cases divided by the total number of predictions (Foody, 2020), according to Equation (3).

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N} \qquad \text{(Eq. 3)}$$

Where TP refers to the correctly classified instances of class i and C is the total number of MUs, divided by the total number of cases N.

2.3.2.1 Pixel-by-Pixel Spatial Prediction Performance on the Legacy Map

The digital maps produced by the tested algorithms were compared pixel-by-pixel with the legacy soil map using the Semi-Automatic Classification Plugin (SCP) (Congedo, 2021) implemented in QGIS Desktop 3.34.8. Model performance was evaluated through a confusion matrix, User's accuracy (UA), Producer's accuracy (PA), overall accuracy (OA), and Kappa index $(\hat{k})$ (Congalton; Green, 2019).

Overall accuracy (OA) measures the proportion of correctly classified pixels relative to the reference data (Equation (4)).

$$OA = \frac{\sum_{i=1}^{k} n_{ii}}{N} \quad \text{(Eq. 4)}$$

Where n_{ii} is the number of correctly classified samples for MU i, corresponding to the diagonal of the confusion matrix; k is the total number of MUs; and N is the total number of samples (sum of all elements in the matrix).

To properly assess UA and PA, the F1 score (Equation (5)) was calculated, which represents the harmonic mean of these two metrics. The F1 score was computed individually for each MU class and ranges from 0 to 1, with values closer to 1 indicating higher predictive accuracy (Goutte; Gaussier, 2005).

$$F1\ score = 2\left(\frac{UA \times PA}{UA + PA}\right) \quad \text{(Eq. 5)}$$

where UA is the percentage of samples correctly predicted for a given class relative to the total number of samples predicted for that MU, also known as precision, and PA (or recall) is the proportion of samples correctly predicted for a given MU.

2.3.2.2 Uncertainty of Algorithms Based on Soil Mapping Unit Classification

To estimate the prediction error of the generated digital soil maps—i.e., how close the predicted values are to the true values, also referred to as uncertainty—the probabilistic prediction of class occurrence for each MU was calculated using the numbers of votes from the n = 100 maps generated during algorithm execution. This assessment, however, is only of practical value to the end user if it is accurate and reliable (Schmidinger; Heuvelink, 2023).

The confusion index (CI), which quantifies the uncertainty between dominant and subdominant soil classes in the mapping units (MUs) (Equation (6)) (Shannon, 1948), and the normalized Shannon entropy [64], which measures the spatial prediction uncertainty (Equation (7)), are defined as follows:

$$CI = [1 - (\mu_{max\,i} - \mu_{(\max-1)i})] \quad \text{(Eq. 6)}$$

Where $\mu_{max\,i}$ is the probability of the most likely MU for soil sampling point i,, and $\mu_{(\max-1)i}$ is the probability of the second most likely MU for soil sampling point i,. CI values range from 0 to 1, with higher values indicating greater uncertainty.

$$Entropy = \frac{-\sum_{k=1}^{n} pk\,.log_2(pk)}{log_2(n)} \quad \text{(Eq. 7)}$$

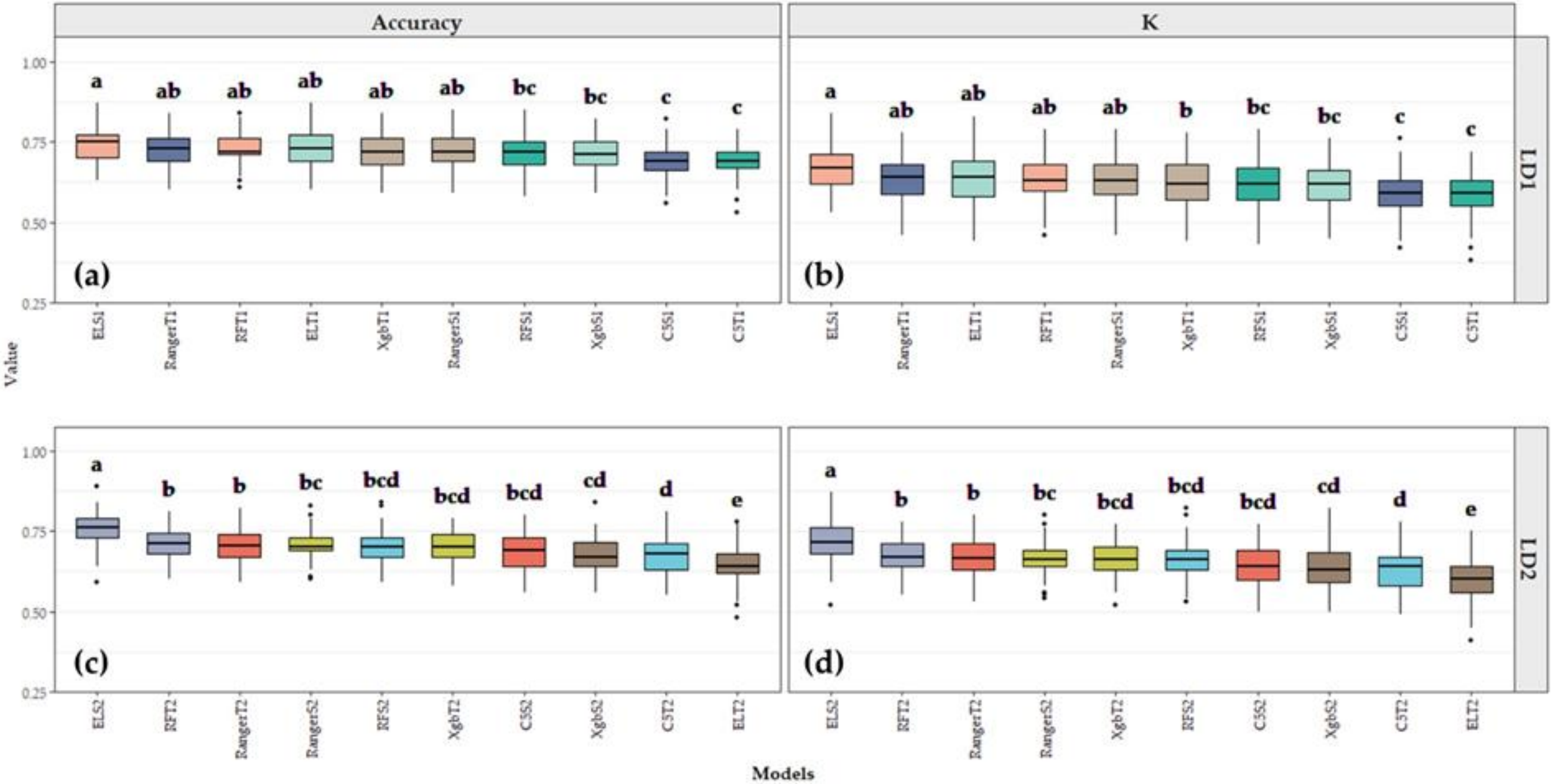
Where pk is the probability of class k and n is the total number of MUs. The entropy values were normalized, where 0 indicates maximum confidence in the predicted class and 1 indicates maximum uncertainty in the predictions.

2.4 Results

2.4.1 Model performance

We compared the performance of the models in predicting the MUs in LD1 and LD2 (Figure 5a–d). All models presented a “very good” k (0.61–0.80), except for models C5S1 and C5T1 (LD1) and ELT2 (LD2), which obtained a “moderate” k (0.41–0.60), according to the scale proposed by (Landis; Koch, 1977).

Figure 5 - Accuracy and Kappa index (k) results of predictive models. Boxplots of n = 100 predicted maps. (a) and (b) Accuracy and Kappa of the models in LD1, respectively; and (c) and (d) Accuracy and Kappa of the models in LD2, respectively. Dataset derived from digital surface (S) and terrain (T) models. Different letter types between models within the same metric and LD indicate a statistically significant difference at the 1% level according to Tukey's HSD test.

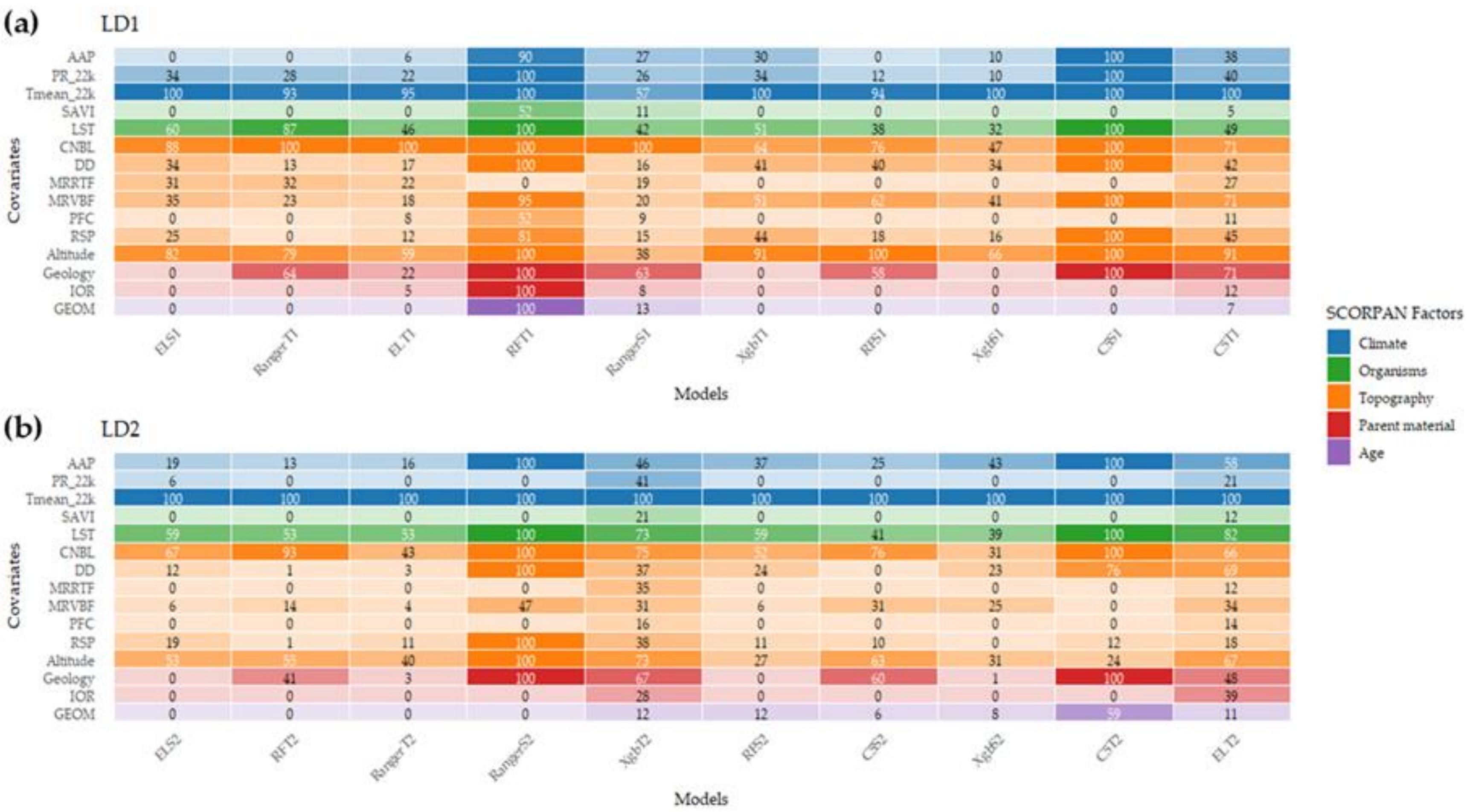


Source: The author (2026).

In LD1, the ELS1, RangerT1, RFT1, ELT1, and RangerS1 models showed the highest average accuracy and k values, ranging from 0.74 to 0.72 and 0.67 to 0.63, respectively. In LD2, only the ELS2 model showed a higher average than the others, with 0.75 and 0.72 for accuracy and k, respectively. On the other hand, the C5 algorithms showed the lowest performance in both LD1 and LD2.

We evaluated the weight of the 15 environmental covariates in all models studied and in the prediction of soil mapping units in LD1 and LD2 (Figure 6a, b). When considering the ELS1 and ELS2 models, the environmental covariates Tmean_22k, CNBL, Altitude, and LST (ELS1) and Tmean_22k, CNBL, LST, and Altitude (ELS2) were the most important for the models.

Figure 6 - Importance of environmental covariates in the evaluated models: (a) models using level of detail LD1 and (b) models using level of detail LD2. The x-axis represents the machine learning models, with S indicating models using covariates derived from the DSM and T indicating models using covariates derived from the DTM. The y-axis shows the environmental covariates selected by Recursive Feature Elimination (RFE) for digital soil mapping. The intensity of the color represents the level of importance, where 0 indicates low importance (lightest shade) and 100 indicates high importance (darker shade), varying according to the SCORPAN factor legend.



Source: The author (2026).

2.4.1.1 Internal Validation

We compared the predicted map with the legacy (reference) map, pixel-by-pixel, based on the boundaries of the predicted taxonomic units. The OA and k̂ values (Table 5) showed very little variation when comparing all of the models tested, in LD1 and LD2. Despite the reduction in the number of soil mapping units from nine in LD2 to five in LD1, we observed that the average k̂ decreased from 0.65 to 0.62.

Table 5 - Overall accuracy (OA) and Kappa index (k̂) for the models at two levels of detail (LD1 and LD2).

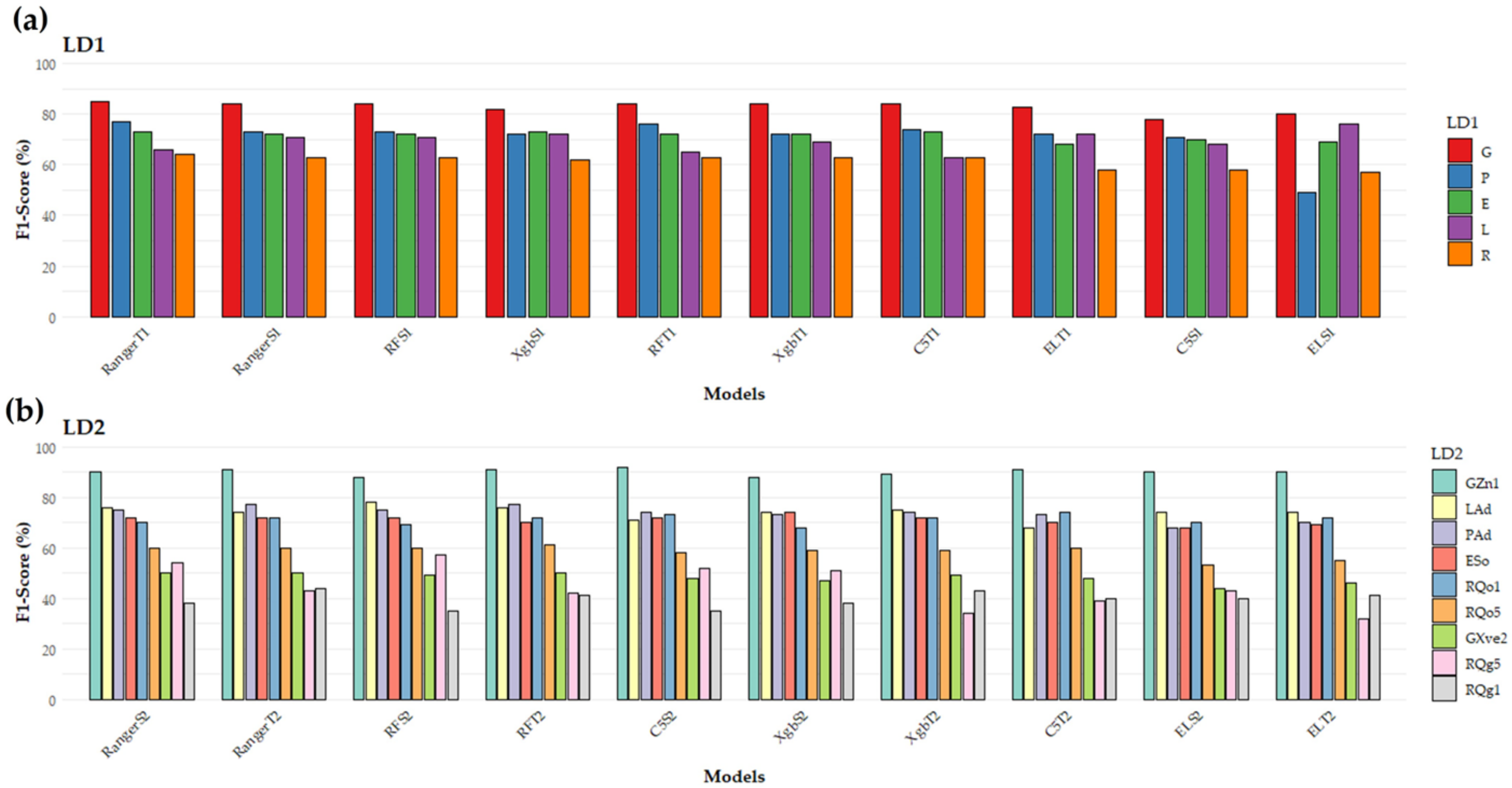
Levels of Detail					
LD1			**LD2**		
Models	OA	$\hat{k}$	Models	OA	$\hat{k}$
RFS1	0.72	0.63	RFS2	0.71	0.66
RangerS1	0.72	0.63	RangerS2	0.71	0.66
XgbS1	0.72	0.62	XgbS2	0.69	0.64
C5S1	0.68	0.58	C5S2	0.70	0.64
ELS1	0.69	0.59	ELS2	0.68	0.62
RFT1	0.71	0.62	RFT2	0.72	0.67
RangerT1	0.72	0.63	RangerT2	0.71	0.66
XgbT1	0.72	0.63	XgbT2	0.70	0.65
C5T1	0.71	0.61	C5T2	0.69	0.63
ELT1	0.71	0.62	ELT2	0.69	0.63

Source: The author (2026).

In Figure 7, we observe that the RangerT1, RangerS1, and RFS1 models in LD1 and the RangerS2, RangerT2, and RF2 models in LD2 presented the highest absolute values of PA and UA (Figure 7a, b). We observed that RQg5 and GXve2 presented the lowest PA metrics, frequently below 50% in several models, highlighting difficulties in spatial discrimination.

GZn1 showed the highest UA in almost all models—RangerS2, RangerT2, RFS2, RFT2, and C5S2—reaching 96%. The PA was also consistently high, reaching 89% in the C5S2 model. It is also worth highlighting that the high F1-score in GZn1 revealed that the climate (Tmean_22k) and topography (CNBL) factors were the most important predictive covariates in predicting this unit, which indicates soils in hydromorphic environments.

Figure 7 - F1-score results for MUs predicted by the evaluated machine learning models at two levels of detail: LD1 (a) and LD2 (b). The x-axis represents the machine learning models, and the y-axis shows the F1-score (%). Different colors represent the predicted soil mapping units within each level of detail.

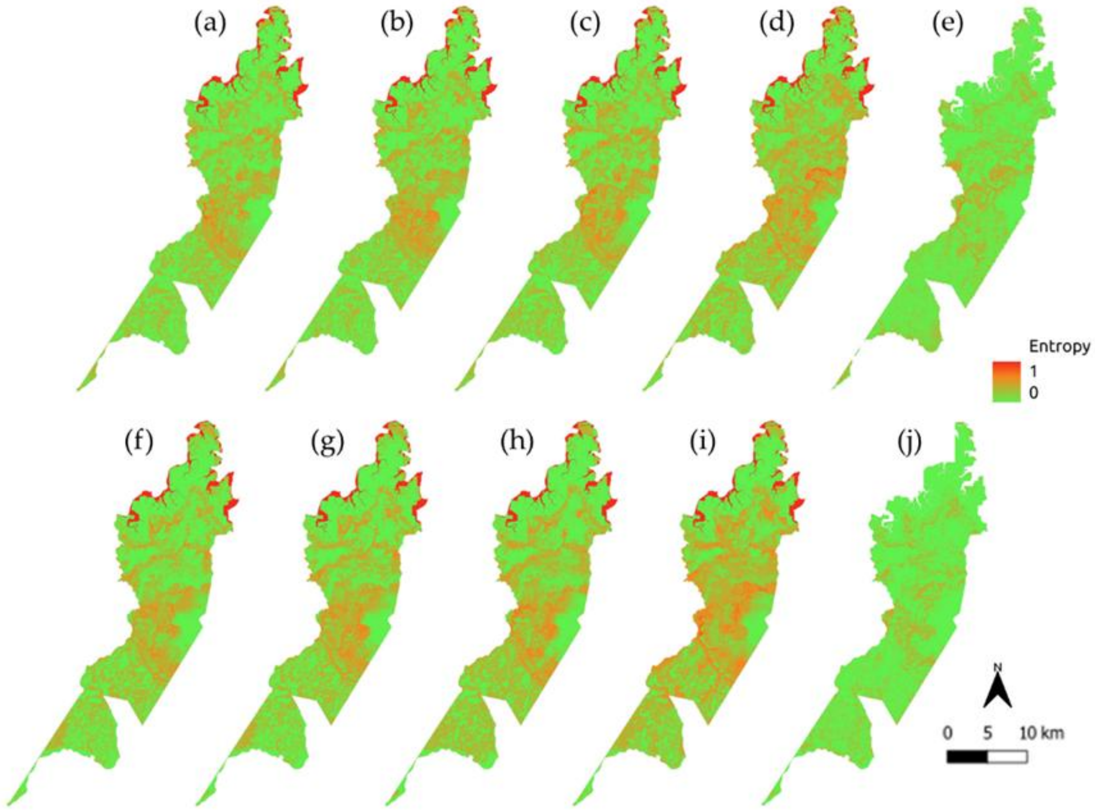


Source: The author (2026).

2.4.2 Uncertainty of Spatial Predictions

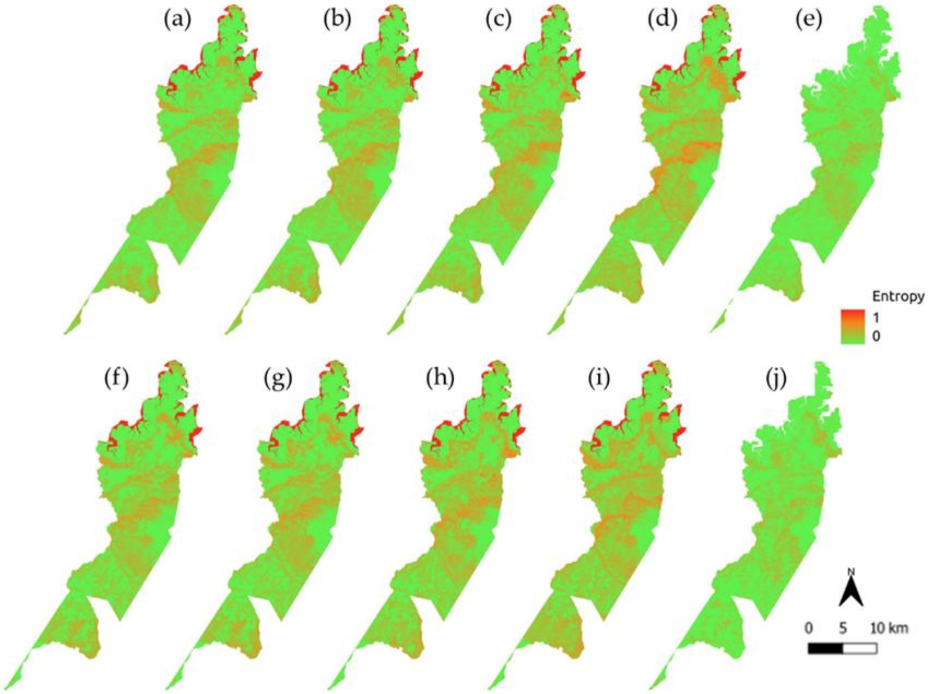
In Figures 8 and 9 (entropy maps) and Figures 10 and 11 (CI maps), we observe that the ELS and ELT models for both LD1 and LD2 similarly presented a larger area with predicted MUs with less uncertainty (entropy and CI close to 0) when compared to other conventional models.

Figure 8 - Entropy of all models tested at LD1. Model arrangement: (a) RFS1, (b) RangerS1, (c) XgbS1, (d) C5S1, (e) ELS1, (f) RFT1, (g) RangerT1, (h) XgbT1, (i) C5T1, and (j) ELT1.



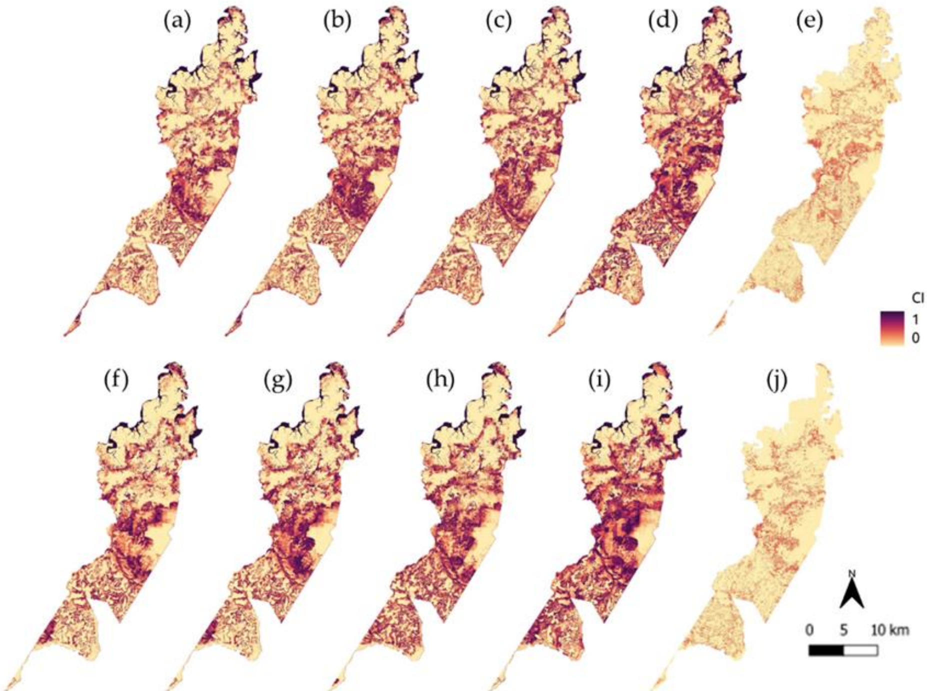
Source: The author (2026).

Figure 9 - Entropy of all models tested at LD2. Model arrangement: (a) RFS2, (b) RangerS2, (c) XgbS2, (d) C5S2, (e) ELS2, (f) RFT2, (g) RangerT2, (h) XgbT2, (i) C5T2, and (j) ELT2.



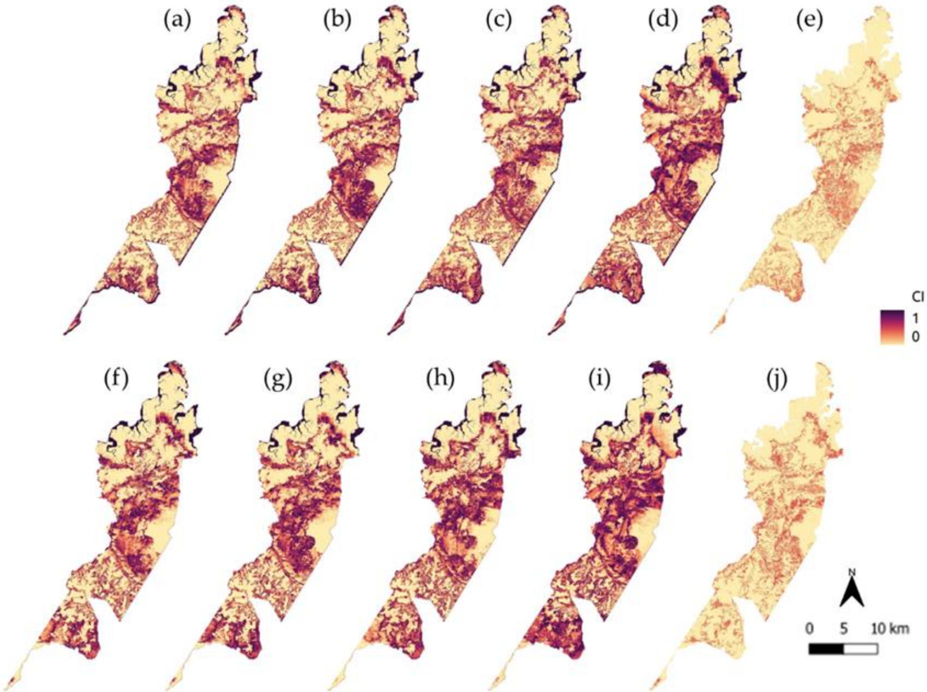
Source: The author (2026).

Figure 10 - Confusion index of all models tested at LD1. Model arrangement: (a) RFS1, (b) RangerS1, (c) XgbS1, (d) C5S1, (e) ELS1, (f) RFT1, (g) RangerT1, (h) XgbT1, (i) C5T1, and (j) ELT1.



Source: The author (2026).

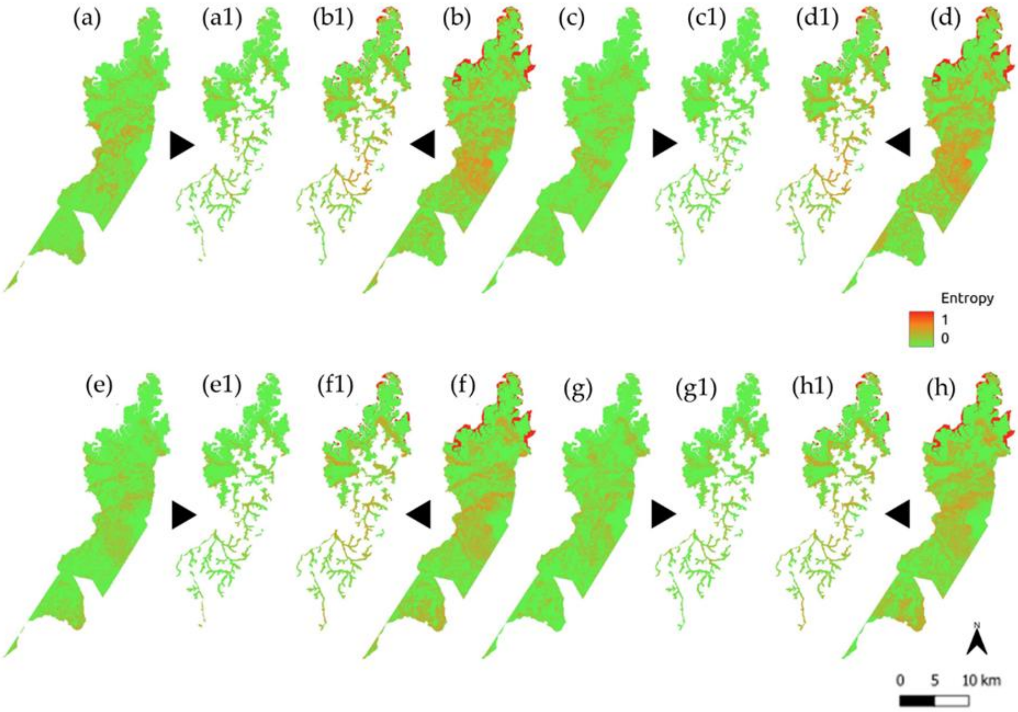
Figure 11 - Confusion index of all models tested at LD2. Model arrangement: (a) RFS2, (b) RangerS2, (c) XgbS2, (d) C5S2, (e) ELS2, (f) RFT2, (g) RangerT2, (h) XgbT2, (i) C5T2, and (j) ELT2.



Source: The author (2026).

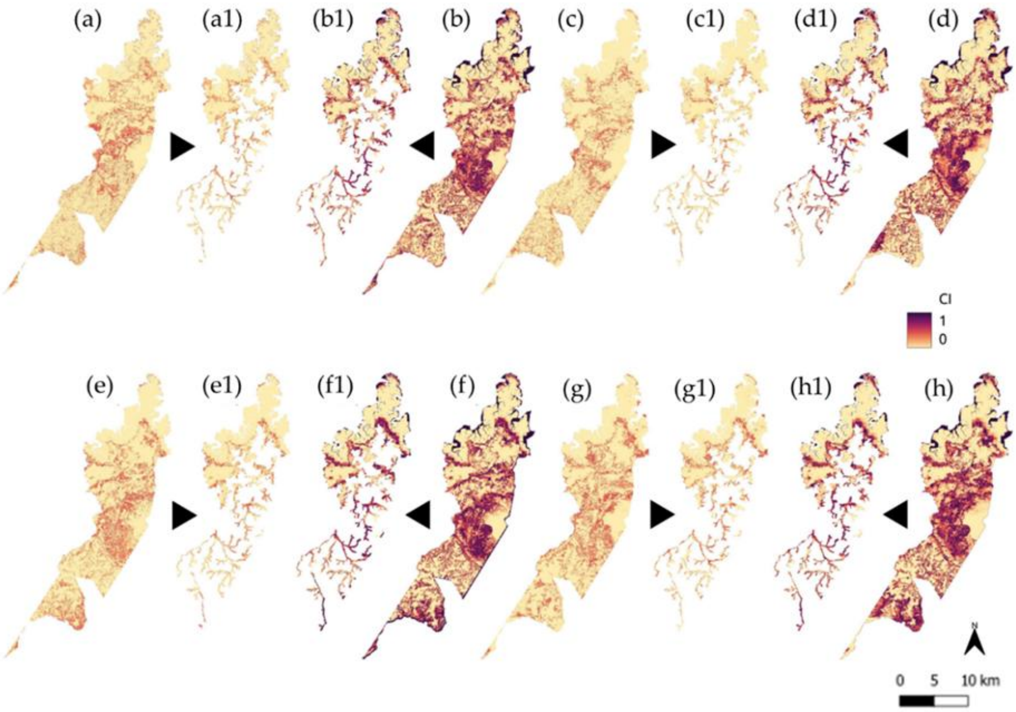
The spatial uncertainty of the predictive models was further evaluated for soil mapping units under hydromorphic conditions, which are characteristic of the carbon-rich valley bottoms and alluvial plains of the Amazon. A comparative analysis of the entropy and confusion index (CI) across mapping units LD1 and LD2 is illustrated in Figures 12 and 13.

Figure 12 - Comparison of entropy, highlighting areas under hydromorphic conditions between ELS1 and RangerS1 between ELS1 and RangerS1 (a,a1,b1,b), ELT1 and RangerT1 (c,c1,d1,d), ELS2 and RangerS2 (e,e1,f1,f), and ELT2 and RangerT2 (g,g1,h1,h) models.



Source: The author (2026).

Figure 13 - Comparison of confusion index (CI), highlighting areas under hydromorphic conditions between ELS1 and RangerS1 (a,a1,b1,b), ELT1 and RangerT1 (c,c1,d1,d), ELS2 and RangerS2 (e,e1,f1,f), and ELT2 and RangerT2 (g,g1,h1,h) models.



Source: The author (2026).

2.5 Discussion

2.5.1 Evaluation of Modeling Metrics

In LD1, the stability in the prediction (Accuracy and k) of the MUs of the ELS1, RangerT1, ELT1, RFT1, and RangerS1 models was statistically higher and equal to that of the C5S1 and C5T1 models. In LD2, the stability in the prediction of the ELS2 model was statistically superior to all individual predictive models, including the ELT2 model (Figure 5).

The greater complexity of the spatial distribution of the MUs in LD2 may be related to the difference between the larger number of predictive models with higher precision and k values in LD1 compared to LD2 (Figure 5). This may be related to the potential of the EL approach to mitigate errors, mainly by using a conventional soil map as a reference. These results are consistent with those of (Azizi et al., 2023; Li et al., 2020).

Unlike our hypothesis, in general, the use of terrain attributes derived from ANADEM (DTM) and the aggregation of soil units in LD1 did not improve performance when used in modeling. This result demonstrated that the use of environmental covariates derived from DSM was sufficient in predicting soil units. The aggregation of soil units from a legacy map did not prove to be a strategy capable of improving model performance, as there was a 5-percentage point decrease in the k of ELS2 compared to ELS1 (Figure 5). This proves that despite the challenges regarding the high uncertainty implicit in the legacy map, these provide valuable information on the spatial distribution of soils (Medeiros et al., 2024), particularly in the Amazon region.

These results demonstrate the ability of EL to aggregate the strengths of individual algorithms. There was a 4-percentage-point difference in the k between ELS1 (0.67) and ELS2 (0.72), showing that ELS2 was more balanced than ELS1 (Figure 5), despite disagreeing with the results of Mello et al. (2021), which justified the reduction in model performance due to the increased complexity of the MUs. In our study, the reduction in ELS1 performance compared to ELS2 may be related to error amplification by aggregating LD2 MUs to LD1 from the conventional map (reference).

The small differences in accuracy and k (Figure 5), especially in the models EL, Ranger, and RF in LD1, may be related to the fact that these algorithms all use a bagging structure, while Xgb uses a boosting structure (Adeniyi et al., 2023; Taghizadeh-Mehrjardi et al., 2019, 2021).

The potential improvement in k values achieved by EL compared to conventional models for soil class prediction has been poorly reported in the literature (Medeiros et al., 2024). Our results are consistent with those of (Azizi et al., 2023; Li et al., 2020), who also reported superior performance of ensemble models compared to conventional approaches. These results indicate that ELS mitigated the spatial prediction errors of individual algorithms, mainly when using a conventional soil map as a reference.

Our findings showed that paleoclimate, represented by the covariate Tmean_22k (Figure 6), was more important than the covariates representing the current climate. Similar results were found by (Carvalho Monteiro et al., 2023), who observed the estimated precipitation from 22,000 years ago in the south of the state of Minas Gerais. During this period, it is believed that, in Brazil, a warmer and rainier climate favored erosive processes related to pedogenesis (Carvalho Monteiro et al., 2023). Even in gentler relief conditions, both water infiltration into the soil profile and surface runoff can accelerate weathering (Carvalho Monteiro et al., 2023). Particularly in the development area of our study, characterized by its proximity to the Atlantic Ocean, this may justify the distribution of the MUs in the area.

The terrain attribute CNBL was one of the most important covariates in the ELS1 and ELS2 models (Figure 6); as a covariate derived from the drainage network, it is frequently associated with soil erosion and sedimentation and plays a key role in soil–landscape stratification (Adeniyi et al., 2023; Mello et al., 2021; Taghizadeh-Mehrjardi et al., 2014). The CNBL calculates the vertical distance of drainage channels, capable of separating soils under hydromorphic influence from soils in plateau areas (Campos et al., 2019). In our work, approximately 40% of the mulches are under hydromorphic influence, which justifies this covariate being among the most important. Based on these results, it was observed that the selection of uncorrelated covariates that encompass all SCORPAN factors may be related to the higher Kappa indices and precision.

These findings support the potential integration of tacit knowledge concerning soil-landscape modeling and expert mental maps of soil distribution with machine learning. However, this should be approached with caution, since pedological knowledge, in addition to being reflected in the selection of relevant covariates, requires subsequent interpretation and analysis of the resulting patterns, both from a spatial prediction and soil science perspective (Rentschler; Scholten, 2025; Wadoux et al., 2020).

We compared all models pixel by pixel with the conventional map (Table 4) in terms of the relationship between the reference and predicted MUs and by means of a confusion matrix (Tables A1–A15 and Figure 7). The average value of OA in LD1 and LD2 was 0.71 and 0.70, respectively; however, the average kˆ value was 0.62 in LD1 and 0.65 in LD2. These results lead us to not indicate the aggregation of MUs. In the analysis of the F1 score (Figure 7), we observed that the individual Ranger and RF models, as well as all other models trained, in general, as MUs G in LD1 and GZn1 in LD2, obtained the highest values, indicating that all models were able to predict them.

Regarding Tables 1 and 2, we calculated the sampling density, from which we observed that the Ranger achieved the highest F1 scores for G and GZn1, even though their densities of 0.27 and 0.18 points/km2, respectively, were the second lowest among the classes studied, represented by LD1 and LD2. In our work, the Ranger model was able to capture soil variability even with low sampling density, corroborating the findings of (Esfandiarpour-Boroujeni et al., 2020), who compared the potential of decision tree and vector learning quantization models for soil class prediction, finding greater accuracy for decision tree models in areas with low altitude variation.

2.5.2 Uncertainty and Practical Applications

The quantified uncertainty in PDSM is one of the main advantages over conventional soil surveys, in which the spatial distribution of mapping units or soil classes is a source of error. This reflects classification uncertainty and may be related, for example, to combined mapping units whose evaluation criteria are subjective, or to estimation models composed of coexisting soils or other generalized combinations of dominant soils (Lamichhane; Kumar; Adhikari, 2021; Sarmento et al., 2017).

In the study area, mapping units of hydromorphic soil classes cover more than 40% of the municipality of Tracuateua and present specific challenges, such as limited accessibility for the soil survey team. Incorporating uncertainty into the existing map can substantially assist future field surveys, reducing the time and resources required between planning and map production.

Locations with a higher entropy and confusion index, both close to 1 (Figures 12 and 13), are associated with greater difficulty in pedogenesis modeling, requiring a larger number of samples to improve the quality of the soil map (Medeiros et al., 2024). This demonstrates, in the present study, that models related to EL can improve the prediction of soil mapping units in areas where individual algorithms presented greater uncertainty.

Commonly represented by means of maps, the incorporation of uncertainty continues to be a major challenge in pedometrics (Courteille et al., 2025; Wadoux et al., 2021). In the joint analysis of the F1 score (Figure 7) and Figures 12 and 13, in which uncertainty (entropy and

IC) was compared, lower uncertainty values for MUs under hydromorphic conditions—G, E, RQg5, Eso, GXve2, RQg1, and GZn1—were obtained using models with the EL approach compared to Ranger. These results indicate that the overall accuracy of the map should be evaluated in conjunction with the uncertainty (Schmidinger; Heuvelink, 2023).

Valley bottom areas and alluvial plains, as shown in Figures 12 and 13, are extremely common and vital geomorphic environments in the Amazon, concentrating significant carbon stocks in soils (Amendola et al., 2018). However, we believe that pedogenesis in these environments is intrinsically complex, being controlled by multiple interactive processes. According to (Phillips, 2004), this complexity often reflects polygenesis, in which soils evolve through successive or overlapping pedogenetic processes that occur at different times or under different environmental conditions within the same landscape. This intrinsic complexity, combined with the lack of detailed soil maps for decision-making, represents significant challenges for PDSM in the region, since hydromorphic mapping units have subtle boundaries and high spatial variability, requiring high-resolution terrain co-variates. Ref. (Meier et al., 2018) found that soils in hydromorphic lowlands (GXbd) were classified more accurately by some machine learning algorithms, highlighting the need for fine-scale terrain covariates. Therefore, the need to use high-resolution terrain covariates to map hydromorphic pedological units is supported by the recent review in (Adeniyi; Ba-ture; Mearker, 2024).

In comparing MUs under hydromorphic conditions using the EL and Ranger models (Figures 12 and 13), despite the low sample density for G and GZn1 (0.27 and 0.18 points/km2, respectively) compared to the other classes, lower entropy and CI values were observed for the EL approach. This indicates that the joint learning approach incorporated higher predictive quality in hydromorphic environments, i.e., lower uncertainty even with a reduced number of samples, corroborating the findings of (Carvalho Monteiro et al., 2023) in their assessment of uncertainty in digital mapping of soil classes in the southern region of the state of Minas Gerais.

Given the challenges in the Amazon biome—such as budget constraints and difficult field access, particularly in hydromorphic areas—this study has shown promise for PDSM in the region. Another promising application of PDSM is predictive extrapolation, which could significantly support ongoing soil surveys with higher resolution, thus helping to fill existing gaps in soil maps throughout the Amazon (Grinand et al., 2008; Lemercier et al., 2012; Nenkam et al., 2022).

2.6 Conclusion

The results of this study confirm the initial hypothesis, demonstrating that the EL approach effectively discriminates between MUs, even under complex hydromorphic conditions. Considering the polygenetic nature of tropical soils, the most important covariates (Tmean_22k, CNBL, LST, and altitude) proved adequate for modeling MUs in northeastern Pará.

Despite the scarcity of validation points for some MUs in the legacy soil map of Tracuateua, the performance metrics indicate that the EL approach provided greater predictive stability than individual machine learning models, particularly in LD2 when dealing with MUs imbalance.

The results support the hypothesis that EL can discriminate MUs under hydromorphic conditions. In addition, spatial estimates of prediction uncertainty based on entropy and CI provide useful tools for guiding field surveys and improving soil classification in difficult to access hydromorphic areas.

Furthermore, reducing the level of detail of MUs in older maps to simplify modeling did not improve predictive performance and may increase the implicit uncertainty associated with conventional soil maps. Furthermore, despite the slightly higher use of terrain attributes derived from DSM in terms of learning model performance compared to DTM, the greater uncertainty (Entropy and CI) observed in the latter digital model may be related to the more accurate representation of terrain elevation.

However, our approach provides a framework for identifying the most appropriate use of legacy soil class data under regional constraints. The results highlight that soils, directly associated with natural landscape patterns, can be predicted more effectively by machine learning models, exhibiting more accurate spatial patterns of soil variability through the use of appropriate environmental covariates in digital soil mapping models.

REFERENCES

ADENIYI, Odunayo David *et al.* Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. **Land**, v. 12, n. 2, p. 494, 16 fev. 2023.

ADENIYI, Odunayo David; BATURE, Hauwa; MEARKER, Michael. A Systematic Review on Digital Soil Mapping Approaches in Lowland Areas. **Land**, v. 13, n. 3, p. 379, 17 mar. 2024.

ALVARES, Clayton Alcarde *et al.* Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.

AMENDOLA, D. *et al.* Soil hydromorphy and soil carbon: A global data analysis. **Geoderma**, v. 324, p. 9–17, ago. 2018.

ARIF, Muhammad. Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review. **International Journal of Database Theory and Application**, v. 11, n. 1, p. 1–8, 31 mar. 2018.

AZIZI, Kamran *et al.* Integration of Sentinel-1/2 and topographic attributes to predict the spatial distribution of soil texture fractions in some agricultural soils of western Iran. **Soil and Tillage Research**, v. 229, p. 105681, maio 2023.

BAUMGARDNER, Murray *et al.* Reflectance properties of soils. *In*: BRADY, Nyle (Org.). **Advances in Agronomy**. San Diego, CA, USA: Academic Press, 1985. v. 38 p. 1–44.

BAZAGLIA FILHO, Osmar *et al.* Comparison between detailed digital and conventional soil maps of an area with complex geology. **Revista Brasileira de Ciência do Solo**, v. 37, n. 5, p. 1136–1148, out. 2013.

BIELSKI, Conrad *et al.* Novel Approach for Ranking DEMs: Copernicus DEM Improves One Arc Second Open Global Topography. **IEEE Transactions on Geoscience and Remote Sensing**, v. 62, p. 1–22, 2024.

INSTITUO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Geologia 1:250.000**.

BREIMAN, Leo. Bagging Predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, ago. 1996.

CAMPOS, Alcinei Ribeiro *et al.* Selection of Environmental Covariates for Classifier Training Applied in Digital Soil Mapping. **Revista Brasileira de Ciência do Solo**, v. 42, n. 0, 7 jan. 2019.

CARVALHO MONTEIRO, Maria Eduarda *et al.* Digital mapping of soil classes in Southeast Brazil: environmental covariate selection, accuracy, and uncertainty. **Journal of South American Earth Sciences**, v. 132, p. 104640, dez. 2023.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost. *In*: New York, NY, USA: ACM, 13 ago. 2016.

CHRISTOFOLETTI, A. **Geomorfologia**. 2nd ed ed. São Paulo, SP: Blucher, 1980.

COHEN, Jacob. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1 abr. 1960.

CONGALTON, Russell G.; GREEN, Kass. **Assessing the Accuracy of Remotely Sensed Data**. *[S.l.]*: CRC Press, 2019.

CONGEDO, Luca. Semi-Automatic Classification Plugin: A Python tool for the download and processing of remote sensing images in QGIS. **Journal of Open Source Software**, v. 6, n. 64, p. 3172, 27 ago. 2021.

CONRAD, O. *et al.* System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. **Geoscientific Model Development**, v. 8, n. 7, p. 1991–2007, 7 jul. 2015.

COSTA, E. J. de S.; ALMEIDA, H. G. de. **Mapa geológico e mapa de favorabilidade para tipos de jazimentos minerais - município de Tracuateua, PA**. Belém, PA: 1998.

COURTEILLE, Léa *et al.* What is the best way to communicate the uncertainty of a digital soil mapping product? Some lessons from an end-users survey. **Geoderma**, v. 459, p. 117302, jul. 2025.

ESFANDIARPOUR-BOROUJENI, I. *et al.* Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations. **Catena**, v. 191, p. 104581, ago. 2020.

FERNANDES FILHO, Elpídio I. *et al.* The Future of Brazilian Pedology: Pedometrics and Advanced Methods for Soil Survey. *In*: The Soils of Brazil. p. 423–433. 2023.

FLORINSKY, Igor V. Influence of Topography on Soil Properties. *In*: **Digital Terrain Analysis in Soil Science and Geology**. *[S.l.]*: Elsevier, 2012. p. 145–149.

FOODY, Giles M. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. **Remote Sensing of Environment**, v. 239, p. 111630, mar. 2020.

GALLANT, John C.; DOWLING, Trevor I. A multiresolution index of valley bottom flatness for mapping depositional areas. **Water Resources Research**, v. 39, n. 12, 16 dez. 2003.

GORELICK, Noel *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, dez. 2017.

GOUTTE, Cyril; GAUSSIER, Eric. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *In*: Advances in Information Retrieval. p. 345–359. 2005.

GRINAND, Clovis *et al.* Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, v. 143, n. 1–2, p. 180–190, jan. 2008.

GRUNWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. **Geoderma**, v. 152, n. 3–4, p. 195–207, set. 2009.

HENGL, Tomislav *et al.* Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 29 ago. 2018.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, v. 25, n. 3, p. 295–309, ago. 1988.

JASIEWICZ, Jarosław; STEPINSKI, Tomasz F. Geomorphons — a pattern recognition approach to classification and mapping of landforms. **Geomorphology**, v. 182, p. 147–156, jan. 2013.

JENNY, H. **Factors of Soil Formation: a System of Quantitative Pedology**. New York: Dover Publications, 1941.

KAMPF, N.; CURI, N. Formação e Evolução do Solo (Pedogênese). *In*: KER, J. C. ;. *et al.* (Orgs.). **Pedologia: Fundamentos**. 1ª ed. Viçosa, MG: Sociedade Brasileira de Ciência do Solo - SBCS, 2012. p. 207–302.

KEMPEN, Bas *et al.* Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. **Soil Science Society of America Journal**, v. 76, n. 6, p. 2097–2115, 19 nov. 2012.

KUENZER, C.; DECH, S. Theoretical Background of Thermal Infrared Remote Sensing. *In*: KUENZER, Claudia; DECH, Stefan (Orgs.). **Thermal Infrared Remote Sensing**. Dordrecht: Springer Netherlands, 2013. v. 17 p. 1–26.

LAGACHERIE, P.; MCBRATNEY, A. B. Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. *In*: Developments in Soil Science. p. 3–22. 2006.

LAIPELT, Leonardo *et al.* ANADEM: A Digital Terrain Model for South America. **Remote Sensing**, v. 16, n. 13, p. 2321, 25 jun. 2024.

LAMICHHANE, Sushil; KUMAR, Lalit; ADHIKARI, Kabindra. Updating the national soil map of Nepal through digital soil mapping. **Geoderma**, v. 394, p. 115041, jul. 2021.

LANDIS, J. Richard; KOCH, Gary G. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, v. 33, n. 1, p. 159, mar. 1977.

LEMERCIER, Blandine *et al.* Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. **Geoderma**, v. 171–172, p. 75–84, fev. 2012.

LI, Xinchuan *et al.* Improving Soil Thickness Estimations Based on Multiple Environmental Variables with Stacking Ensemble Methods. **Remote Sensing**, v. 12, n. 21, p. 3609, 3 nov. 2020.

LIN, Yiqi *et al.* Mapping soil parent materials in a previously glaciated landscape: Potential for a machine learning approach for detailed nationwide mapping. **Geoderma Regional**, v. 40, p. e00905, mar. 2025.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1–2, p. 3–52, nov. 2003.

MEDEIROS, Beatriz Macêdo *et al.* Soil legacy data: An opportunity for digital soil mapping. **Revista Brasileira de Ciência do Solo**, v. 48, 19 jun. 2024.

MEIER, Martin *et al.* Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. **Revista Brasileira de Ciência do Solo**, v. 42, n. 0, 14 nov. 2018.

MELLO, Fellipe *et al.* Expert-based maps and highly detailed surface drainage models to support digital soil mapping. **Geoderma**, v. 384, p. 114779, fev. 2021.

MIRCHOOLI, Fahimeh *et al.* Spatial distribution dependency of soil organic carbon content to important environmental variables. **Ecological Indicators**, v. 116, p. 106473, set. 2020.

MOQUEDACE, Cássio Marques *et al.* High-resolution mapping of soil carbon stocks in the western Amazon. **Geoderma Regional**, v. 36, p. e00773, mar. 2024.

MUKUMBUTA, I. *et al.* Accessing and assessing legacy soil information, an example from two provinces of Zambia. **Geoderma**, v. 420, p. 115874, ago. 2022.

NENKAM, Andree M. *et al.* Using homosoils for quantitative extrapolation of soil mapping models. **European Journal of Soil Science**, v. 73, n. 5, 2 set. 2022.

NEYESTANI, Mehrnaz *et al.* Digital mapping of soil classes using spatial extrapolation with imbalanced data. **Geoderma Regional**, v. 26, p. e00422, set. 2021.

ODEGUA, R. An empirical study of ensemble techniques (bagging, boosting and stacking). *In*: Nairobi, Kenya: IndabaXAt, 2019.

OLAYA, V.; CONRAD, O. Chapter 12 Geomorphometry in SAGA. *In*: *[S.l.: S.n.]*. p. 293–308.

OLIVEIRA JUNIOR, Raimundo *et al.* **Zoneamento agroecológico do município de Tracuateua, Estado do Pará (Documentos - Número, 15)**. Belém: *[S.n.]*. Disponível em:

<https://www.infoteca.cnptia.embrapa.br/bitstream/doc/378660/1/OrientalDoc15.pdf>. Acesso em: 6 mar. 2026.

PADARIAN, José; MINASNY, Budiman; MCBRATNEY, Alex B. Machine learning and soil sciences: a review aided by machine learning tools. **SOIL**, v. 6, n. 1, p. 35–52, 6 fev. 2020.

PAVÃO, Quésia Sá *et al.* Texture prediction of natural soils in the Brazilian Amazon through proximal sensors. **Geoderma Regional**, v. 37, p. e00813, jun. 2024.

PHILLIPS, Jonathan D. Geogenesis, pedogenesis, and multiple causality in the formation of texture-contrast soils. **Catena**, v. 58, n. 3, p. 275–295, dez. 2004.

POGGIO, Laura *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. **SOIL**, v. 7, n. 1, p. 217–240, 14 jun. 2021.

POLIDORO, J. C. *et al.* **Programa Nacional de Solos do Brasil (PronaSolos) - (Documentos - Número, 183)**. Embrapa Solos. Rio de Janeiro: *[S.n.]*. Disponível em: <infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1054924/1/Doc183ProgramaNacionaldeSolosdoBrasil.pdf>. Acesso em: 6 mar. 2026.

QI, Feng; ZHU, A. Xing. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. **Computers & Geosciences**, v. 37, n. 9, p. 1425–1436, set. 2011.

QUINLAN, J. R. .. **C4.5 : programs for machine learning**. *[S.l.]*: Morgan Kaufmann Publishers, 1993.

RENTSCHLER, Tobias; SCHOLTEN, Thomas. A Note on Spurious Correlations and Explainable Machine Learning in Digital Soil Mapping. **European Journal of Soil Science**, v. 76, n. 4, 8 jul. 2025.

SANTOS, H. G. dos *et al.* **Sistema Brasileiro de Classificação de Solos**. 6° Edição ed. Brasília, DF: Embrapa, 2025.

SARMENTO, Eliana Casco *et al.* Disaggregating conventional soil maps with limited descriptive data: A knowledge-based approach in Serra Gaúcha, Brazil. **Geoderma Regional**, v. 8, p. 12–23, mar. 2017.

SAYÃO, Veridiana Maria; DEMATTÊ, José A. M. Soil texture and organic carbon mapping using surface temperature and reflectance spectra in Southeast Brazil. **Geoderma Regional**, v. 14, p. e00174, set. 2018.

SCHMIDINGER, Jonas; HEUVELINK, Gerard B. M. Validation of uncertainty predictions in digital soil mapping. **Geoderma**, v. 437, p. 116585, set. 2023.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, jul. 1948.

SINGER, Robert B. Near-infrared spectral reflectance of mineral mixtures: Systematic combinations of pyroxenes, olivine, and iron oxides. **Journal of Geophysical Research: Solid Earth**, v. 86, n. B9, p. 7967–7982, 10 set. 1981.

SOIL SCIENCE DIVISION STAFF. **Soil Survey Manual**. 18th ed. ed. United States: Soil Science Division Staff, 2017.

STRAHLER, A. N. Hypsometric Analysis of Erosional Topography. **Bulletin of the Geological Society of America**, v. 63, p. 1117–1142, 1952.

TAGHIZADEH-MEHRJARDI, R. *et al.* Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran. **Arid Land Research and Management**, v. 28, n. 2, p. 147–168, 3 abr. 2014.

TAGHIZADEH-MEHRJARDI, Ruhollah *et al.* Digital Mapping of Soil Classes Using Ensemble of Models in Isfahan Region, Iran. **Soil Systems**, v. 3, n. 2, p. 37, 28 maio 2019.

TAGHIZADEH-MEHRJARDI, Ruhollah *et al.* Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. **Geoderma**, v. 399, p. 115108, out. 2021.

TAYLOR, Karl E.; STOUFFER, Ronald J.; MEEHL, Gerald A. An Overview of CMIP5 and the Experiment Design. **Bulletin of the American Meteorological Society**, v. 93, n. 4, p. 485–498, 1 abr. 2012.

WADOUX, Alexandre M. J. -C. *et al.* A note on knowledge discovery and machine learning in digital soil mapping. **European Journal of Soil Science**, v. 71, n. 2, p. 133–136, 19 mar. 2020.

WADOUX, Alexandre M. J. C. *et al.* Ten challenges for the future of pedometrics. **Geoderma**, v. 401, p. 115155, nov. 2021.

WADOUX, Alexandre M. J. -C. Artificial intelligence in soil science. **European Journal of Soil Science**, v. 76, n. 2, 14 mar. 2025.

WADOUX, Alexandre M. J. -C.; HEUVELINK, Gerard B. M. Uncertainty of spatial averages and totals of natural resource maps. **Methods in Ecology and Evolution**, v. 14, n. 5, p. 1320–1332, 10 maio 2023.

WRIGHT, Marvin N.; ZIEGLER, Andreas. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 77, n. 1, 2017.

ZHU, Changda *et al.* Extracting Typical Samples Based on Image Environmental Factors to Obtain an Accurate and High-Resolution Soil Type Map. **Remote Sensing**, v. 16, n. 7, p. 1128, 23 mar. 2024.

Statements & Declarations

Acknowledgements: We thank the members of Geotechnologies and Pedometrics Research Group – GEOP (https://geopufra.com /equipe/) for the support of the team. This study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Funding: This thesis was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Chamada CNPq/MCTI/FNDCT N° 18/2021 - Universal 2021 – Process: 406838/2021-6.

CHAPTER 3

3 TRANSFERRING SOIL–LANDSCAPE RELATIONSHIPS IN DATA-LIMITED REGIONS OF THE AMAZON: A KNOWLEDGE-GUIDED DIGITAL SOIL MAPPING APPROACH BASED ON REFERENCE AREAS

3.1 Abstract

Digital soil mapping has become established as a promising alternative for areas lacking conventional soil surveys. The integration of predictive models based on soil-forming factors, environmental covariates, and information on the spatial distribution of soils from reference areas can be extrapolated to other similar regions. In this study used five machine learning algorithms: Random Forest, Ranger, XGBoost, C5.0, and an ensemble of these algorithms. We used 15 covariates related to the SCORPAN model factors, whose terrain attributes were derived from two digital elevation models: a digital surface model (DSM) and a digital terrain model (DTM). As a reference area, we used information from a legacy map at a 1:100,000 scale of the municipality of Tracuateua, State of Pará, originally developed with nine mapping units (MUs) using conventional methods up to the third categorical level (Great Group) according to the Brazilian Soil Classification System (SiBCS), which we refer to as the LD2 level of detail. We also applied an aggregation approach to five MUs at the LD1 level of detail. We evaluated model performance in internal validation using accuracy and the Kappa coefficient, as well as the relative importance of the covariates. In external validation, using an error matrix based on 381 field observation points, we assessed overall accuracy (OA) and Kappa, and also calculated the purity of the predicted mapping units. The ensemble approaches—ELS1, ELT1, and ELS2—achieved superior stability and accuracy. In general, environmental covariates such as: Altitude, Channel Network Base Level (CNBL), and paleoclimate (mean annual temperature estimated at 22,000 years ago, Tmean_22k) drove the extrapolation of Gleysols ("G") and Sodic Salic Gleysols (GZn1). The ensemble approach extrapolated mapping units under hydromorphic environments with robust accuracy and purity. The extrapolation maps were successfully produced and were consistent with the observed landscape reality.

Keywords: pedology, soil survey, pedometrics, pedological cartography, ensemble learning, prediction of soil units, artificial intelligence, amazonian soils.

3.2 Introdution

The use of digital soil mapping (DSM) and machine learning techniques represents an important strategy for soil surveys in the Amazon, with a focus on the sustainability of this biome (Pavão et al., 2024). In this context, increased detail in the spatial distribution of these soils provides the foundation for decision-making regarding the most appropriate management practices to be adopted, aiming, among other objectives: carbon sequestration from the atmosphere, mitigation of climate change, and the reduction of erosion processes, especially in ecologically sensitive biomes such as the Brazilian Amazon tropical forest (Lamichhane et al., 2021; Moquedace et al., 2024).

However, a persistent gap remains in the availability of high-resolution data (Hartemink and McBratney, 2008), as is the case in the Amazon biome, where available data are predominantly characterized by generalized legacy maps at a coarse scale of 1:250,000.

Conventional soil surveys in the Amazon region, such as those derived from the RADAMBRASIL project, which are still used today, represent pedological knowledge of fundamental importance. These legacy data are capable of organizing and materializing the mental models of experienced pedologists (soil maps), while also accounting for spatial distribution based on soil–landscape relationships (Medeiros et al., 2024).

The integration of quantitative machine learning methods (mathematical and statistical) through the SCORPAN model (McBratney et al., 2003) and DSM has enabled the prediction of soil distribution patterns across the landscape, as well as the quantification of associated error through accuracy metrics (Carvalho Monteiro et al., 2023). This strategy has the potential to facilitate the dissemination of more detailed surveys, which are still scarce or limited to small areas in Brazil (Mendonça Santos and Santos, 2006), particularly in the Amazon region.

Among machine learning approaches, Favrot et al. (1996) and Lagacherie et al. (1996; 2001) proposed the reference area method, based on the hypothesis that most soil classes present in a given region can be sampled within a reference area. This transfer of information on the spatial variability of soils between a reference area and another unsampled region has been widely studied (Machado et al., 2019; Ferreira et al., 2023; Nenkam et al., 2023).

Machine learning represents the interface between conceptual pedological knowledge and quantitative soil-forming factors, which are represented by environmental covariates (Jenny, 1941). However, there is no consensus regarding the set of environmental covariates and learning algorithms capable of capturing the spatial associations addressed in extrapolation modeling, whose pedogenetic processes are highly variable across the landscape (Arrouays et al., 2020).

Regarding machine learning, empirical models using individual algorithms have been applied for the extrapolation of digital soil mapping models (Angelini et al., 2020; Gelsleichter et al., 2023; Ye et al., 2026). However, ensemble learning, which integrates multiple algorithms, can increase predictive stability and reduce uncertainty in predictive models (Minasny and Hartemink, 2011; Taghizadeh-Mehrjardi et al., 2021).

Although many studies have investigated the use of the reference area approach and legacy data (Sulieman et al., 2023; Khosravani et al., 2024; Rodrigues et al., 2025), we found no studies that integrate, in addition to the reference area approach, ensemble learning and external validation in the Amazon biome.

In this context, our hypothesis is: soil knowledge is transferable in the Amazonian environment through ensemble learning algorithms that incorporate regional environmental covariates to capture complex soil-landscape relationships typical of hydromorphic environments. To address this question, our study established the following objectives: (i) to identify the most suitable machine learning models and environmental covariates for representing soil–landscape relationships in a reference area; (ii) to evaluate the transferability of soil-landscape relationships learned in a 1:100,000 reference area to an extrapolation area with a current scale of 1:250,000, aiming to increase spatial resolution and improve the delimitation and prediction of soil mapping units; and (iii) to assess the effectiveness of the reference area approach for digital soil mapping in data-scarce regions of the Eastern Amazon.

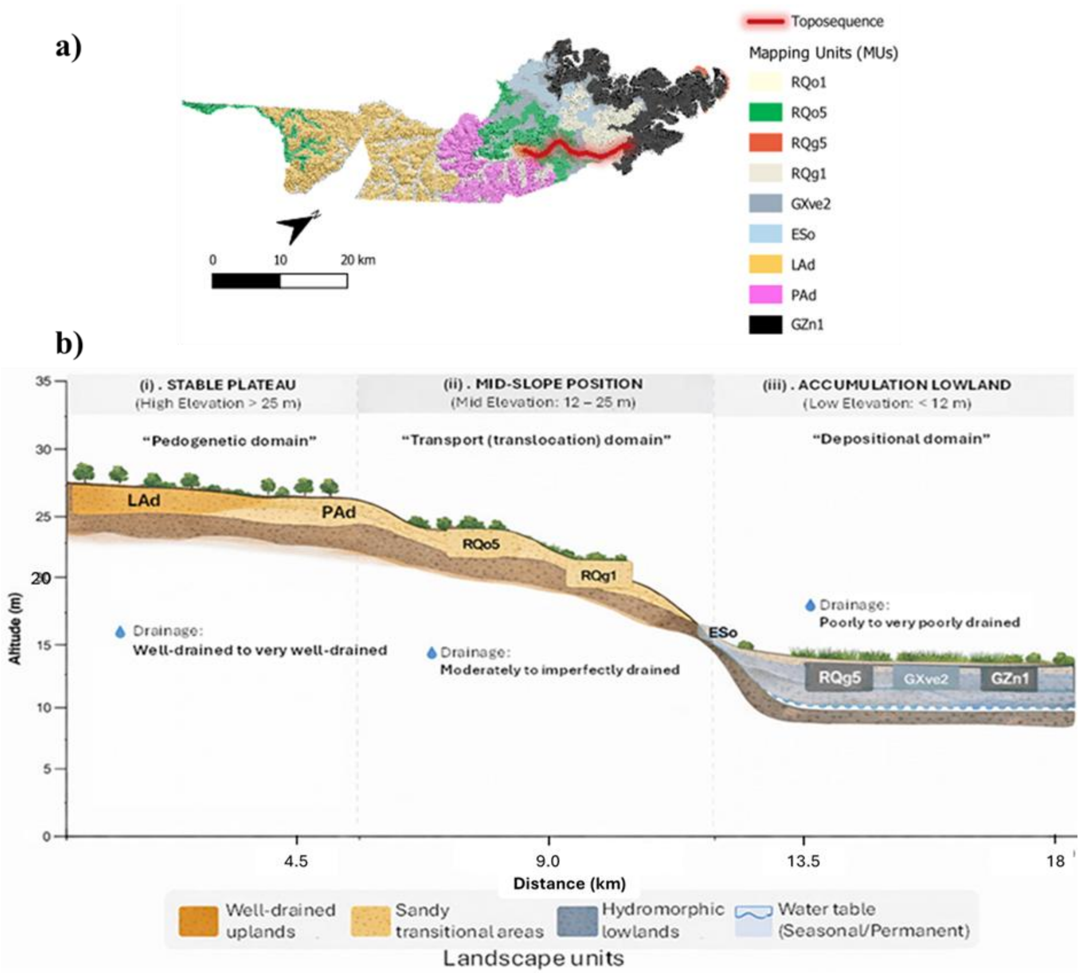
3.3 Materials and methods

3.3.1 Basic idea

This study is based on the possibility of transferring soil–landscape models from a reference area, which has a conventional soil survey, to another area in the region that lacks a more detailed soil survey using the knowledge on soil pattern of a previously mapped reference area.

The rules applied to the modeling carried out in this study were derived from an analysis of local soil-landscape relationships as a conceptual pedological model developed by Farias et al. (2026), which were confirmed and predefined, thus allowing these mapping units (MUs) to be extrapolated to areas with similar environmental conditions. Within this framework, based on the environmental characteristics. Based on a toposequence encompassing the main soil MUs, it was possible to delineate the distinct soil–landscape relationships (Fig. 14a). The LAd and PAd, the following landscape units can be distinguished: the Flattened Plateau Surface, which represents the main elevated core (flat summit) with predominantly flat relief and altitudes above 25 m; the Plateau Slopes, which comprise irregular surfaces, often strongly eroded, forming the margins of the plateau along its eastern boundaries or extending into its valleys, with predominantly sandy soils and altitudes ranging from 12 to 25 m; and the Lowlands, which consist of depressed areas shaped as gentle troughs dissecting the flat summit plateau, with altitudes below 12 m and characterized by hydromorphic soils (Gleissolos and NEOSSOLOS QUARTZARÊNICOS Hidromórficos), as summarized in Fig. 14b.

Figure 14 - Schematic representation of the studied toposequence and its distinct soil–landscape relationships across the landscape units in northeastern Pará, Eastern Amazon. a) location of the toposequence in the different UMs. b) The diagram of soil–landscape relationships derived from the reference area can be interpreted as a three-domain catena: (i) a stable plateau dominated by pedogenesis, (ii) a transitional slope characterized by lateral material transport, and (iii) a lowland accumulation zone controlled by depositional and hydromorphic processes.



Source: The author (2026).

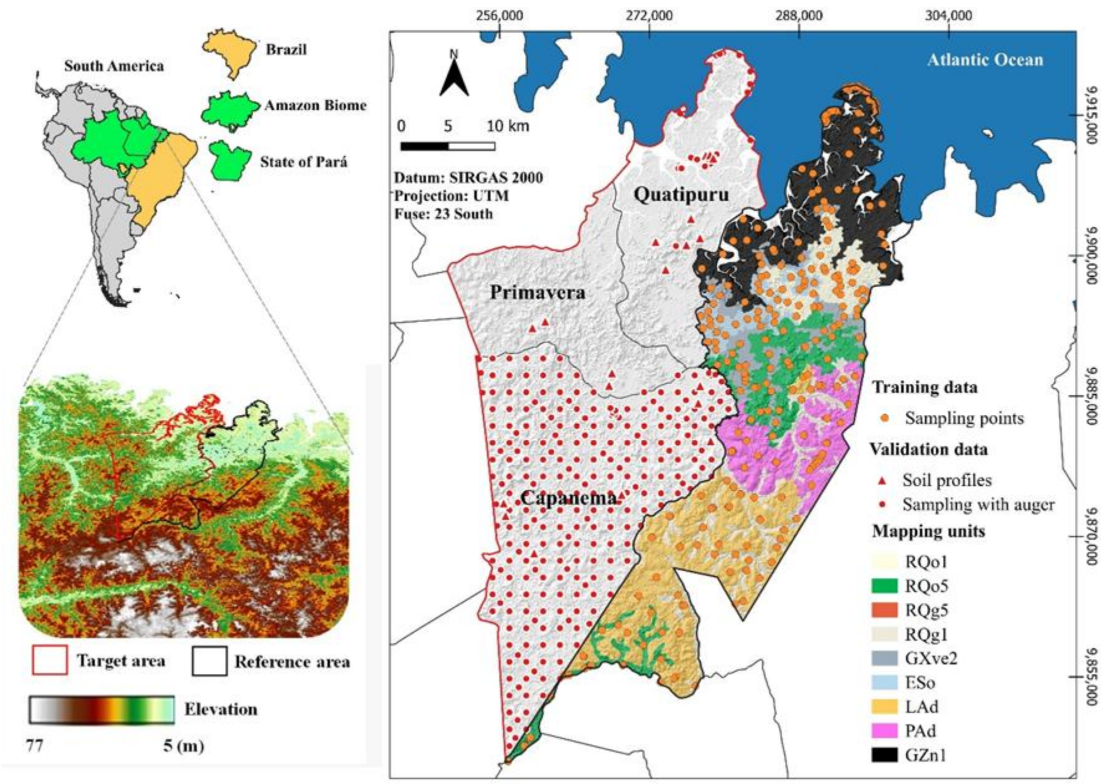
Based on this framework, and on a methodology involving the evaluation of environmental covariate importance, performance and uncertainty metrics (internal validation), and external validation using field observation points in the extrapolated area, we propose that the ensemble learning approach has strong potential for predictive modeling, achieving higher accuracy in mapping units associated with hydromorphic environments.

External validation was employed to evaluate the feasibility of spatial transferability between the reference and extrapolation areas, testing whether supervised classification and predictive modeling can effectively leverage the tacit knowledge embedded in the legacy map and transfer it with acceptable accuracy.

3.3.2 Knowledge base and reference area

The study area is located in northeastern State of Pará, within the Amazon biome. The municipality of Tracuateua, with an area of 868 km², constituted the reference area (model training area), while the extrapolation area covers 1,181 km² and includes the municipalities of Capanema, Primavera, and Quatipuru, as shown in Fig. 15.

Figure 15 - Location of the study area. Detail of the reference area (Tracuateua) and the target area (Capanema, Primavera, and Quatipuru).



Source: The author (2026).

The climate of the study area is classified as Am, tropical monsoon climate according to the Köppen classification, with annual precipitation varying between 1,680 mm from January to May (Amazonian winter) and 838 mm from June to December (Amazonian summer), with an average of 2,581 mm. Mean temperatures range between 26.6 °C and 28.1 °C in the Amazonian winter and summer, respectively (Alvares et al., 2013).

Elevation in the study area ranges from −11 to 66 m, with predominantly flat to gently undulating relief. The geomorphology is characterized by flat floodplain areas (hydromorphic environments) and fluvio-marine plains, as well as flat and gently undulating terrains, classified as a low Amazon plateau (Oliveira Junior et al., 1999).

The representative geological formations common to both the reference and extrapolation areas correspond to the Barreiras Group, sedimentary deposits, and the Tracuateua Formation. Among these parent materials, areas under the Barreiras Group are notable, where Latossolos and Argissolos predominate, as well as sedimentary deposits represented by recent alluvial deposits that occur as narrow and sometimes discontinuous strips along drainage channels. In these environments, soils such as Espodossolo, Gleissolos, and hydromorphic Neossolos are formed under predominantly hydromorphic conditions, according to the Brazilian Soil Classification System (SiBCS) (Santos et al., 2025).

We used legacy data from a conventional soil survey at a 1:100,000 scale (Oliveira Júnior et al., 1999), at the third taxonomic level (Great Group), grouped into nine mapping units (MUs) (Table 6), according to the Brazilian Soil Classification System (SiBCS) (Oliveira Junior et al., 1999; Santos et al., 2025).

Table 6 - Mapping units (MU) - WRB/FAO correspondence and SiBCS legend, spatial extent of the mapping units in the reference area (legacy map), levels of detail (LD1, LD2), and number of sampling pixels.

MU	**WRB/FAO correspondence**	**Area (km²)**	**LD 1**		**LD 2**	
			SiBCS legend	**Sampling pixels number**	**SiBCS legend**	**Sampling pixels number**
NEOSSOLOS QUARTZARÊNICOS Órticos	Arenosols	58.18	R	109	RQo1	28
NEOSSOLOS QUARTZARÊNICOS Órticos + ARGISSOLOS VERMELHO-AMARELOS Distróficos	Arenosols + Acrisols	107.79			RQo5	36
NEOSSOLOS QUARTZARÊNICOS Órticos + NEOSSOLOS QUARTZARÊNICOS Hidromórficos + GLEISSOLOS HÁPLICOS Tb Distróficos	Arenosols + Gleysols	3.27			RQg5	23
NEOSSOLOS QUARTZARÊNICOS Hidromórficos	Fluvisols	74.17			RQg1	22
GLEISSOLOS HÁPLICOS Ta Eutróficos	Gleysols	56.18	G	63	GXve2	30
GLEISSOLOS SÁLICOS Sódicos	Solonchaks	177.33			GZn1	33
ESPODOSSOLOS FERRILÚVICOS Órticos	Podzols	45.59	E	36	ESo	36
LATOSSOLOS AMARELOS Distróficos	Ferralsols	212.26	L	30	LAd	30
ARGISSOLOS AMARELOS Distróficos	Acrisols	102.94	P	32	PAd	32

Source: The author (2026).

Based on the legacy map, we implemented a taxonomic generalization for the MUs, reducing them from nine to five, aiming to improve the performance of the prediction models under the hypothesis of better correspondence with the landscape. Thus, two levels of detail (LD) were adopted: LD1, comprising five MUs that include simple and compound taxonomic units at the Order level; and LD2, comprising the nine MUs represented in the legacy map.

For better understanding, for both simple and compound mapping units (MUs), we adopted the identification nomenclature used in the Brazilian Soil Classification System (SiBCS), for both LD1 and LD2. In LD1, the MUs represent the taxonomic level of Order: NEOSSOLO (R), GLEISSOLO (G), ESPODOSSOLO (E), LATOSSOLO (L), and ARGISSOLO (P). In LD2, the MUs represent the taxonomic level of Great Group: NEOSSOLOS QUARTZARÊNICOS Órticos (RQo1), NEOSSOLOS QUARTZARÊNICOS Órticos (RQo5), NEOSSOLOS QUARTZARÊNICOS Órtico (RQg5), NEOSSOLOS QUARTZARÊNICOS Hidromórficos (RQg1), GLEISSOLOS HÁPLICOS Ta Eutróficos (GXve2), GLEISSOLOS SÁLICOS Sódicos (GZn1), ESPODOSSOLOS FERRILÚVICOS

Órticos (ESo), LATOSSOLOS AMARELOS Distróficos (LAd) e ARGISSOLOS AMARELOS Distróficos (PAd) (Table 6).

To ensure the representativeness of all different mapping units (MUs) contained in the legacy map for model training, a stratified sampling approach was implemented. A total of 270 pixels were selected to encompass every individual MU present in the study area, with sample points distributed across each soil class stratum to capture their full spatial diversity (Table 6). Consequently, each selected pixel directly represents the specific MU associated with its location, following Cremon et al. (2021). To understand soil–landscape relationships in the training area, field surveys were conducted at the locations corresponding to the 270 sampled pixels, using an auger to a depth of 1.50 m. To avoid transition zones, sampling points were allocated at a minimum distance of 60 m from the boundaries between MUs.

3.3.3 Environmental covariates

As input data for model training, we used 15 environmental covariates based on the SCORPAN model (McBratney et al., 2003), namely: climate, organisms, topography, parent material, and time. To ensure spatial consistency across all datasets, these covariates were resampled using the nearest neighbor method and standardized to a uniform spatial resolution of 30 x 30 m, matching the pixel size of the primary digital elevation model.

For the climate covariate, we used mean annual precipitation (AAP), obtained from the Google Earth Engine (GEE) using the UCSB-CHG/CHIRPS/DAILY collection, with a time series from 1981 to 2023 (Gorelick et al., 2017); and paleoclimate variables, including annual precipitation (PR_22k) and mean annual temperature (Tmean_22k) estimated for 22,000 years ago, with spatial resolutions of approximately 1 km and 5 km, respectively.

Representing the organism factor, we used the Soil Adjusted Vegetation Index (SAVI), as proposed by Huete (1988), and daytime Land Surface Temperature (LST), derived from multispectral data from the Landsat 8 OLI sensor (USGS) and from the daily time series (2000–2024) of the MODIS/061/MOD11A1 collection available on the GEE platform (Gorelick et al., 2017).

As terrain attributes derived from two digital elevation models (DEMs), we used a digital surface model (DSM) with 30 m resolution, developed by Copernicus DEM – GLO-30 (Ecosystem, 2024), and a digital terrain model (DTM) for South America (ANADEM), also with 30 m resolution (Laipelt et al., 2024). From these datasets, and with the aid of SAGA GIS

9.3.1 software (Conrad et al., 2015), we generated terrain-related variables representing soil-forming factors.

The terrain-related covariates included Channel Network Base Level (CNBL) (Bock and Köthe, 2008), Drainage Density (DD) (Christofoletti, 1980), Multiresolution Index of Ridge Top Flatness (MRRTF), Multiresolution Index of Valley Bottom Flatness (MRVBF) (Gallant and Dowling, 2003), Profile Curvature (PFC) (Florinsky, 2012), Relative Slope Position (RSP) (Conrad, 2005), and elevation (Florinsky, 2012). In addition, Geomorphons (GEOM) (Jasiewicz and Stepinski, 2013) was used to represent the time (age) covariate.

Geology and the Iron Oxide Ratio (IOR), calculated according to Singer (1981), were used to represent the parent material. These covariates were obtained, respectively, from a geological map at a 1:250,000 scale (Instituto Brasileiro de Geografia e Estística – IBGE, 2023) and from multispectral data from the Landsat 8 OLI sensor (U.S. Geological Survey – USGS) (Landsat 8 | U.S. Geological Survey, 2023).

3.3.4 Dataset preparation

This study used four datasets, based on two different digital elevation models and two taxonomic levels resulting in the following combinations:

(a) Dataset 1: level of detail 1 (LD1), representing the taxonomic Order level of the MUs, using a digital surface model (DSM) to derive terrain-related environmental covariates;

(b) Dataset 2: level of detail 1 (LD1), representing the taxonomic Order level of the MUs, using a digital terrain model (DTM) to derive terrain-related environmental covariates;

(c) Dataset 3: level of detail 2 (LD2), representing the taxonomic Great Group level of the MUs, using a digital surface model (DSM) to derive terrain-related environmental covariates; and

(d) Dataset 4: level of detail 2 (LD2), representing the taxonomic Great Group level of the MUs, using a digital terrain model (DTM) to derive terrain-related environmental covariates.

A total of 51 covariates related to the soil-forming factors of the predominant soil classes in the study area were initially considered (Appendix Quadro A1). However, in order to reduce the number of covariates to the most appropriate set and facilitate the interpretation of the results, we applied two data mining steps to select the most parsimonious set of environmental predictors, resulting in lower prediction error by avoiding overfitting.

3.3.5 Feature engineering

In the first step, to avoid the inclusion of redundant predictor covariates in the model, a correlation analysis was performed to address multicollinearity among continuous covariates. Using the "findCorrelation" function (Kuhn, 2007), with a pairwise correlation coefficient (Spearman's corretalation) $r \geq |0.95|$, the covariates were evaluated, and those with the highest absolute correlation were removed.

In the second step, the Recursive Feature Elimination (RFE) algorithm, implemented in the caret package (Kuhn, 2007), was used to identify the optimal subset of predictor covariates based on their importance (predictive power). Based on the principle of backward selection, RFE iteratively trains a Random Forest model, ranks the covariates according to their importance, and systematically eliminates the least significant variables (Kuhn and Johnson, 2013). This iterative refinement continues until a threshold of the ten most influential covariates is reached.

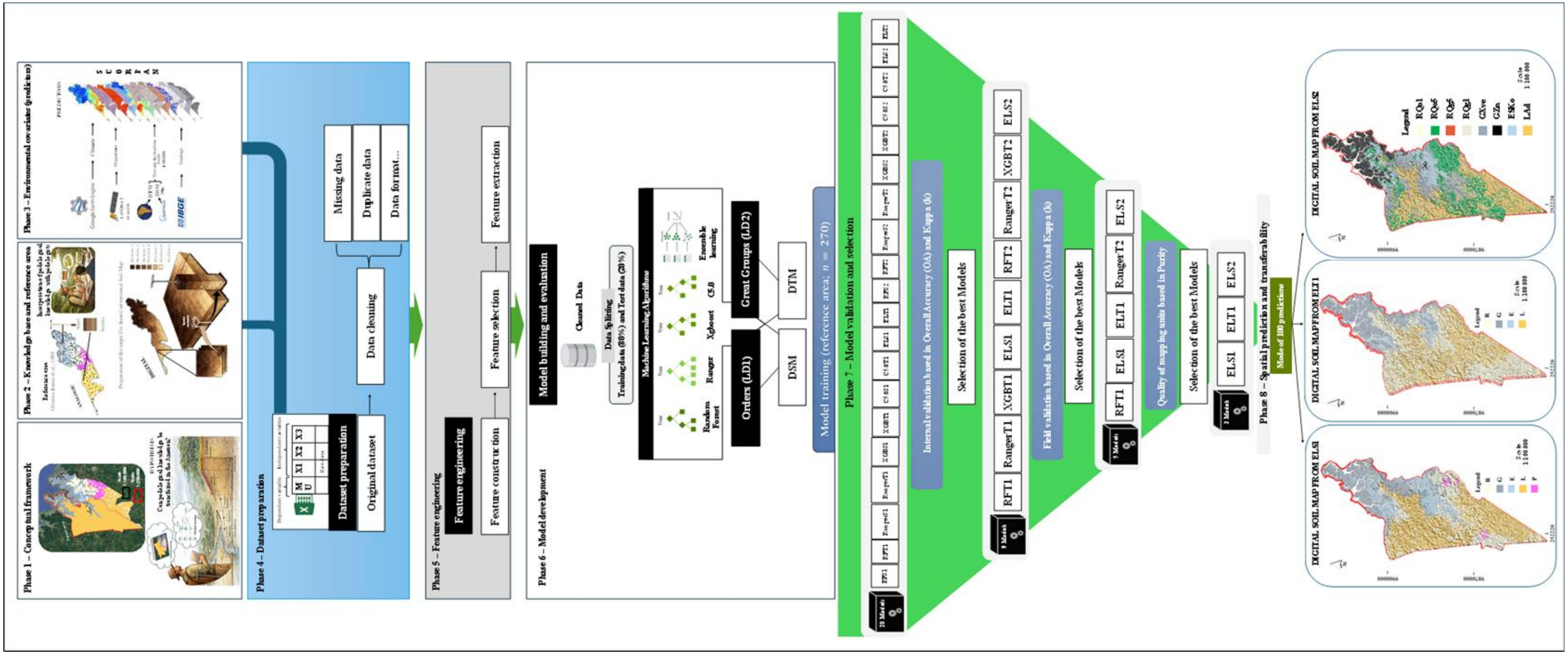
3.3.6 Hyperparameter optimization

The optimization of machine learning algorithms involved the systematic tuning of key hyperparameters to improve the overall accuracy of the different algorithms evaluated. Despite differences in nomenclature, these models share fundamental structural parameters, such as those related to the complexity of individual nodes. For the Random Forest and Ranger implementations, the focus was on defining the number of trees (ntree = 1000 or num.trees = 1000, respectively) and the number of variables randomly sampled at each split (mtry). The latter hyperparameter was determined through a comparative tuning process, in which the Kappa index was used as the primary selection criterion due to its sensitivity in assessing agreement among classes.

In the XGBoost and C5.0 algorithms, similar concepts are applied through the number of boosting iterations (nrounds = 1000 or trials = 100) and tree depth controls, which, analogous to the mtry parameter, regulate the model's ability to capture nonlinear interactions between soil and landscape.

The modeling phases carried out in this study were as follows: i) Conceptual framework, ii) Knowledge base and reference area, iii) Environmental covariates (predictors), iv) Dataset preparation, v) Feature engineering, vi) Model development, vii) Model validation and selection e viii) Spatial prediction and transferability são apresentadas por meio de um flowchart (Fig. 16)

Figure 16 - Methodological flowchart illustrating the workflow adopted for digital soil mapping based on a reference area approach. The framework includes: (i) conceptual framework and definition of the hypothesis of soil–landscape knowledge transfer; (ii) characterization of the reference area and incorporation of pedological knowledge; (iii) acquisition and processing of environmental covariates; (iv) dataset preparation, including data cleaning and standardization; (v) feature engineering (feature construction, selection and extraction); (vi) model development using machine learning and ensemble techniques; (vii) model validation and selection based on internal validation, field validation, and mapping unit purity; and (viii) spatial prediction and transferability to a neighboring area, resulting in the final digital soil map.



Source: The author (2026).

3.3.7 Model development

The predictions and spatial extrapolations of the target mapping units (MUs) were performed using the R programming language (version 4.3.3; “R Core Team,” 2024b), employing the Random Forest, Ranger (an implementation of Random Forest), XGBoost, and C5.0 algorithms, as well as an ensemble learning (EL) approach combining these algorithms.

The 270 sampled pixels were divided into two datasets, with 80% used for training with 5-fold cross-validation and 20% reserved for testing, using the createDataPartition function from the caret package, which performs a more representative random sampling.

Random Forest is an algorithm based on an ensemble of decision trees, developed as an extension of the Classification and Regression Trees (CART) method to optimize the predictive performance of discrete and categorical variables (Breiman, 2001). The algorithm uses predefined environmental covariates to establish complex relationships and predict target variables (such as soil classes) in unsampled locations. Unlike conventional CART, Random Forest constructs multiple trees independently without pruning, where each tree is generated from subsets of data selected through the bootstrap method (Efron and Gong, 1983). At the end of the process, each tree in the forest classifies the target, and the model determines the final prediction through a voting system, selecting the class by majority vote (classification).

The Ranger package is a high-performance implementation of the Random Forest algorithm, designed for computational efficiency with high-dimensional data by using an accelerated parallel processing approach within the context of decision trees (Wright, 2015).

XGBoost algorithm work through an incremental learning process, in which each subsequent decision tree is specifically designed to correct the residuals and prediction errors of the preceding ensemble. In addition, this algorithm introduces advanced features such as built-in regularization to prevent overfitting, native parallelization for high computational efficiency, and automated handling of missing values (Chen and Guestrin, 2016).

Among the strengths of the C5.0 algorithm is its inherent robustness in handling noise and missing values in environmental datasets, which are common in large-scale soil surveys. By allowing differential weighting of classification errors, the model prioritizes the identification of soil units that are pedologically critical or spatially rare, ensuring that the final map preserves landscape coherence and technical usefulness rather than focusing solely on statistical optimization (Arif, 2018; Quinlan, 2014).

Although predictive digital soil mapping (PDSM) conventionally employs individual machine learning algorithms, the ensemble learning (EL) approach can leverage the combined strengths of multiple models (Ganaie et al., 2022). In this study, a bagging (Bootstrap Aggregating) technique was used to enhance model robustness and reduce variance by combining similar tree-based models (Maclin and Opitz, 1999). To obtain a single map per base classifier, a modal was performed on the 100 generated maps, thus providing a more refined level of confidence for soil class assignments.

For each algorithm (Fig. 16), 100 iterations were generated, with each map resulting from an independent sampling process. The best model was selected based on the highest Kappa index, and the final PDSM was obtained by calculating the mode across all 100 maps).

3.3.8 Model validation and seletion

3.3.8.1 Accuracy and uncertainty assessment and model validation

After extrapolating the PDSM (n = 100 maps) to the study area (reference area and extrapolation area), we evaluated the optimal number of maps generated using the elbow method, based on the agreement index of the predicted mapping units (MUs) across the maps.

We evaluated model performance on the test dataset using the mean values of the Kappa index (k) (Cohen, 1960) (Eq. 1) and Accuracy (Eq. 2) calculated from the n = 100 PDSM generated at the end of the processing of each model. The Kappa index assesses the model's performance by quantifying the level of agreement beyond that expected from purely random factors, allowing for a more robust analysis of the overall certainty between predicted and observed model units, whereas Accuracy represents the degree of agreement between correctly classified MUs and the total number of observations in the test dataset (270 point sampling).

$$k = \frac{P_o - P_a}{1 - P_a} \qquad \text{(Eq. 1)}$$

Where, P_o is the proportion of correctly classified samples, and P represents the probability of random agreement. The results range from −1 to 1, indicating increasing accuracy as the values approach 1.

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N} \qquad \text{(Eq. 2)}$$

Where, TP refers to the correctly classified MUs of class i, C is the total number of MUs, and N is the total number of observations in the test dataset.

From the n = 100 PDSM generated for each model, we evaluated the statistical significance of the Accuracy and Kappa (k) metrics using the Tukey Honest Significant Difference (HSD) test (Eq. 3) at a significance level of $p < 0.01$.

$$Tukey's\ HSD = q\sqrt{\frac{MSE}{n_c}} \quad \text{(Eq. 3)}$$

Where, HSD is the minimum difference between two group means required for statistical significance; q is the critical value from the table; MSE is the mean squared error; and n_c is the sample size for each group.

We also evaluated the importance of environmental covariates for each predictive model using the varImp method from the caret package using R software (R Core Team, 2024a). This metric is calculated based on the increase in out-of-bag (OOB) error after removing a predictor variable, while all other variables are kept constant, as described by Carvalho Monteiro et al. (2023).

We also evaluated the performance of the prediction models using field data, in which soils were classified according to the Brazilian Soil Classification System (SiBCS) (Santos et al., 2025), based on an error matrix constructed from 381 validation points (19 soil profiles and 362 auger observations), using the R programming language (R Core Team, 2024a). For the allocation of validation points, the following criteria were used: (i) coverage of the MUs; and (ii) selection of points outside the training area, based on Grinand et al. and Mello et al. (2008; 2021).

From the error matrix, the F1-score (Eq. 4), which is the harmonic mean of user's accuracy (UA) and producer's accuracy (PA), overall accuracy (OA) (Eq. 5), representing the percentage of samples whose soil classes identified in the field correspond to those presented in the prediction map, and the Kappa index were calculated.

$$F1\ score = 2\left(\frac{UA \times PA}{UA + PA}\right) \quad \text{(Eq. 4)}$$

Where, UA (commission error) is the percentage of samples in a predicted MU on the map that correspond to the same MU in the field, and PA (omission error) expresses the probability that an MU is correctly classified in the predicted map.

$$OA = \frac{\sum_{i=1}^{k} n_{ii}}{N} \quad \text{(Eq. 5)}$$

Where n_{ii} is the number of correctly classified samples for class i, corresponding to the diagonal of the confusion matrix; k is the total number of MUs; and N is the total number of samples (sum of all elements in the matrix).

3.3.8.2 Post-processing: Minimum Mapping Area (MMA)

The transition from pixel-based digital predictions to a functional soil map requires spatial generalization consistent with both pedometric principles and conventional soil survey standards. This transition is necessary because machine learning models evaluate and predict MUs independently for each individual pixel. In contrast, practical soil conservation and land-use planning require continuous, cohesive cartographic polygons – MUs - rather than fragmented cellular outputs. Therefore, spatial generalization acts as a bridge to convert independent grid-cell data into interpretable and management-oriented soil polygons. While the ensemble models produced predictions at a spatial resolution of 30 m (pixel area of 900 m²), such fine-resolution outputs commonly exhibit high-frequency spatial noise (i.e., “salt-and-pepper” patterns), which are not compatible with the conceptual definition of soil mapping units at medium scales, according to the Minimum Mapping Area (MMA) definition in the IBGE Technical Manual of Pedology (IBGE, 2015).

In the framework of digital soil mapping, this step represents a necessary reconciliation between prediction support and mapping support, as discussed by McBratney et al. (2003) and further formalized by Wadoux et al. (2020). Pixel-based predictions capture local variability, whereas soil maps require spatially coherent units that reflect pedogenetic processes and landscape organization.

Following cartographic conventions adopted by the Brazilian Soil Classification System (Santos et al., 2025) and the IBGE (2015), the MMA was defined as 40 ha. This threshold is derived from the conventional minimum delineation limit of 0.4 cm² on the map surface, which corresponds to 400,000 m² (40 ha) at a scale of 1:100,000, according to Eq. (6).

$$MMA\ (ha) = 0.4 \times \left(\frac{Scale}{10{,}000}\right)^2 \quad \text{(Eq. 6)}$$

Although this MMA is larger than those adopted in some digital soil mapping studies (Dalmolin et al. 2020; Giassson et al., 2015; Rossiter, 2000), it reflects the need to ensure compatibility between high-resolution predictive outputs and the operational requirements of soil survey mapping units at this scale. This choice prioritizes cartographic legibility, pedological interpretability, and consistency with legacy soil maps.

At the adopted spatial resolution, each MMA contains approximately 444 pixels, providing a robust statistical basis for spatial aggregation. A Majority Spatial Filter was applied, whereby the final class assigned to each mapping unit corresponds to the most frequent class among its constituent pixels. This approach acts as a spatial consensus mechanism, reducing local prediction variance while preserving dominant soil–landscape patterns.

Importantly, this generalization procedure does not merely smooth the data but aligns the spatial structure of predictions with the concept of soil bodies as continuous landscape features. As highlighted by Wadoux et al. (2020), such post-processing steps are essential to ensure that digital soil mapping outputs are both statistically sound and geomorphologically meaningful.

The sampling design further supports the reliability of the resulting mapping units. The overall sampling density corresponds to a High-Intensity Reconnaissance Survey according to EMBRAPA (1995), with approximately one observation per 310 ha. Additionally, the density of soil profiles (~1 per 5,624 ha) exceeds standard recommendations for this survey level, reinforcing the pedological representativeness of the dataset.

Together, the integration of high-resolution digital predictions, statistically supported aggregation, and established cartographic criteria enables the transformation of raw classification outputs into a consistent and reliable soil map suitable for regional-scale applications.

3.3.8.3 Map purity assessment

To evaluate the reliability and taxonomic consistency of the soil mapping units (SMUs), the Map Purity (MP) (Eq. 7) was calculated as a measure of the predictive uncertainty, following the framework established by Lark et al. (2022). The MP was derived from the Confusion Index (CI) (Esfandiarpour-Boroujeni et al., 2020), which assesses the model's "hesitation" between the two most likely classes for each pixel. The MP was calculated as follows:

$$MP = 1 - CI \qquad \text{(Eq. 7)}$$

The Mean Map Purity (MPm) for each soil mapping unit was then aggregated within the Minimum Mapping Area (MMA) of 40 ha using zonal statistics. This metric quantifies the taxonomic consistency of the final map, where values closer to 1.0, or expressed as a percentage (100%), indicate high predictive certainty and minimal inclusion of subdominant classes within the delineated polygons.

The calculated MPm values were qualitatively interpreted based on an adaptation of the framework proposed by Bregt (1992). The soil mapping units were categorized into four reliability classes according to their percentage purity: High (> 80 %), Moderate (60–80 %), Low (40–60 %), and Very Low (< 40 %). Pixels or areas where purity could not be determined due to edge effects or lack of predictive support were designated as “Not suitable.” This classification enables a standardized assessment of the map’s taxonomic integrity, providing end users with a clear indication of the spatial confidence associated with each delineated soil polygon.

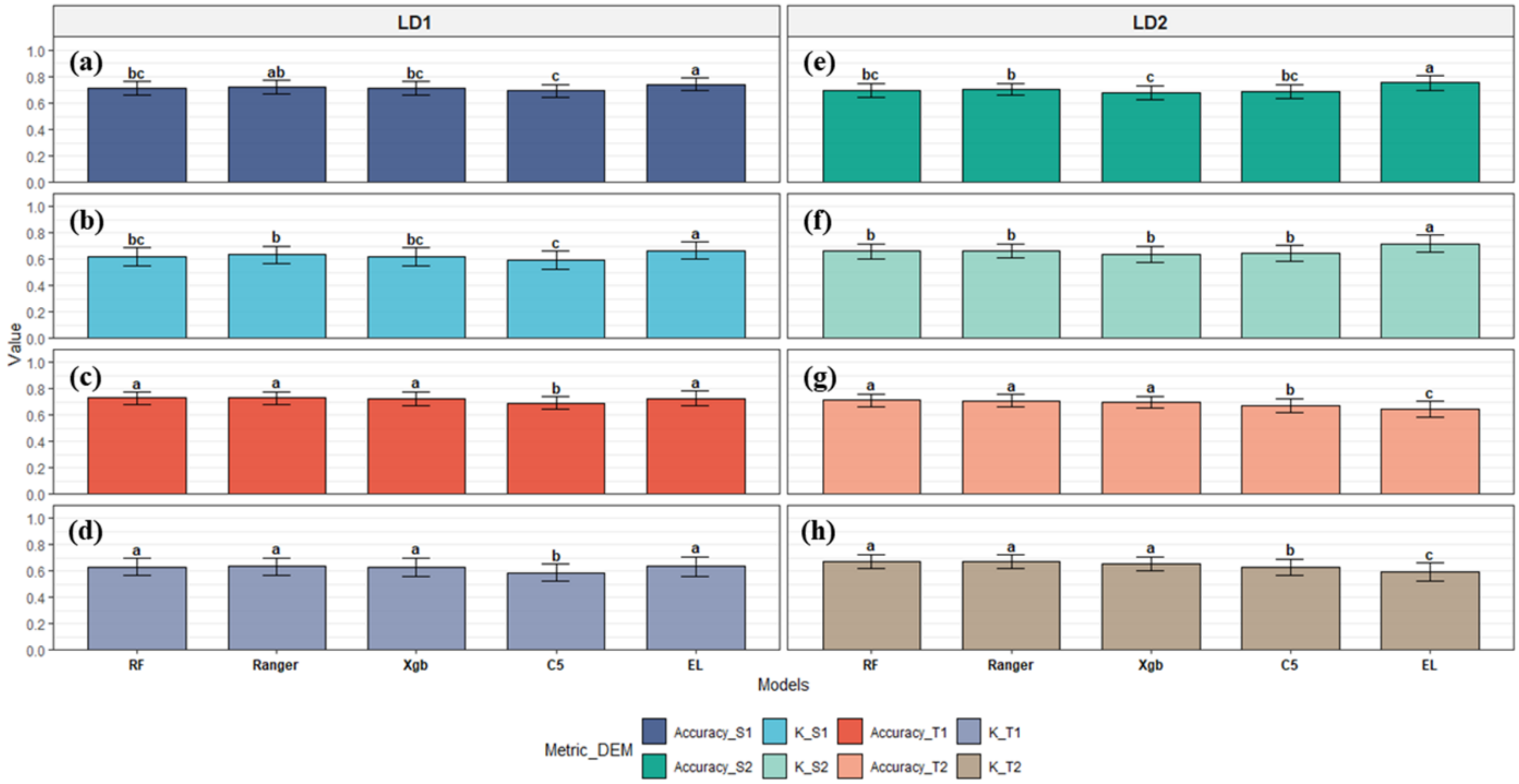
3.4 Results and discussion

3.4.1 Internal validation and performance evaluation of the models

The elbow method was used to evaluate the stability of the agreement index among the 100 maps generated at the end of each processing step for the algorithms studied, as shown in Appendix A2. This analysis indicated that the agreement index reached stability within a range of 30 to 50 generated maps. Thus, although the execution of 100 maps was computationally more demanding—resulting in increased processing time and greater memory requirements—it provided greater stability of agreement for each predictive model evaluated.

Following this stabilization, the mean Accuracy and Kappa (k) values were compared across the 100 PDSM realizations generated for MU prediction (Fig. 4). According to the statistical test (Tukey HSD test, $p < 0.01$), the highest mean values of Accuracy and Kappa were obtained for the following models: ELS1 (0.74 and 0.67), ELT1 (0.73 and 0.63), RFT1 (0.73 and 0.63), RangerT1 (0.73 and 0.63), XgbT1 (0.72 and 0.63), ELS2 (0.75 and 0.72), RFT2 (0.71 and 0.67), RangerT2 (0.71 and 0.67), and XgbT2 (0.70 and 0.66) (Fig. 17).

Figure 17 - Machine learning models for predicting soil mapping units at LD1 and LD2 levels of detail. (a) and (b) mean accuracy and Kappa (κ) of the predictive models using terrain attributes derived from DSM (S) at LD1; (c) and (d) mean accuracy and Kappa (κ) of the predictive models using terrain attributes derived from DTM (T) at LD1; (e) and (f) mean accuracy and Kappa (κ) of the predictive models using terrain attributes derived from DSM (S) at LD2; (g) and (h) mean accuracy and Kappa (κ) of the predictive models using terrain attributes derived from DTM (T) at LD2, after n = 100 PDSM generated for each model. Different letters among models indicate statistically significant differences at $p < 0.01$ according to the Tukey HSD test.



Source: The author (2026).

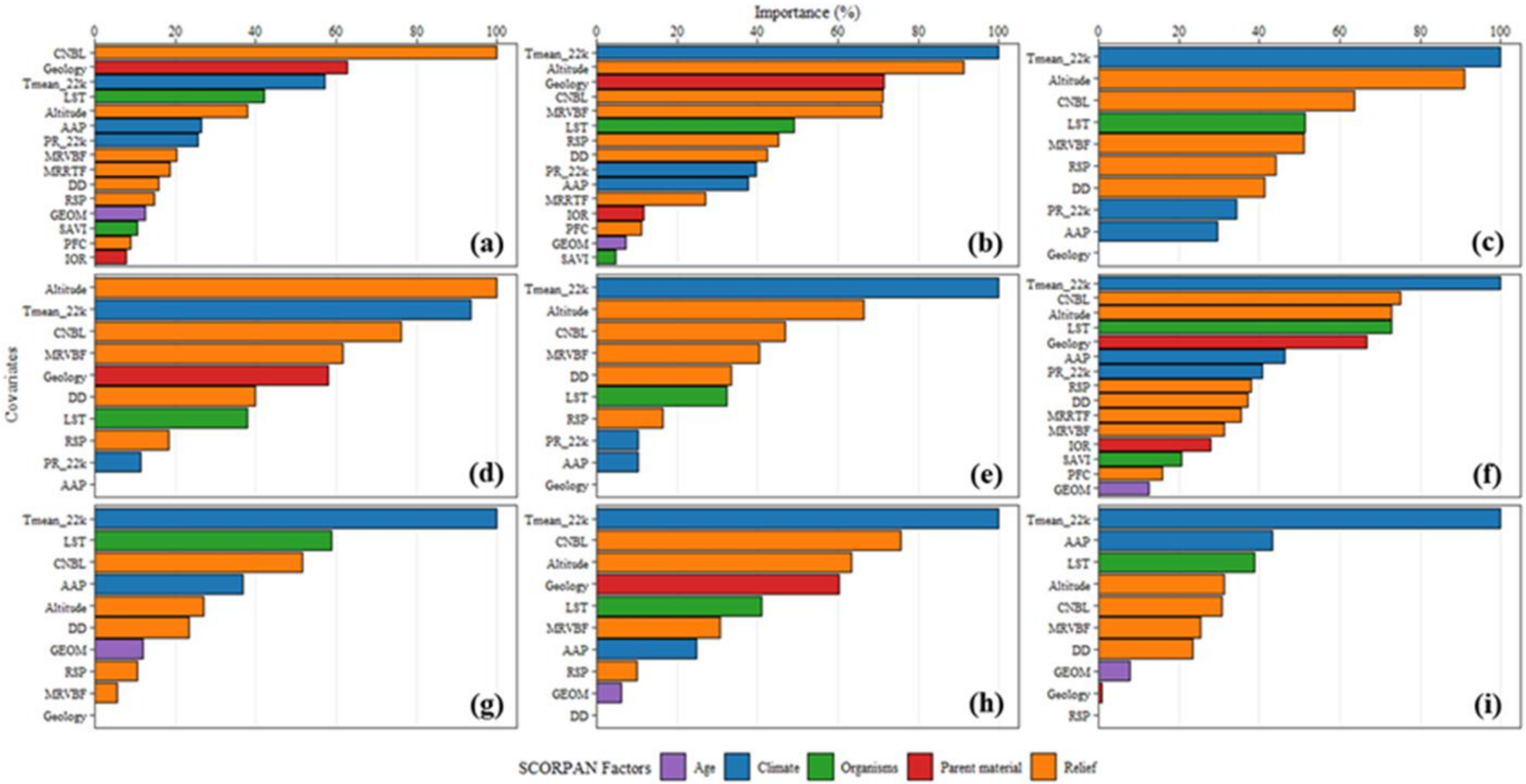
A key behavioral divergence emerged regarding the digital elevation inputs (Fig. 4). Models based on individual machine learning algorithms showed superior performance when using terrain attributes derived from the digital terrain model (DTM). Conversely, ensemble learning (EL) frameworks achieved their peak performance when coupled with terrain attributes derived from the digital surface model (DSM), across both LD1 and LD2 categorical levels. This indicates that EL methods possess a superior capacity to capture complex soil–landscape relationships and process surface canopy reflections effectively, aligning with the findings of Adeniyi et al. (2023), who reported that ensemble approaches consistently mitigate standalone algorithmic constraints.

This critical role of terrain geometry in driving soil spatial predictions echoes the fundamental paradigms established by Gessler et al. (1995). Furthermore, our internal validation metrics (Accuracy: 0.75; Kappa: 0.72) demonstrate a robust extrapolation capacity. These metrics are highly comparable to the values reported by Machado et al. (2008, 2019), who achieved an Accuracy of 0.82 and a Kappa of 0.75, despite their utilization of a much finer reference area scale of 1:10,000.

3.4.2 Importance of covariates

After isolating the nine best-performing models, the relative importance of environmental covariates for predicting MUs at both detail levels was extracted (Fig. 18). Overall, a dominant trend emerged where paleoclimate (Tmean_22k), catchment area (CNBL), and Altitude stood out as the primary driving variables.

Figure 18 - Covariate importance in the models: (a) ELS1, (b) ELT1, (c) RFT1, (d) RangerT1, (e) XgbT1, (f) ELS2, (g) RFT2, (h) RangerT2, and (i) XgbT2.



Source: The author (2026).

These covariates directly mirror the classic soil-forming factors of climate and relief, capturing how regional pedons reflect intensive tropical pedogenetic processes such as transformation, translocation, and desilication (Carvalho Monteiro et al., 2023). Notably, the higher relevance of paleoclimate (Tmean_22k) compared to current climatic covariates suggests that the modern soil architecture was heavily molded by the more intense hot and humid conditions of the Last Glacial Maximum approximately 22,000 years ago. Furthermore, in areas characterized by low topographic relief, these paleoclimatic erosional and depositional dynamics are often accelerated (Kämpf and Curi, 2012; Resende et al., 2014), reinforcing why terrain-derived attributes consistently dominate predictive soil modeling in highly weathered landscapes (Chen et al., 2022; Pahlavan-Rad et al., 2020).

3.4.3 Accuracy assessment and field validation

External validation yielded peak Overall Accuracy (OA) and Kappa values of 34.4% and 0.11 for ELT1, 34.1% and 0.13 for ELS1, and 33.9% and 0.20 for ELS2 (Table 7). While these absolute values appear low, a comparison with similar reference-area studies across Brazil (Table 8) reveals that such metrics are typical when mapping complex tropical soil taxonomic units characterized by highly variable parent materials, landforms, and localized microclimates (Carvalho Monteiro et al., 2023).

Table 7 - Overall accuracy (OA) and Kappa metrics derived from the error matrices between the predicted map (machine learning) and the reference area (legacy map).

Models	**OA (%)**	**kappa**	**Field sample**
RFT1	33.3	0.10	381
RangerT1	31.2	0.07	
XgbT1	32.0	0.08	
ELS1	34.1	0.13	
ELT1	34.4	0.11	
RFT2	26.1	0.09	380
RangerT2	29.5	0.14	
XgbT2	21.8	0.05	
ELS2	33.9	0.20	

Source: The author (2026).

Under these constraints, a deep understanding of local soil–landscape relationships remains a vital factor in contextualizing accuracy results within the Amazonian biome. This underscores that even when traditional validation metrics like OA and Kappa are low, the underlying pedological consistency of the spatial predictions fundamentally relies on how effectively the input covariates capture the true geomorphological heterogeneity and landscape representativeness (Rossiter and Poggio, 2025).

3.4.3.1 Accuracy assessment of the confusion matrix at LD1

At the first level of detail (LD1), the ELT1 model achieved an OA of 34.4% and a Kappa of 0.11, while the ELS1 model reached 34.1% and 0.13 (Table 7). Although both models displayed comparable performance, the DTM-derived terrain attributes exhibited a superior capacity to capture soil–relief interactions during calibration. This highlights the importance of matching spatial input signals with local terrain variations to capture complex landscape discontinuities (Mello et al., 2021).

Table 8 - Relations between reference area, total area, and OA and Kappa metrics derived from predictive digital soil mapping studies conducted in Brazil.

Study	**References**	**Predominant relief**	**RA (%)**	**RA (km^2)**	**TA (km^2)**	**OA (%)**	**kappa**	**Mapping scale**
1	Arruda et al. (2016)	Undulating	4.5	5	111	83.7	0.81	1:10,000
2	Machado et al. (2019)	Undulating	6.4	1.75	27.19	82	0.75	1:10,000
3	Arruda et al. (2013)	Undulating	10	12	120	77.5	0.74	1:10,000
4	Villela (2023)	Flat	11	80	730	74.6 to 88.8	0.68 to 0.85	1:10,000
5	Silva et al. (2016)	Undulating	0.27	4.85	1,771.9	70.97	0.55	1:12,500
6	Rodrigues et al. (2025)	Undulating	23.5	212	901	75	0.50	1:250,000
7	Wolski et al. (2017)	Gently undulaing	1.5	10	678	66.1	0.36	1:50,000
8	Silva et al. (2013)	Undulating	-	-	1,535	66	0.43	1:50,000
9	Rizzo et al. (2020)	Undulating	-	-	478.8	54.46	0.35	1:100,000
10	Giasson et al. (2006)	Flat	-	-	253	48	0.36	1:50,000
11	ten Caten et al. (2011)	Heterogeneous	16.4	143	874	39.3	0.21	1:50,000
12	Mello et al. (2021)	Undulating	-	-	538	38 to 56	0.25 to 0.31	1:20,000
13	Present study	Gently undulaing	42.4	868.3	2,049	33.9 to 34.4	0.11 to 0.20	1:100,000

RA, reference area; TA, total area; and OA, overall accuracy.

The highest F1-Scores were obtained for the mapping units "G" (Gleissolos) and "L" (Latossolos), reaching 51.8% and 42.4% under ELT1, and 50.0% and 44.4% under ELS1, respectively (Tables 9 and 10). These units also exhibited optimal Producer's Accuracy (PA) values: 57.3% ("G") and 73.0% ("L") for ELT1, and 51.7% ("G") and 50.5% ("L") for ELS1. The strong contribution of the CNBL covariate explains this high discriminative capacity, as it successfully isolates the distinct topographic zones separating very poorly drained hydromorphic environments ("G") in lowlands from stable, well-drained plateaus ("L") on flat summits, a toposequence distribution pattern supported by Bócoli et al. (2025).

Table 9 - Error matrix and accuracy parameters between the predicted mapping units map ELT1 (MUs) and external validation. Field validation using 381 observation points in the extrapolation area.

		Validation points							
	MUs	R	G	E	L	P	∑ rows	UA (%)	F1 Score
Digital soil map	R	33	21	8	55	39	156	21.2	25.3
	G	47	51	2	5	3	108	47.2	51.8
	E	1	2	1	1	2	7	14.3	9.5
	L	24	15	3	46	22	110	41.8	42.4
	P	0	0	0	0	0	0	-	-
	∑ columns	105	89	14	107	66	381		
	PA (%)	31.4	57.3	7.1	43.0	0.0			
								OA 34.4%	k = 0.11

R, NEOSSOLOS QUARTZARÊNICOS (Arenosols); G, GLEISSOLOS (Gleysols); E, ESPODOSSOLOS (Podzols); L, LATOSSOLOS (Ferralsols); P, ARGISSOLOS (Acrisols); UA, user accuracy; PA, producer accuracy; OA, Overall Accuracy; K, kappa index.

Table 10 - Error matrix and accuracy parameters between the predicted mapping units map ELS1 (MUs) and external validation. Field validation using 381 observation points in the extrapolation area.

		Validation points							
	MUs	R	G	E	L	P	∑ rows	UA (%)	F1 Score
Digital soil map	R	23	14	4	41	22	104	22.1	22.0
	G	37	46	1	6	5	95	48.4	50.0
	E	17	9	5	4	4	39	12.8	18.9
	L	28	17	4	54	33	136	39.7	44.4
	P	0	3	0	2	2	7	28.6	5.5
	∑ columns	105	89	14	107	66	381		
	PA (%)	21.9	51.7	35.7	50.5	3.0			
								OA = 34.1%	k = 0.13

R, NEOSSOLOS QUARTZARÊNICOS (Arenosols); G, GLEISSOLOS (Gleysols); E, ESPODOSSOLOS (Podzols); L, LATOSSOLOS (Ferralsols); P, ARGISSOLOS (Acrisols); UA, user accuracy; PA, producer accuracy; OA, Overall Accuracy; K, kappa index.

Conversely, the mapping unit "R" (Neossolos) exhibited poor User's Accuracy (UA), yielding only 21.2% (ELT1) and 22.1% (ELS1) due to a pronounced commission error with MU "L" (Tables 9 and 10). This indicates a systemic tendency of the algorithms to misclassify Latossolos as Neossolos, creating false positives. In the local landscape, these soil classes occur in close spatial proximity, with Latossolos dominating ridge positions and Neossolos occurring along shoulders and steep edges (Santos et al., 2025). Consequently, the 30 m grid resolution represents a spatial bottleneck; utilizing finer-resolution terrain attributes could significantly improve the detection of these abrupt slope-to-ridge soil transitions.

3.4.3.2 Accuracy assessment of the confusion matrix at LD2

At the third categorical level of detail (LD2), ELS2 emerged as the top-performing framework, achieving an OA of 33.9% and a Kappa of 0.20 (Table 11). Within this model, the unit "RQg5" (Neossolos Quartzarênicos hidromórficos) achieved exceptional metrics, with a PA of 90.0%, a UA of 100.0%, and an F1-Score of 94.7%. The observed reduction in overall performance metrics from LD1 to LD2 is consistent with trends documented by Rizzo et al. (2020) and Manteghi et al. (2024), who noted that predictive performance naturally decays as the spatial and taxonomic variance of the target MUs increases.

Table 11 - Error matrix and accuracy parameters between the predicted map ELS2 (MUs) and external validation. Field validation using 380 observation points outside the model training areas at the LD2 level.

		Validation points											
	MUs	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXVe2**	**Eso**	**LAd**	**PAd**	**GZn1**	**∑ rows**	**UA (%)**	**FI-Score**
Digital soil map	RQo1	6	0	0	0	0	0	0	0	0	6	100.0	66.7
	RQo5	1	11	0	2	12	3	42	23	9	103	10.7	12.9
	RQg5	0	0	9	0	0	0	0	0	0	9	100.0	94.7
	RQg1	0	0	0	2	3	0	1	0	0	6	33.3	19.0
	GXVe2	1	2	0	6	12	3	10	13	0	47	25.5	26.7
	ESo	1	12	0	1	2	4	2	2	3	27	14.8	19.5
	LAd	1	22	0	4	12	4	52	28	0	123	42.3	45.2
	PAd	0	0	0	0	0	0	0	0	0	0	-	-
	GZn1	2	21	1	0	2	0	0	0	33	59	55.9	63.5
	∑ columns	12	68	10	15	43	14	107	66	45	380		
	PA (%)	50.0	16.2	90.0	13.3	27.9	28.6	48.6	0.0	73.3			
												OA = 33.9%	k = 0.20

Source: The author (2026).

Similarly, the hydromorphic unit "GZn1" displayed highly robust metrics, with a PA of 73.3%, a UA of 55.9%, and an F1-Score of 63.5% (Table 11). This confirms that the environmental signature of hydromorphic lowlands remains highly distinct within an ensemble learning approach, reinforcing the findings of Sobrinho et al. (2026) regarding algorithm-driven updates of legacy soil data. Geomorphon logfics show that both "RQg5" and "GZn1" (composing 0.4% and 21% of the reference area, respectively) are restricted to valley contexts under a constant or seasonal water table influence.

In contrast, the mapping unit "RQo5"—a complex association of Neossolos Quartzarênicos Órticos and Argissolos Vermelho-Amarelos—yielded the lowest UA (10.7%). The 30 m pixel resolution failed to resolve the narrow transitional mid-slope zones (altitudes between 25 and 29 m) where the Quartzipsamments predominate, frequently misclassifying them into the stable plateau domain (altitudes approx 30 m) dominated by "LAd" (Yellow Argisols/Oxisols). Indeed, the confusion matrix confirms this limitation: out of 103 observations labeled as "RQo5", 42 points were incorrectly assigned to "LAd" (Table 11). This boundary confusion mirrors the challenges faced by Arruda et al. (2016), who reported lower neural network performance when trying to differentiate transitional Argisol-Latossol zones in tropical toposequences.

Ultimately, these digital patterns align with conventional soil mapping rules, where units are differentiated primarily based on landform positions and field-validated soil–landscape relationships (Bazaglia Filho et al., 2013). In tropical climates, relief remains the most critical covariate for predictive modeling (Carvalho Monteiro et al., 2023; Sobrinho et al., 2026), as it directly dictates pedogenetic processes like sediment addition and water logging in low-development hydromorphic zones.

3.4.4 Reliability of predicted maps

Map reliability within the reference area was assessed using map purity spatial metrics, where the ELS1, ELT1, and ELS2 models yielded the highest coverage of the "High" purity class (> 80%). In ELS1, units "E" and "L" achieved purity values of 80.9% and 89.8%, respectively. Under ELT1, purity values reached 86.5% for "R", 83.6% for "E", and peaked at 93.3% for the hydromorphic unit "G" (Tables 12 and 13).

Table 12 - Purity and area of mapping units based in LD1 level on reference area for the best models.

Models	Mapping Unit	Area (km²)	Area (%)	Purity (%)	Purity Class
ELS1	R	266.42	23.4	77.9	Moderate
ELS1	G	227.17	20.0	78.6	Moderate
ELS1	E	143.95	12.7	80.9	High
ELS1	L	484.87	42.6	89.8	High
ELS1	P	15.24	1.3	60.6	Moderate
ELT1	R	527.09	46.6	86.5	High
ELT1	G	213.43	18.9	93.5	High
ELT1	E	54.96	4.9	83.6	High
ELT1	L	336.20	29.7	88.4	High
ELT1	P	0.32	0.03	56.5	Low
RangerT1	R	691.85	60.8	70.9	Moderate
RangerT1	G	235.92	20.7	78.7	Moderate
RangerT1	E	35.75	3.1	53.2	Low
RangerT1	L	169.11	14.9	68.9	Moderate
RangerT1	P	5.03	0.4	20.5	Very low
RFT1	R	686.72	60.4	68.7	Moderate
RFT1	G	229.57	20.2	79.1	Moderate
RFT1	E	37.08	3.3	55.4	Low
RFT1	L	183.12	16.1	64.7	Moderate
RFT1	P	1.17	0.1	13.8	Very low
XGBT1	R	682.22	60.3	67.3	Moderate
XGBT1	G	215.84	19.1	76.1	Moderate
XGBT1	E	36.42	3.2	58.8	Low
XGBT1	L	197.58	17.5	53.5	Low
XGBT1	P	0.00	0.0	0.0	Not suitable

R, NEOSSOLOS QUARTZARÊNICOS (Arenosols); G, GLEISSOLOS (Gleysols); E, ESPODOSSOLOS (Podzols); L, LATOSSOLOS (Ferralsols); P, ARGISSOLOS (Acrisols).

Table 13 - Purity and area of mapping units based in LD2 level on reference area for the best models.

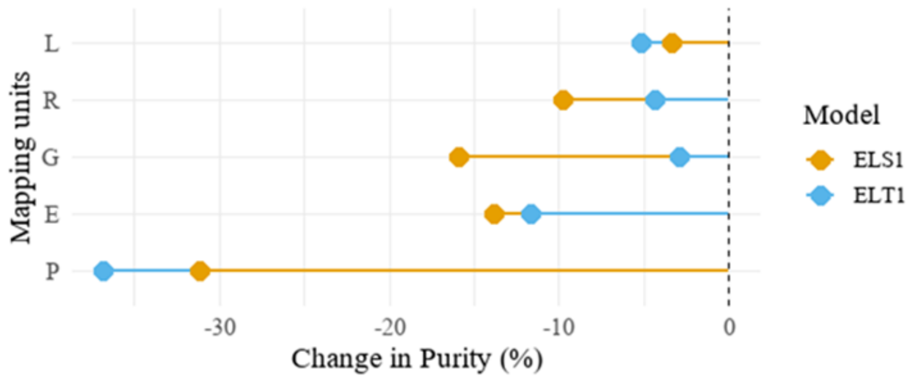
Algorithms	Mapping Unit	Area (km²)	Area (%)	Purity (%)	Purity Class
ELS2	RQo1	3.25	0.3	55.6	Low
ELS2	RQo5	278.42	24.5	77.9	Moderate
ELS2	RQg5	1.90	0.2	75.4	Moderate
ELS2	RQg1	12.31	1.1	72.4	Moderate
ELS2	GXve2	218.39	19.2	76.0	Moderate
ELS2	ESo	86.11	7.6	82.3	High
ELS2	LAd	382.78	33.6	87.7	High
ELS2	PAd	0.20	0.02	61.6	Moderate
ELS2	GZn1	154.30	13.6	93.0	High
RangerT2	RQo1	55.49	4.9	28.6	Very low
RangerT2	RQo5	342.21	30.1	56.7	Low
RangerT2	RQg5	0.49	0.04	33.7	Very low
RangerT2	RQg1	21.64	1.9	39.6	Very low
RangerT2	GXve2	55.35	4.9	26.5	Very low
RangerT2	ESo	81.97	7.2	53.7	Low
RangerT2	LAd	416.31	36.6	60.7	Moderate
RangerT2	PAd	8.38	0.7	26.8	Very low
RangerT2	GZn1	155.80	13.7	82.8	High

Algorithms	Mapping Unit	Area (km²)	Area (%)	Purity (%)	Purity Class
RFT2	RQo1	16.44	1.4	30.8	Low
RFT2	RQo5	431.35	37.9	56.6	Low
RFT2	RQg5	0.46	0.04	48.4	Low
RFT2	RQg1	8.42	0.7	38.0	Very low
RFT2	GXve2	61.80	5.4	31.0	Very low
RFT2	ESo	74.41	6.5	57.2	Low
RFT2	LAd	396.52	34.9	58.0	Low
RFT2	PAd	4.31	0.4	26.9	Very low
RFT2	GZn1	143.94	12.7	83.9	High
XGBT2	RQo1	0.03	0.0	1.8	Very low
XGBT2	RQo5	597.94	52.8	61.1	Moderate
XGBT2	RQg5	0.76	0.1	23.1	Very low
XGBT2	RQg1	5.47	0.5	23.1	Very low
XGBT2	GXve2	109.60	9.7	28.2	Very low
XGBT2	ESo	54.72	4.8	56.2	Low
XGBT2	LAd	238.92	21.1	49.9	Low
XGBT2	PAd	0.00	0.0	0.0	Not suitable
XGBT2	GZn1	124.57	11.0	82.9	High

RQo1, NEOSSOLOS QUARTZARÊNICOS Órticos + ARGISSOLOS VERMELHO-AMARELOS Distróficos (Arenosols + Acrisols); RQo5, NEOSSOLOS QUARTZARÊNICOS Órticos + NEOSSOLOS QUARTZARÊNICOS Hidromórficos + GLEISSOLOS HÁPLICOS Tb Distróficos (Arenosols + Gleysols); RQg5, NEOSSOLOS QUARTZARÊNICOS Hidromórficos (Fluvisols); RQg1 NEOSSOLOS QUARTZARÊNICOS Hidromórficos (Fluvisols); GXve2, GLEISSOLOS HÁPLICOS Ta Eutróficos (Gleysols); GZn1, GLEISSOLOS SÁLICOS Sódicos (Solonchaks); ESo, ESPODOSSOLOS FERRILÚVICOS Órticos (Podzols); LAd, LATOSSOLOS AMARELOS Distróficos (Ferralsols); PAd, ARGISSOLOS AMARELOS Distróficos (Acrisols).

High spatial reliability was also maintained in ELS2, where "GZn1" (93.0%), "LAd" (87.7%), and "Eso" (82.3%) were classified within the "High" purity tier. As established by Kempen et al. (2009) and Brus et al. (2011), map purity quantifies the probability that the predicted dominant unit matches actual field configurations. Our > 80% purity threshold significantly exceeds the overall purity of 67% reported by Kempen et al. (2009) for the Drenthe map update in the Netherlands. This validates our strategy: selecting refined digital data and regionalized covariates drastically enhances map certainty, particularly for hydromorphic ("G") and thin ("R") soils (Figs. 19, 21a, 21b).

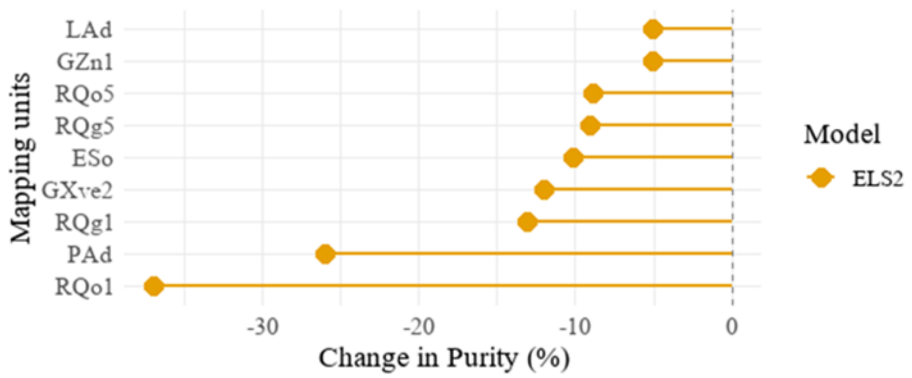
Figure 19 - Change in the purity (%) of mapping units at the LD1 level for the predictive models ELS1 and ELT1 between the reference area (Tracuateua) and the extrapolation area.



Source: The author (2026).

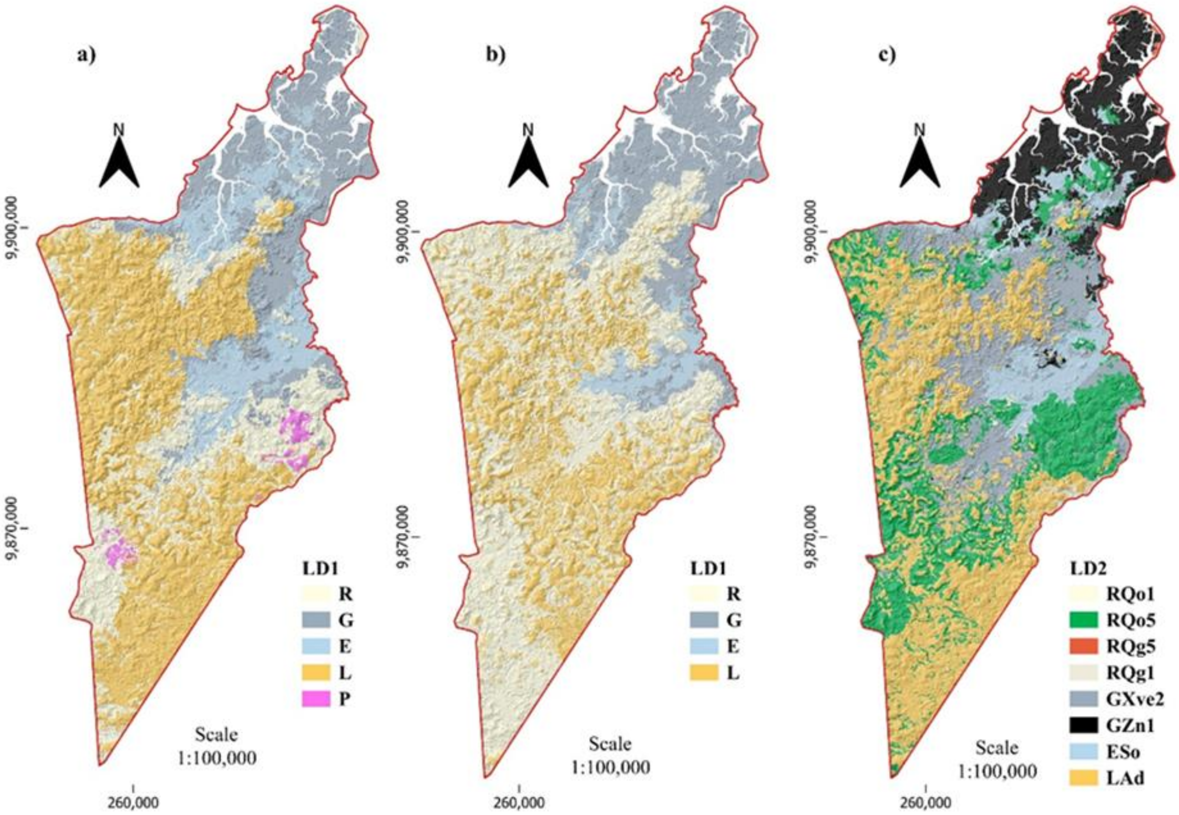
Furthermore, "LAd" and "GZn1" exhibited minimal spatial variation in purity when transitioned into the ELS2 extrapolation domain (Figs. 20, 21c). This high stability suggests that extrapolated zones mapped under these units can reliably support field validation schemes for regional soil surveys, a link between low purity and taxonomic misclassification previously warned about by Bargaoui et al. (2019). Conversely, at the LD2 level, unit "RQo1" exhibited the highest spatial impurity due to its highly specific and fragmented landscape position. The 30 m grid base was insufficient to capture these narrow slope segments where topography exerts an acute micro pedological control.

Figure 20 - Change in the purity (%) of mapping units at the LD2 level for the predictive model ELS2 between the reference area (Tracuateua) and the extrapolation area.



Source: The author (2026).

Figure 21 - Best digital soil maps of the extrapolation area generated by the models: (a) ELS1 and (b) ELT1 at the LD1 level, and (c) ELS2 at the LD2 level. Colors represent soil mapping units as indicated in the legend.



Source: The author (2026).

Nevertheless, these results demonstrate the immense potential of ensemble stacking—as advocated by Wadoux, Minasny, and McBratney (2020)—to overcome classical survey challenges in the Amazon biome. From a resource-allocation perspective, our framework proves that the ELT1 model exhibits a powerful predictive capability for broad hydromorphic soils ("G"), whereas ELS2 is optimized for detailed subgroups ("GZn1"). This operational distinction is vital; as Bregt (1992) noted, producing spatial data is insufficient if it lacks localized reliability, since some units may replicate across extrapolation borders but fail in structural quality.

Crucially, mapping these hydromorphic soil profiles under severe accessibility constraints holds significant value for quantifying critical ecosystem services, such as wetland carbon sequestration and conservation monitoring (Goldman et al., 2020). Our results show that while DTM attributes are optimal for mapping lowlands ("G") and ridges ("R"), DSM layers excels at capturing subtle canopy-topography variations to delineate highly weathered plateau units ("LAd" and "GZn1") at the LD2 level, matching the findings of Bohn and Miller (2025).

This counters the limitations reported by Silva et al. (2016b), who noted that standard DTMs often fail to capture subsurface variations (such as the Bt textural horizon of Ultisols) on identical topographic surfaces, an obstacle our hybrid DSM-Ensemble approach successfully circumvented.

3.4.5 Predictive digital soil map (PDSM)

The final predictive soil maps were generated using the ELS1, ELT1, and ELS2 configurations (Figs. 21a, b, c). Post-processing involved a spatial smoothing procedure to remove isolated noise pixels falling below the Minimum Mapping Area (MMA), enhancing spatial coherence (Poppiel et al., 2019). The resulting PDSMs represent taxonomic units classified at the LD1 level (Order) for ELS1/ELT1 and the LD2 level (Great Group) for ELS2, in strict accordance with the Brazilian Soil Classification System (Santos et al., 2025).

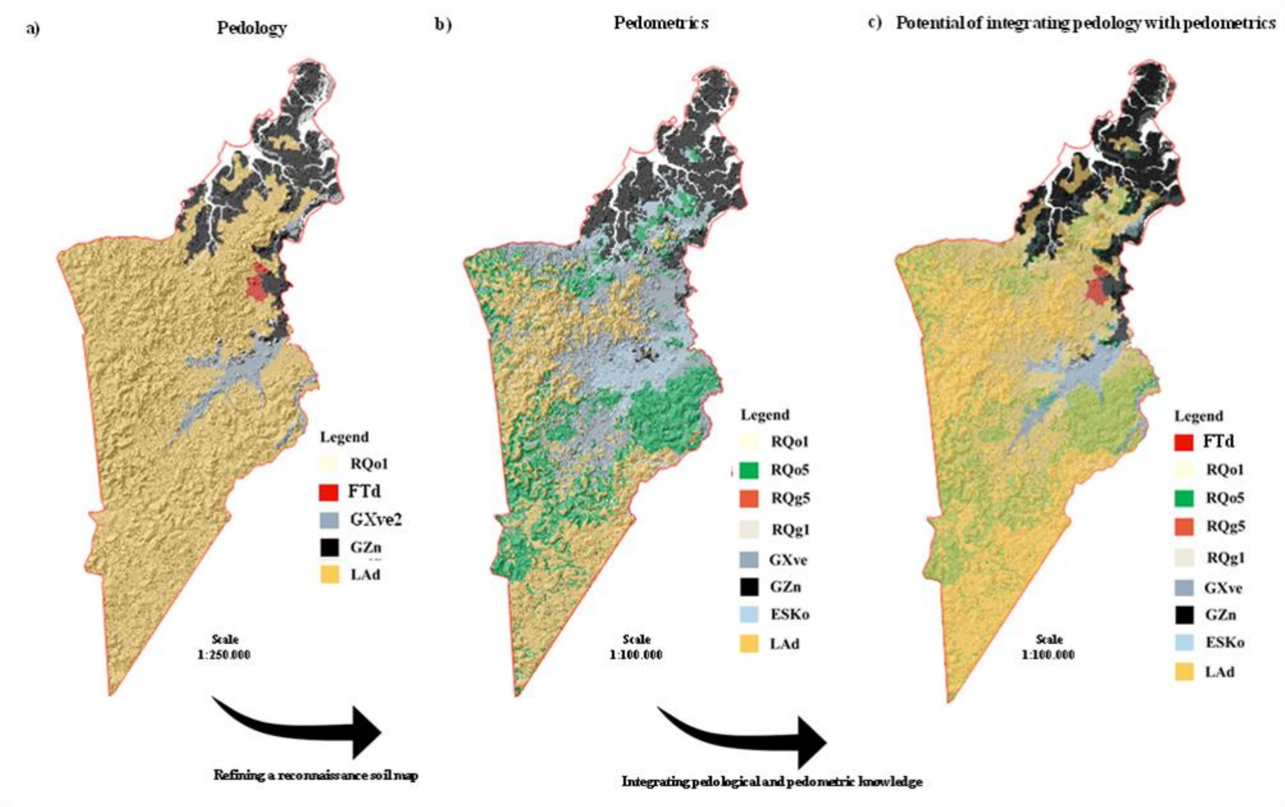
Field validation across the extrapolation window (Fig. 22) confirmed the soil–landscape patterns identified during the reference area training phase. Morphological descriptions of the independent validation profiles showed strong spatial agreement with the top-performing ensemble models, accurately predicting the transitions between well-drained upland pedons and poorly drained lowlands. This evolution from polygon-bound legacy maps to pixel-based continuous predictions represents a major spatial refinement for soil suborder distributions (Figs. 23a, b, c).

Figure 22 - Field validation in the extrapolation area, showing the spatial distribution of sampling points used to evaluate the predicted soil mapping units. Representative soil profiles and auger observations are presented for the main soil classes (Latossolos, Argissolos, Neossolos, Gleissolos, Plintossolos, and Espodossolos), supporting the interpretation of soil–landscape relationships.

Source: The author (2026).

To illustrate this integration, an overlay was executed between the conventional 1:250,000 legacy map (Fig. 10a) and the ELS2 PDSM (Fig. 10b). The resulting synthesis (Fig. 10c) clearly proves that utilizing local pedogenetic knowledge to structure environmental covariates allows the models to capture authentic soil variability far more effectively than evaluating statistical accuracy metrics alone (Angelini et al., 2020).

Figure 23 - Comparison between the legacy reconnaissance soil map (a) and the refined digital soil map generated using ensemble learning (b). The transition from polygon-based mapping to pixel-based predictive modeling highlights the improved spatial resolution and refinement of soil suborder distribution. The digital soil map (b), derived from DSM at 30 m spatial resolution, contrasts with the conventional map (a) at a scale of 1:250,000. Panel (c) illustrates the potential of integrating pedological knowledge with pedometric approaches through a hybrid framework, combining conventional high-intensity soil survey with Digital Soil Mapping (DSM). The resulting hybrid soil map is presented at a scale of 1:100,000, enhancing spatial inference and reducing mapping uncertainty compared to conventional survey methods.



Source: The author (2026).

Significantly, this hybrid integration successfully resolves a notorious limitation of standard digital soil mapping: the inability to predict taxonomic units absent from the input reference training dataset. Through this map-overlay framework, the unit Plintossolos (FTd)—which was completely absent from the reference zone—was mapped across the extrapolation landscape with high pedological consistency. Independent field validation confirmed actual patches of FTd clustered directly near the predicted boundary domains of this legacy unit (Figs. 22, 23c).

This outcome validates our core hypothesis. By representing the soil "continuum" while simultaneously mapping prediction errors (Terribile et al., 2011), the PDSM framework successfully bypasses the historical constraints of conventional soil surveys. Ultimately, given the strong performance of the ELT1 and ELS2 models in characterizing hydromorphic environments within geographically inaccessible regions of the Amazon biome, this approach effectively overcomes the historical defects of training models with outdated legacy data (Wadoux and McBratney, 2021), proving highly viable for regional mapping across data-scarce tropical biomes.

3.5 Challenges and limitations

In addition to the data gap hindering advances in digital soil mapping in the Amazon, available legacy maps present intrinsic limitations, such as the lack of more detailed conventional soil surveys. The reliance on historical surveys introduces spatial uncertainties that are propagated during the model training phase, making it difficult to distinguish mapping units.

Inherent characteristics of the Amazon biome pose challenges to necessary field validation, lack of topographic base detailed, including the difficulty of access, which may preclude soil sampling for characterization purposes, and severe financial constraints given the extensive areas that must be traversed and mapped.

A critical factor identified is the shortage of human resources. Digital soil mapping depends on the formalization of knowledge from experienced soil scientists; however, there has been a drastic reduction in the number of active and available professionals to contribute to the structuring of rule-based systems (tacit knowledge).

A limitation of digital soil mapping is the inability to predict MUs not present in the reference area. Considering pedogenetic variability, soil–landscape relationships, and the occurrence of hydromorphic environments—characteristic features of the Amazon biome—this limitation requires further investigation.

3.6 Conclusion

In conclusion, the results of this thesis confirm the hypothesis that knowledge about soil is transferable in the Amazonian environment through ensemble learning algorithms and the use of regional environmental covariates to capture the complex soil-landscape relationships typical of hydromorphic environments. By integrating conceptual frameworks based on expert knowledge with robust predictive modeling, the proposed approach successfully overcame the limitations of conventional soil maps and data scarcity.

ELT and ELS models, particularly at LD1 and LD2 levels, achieved the highest predictive performance, indicating that ensemble-based approaches combined with carefully selected covariates effectively capture soil–landscape relationships for spatial extrapolation, produced the best predictive maps for the extrapolation area for the MUs “G” and “GZn1”. The use of input data as appropriate covariates related to the most important soil-forming factors and pedogenetic processes in the extrapolation region was the key to achieving high-quality results.

Considering the scarcity of detailed soil surveys in the Amazon biome, our study updated the soil map of the extrapolation area from 1:250,000 to 1:100,000, using a reference area. The reference-area framework combined with ensemble learning proved effective for digital soil mapping in data-scarce regions of the Eastern Amazon, but model selection should integrate accuracy metrics, internal and external validation, and spatial uncertainty metrics (e.g., map purity) to support robust decision-making.

However, for decision-making regarding the best digital map, it is necessary to evaluate not only performance metrics and the confusion matrix, including internal and external validation, but also uncertainty metrics such as map purity.

This study highlights the potential of a hybrid digital soil mapping framework that integrates expert pedological knowledge, legacy soil maps, and ensemble learning algorithms as an effective and scalable strategy for soil mapping improvement in data-limited Amazonian environments.

REFERENCES

ADENIYI, Odunayo David et al. Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. **Land**, v. 12, n. 2, 15 fev. 2023.

ALVARES, Clayton Alcarde et al. Köppen’s climate classification map for Brazil. **Meteorologische Zeitschrift**, p. 711–728, 1 dez. 2013.

ANGELINI, M. E. et al. Extrapolation of a structural equation model for digital soil mapping. **Geoderma**, v. 367, p. 114226, 15 maio 2020.

ARIF, Muhammad. Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review. **International Journal of Database Theory and Application**, v. 11, n. 1, p. 1–8, 31 mar. 2018.

ARROUAYS, Dominique et al. Impressions of digital soil maps: The good, the not so good, and making them ever better. **Geoderma Regional**, v. 20, p. e00255, 1 mar. 2020.

ARRUDA, Gustavo Pais De *et al*. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, v. 73, n. 3, p. 266–273, jun. 2016.

ARRUDA, Gustavo Pais De; *et al*. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. **Revista Brasileira de Ciência do Solo**, v. 37, n. 2, p. 327–338, abr. 2013.

BARGAOUI, Yosra et al. Validation of digital maps derived from spatial disaggregation of legacy soil maps. **Geoderma**, 2019.

BAZAGLIA FILHO, Osmar et al. Comparison between detailed digital and conventional soil maps of an area with complex geology. **Revista Brasileira de Ciência do Solo**, v. 37, p. 1136–1148, 2013.

BOCK, Michael; KÖTHE, Rüdiger. Predicting the depth of hydromorphic soil characteristics influenced by ground water. **SAGA-seconds** Out, v. 19, 1 jan. 2008.

BÓCOLI, Fernanda Almeida *et al*. On the importance of digital soil mapping scale for ecosystem service assessment and policy - A study involving soil sediments dynamic and direct aquifer recharge. **Geoderma Regional**, v. 42, p. e00999, set. 2025.

BOHN, Meyer P.; MILLER, Bradley A. Digital soil mapping via machine learning of agronomic properties for the full soil profile at within-field resolution. **Agronomy Journal**, v. 117, n. 5, p. e70144, 2025.

BREGT, Arnold. **Processing of soil survey data**. Doctoral thesis—Wageningen, The Netherlands,: Wageningen Agricultural University, maio 1992.

BREIMAN, Leo. Random Forests. Machine Learning, v. 45, n. 1, p. 5–32, 1 out. 2001.
BRUS, D. J.; KEMPEN, B.; HEUVELINK, G. B. M. Sampling for validation of digital soil maps. **European Journal of Soil Science**, 2011.

CARVALHO MONTEIRO, Maria Eduarda et al. Digital mapping of soil classes in Southeast Brazil: environmental covariate selection, accuracy, and uncertainty. **Journal of South American Earth Sciences**, v. 132, p. 104640, 1 dez. 2023.

CHEN, Songchao et al. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. **Geoderma**, v. 409, p. 115567, mar. 2022.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: 13 ago. 2016. Disponível em: <http://arxiv.org/abs/1603.02754>. Acesso em: 14 mar. 2026

CHRISTOFOLETTI, Antonio. Geomorfologia. [S.l.]: Editora Blucher, 1980.

COHEN, Jacob. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, p. 37–46, 1960.

CONRAD, O. Digitale Reliefanalyse in der multimedialen Lehre. In: [S.l.: S.n.]. p. 37–47.

CONRAD, O. et al. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. **Geoscientific Model Development**, v. 8, n. 7, p. 1991–2007, 7 jul. 2015.

CREMON, Édipo H. et al. Geological and terrain attributes for predicting soil classes using pixel- and geographic object-based image analysis in the Brazilian Cerrado. **Geoderma**, v. 401, p. 115315, 1 nov. 2021.

DALMOLIN, R. S. D. et al. How is the learning process of digital soil mapping in a diverse group of land use planners? **Revista Brasileira de Ciência de Solo**, v. 44, n e0190037, 2020.

ECOSYSTEM, Copernicus Data Space. **Copernicus DEM - Global and European Digital Elevation Model**. Disponível em: <https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM>. Acesso em: 14 mar. 2026.

EFRON, Bradley; GONG, Gail. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. **The American Statistician**, v. 37, n. 1, p. 36–48, 1 fev. 1983.

ESFANDIARPOUR-BOROUJENI, I. et al. Assessment of different digital soil mapping methods for prediction of soil classes in the Shahrekord plain, Central Iran. **Catena**, v. 193, p. 104648, 1 out. 2020.

FARIAS, Fabrício do Carmo *et al.* Digital soil mapping in the Eastern Amazon: algorithm performance and classification uncertainty in hydromorphic and well-drained landscapes. **Soil Research**, v. 64, n. 4, p. SR25209, 11 maio 2026.

FAVROT, J. et al. Secteurs de références et connaissance du milieu physique : sols, climat, ressources en eau. v. 82, n. C.R. **Acad. Agric**., p. 37–54, 1996.

FERREIRA, Ana Carolina de S. et al. Use of Airborne Radar Images and Machine Learning Algorithms to Map Soil Clay, Silt, and Sand Contents in Remote Areas under the Amazon Rainforest. **Remote Sensing**, v. 14, n. 22, 2022.

FERREIRA, Ana Carolina S. et al. Predicting soil carbon stock in remote areas of the Central Amazon region using machine learning techniques. **Geoderma Regional**, v. 32, p. e00614, 1 mar. 2023.

FLORINSKY, I. V. **Digital Terrain Analysis in Soil and Geology**. [S.l.]: Elsevier, 2012.

GALLANT, John C.; DOWLING, Trevor I. A multiresolution index of valley bottom flatness for mapping depositional areas. **Water Resources Research**, v. 39, n. 12, 2003.

GANAIE, M. A. et al. Ensemble deep learning: A review. **Engineering Applications of Artificial Intelligence**, v. 115, p. 105151, out. 2022.

GELSLEICHTER, Yuri Andrei et al. Enhancing Soil Mapping with Hyperspectral Subsurface Images generated from soil lab Vis-SWIR spectra tested in southern Brazil. **Geoderma Regional**, v. 33, p. e00641, 1 jun. 2023.

GESSLER, P. E. et al. Soil-landscape modelling and spatial prediction of soil attributes. **International journal of geographical information systems**, v. 9, n. 4, p. 421–432, 1 jul. 1995.

GIASSON, Elvio et al. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, v. 63, p. 262–268, 2006.

GIASSON, E. et al. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. **Ciência Rural**, v.45, n.9, p.1592-1598, 2015.

GOLDMAN, Margaret A. et al. Digital soil mapping in a low-relief landscape to support wetland restoration decisions. **Geoderma**, v. 373, p. 114420, 15 ago. 2020.

GORELICK, Noel et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment, Big Remotely Sensed Data: tools, applications and experiences**. v. 202, p. 18–27, 1 dez. 2017.

GRINAND, Clovis et al. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, v. 143, n. 1, p. 180–190, 15 jan. 2008.

HARTEMINK, Alfred E.; MCBRATNEY, Alex. A soil science renaissance. **Geoderma**, v. 148, n. 2, p. 123–129, 15 dez. 2008.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, v. 25, n. 3, p. 295–309, 1 ago. 1988.

JASIEWICZ, Jarosław; STEPINSKI, Tomasz F. Geomorphons — a pattern recognition approach to classification and mapping of landforms. **Geomorphology**, v. 182, p. 147–156, 15 jan. 2013.

JENNY, Hans. **Factors of soil formation: a system of quantitative pedology**. First ed. New York: Soil Science, 1941.

KÄMPF, N.; CURI, Nilton. Formação e Evolução Do Solo (Pedogênese). In: **Pedologia: Fundamentos**. [S.l.]: SBCS, 2012. p. 343.

KEMPEN, Bas et al. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. **Geoderma**, 2009.

KHOSRAVANI, Pegah et al. Assessing the Role of Environmental Covariates and Pixel Size in Soil Property Prediction: A Comparative Study of Various Areas in Southwest Iran. **Land**, v. 13, n. 8, 2024.

KUHN, Max. **caret: Classification and Regression Training**. 5 out. 2007. Disponível em: <https://CRAN.R-project.org/package=caret>. Acesso em: 14 mar. 2026

KUHN, Max; JOHNSON, Kjell. Applied Predictive Modeling. New York, NY: Springer New York, 2013.

LAGACHERIE, P. et al. Mapping of reference area representativity using a mathematical soilscape distance. **Geoderma**, v. 101, n. 3, p. 105–118, 1 abr. 2001.

LAIPELT, Leonardo et al. ANADEM: A Digital Terrain Model for South America. **Remote Sensing**, v. 16, n. 13, 24 jun. 2024.

LAMICHHANE, Sushil; KUMAR, Lalit; ADHIKARI, Kabindra. Updating the national soil map of Nepal through digital soil mapping. **Geoderma**, 2021.

Landsat 8 | U.S. Geological Survey. Disponível em: <https://www.usgs.gov/landsat-missions/landsat-8>. Acesso em: 14 mar. 2026.

LARK, R. M.; CHAGUMAIRA, C.; MILNE, A. E. Decisions, uncertainty and spatial information. **Spatial Statistics**, Special Issue: The Impact of Spatial Statistics. v. 50, p. 100619, 1 ago. 2022.

MACHADO, Diego Fernandes Terra et al. Soil type spatial prediction from Random Forest: different training datasets, transferability, accuracy and uncertainty assessment. **Scientia Agricola**, v. 76, p. 243–254, 2019.

MACLIN, R.; OPITZ, D. Popular Ensemble Methods: An Empirical Study. **Journal of Artificial Intelligence Research**, v. 11, p. 169–198, 1 ago. 1999.

MANTEGHI, Shaho et al. Digital soil mapping for soil types using machine learning approaches at the landscape scale in the arid regions of Iran. **Advances in Space Research**, v. 74, n. 1, p. 1–16, 1 jul. 2024.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1, p. 3–52, 1 nov. 2003.

MEDEIROS, Beatriz Macêdo et al. Soil legacy data: An opportunity for digital soil mapping. **Revista Brasileira de Ciencia do Solo**, v. 48, 2024.

MELLO, Danilo César de et al. Mapping soil drainage classes: Comparing expert knowledge and machine learning strategies. **Soil Advances**, v. 3, p. 100028, 1 jun. 2025.

MELLO, Fellipe et al. Expert-based maps and highly detailed surface drainage models to support digital soil mapping. **Geoderma**, v. 384, p. 114779, 15 fev. 2021.

MENDONÇA SANTOS, M. L.; SANTOS, H. G. dos. Chapter 3 The State of the Art of Brazilian Soil Mapping and Prospects for Digital Soil Mapping. In: **Developments in Soil Science**. [S.l.]: Elsevier, 2006. v. 31 p. 39–54.

MINASNY, Budiman; HARTEMINK, Alfred E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v. 106, n. 1, p. 52–62, 1 maio 2011.

MOQUEDACE, Cássio Marques et al. High-resolution mapping of soil carbon stocks in the western Amazon. **Geoderma Regional**, v. 36, p. e00773, 1 mar. 2024.

NENKAM, Andree M. et al. Using homosoils to enrich sparse soil data infrastructure: An example from Mali. **Catena**, v. 223, p. 106862, 1 abr. 2023.

OLIVEIRA JUNIOR, R. C. de et al. Zoneamento agroecológico do município de Tracuateua, Estado do Pará. 1999.

PAHLAVAN-RAD, Mohammad Reza et al. Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. **Catena**, v. 194, p. 104715, nov. 2020.

PAVÃO, Quésia Sá et al. Texture prediction of natural soils in the Brazilian Amazon through proximal sensors. **Geoderma Regional**, v. 37, p. e00813, 1 jun. 2024.

POLIDORO, J. C. et al. Programa Nacional de Solos do Brasil (PronaSolos). 2016.

POPPIEL, Raúl R. et al. Pedology and soil class mapping from proximal and remote sensed data. **Geoderma**, v. 348, p. 189–206, 15 ago. 2019.

QUINLAN, J. Ross. **C4.5: Programs for Machine Learning**. [S.l.]: Elsevier, 2014.

R Core Team. Disponível em: <https://www.R-project.org/>. Acesso em: 14 mar. 2026a.

R CORE TEAM. A Language and Environment for Statistical Computing. Vienna, AustriaR Foundation for Statistical Computing, , 2024b. Disponível em: https://www.R-project.org/

RESENDE, Mauro et al. **Pedologia â Base para Distinção de Ambientes**. 6. ed. [S.l.]: UFLA, 2014.

RIZZO, Rodnei et al. Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in southeastern Brazil. **Geoderma**, 2020.

RODRIGUES, Hugo et al. autoRA: An Algorithm to Automatically Delineate Reference Areas—A Case Study to Map Soil Classes in Bahia, Brazil. **Land**, v. 14, n. 3, 2025a.

RODRIGUES, Hugo et al. autoRA: An Algorithm to Automatically Delineate Reference Areas—A Case Study to Map Soil Classes in Bahia, Brazil. **Land**, v. 14, n. 3, 12 mar. 2025b.

ROSSITER, DG. **Methodology for soil resource inventories**. 2nd ed. Enschede: International Institute for Aerospace Survey & Earth Sciences; 2000.

ROSSITER, David G.; POGGIO, Laura. Representing soil landscapes from digital soil mapping products – helping the map to speak for itself. **Soil**, v. 11, n. 2, p. 849–881, 21 out. 2025.

SANTOS, H. G. dos et al. **Procedimentos normativos de levantamentos pedológicos**. [S.l.]: Brasília, DF: EMBRAPA-SPI; Rio de Janeiro: EMBRAPA-CNPS, 1995., 1995.
Instituto Brasileiro de Geografia e Estatística (IBGE). **Manual técnico de pedologia**. 3 ed. Rio de Janeiro, 2015.

SANTOS, H. G. dos et al. **Sistema Brasileiro de Classificação de Solos**. [S.l.]: Brasília, DF: Embrapa, 2025., 2025.

SILVA, Cristiano Cassiano da et al. Mapeamento pedológico digital da folha Botucatu (SF-22-Z-B-VI-3): treinamento de dados em mapa tradicional e validação de campo. **Revista Brasileira de Ciência do Solo**, v. 37, p. 846–857, 2013.

SILVA, Sérgio Henrique Godinho et al. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. **Geoderma**, v. 267, p. 65–77, 1 abr. 2016a.

SILVA, Sérgio Henrique Godinho et al. Proximal Sensing and Digital Terrain Models Applied to Digital Soil Mapping and Modeling of Brazilian Latosols (Oxisols). **Remote Sensing**, v. 8, n. 8, 24 jul. 2016b.

SINGER, Robert B. Near-infrared spectral reflectance of mineral mixtures: Systematic combinations of pyroxenes, olivine, and iron oxides. **Journal of Geophysical Research: Solid Earth**, v. 86, n. B9, p. 7967–7982, 1981.

SOBRINHO, Rômulo José Alencar et al. Integrating Tacit Knowledge and AI for Digital Soil Mapping in Eastern Amazonia: Ensemble Learning, Model Performance, and Uncertainty Incorporation. **Soil Systems**, v. 10, n. 3, 16 mar. 2026.

SULIEMAN, Magboul M. et al. Application of Machine Learning Algorithms for Digital Mapping of Soil Salinity Levels and Assessing Their Spatial Transferability in Arid Regions. **Land**, v. 12, n. 9, 2023.

TAGHIZADEH-MEHRJARDI, Ruhollah et al. Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. **Geoderma**, v. 399, p. 115108, 1 out. 2021.

TAGHIZADEH-MEHRJARDI, Ruhollah et al. Semi-supervised learning for the spatial extrapolation of soil information. **Geoderma**, v. 426, 2022.

TEN CATEN, Alexandre et al. Regressões Logísticas Múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v. 35, n. 1, p. 53–62, fev. 2011.
TERRIBILE, F. et al. Potential and limitations of using soil mapping information to understand landscape hydrology. **Hydrology and Earth System Sciences**, v. 15, n. 12, p. 3895–3933, 22 dez. 2011.

VILLELA, André Luis Oliveira. **Mapeamento digital de solos da formação solimões sob floresta tropical amazônica**. Rio de Janeiro: Universidade Federal Rural do Rio de Janeiro, ago. 2023.

WADOUX, Alexandre M. J. C.; MCBRATNEY, Alex B. Hypotheses, machine learning and soil mapping. **Geoderma**, v. 383, p. 114725, 1 fev. 2021.

WADOUX, Alexandre M. J. C.; MINASNY, Budiman; MCBRATNEY, Alex B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews**, 2020.

WOLSKI, Mario Sergio et al. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 633–642, ago. 2017.

WRIGHT, Marvin N. **ranger: A Fast Implementation of Random Forests**. , 28 jul. 2015. Disponível em: <https://CRAN.R-project.org/package=ranger>. Acesso em: 14 mar. 2026

YE, Mingliang et al. Soil class mapping with severely limited sample data: an explicit rule of soil-landscape relationships in complex terrain. **Pedosphere**, 13 fev. 2026.

ZERAATPISHEH, Mojtaba et al. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. **Geomorphology**, 2017.

4 GENERAL CONCLUSION

Soil mapping is fundamental to achieving the Sustainable Development Goals (SDGs), especially SDG 15 (Life on Land), focusing on sustainable management, combating illegal logging, and ensuring food security. Conventional soil survey methods are well-established but present major drawbacks within the Amazon biome, such as high operational costs, restricted field access, and a shortage of qualified human resources capable of formalizing pedological expertise. The findings of this thesis demonstrated that a hybrid Digital Soil Mapping approach—integrating experts' tacit knowledge of soil–landscape relationships, legacy maps, and ensemble learning algorithms—constitutes an efficient and scalable strategy to overcome data scarcity in complex environments of the Eastern Amazon. The results of this study epresent a significant scientific advancement for digital soil mapping in the Amazon, establishing a methodology capable of supporting the sustainable management and conservation of soils in hydromorphic environments, which represent spatially extensive ecosystems and vital providers of multiple ecosystem services.

Acknowledgments

We thank the members of Geotechnologies and Pedometrics Research Group – GEOP (https://geopufra.com /equipe/) for the support of the team. This study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Funding

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Chamada CNPq/MCTI/FNDCT N° 18/2021 - Universal 2021 – Process: 406838/2021-6

.

APPENDIX

Table A1 - Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RFS1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	188,925	41,257	5,626	72,208	22,277	330,293	57
G	19,300	194,518	3019	0	662	217,499	89
E	9585	12,833	41,022	0	0	63,440	65
L	33,780	0	0	149,874	5187	188,841	79
P	21,924	227	0	10,971	84,462	117,584	72
Total	273,514	248,835	49,667	233,053	112,588	917,657	
PA %	69	79	82	64	75		
	OA = 0.72	Kappa = 0.63					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A2. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RFT1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	193,479	33,912	4,569	95,203	16,450	343,613	56
G	24,428	205,370	6,183	0	1,844	237,825	86
E	9,147	10,645	38,915	0	0	58,707	66
L	23,589	0	0	127,329	5,272	156,190	82
P	22,844	281	0	10,510	89,017	122,652	73
Total	273,487	250,208	49,667	233,042	112,583	918,987	
PA %	71	83	78	55	79		
	OA = 0.71	Kappa = 0.62					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A3. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RangerS1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	191,271	38,602	5,786	72,568	24,527	332,754	57
G	20,793	196,551	2,830	0	736	220,910	89
E	9,767	13,566	41,051	0	0	64,384	64
L	37,883	0	0	154,306	10,661	202,850	76
P	13,800	116	0	6179	76,664	96,759	79
Total	273,514	248,835	49,667	233,053	112,588	917,657	
PA %	70	80	82	66	68		
	OA = 0.72	Kappa = 0.63					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A4. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by RangerT1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	198,735	31,733	4,530	95,211	18,447	348,656	57
G	25,005	208,141	6,067	0	2,037	241,250	86
E	8,444	10,270	39,070	0	0	57,784	68
L	24,610	0	0	129,892	6,670	161,172	81
P	16,693	64	0	7,939	85,429	110,125	78
Total	273,487	250,208	49,667	233,042	112,583	918,987	
PA %	72	84	78	56	76		
	OA = 0.72	Kappa = 0.63					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A5. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the XgbS1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	183,473	42,970	5049	65,142	21,162	317,796	58
G	23,687	193,069	4,355	222	1,152	222,485	87
E	8,091	12,563	40,263	0	0	60,917	66
L	35,862	3	0	155,065	6,706	197,636	78
P	22,401	230	0	12,624	83,568	118,823	70
Total	273,514	248,835	49,667	233,053	112,588	917,657	
PA %	67	78	81	66	74		
	OA = 0.72	Kappa = 0.62					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A6. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the XgbT1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	191,145	35,753	4,883	77,416	21,197	330,394	58
G	24,807	203,672	6,193	0	2,436	237,108	86
E	8009	10,570	38,591	0	0	57,170	68
L	29,316	0	0	142,863	6069	178,248	80
P	20,210	213	0	12,763	82,881	116,067	71
Total	273,487	250,208	49,667	233,042	112,583	918,987	
PA %	70	82	77	61	74		
	OA = 0.72	Kappa = 0.63					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A7. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the C5S1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	170,828	46,142	5,554	74,504	13,460	310,488	55
G	32,521	186,297	2,601	1,507	7,008	229,934	81
E	11,649	16,126	41,512	0	226	69,513	60
L	38,834	115	0	145,647	11,675	196,271	74
P	19,682	155	0	11,395	80,219	111,451	72
Total	273,514	248,835	49,667	233,053	112,588	917,657	
PA %	62	75	83	62	71		
	OA = 0.68	Kappa = 0.58					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A8. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the C5T1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	195,243	31,002	4,360	99,585	15,934	346,124	56
G	27,938	209,649	7,074	61	2,788	247,510	85
E	6,927	9,507	38,233	0	60	54,727	70
L	24,723	0	0	121,181	8,676	154,580	78
P	18,656	50	0	12,215	85,125	116,046	73
Total	273,487	250,208	49,667	233,042	112,583	918,987	
PA %	71	84	77	52	76		
	OA = 0.71	Kappa = 0.61					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A9. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the ELS1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	136,555	31,631	2,981	32,173	14,456	217,796	63
G	28,668	196,765	2,921	637	3,771	232,762	85
E	18,405	17,064	43,374	0	0	78,843	55
L	58,218	408	1	189,792	29,048	277,467	68
P	19,533	382	0	5,083	61,382	86,380	71
Total	261,379	246,250	49,277	227,685	108,657	893,248	
PA %	52	76	91	86	37		
	OA = 0.69	Kappa = 0.59					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A10. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the ELT1 (MU) model and the legacy map (Reference), in LD1.

	Reference						
MU [a]	**R**	**G**	**E**	**L**	**P**	**Total**	**UA %**
R	151,403	20,900	4,604	57,041	15,889	249,837	61
G	41,468	215,041	8,094	102	2,554	267,259	80
E	8,163	12,920	36,930	0	0	58,013	64
L	47,344	93	1	165,135	12,029	224,602	74
P	23,187	392	0	10,116	81,449	115,144	71
Total	271,565	249,346	49,629	232,394	111,921	914,855	
PA %	55	87	74	71	73		
	OA = 0.71	Kappa = 0.62					

MU[a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A11. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RFS2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	54,138	3,556	0	194	4,512	3,477	122	0	24,575	90,574	60
RQo5	340	65,847	0	3,655	8,825	566	6,428	9,116	1,735	96,512	68
RQg5	0	0	1,132	0	0	0	0	0	1,254	2,386	47
RQg1	0	4,639	0	26,567	6	4	31,941	2,206	2	65,365	41
GXve2	1,881	20,538	0	1,129	30,767	632	71	3,340	3,945	62,303	49
ESo	8,655	2,985	0	0	13,197	43,817	0	0	3,266	71,920	61
LAd	0	17,427	0	31,391	6	0	183,408	6,050	0	238,282	77
PAd	0	6,307	0	22,069	309	0	11,083	91,888	0	131,656	70
GZn1	594	0	458	0	4,474	1,171	0	0	151,962	158,659	96
Total	65,608	121,299	1,590	85,005	62,096	49,667	233,053	112,600	186,739	917,657	
PA %	82	54	72	31	49	88	79	82	82		
	OA = 0.71	Kappa = 0.66									

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A12. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RFT2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	53,149	4,705	0	0	5,307	3,321	0	0	15,898	82,380	65
RQo5	30	66,391	0	2,094	6,583	620	16,728	3,814	175	96,435	69
RQg5	0	0	1,124	0	0	0	0	0	2,572	3,696	30
RQg1	0	3,395	0	32,852	16	0	36,161	2,279	0	74,703	44
GXve2	2,284	24,644	0	1,131	33,417	871	101	5,554	2,290	70,292	48
ESo	9,109	2,857	0	0	12,147	42,694	0	0	5,230	72,037	59
LAd	0	12,225	0	28,539	0	0	171,097	6,474	0	218,335	78
PAd	0	7,039	0	20,381	354	0	8,955	94,473	0	131,202	72
GZn1	1,038	0	489	0	4,278	2,161	0	0	161,941	169,907	95
Total	65,610	121,256	1,613	84,997	62,102	49,667	233,042	112,594	188,106	918,987	
PA %	81	55	70	39	54	86	73	84	87		
	OA = 0.72	Kappa = 0.67									

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A13. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RangerS2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	53,729	4,683	0	749	4,843	3,552	786	1	19,137	87,480	61
RQo5	54	66,020	0	3,822	7,695	603	11,993	9,103	77	99,367	66
RQg5	0	0	1,107	0	0	0	0	0	1,394	2,501	44
RQg1	0	2,344	0	28,351	5	0	33,080	2,180	0	65,960	43
GXve2	1,973	21,940	0	1,044	31,541	524	32	3,343	3,576	63,973	49
ESo	8,802	2,833	0	0	13,236	43,229	0	0	2,982	71,082	61
LAd	0	17,862	0	32,384	14	0	178,797	9,860	0	238,917	75
PAd	0	5,613	0	18,598	204	0	8,352	88,113	0	120,880	73
GZn1	1,050	4	483	57	4,558	1,759	13	0	159,573	167,497	95
Total	65,608	121,299	1,590	85,005	62,096	49,667	233,053	112,600	186,739	917,657	
PA %	82	54	71	33	51	87	77	78	86		
	OA = 0.71Kappa = 0.66										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A14. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the RangerT2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	53,615	5,819	0	0	5,664	3,541	0	0	14,057	82,696	65
RQo5	0	66,310	0	1,987	6,075	665	21,046	4,319	2	100,404	66
RQg5	0	0	1,131	0	0	0	0	0	2,497	3,628	31
RQg1	0	2,742	0	36,141	3	0	38,462	3,636	0	80,984	45
GXve2	2,161	27,066	0	1,049	33,881	782	97	6,083	1,809	72,928	46
ESo	7,508	2,811	0	0	11,580	42,173	0	0	4,060	68,132	62
LAd	0	11,067	0	29,119	0	0	167,273	10,527	0	217,986	77
PAd	0	5,440	0	16,701	268	0	6,164	88,029	0	116,602	75
GZn1	2,326	1	482	0	4,631	2,506	0	0	165,681	175,627	94
Total	65,610	121,256	1,613	84,997	62,102	49,667	233,042	112,594	188,106	918,987	
PA %	82	55	70	43	54	85	72	78	89		
	OA = 0.71Kappa = 0.66										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A15. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the XgbS2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	53,761	2,446	0	1,829	4,349	2,599	4,942	0	23,271	93,197	58
RQo5	599	66,024	0	2,461	8,977	573	13,897	8,119	566	101,216	65
RQg5	45	0	1,061	0	0	0	0	0	1,476	2,582	41
RQg1	0	1,897	0	28,443	102	0	30,407	4,527	0	65,376	44
GXve2	3,216	23,921	0	1,479	31,883	2,476	20	3,355	6,051	72,401	44
ESo	7,093	2,359	0	0	11,325	42,870	0	0	2,867	66,514	64
LAd	158	17,913	0	30,212	66	0	170,003	7,728	18	226,098	75
PAd	0	6,739	0	20,581	416	0	13,784	88,871	0	130,391	68
GZn1	736	0	529	0	4,978	1,149	0	0	152,490	159,882	95
Total	65,608	121,299	1,590	85,005	62,096	49,667	233,053	112,600	186,739	917,657	
PA %	82	54	68	33	51	86	73	79	82		
	OA = 0.69Kappa = 0.64										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A16. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the XgbT2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	53,359	2,086	0	91	4,572	3,195	0	0	19,110	82,413	65
RQo5	929	65,839	0	2,683	7,391	868	17,463	4,910	1,309	101,392	65
RQg5	47	0	1,136	0	0	0	0	0	3,807	4,990	23
RQg1	0	2,931	0	36,139	61	0	37,090	6726	0	82,947	44
GXve2	3,002	28,726	0	821	35,478	2,183	18	6,038	5,232	81,498	44
ESo	7,350	2,900	0	0	9,738	41,553	0	0	4,041	65,582	63
LAd	62	12,237	0	26,556	64	1	166,412	7,024	33	212,389	78
PAd	0	6,537	0	18,707	492	0	12,059	87,896	0	125,691	70
GZn1	861	0	477	0	4,306	1,867	0	0	154,574	162,085	95
Total	65,610	121,256	1,613	84,997	62,102	49,667	233,042	112,594	188,106	918,987	
PA %	81	54	70	43	57	83	71	78	83		
	OA = 0.70Kappa = 0.65										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A17. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the C5S2 (MU) model and the legacy map (reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	51,696	2,738	0	9	4,875	3,992	42	0	11,932	75,284	69
RQo5	902	72,154	0	5,445	10,809	1,528	27,894	8,641	487	127,860	56
RQg5	0	0	1,135	0	0	0	0	0	1,702	2,837	40
RQg1	0	1,257	0	25,207	9	0	30,611	2,208	0	59,292	43
GXve2	1,536	18,622	0	1,426	29,497	545	62	5,285	3,198	60,171	49
ESo	7,453	3,744	0	0	12,321	42,640	0	0	3,189	69,347	61
LAd	0	18,048	0	32,769	4	0	162,657	9,284	0	222,762	73
PAd	0	4,736	0	20,148	122	0	11,779	87,143	0	123,928	70
GZn1	4,021	0	455	0	4,459	962	0	0	166,231	176,128	94
Total	65,608	121,299	1,590	85,004	62,096	49,667	233,045	112,561	186,739	917,609	
PA %	79	59	73	30	47	86	70	77	89		
	OA = 0.70Kappa = 0.64										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A18. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the C5T2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	49,802	2,809	0	0	4,610	3408	0	0	8,925	69,554	72
RQo5	622	72,464	0	3,700	10,022	1,688	27,088	3,923	108	119,615	61
RQg5	0	0	1,284	0	0	0	0	0	3,718	5,002	26
RQg1	116	2,110	0	34,514	68	0	46,856	2,851	1,444	87,959	39
GXve2	2,288	22,894	0	1,175	32,027	1,349	131	7915	3,209	70,988	45
ESo	6,982	3,794	0	0	11,533	41,273	0	0	4,047	67,629	61
LAd	12	13,031	0	26,214	0	0	148,477	12,780	0	200,514	74
PAd	0	4,154	0	19,394	101	0	10,490	85,125	2	119,266	71
GZn1	5,788	0	329	0	3,741	1,949	0	0	166,653	178,460	93
Total	65,610	121,256	1,613	84,997	62,102	49,667	233,042	112,594	188,106	918,987	
PA %	76	60	80	41	51	83	64	76	89		
	OA = 0.69Kappa = 0.63										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A19. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the ELS2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	51,727	4,349	0	451	4,749	2,036	1,434	0	17,950	82,696	63
RQo5	167	55,305	0	3,067	7,181	417	19,355	8142	343	93,977	59
RQg5	0	0	632	0	0	0	0	0	1,108	1,740	36
RQg1	0	2,760	0	29,751	126	0	26,572	8,280	0	67,489	44
GXve2	2,093	27,131	0	1,724	28,767	1,864	95	5,685	1,979	69,338	41
ESo	10,538	3,232	3	0	15,315	43,106	0	0	5,511	77,705	55
LAd	7	15,648	0	28,836	33	1	166,439	12,817	11	223,792	74
PAd	0	5,112	0	16,945	292	0	13,546	73,651	0	109,546	67
GZn1	896	113	595	87	5,030	1,799	228	0	157,840	166,588	95
Total	65,428	113,650	1,230	80,861	61,493	49,223	227,669	108,575	184,742	892,871	
PA %	79	49	53	37	47	87	73	68	86		
	OA = 0.68Kappa = 0.62										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Table A20. Pixel-by-pixel internal validation confusion matrix between the predicted map generated by the ELT2 (MU) model and the legacy map (Reference), in LD2.

	Reference										
MU [a]	**RQo1**	**RQo5**	**RQg5**	**RQg1**	**GXve2**	**ESo**	**LAd**	**PAd**	**GZn1**	**Total**	**UA %**
RQo1	51,998	3,515	0	0	4,438	2,792	0	0	16,186	78,929	66
RQo5	12	56,228	0	2,737	8,501	672	6,270	8,043	35	82,498	68
RQg5	59	17	1,004	0	0	0	0	0	3,768	4,848	21
RQg1	22	7,177	0	36,231	100	0	43,762	6,379	1	93,672	39
GXve2	6,689	31,827	0	510	34,703	6,174	56	2,867	5,389	88,215	39
ESo	5,524	1,753	0	0	9,625	36,921	0	0	2,675	56,498	65
LAd	3	14,609	0	30,065	4	0	175,185	19,718	1	239,585	73
PAd	0	5,024	0	14,864	200	0	7,121	74,914	0	102,123	73
GZn1	1,278	9	410	0	4,444	3070	0	0	159,276	168,487	95
Total	65,585	120,159	1,414	84,407	62,015	49,629	232,394	111,921	187,331	914,855	
PA %	79	47	72	43	56	74	75	67	86		
	OA = 0.69Kappa = 0.63										

MU [a]: Soil classification according to SiBCS (Santos et al, 2025). UA: User's accuracy; PA: Producer's accuracy; OA: overall accuracy.

Figure A1. Field prospection conducted to validate legacy soil mapping units to guide the placement of several representative pedons sampling points. avoiding transition zones indicated in Figure 3.

Source: Farias (2023).

Quadro A1. Covariates used in modeling for digital soil mapping in this study Tracuateua. Capanema. Primavera. Quatipuru. Pará. Brazil.

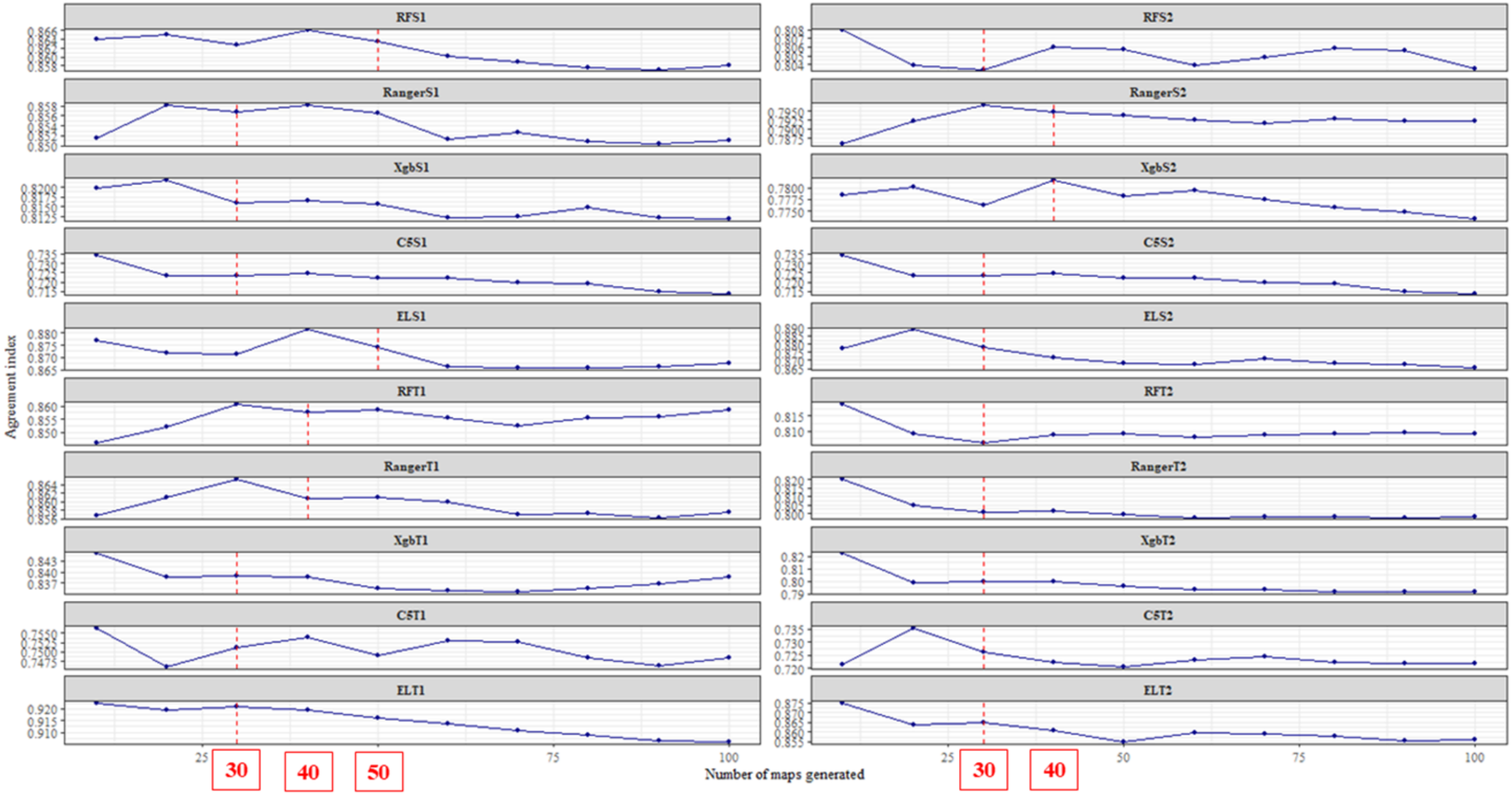
Covariates	Descriptions
Analytical Hillshading (AH)	Provides visualization of the alignments and textural density of the relief elements present in the area.
Aspect (AS)	Grid of slopes
Channel Network Base Level (CNBL)	Vertical distance to the base level of the channel network.
Closed Depressions (CD)	Enclosed area that has no surface drainage outlet and from which water escapes only by evaporation or subsurface drainage.
Convergence Index (CI)	Calculates the convergence/divergence index with respect to surface runoff.
Cross Sectional Curvature (CSC)	Curvature relative to the vertical plane of the level differences.
Flow Accumulation (FA)	Areas where the flow is concentrated and allows identification of the paths or flow of water.
General Curvature (GC)	General Curvature
Gradient (GD)	This corresponds to the hydraulic gradient
Local Curvature (LCU)	Local Curvature

Covariates	Descriptions
Longitudinal Curvature (LGC)	Rate of change of slope represents the deviation of the gradient along the flow
LS Factor (LSF)	This represents the calculation of the slope length factor
Mass Balance Index (MBI)	Mass Balance Index
Multiresolution Index of Ridge Top Flatness (MRRTF)	Valley Top Multiresolution Index
Multiresolution Index of Valley Bottom Flatness (MRVBF)	Valley Bottom Multiresolution Index
Plan Curvature (PC)	Plane curvature of the site
Profile Curvature (PFC)	Shapes of landscape features
Relative Slope Position (RSP)	It represents the slope position of the cell and its relative position between the valley and the summit.
Slope (SP)	The slope of the ground surface relative to the horizontal plane
Tangential Curvature (TC)	Describes the first accumulation mechanism
Terrain Ruggedness Index (TRI)	Difference in elevation values from a central cell and the eight neighboring cells
Texture (TX)	Terrain surface texture
Topographic Position Index (TPI)	Topographic index based on geomorphology
Topographic Wetness Index (TWI)	Relationship between local slope and upstream specific contributing area
Total Insolation (TI)	Potential for solar radiation input
Upslope Curvature (USC)	Local distance-weighted average curvature in the area of contribution of a cell based on the flow direction multiple
Valley Depth (VD)	Valley depth refers to the vertical distance to a base level of the channel network
Vector Ruggedness Measure (VRM)	Surface roughness
Vertical Distance to Channel Network (VDCN)	Vertical distance to the base level of the channel network
Geomorphons (GEOM)	This tool derives so called geomorphons. which represent categories of terrain forms. from a digital elevation model using a machine vision approach
Geology	Geological map of the study area
Average Annual Precipitation (AAP)	Contemporary average annual precipitation
Annual Precipitation 22.000 Years Ago (PR_22k)	Annual Precipitation 22.000 Years Ago
Temperature 22.000 Years Ago (Tmean_22k)	Temperature 22.000 Years Ago
Ground Surface Temperature during the day (LST)	Ground Surface Temperature during the day
Drainage Density (DD)	The drainage density correlates the total length of the drainage channels with the area of the watershed
Altitude (m)	Vertical distance measured between a point and mean sea level
Horizontal Overland Flow Distance (HOFD)	Calculates overland flow distances to a channel network based on gridded digital elevation data and channel network information
Band 2-Blue (B2)	Used to detect bodies of water
Band 3-Green (B3)	Detection and analysis of vegetation
Band 4-Red (B3)	Contrast between areas occupied by vegetation. exposed areas. roads and urban areas
Band 5-NIR (B4)	Characteristics of the Earth's surface

Covariates	Descriptions
Band 6 - SWIR1 (B6)	Related to the object's temperature
Band 7 - SWIR2 (B7)	Terrain morphology
Clay Minerals Ratio (CMR)	Highlights rocks hydrothermally altered containing clay and alunite
Enhanced Vegetation Index (EVI)	Vegetation signal with enhanced sensitivity in high biomass regions and vegetation monitoring
Ferrous Minerals Ratio (FMR)	Iron-bearing minerals. It uses the relationship between the SWIR band and the NIR band
Iron Oxide Ratio (IOR)	This causes areas with strong iron alteration to appear bright
Normalized Difference Vegetation Index (NDVI)	Spectral response of plants in the red and near-infrared bands
Simple Ratio (SR)	The ratio between the near-infrared (NIR) band and the red (RED) band. Indicates the presence of vegetation
Soil Adjusted Vegetation Index (SAVI)	A vegetation index that attempts to minimize the influences of soil brightness using a soil brightness correction factor

Source: The author (2026).

Figure A2. Stabilization curve between the number of maps generated in the processing of predictive models and the concordance index among the 100 maps generated.



Source: The author (2026).

Table A22. Soil profiles descriptions and physicochemical analysis.

Profile (n°)	FERRALSOL (06)
Classification (SiBCS[1])	LATOSSOLO AMARELO Distrófico
Date	02/08/2025
Coordinates UTM	256,623.904 x 9,872,230.132 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Barreiras
Relief [2]	very gently sloping
Elevation (m)	78.94 m
Erosion	Moderate
Permeability	well drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

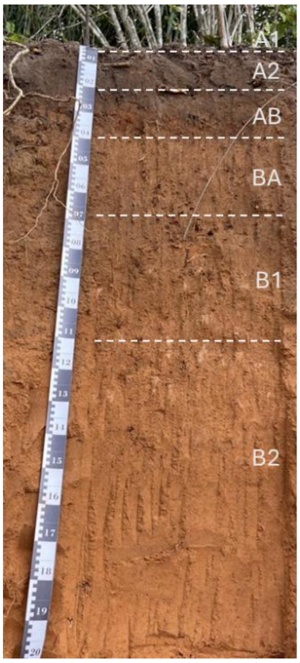


Table A22.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	141	221	638
A2	167	243	590
AB	189	159	652
BA	209	99	692
Bw1	236	148	616
Bw2	235	197	568

Table A22.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm^3---				
A1	4.0	1.0	8	5	4.3	44
A2	4.1	0.2	2	1	12.3	41
AB	4.2	0.6	3	1	5.4	23
BA	4.3	0.1	7	1	3.4	14
Bw1	4.4	0.3	7	1	1.9	17
Bw2	4.4	0.1	7	1	1.3	15

Table A22.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm^3---		------%------		g/kg	mg/dm^3
A1	14.1	58.1	24	23	33	5
A2	3.2	44.5	7	79	24	4
AB	4.6	28.0	17	54	15	4
BA	8.1	22.4	36	30	12	3
Bw1	8.3	25.2	33	19	12	3
Bw2	8.1	22.8	36	14	11	3

[1] SiBCS - Brazilian Soil Classification System

Table A23 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	ARENOSOL (07)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	02/08/2025
Coordinates UTM	2,59,491.191 x 9,892,312.585 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	gently sloping
Elevation (m)	24.79 m
Erosion	Laminar
Permeability	Excessively drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A23.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A	150	136	714
CA	122	114	764
C	140	110	750

Table A23.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A	4.3	0.4	2	1	9.9	50
CA	4.4	0.2	2	1	5.8	31
C	4.6	0.1	2	1	3.7	26

Table A23.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A	3.4	53.8	6	74	19	7
CA	3.2	34.6	9	64	17	4
C	3.1	29.4	11	54	14	4

Table A24 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	FLUVISOL (41)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Hidromórfico
Date	02/22/2025
Coordinates UTM	272,665.072 x 9,901,503.765 m. fuse 23. Datum SIRGAS 2000
Land use	Secundary forest
Parent Material	recent alluvial deposits
Relief [2]	depressional to level
Elevation (m)	8.78 m
Erosion	Laminar
Permeability	poorly drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A24.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	140	160	700
A2	108	62	830
C	91	69	840

Table A24.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A1	3.0	2.5	4	2	15.3	215
A2	3.3	0.3	2	1	8.4	71
C	3.4	0.1	2	1	5.3	16

Table A24.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A1	8.7	223.4	4	64	65	5
A2	3.3	73.8	5	72	19	3
C	3.1	19.1	16	63	6	4

Table A25 Soil profiles descriptions and physicochemical analysis

Profile (n°)	FERRASOL (50)
Classification (SiBCS[1])	LATOSSOLO AMARELO Distrófico
Date	02/22/2025
Coordinates UTM	275,941.809 x 9,901,176.843 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Barreiras
Relief [2]	depressional to level
Elevation (m)	4.75 m
Erosion	Not apparent
Permeability	well drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A25.1 Granulometric analysis.

Horizon	Particle Size Analisys Clay	Silt	Sand
	---g/kg---		
A1	161	165	674
A2	280	94	626
BA	138	210	652
BW1	287	53	660
BW2	325	35	640
BW3	331	69	600

Table A25.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A1	3.9	0.5	2	1	11.3	39
A2	3.8	0.1	2	1	14.5	36
BA	4.1	0.1	2	1	12.1	34
BW1	4.0	0.1	2	1	11.2	26
BW2	4.1	0.3	2	1	9.4	23
BW3	4.0	0.2	2	1	9.6	23

Table A25.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

Horizon	SB	T	V	m	O.M.	P
	---mmol/dm³---		------%------		g/kg	mg/dm³
A1	3.6	42.8	8	76	13	3
A2	3.2	39.5	8	82	9	3
BA	3.1	36.9	9	79	7	3
BW1	3.1	29.4	11	78	4	3
BW2	3.3	26.7	12	74	4	4
BW3	3.2	26.1	12	75	3	3

Table A26 Soil profiles descriptions and physicochemical analysis

Profile (n°)	ARENOSOL (51)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	03/15/2025
Coordinates UTM	277,395.510 x 9,901,906.664 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	very gently sloping
Elevation (m)	90.64 m
Erosion	Extremely strong
Permeability	Excessively drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A26.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	110	84	806
A2	172	90	738
CA	140	262	598
C1	160	60	780
C2+	180	178	642

Table A26.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A1	3.8	0.2	2	1	8.9	44
A2	4.1	0.1	2	1	10.3	36
CA	4.2	0.1	2	1	8.9	31
C1	4.2	0.1	2	1	9.0	23
C2+	4.2	0.1	2	1	8.4	22

Table A26.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A1	3.2	47.2	7	73	15	8
A2	3.1	39.1	8	77	13	5
CA	3.1	33.9	9	74	9	3
C1	3.1	25.8	12	74	7	3
C2+	3.1	24.6	13	73	11	5

Table A27 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	ARENOSOL (57)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	04/03/2025
Coordinates UTM	277,214.180 x 9,884,263.368 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	Formation Tracuateua
Relief [2]	gently sloping
Elevation (m)	32.12 m
Erosion	not apparent
Permeability	well drained
Described by	Sobrinho. R. J. A;

Table A27.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A	98	64	838
CA	113	57	830
C1	172	392	436
C2	176	96	728
C3+	188	138	674

Table A27.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A	3.8	0.1	2	1	12.0	49
CA	4.0	0.1	2	1	15.6	46
C1	4.1	0.1	2	1	11.2	40
C2	4.2	0.1	4	1	6.5	25
C3+	4.2	0.1	4	1	3.6	19

Table A27.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A	3.1	51.9	6	79	24	5
CA	3.1	49.4	6	83	16	3
C1	3.1	42.7	7	78	9	3
C2	5.1	30.3	17	56	6	4
C3+	5.1	24.5	21	41	6	4

Table A28 Soil profiles descriptions and physicochemical analysis

Profile n°	GLEYSOL (58)
Classification (SiBCS[1])	GLEISSOLO MELÂNICO Ta Distrófico
Date	04/03/2025
Coordinates UTM	277,414.012 x 9,886,113.866 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	recent alluvial deposits
Relief [2]	depressional to level
Elevation (m)	12.87 m
Erosion	not apparent
Permeability	Very poorly drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A28.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A	181	386	433
Cg1	567	161	272
Cg2	432	172	396
Cg3	343	338	319

Table A28.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A	3.8	0.7	9	5	45.8	183
Cg1	3.5	1.2	5	4	109.6	244
Cg2	3.5	1.9	8	5	60.5	259
Cg3	3.3	1.0	6	3	57.1	224

Table A28.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A	15.0	198.4	8	75	31	6
C1g	10.5	254.1	4	91	18	4
C2g	15.1	274.5	6	80	14	3
C3g	10.3	234.2	4	85	12	3

Table A29 Soil profiles descriptions and physicochemical analysis.

Profile n°	ACRISOL (71)
Classification (SiBCS[1])	ARGISSOLO AMARELO Distrófico
Date	03/15/2025
Coordinates UTM	273,728.102 x 9,898,540.568 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Barreiras
Relief [2]	gently sloping
Elevation (m)	8.38 m
Erosion	not apparent
Permeability	well drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A29.1 Granulometric analysis.

	Particle Size Analisys		
	Clay	Silt	Sand
Horizon	---g/kg---		
A1	127	99	774
A2	145	133	722
AB	189	227	584
BA	274	218	508
BW1	290	108	602
BW2	339	83	578
BW3	359	159	482

Table A29.2 Soil pH. sorption complex and extractable acidity.

	pH	K	Ca	Mg	Al	H+Al
Horizon		---mmolc/dm³---				
A1	4.0	0.3	7	4	5.2	40
A2	3.9	0.1	2	1	6.9	46
AB	4.1	0.2	2	1	7.4	40
BA	4.0	0.1	2	1	10.7	45
BW1	4.0	0.1	2	1	12.9	35
BW2	4.0	0.1	4	1	5.7	30
BW3	4.2	0.1	2	1	4.5	20

Table A29.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A1	11.3	51.3	22	32	25	7
A2	3.1	49.4	6	69	22	5
AB	3.2	42.8	8	70	16	4
BA	3.1	48.0	7	78	14	4
BW1	3.1	37.6	8	81	12	6
BW2	5.1	34.6	15	53	9	4
BW3	3.1	23.3	13	59	6	4

Table A30 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	ARENOSOL (73)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	07/11/2023
Coordinates UTM	278,014.110 x 9,910,723.180 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	depressional to level
Elevation (m)	
Erosion	moderate
Permeability	well drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

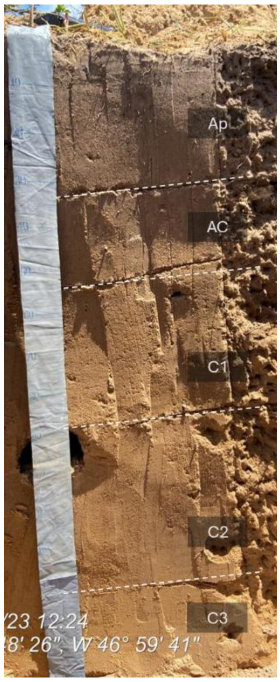


Table A30.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
Ap	75	2	902
AC	75	27	897
C1	49	22	929
C2	74	19	906
C3	62	25	913

Table A30.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	3.97	<0.2	0.5	0.6	9.5	33.1
AC	4	<0.2	<0.5	1.2	9.7	33.5
C1	4.33	<0.2	<0.5	0.3	7.8	29.8
C2	4.42	<0.2	<0.5	0.4	7.1	26.8
C3	4.42	<0.2	<0.5	0.4	6.4	22.4

Table A30.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	1.3	34.4	4	88	18.2	13.9
AC	1.8	35.3	5	84	16	10.8
C1	0.6	30.4	2	93	11.7	1.8
C2	0.7	27.5	2	91	10	1.2
C3	0.8	23.2	3	88	8.7	1.6

Table A31 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	SOLONCHAKS (74)
Classification (SiBCS¹)	GLEISSOLO SÁLICO Sódico
Date	08/15/2023
Coordinates UTM	278,477.960 x 9,910,784.850 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	recent alluvial deposits
Relief ²	depressional to level
Elevation (m)	6.7 m
Erosion	not apparent
Permeability	very poorly drained
Described by	Sobrinho. R. J. A;

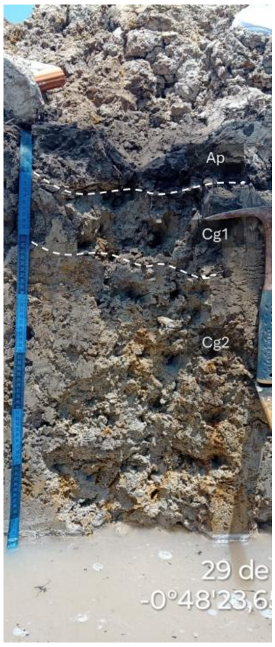


Table A31.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
Ap	406	113	481
Cg1	303	14	683
Cg2	279	66	655

Table A31.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.75	9.44	22.9	12.8	0.4	58.1
Cg1	4.04	4.71	8.6	51.8	2.2	14.7
Cg2	3.56	4.29	12.1	65.3	3.9	18.8

Table A31.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	422.8	480.9	88	0	44.4	10.3
Cg1	208.7	223.4	93	1	8.1	6.2
Cg2	315.2	334	94	1	7.0	6.9

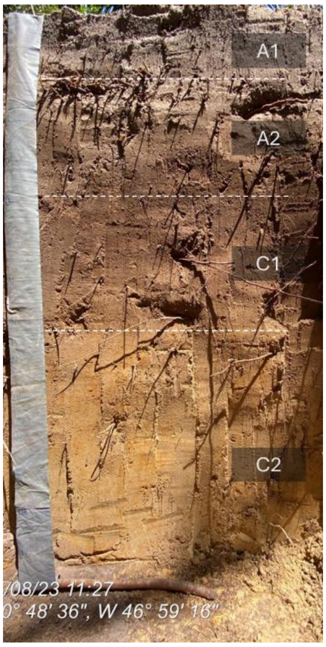


Table A32 Soil profiles descriptions and physicochemical analysis

Profile (n°)	ARENOSOL (75)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	08/15/2023
Coordinates UTM	278,818.320 x 9,910,416.350 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	depressional to level
Elevation (m)	
Erosion	moderate
Permeability	well drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A32.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	75	26	900
A2	100	17	883
C1	150	20	830
C2	87	21	892

Table A32.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm^3---				
A1	3.87	0.31	1	1.4	11	36.4
A2	4.46	<0.2	0.5	0.9	8.1	44.9
C1	4.6	<0.2	<0.5	0.6	5.4	34.7
C2	4.4	<0.2	0.5	0.9	4.8	17.4

Table A32.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm^3---		------%------		g/kg	mg/dm^3
A1	4.2	40.6	10	74	17.8	2.1
A2	1.8	46.7	4	82	14.4	2.1
C1	1.2	34.7	3	82	14.1	4.2
C2	1.6	17.4	9	75	22.3	1.5

Table A33 Soil profiles descriptions and physicochemical analysis.

Profile n°	ARENOSOL (90)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	01/03/2026
Coordinates UTM	276,439.000 x 9,903,968.000 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	very gently sloping
Elevation (m)	28.25 m
Erosion	well drained
Permeability	Excessively drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A33.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	217	133	650
AC	265	139	596
C	239	99	662

Table A33.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
			---mmolc/dm³---			
A1	4.1	0.1	5	2	12.0	53
AC	4.2	0.1	2	1	12.8	47
C	4.2	0.1	2	1	9.3	28

Table A33.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A1	7.1	59.6	12	63	18	4
AC	3.1	49.9	6	81	11	3
C	3.1	31.4	10	75	4	4

Table A34 Soil profiles descriptions and physicochemical analysis.

Profile n°	PODZOL (106)
Classification (SiBCS[1])	ESPODOSSOLO HIDROMÓRFICO Órtico
Date	11/29/2024
Coordinates UTM	267,681.000 x 9,886,154.000 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	recent alluvial deposits
Relief [2]	depressional to level
Elevation (m)	4.75 m
Erosion	moderate
Permeability	poorly drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

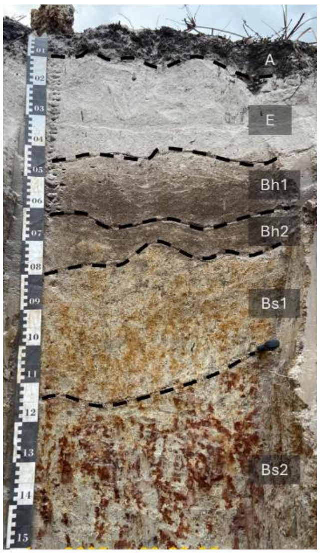


Table A34.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A	77	79	844
E	43	99	858
Bh1	93	81	826
Bh2	206	164	630
Bs1	293	201	506
Bs2	345	201	454

Table A34.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A	4.7	0.2	10	3	0.0	18
E	4.2	0.1	2	1	1.7	12
Bh1	4.1	0.3	2	1	10.3	28
Bh2	4.0	0.2	4	2	36.5	76
Bs1	3.9	0.1	4	1	54.6	105
Bs2	3.8	0.3	2	1	54.2	135

Table A34.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
A	13.3	31.1	43	0	25	6
E	3.1	15.0	21	35	3	3
Bh1	3.3	31.0	11	76	5	6
Bh2	6.3	82.2	8	85	5	4
Bs1	5.2	110.3	5	91	5	4
Bs2	3.3	138.6	2	94	3	7

Table A35 Soil profiles descriptions and physicochemical analysis

Profile n°	ARENSOL (107)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	11/29/2024
Coordinates UTM	267,987.000 x 9,887,491.000 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	Formation Tracuateua
Relief [2]	depressional to level
Elevation (m)	6.86 m
Erosion	Strong
Permeability	Excessively drained
Described by	Sobrinho. R. J. A;

Table A35.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
A1	77	77	846
C1	121	47	832
C2	108	142	750
C3	132	80	788

Table A35.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
A1	4.2	0.1	2	1	4.8	27
C1	4.3	0.1	2	1	5.3	30
C2	4.4	0.4	2	1	4.0	20
C3	4.3	0.1	2	1	6.7	20

Table A35.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm^3---		------%------		g/kg	mg/dm^3
A1	3.1	30.2	10	61	10	4
C1	3.1	32.6	10	63	7	4
C2	3.4	23.2	15	54	5	3
C3	3.1	23.1	13.5	68	3	4

Table A36 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	FERRASOL (109)
Classification (SiBCS[1])	LATOSSOLO AMARELO Distrófico
Date	12/08/2025
Coordinates UTM	259,678.000 x 9,868,236.000 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	Formation Barreiras
Relief [2]	gently sloping
Elevation (m)	48.77 m
Erosion	not apparent
Permeability	well drained
Described by	Sobrinho. R. J. A;

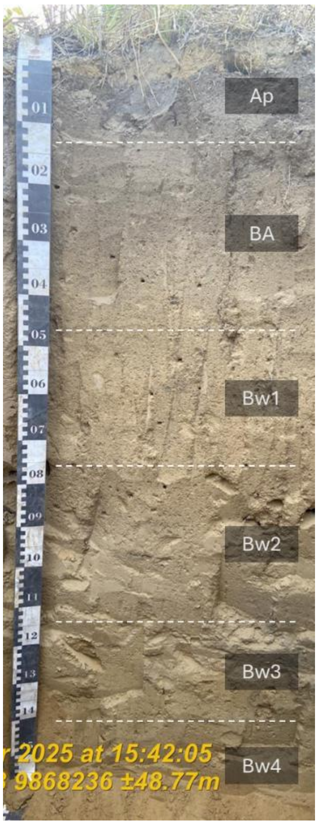


Table A36.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
Ap	178	100	722
BA	168	140	692
Bw1	178	118	704
Bw2	99	259	642
Bw3	185	71	744
Bw4	213	75	712

Table A36.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.0	0.6	4	1	14.2	48
BA	4.2	0.5	2	1	15.1	57
Bw1	4.1	0.5	2	1	15.4	40
Bw2	4.2	0.5	2	1	10.3	36
Bw3	4.2	0.1	2	1	8.1	23
Bw4	4.2	0.4	2	1	13.2	29

Table A36.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	5.7	53.5	11	71	15	5
BA	3.5	60.1	6	81	9	4
Bw1	3.5	43.9	8	81	8	3
Bw2	3.5	39.1	9	75	6	4
Bw3	3.1	26.3	12	72	5	3
Bw4	3.4	32.6	11	80	5	4

Table A37 Soil profiles descriptions and physicochemical analysis.

Profile (n°)	ARENOSOL (110)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	12/13/2025
Coordinates UTM	260,860.000 x 9,892,987.000 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	gently sloping
Elevation (m)	25.21 m
Erosion	moderate
Permeability	well drained
Described by	Sobrinho. R. J. A;

Table A37.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
Ap	268	68	664
C1	239	183	578
C2	233	231	536

Table A37.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.1	1.4	9	3	5.1	33
C1	4.2	0.5	2	1	7.3	30
C2	4.1	0.3	2	1	7.2	21

Table A37.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	13.4	46.2	29	28	18	7
C1	3.5	33.3	11	68	8	5
C2	3.3	24.1	14	69	5	11

Table A38 Soil profiles descriptions and physicochemical analysis

Profile n°	ARENOSOL (113)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	01/03/2026
Coordinates UTM	256,420.000 x 9,873,571.000 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Tracuateua
Relief [2]	depressional to level
Elevation (m)	10.61 m
Erosion	Slight
Permeability	Excessively drained
Described by	Sobrinho. R. J. A; Silva Júnior. J. F. da

Table A38.1 Granulometric analysis.

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---g/kg---		
Ap	116	74	810
AC1	171	39	790
C1	270	44	686
C2	248	50	702

Table A38.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.0	0.1	2	1	7.4	41
AC1	4.2	0.8	3	1	5.7	34
C1	4.2	0.1	4	1	5.3	31
C2	4.0	0.3	2	1	9.2	30

Table A38.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

Horizon	SB	T	V	m	O.M.	P
	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	3.1	44.4	7	70	12	5
AC1	4.8	39.0	12	54	10	4
C1	5.2	36.6	14	51	7	5
C2	3.4	33.8	10	73	6	4

Table A39 Soil profiles descriptions and physicochemical analysis.

Profile n°	FERRALSOL (114)
Classification (SiBCS[1])	LATOSSOLO AMARELO Distrófico
Date	01/06/2026
Coordinates UTM	278,536.000 x 9,880,220.000 m. fuse 23. Datum SIRGAS 2000
Land use	no tillage
Parent Material	Formation Barreiras
Relief [2]	gently sloping
Elevation (m)	52.03 m
Erosion	Strong
Permeability	Well drained
Described by	Sobrinho. R. J. A;

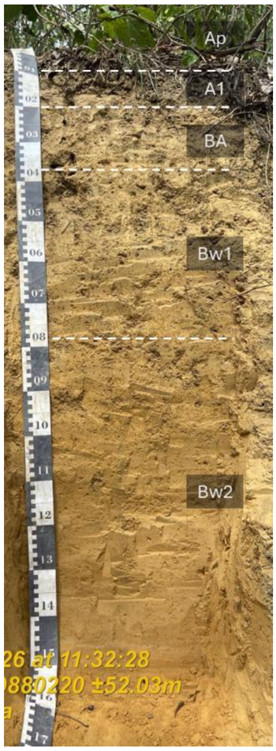


Table A39.1 Granulometric analysis.

Horizon	Particle Size Analisys		
	Clay	Silt	Sand
	---g/kg---		
Ap	203	89	708
A1	246	92	662
BA	281	89	630
Bw1	302	16	682
Bw2	280	160	560

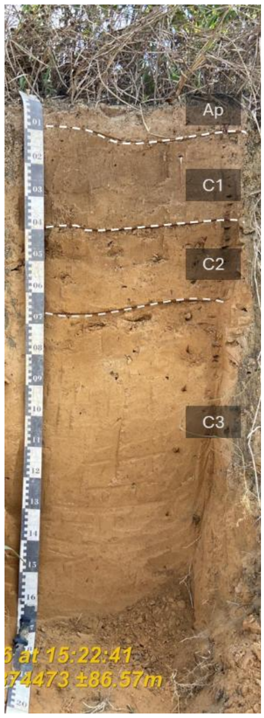
Table A39.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.1	2.7	4	3	8.0	40
A1	4.0	2.5	2	1	13.8	46
BA	4.0	0.6	2	1	15.4	53
Bw1	4.0	0.5	2	1	14.3	39
Bw2	4.0	0.3	2	1	11.4	32

Table A39.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	9.8	50.2	20	45	17	8
A1	5.6	51.9	11	71	14	6
BA	3.7	56.2	7	81	10	6
Bw1	3.6	42.3	8	80	9	7
Bw2	3.4	35.5	10	77	6	6

Table A40 Soil profiles descriptions and physicochemical analysis.



Profile n°	ARENOSOL (121)
Classification (SiBCS[1])	NEOSSOLO QUARTZARÊNICO Órtico
Date	01/06/2026
Coordinates UTM	269,024.000 x 9,874,473.000 m. fuse 23. Datum SIRGAS 2000
Land use	Pasture
Parent Material	Formation Tracuateua
Relief [2]	gently sloping
Elevation (m)	86.57 m
Erosion	moderate
Permeability	well drained
Described by	Sobrinho. R. J. A;

Table A40.1 Granulometric analysis.

	Particle Size Analisys		
Horizon	Clay	Silt	Sand
	---g/kg---		
Ap	107	55	838
C1	150	40	810
C2	134	42	824
C3	135	13	852

Table A40.2 Soil pH. sorption complex and extractable acidity.

Horizon	pH	K	Ca	Mg	Al	H+Al
		---mmolc/dm³---				
Ap	4.1	0.2	2	1	7.7	35
C1	4.2	0.1	2	1	9.1	32
C2	4.2	0.2	2	1	8.0	24
C3	4.3	0.1	2	1	7.0	21

Table A40.3 Sum of bases (SB). Cation exchange capacity (CEC at pH 7.0). base saturation (V). aluminum saturation (m). organic matter and remaining phosphorus assimilable.

	SB	T	V	m	O.M.	P
Horizon	---mmol/dm³---		------%------		g/kg	mg/dm³
Ap	3.3	37.8	9	70	12	5
C1	3.1	35.5	9	74	10	5
C2	3.2	26.9	12	71	7	6
C3	3.1	23.9	13	69	5	5

[1] SiBCS - Brazilian Soil Classification System
[2] Relief (Slope %): depressional to level (0-0.5); very gently sloping (0.5-2.0); gently sloping (2-5); moderately sloping (5-9); strongly sloping (9-15); steeply sloping (15-30); very steeply sloping (30-60); extremely sloping (over 60)