

UFRRJ

PRÓ-REITORIA DE EXTENSÃO
CURSO DE ESPECIALIZAÇÃO EM GEOPROCESSAMENTO,
LEVANTAMENTO E INTERPRETAÇÃO DE SOLOS

TRABALHO DE CONCLUSÃO DE CURSO

Mapeamento Digital de Solos via Aprendizado de
Máquina na Bacia do Alto Taquari (MS/MT)

Gabriela da Rocha Saldanha

2025





UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PRÓ-REITORIA DE EXTENSÃO
CURSO DE ESPECIALIZAÇÃO EM GEOPROCESSAMENTO,
LEVANTAMENTO E INTERPRETAÇÃO DE SOLOS

Mapeamento Digital de Solos via Aprendizado de Máquina na Bacia do
Alto Taquari (MS/MT)

GABRIELA DA ROCHA SALDANHA

Sob a Supervisão do(a) Professor(a)
Helena Saraiva Koenow Pinheiro

Trabalho de Conclusão de Curso
submetido como requisito parcial para
obtenção do título de **Especialista** em
Geoprocessamento, Levantamento e
Interpretação de Solos.

Seropédica, RJ
Julho de 2025



Universidade Federal Rural do Rio de Janeiro
Biblioteca Central/Seção de Processamento Técnico

Ficha catalográfica elaborada
Com os dados fornecidos pelo(a) autor(a)

dm da Rocha Saldanha, Gabriela, 1995-
Mapeamento Digital de Solos via Aprendizado de
Máquina na Bacia do Alto Taquari (MS/MT) / Gabriela
da Rocha Saldanha. - Rio de Janeiro, 2025.
38 f.

Orientador: Helena Saraiva Koenow Pinheiro.
Trabalho de conclusão de curso(Graduação). --
Universidade Federal Rural do Rio de Janeiro, Curso
de Especialização em Geoprocessamento, Levantamento e
Interpretação de Solos, 2025.

1. Aprendizado de máquina. 2. Random Forest. 3.
Unidades de Mapeamento. I. Saraiva Koenow Pinheiro,
Helena, 1985-, orient. II Universidade Federal Rural
do Rio de Janeiro. Curso de Especialização em
Geoprocessamento, Levantamento e Interpretação de
Solos III. Título. |

**O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de
Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001**

É permitida a cópia parcial ou total deste documento, desde que seja citada a fonte.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PRÓ-REITORIA DE EXTENSÃO
CURSO DE ESPECIALIZAÇÃO EM GEOPROCESSAMENTO, LEVANTAMENTO
E INTERPRETAÇÃO DE SOLOS

Folha de aprovação

GABRIELA DA ROCHA SALDANHA

Este Trabalho de Conclusão de Curso (TCC) foi submetido como requisito parcial para obtenção do título de Especialista em Geoprocessamento, Levantamento e Interpretação de Solos.

TCC APROVADO EM 15 de JULHO de 2025.

Profa. Dra. Helena Saraiva Koenow Pinheiro
(Supervisor(a))

Prof. Dr. Leonardo Durval Duarte Guimarães
UFRRJ

Dra. Ana Carolina de Souza Ferreira
UFRRJ

RESUMO

Mapeamento Digital de Solos via Aprendizado de Máquina na Bacia do Alto Taquari (MS/MT).

Saldanha, Gabriela. **Mapeamento de Classes de Solos da Bacia Hidrográfica do Alto Taquari (MS/MT)**. 2025. 38f. Trabalho de Conclusão de Curso (Especialização em Geoprocessamento, Levantamento e Interpretação de Solos). Pró-reitoria de Extensão, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

O mapeamento de classes de solos é importante para que decisões confiáveis sejam tomadas quanto ao uso e sustentabilidade ambiental. Nesse contexto, o objetivo do trabalho foi gerar um mapa preditivo de Unidades de Mapeamento (UM) de solo da Bacia Hidrográfica do Alto Taquari (BHAT), utilizando o modelo *Random Forest* (RF). Para isso foram utilizados 329 perfis de solos, oriundas de bases legadas, e alocadas em UMs de acordo com semelhanças taxonômicas. Para equilíbrio do número de amostras nas UMs, foram coletadas amostras sintéticas, com auxílio do *software* ArcGIS. Para predição foram utilizadas covariáveis de terreno, espectrais ópticas e de radar, e o modelo *machine learning* RF. Com relação as amostras coletadas, a UM com maior representação foi a dos Neossolos Quartzarênicos (153), seguido dos Latossolos Vermelhos (46), Latossolos Vermelhos psamíticos e Argissolos Vermelho-Amarelos (39). Segundo a modelagem, a distribuição dos solos está relacionada a distribuição de umidade, composição granulométrica, e presença de componentes ferruginosos, captados pelas covariáveis de terreno e índices espectrais. Entretanto, o modelo apresentou baixo desempenho preditivo, apresentando Kappa e Acurácia Global iguais a 0,29 e 37,54, respectivamente. O mapa produzido apresentou predominância das UMs com Neossolos Quartzarênicos e Litólicos, e baixa frequência e/ou ausência das UMs com Cambissolos Háplicos, Latossolos Vermelho e Latossolo Vermelho-Amarelos, ambos psamíticos, na área, reafirmando a baixa acurácia do modelo, visto que foram classes de solos frequentes na área de estudo. Esse resultado pode estar relacionado com a dificuldade do modelo em encontrar padrões nas covariáveis utilizadas e/ou a coleta de amostras sintéticas não muito eficiente, podendo ser aprimorada, sendo essas providências a serem tomadas para melhoria do modelo. Entretanto, considerando o mapa obtido, indicando predomínio das classes de solos Neossolos Quartzarênicos e Litólicos, que possuem ocorrências associadas a, principalmente, elevação, distância dos canais de drenagem e presença de minerais ferruginosos. A partir dessas informações ressalta-se a necessidade de práticas de manejo mais sustentáveis na bacia.

Palavras-chave: Aprendizado de máquina. *Random Forest*. Unidades de Mapeamento.

ABSTRACT

Digital Soil Mapping via Machine Learning in the Alto Taquari Basin (MS/MT)

Saldanha, Gabriela. **Digital Soil Mapping via Machine Learning in the Alto Taquari Basin (MS/MT)**. 2025. 38f. Course Completion Work (Specialization in Geoprocessing, Survey and Soil Interpretation). Pro-rectorate of Extension, Federal Rural University of Rio de Janeiro, Seropédica, RJ, 2025.

Soil class mapping is essential for making reliable decisions regarding environmental use and sustainability. In this context, the objective of this study was to generate a predictive map of soil Mapping Units (MU) in the Alto Taquari Watershed (BHAT) using the Random Forest (RF) model. For this purpose, 329 soil profiles from legacy databases were used and allocated to MUs according to taxonomic similarities. To balance the number of samples in the MUs, synthetic samples were collected using ArcGIS software. Terrain covariates, optical and radar spectral data, and the RF machine learning model were used for prediction. Regarding the samples collected, the MU with the greatest representation, according to the Brazilian soil classification system (SIBCS), was that of *Neossolos Quartzarênicos* (153), followed by *Latossolos Vermelhos* (46), *Latossolos Vermelhos psamíticos*, and *Argissolos Vermelho-Amarelos* (39). According to the modeling, soil distribution is related to moisture distribution, particle size composition, and the presence of ferruginous components, captured by terrain covariates and spectral indices. However, the model presented poor predictive performance, with Kappa and Global Accuracy values of 0.29 and 37.54, respectively. The map produced showed a predominance of MUs, with *Neossolos Quartzarênicos* e *Litólicos*, and a low frequency and/or absence of MUs with *Cambissolos Háplicos*, *Latossolos Vermelhos*, and *Latossolos Vermelho-Amarelos* (SIBCS), both psammitic, in the area, reaffirming the model's low accuracy, as these were common soil classes in the study area. This result may be related to the model's difficulty in finding patterns in the covariates used and/or the inefficient collection of synthetic samples, which could be improved. These measures should be taken to improve the model. However, considering the map obtained, which indicates a predominance of the *Neossolos Quartzarênicos* e *Litólicos* soil classes, which have occurrences associated primarily with elevation, distance from drainage channels, and the presence of ferruginous minerals, this information highlights the need for more sustainable management practices in the basin.

Keywords: Machine Learning. Random Forest. Soil Mapping Units.

LISTA DE FIGURAS

Figura 1. Limite da Bacia Hidrográfica do Alto Taquari.	8
Figura 2. Amostras de controle de campo e amostras sintéticas alocadas a partir das ROICs definidas.	11
Figura 3. Amostras de controle coletadas pela BHAT.	14
Figura 4. Graficos de importância das covariáveis de acordo com o modelo RF. A) Importância na Redução da Acurácia; B) Importância em relação a pureza dos nos (Índice Gini).	16
Figura 5. Erro do modelo RF considerando diferentes números de árvores.	18
Figura 6. Mapa com as predição das Unidades de Mapeamento (UMs) da Bacia Hidrográfica do Alto Taquari (BHAT).	22

LISTA DE TABELAS

Tabela 1. Unidades de mapeamento (UM), classes de solos e número de amostras coletadas.	9
Tabela 2. Amostras controle e sintéticas obtidas na BHAT.	12
Tabela 3. Covariáveis morfométricas e hidrológicas produzidas pelo software SAGA GIS para a área da BHAT.	13
Tabela 4. Parâmetros de validação do modelo <i>Random Forest</i> (RF) para os subconjuntos de treino e teste das amostras de controle e sintéticas da BHAT.	18
Tabela 5. Matriz de confusão das amostras de teste.	20

SUMÁRIO

1	INTRODUÇÃO.....	1
2	REVISÃO DE LITERATURA.....	3
2.1	Bacia Hidrográfica do Alto Taquari.....	3
2.2	Mapeamento Digital de Solos	4
2.3	Aprendizado de Máquina	5
2.4	Balanceamento de Amostras de Solos.....	6
3	MATERIAL E MÉTODOS.....	8
3.1	Área de estudo.....	8
3.2	Amostras de Perfis de Solo	8
3.3	Classes de Solos da Bacia Hidrográfica do Alto Taquari (BHAT)	9
3.4	Covariáveis Utilizadas na Predição.....	12
3.5	Seleção de Covariáveis	13
3.6	Modelagem e Avaliação dos Modelos	13
4	RESULTADOS E DISCUSSÃO.....	14
4.1	Avaliação da Predição de Unidades de Mapeamento (UMs) na BHAT	15
4.2	Mapa Preditivo de UMs Obtido pelo Modelo Random Forest.....	22
5	CONCLUSÕES	24
6	REFERÊNCIAS BIBLIOGRÁFICAS.....	25

1 INTRODUÇÃO

Mapas de solo apresentam papel de destaque em vários campos, fornecendo informações relevantes quanto a distribuição de propriedades e classes do solo. A confecção desses produtos atualmente é realizada de forma digital, através do aprendizado de máquina (*machine learning* – ML), que tem gradativamente substituído os modelos convencionais estatísticos devido sua maior flexibilidade e satisfatório desempenho preditivo (VAN DER WESTHUIZEN et al., 2024). Com o auxílio do cruzamento de informações do solo e covariáveis ambientais, modelos digitais de elevação (MDE) e imagens digitais de satélites, os mapas de solo têm se tornado cada vez mais detalhados, auxiliando a adoção de um manejo mais sustentável (MANTEGHI et al., 2024).

Observando todos os benefícios fornecidos pelas por essas tecnologias voltadas à ciência do solo, o conhecimento espacial de classes de solos poderiam ser úteis para estratégias de planejamento territorial e recuperação da Bacia Hidrográfica do Alto Taquari (MS/MT), visto os problemas ambientais que assolam essa área, devido uso intensivo do solo (RABELO E SILVA, 2021), que associado as características geomorfológicas e pedológicas, tem gerado graves problemas de erosão, e ocasionado o assoreamento do rio Taquari, que origina o nome da bacia.

Os avanços obtidos na ‘Era da Informação’ têm gerado diversos dados digitais ambientais, que associados a inteligência artificial (IA) e aprendizado de máquina ampliaram nossa capacidade de encontrar padrões complexos e otimizar processamento de grandes bases de dados, entretanto, esse conjunto volumoso de dados não está isento de desafios (BOHN E MILLER, 2024). Como os modelos ML são embasados em encontrar padrões entre dados de solo e covariáveis ambientais, e muitos deles são insensíveis ao número de covariáveis para explicar as relações encontradas (KASRAEI et al., 2024), há um interesse crescente em explicar como os modelos realizam a predição de atributos ou classes, visto que muito são ‘caixa preta’ (MINASNY et al., 2024). Weindorf e Chakraborty (2024) apontam um equilíbrio entre a modelagem ML e a percepção e o conhecimento pedológico, para que seja garantida a completude, validade e interpretação de resultados gerados digitalmente (BOHN E MILLER, 2024), afim de que sejam gerados mapas de solos, com precisão conhecida, e dependendo do modelo, sua incerteza (MANTEGHI et al., 2024).

Para o mapeamento de classes taxonômicas de solos, tão importante quanto o produto final, é entender as relações entre os fatores de formação, como a topografia, material de origem, e as classes de solos existentes, para que seja útil em subsidiar decisões importantes quanto à utilização e manejo do solo que deve ser empregado na área, assim como, para a melhoria dos modelos de predição através do aprendizado de máquinas, no que tange a seleção das covariáveis de *input* (BRUNGARD et al., 2015). Com o desenvolvimento sustentável em alta, diferentes modelos de ML estão sendo empregados para tal fim, entre eles o modelo *Random Forest* (RF), que tem apresentado bons resultados para predição de classes de solos ao redor do mundo (CAHYANA et al., 2023; JEUNE et al., 2018). Esse sucesso está relacionado a capacidade de lidar com dados desbalanceados, maior precisão na seleção de covariáveis durante a modelagem, reduzindo o *overfitting* por meio do treinamento de várias árvores,

aumentando sua capacidade de generalização dos dados (SALMAN; KALAKECH; STEITI, 2024).

Mesmo com bons classificadores, a precisão do modelo depende da representatividade e distribuição equilibrada das amostras de solo, sendo então um obstáculo o número desequilibrado de observações de amostras de solo entre as diferentes classes taxonômicas (WANG et al., 2023). Segundo os mesmos autores, esse desbalanceamento pode ocasionar baixa precisão do modelo, assim como a omissão de classes minoritárias. Como forma de resolver essa problemática diversas estratégias são estudadas para reduzir o erro que pode ser ocasionado pelos modelos, como as técnicas de reamostragem feitas ao nível dos dados, por serem mais simples de realizar o balanceamento amostral (SHARIFIFAR et al., 2019).

O método desenvolvido por Carvalho Junior et al. (2020), com a delimitação de polígonos irregulares ao redor de pontos amostrais de controle (ROICs), tem como finalidade balancear o número de classes minoritárias com as demais com maior número de amostras, além de produzir amostras sintéticas espacialmente adequada, e consequentemente, aumentar o número pontos coletados em áreas com baixa densidade amostral. Por meio desse método, os autores obtiveram bons resultados na área da Bacia Hidrográfica do Bonfim (RJ), sendo essa estratégia adotada para balanceamento amostral das classes encontradas na Bacia Hidrográfica do Alto Taquari (BHAT - MS/MT), área abordada pelo presente estudo.

Nesse contexto, como o modelo RF tem apresentado bons resultados para a predição de classes em diferentes regiões do mundo, conforme relatado na literatura, devido sua capacidade de captar relações lineares e não lineares, hipotetiza-se que o modelo consiga predizer, de forma satisfatória, as UMs presentes na BHAT, a fim de que sejam produzidas mapas pedológicos que possam embasar recomendações de manejo mais adequadas para a área.

Este trabalho tem como objetivo geral realizar a predição de classes taxonômicas de solos da BHAT (MS/MT) por meio do modelo *Random Forest* (RF).

2 REVISÃO DE LITERATURA

2.1 Bacia Hidrográfica do Alto Taquari

A Bacia Hidrográfica do Alto Taquari (BHAT), compreende a área do planalto drenado pelo Rio Taquari e seus afluentes até a escarpa cuneiforme da Bacia Sedimentar do Paraná, próxima à Cidade de Coxim (GALDINO et al., 2005). Sua superfície é de aproximadamente 27.839 km², com cerca de 13,8% (3.823 km²) da área no MT e 86,2% (24.016 km²) no MS. Os municípios que a compõem são: Alta Araguaia e Alto Taquari (MT), Pedro Gomes, Alcinoópolis, Costa Rica, Coxim, Figueirão, Rio Verde de Mato Grosso, São Gabriel do Oeste, Camapuã e pequena porção de Sonora (MS).

Segundo GALDINO E PELLEGRIN (2003), por meio de cruzamento de informações cartográficas de Geologia, Geomorfologia, Vegetação e Clima, forneceram as Unidades Ambientais Naturais da BHAT, denominadas como: Planalto do rio Correntes, Chapada das Emas, Planalto do Taquari, Chapada de São Gabriel, Coxim e Campo Grande, Planalto de Maracaju – Campo Grande. Segundo os mesmos autores, as cinco primeiras unidades, apresentam predomínio da formação Bauru e Botucatu, sobretudo no Planalto do Taquari, que corresponde a 18.457 km² (65,9% da área). O Planalto Taquari é formado, principalmente, pelas formações citadas, e em menor proporção pela Pirambóia, Palermo e Aquidauana; originando solos principalmente das classes Areia Quartzosas (atuais Neossolos Quartzarênicos) e Podzólicos Vermelho-Amarelo (atuais Argissolos Vermelho-Amarelo) com textura arenosa/média, ambos apresentando caráter álico. Em menor proporção, também podem ser encontrados classes como Latossolos Vermelho e Vermelho-Amarelos de textura média.

Sendo um dos seguimentos da Bacia do Alto Paraguai, a Bacia Hidrográfica do Rio Taquari (BHRT), apresenta área, majoritariamente, de colinas de topos amplos e convexos, e baixa densidade de canais de drenagem, entretanto, os processos erosivos pluviais são elevados (GALDINO et al., 2005). Esse desequilíbrio ambiental pode estar associado as características de relevo e solo, assim como o material originário, que formam solos com baixíssima resistência a erosão (FERREIRA E AQUINO, 2007), como os citados anteriormente.

Os processos erosivos da BHAT é um assunto estudado desde os anos de 1970, onde vem sendo documentado os prejuízos ambientais, sociais e econômicos trazidos pelo intenso arraste de partículas para a parte baixa da BHRT, que além das características ambientais favoráveis, também está associada a substituição de grandes áreas de mata nativa, em áreas de desenvolvimento agropecuário (SILVA; MEDEIROS; CARVALHO, 2007). Tal fenômeno associou a BHAT a uma grande fonte de sedimentos, carregada de produtos químicos, fertilizantes e pesticidas residuais, que tem ocasionado o assoreamento do Rio Taquari e mudança do curso do rio, levando a inundação permanente de milhares de quilômetros quadrados (RABELO E SOUZA, 2021), assim como, a expulsão de famílias das pantaneiras, e aumentando o isolamento físico e comercial das comunidades da área (CURADO, 2004).

Diante desse cenário, o mapeamento digital de solos pode ser útil fornecendo informações das classes taxonômicas dos solos, a fim de que esses dados possam ser utilizados futuramente para planos de manejo, auxiliando o uso racional e sustentável dos solos, visto constatada fragilidade das características dos solos frente as mudanças de uso e ocupação.

2.2 Mapeamento Digital de Solos

Os diferentes modos para produção de mapas de solos, convencional ou digital, com o objetivo de produzir mapas categóricos de solos, apresentam como principais diferenças o custo-benefício, o tamanho amostral, a precisão e o nível taxonômico desejado (ZERAATPISHEH et al., 2017). O método convencional exige densa amostragem, aquisição de informações amplas sobre o solo e análises laboratoriais, perde-se muito tempo e apresentando considerável custo (NEYESTANI et al., 2021). Além disso, são produzidos a partir da interpretação do pedólogo, que correlaciona observações morfológicas de campo e dados analíticos, a fim de alocar as classes e atributos do solo em mapas, considerando os sistemas taxonômicos preexistentes, o que dificulta a reprodutibilidade por outros profissionais da área (TAGHIZADEH-MEHRJARDI et al., 2019). Dessa forma, abre-se espaço para que novas técnicas mais rápidas e econômicas, como o mapeamento digital de solos (MDS), se desenvolvam e sejam aplicadas nas diferentes partes do mundo (NEYESTANI, et al., 2021).

A reformulação da função que explica a ocorrência dos solos e suas propriedades realizada por McBratney et al. (2003) preconizada como S.C.O.R.P.A.N., com base nos fatores de formação (JENNY., 1941), considerando a localização espacial (“n”) e informações pretéritas dos tipos de solos existentes na área em questão (“s”), além do clima, relevo, material de origem e tempo, abriu novo horizonte para a formulação e produção de mapas de solos. Com os avanços que ocorreram nos ramos da computação e da tecnologia da informação, ciência do solo, assim como áreas de atuação, foi consideravelmente beneficiada com a alimentação contínua de bancos de dados, desde níveis regionais a mundial, e com o desenvolvimento de ferramentas como sistemas de informação geográfica (SIG), GPS, sensores remotos e proximais; e fontes de dados, como os modelos digitais de elevação (MDEs) (MCBRATNEY; SANTOS; MINASNY, 2003; WADOUX; MINASNY; MCBRATNEY, 2020) e suas covariáveis ambientais derivadas (*slope*, *Topography Wetness Index* (TWI), LS fator, entre outras). Tais avanços ocorreram em associação com os algoritmos de inteligência artificial (IA) e aprendizado de máquina, que possibilitaram, a partir do cruzamento de dados amostrais com variáveis ambientais, compreender e prever mapas de atributos e classes de solos (BOHN e MILLER, 2024).

As tecnologias que envolvem o MDS estão possibilitando a substituição de mapas produzidos convencionalmente, baseados em polígonos, não só para predição de propriedades contínuas (textura, carbono orgânico, fertilidade, etc.) e características descontínuas e horizontes de solo, com a vantagem de ter suas estimativas de incerteza associadas (ZERAATPISHEH et al., 2020), mas também é apropriado para aperfeiçoamento da qualidade de mapeamentos anteriores (GRUNWALD, 2009), sem muita amostragem e análises laboratoriais adicionais, visto que mapas convencionais contém precioso conhecimento das relações solo-paisagem; entretanto, apresentam limitações quanto ao nível de detalhamento espacial, e conseqüentemente, da previsão do objeto mapeado (YANG et al., 2011). Entretanto, vale ressaltar que o MDS apresenta componentes estruturantes, que segundo Lagacherie e McBratney (2006) são eles: i) dados de entrada, com dados amostrais e laboratoriais, incluindo técnicas de amostragem estatística; ii) algoritmos preditivos, que buscam padrões dentro dos dados de campo e covariáveis ambientais ou fatores scorpan; e iii) produto final (saída), onde se originam os produtos finais de predição (rasters preditivos), com suas incertezas associadas.

O método digital é mais robusto de mapeamento, desenvolveu-se em uma época de grande busca global por dados e informações sobre o solo, para seu monitoramento e modelagem ambiental (MCBRATNEY; SANTOS; MINASNY, 2003). A partir de informações do solo de boa qualidade e alta resolução espacial, as sequências dos procedimentos de mapeamento foram implementadas e evoluiu a ponto de aumentar consideravelmente a precisão da metodologia e dos mapas gerados (KUMARAPERUMAL et al., 2022), o que é essencial para a sustentação de práticas apropriadas de manejo e uso da terra e seus recursos, bem como do equilíbrio ambiental (ZERAATPISHEH et al., 2017).

2.3 Aprendizado de Máquina

Diversos fatores contribuíram intensamente para a popularização do MDS em escala global, entre eles, o crescente interesse por dados dos solos, acúmulo de banco de dados com propriedades do solo, acessibilidade a covariáveis ambientais amplamente conhecidas, e o desenvolvimento de modelos numéricos e suporte computacional que pudessem cruzar e explorar os dados do solo (WADOUX; MINASNY; MCBRATNEY, 2020). Tais avanços alavancaram a migração do método convencional para o digital de mapeamento, na virada do século XX e XXI, com o objetivo de tornar mais eficientes e precisos os produtos finais gerados (SULEYMANOV; ARROUAYS; SAVIN, 2024).

Incontáveis métodos se desenvolveram para prever a distribuição espacial do solo nas últimas décadas, sendo os principais baseados em interpolação geoestatística e aprendizado de máquina (*machine learning* - ML) (ZHANG et al., 2021). Segundo Wadoux; Minasny; McBratney (2020) os modelos geoestatísticos, apresentam diversas vantagens, como solidez estatística; autocorrelação espacial e apuração explícita de incerteza. Entretanto, apresenta também limitações, como resíduos distribuídos normalmente; necessidade de estimar diversos parâmetros, tornando mais complexos a modelagem; e em áreas homogêneas, o modelo de variação espacial não captura mudanças graduais e abruptas. Por outro lado, o ML é principalmente destinado a mineração de dados e reconhecimento de padrões, por meio de amplo leque de algoritmos não lineares, sendo esse método frequentemente utilizado para dados contínuos e categóricos. Ou seja, comparativamente, o ML não precisa admitir rigorosa estatística sobre a distribuição espacial das amostras de solo, sendo o algoritmo orientado por dados que constroem relações solo-paisagem a partir da correlação entre amostras de solo e covariáveis ambientais (ZHANG et al., 2021).

Entre os modelos de ML, até 2011 os mais utilizados para classificação eram as redes neurais, regressão logística e árvore de decisão, entretanto, nos anos seguintes com o compartilhamento gratuito de códigos utilizados em plataformas de código aberto (software R), diferentes modelos puderam ser desenvolvidos, e popularizados, em pesquisas relacionadas ao MDS (MEIER et al., 2018). Para predição, tais modelos se baseiam na correlação de dados de laboratório e covariáveis ambientais e/ou geoespaciais, podendo estas últimas serem ranqueadas a partir de sua importância (MA et al., 2019). Entretanto, é observado que diferentes modelos, com um mesmo conjunto de dados amostrais e covariáveis ambientais, podem apresentar consideráveis diferenças entre parâmetros de validação, sendo interessante a análise de diferentes modelos preditivos (ZHANG et al., 2021; PADARIAN; MINASNY; MCBRATNEY, 2020; HEUNG et al., 2016). Vale ressaltar que os produtos oriundos do

processamento ML não devem ser traduzidos de forma direta como conhecimento pedológico, visto que para ser considerado dessa forma, requer mais do que reconhecimento de padrões, é necessário também a pré-seleção e pré-processamento das covariáveis pedologicamente relevantes, e posterior interpretação e avaliação dos padrões reconhecidos (WADOUX et al., 2019).

Entre os modelos desenvolvidos e amplamente utilizados KHALEDIAN E MILLER (2020) citam *Multiple Linear Regression* (MLR), *K-Nearest Neighbors* (KNN), *Support Vector Regression* (SVR), *Cubist*, *Random Forest* (RF), e *Artificial Neural Networks* (ANN), sobretudo o RF que na última década ganhou ampla popularidade, visto que esse algoritmo, e Cubist, são baseados em árvores de decisão/reressao, com fácil implementação no que tange a configuração de parâmetros de treinamento. O RF apresenta relevante robustez devido a utilização de diversas árvores de aprendizado, evitando um aprendizado superficial das amostras de treinamento, e consequentemente, aumentando sua capacidade de extrapolação eficiente dos dados, além de conseguir resolver problemas de relacionamentos lineares e não lineares, sendo esses pontos relevantes para sua aplicabilidade frente a outros modelos (SALMAN; KALAKECH; STEITI, 2024).

Khaledian e Miller (2020) afirmam ainda que alguns modelos mesmo apresentando resultados satisfatórios em diferentes trabalhos, devido às peculiaridades que embasam cada modelo, levando em consideração seus pontos fortes e fracos, o usuário deve ter bem delimitado o propósito da pesquisa e as limitações existentes, para que seja reconhecido o algoritmo que melhor se adeque às expectativas e condições de mapeamento.

Dentre os modelos que são amplamente utilizados, o RF é um algoritmo conceitualmente embasado em árvores de decisão/regressão, e consequentemente, proporciona as mesmas vantagens, entretanto, múltiplas árvores são treinadas e os resultados são oriundos da previsão de diversas árvores individuais (BREIMAN, 2001). Dentro de modelos que tem árvores em suas estruturas, há a associação de um nó, ou folha, de uma árvore com as partições de covariáveis realizadas anteriormente, dessa forma, as árvores são ajustadas por divisão binária recursiva, a fim de que sejam reduzidas as impurezas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Segundo Westhuizen, Heuvelink e Hofmeyr (2023) o alto grau de impureza das folhas está associada a respostas apresentarem alto grau de dispersão, sendo por isso determinada pela relação da distribuição da variável de resposta relacionada aos vetores de covariância nela relacionada. Os ajustes para obtenção de bons resultados do modelo ocorre por meio do treinamento a partir de amostragem *bootstrap* aleatória para conjunto de treinamento, assim como o subconjunto de preditores usados para as regras de divisão de nós, o que torna o modelo mais robusto (HEUNG et al., 2016). Devido a tal estrutura, o sucesso do modelo para o MDS tem se tornado mais proeminente (HEUNG ET AL., 2016), tanto para atributos, como argila, areia (DHARUMARAJAN E HEGDE, 2020; NGU et al., 2024), carbono orgânico (EMADI et al., 2020), assim como para classes de solos (BRUNGARD et al., 2015; FRANCOS et al., 2025).

2.4 Balanceamento de Amostras de Solos

Mapas que apresentem a distribuição espacial das classes taxonômicas do solo são de suma importância em diferentes partes do mundo, e em diferentes escalas, pois a partir desse

conhecimento, decisões de uso e manejo de solo são determinados (BRUNGARD et al., 2015). Nesse contexto o MDS relacionando características de solos com covariáveis ambientais, os modelos são ajustados a partir das amostras e, posteriormente, utilizados para estimar atributos em áreas não amostradas (HEUNG et al., 2016). Entretanto, no caso de previsões qualitativas, como as unidades taxonômicas, afim de realizar a classificação (MCBRATNEY; SANTOS; MINASNY, 2003), um importante ponto a ser observado com relação a modelagem de classes de solos é o desbalanceamento amostral, visto que, o desequilíbrio amostral pode, tanto superestimar as classes majoritárias, quanto subestimar ou perder classes minoritárias, excluindo essas classes menos amostradas do mapa final (SHARIFIFAR et al., 2019).

A diversidade de classes em uma área é associada a função da complexidade ambiental e da origem do sistema de classificação taxonômica adotado, correspondente a área (SHARIFIFAR et al., 2019). Atualmente os modelos ML, e simulações, presumem que há distribuição equilibrada de entrada de dados, apresentando baixa precisão quando treinados com classes desequilibradas, dessa forma o equilíbrio amostral de classes de solo é um problema que afeta a predição desses mapas (SHARIFIFAR E SARMANDIAN, 2023). Segundo os mesmos autores, tal problemática não foi bem estudado na ciência do solo, resultando em perda de resultado, e consequentemente, exclusão de classes minoritárias, produzindo mapas incertos ou enganosos. Para resolução desse problema, algumas abordagens foram realizadas, incluindo a técnica de reamostragem (aumentando e diminuindo sinteticamente amostras minoritárias e majoritárias, respectivamente), *ensemble learning* (combinação de diferentes classificadores para obter um novo e melhor) (GALAR et al., 2012) e ajuste de algoritmo (alteração das funções dos modelos a fim de valorizar classes minoritárias, principalmente por modelos sensível a custo) (SIERS E ISLAM, 2018).

Entre todos os métodos, e como apontado por SHARIFIFAR et al. (2019) o ajuste em nível de dados é o mais simples de ser realizado, e tem sido eficaz para a predição de classes. Além do ajuste, a utilização de modelos adequados, como os modelos baseados em árvores (WANG et al., 2019; CARVALHO JUNIOR ET AL., 2020) e que apresentam maior complexidade (BRUNGARD et al., 2015), podem otimizar e melhorar os resultados obtidos por meio dos mapas finais. Entre as alternativas de ML até então desenvolvidas, o RF tem sido utilizado e apresentado bons resultados para a previsão de classes de solos, quando comparados a outros algoritmos, além de fornecer a importância a relativa de cada covariável utilizada no modelo (TAGHIZADEH-MEHRJARDI et al., 2019)

Como forma de adequar as classes de solos encontradas área, utilizando o ajuste a nível de dados, CARVALHO JUNIOR et al. (2020) propôs o método denominado ROICs (delimitação de áreas irregulares ao redor de pontos amostrais de controle), afim de que fossem equiparados os números amostrais dentro de cada unidade de mapeamento (UM), a partir da coleta de amostras sintéticas espacialmente consistentes. Essa técnica foi inicialmente utilizada na Bacia Hidrográfica do Bonfim (RJ), realizada a partir da definição de polígonos (áreas) irregulares (ROICs) ao redor dos pontos de controle de solo. Os resultados obtidos com a predição das classes da área em questão se mostraram animadores, indicando a técnica promissora para utilização em outras áreas, inspirando o presente estudo. Dessa forma, o método ROICs foi utilizado para coleta de amostras sintéticas na área da BHAT, a fim de

equilibrar o número de amostras dentro de cada UM, favorecendo que classes taxonômicas menos amostradas fossem igualmente favorecidas para melhor predição.

3 MATERIAL E MÉTODOS

3.1 Área de estudo

A área de estudo corresponde à Bacia Hidrográfica do Alto Taquari (BHAT), com aproximadamente 27.839 km², localizada entre 17°14'42.5"S 19°38'08.5"S e 53°08'14.8"O, 55°05'30.4"O (Figura 1), majoritariamente no estado do Mato Grosso do Sul. O rio Taquari, que possui 801 km de extensão, tem sua nascente no Planalto Central, a 900 m de altitude, e sua foz em Coxim (MS), com altitude reduzida para 200 m.

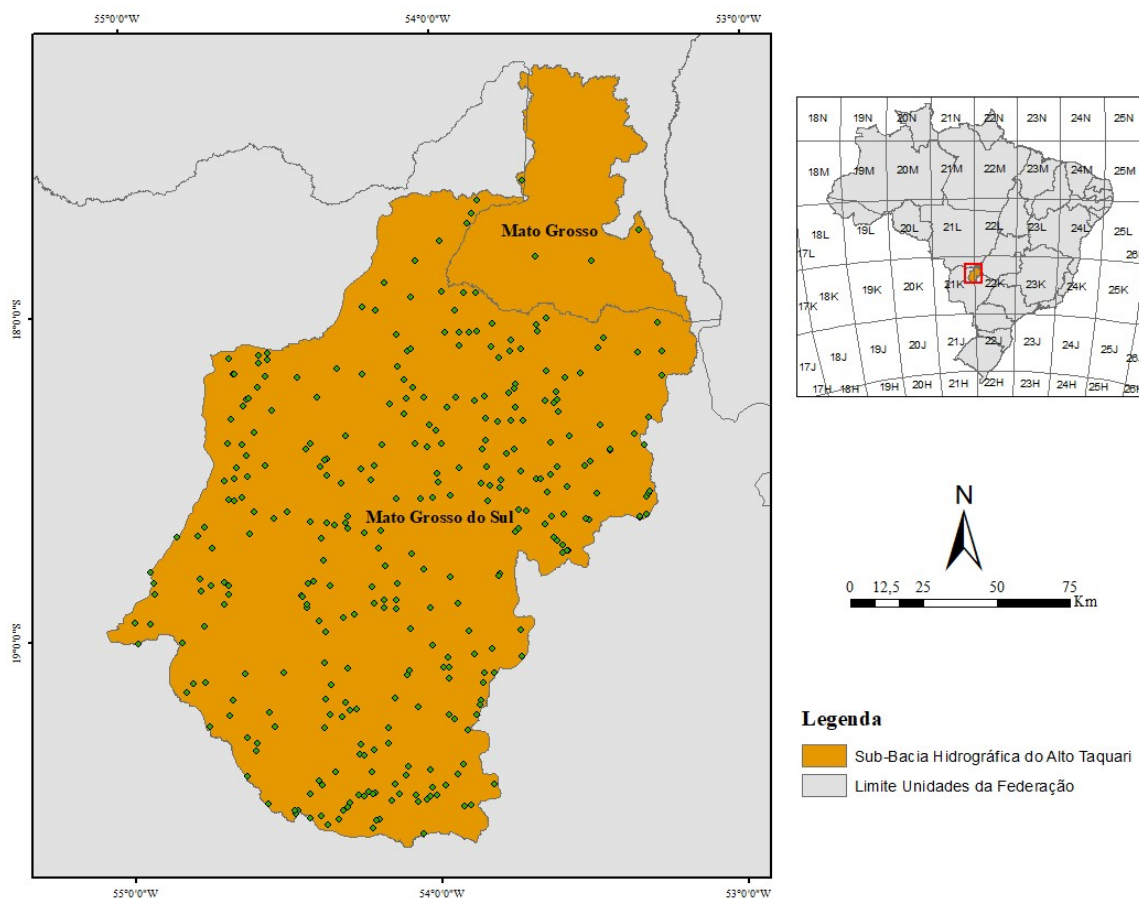


Figura 1. Limite da Bacia Hidrográfica do Alto Taquari.

3.2 Amostras de Perfis de Solo

Foram utilizados 328 perfis de solo nos procedimentos de mapeamento digital. Tais perfis foram oriundos do Projeto de Zoneamento Agroecológico do Mato Grosso do Sul, realizado pela Embrapa Solos e Governo do Estado do Mato Grosso do Sul; e do levantamento da Folha SE.22 Goiânia, volume 31 do mapeamento pedológico realizado pelo Projeto RADAMBRASIL (BRASIL, 1983).

3.3 Classes de Solos da Bacia Hidrográfica do Alto Taquari (BHAT)

Foram definidas UMs com base na classe taxonômica de solo dominante em cada unidade. Dessa forma, as UMs apresentadas continham as seguintes classes de solos:

Tabela 1. Unidades de mapeamento (UM), classes de solos e número de amostras coletadas.

UMs	Classes de solos	Nº amostras
HIDROMÓRFICOS	NEOSSOLO QUARTZARÊNICO Hidromórfico típico	4
	GLEISSOLO HÁPLICO Tb Distrófico	1
	PLANOSSOLO NÁTRICO Órtico típico	1
CX_TÍPICO	CAMBISSOLO HÁPLICO Tb distrófico	6
	CAMBISSOLO HÁPLICO Ta aluminico	3
	CAMBISSOLO HÁPLICO Tb eutrófico	1
RL_TIPICO	NEOSSOLO LITÓLICO Distrófico típico	3
	NEOSSOLO LITÓLICO Eutrófico fragmentário	3
	CAMBISSOLO HÁPLICO Tb Distrófico saprolítico	1
	NEOSSOLO LITÓLICO Chernossólico fragmentário	1
	NEOSSOLO LITÓLICO Chernozêmico típico	1
	NEOSSOLO LITÓLICO Distrófico fragmentário	1
	NEOSSOLO LITÓLICO Eutrófico típico	1
	NEOSSOLO REGOLÍTICO Eutrófico típico	1
LVA_PSAM	NEOSSOLO REGOLÍTICO Psamítico típico	1
	LATOSSOLO VERMELHO-AMARELO Distrófico psamítico	13
	LATOSSOLO AMARELO Distrófico psamítico	7
	LATOSSOLO AMARELO Aluminico psamítico	1
LV_PSAM	LATOSSOLO VERMELHO-AMARELO Aluminico psamítico	1
	LATOSSOLO VERMELHO Distrófico psamítico	28
	NEOSSOLO QUARTZARÊNICO Órtico latossólico	9
	LATOSSOLO VERMELHO Ácrico psamítico	1
PVA_TIPICO	LATOSSOLO VERMELHO Eutrófico psamítico	1
	ARGISSOLO VERMELHO-AMARELO Distrófico típico	10
	ARGISSOLO AMARELO Distrófico típico	4
	ARGISSOLO VERMELHO-AMARELO Eutrófico abruptico	3
	ARGISSOLO VERMELHO-AMARELO Ta Aluminico típico	3
	ARGISSOLO VERMELHO Distrófico latossólico	2
	ARGISSOLO VERMELHO Eutrófico típico	2
	ARGISSOLO VERMELHO-AMARELO Ta Distrófico arênico	2
	ARGISSOLO AMARELO Eutrófico típico	1
	ARGISSOLO AMARELO Ta Aluminico abruptico	1
	ARGISSOLO VERMELHO Distrófico abruptico	1
	ARGISSOLO VERMELHO Distrófico típico	1
	ARGISSOLO VERMELHO Eutrófico arênico	1
	ARGISSOLO VERMELHO-AMARELO Distrófico abruptico	1



	ARGISSOLO VERMELHO-AMARELO Distrófico arênico	1
--	---	---

Continua...

Continuação...

UMs	Classes de solos	Nº amostras
PVA_TÍPICO	ARGISSOLO VERMELHO-AMARELO Distrófico espessarênico	1
	ARGISSOLO VERMELHO-AMARELO Distrófico plintossólico	1
	LUVISSOLO CRÔMICO Pálico abrupto	1
	LUVISSOLO CRÔMICO Pálico espessarênico	1
	LUVISSOLO CRÔMICO Pálico plintossólico	1
	LUVISSOLO CRÔMICO Pálico típico	1
LV_TÍPICO	LATOSSOLO VERMELHO Distrófico típico	19
	LATOSSOLO VERMELHO Ácrico típico	3
	LATOSSOLO VERMELHO-AMARELO Distrófico típico	3
	LATOSSOLO AMARELO Distrófico típico	2
	LATOSSOLO VERMELHO Distrófico argissólico	2
	LATOSSOLO VERMELHO-AMARELO Distrófico argissólico	2
	LATOSSOLO AMARELO Distrófico argissólico	1
	LATOSSOLO AMARELO Distrófico plintossólico	1
	LATOSSOLO VERMELHO Acriférico típico	1
	LATOSSOLO VERMELHO Distrófico cambissólico	1
	LATOSSOLO VERMELHO Escuro álico	1
	LATOSSOLO VERMELHO Escuro distrófico	1
	LATOSSOLO VERMELHO Eutrófico típico	1
	LATOSSOLO VERMELHO-AMARELO Ácrico argissólico	1
	LATOSSOLO VERMELHO-AMARELO Ácrico típico	1
	LATOSSOLO VERMELHO-AMARELO Distrófico petroplântico	1
	NITOSSOLO VERMELHO Distrófico típico	1
	PLINTOSSOLO HÁPLICO Distrófico petroplântico	1
	PLINTOSSOLO PÉTRICO Concrecionário argissólico	1
	PLINTOSSOLO PÉTRICO Concrecionário típico	1
	PLINTOSSOLO PÉTRICO Litoplintico típico	1
RQ_TÍPICO:	NEOSSOLO QUARTZARÊNICO Órtico típico	148
	NEOSSOLO QUARTZARÊNICO Órtico eútrico	3
	NEOSSOLO QUARTZARÊNICO álico	1
	NEOSSOLO QUARTZARÊNICO Órtico solódico	1

Seguindo a metodologia proposta por Carvalho Júnior et al. (2020), foram definidas ROICs, para posterior obtenção de amostras sintéticas, como demonstrado na Figura 2. As amostras de controle são amostras obtidas em campo, que embasam a coleta das amostras sintéticas, respeitando características similares de paisagem, o que embasa a definição das Unidade de Mapeamento (UM) (IBGE, 2015).

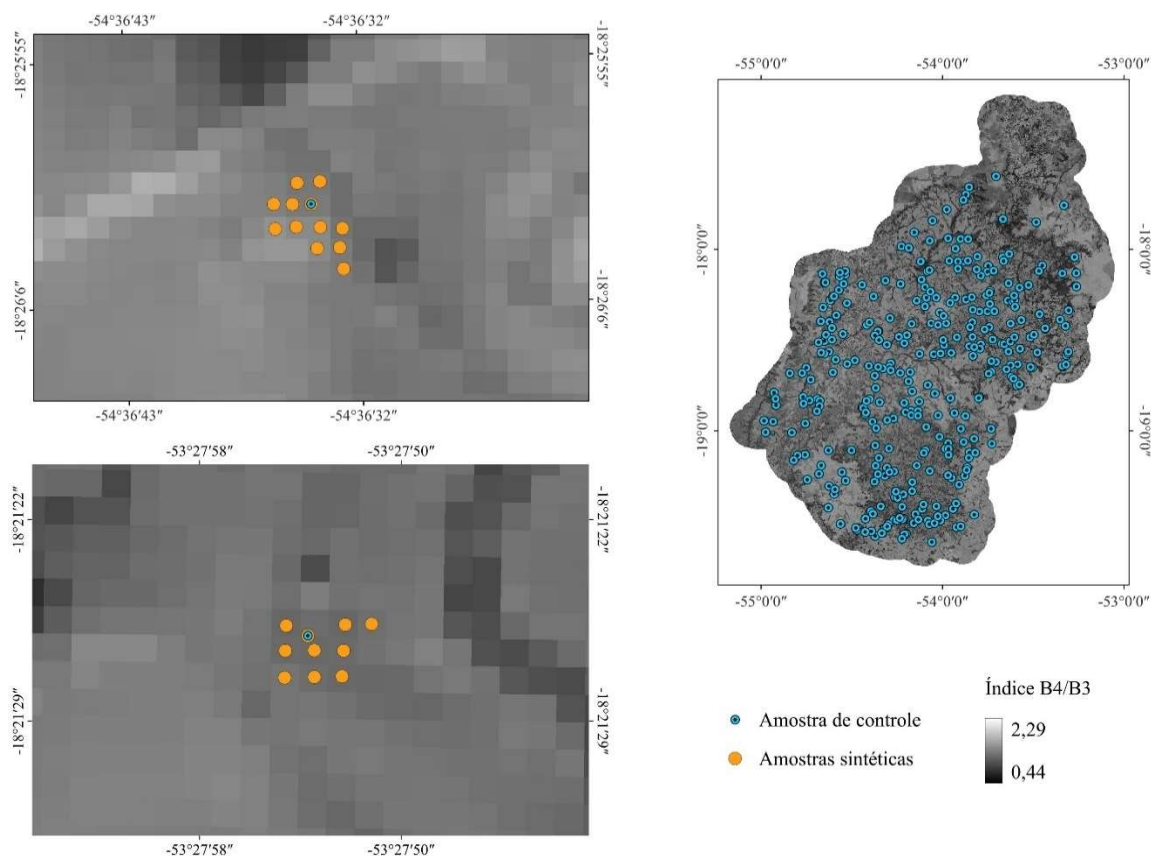


Figura 2. Amostras de controle de campo e amostras sintéticas alocadas a partir das ROICs definidas.

Para a coleta das amostras sintéticas das UMs minoritárias, foi considerado o número de amostras da UM que apresentou maior amostras, no caso RQ_TÍPICO que tinha 153 amostras. Esse valor era subtraído do valor de amostras da UM que teria que ser balanceada, para que fosse conhecida quantas amostras sintéticas deveriam ser coletadas, e posteriormente, dividido pelo número de amostras de controle, para que fossem coletadas amostras sintéticas de forma equilibrada entre os pontos de controle. Ex.: Para equilibrar a UM LV_PSAM:

$$153 - 39 \approx 114 \text{ amostras sintéticas deveriam ser coletadas}$$

$$114/39 \approx 3 \text{ amostras sintéticas deveriam ser coletadas ao redor de cada ponto de controle.}$$

Na Tabela 1 são apresentados os números de amostras de controle e sintéticas obtidas na BHAT.

Tabela 2. Amostras controle e sintéticas obtidas na BHAT.

UM	Amostras Controle	Amostras Sintéticas	Total
LV_PSAM	39	138	177
LVA_PSAM	22	149	171
RL_TÍPICO	13	140	153
PVA_TÍPICO	39	140	179
HIDROMÓFICOS	6	150	156
LV_TÍPICO	46	118	164
CX_TÍPICO	10	170	180
RQ_TÍPICO	153	-	153
TOTAL AMOSTRAS	328	1.005	1.333

3.4 Covariáveis Utilizadas na Predição

Todas as covariáveis foram reamostradas para a resolução espacial de 30 m e reprojetaadas para o sistema de referência de coordenadas Universal Transversa de Mercator (UTM) 22S (EPSG: 31982) com o auxílio do software ArcGIS (v. 10.5).

O MDE utilizado foi obtido na base de dados *University of Bristol (UoB), Forest And Buildings removed Copernicus DEM (FABDEM) V1- 0* (Hawker e Neal, 2021). Como forma de pré-processamento foi utilizado a ferramenta *Fill* para preencher coletores presentes ('depressões espúrias' – *sinks*) para eliminação de pequenas imperfeições. Por meio desse MDE foram geradas todas as covariáveis morfológicas e hidrométricas, através do *software* Sistema para Análises Geocientíficas Automatizadas - SAGA (v. 9.5.1), pela ferramenta *Terrain Analysis*.

As imagens de satélite foram obtidas pela plataforma Google Earth Engine (GEE), do satélite Landsat-8 OLI, coleção 2, entre os meses de junho-setembro (período de inverno), para cada ano, considerando o período entre 2013 e 2022, onde foram obtidas as bandas B1:B7. As imagens utilizadas correspondem ao período seco pois há menor cobertura de nuvens, menor interferência atmosférica e redução da vegetação herbácea, favorecendo a expressão espectral das feições edáficas.

Além das bandas espectrais, foram calculadas relações entre bandas (B4/B3 e B4/B7), e os índices *Normalized Difference Vegetation Index* - NDVI (ROUSE et al., 1974), *Normalized Difference Water Index* - NDWI (MCFEETERS, 1996), *Clay Minerals Index* (BOETTINGER et al., 2008) e *Soil Adjusted Vegetation Index* - SAVI (HUETE, 1988).

Tabela 3. Covariáveis morfométricas e hidrológicas produzidas pelo software SAGA GIS para a área da BHAT.

Morfométricas		
1. MDE	6. Plan curvature	11. Minimal curvature
2. Slope	7. Total curvature	12 Tangencial curvature
3. MRRTF	8. Sediment balance	13 Mass balance index
4. MRVBF	9. Flow line curvature	14 Longitudinal curvature
5. Texture	10. Maximal curvature	15. Modified catchment area
Hidrológicas		
1. Convexity	3. Catchment area	5. Convergence index
2. LS factor	4. Catchment slope	6. Topographic wetness index (TWI)
Canais de Drenagem		
1. Valley Depth	2. Vertical distance to channel network	

As imagens de sintéticas de radar (SAR) foram obtidas através do site *Alaska Satellite Facility* - ASF (<https://asf.alaska.edu/>). As imagens são oriundas do satélite Sentinel-1 (S1), entre o período de junho-setembro, dos anos de 2020 e 2021. O tipo de imagem foi *Single Look Complex* (SLC), de órbita descendente, com dados de polarização dupla, transmissão e recepção vertical (VV) e transmissão vertical e recepção horizontal (VH). Além das bandas unitárias foi desenvolvido o índice *Dual-polarization SAR Vegetation Index* (DPSVI) (PERIASAMY, 2018).

3.5 Seleção de Covariáveis

Primeiramente foi adotada a correlação de *Spearman*, estabelecendo um limite de significância de 95%, utilizando a função *findcorrelation* do pacote *caret* (KUHN, 2021). Nessa etapa, variáveis explicativas que possuem valores de correlação significativas maiores que $|0,95|$ foram detectadas e uma das duas variáveis foi removida, do conjunto total, conforme a maior correlação média absoluta.

3.6 Modelagem e Avaliação dos Modelos

As diferentes classes taxonômicas, e suas respectivas amostras sintéticas presentes nos ROICs, foram submetidas a uma divisão de 70% para treino e 30% validação. Foi assegurado que cada amostras, com suas respectivas amostras sintéticas ficassem no mesmo grupo, ou treino ou teste, para maior confiabilidade no poder preditivo do modelo. A separação dos subconjuntos foi realizada pela função *createDataPartition* do pacote *caret*.

O modelo RF foi executado pelo pacote *randomForest*, através do *software* R (v. 4.4.2). Os hiperparâmetros exigidos e adotados para a modelagem com o RF, *mtry* e *ntree*, foram determinados como 4 e 300, respectivamente.

O mapa gerado ao final teve seu *layout* produzido a partir do *software* ArcGIS (v. 10.5).

4 RESULTADOS E DISCUSSÃO

Na Figura 3 estão apresentados os números de amostras de controle (originais), correspondentes a cada UM, na BHAT.

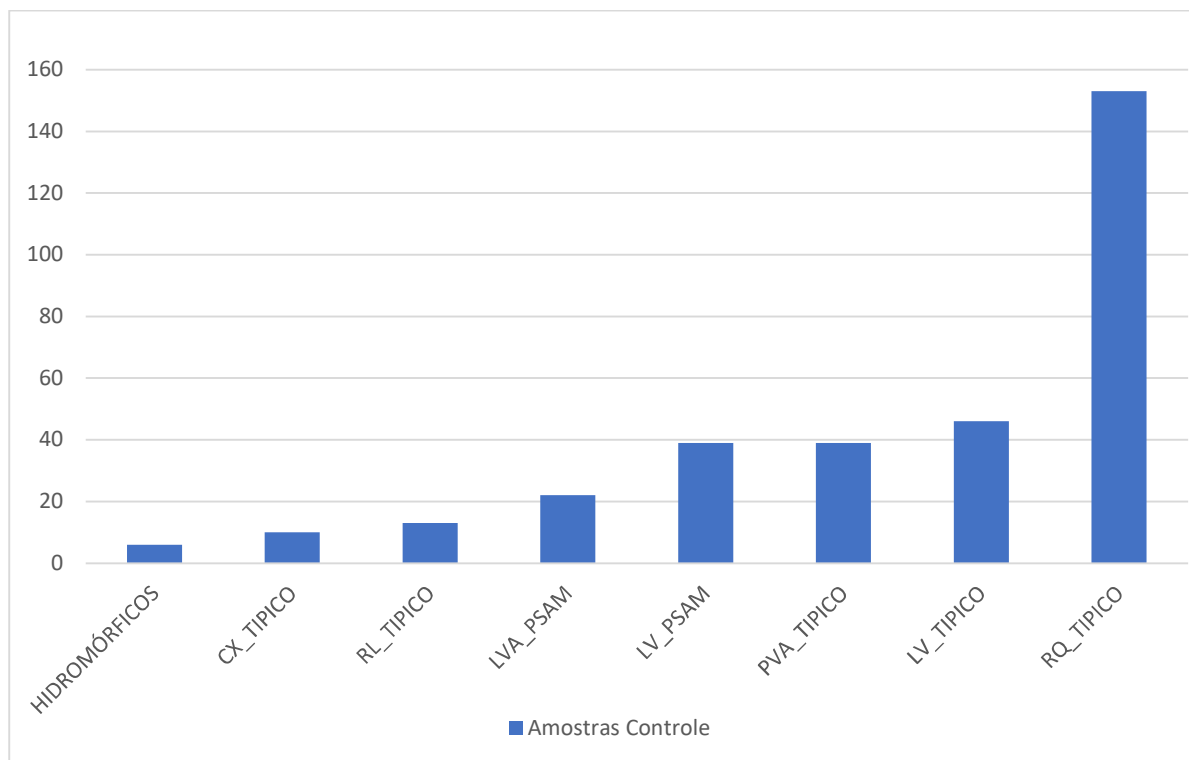


Figura 3. Amostras de controle coletadas pela BHAT.

Como pode ser observado a ordem crescente de amostras por UM é: HIDROMÓRFICOS (6); CX_TÍPICO (10); RL_TÍPICO (13); LVA_PSAM (22); PVA_TÍPICO e LV_PSAM (39); LV_TÍPICO (46) e RQ_TÍPICO (153).

A UM RQ_TÍPICO, a UM mais abundante, com predomínio da classe NEOSSOLO QUARTZARÊNICO Órtico típico, indicando solos com predomínio de textura areia ou areia franca, até no mínimo 150 cm, formados essencialmente por minerais como o quartzo ($\geq 5\%$), com preponderância das frações areia grossa e areia fina (SANTOS et al., 2018). Em sequência observamos, principalmente, presença de LATOSSOLOS VERMELHOS Distróficos típicos, ARGISSOLOS VERMELHO-AMARELO Distrófico típico, LATOSSOLO VERMELHO Distrófico psamítico e LATOSSOLO VERMELHO-AMARELO Distrófico psamítico, apresentando o primeiro, classe textura argilosa, o segundo textura arenosa/média, e o terceiro e quarto textura média.

As características indicadas pelas texturas entre média e arenosas, com exceção dos Latossolos Vermelhos, das classes encontradas está relacionada com a litologia presente na área, como indicado por Galdino e Pellegrin (2003), com predomínio do Grupo Bauru e formação Botucatu, e em menor proporção as formações Pirambóia, Palermo e Aquidauana. Segundo Lacerda Filho et al. (2006) o grupo Bauru, formado no MS, pelas formações Vale do Rio do Peixe e Marília, são constituídos essencialmente, o primeiro grupo formado por camadas tabulares de arenitos muito finos a finos, de coloração marrom, rosa e alaranjada, e o segundo

por arenitos imaturos, grossos e finos, amarelo pálido a avermelhados (*Membro Serra da Galga*), associado ao *Membro Serra da Galga*, com variações (aumento) de proporções de cimento carbonático, com calcários arenosos maciços, calcários conglomeráticos, de matriz arenosa, e calcários finos fragmentados (*Membro Ponte Alta*); e arenitos finos a médios imaturos, contendo cimento e nódulos de carbonato, com estrados tabulares maciços, bege a rosa (*Membro Echaporã*). Acompanhando as mesmas características do último membro, os arenitos presentes na formação Botucatu são finos e médios., avermelhados e friáveis.

A formação Pirambóia é constituída arenitos médios a finos, siltosos, brancos, vermelhos e arroxeados, com estratificação cruzada (ROISENBERG et al., 2008); formação Palermo apresenta arenitos cinza-arroxeados, finos a muito finos, siltitos amarelados a avermelhados, arenosos, finamente estratificados, muito silicatados; a Aquidauana é constituída por conglomerados basal seguido de arenitos vermelhos a róseos, siltitos folheados e arenitos finos, vermelhos a róseos, e arenitos vermelhos, com estratificação cruzada (LACERDA FILHO et al., 2006).

A decomposição de todo material litológico citado, predominantemente formado por arenitos, assim como o relevo com declividades pouco acentuadas, de aproximadamente 1%, fornecem material necessário para a formação de solos como Neossolos Quartzarênicos, e em menor proporção de Latossolos e Argissolos. Além das classes, a textura associada com a profundidade caracteriza os Latossolos Vermelho e Vermelho-Amarelo como psamíticos, indicando solos profundos (≥ 150 cm), são produtos da associação das tais características litogênicas e geomorfológicas (GALDINO E PELLEGRIN, 2003). Segundo Donagemma et al. (2016), os solos encontrados poderiam se enquadrar como solos leves, que apresentam classes texturais areia e areia franco-argilosa ou franco-arenosa, até 0,75 m de profundidade ou mais, assim como, sua susceptibilidade a degradação, como erosão, e seu potencial agrícola.

4.1 Avaliação da Predição de Unidades de Mapeamento (UMs) na BHAT

Após a submissão das covariáveis a análise de variância e correlação, utilizou-se o algoritmo RF do pacote *randomForest*, por meio do *software* R. Na Figura 4, estão indicadas as importâncias relativas das covariáveis (4A) e importância pela pureza dos nós das árvores contidas no modelo (4B).

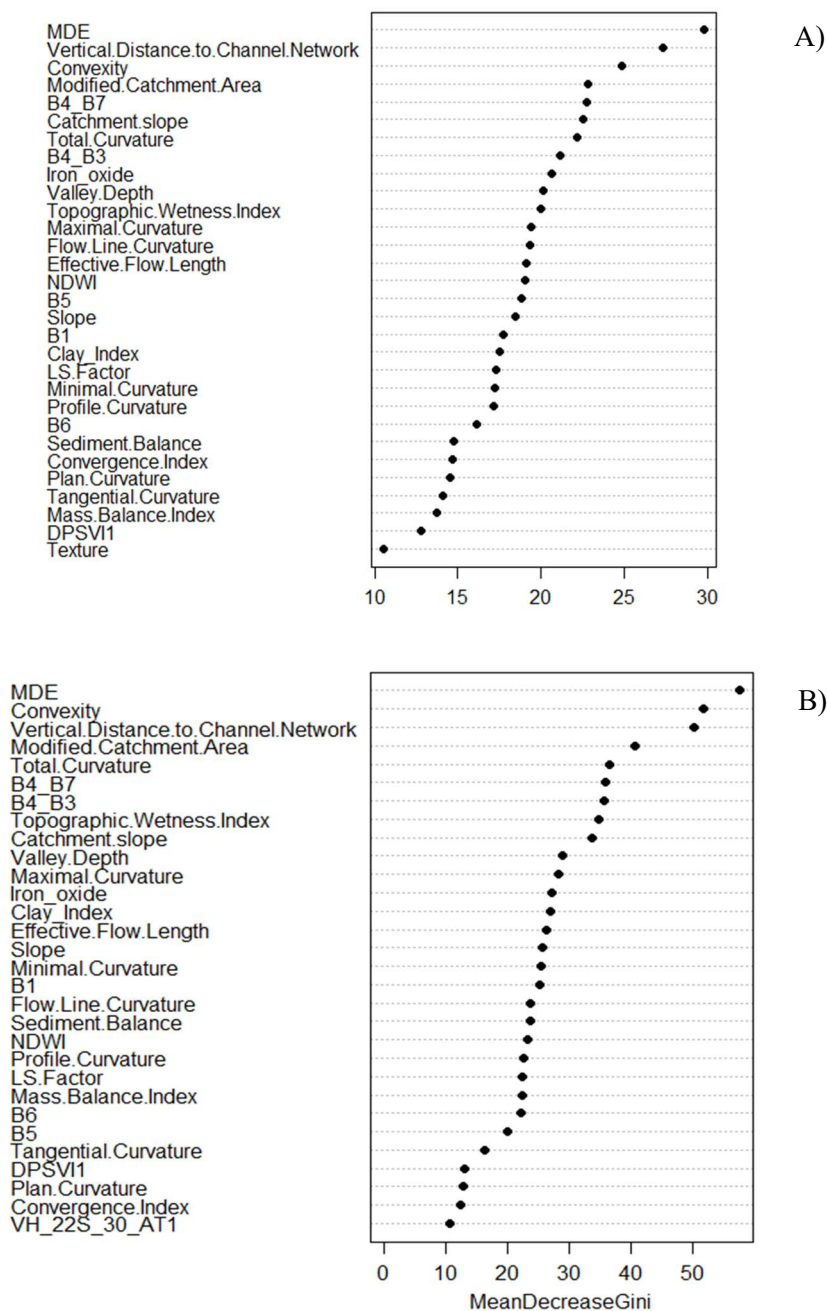


Figura 4. Graficos de importância das covariáveis de acordo com o modelo RF.
 A) Importância na Redução da Acurácia; B) Importância em relação a pureza dos nós (Índice Gini).

Como pode ser observado na Figura 4A, estão descritas de forma decrescente a importância das covariáveis que foram utilizadas pelo modelo para predição das classes de solos. Com relação as covariáveis, dentre as dez mais importantes, oito são oriundas morfológicas, obtidas a partir do MDE, e que refletem relações diretas entre os tipos de solo e características do relevo. Com destaque para o próprio MDE, que reflete a altimetria, *Vertical Distance to Channel Network* e *Convexity*, *Modified Catchment Area*, *Catchment Slope* e *Total Curvature*, que estão relacionados com o fator de formação “Relevo”, mensurado a diferença de altitude entre um ponto e um ponto mais próximo da rede de drenagem, curvatura do terreno (altimetria da vizinhança de um pixel), distribuição de água no terreno considerando depressões

ou barreiras de fluxo, e média da inclinação do terreno, respectivamente. Além das covariáveis supracitadas, as características do relevo foram as que melhor contribuíram para a performance do modelo.

Esse fato pode estar relacionado ao pressuposto das relações existentes entre solo-relevo, como indicado pelo modelo de Jenny et al. (1941) que indica que a formação do solo ocorre pela interação do clima, organismos, relevo, material de origem e tempo (*clorpt*), e atualizado por McBratney et al. (2003), que incorporou a localização espacial do solo na equação *scorpan*. Covariáveis de relevo, como o próprio MDE, e suas derivadas, são importantes para a classificação de solos, proporcionando bons resultados para o mapeamento digital de solos. Bodaghabadi et al. (2015) demonstraram que a utilização de covariáveis de terreno foram suficientes para predição satisfatória das diferentes classes de solos, a partir de formas do relevo, através de redes neurais artificiais, no Irã Central.

As covariáveis supracitadas estão intimamente relacionadas ao carregamento de partículas e distribuição de umidade que ocorre na área de estudo, que dependem diretamente da composição granulométrica dos solos. Brungard et al. (2015) que utilizou aprendizado de máquina para prever três áreas no semiárido do oeste dos Estados Unidos da América (EUA), observou que no nordeste do Wyoming, com características mais aplainadas de relevo, indicaram que covariáveis relacionadas ao carregamento de partículas e fluxo de umidade, provavelmente controlam o desenvolvimento do solo na área, visto que o local de estudo era uma planície alta e ondulada.

A refletância espectral obtida a partir do solo é uma função das peculiaridades físicas e químicas, assim como de sua estrutura interna, fornecendo novas informações que auxiliam na distribuição espacial de classes de solos (ZHANG et al., 2018). Esse fato pode ser constatado com o ranqueamento dos índices B4/B7 e B4/B3 que foram, relevantemente, mais importantes que as imagens de satélite individualizadas e demais índices. Isso pode estar relacionado com os componentes do solo que tais índices mais respondem. Segundo Saunders e Boettinger (2006) essas relações de melhoria de solo mais comumente utilizadas, que tem como finalidade acentuar radicais de ferro, em solos expostos. Chagas et al. (2016) também obteve alta correlação positiva entre areia e argila com a relação entre as bandas B4/B7, indicando afinidade entre o índice e partículas do solo. Visto a afinidade dessas bandas com componentes ferruginosos e suas relevantes presenças em solos mais intemperizados, como os Latossolos, a importância dessas covariáveis podem estar relacionadas na diferenciação dos solos mais arenosos dos mais argilosos, contribuindo assim para a determinação da classe de solo.

Com relação ao índice de pureza dos nós (Figura 4B), as covariáveis mais importantes são as mesmas observadas para importância relativa (Figura 4^a), confirmado que formas do relevo que regem umidade e carregamento de partículas, assim como, covariáveis que diferem quanto a presença de radicais ferrosos, presentes no solo, são relevantes e auxiliam na separação das amostras dentro do modelo.

Na Figura 5 estão ilustrados o erro de validação *Out-of-Bag* (BBO) em relação aos números de árvores testadas na calibração do modelo, o que justificou a adoção de 300 árvores para predição. Há estabilização do erro a partir de 200 árvores, indicando que considerar mais árvores não acrescentaria mais assertividade pelo modelo. Entretanto, de forma conservadora,

para que houvesse maior confiabilidade na predição do modelo, foi considerado um número maior de árvores (300).

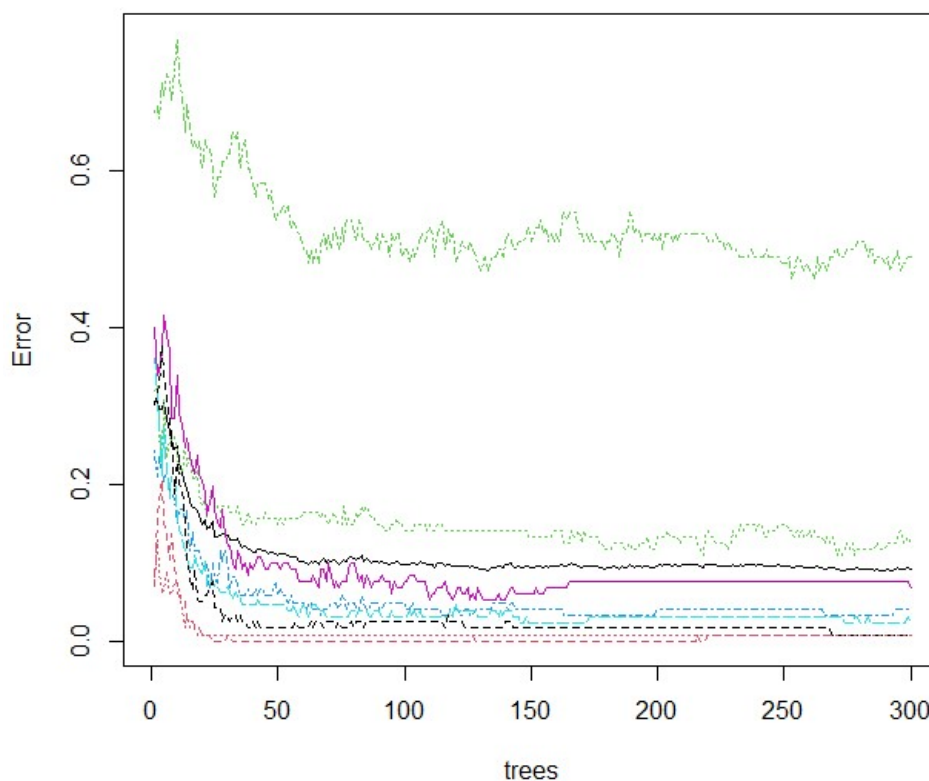


Figura 5. Erro do modelo RF considerando diferentes números de árvores.

Na Tabela 4 estão indicados os parâmetros de avaliação do modelo RF para treino e teste. Observa-se que para as amostras de treinamento há indícios de que o modelo está adequadamente calibrado, com valor Kappa de 0,90 e Acurácia Global (OA) de 90,86. Entretanto, referentes as amostras de teste, para validação do modelo, observam-se que não houve acertos pelo modelo de forma satisfatória, fornecendo um Kappa e OA iguais a 0,29 e 37,54, respectivamente. Segundo Landis e Koch (1977) o valor Kappa significa uma concordância razoável entre os dados originais e os preditos. Esses resultados são insatisfatórios e contrapõem outros trabalhos, quando comparado com a predição de classes de solos realizada pelo mesmo modelo (SHUKLA et al., 2018; DORNIK; DRĂGUȚ; URDEA, 2018)

Tabela 4. Parâmetros de validação do modelo *Random Forest* (RF) para os subconjuntos de treino e teste das amostras de controle e sintéticas da BHAT.

Treino			Teste		
Kappa	Accuracy	OA ¹	Kappa	Accuracy	OA
0,9	0,91	90,86	0,29	0,38	37,54

¹Acurácia Global (OA)

Mesmo apresentando baixa precisão na predição das classes de solo da BHAT, o RF tem sido um dos algoritmos mais frequentemente utilizados, e que tem apresentado bons resultados

para classificação de solos (CAHYANA et al., 2023). Como observado por Jeune et al. (2018) que alcançou valores Kappa de 0,55, indicando maior relevante capacidade de generalização do modelo para solos no oeste do Haiti. Da mesma forma Shukla et al. (2018) que comparando o RF com modelos *Classification and Regression Trees* (CART) e CART ensemble bagger (CEB), obteve melhor predição do RF, e maior resistência ao ruído, mostrando alta estabilidade contra a redução do conjunto de dados. Tais resultados se apresentam em contraponto ao encontrado no trabalho, mas considerando a ampla confiabilidade do modelo, outros aspectos preditivos podem ser colocados a prova, como o número de covariáveis e ajuste dos parâmetros do modelo.

Segundo Bodaghabadi et al. (2015), mesmo os atributos do solo, obtidos a partir do MDE, foram suficientes para identificar classes de solos, os mesmos não foram suficientes para aplicar o modelo *machine learning* para predição de áreas não amostradas, indicando que outros fatores de formação devam ser levados em consideração na modelagem, como mais covariáveis geomorfológicas, litologia e/ou climáticas. Dessa forma, vale considerar outras covariáveis para treinamento do modelo, assim como, calibração dos hiperparâmetros dos modelos utilizados e aprimorar a coleta de amostras sintéticas para a predição de classes.

A Tabela 5 ilustra a matriz de confusão obtida a partir das amostras de treino para classificação de solos pelo modelo RF.



Tabela 5. Matriz de confusão das amostras de teste.

Classes	CX	LVP	LVT	LVAP	PVAT	RLT	HIDRO	RQT	Total	Usuário (%)
CX	4	1	0	1	0	3	0	1	10	40
LVP	0	24	2	0	4	0	0	6	36	66,67
LVT	9	1	30	4	1	0	0	1	46	65,22
LVAP	0	5	5	6	1	0	0	6	23	26,09
PVAT	11	0	2	5	18	8	6	7	57	31,58
RLT	22	0	0	7	11	20	0	3	63	31,74
HIDRO	1	0	0	0	5	0	8	0	14	57,14
RQT	4	18	4	21	8	3	21	21	100	21
Total	51	49	43	44	48	34	35	45	349	
Produtor (%)	7,84	48,98	69,77	13,64	37,50	58,82	22,86	46,47		

UMs:

- **CX_TÍPICO (CX):** Cambissolo Háplico;
- **LV_PSAM (LVP):** Latossolo Vermelho psamítico;
- **LV_TÍPICO (LVT):** Latossolo Vermelho típico;
- **LVA_PSAM (LVAP):** Latossolo Vermelho-Amarelo psamítico;
- **PVA_TÍPICO (PVAT):** Argissolo Vermelho-Amarelo típico;
- **RL_TÍPICO (RLT):** Neossolo Litólico típico;

- **HIDROMÓRFICOS (HIDRO):** Neossolo Quartzarênico Hidromórfico, Gleissolo Háplico, Planossolo Nátrico;
- **RQ_TÍPICO (RQT):** Neossolo Quartzarênico típico;



Observa-se que os maiores erros (Tabela 5) estão nas alocações das classes CX_TÍPICO, LV_PSAM, LVA_PSAM. Os Latossolos Vermelhos e Vermelho-Amarelos, ambos psamíticos, podem estar presentes em áreas muito parecidas com os Neossolos Quartzarênicos, visto que esses últimos, por mais que apresentem as características para serem classificadas como tal, apresentam relevante profundidade (≥ 100 cm), podendo estar situados em localidades na paisagem parecidas com áreas de ocorrência dos Latossolos psamíticos. Essa justificativa vale também para a confusão entre Cambissolos Háplicos e Neossolos Quartzarênicos. Dessa forma, a adição de covariáveis, assim como, a reformulação das UM e nova coleta de amostras sintéticas podem auxiliar a predição das classes da BHAT.

4.2 Mapa Preditivo de UMs Obtido pelo Modelo *Random Forest*

Na Figura 6 está apresentado o mapa com as UMs na BHAT predito pelo modelo Random Forest.

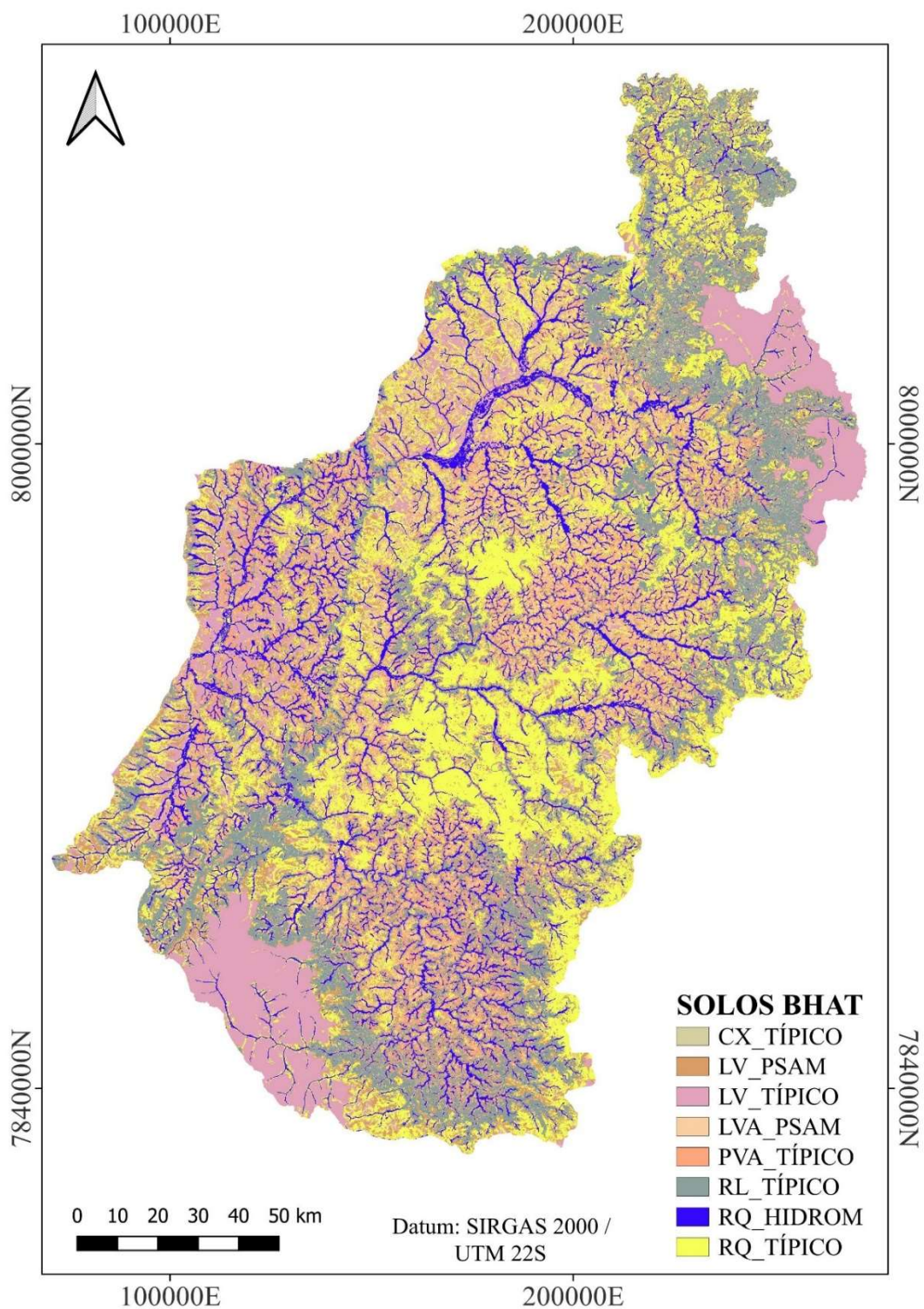


Figura 6. Mapa com as predição das Unidades de Mapeamento (UMs) da Bacia Hidrográfica do Alto Taquari (BHAT).

Através do mapa obtido pela generalização do modelo RF, podemos observar predomínio das UMs RL_TÍPICO e RQ_TÍPICO, que destaca a ocorrência de Neossolos Litólicos e Quartzarênicos, respectivamente. Os resultados obtidos diferem ligeiramente dos observados por Galdino e Pellegrin (2003), visto que constataram predomínio dos Neossolos Quartzarênicos. A melhor predição desses solos pode estar relacionada a ausência de amostras sintéticas, visto que a UM dos Neossolos Quartzarênicos apresentou maior número de amostras coletadas em campo, demonstrando predomínio de solos mais jovens e arenosos na área. Esse resultado indica que a obtenção de amostras sintéticas pode ser aprimorada, possivelmente ocasionando alocação mais precisa de amostras minoritárias e menor confusão do modelo entre as UMs.

A UM LV_TÍPICO, que representa os Latossolos Vermelhos típicos, foram encontradas de forma mais isolada na porção extrema a nordeste e sudoeste da área. Os RQ_HIDROM, com presença dos Neossolos Quartzarênicos Hidromórficos, Gleissolo Háptico e Planossolo, foram encontrados próximos aos canais de drenagem, ou seja, condizente com as áreas de ocorrência desses solos, próximos aos PVA_TÍPICO (Argissolos Vermelho-Amarelos), que ocupam predominantemente os tercos inferiores de encontra em transição para planícies.

A ocorrência das classes de solos CX_TÍPICOS, LV_PSAM e LVA_PSAM, não ficaram claramente distribuídas na área, ou não foram generalizadas de forma condizente no mapa. Essa falta de precisão, pode estar relacionada também com a coleta de amostras sintéticas, demonstrando que pode ser melhor aplicada a técnica, assim como, a adição de covariáveis ao modelo que possam auxiliar na obtenção de padrões que poderão ser utilizados pelo modelo, e consequentemente, melhor predizer as UMs.

No geral, foi obtido um mapa com a distribuição espacial das UMs da BHAT, que por meio dos parâmetros de avaliação, pôde ser constatado que melhorias como a adição de covariáveis que possam representar melhor os fatores de formação dos solos, além de, melhor aquisição de amostras sintéticas e ajustes dos hiperparâmetros podem ser realizados, a fim de que seja produzido um mapa. Entretanto, a partir do mapa produzido, observa-se que os solos da área apresentam fragilidades que os tornam propensos a processos erosivos, o que justifica a adoção de uso racional e sustentável da área.

5 CONCLUSÕES

O mapeamento de classes taxonômicas da BHAT, utilizando o método ROIC para balanceamento amostral, e o modelo RF, sendo este um modelo robusto e amplamente utilizado na ciência, não atingiu os resultados esperados para predição das classes de solos da BHAT, apresentando OA e Kappa iguais a 37,24 e 0,29, respectivamente. Esse resultado, aponta medidas que podem ser adotadas para melhorar a predição, como a adição e escolha de novas covariáveis, que representem melhor os fatores de formação do solo, assim como, a melhor seleção de amostras sintéticas para equilibrar as UM e/ou reformular as UM para melhor interpretação do RF e de outros modelos.

Para a predição das UMs o modelo RF detectou que as covariáveis mais importantes foram relacionadas a características do relevo, como *Vertical Distance to Channel Network* e *Convexit*, *Modified Catchment Area*, *Catchment Slope* e *Total Curvature*, que influenciam o arraste de partículas e fluxo de umidade. Além delas, os índices espectrais ópticos B4/B7 e B4/B3, que se relacionam com acentuar radicais de ferro, em solos expostos. Entretanto, pôde ser observado que, para as UMs que mais possuam amostras sintéticas, ocorreram confusões em suas predições espaciais, explicitando as limitações dessas covariáveis para a alocação das ROICs, e consequentemente, obtenção de amostras sintéticas.

Considerando os dados até aqui alcançados, pode-se observar que as UM que continham dados com Neossolos Litólicos e Quartzarênicos, se mostraram predominantes na área, indicando a fragilidade dos solos da bacia, reforçando que manejos mais sustentáveis devem ser adotados para a área, que vem ao longo dos anos sofrendo com problemas relacionados a erosão do solo.

6 REFERÊNCIAS BIBLIOGRÁFICAS

- BODAGHABADI, M. B; MARTÍNEZ-CASASNOVAS, J. A; SALEHI, M. H; MOHAMMADI, J; BORUJENI, I. E; TOOMANIAN, N; GANDOMKAR, A. **Digital soil mapping using artificial neural networks and terrain-related attributes**. *Pedosphere*, v. 25, n. 4, p. 580-591, 2015.
- BOETTINGER, J. L; RAMSEY, R. D; BODILY, J. M; COLE, N. J; KIENAST-BROWN, S; NIELD, S. J; SAUNDERS, A. M; STUM, A. K. **Landsat spectral data for digital soil mapping**. In: *Digital soil mapping with limited data*. Dordrecht: Springer Netherlands, p. 193-202, 2008.
- BOHN, M P.; MILLER, B. **Locally enhanced digital soil mapping in support of a bottom-up approach is more accurate than conventional soil mapping and top-down digital soil mapping**. *Geoderma*, v. 442, p. 116781, 2024.
- BRASIL. PROJETO RADAMBRASIL. Programa de Integração Nacional. **Levantamento dos recursos potenciais naturais da terra**. v. 31, Folha SE. 22, 1983.
- BREIMAN, L. **Random forests**. *Machine learning*. v. 45, p. 5-32, 2001.
- BRUNGARD, C. W; BORTTINGER, J. L; DUNIWAY, M. C; WILLS, S. A; EDWARDS Jr, T. C. **Machine learning for predicting soil classes in three semi-arid landscapes**. *Geoderma*, v. 239, p. 68-83, 2015.
- CAHYANA, D; SULAEMAN, Y; BARUS, B; DARMAWAN; MULYANTO, B. **Improving digital soil mapping in Bogor, Indonesia using parent material information**. *Geoderma Regional*, v. 33, p. e00627, 2023.
- CARVALHO JÚNIOR, W; PEREIRA, N. R; FERNANDES FILHO, E. I; CALDERANO FILHO, B; PINHEIRO, H. S. K; CHAGAS, C. S; BHERING, S. B; PEREIRA, V. R; LAWALL, S. **Sample design effects on soil unit prediction with machine: randomness, uncertainty, and majority map**. *Revista Brasileira de Ciência do Solo*, v. 44, p. e0190120, 2020.
- CHAGAS, C. S; CARVALHO JUNIOR, W; BHERING, S. B; CALDERANO FILHO, B. **Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions**. *Catena*, v. 139, p. 232-240, 2016.
- CURADO, F. F. **Considerações sócio-econômicas e ambientais relacionadas ao "arrombados" na planície do rio Taquari - MS**. 2004.
- DHARUMARAJAN, S; HEGDE, R. **Digital mapping of soil texture classes using Random Forest classification algorithm**. *Soil Use and Management*, v. 38, n. 1, p. 135-149, 2020.
- DONAGEMMA, G. K; FREITAS, P. L; BALIEIRO, F. C; FONTANA, A; SPERA, S. T; LUMBRERAS, J. F; VIANA, J. H. M; ARAÚJO FILHO, J. C; SANTOS, F. C; ALBUQUERQUE, M. R; MAMCEDO, M. C.M; TEIXEIRA, P. C; AMARAL, A. J; BORTOLON, E; BORTOLON, L. **Characterization, agricultural potential, and perspectives for the management of light soils in Brazil**. *Pesquisa Agropecuária Brasileira*, v. 51, n. 9, p. 1003-1020, 2016.
- DORNIK, A; DRĂGUT, L; URDEA, P. **Classification of soil types using geographic object-based image analysis and random forests**. *Pedosphere*, v. 28, n. 6, p. 913-925, 2018.

- EMADI, M; TAGHIZADEH-MEHRJARDI, R; CHERATI, A; DANESH, M; MOSAVI, A; SCHOLTEN, T. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing*, v. 12, n. 14, p. 2234, 2020.
- FERREIRA, M. R; AQUINO, L. B. C. **O Transporte de Sedimentos nas Encostas e a Dinâmica Hidrológica do Alto Taquari**. PANTANEIRA, p. 7, 2007.
- FRANCOS, N; KARASIK, E; MYERS, M; BEM-DOR, E. **Soil type classification using Landsat 8: A comparison between the USDA and a local system in Israel**. *International Soil and Water Conservation Research*, 2025.
- GALAR, M; FERNANDEZ, A; BARRENECHEA, E; BUSTINCE, H; HERRERA, F. **A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches**. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 42, n. 4, p. 463-484, 2012.
- GALDINO, S; PELLEGRIN, L. A. **Diagnóstico e diretrizes do PCBAP relativos à erosão na Bacia do Alto Taquari**. EMBRAPA. 2003.
- GALDINO, S; VIEIRA, L. M; PELLEGRIN, L. A. **Impactos ambientais e socioeconômicos na Bacia do Rio Taquari-Pantanal**. Corumbá: Embrapa Pantanal, 2005.
- GRUNWALD, S. **Multi-criteria characterization of recent digital soil mapping and modeling approaches**. *Geoderma*, 152(3-4), 195-207, 2009.
- HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. **The elements of statistical learning**. 2009.
- HAWKER, L; NEAL, J. **FABDEM - Forest And Buildings removed Copernicus DEM, V1-0**. University of Bristol, 2021.
- HEUNG, B; HO, H. C; ZHANG, J; KNUDBY, A; BULMER, C. E; SCHMIDT, M. G. **An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping**. *Geoderma*, v. 265, p. 62-77, 2016.
- HUETE, A. R. **A soil-adjusted vegetation index (SAVI)**. *Remote sensing of environment*, v. 25, n. 3, p. 295-309, 1988.
- IBGE – Instituto Brasileiro de Geografia e Estatística. **Manual Técnico de Pedologia: Guia Prático de Campo**. Rio de Janeiro, 133 p, 2015.
- JENNY, H. **Factors of Soil Formation: A System of Quantitative Pedology**. Nova Iorque: McGraw-Hill, 1941.
- JEUNE, W; FRANCELINO, M. R; SOUZA, E; FERNANDES FILHO, E. I; ROCHA, G. C. **Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti**. *Revista Brasileira de Ciência do Solo*, v. 42, p. e0170133, 2018.
- KASRAEI, B; SCHNIDT, M. G; ZHANG, J; BULMER, C. E; FILATOW, D. F; ARBOR, A; PENNEL, T; HEUNG, B. **A framework for optimizing environmental covariates to support model interpretability in digital soil mapping**. *Geoderma*, v. 445, p. 116873, 2024.
- KHALEDIAN, Y; MILLER, B. A. **Selecting appropriate machine learning methods for digital soil mapping**. *Applied Mathematical Modelling*, v. 81, p. 401-418, 2020.
- KUHN, M. **Caret: classification and regression training**. R package version, v. 3, 2021.
- KUMARAPERUMAL, R; PAZHANIVELAN, S; GEETHALAKSMI, V; RAJ, M. N; MUTHUMANICKAM, D; KALIAPERUMA, R; SHANKAR, V; NAIR, A. M; YADAV, M.



K; KSHATRIYA, T. V. T. **Comparison of machine learning-based prediction of qualitative and quantitative digital soil-mapping approaches for Eastern Districts of Tamil Nadu, India.** Land, v. 11, n. 12, p. 2279, 2022.

LACERDA FILHO, J V; BRITO, R. S. C; SILVA, M. G; OLIVEIRA, C. C.; MORETON, L. C; MARTINS, E. G; LOPES, R. C; LIMA, T, M; LARIZZATTI, J. H; VALENTE, C. R. **Geologia e recursos minerais do estado de Mato Grosso do Sul.** CPRM; SEPROTUR/MS; EGRHP/MS, 2006.

LAGACHERIE, P; MCBRATNEY, A. B. **Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping.** Developments in soil science, 31, 3-22, 2006.

LANDIS, J. R; KOCH, G. G. **The measurement of observer agreement for categorical data.** biometrics, p. 159-174, 1977.

MA, Y; MINASNY, B; MALONE, B. P; McBRATNEY, A. B. **Pedology and digital soil mapping (DSM).** European Journal of Soil Science, v. 70, n. 2, p. 216-235, 2019.

MANTEGHI, S; MOURAVEJ, K; MOUSAVI, S. R; DELAVAR, M. A; MASTINU, A. **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, 2024.

MCBRATNEY, A. B; SANTOS, M. M; MINASNY, B. **On digital soil mapping.** Geoderma, 117(1-2), 3-52, 2003.

McFEETERS, S. K. **The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features.** International journal of remote sensing, v. 17, n. 7, p. 1425-1432, 1996.

MEIER, M; SOUZA, E; FRANCELINO, M. R; FERNANDES FILHO, E. I; SCHAEFER, C. E. G. R. **Digital soil mapping using machine learning algorithms in a tropical mountainous area.** Revista Brasileira de Ciência do Solo, v. 42, p. e0170421, 2018.

MINASNY, B; BANDAI, T; GHEZZEHEI, T. A; HUANG, Y-C; MA, Y; McBRATNEY, A. B; NG, W; NOROUZI, S; PADARIAN, J; RUDIYANTO; SHARIFIFAR, A; STYC, Q; WIDYASTUTI, M. **Soil Science-Informed Machine Learning.** Geoderma, v. 452, p. 117094, 2024.

NEYESTANI, M; SARMANDIAN, F; JAFARI, A; KESHAVARZI, A; SHARIFIFAR, A. **Digital mapping of soil classes using spatial extrapolation with imbalanced data.** Geoderma Regional, v. 26, p. e00422, 2021.

NGU, N. H; THANH, N. N; DUC, T. T; NON, D. Q; NA, N. T. T; CHOTPANTARAT, S. **Active learning-based random forest algorithm used for soil texture classification mapping in Central Vietnam.** Catena, v. 234, p. 107629, 2024.

PADARIAN, J; MINASNY, B; MCBRATNEY, A. B. **Machine learning and soil sciences: A review aided by machine learning tools.** Soil, v. 6, n. 1, p. 35-52, 2020.

PERIASAMY, S. **Significance of dual polarimetric synthetic aperture radar in biomass retrieval: An attempt on Sentinel-1.** Remote sensing of environment, v. 217, p. 537-549, 2018.

RABELO, A. P. C; SOUZA, M. G. **Bacia do Alto Paraguai: uma viagem no tempo.** 2021.

ROISENBERG, A; MARQUES, J. C; BACHI, F. A; FRANTZ, J. C. **Geologia da folha Gravataí SH. 22-XCV**. CPRM, 2008.

ROUSE, J. W; HAAS, R. H; SCHELL, J. A; DEERING, D. W. **Monitoring vegetation systems in the Great Plains with ERTS**. NASA Spec. Publ, v. 351, n. 1, p. 309, 1974.

SANTOS, H. G; JACOMINE, P. K. T; ANJOS, L. H. C; OLIVEIRA, V. A; LUMBRERAS, J. F; COELHO, M. R; ALMEIDA, J. A; ARAUJO FILHO, J. C; OLIVEIRA, J. B; CUNHA, T. J. F. **Sistema Brasileiro de Classificação de Solos**. Brasília, DF: Embrapa, 2018.

SAUNDERS, A. M.; BOETTINGER, J. L. **Incorporating classification trees into a pedogenic understanding raster classification methodology, Green River Basin, Wyoming, USA**. Developments in Soil Science, v. 31, p. 389-620, 2007.

SHARIFIFAR, A; SARMADIAN, F; MALONE, B. P; MINASNY, B. **Addressing the issue of digital mapping of soil classes with imbalanced class observations**. Geoderma, v. 350, p. 84-92, 2019.

SHUKLA, G; GARG, R. D; SRIVASTAVA, H. S; GARG, P. K. **An effective implementation and assessment of a random forest classifier as a soil spatial predictive model. International. Journal of Remote Sensing**, v. 39, n. 8, p. 2637-2669, 2018.

SIERS, M J.; ISLAM, M. Z. **Novel algorithms for cost-sensitive classification and knowledge discovery in class imbalanced datasets with an application to NASA software defects**. Information Sciences, v. 459, p. 53-70, 2018.

SILVA, G; MEDEIROS, S; CARVALHO, E. M. **Impactos Ambientais Ocasionados pela Movimentação de Sedimentos na Bacia do Rio Taquari**. Revista Pantaneira, 2007.

SULEYMANOV, A; ARROUAYS, D; SAVIN, I. **Digital soil mapping in the Russian Federation: A review**. Geoderma Regional, v. 36, p. e00763, 2024.

TAGHIZADEH-MEHRJARDI, R; MINASNY, B; TOOMANIAN, N; ZERAATPISHEH, M; AMIRIAN-CHAKAN, A; TRIANTAFILIS, J. **Digital mapping of soil classes using ensemble of models in Isfahan region, Iran**. Soil Systems, v. 3, n. 2, p. 37, 2019.

VAN DER WESTHUIZEN, S; HEUVELINK, G. B. M; GARDNER-LUBBE, S; CLARKE, C. E. **Biplots for understanding machine learning predictions in digital soil mapping**. Ecological Informatics, v. 84, p. 102892, 2024.

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

WADOUX, A. M. J-C; SAMUEL-ROSA, A/ POGGIO, L; MULDER, V. L. **A note on knowledge discovery and machine learning in digital soil mapping**. European Journal of Soil Science, v. 71, n. 2, p. 133-136, 2019.

WANG, L; WANG, X; KOOCH, Y; SONG, K; WU, D. **Improvement of data imbalance for digital soil class mapping in Eastern China**. Computers and Electronics in Agriculture, v. 214, p. 108322, 2023.

WANG, X; ZHANG, X; LI, H; ZHANG, X; LIU, H; DOU, X; YU, Z. **The minimum level for soil allocation using topsoil reflectance spectra: Genus or species?** Catena, v. 174, p. 36-47, 2019.



WEINDORF, D. C; CHAKRABORTY, S. **Balancing machine learning and artificial intelligence in soil science with human perspective and experience.** *Pedosphere*, v. 34, n. 1, p. 9-12, 2024.

YANG, L; JIAO, Y; FAHMY, S; ZHU, A. X; HANN, S; BURT, J. E; QI, F. **Updating conventional soil maps through digital soil mapping.** *Soil Science Society of America Journal*, 75(3), 1044-1053, 2011.

ZERAATPISHEH, M; AYOUBI, S; JAFARI, A; FINKE, P. **Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran.** *Geomorphology*, v. 285, p. 186-204, 2017.

ZERAATPISHEH, M; JAFARI, A; BODAGHABADI, M. B; AYOUBI, S; TAGHIZADEH-MEHRJARDI, R; TOOMANIAN, N; KERRY, R; XU, M. **Conventional and digital soil mapping in Iran: Past, present, and future.** *Catena*, v. 188, p. 104424, 2020.

ZHANG, L; YANG, L; MA, T; SHEN, F; CAI, Y; ZHOU, C. **A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data.** *Geoderma*, v. 384, p. 114809, 2021.

ZHANG, X; LIU, H; ZHANG, X; YU, S; DOU, X; XIE, Y; WANG, N. **Allocate soil individuals to soil classes with topsoil spectral characteristics and decision trees.** *Geoderma*, v. 320, p. 12-22, 2018.

SALMAN, H. A; KALAKECH, A; STEITI, A. **Random forest algorithm overview.** *Babylonian Journal of Machine Learning*, v. 2024, p. 69-79, 2024.



ATA Nº 3593/2025 - CEAD (12.28.01.36)

Nº do Protocolo: 23083.040225/2025-98

Seropédica-RJ, 23 de julho de 2025.

ATA DE DEFESA

Aos quinze dias do mês de julho do ano de dois mil e vinte e cinco, às dezesseis horas, através de web conferência, instalou-se a banca examinadora de Trabalho de Conclusão do Curso de Especialização em Geoprocessamento, Levantamento e Interpretação de Solos, da cursista Gabriela da Rocha Saldanha sob a orientação da professora Helena Saraiva Koenow Pinheiro. A banca examinadora foi composta pelos pesquisadores Leonardo Durval Duarte Guimarães e Ana Carolina de Souza Ferreira. A defesa do TCC intitulado "Mapeamento de Classes de Solos da Bacia Hidrográfica do Alto Taquari (MS/MT)", foi iniciada às dezesseis horas e teve a duração de vinte minutos de apresentação seguida da avaliação pela banca. A cursista, após avaliada pela banca examinadora obteve o resultado:

(X) APROVADA, devendo a cursista proceder a eventual revisão solicitada pela supervisora e /ou pela banca, e entregar a versão final em até 15 dias à coordenação do Curso.

() NÃO APROVADA.

Seropédica, 15 de julho de 2025.

Leonardo Durval Duarte Guimarães
Primeiro Examinador

Ana Carolina de Souza Ferreira
Segunda Examinadora

Helena Saraiva Koenow Pinheiro
Presidente

Gabriela da Rocha Saldanha
Cursista

(Assinado digitalmente em 23/07/2025 12:25)
HELENA SARAIVA KOENOW PINHEIRO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptoS (12.28.01.00.00.33)
Matricula: ###236#8

(Assinado digitalmente em 23/07/2025 15:35)
LEONARDO DURVAL DUARTE GUIMARAES
PROFESSOR FMS BASICO TECN TECNOLÓGICO
DeptoS (12.28.01.00.00.33)
Matricula: ###205#1

(Assinado digitalmente em 24/07/2025 15:05)
ANA CAROLINA DE SOUZA FERREIRA
ASSINANTE EXTERNO
CPF: ###.###.607-##

(Assinado digitalmente em 27/07/2025 19:47)
GABRIELA DA ROCHA SALDANHA
DISCENTE
Matricula: 2024#####4